SISTEMA DE RECONOCIMIENTO DE VOZ BASADO EN UN MÉTODO DE APRENDIZAJE SUPERVISADO Y LA CORRELACIÓN DE PEARSON

K-NN ALGORITHM AND PEARSON CORRELATION-BASED A VOICE RECOGNITION SYSTEM

Anel Ramírez Álvarez
Benemérita Universidad Autónoma de Puebla, México
anel.ramirez.al@gmail.com

Luz A. Sánchez Gálvez
Benemérita Universidad Autónoma de Puebla, México
sanchez.galvez@correo.buap.mx

Mario Anzures García
Benemérita Universidad Autónoma de Puebla, México
mario.anzures@correo.buap.mx

Sully Sánchez Gálvez
Benemérita Universidad Autónoma de Puebla, México
ssanchez@cs.buap.mx

Mariano Larios Gómez
Benemérita Universidad Autónoma de Puebla, México
mlarios77@gmail.com

Recepción: 30/octubre/2020 **Aceptación:** 10/diciembre/2020

Resumen

El reconocimiento automático de voz es una disciplina de la inteligencia artificial, que tiene como objetivo permitir la comunicación hablada entre seres humanos y computadoras. Este artículo propone un sistema de reconocimiento de voz, basado en la extracción de características distintivas de la voz y el método de aprendizaje supervisado, denominado algoritmo k-NN (*k-Nearest Neighbors*), que requiere del entrenamiento del sistema. Así como se plantea calcular K automáticamente por medio de la correlación de Pearson, para que el sistema de reconocimiento de voz sea más del algoritmo k-NN. Finalmente, se evalúa el sistema con voces de personajes conocidos para centrarse en la eficiencia del sistema.

Palabras Claves: Algoritmo K-vecinos más cercanos, correlación de Pearson, entrenamiento, extracción de características, sistema de reconocimiento de voz.

Abstract

Automatic speech recognition or automatic voice recognition is a discipline of artificial intelligence, which aims to allow spoken communication between humans and computers. This paper proposes a speech recognition system, based on the extraction of distinctive characteristics of the voice, and the k-NN (k-Nearest Neighbors) algorithm, which requires training of the system. As well as, it, presents the calculation of K through Pearson's correlation, in this way k will not be fixed, and the speech recognition will be most efficient. Finally, the system is evaluated; by using known characters for it focuses on the efficiency of such system.

***Keywords**: Feature extraction, k-Nearest neighbors Pearson correlation, training, voice recognition system*.

1. Introducción

La biometría realiza estudios de reconocimiento de humanos basados en rasgos conductuales o físicos intrínsecos y particulares: identificando y verificando la anatomía o el comportamiento de una persona [Ortega, 2013]: 1) *Biometría estática*, que define algo que el usuario es, un rasgo físico o anatómico, característico y ´único en cada ser humano; 2) *Biometría dinámica*, que se refiere a la conducta o comportamiento, algo que el humano hace, como su escritura o su propia voz. Tomando como base la biometría es posible llevar a cabo estudios de reconocimiento de humanos fundamentados en rasgos conductuales o físicos intrínsecos y particulares de cada persona; permitiendo autenticar individuos, mediante [About, 2011]:

- La identificación. Dice quién es una persona, dependiendo de sus características físicas o de su conducta.
- La verificación. Aclara si una persona es quien dice ser, partiendo de análisis biométricos y realizando comparaciones con otros candidatos.

La biometría tiene dos ámbitos de aplicación: la salud y la seguridad. Hasta ahora estos sistemas se separan en dos grandes módulos [About, 2011]: 1) Algo que el usuario sabe; y 2) Algo que el usuario tiene.

Este trabajo de investigación, se centra en la biometría dinámica, que es algo que un ser humano hace. Además, se considera que en el desarrollo del reconocimiento automático del habla intervienen diversas disciplinas, tales como: la fisiología, la acústica, la lingüística, el procesamiento de señales, la inteligencia artificial y la ciencia de la computación [Tordera, 2011].

Los trabajos de reconocimiento de voz datan de la mitad del siglo XX: [Furui, 1972] que buscaba las características en parámetros estadísticos o predictivos, matrices de covarianza, histogramas de frecuencia, etc.; [Rabiner, 1993] que diseño métodos elementales de normalización de tiempo; [Matsui, 1993] que propone un sistema que eliminaba enunciados con texto diferente al estudiado y reconocía al locutor con un vocabulario limitado, y [Juang, 1998] que utilizó métodos de programación dinámica para conseguir el alineamiento temporal de los pares de realizaciones de habla. En los últimos años, los trabajos se centran, en la identificación y verificación del locutor. En el segundo caso y con respecto a la clasificación, se usan métodos de aprendizaje supervisado como el algoritmo de k-vecinos más cercanos (k-NN, *k-Nearest Neighbors*), dando buenos resultados mediante el establecimiento de un k fijo [Arias, 2018]; sin embargo, se recomienda variar el valor de K algo que se hace de forma manual para mejorar la eficiencia del sistema.

Por tanto, se presenta un sistema de reconocimiento de voz que determina quién es el personaje que habla; considerando las voces de cinco personajes de “Los Simpson” porque la idea es probar la eficiencia del sistema propuesto tomando k de manera automática y utilizando el procedimiento de reconocimiento de voz, que consta de [Big, 2017]: adquisición de voz, espectrograma y gráfica en el tiempo, extracción de características, entrenamiento y verificación. También se plantea utilizar la correlación Pearson para determinar el valor de k automáticamente en el algoritmo de para mejorar la eficiencia de dicho sistema.

Este artículo se encuentra organizado de la siguiente manera: Sección 2 explica el método que fundamenta el sistema de reconocimiento de voz. Sección 3 describe

los resultados del sistema de reconocimiento de voz. Sección 4 presenta la discusión. Finalmente, Sección 5 detalla las conclusiones y el trabajo futuro.

2. Método

Este sistema de reconocimiento de voz comienza con la adquisición de los audios a procesar por el sistema y se entrena con las muestras de audio a reconocer. En la segunda etapa (figura 1), comienza el algoritmo K-NN, primero adquiriendo la señal con el uso de la librería scipy, después normalizando y preparando la señal para la extracción de parámetros en tiempo y frecuencia. Para el tiempo se calcula la Energía y Raíz Media Cuadrática (RMS) de las muestras; y en la Frecuencia el Centroide Espectral y el Mel-Frequency Cepstral Coeficients (MFCC). Esto se realiza para entrenar el sistema a partir de la agrupación de lso sonidos por similitud en base a descriptores o características particulares de cada personaje, clasificándolos en grupos (Figura 2). Se aplican evaluaciones y predicciones [Big, 2017] con la incorporación de técnicas de Inteligencia Artificial (IA) y de análisis de metadatos obtenidos para las muestras de la base de datos.

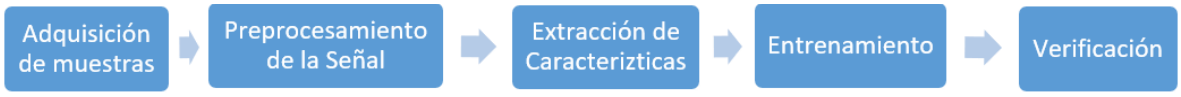

Figura 1 Sistema de reconocimiento de voz.

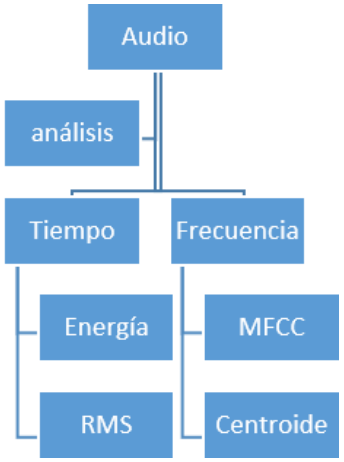

Figura 2 Tipos de análisis para la señal de audio.

Luego se realiza la clasificación, al juntar todas las clases o grupos de características de cada uno de los 5 personajes. Estas clasificaciones sirven en el entrenamiento, y posterior evaluación basada en el algoritmo de k-NN para predecir a que clase pertenece un audio externo ingresado. Finalmente, para la evaluación se usa el algoritmo k-NN, que busca los k vecinos más cercanos y además se utiliza la correlación como herramienta para determinar el valor de k que proporcione mayor eficiencia al sistema de reconocimiento de voz.

Adquisición de la señal de audio

En la adquisición de las señales de audio se realizan grabaciones desde cero o se usan audios pregrabados. En este trabajo, se emplean audios pregrabados de capítulos de “Los Simpson” en español, de 5 personajes distintos, con una duración promedio de 3 segundos, tomando 18 muestras por personaje; 15 para el entrenamiento y 75 para la base de datos usada en el procesamiento y reconocimiento. Tres muestras de cada personaje, en total 15 para las pruebas de reconocimiento. Los audios se tomaron de un capítulo en específico por cada personaje, para trabajar con una muestra contralada y así centrarse en la exactitud del sistema de reconocimiento de voz.

Procesamiento de la señal

La recuperación de información musical (Music Information Retrieval, MIR) [MIR, 2011], agrupa una serie de algoritmos que permiten obtener metadatos del audio, calculando valores que lo describan de alguna forma de interés. Estos valores se denominan descriptores o características de un sonido, haciendo que el usuario piense en términos de descriptores MIR [Li-Chung, 2017]. Los ejemplos más utilizados son: Beats Por Minuto o tiempo (BPM), el centro de gravedad espectral (spectral centroid), contenido en alta o baja frecuencia, clave musical, la ubicación temporal de los eventos musicales (onsets), cantidad de disonancia armónica e índice de complejidad espectral (spectral complexity) [Big, 2017].

En este artículo, se hace uso de las librerías numpy, scipy, math, librosa, mpl_toolkits y matplotlib. para el pre-procesamiento y procesamiento en Python, que

fue el lenguaje para implementar el sistema. En el pre-procesamiento se obtuvo el espectrograma de las muestras (Figura 3) y la transformada de Fourier Discreta; que da como resultados una secuencia de números complejos, que constan de una parte real, imaginaria, fase y magnitud. Para el procesamiento en el dominio del tiempo se toman los valores absolutos de los *frames* de la señal de audio, en la figura 4 se muestran estos valores en la envolvente de la señal de audio; con los máximos locales de los valores absolutos de la gráfica de audio normalizada.

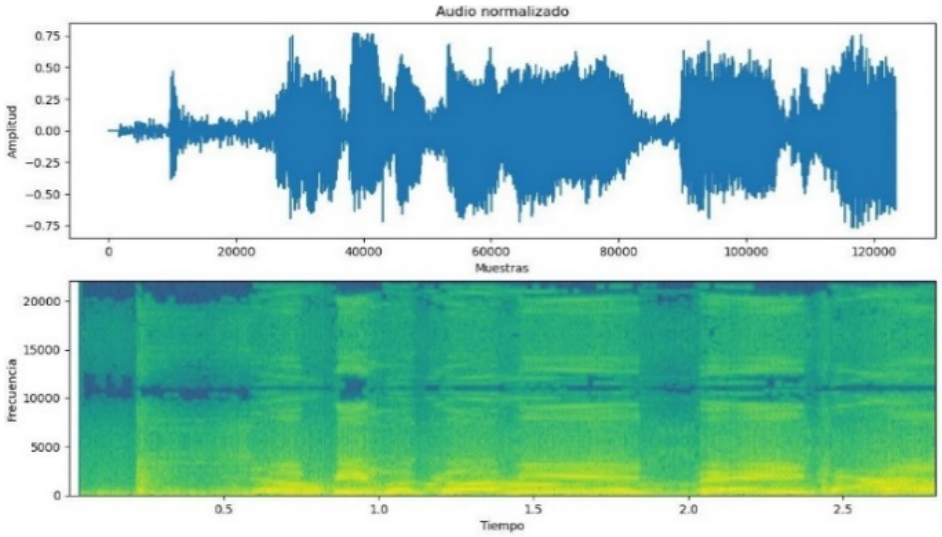

Figura 3 Audio y espectrograma de una señal de audio.

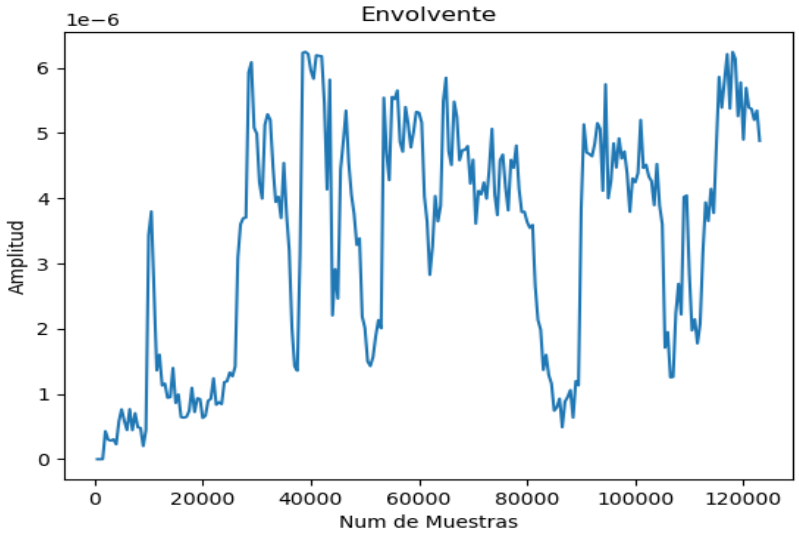

Figura 4 Envolvente de la señal de audio normalizada.

Extracción de características

En esta fase se agrupan, los sonidos por similitud en base a descriptores, lo que es posible al incorporar técnicas de Machine Learning al análisis de los metadatos obtenidos para cada muestra de la base de datos. El sonido se agrupa de acuerdo a sus descriptores como el timbre (centroide espectral, MFCC, etc.), características relacionadas con la dinámica (volumen en una señal acústica particular, el nivel medio, etc.) y características relacionadas con el pitch. Se considera un sonido como un grupo de características, cada una tiene un valor numérico [Chu, 2012].

Energía

La energía representa el volumen del audio, esta es calculada a partir de frames en el tiempo o la frecuencia. Los frames en el dominio del tiempo, se obtienen sumando los cuadrados de sus magnitudes para una señal continúa, ecuación 1.

$$\mathrm{E_x} = \int_{-\infty}^{\infty} |\mathrm{x(t)}|^2 \mathrm{dt} \tag{1}$$

Donde t es el tiempo y x la magnitud de la señal en función del tiempo. Para una señal discreta, se tiene la ecuación 2.

$$\mathrm{E_x} = \sum_{\mathrm{n=1}}^{\mathrm{N}} |\mathrm{x[n]}|^2 \tag{2}$$

Donde n es el número i-ésimo de la muestra (frame) y x la magnitud de la señal en función del número de frame. La sumatoria va desde la primera muestra hasta el N-ésimo número de muestras total de la señal.

Raíz media cuadrada

La raíz media cuadrada o RMS por sus siglas en inglés, representa la presión sonora del audio y es otra forma de visualizar la energía. Esta es la raíz cuadrada del promedio ponderado de la energía, que se calcula con la ecuación 3.

$$RMS_X = \sqrt{\frac{1}{N^2} \sum_{n=1}^{N} |x[n]|^2} \tag{3}$$

Implementación gráfica: energía vs RMS

Al considerar el sonido como un grupo de características, cada una con un valor numérico; como la energía y RMS, estás pueden expresarse en una dimensión, en gráficas del tipo *X* vs *Y* (Figura 5). Para probar esta idea se propusieron tres audios grabados con distintas palabras perro, agua y casa: 'casa.wav', 'perro.wav' y 'agua.wav'. Aplicando las operaciones de Energía y RMS se obtuvo:

- **Casa**: Energía= 584.3195553522062, RMSx= 0.0006043175672567602.
- **Perro**: Energía= 830.0490110166623, RMSx= 0.0007202642791957782.
- **Agua**: Energía= 2517.8710373610616, RMSx= 0.0012544598034017126.

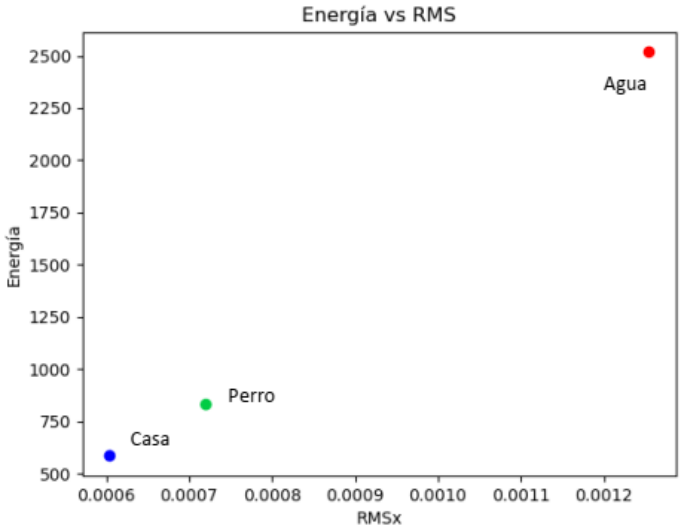

Figura 5 Características en el dominio del tiempo: Energía vs RMS.

Se agregó un cuarto audio llamado 'prueba.wav' donde la palabra dicha fue *Perro*. Obteniendo los siguientes datos: **Prueba**: Energía= 732.068258819977, RMSx= 0.000676418998670562. Se grafican estos datos, asignando en las ordenadas los datos de Energía y en las abscisas los RMS. En la figura 6, se aprecia que se obtuvieron datos correctos, ya que como se mencionó Perro fue la palabra de 'prueba.wav', y la palabra más cercana a ella, fueron los datos del audio de Perro. Sin embargo, con más palabras en la base de datos o grabaciones de las mismas palabras con un volumen más bajo o alto se podría generar resultados incorrectos. Por tanto, aunque la Energía y el RMS son parámetros muy necesarios a valorar,

se incluyen características en el dominio de la frecuencia, que aportan información de los audios, sin que estén sujetos a cambios tan sutiles como el volumen.

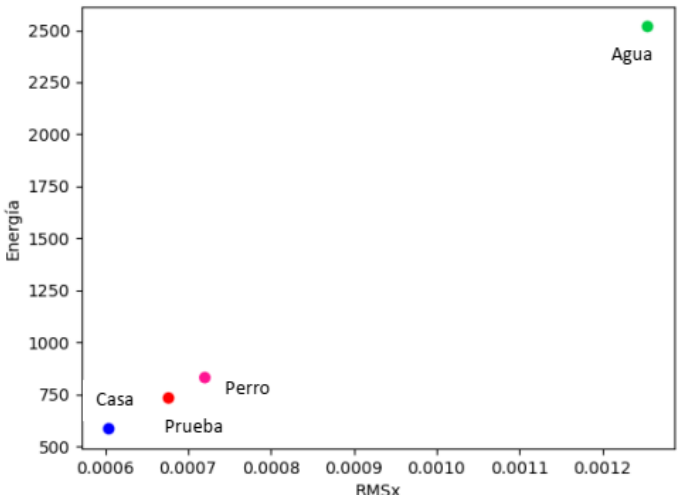

Figura 6 Características en el dominio del tiempo: Energía vs RMS.

Centroide espectral

Es una característica para definir la forma espectral de un sonido, indicando la parte más concentrada del espectro. Perceptualmente, se relaciona con la claridad que puede tener un sonido. Se calcula como la media ponderada de las frecuencias presentes en la señal, con sus magnitudes como lo pesos, ecuación 4.

$$Centroide = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \quad (4)$$

Donde $x(n)$ es la magnitud n- ésima y $f(n)$ representa la frecuencia n-ésima.

Escala de Mel

Usando la librería librosa y a partir de la transformada de la señal se obtiene la escala de Mel (Figura 7).

Mel-frecuency Cepstral Coefficients (MFCC)

MFCC es una representación de la magnitud espectral y se resuelve partiendo de la transformada del coseno del logaritmo de su magnitud espectral en una escala no lineal, llamada escala de Mel.

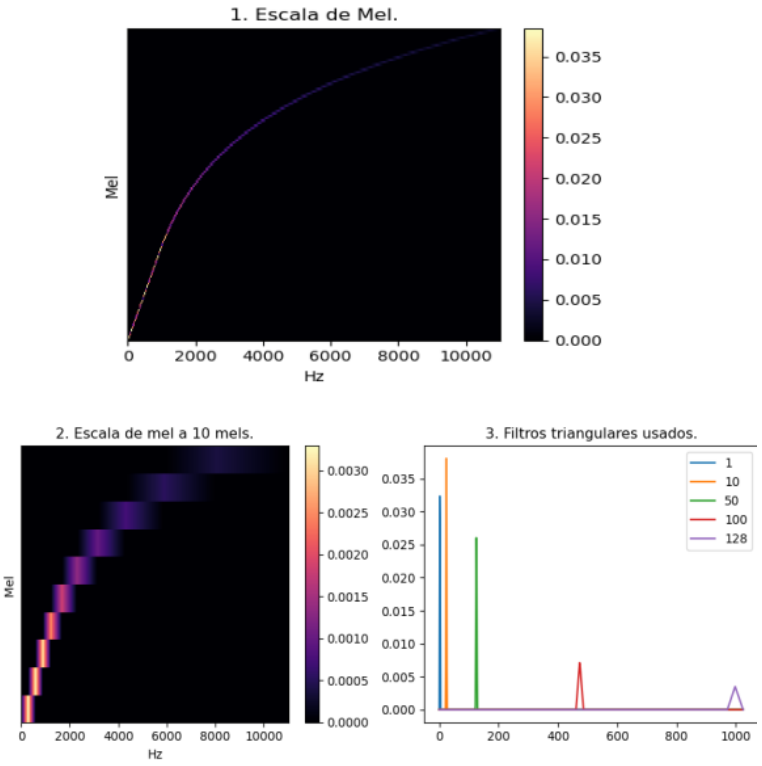

Figura 7 Características en el dominio del tiempo: Energía vs RMS.

La ecuación 5 muestra como el espectro completo $xl[k]$ es multiplicado por un banco de filtros. Así, cada frecuencia que sea dependiente de la *escala Mel* cambia. El objetivo es hacer más perceptiva la magnitud espectral del resultado de la FFT, después se aplica el logaritmo y por último la DCT. La extracción de coeficientes MFCC [Salomón, 2011] es la técnica de parametrización más utilizada en el área de verificación de locutor.

$$mfcc_l = DCT\left(log_{10}\left(\sum_{k=0}^{N/2}|X_l(k)|H_l(k)\right)\right) \quad (5)$$

Donde $\|X_l(k)\|$ es la parte positiva de la magnitud espectral, $H_l(k)$ es el banco de filtros de la escala de Mel y $DCT(m) = \sum_{n=0}^{N-1} f(n)\cos(\pi/N(n+1/2)m)$.

Implementación gráfica: MFCC vs centroide espectral

Después de calcular el Coeficiente de MFCC y el Centroide Espectral, se grafican y someten a la misma prueba de X vs Y con audios. La palabra de 'prueba.wav' fue

Agua, y resulto que la palabra más cercana fue Agua (Figura 8), igual sucedió para otros casos de prueba.

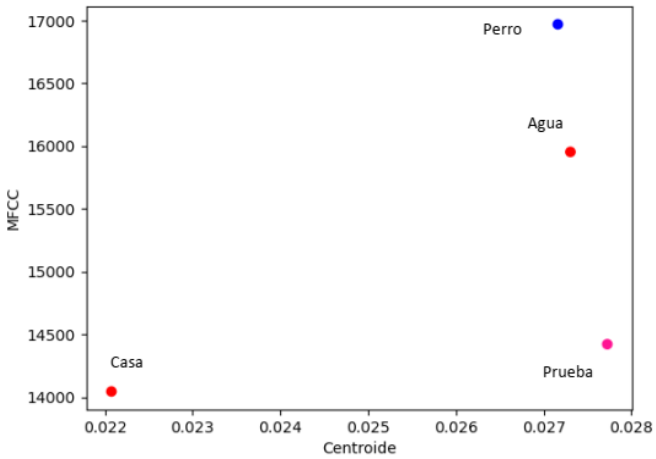

Figura 8 MFCC vs Centroide Espectral.

Entrenamiento del sistema

Para el entrenamiento del sistema se tomaron muestras de voz de 5 personajes de “Los Simpson”: Homero, March, Bart, Lisa y el Sr. Burns, del mismo capítulo; en total se tomaron 90 audios, 15 para la base de datos del sistema y tres de cada personaje para las pruebas.

Entrenamiento: características en el tiempo

Se realizó el entrenamiento calculando las características de Energía y RMS de las 75 muestras, marcando los rangos distintivos de cada personaje en una gráfica de una dimensión. En la figura 9, se muestran los datos del personaje de Bart.

Entrenamiento: características en la frecuencia

Se identifican las características en el dominio de la frecuencia de los audios, con los Coeficientes de MFCC y Centroide Espectral (Figura 10).

Clasificación

Para la clasificación se consideran las muestras de los 5 personajes, según sus características para ver las zonas o rangos específicos en las que se encuentran.

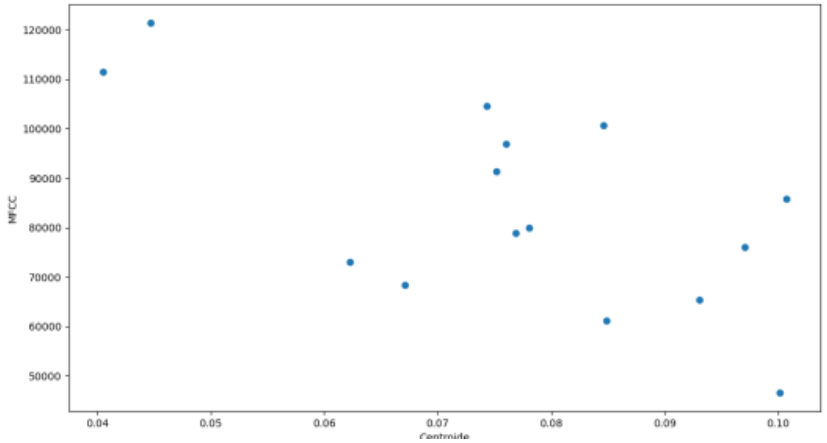

Figura 10 Caracterización de Bart en frecuencia: MFCC vs Centroide Espectral.

En la figura 11, se aprecia que las características de cada personaje relacionadas con la Energía (análogo al volumen de la voz) y RMS (análogo a la presión sonora), los sitúan en distintas zonas en la gráfica. March (puntos amarillos) se sitúa en la parte inferior izquierda, porque ambas características son bajas. Lisa (puntos rojos) con valores medianos de RMS y bajos en Energía; Bart en azul, Homero en verde y finalmente, en la parte superior derecha, el Sr. Burns con valores altos en Energía y RMS.

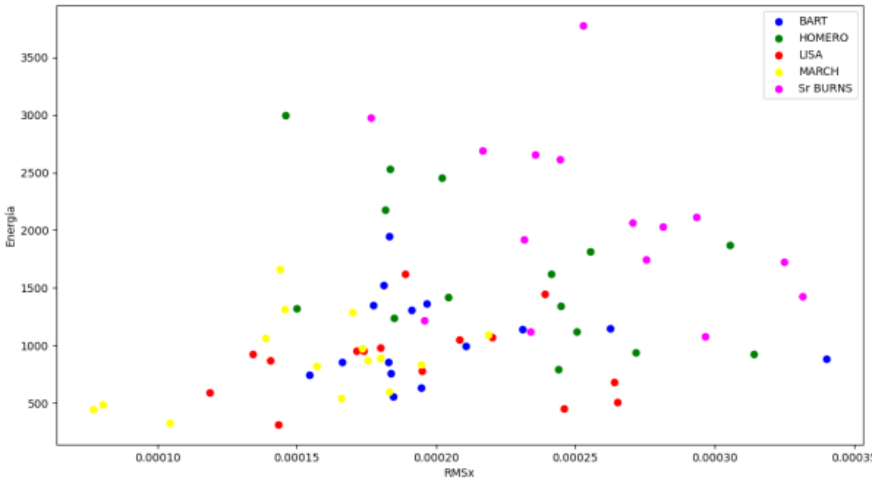

Figura 11 Energía vs RMS de todos los personajes.

Estableciendo que la Energía y RMS fueron muy efectivos para distinguir las zonas en las que se encuentra un personaje. Por tanto, se tiene un aporte de información

para detectar a quién pertenece la voz. Como los puntos se encuentran muy cerca en los rangos de RMS [0.00015 – 0.00025] y de Energía [500 - 1500], esto generaría ruido, por lo que se agrega la característica de frecuencia para tener información más concluyente. En esta gráfica, el parámetro que proporciona más información es la Presión Sonora (RMS) ya que presento mayor ritmo de cambio. En la figura 12, se aprecia que el parámetro MFCC y el *Centroide* Espectral de cada personaje también los coloca en distintas zonas: Bart en la parte superior, seguido del Sr. Burns, Lisa y en la parte inferior Homero y March. En general, el MFCC proporciona más información que el *Centroide* al aportar un mayor ritmo de cambio.

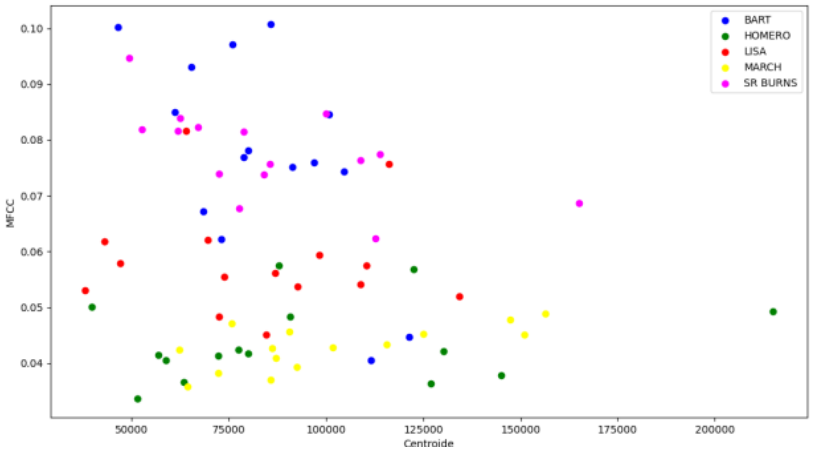

Figura 12 MFCC vs Centroide de todos los personajes.

Para visualizar mejor las zonas de dispersión de cada personaje se realizó una gráfica en 3D (Figura 13); con las características Energía, RMS y MFCC, mostrando un mayor ritmo de cambio, para los 15 audios de todos los personajes.

Evaluación

Con la clasificación por zonas de los personajes presentados en las figuras 11 y 12 se propone un algoritmo que determina a qué conjunto de datos pertenece un audio, por su posición en las gráficas (los valores cercanos al mismo).

Este sistema devolverá resultados en porcentajes, lo que significa que el mayor valor indica la más alta probabilidad de que pertenezca a algún personaje (en función del grupo de características) y valores menores caso contrario, siempre en

función de la posición en la gráfica del sonido externo. Consecuentemente, el *Resultado con Mayor Porcentaje* será tomado como la *Respuesta Final del Sistema*. En la figura 14, se muestra un audio de prueba ingresado con los audios de todos los personajes, se observa como el audio de prueba se sitúa en una zona en específico de la gráfica. Es importante mencionar que el algoritmo utilizado para predecir a que personaje pertenece este audio externo debe considerar la zona en la que se encuentra, es decir, los valores más cercanos a él.

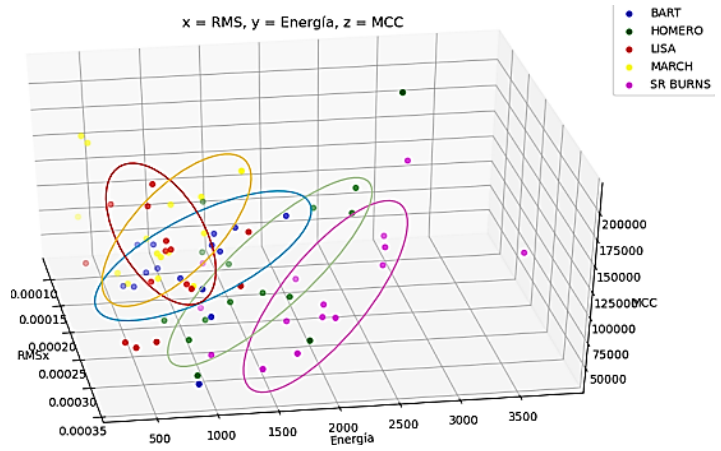

Figura 13 Gráfica 3D de todos los personajes.

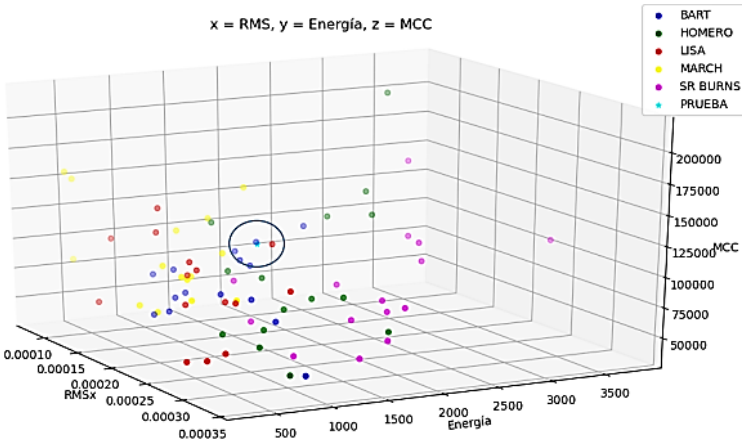

Figura 14 RMS vs Energía vs MFCC ingresando un audio de prueba.

Evaluación: algoritmo k-NN y correlación de Pearson

Se propone resolver el problema planteado con un algoritmo de aprendizaje supervisado, ya que se tienen datos de entrenamiento y se pretende deducir o

predecir el resultado en función de los mismos. El algoritmo propuesto para resolver este problema, es k-NN, que es uno de los algoritmos de clasificación básicos y esenciales en Machine Learning. Este algoritmo predice a qué conjunto de datos o clustering pertenece un dato ingresado, sin tener información adicional de los clustering. Dicho algoritmo está en función de determinar distancias, se utiliza la fórmula para la distancia entre dos puntos en un plano cartesiano (Ecuación 6).

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (6)$$

Para aplicar esta fórmula, se convirtieron los datos a la misma escala en ambos ejes de las gráficas 11 y 12. KNN sigue los pasos básicos:

- Calcular distancias.
- Encontrar sus vecinos más cercanos (las K distancias mínimas al punto a evaluar).
- Calcular grados de pertenencia a las etiquetas (asignar porcentajes).

El número de vecinos en KNN

El número de K vecinos depende del tipo de conjuntos de datos, cada conjunto de datos tiene sus propios requisitos [Li-Chun, 2020]. En este artículo, se han considerado el tiempo y la frecuencia, para lo cual, se propuso la Correlación de Pearson para determinar el valor de k a utilizar. Por tanto, en vez de usar un valor fijo de K (lo usual en el uso de K-NN), se propone ecuación 7 para la elección de k.

$$k(r) = \max(r_1, r_2, \dots, r_n) \quad (7)$$

Donde r es la Correlación de Pearson, ecuación 8, que puede tomar valores de 0 a 1 donde a mayor valor mejor relación entre los datos.

$$r_i = \frac{N\sum_{l=0}^{N} xy - (\sum_{l=0}^{N} x)(\sum_{l=0}^{N} y)}{\sqrt{(N\sum_{l=0}^{N} x^2 - (\sum_{l=0}^{N} x)^2)(N\sum_{l=0}^{N} y^2 - (\sum_{l=0}^{N} y)^2)}} \quad (8)$$

Donde N es el número total de datos, X el primer arreglo de datos y Y el segundo arreglo de datos a comparar respecto a X. **Procedimiento de la evaluación**:

- **Aplicar el algoritmo K-NN.** Calcular la distancia del valor a evaluar con los demás datos y después calcular los K vecinos más cercanos. Considerando

$Dom(t)$ y $Dom(f)$, variando K desde el 10% de los datos (k = 8 para este caso) y 20% (k = 15) en pasos de 1, (k [8,15]). Los porcentajes se fijaron por experimentación, al ver el comportamiento de los datos para esos valores y se concluyó que en estos los resultados son los mejores.

- **Calcular correlación de Pearson.** Del paso anterior, se tienen 7 cadenas para calcular la correlación, donde "x", ecuación 8, son los datos en el $Dom(t)$ y "y" serán los datos en el $Dom(f)$. Asi se obtiene una cadena de 7 correlaciones en función de variar K siete veces de 8 a 15, para este caso.
- **Elección de K.** De la cadena de tamaño 7 elegiremos usar la K con el valor de correlación r máximo en la cadena, ecuación 7.
- **Evaluar.** Fijando ya el resultado en el paso anterior, se asignan porcentajes en función del número de datos arrojados para cada personaje (asignación de etiquetas, cada etiqueta es un personaje), el porcentaje máximo de la evaluación será nuestro Resultado Final de Predicción.

3. Resultados

El sistema de reconocimiento de voz se sometió a 15 pruebas, dos de las mismas se muestran a continuación. La primera, correspondiente a la voz de Lisa resulto como mejor correlación entre los datos $Dom(t)$ y $Dom(f)$, para k = 8 (figura 15 y 16).

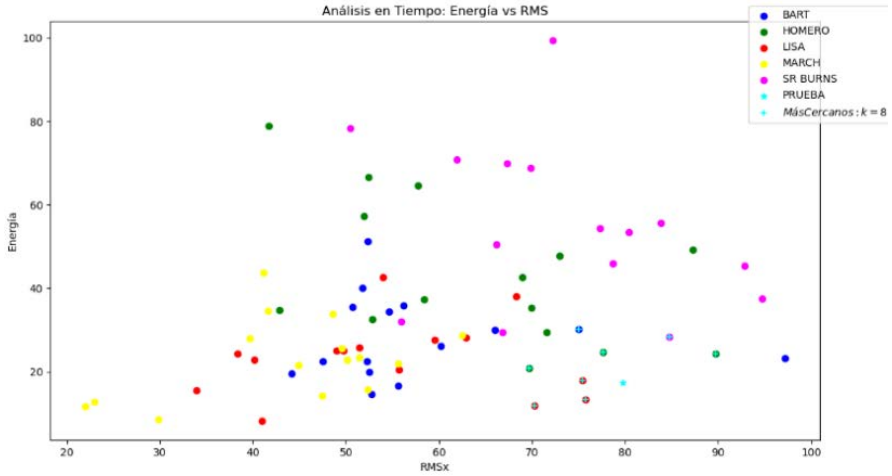

Figura 15 Prueba Audio Uno. $Dom(t)$: $K = 8$ vecinos más cercanos.

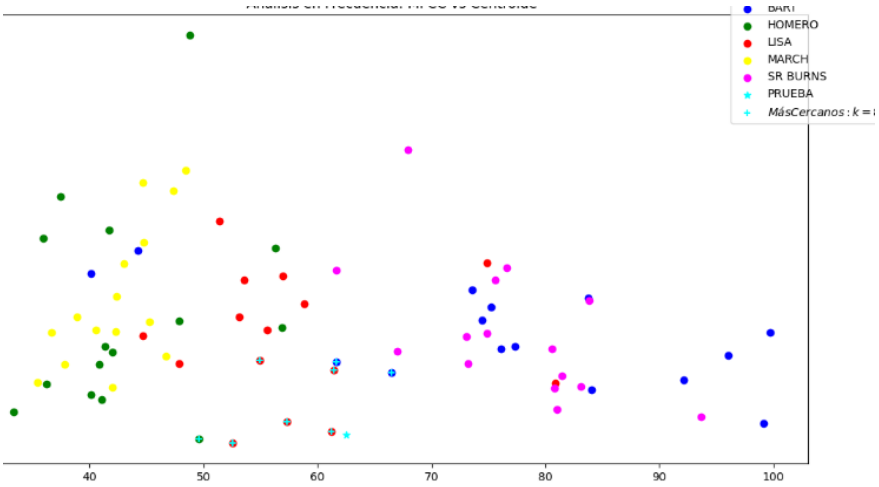

Figura 16 Prueba Audio Uno. $Dom(f)$: $k = 8$ vecinos más cercanos.

Los resultados de la evaluación por el algoritmo K-NN y la Correlación de Pearson son:

$$r(k[8,15]) = [0.6469966392206304, 0.4587596979470412, 0.4225771273642583, 0.42059978233966894, 0.37234154758203314, 0.5086989586601017, 0.3653325655726347].$$

Con la ecuación 7 se elige k = 8, se tiene la solución en la figura 17, que indica que el audio de Prueba Uno pertenece a Lisa, lo cual es **correcto**.

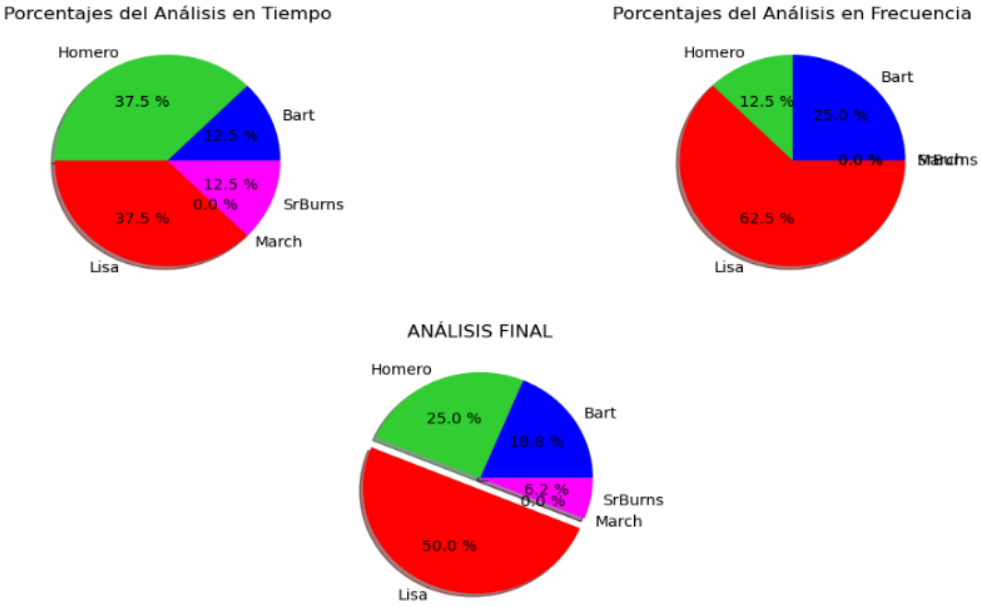

Figura 17 Gráficas de Pastel de predicciones con k = 8.

En la segunda prueba correspondiente a la voz de Bart, el algoritmo eligió k = 11 (véase la figura 18 y 19), como mejor correlación con r=0.2762137:

$$r(k[8,15]) = [0.16333965194414124, 0.15002479953724476, 0.11396057645963796, 0.2762137969161919, 0.2563137711111046, 0.192327314540518, 0.1468293986405344].$$

Los resultados para k = 11 se muestran en la figura 20, dando como resultado que el audio pertenece a Bart. Lo cual es **correcto**.

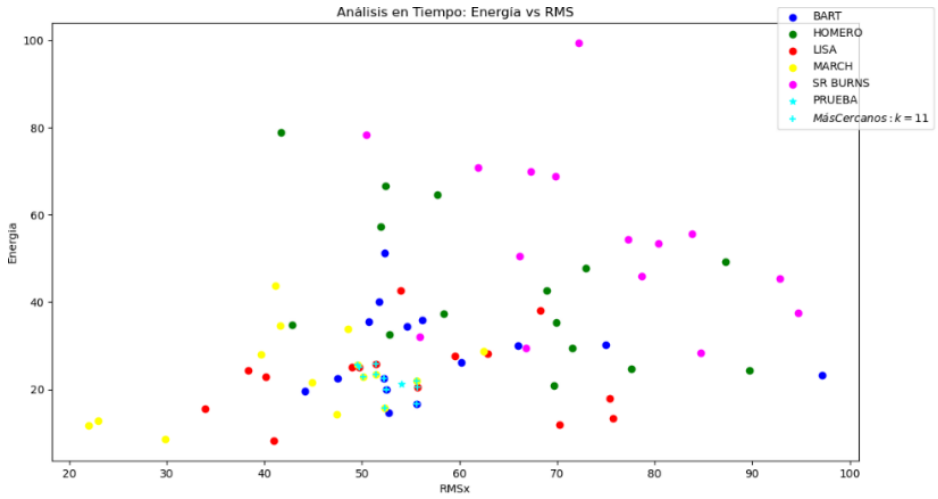

Figura 18 Prueba Audio Dos. $Dom(t)$: $K = 11$ vecinos más cercanos.

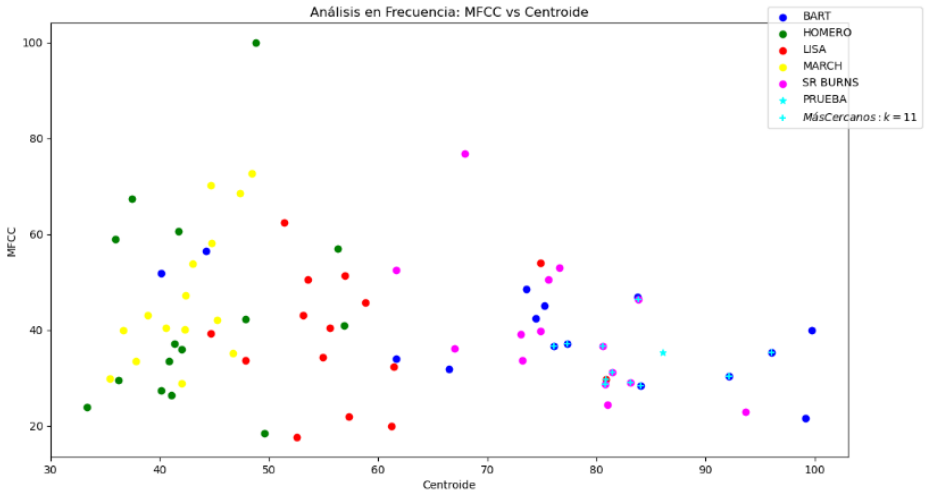

Figura 19 Prueba Audio Dos. $Dom(f)$: $k = 11$ vecinos más cercanos.

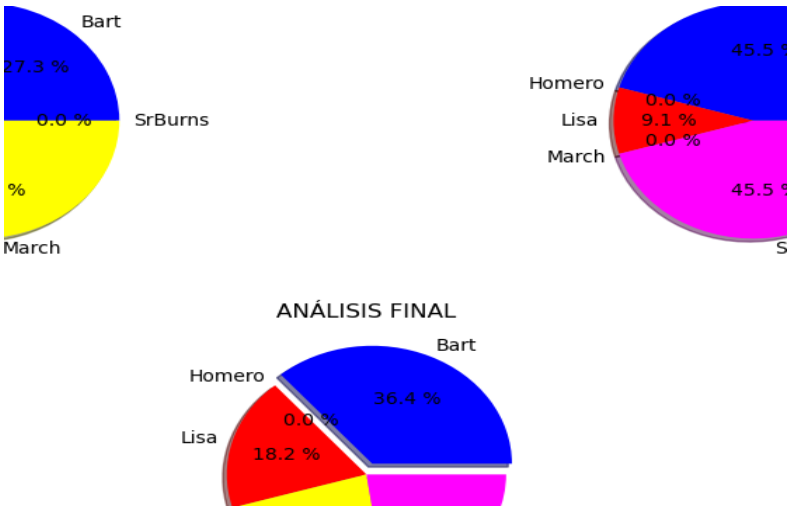

Figura 20 Gráficas de Pastel de predicciones con k = 11.

4. Discusión

El sistema fue probado con los 15 audios de prueba, 3 por cada personaje. Los resultados se muestran en la tabla 1, utilizando un peso de 1 para correcto, 0.5 empate y 0 incorrecto.

Tabla 1 Resultados de las 15 pruebas al sistema de reconocimiento de voz.

Personaje	**Respuesta para k fijo k = 8**	**Respuesta para k variable k = [8,15]**	
		k(Max.correlación)	
Bart	correcto	11	correcto
	correcto	11	correcto
	correcto	13	correcto
Homero	correcto	8	correcto
	correcto	13	correcto
	correcto	9	correcto
Lisa	correcto	8	correcto
	correcto	9	correcto
	correcto	13	correcto
March	correcto	8	correcto
	incorrecto	11	empate
	empate	14	correcto
Sr Burns	correcto	12	correcto
	correcto	14	correcto
	correcto	14	correcto
Aciertos	13.5		14.5
% de eficiencia	90%		96.67%

Los resultados muestran que la propuesta en este artículo de variar k del algoritmo k-NN en función de la correlación de Pearson proporciona un sistema de reconocimiento de voz más exacto que con el parámetro k fijo, es decir, se observa como el sistema mejora su respuesta buscando la máxima correlación posible. El sistema obtuvo la predicción de empate (0.5) para un personaje, esto es, encontró la misma relación entre dos personajes; esto sugiere ingresar más datos en la etapa de entrenamiento para una respuesta concluyente.

Es importante mencionar que la elección del 10% de los datos (k = 8) se hizo ya que por lo general es la mejor opción. También es primordial decir que los resultados usado un k fijo (al 10 %) es más preciso, ya que arroja siempre mayores porcentajes a los personajes ganadores, pero es menos exacto ya que se equivoca un poco más. En contraste, k variable en función de la correlación es más exacto pues se equivoca menos, pero es menos preciso ya que arroja Porcentajes Menores a los personajes Ganadores. Esto último, no afecta los resultados finales ya que la respuesta del algoritmo lo asigna al personaje que le corresponde. En consecuencia, calcular el valor de k con la correlación de Pearson permite **lograr mayor exactitud** al sistema de reconocimiento de voz realizado en este trabajo**.**

5. Conclusiones

Se logró implementar un Sistema de Reconocimiento de Voz basado en el método de aprendizaje supervisado, denominado algoritmo k-NN, que suministra mayor exactitud, gracias a la propuesta de calcular k mediante la correlación de Pearson. Para ello, se caracterizó la voz en función de parámetros especiales del audio como sonoridad, volumen y componentes de frecuencia, así como se clasificó en zonas a cada personaje de acuerdo a dichas características. También, se probó bajo las mismas condiciones, pero con un k fijo dicho sistema, obteniendo un 90% de exactitud; mientras que variando el parámetro de k en función de la mejor correlación posible se alcanzó un 97% de exactitud. Con lo cual la propuesta algorítmica funcionó correctamente para mejorar el rendimiento del Sistema de Reconocimiento de Voz, que era lo que se buscaba y por ello se utilizaron personajes bien conocidos. Como trabajo futuro, se utilizarán más voces y con

diferentes tonos (graves y agudos para la misma persona) buscando aumentar la complejidad, pero ofreciendo el mismo nivel de exactitud de nuestro sistema.

6. Bibliografía y Referencias

[1] About I, and Denis V. Historia de la identificación de las personas, 2011.

[2] Aguirrezabala M. Estudio de verificación biométrica de voz. Tesis de Maestría, 2015.

[3] Arias A., Rubiano D. Método automático de reconocimiento de voz para la clasificación de vocales al lenguaje de señas colombiano, Disertación, 2018.

[4] Big, Aproximación de Big Data a las Colecciones Musicales. 5to Congreso Nacional de Ingeniería, Informática/Sistemas de Información. Aplicaciones Informáticas y de Sistemas de Información. Noviembre 2017.

[5] Chu S., Narayanan S. and Jay Kuo C. Environmental sound recognition with time-frequency audio features. IEEE Trans. Audio, Speech and Lang. pp. 1142-1158, 2012.

[6] Fix E., Hodges, J. L. An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). International Statistical Review / Revue Internationale de Statistique 57 (3): 233-238, 1989.

[7] Furui S. Talker recognition by long time averaged speech spectrum Electronics and Communications, Japan, 1972.

[8] Juang B. H. The Past, Present, and Future of Speech Processing, IEEE Signal Processing Magazine, mayo 1998.

[9] Li-Chun W., (2020). An Industrial-Strength Audio Search Algorithm. Shazam Entertainment, Ltd: https://www.ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf.

[10] Matsui T., Furui S. Concatenated phoneme models for text-variable speaker recognition, Proc. ICASSP,1993.

[11] MIR, (2020). Music Information Retrieval: Part 2. Feature Extraction. Alexander Schindler: http://www.ifs.tuwien.ac.at/~schindler/lectures/MIR_ Feature _Extraction.html.

[12] Ortega M. Introducción a la biometría. técnicas avanzadas de procesado de imagen, 2013.

[13] Rabiner L. R., Juang B. H. Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, 1993.

[14] Salamón J., Gómez E., Bonada J. Sinusoid Extraction and salience function design for predominant melody stimation.Music Technology Group Universitat Pompeu Fabra, Barcelona, 2011.

[15] Tordera J. C., (2101). Lingüística computacional. Tratamiento del habla. Valencia: Universtitat de València: https://es.wikipedia.org/wiki/ Reconocimiento_del_habla.

[16] Weisstein E. W., Fast Fourier Transform. Weisstein, Eric W., ed. MathWorld Wolfram Researc, 2015.