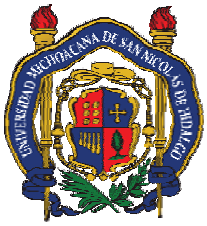


UNIVERSIDAD MICHOACANA DE SAN NICOLÁS DE HIDALGO

Facultad de Ingeniería Eléctrica

División de Estudios de Posgrado

IDENTIFICACIÓN DE PARLANTES INDEPENDIENTEMENTE DEL TEXTO UTILIZANDO LOS FORMANTES DE LAS VOCALES

TESIS

Que para obtener el grado de

MAESTRA EN CIENCIAS EN INGENIERÍA ELÉCTRICA

presenta

Monserrat Aranzazu Castro Coria

Dr. José Antonio Camarena Ibarrola

Director de Tesis

Morelia, Michoacán Agosto 2019

"Lo único imposible es aquello que no se intenta"

.

Agradezco a mi director de tesis:
Dr. José Antonio Camarena Ibarrola, por todo su apoyo, tiempo y confianza, por sus consejos y correcciones, por su compromiso hacia la investigación pero sobre todo por creer en mi.
A mis maestros:
Dr. Juan José Flores Romero y Dr. Jaime Cerda Jacobo, por creer en mi, por impulsarme a realizar más cosas de las que creía que podía hacer.
Dr. Félix Calderón Solorio y Dr. Luis Valero Elizondo por sus comentarios y valiosas aportaciones a esta tesis.
A mi esposo:
Por sus desvelos acompañandome en este proyecto, por su amor y apoyo incondicional, por ser uno conmigo.
A mis padres:
Porque son mi más grande inspiración, porque formaron una hija que ama estudiar, porque me impulsan y están siempre en todos mis proyectos.
A mis hijos:
Bryan e Hiram, ustedes son mi motivo de crecer, porque me han hecho aprender que si la palabra guía, el ejemplo arrastra.
A mi cuñado y su hermosa familia:
Por su apoyo y cariño.
A mi hermano, a mi equipo, amigos y compañeros, porque compartieron su tiempo y conocimientos en esta etapa.
A la Universidad Michoacana de San Nicolás de Hidalgo, porque me ha dado tanto y me ha hecho ser quien soy.

Lista de Publicaciones

"Cloud Point Matching for Text-Independent Speaker Identification"
Antonio Camarena-Ibarrola, Monserrat Castro-Coria and Karina Figueroa
Publicado en 2018 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), pages 1-6

Best Computing Track Paper Award (ROPEC 2018)

"An efficient method to obtain bifurcation diagrams based on PSO algorithms"
Bryan E. Martínez, Monserrat Castro-Coria, Jaime Cerda and Alberto Avalos
Publicado en Proceedings of the World Congress on Engineering and Computer Science 2018 Vol. I, pages 73-78

Certificate of Merit for International Conference on Computer Science and Applications 2018 (WCECS 2018)

"Parallel mining of frequent patterns for school records analytics at the Universidad Michoacana"
Juan J. Flores, J. Luis Garcia-Nava, Monserrat A. Castro-Coria, Victor M. Tellez, Emanuel Huerta B., Josue Espinosa-Romero and Felix Calderon
Publicado en 2017 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), pages 1-6

Resumen

La identificación de parlantes independientemente del texto es la forma de reconocer a un individuo por su voz sin que el locutor pronuncie una o varias palabras en específico. Esta forma de identificación suele ser utilizada para reconocimiento forense o procesos judiciales, donde el individuo a identificar no tiene que estar presente o no quiera colaborar, con solo tener una grabación de su voz es posible identificarlo siempre y cuando se tengan características de su voz en un diccionario de datos.

Existen muchas formas para extraer diferentes características de la señal de voz, esta tesis implementa la detección de los primeros tres formantes de las vocales a partir de los coeficientes de predicción lineal para formar un vector de características de tamaño n por 3, donde n es el número de vocales pronunciadas en una frase y 3 los formantes que se obtienen por cada vocal, por lo que se obtiene un vector de características para cada individuo de las bases de datos. Para identificar a un individuo, se mide la similitud entre vectores de un individuo desconocido con los vectores que pertenecen a los individuos conocidos.

Se realizaron pruebas con dos bases de datos en español y dos bases de datos en inglés de diferente número de individuos así como con diferente número de vocales pronunciadas, logrando resultados favorecedores. Además se realizó una prueba buscando las diferencias entre la voz de una persona real y un imitador encontrando incluso diferencias entre estos. Con ello se logro una robustez en el algoritmo para dejar abierto posibilidades a futuros trabajos.

Palabras clave: Identificación de individuos, Texto independiente, LPC, Formantes, Identificación de vocales.

Abstract

The identification of speakers regardless of the text-independent is the way to recognize an individual by his/her voice without the speaker pronouncing one or several specific words. This form of identification is usually used for forensic recognition or legal proceedings, where the individual to be identified does not have to be present or does not want to collaborate, just by having a recording of their voice it is possible to identify it as long as they have characteristics of his/her voice in a data dictionary.

There are many ways to extract different characteristics of the voice signal, this thesis implements the detection of the first three vowel formants, from the linear prediction coefficients, to form a vector of characteristics of size n by 3, where n is the number of vowels pronounced in a sentence and 3 the formants that are obtained by each vowel, so that a vector of characteristics is obtained for each individual of the databases. To identify an individual, the similarity between vectors of an unknown individual is measured with vectors belonging to known individuals.

Tests were carried out with two databases in Spanish and two databases in English of different numbers of individuals as well as with different numbers of pronounced vowels, achieving favorable results. In addition, a test was conducted looking for the differences between the voice of a real person and an imitator, finding differences between them. This achieved a robustness in the algorithm to leave open possibilities for future work.

Contenido

Lista de Figuras

Lista de Tablas

Lista de Símbolos

F_k	K-ésimo formante.
Hz	Unidad de Frecuencia del Sistema Internacional de Unidades (Hertz).
$R_n(k)$	Correlación.
$\hat{R}_n(k)$	Correlación modificada.
$S(n)$	Señal de audio con respecto de n muestras.
$w(n)$	Ventana de una señal con respecto de n muestras.
$y(n)$	Filtro pre-énfasis con respecto de n muestras.

Lista de Acrónimos

dB	Unidad de decibel.
FN	Falso negativo (False negative).
FP	Falso positivo (False positive).
FPR	Tasa de falsos positivos (False positive rate).
LPC	Coeficientes de Predicción Lineal (Linear Predict Coeficients).
MSACF	Función modificada de autocorrelación de tiempo corto (Modified Short-time Autocorrelation Function).
PARCOR	Correlación Parcial (Partial Correlation)
ROC	Reciever Operating Characteristics.
SACF	Función de autocorrelación de tiempo corto (Short-time Autorrelation Function). (Structural on Maximum a Posteriori Probability)
TISI	Identificación del individuo de forma Texto-Independiente (Text Independent Speaker Identification)
TN	Verdadero negative (True negative).
TP	Verdaderos positives (True positive).
TPR	Tasa de verdaderos positivos (True positive rate).

Capítulo 1

Introducción

"La voz es el propio emblema del hablante, entretejido indeleblemente en el tejido del habla. En este sentido, cada una de nuestras expresiones del lenguaje hablado lleva no solo su propio mensaje, sino que a través del acento, el tono de voz y la calidad de voz habitual, es al mismo tiempo una declaración audible de nuestra membresía en grupos sociales regionales particulares, de nuestra persona. Identidad física y psicológica, y de nuestro estado de ánimo momentáneo."

John Laver

1.1. Planteamiento del Problema

El reconocimiento del parlante es un problema que utiliza la voz de un individuo para identificarlo. Existen dos procesos para la identificación del parlante, el primero dependiente del texto (Text Dependent Speaker Identification, TDSI por sus siglas en inglés), donde la persona que se quiere identificar tiene que pronunciar una palabra o frase preestablecida vista como una "palabra secreta"de su voz (contraseña), el segundo independientemente del texto (Text Independent Speaker Identification, TISI por sus siglas en inglés), donde el individuo es identificado indistintamente de la palabra o frase que pronuncie. Se obtienen características de la frecuencia de la señal de voz, las cuales se comparan de acuerdo a su duración, dinámica, intensidad y tono, después de un procesamiento de la señal, estas características obtenidas se ingresan en un diccionario de datos, por lo que el sistema

determinará de qué parlante se trata.

En la Figura 1.1 se muestra que para identificar un individuo de forma dependiente del texto es necesario que se pronuncie una frase o contraseña para que se pueda verificar, como ejemplo el caso de un banco que para realizar trámites vía telefónica solicita que se repita la frase .En Banco Santander mi voz es mi firma", la frase siempre será la misma todas las veces que se realice un trámite, cuando se identifica un individuo independientemente del texto que se pronuncie se hace buscando características de la señal de voz, sin que el individuo repita una frase ya conocida por el sistema a detectar, esto incluso puede ser mientras la persona se encuentra en un dialogo con otra persona.

Figura 1.1: Formas de identificación de individuos por su voz, donde dependiente del texto implica que el sistema sabe lo que el individuo va a decir (palabra, contraseña, frase, etc.), e independiente del texto donde la persona puede decir cualquier cosa, esto incluso puede ser mientras mantiene una conversación.

La identificación del hablante es aún más difícil que la verificación de éste. Los sistemas de verificación comparan características extraídas de la señal de voz de la persona cuya identidad se está verificando con las características que pertenecen a la persona que dice ser, mientras que los sistemas de identificación no cuentan con información sobre la identidad del individuo y debe reconocerlo como uno de los individuos conocidos de la base de datos. Este problema se vuelve más complejo mientras mas parlantes se encuentran en la base de datos.

1.1.1. Motivación

Ningún individuo suena igual a otro, ya que los órganos que influyen en el proceso de generación de la voz son diferentes. Existe información particular de cada persona en la forma de generar los sonidos de su voz, como lo son el timbre, tono, velocidad y volumen, estos en combinación con el tamaño de la tráquea y la forma y posición de los músculos de la boca que se utilizan para pronunciar los sonidos nos hacen diferentes. Por ejemplo, al recibir una llamada telefónica de un familiar con solo escuchar su voz podemos identificar de quien se trata, incluso de personas con las que tenemos contacto es posible identificarlos por su voz sin necesidad de que se identifiquen con su nombre.

Desde hace muchos siglos se ha buscado poder identificarnos por medio de palabras claves para ingresar a un lugar o ser reconocidos dentro de un grupo social, además el reconocimiento de la voz de un individuo ha sido aceptado dentro de los procesos judiciales como evidencia principal.

La primera evidencia documentada de un testigo auditivo ocurrió en la ciudad de Nueva York en el año de 1861 (Tosi y Tosi, 1979), donde un juez permitió como evidencia el testimonio de una persona que mencionaba podía identificar al perro del acusado reconociendo los ladridos como uno de los perros que habían matado a su oveja (Caso Wilbur vs Hubbard, 1861). El juez determinó que si una persona podía identificar a otra por su voz, también era capaz de identificar el ladrido de un perro.

El primer juicio donde se utilizó la identificación de un individuo por su voz data del año 1932. Charles Lindbergh fué aviador e ingeniero estadounidense, famoso por ser el primer piloto en cruzar el océano Atlántico, y su hijo Charles Lindbergh Jr., por ende, el bebé mas famoso del mundo. En marzo de 1932, Charles Lindbergh Jr. fué secuestrado y posteriormente encontrado muerto. Este caso fué llamado el "crimen del siglo" (Shaver y Acken, 2016) por la popularidad de los involucrados y los hechos en que se suscitó el crimen. A un mes de haber ocurrido el secuestro, Lindbergh aseguró haber escuchado la

voz del secuestrador y dos años y medio después, el juez que llevaba el caso pidió a Lindbergh testificar escuchando nuevamente la voz del presunto responsable del secuestro de su hijo, donde declaró que era la misma voz que había escuchado dos años atrás (Evidencia de Identificación de Voz del Estado vs Hauptmann, 1935), Bruno Hauptmann, un carpintero de origen alemán, fue declarado culpable del secuestro y asesinato de Lindbergh Jr. Este caso fué ampliamente estudiado por Frances McGehee (Yarmey, Yarmey, y Todd, 2008).

En febrero de 1984, Paul Prinzivalli, originario de Long Island y empleado de la Aerolínea Pan American World Airways (Pan Am) fué acusado y arrestado por haber realizado una serie de falsas amenazas telefónicas de bombas a los vuelos de Pan Am. Los funcionarios de la compañía, al escuchar las cintas de audio creyeron que la voz pertenecía a su empleado Prinzivalli porque la voz del audio sonaba como la de él. Fué despedido y encarcelado durante nueve meses, después de este tiempo, el profesor William Labov, experto sociolingüista de Pensilvania, logró identificar las diferencias entre las dos voces al demostrar que las llamadas fueron hechas por un hombre con acento de Boston y Prinzivalli tenía un acento de Nueva York. Se estableció una duda razonable y Prinzivalli fué absuelto de los cargos (Rose, 2003).

En abril de 1993, Anabel Segura fué secuestrada cerca de la ciudad de Madrid, sus familiares recibieron varias llamadas telefónicas pidiendo dinero a cambio de la libertad de la joven, los secuestradores incluso enviaron una cinta con la voz de la joven quien les pedía a sus familiares que pagaran el rescate, supuestamente. Tanto la policía como la familia no dudaron de la autenticidad de la grabación. Investigadores de la Policía Nacional lograron identificar características del acento de los secuestradores cercanas a la ciudad de Toledo. En septiembre de 1995 fueron detenidos los secuestradores quienes confesaron el delito así como la localización del cadáver, dentro de las declaraciones se encontraba que la esposa del secuestrador fué quien había hecho la grabación haciéndose pasar por Anabel (Tola, Catanzaro, Viciano, y Hummel, 2019).

El reconocimiento de individuos por su voz tiene una gran cantidad de usos co-

mo las llamadas telefónicas bancarias, compras por teléfono, acceso a servicios de bases de datos, servicios de información, correos de voz, seguridad y control para información en áreas confidenciales, acceso a computadoras remotas, control de acceso a sitios de Internet, monitoreo en llamadas en la prisión, análisis forense y aplicación de la ley (El-Samie, 2011).

El uso de nuevas tecnologías ha facilitado la realización de trámites por medio de la telefonía, ya que hoy en día se pueden realizar operaciones bancarias, compras por medio del teléfono, operaciones mercantiles e incluso tratos de negocios desde un país a otro sin necesidad de traslados. El uso de diversas cuentas de internet para acceder a un sin número de servicios, como lo son correo electrónico, redes sociales, sitios de búsqueda, entre otros, solicitan identificarnos por medio de un usuario y contraseña, un problema de esto es que muchas veces no recordamos la contraseña para una cuenta en específico cuando se tienen diversas cuentas de los servicios que utilice en internet, tener un reconocedor que identifique al individuo solo usando su voz podría ayudar a acceder a tantas cuentas como sea necesario sin tener que recordar contraseñas. Es útil poder identificarnos por medio de la voz y por ello se propone un identificador de parlantes de modo texto independiente.

1.2. Antecedentes

Tola *et al.* (Tola y cols., 2019) crearon una línea de tiempo con los avances tecnológicos del procesamiento de voz, en los años 40's consistía en la detección de características de la voz como el tono, timbre, calidad de la voz, articulación, dicción, así como ciertas características fonológicas y léxicas escuchando muestras de voz con el oído propio.

Entre los años 50's y 60's se desarrollaron diversas investigaciones para explorar ideas fundamentales sobre los fonemas acústicos, desde las formas para el procesamiento de las señales de voz hasta el uso de tecnologías del computador, utilizando resonancias espectrales y filtros análogos para extraer características de ciertas regiones de las vocales pronunciadas. Sadaoki Furui (Furui, 2005) describe el progreso de los últimos 50 años dentro

de la investigación del reconocimiento automático del habla y del hablante, donde se han tratado diversos métodos para obtener características de los sonidos que son pronunciados por un individuo. Los laboratorios Bell fueron de los primeros en construir un sistema que identificara un individuo utilizando las frecuencias de los formantes de cada región de las vocales. En 1959, en la Universidad de Inglaterra se creó un reconocedor de cuatro vocales y nueve consonantes obteniendo características secuenciales de los fonemas en inglés.

En 1962 (Sakai y Doshita, 1962), en la Universidad de Kyoto, Japón, se construyó un reconocedor de fonemas utilizando un hardware que segmenta los sonidos y determina los cruces por cero para el análisis de los diferentes fonemas. En ese mismo año fue creado un método espectrográfico utilizando la "huella vocal"(Tola y cols., 2019) con los espectrogramas[1] y comparando la similitud entre éstos entre diversos hablantes.

A partir de los años 70's fueron utilizadas numerosas formas de programación dinámica para solucionar problemas del reconocimiento automático de palabras. Durante esta década, laboratorios como Bell, IBM y AT&T buscaron generar nuevos algoritmos para facilitar el reconocimiento de individuos por medio de coeficientes de predicción lineal (LPC) (Itakura, 1975). Rabiner utilizó los coeficientes de correlación parcial (PARCOR) para la identificación de individuos en 1978 (Rabiner y Schafer, 1978), estos coeficientes se relacionan con los coeficientes de reflexión que son la base del modelo de tubos sin pérdida del tracto vocal.

En los años 80's se utilizaron diversas técnicas, como se muestran en la Figura **??**, como Modelos Ocultos de Markov (HMM) (Poritz, 1982), coeficientes Cepstrales y Redes Neuronales, entre otros, para producir secuencias de observaciones y características de las señales de voz que describieran mejor a los individuos para su clasificación e identificación (Naik, Netsch, y Doddington, 1989). Se creó un método acústico fonético (Tola y cols., 2019) para medir parámetros acústicos de la voz como duración, distribución espectral de

[1]Representación gráfica de secuencias de espectros de frecuencia de una señal de audio obtenida de la Transformada de Fourier de tiempo corto (Rabiner y Schafer, 2011)

la energía, frecuencias de formantes vocálicos, trayectoria, dinámica, espectro promedio y la distribución de los formantes de la señal de voz. En 1988, Lieberman estudió la fisiología sobre la sensibilidad del oído humano, explicando que el oído humano puede escuchar sonidos con frecuencias altas hasta 20 KHz, por ello mayormente se trabaja en reconocer individuos en el dominio de la frecuencia y usando escalas logarítmicas como las de Mel o Bark (Lieberman y Blumstein, 1988).

En los años 90's se buscó que varias de las técnicas ya utilizadas se volvieran más robustas respecto a los problemas derivados de la grabación de sonidos de fondo, voces individuales, calidad de los micrófonos, canales de transmisión, reverberación de los cuartos, etc; éstas técnicas incluían la regresión lineal de máxima verosimilitud (MLLR), descomposición del modelo, combinación paralela del modelo (PMC), método estructural máximo a posteriori (SMAP), entre otros (Shinoda y Lee, 2001). Este tipo de tecnologías se utilizaron cada vez más por las compañías telefónicas para automatizar los servicios del operador.

Durante las últimas décadas se ha realizado más investigación así como propuestas de técnicas para el reconocimiento de individuos orientada a la seguridad, especialmente en el mercado empresarial. En 1995, el error del análisis de predicción lineal, que se sabe está relacionado con la forma del pulso glotal, se utilizó para la verificación del hablante independiente del texto (Thévenaz y Hügli, 1995). En el Análisis de Predicción Lineal, el tracto vocal se modela como un filtro digital dinámico de puros polos cuyos parámetros, conocidos como Coeficientes de Predicción Lineal (LPC), se determinan al minimizar el error entre la señal de voz original y la señal de voz como se produce por el modelo del tracto vocal.

1.3. Objetivos de la Tesis

1.3.1. Objetivo general

Implementar un reconocedor de individuos independientemente del texto pronunciado, obteniendo los primeros tres formantes de las vocales articuladas en la señal de voz del parlante, utilizando al menos 3 de las vocales pronunciadas, generando un vector de características de cada individuo de la base de datos para su identificación.

1.3.2. Objetivos particulares

- Implementar un segmentador de audio que separe sonidos vocalizados de sonidos no vocalizados encontrando la región de interés para hacer la extracción de características.
- Extraer características de los segmentos vocalizados con la finalidad de identificar segmentos que contienen vocales.
- Extraer los primeros tres formantes a partir de los coeficientes de predicción lineal de cada vocal pronunciada.
- Generar las combinaciones a partir de 3 vocales tanto para el idioma español como para el idioma inglés, con la finalidad de encontrar los mejores resultados.
- Con los formantes obtenidos y las combinaciones de 3 vocales, crear un vector de características de tamaño $3x3$, tres formantes por tres vocales, y generar un diccionario de características de individuos.
- Medir la similitud entre el vector de características del individuo desconocido con los vectores que pertenecen a los individuos conocidos en el diccionario de datos.
- Realizar diferentes pruebas de acuerdo al número de vocales pronunciadas así como cantidad de individuos para evaluar la efectividad del método.

1.4. Descripción de Capítulos

Esta tesis se organiza de la siguiente manera:

- El Capítulo 2 describe la anatomía de la voz; se hace un análisis de las características que integran la voz y cómo se produce esta dentro del tracto vocal, así como de los tipos de sonidos que existen en idioma español e inglés. Se explica la extracción de características por medio de los coeficientes de predicción lineal (LPC), así como la estimación de los formantes para las vocales que pronuncia un individuo.

- El Capítulo 3 explica la implementación del método, se detalla en este capítulo el procedimiento para la extracción de formantes utilizado y las combinaciones para diferentes número de fonemas pronunciados para la extracción de características tanto para el idioma español como para el idioma inglés.

- En el Capítulo 4 se presentan las pruebas y los resultados obtenidos para cada una de las cuatro bases de datos.

- En el Capítulo 5 señala las conclusiones y trabajo futuro.

Capítulo 2

Producción de la Voz

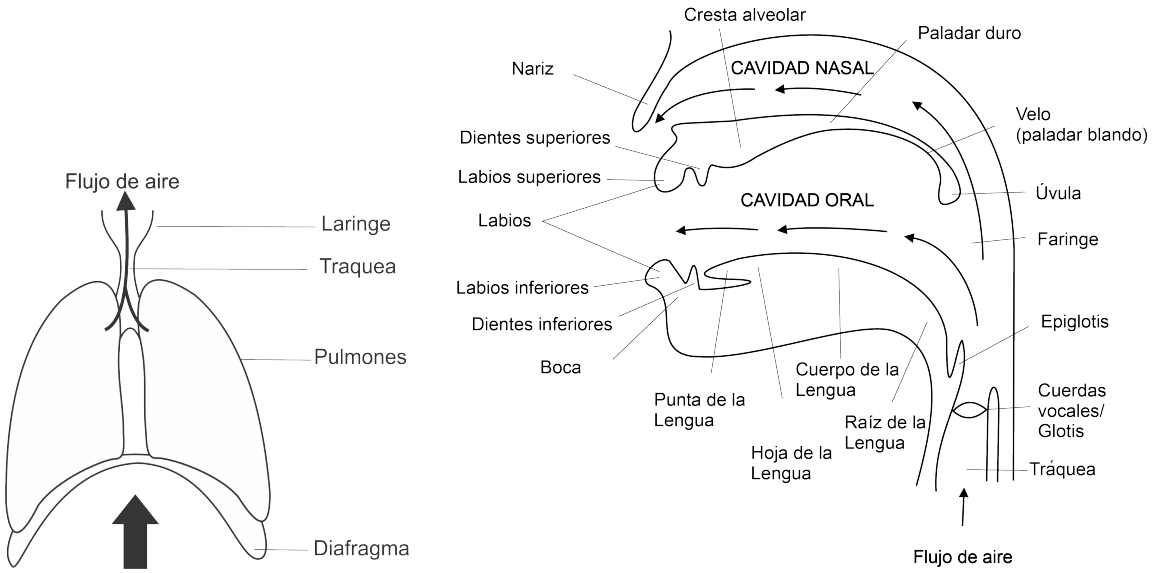
Este capítulo explica cómo se produce la voz así como los factores que permiten hacer identificables a los individuos. En la Sección 2.1 se analiza a detalle cómo se genera la voz mientras que en la Sección 2.2 se detalla la importancia de las vocales en los sonidos emitidos por un individuo.

La voz es la principal forma de comunicación entre los seres humanos; es una forma de comunicar nuestros pensamientos, conocimientos y sentimientos, siendo éste el medio más utilizado en nuestra vida diaria. De ésta se puede obtener información como identificación del idioma, emociones, género del individuo, volumen, timbre y reconocimiento de palabras (Zheng y Li, 2017). Ésta es la información que permite la identificación de individuos.

2.1. Anatomía de la voz

La voz se produce por el paso del aire que proviene de la caja torácica a través de las cuerdas vocales, su calidad depende de tres cualidades (Dias y cols., 2012): tono, volumen y timbre, los cuales son característicos de cada persona. La forma de la boca del individuo influye en cómo se genera la voz. Los sonidos vocalizados se producen cuando vibran las cuerdas vocales, producido por el paso del aire a través de éstas.

El órgano principal de la voz es la laringe, el flujo de aire se produce desde el diafragma y pasa por la tráquea como se muestra en la Figura 2.1 a) (Teng, 2016). La voz es el resultado de esta fuerza que se produce desde el diafragma y los músculos de la lengua y la laringe, que es donde se encuentran las cuerdas vocales. La intensidad de la voz depende de la presión del aire espirado.



(a) Flujo de aire a través del diafragma

(b) Aparato del tracto vocal humano

Figura 2.1: a) El flujo de aire se produce en el diafragma y pasa por la tráquea y la laringe, quien es el órgano principal para la producción de la voz (Teng, 2016). b) Aparato del tracto vocal humano (Teng, 2016), una vez que el flujo de aire pasa por la tráquea y la laringe llega a las cuerdas vocales, produciendo diferentes sonidos de acuerdo a la forma y posición de la boca, utilizando todos los músculos de la boca para la producción de la voz.

La Figura 2.1 b) se muestra el mecanismo de la producción de la voz en el ser humano. El aparato vocal consiste en tres diferentes sistemas: el aparato respiratorio, las cuerdas vocales y el tracto vocal. Para producir la voz se necesitan tres elementos, uno que permita generar una corriente de aire, el segundo que haga vibrar el aire y por último una cámara de resonancia, estos elementos se encuentran en el aparato vocal y son fuelle, vibrador y resonadores:

- Mancha o fuelle: formada por estructuras infraglóticas que determinan la mayor o menor presión del aire espirado que son: laringe, pulmones, tráquea y diafragma.
- Vibrador: formado por las cuerdas vocales de la laringe.
- Resonadores: formado por las cavidades supraglóticas donde el sonido producido es ampliado y modificado en las cuerdas vocales, y esta formado por la faringe, cavidad nasal y cavidad oral.

Todos los seres humanos poseemos diferentes características como tamaño de nuestra tráquea, grosor, forma de abrir y cerrar la boca, el movimiento de la lengua, entre otros rasgos. Todos estos rasgos influyen en la producción de la voz, por ello extrayendo ciertas características de la voz, propias de cada persona, es posible identificar a un individuo con solo su voz.

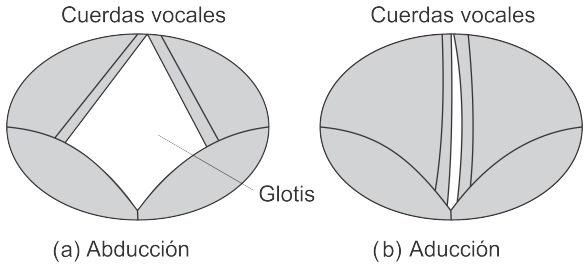


Figura 2.2: Cuerdas vocales (Dias y cols., 2012). a) Cuando las cuerdas vocales se abren se le llama abducción. b) Cuando las cuerdas vocales se cierran se le llama aducción.

Las cuerdas vocales no hacen vibrar el aire, sino que se crean oleadas de aire a través de abrir y cerrar la hendidura glótica, esta interrupción de flujo de aire provoca la vibración acústica. El sonido producido por las cuerdas vocales consta de una frecuencia fundamental y sus armónicos superiores. El tono aumenta cuando estos ciclos que producen las cuerdas vocales son más cortos y se repiten con mayor frecuencia. El tono de la voz está relacionado con la longitud y grosor de las cuerdas vocales de cada individuo. En la Figura 2.2 (Dias y cols., 2012) (a) se muestra cuando las cuerdas vocales se abren para hacer pasar el aire, a este proceso se le llama abducción que es la separación de las cuerdas vocales, mientras que cuando se cierran se le llama aducción (b).

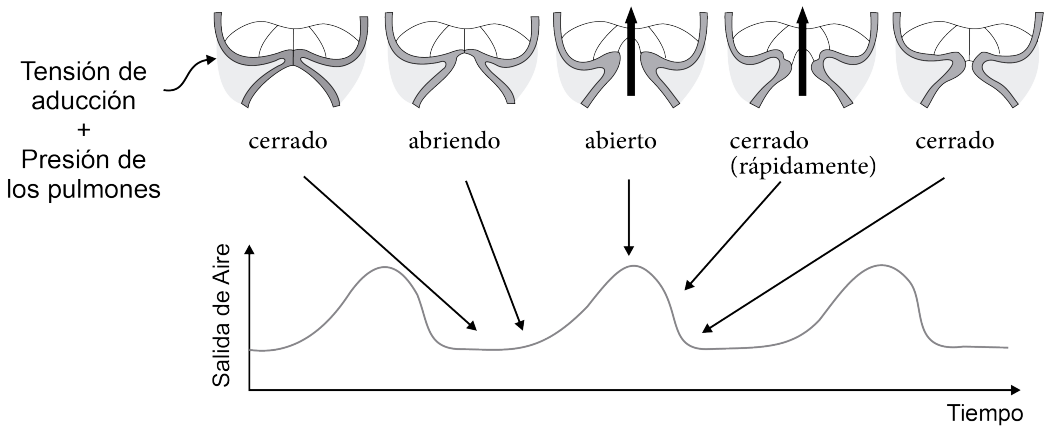
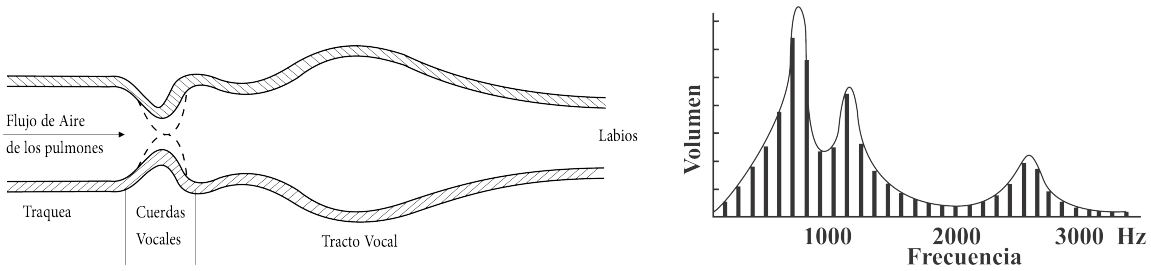


Figura 2.3: Fase glotal (Hirano, 1981). La presión del aire a través de los pulmones genera una tensión en las cuerdas vocales creando que estas se abran y cierren, cuando este proceso ocurre a intervalos idénticos se produce un ciclo glotal, siendo este proceso donde se genera la voz.

Cuando las cuerdas vocales se abren y cierran a idénticos intervalos ocurre un ciclo llamado ciclo glotal, como se muestra en la Figura 2.3 (Hirano, 1981). Cada ciclo está compuesto por 4 fases glotales que son: cerrado, abriendo, abierto y cerrado (rápidamente), donde el aire para a través de las cuerdas vocales y es donde se genera la voz.

2.2. Características del Tracto Vocal

El tracto vocal que produce la voz empieza desde el diafragma, donde se origina el flujo de aire que llega hacia la tráquea desde los pulmones, pasa a través de las cuerdas vocales y termina en los labios. La forma y tamaño del tracto vocal determinan las frecuencias de resonancia, como se muestra en la Figura 2.4 (Rabiner y Schafer, 1978), estas frecuencias de resonancia (b) son frecuencias que se producen por la emisión de sonidos de acuerdo al tamaño y forma del tracto vocal (a), el velo del paladar, así como la lengua, que influyen en cambiar la abertura o cierre del tracto vocal, produciendo así frecuencias naturales que modifican la emisión de los sonidos. Cuando una persona emite un sonido con una frecuencia fundamental de 330 Hz, sus cuerdas vocales abren y cierran 330 veces por segundo (Dias y cols., 2012).



(a) Tracto vocal

(b) Frecuencias de Resonancia del Tracto vocal

Figura 2.4: Tracto vocal (Rabiner y Schafer, 1978). La forma del tracto vocal influye en las frecuencias de resonancia que producen la voz del individuo. Una vez que el flujo de aire pasa de los pulmones, por la tráquea y la laringe y llega a las cuerdas vocales. El tracto vocal produce cinco resonancias importantes, que son las que definan la calidad y tipo de vocal, también llamados formantes.

El tracto vocal posee cinco resonancias importantes, llamados formantes, que definen la calidad y tipo de vocal cuando la onda es audible. Los órganos articulatorios del tracto vocal influyen en la emisión de los sonidos vocálicos; cada sonido emitido por la voz tiene distinta forma de pronunciación en el tracto vocal. Como se puede observar en la Figura 2.5 (Teng, 2016), al pronunciar la palabra *kat* ocurre un cierre del flujo de aire en las consonantes /k/ y /t/, mientras que en /æ/ el flujo de aire pasa a través de la cavidad oral y sale por los labios.

En la Figura 2.6 se muestran los diferentes ciclos producidos por una sola persona pronunciando las cinco vocales del lenguaje español tomando como muestra solo 50 milisegundos de cada señal de voz, lo mismo para la Figura 2.8 para las diez vocales del idioma inglés. Se puede apreciar que cada vocal produce diferentes ciclos. Incluso en un espectrograma[1], como se muestra en la Figura 2.7, es posible identificar una vocal de otra, mientras que una señal en el dominio del tiempo es más complejo identificarla.

[1]Función bidimensional mostrada como imagen en escala de grises, donde en el eje vertical muestra las variaciones de la frecuencia y la amplitud, mientras que en el eje horizontal representa la señal sonora a lo largo del tiempo (Rabiner y Schafer, 2011)

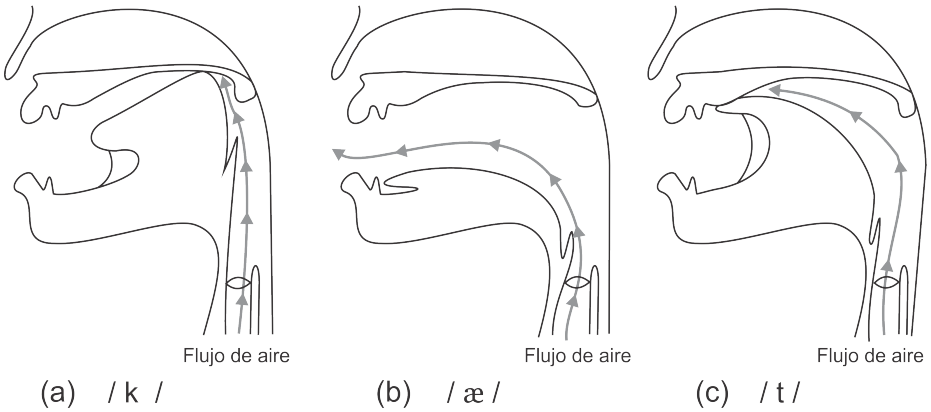


Figura 2.5: Flujo de aire para la palabra *kat* (Teng, 2016). Cuando el flujo de aire pasa a través de las cuerdas vocales y ocurre un cierre en el paso del aire en consonantes como la /k/ ó /t/, en la figura izquierda y derecha, se puede observar que este flujo de aire no pasa por el tracto vocal generando sonidos no vocalizados, mientras que en la pronunciación de la /æ/ pasa el flujo de aire a través del tracto vocal saliendo por los labios, cuando esto ocurre se generan vibraciones en las cuerdas vocales que son los sonidos vocalizados.

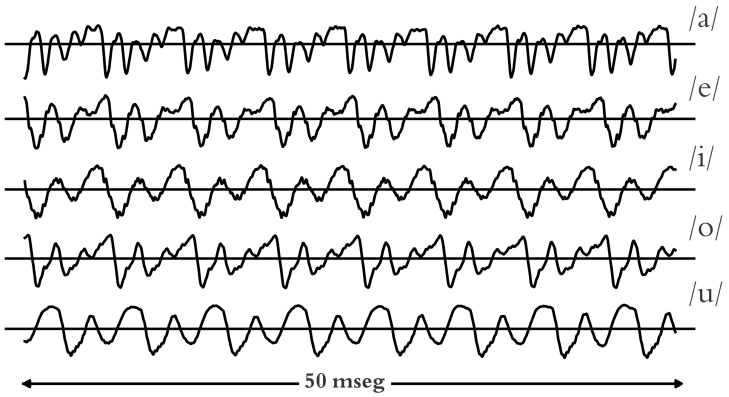


Figura 2.6: Ciclos producidos por las cinco vocales del lenguaje español. Para el idioma español, donde se tienen 5 vocales, se producen diferentes ciclos dependiendo de la vocal pronunciada, vistas desde el dominio del tiempo, siendo cada una de ellas diferente una entre otra en la misma persona, y entre personas. (Figura realizada con voz propia).

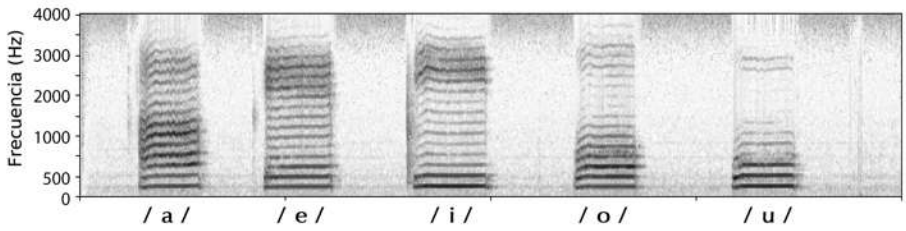


Figura 2.7: Espectrograma de las cinco vocales pronunciadas del idioma español. En un espectrograma se puede visualizar los formantes de cada vocal, para la identificación de estos, donde las franjas obscuras es donde se encuentran los formantes que definen el tipo de vocal. (Figura realizada con voz propia)

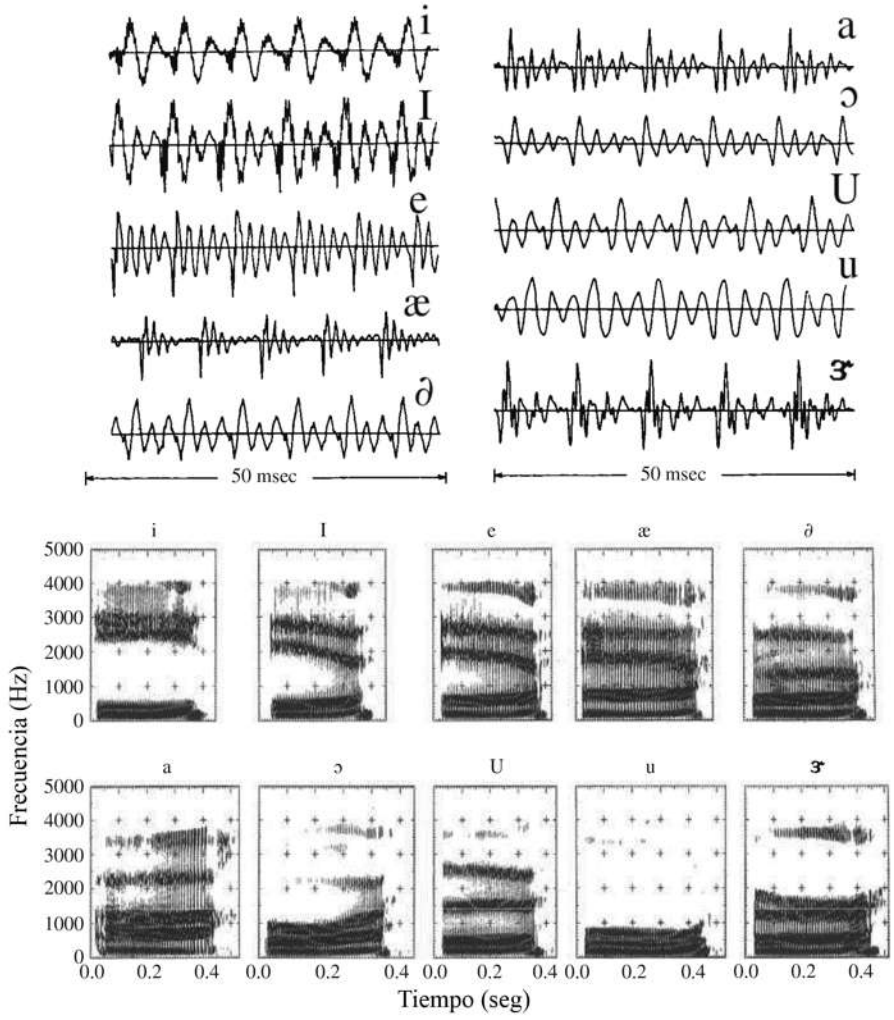


Figura 2.8: Ciclos y espectrogramas producidos por las diez vocales del lenguaje inglés. Para el idioma inglés, donde se tienen 10 vocales, se producen diferentes ciclos dependiendo de la vocal pronunciada, vistas desde el dominio del tiempo, siendo cada una de ellas diferente una entre otra en la misma persona. (Rabiner y Schafer, 2011)

Los sonidos fricativos se producen por la fricción del aire al pasar entre dos órganos bucales, los sonidos oclusivos se producen por el cierre brusco o momentáneo de alguna parte de la boca que impide la salida de aire, los sonidos africados son el resultado de la articulación de la fricción y oclusión y los sonidos nasales se articulan dejando salir aire expirado por la nariz. Las Figuras 2.9 (Rabiner y Schafer, 1978) y 2.10 (Quilis, 1980) muestran los fonemas que se presentan para el lenguaje inglés y español respectivamente, donde se establecen 10

vocales para el idioma inglés y 5 para el idioma español.

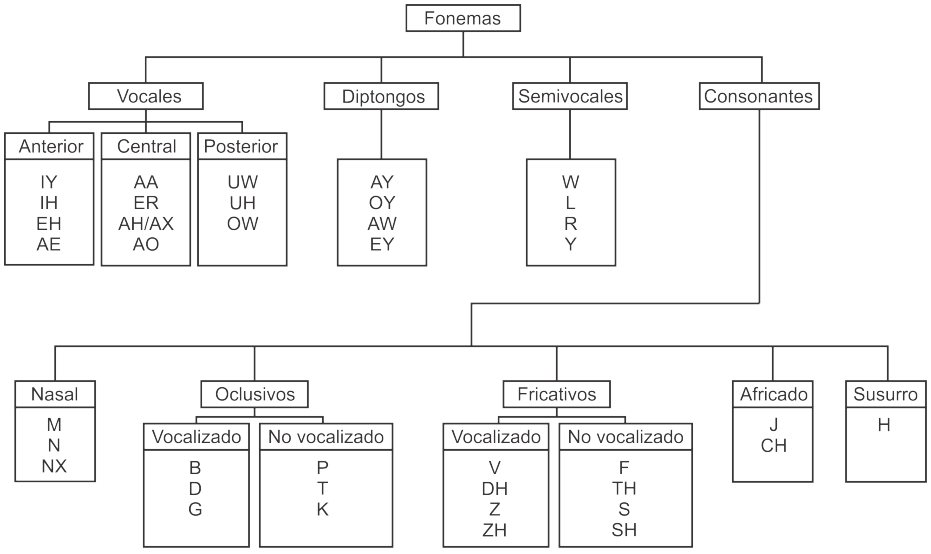


Figura 2.9: Fonemas del idioma inglés (Rabiner y Schafer, 2011)

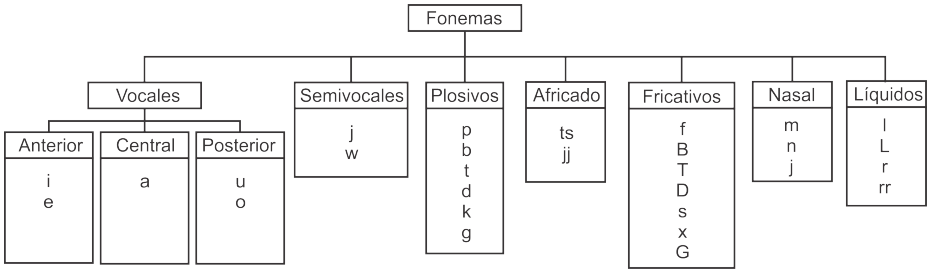
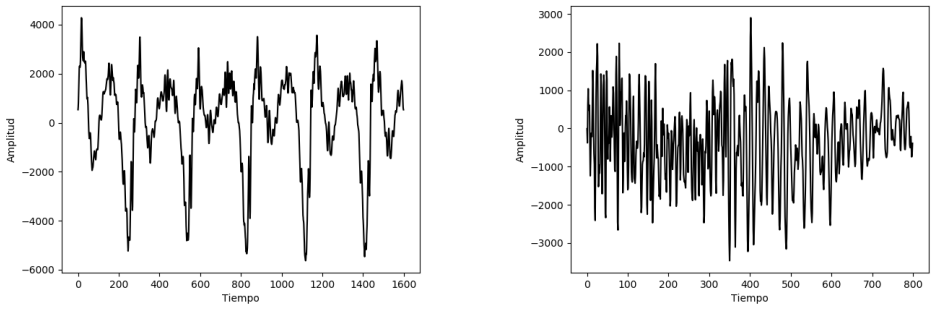


Figura 2.10: Fonemas del idioma español mexicano (Veiga, 2002)

En el idioma español existen cinco fonemas que se conocen como vocales: /a/, /e/, /i/, /o/, /u/. Mientras que Rabiner (Rabiner y Schafer, 1978) señala el sistema vocálico inglés americano con diez fonemas vocálicos: /IY (i)/, /I (ɪ)/, /E (ɛ)/, /AE (æ)/, /UH (ʌ)/, /A (a)/, /OW (ɔ)/, /U (ʊ)/, /OO (u)/, /ER (ɝ)/.

Los sonidos vocalizados son señales pseudo-periódicas en el tiempo, cuando una vocal es pronunciada, tanto para el idioma español como para el inglés, el flujo de aire pasa



(a) Señal de sonido vocalizado (b) Señal de sonido no vocalizado

Figura 2.11: Señal de voz. Una señal de un sonido vocalizado presenta cierta periodicidad en la señal (a), mientras que un sonido no vocalizado se produce sin vibración en las cuerdas vocales (b).

a través de la cavidad vocal hasta salir por los labios. Estos sonidos son resultado de las vibraciones de las cuerdas vocales llamados sonidos vocalizados. En contraste, los sonidos no vocalizados se producen sin la vibración de las cuerdas vocales y por una excitación glótica ruidosa (Dias y cols., 2012), tal como se muestra en la Figura 2.11.

Al pronunciar una vocal, cualquiera que fuese, se produce una vibración mayor en las cuerdas vocales a diferencia de las consonantes; el vocabulario de las personas esta compuesto por consonante y vocal el cual forman sílabas y estas forman palabras. Toda sílaba va acompañada forzosamente de una vocal, es por ello que se eligieron las vocales para identificar a un individuo, ya que en una frase pronunciada se pueden repetir varias veces cualquiera de las vocales, aunque no pronunciamos dos veces igual la misma vocal en una oración, sin embargo encontraremos similitudes entre una u otra, que se detallarán más adelante.

Capítulo 3

Procesamiento de la señal de voz

En este capítulo se analizará en la Sección 3.1 el preprocesamiento de una señal de voz para enfatizar esta, en la Sección 3.2 se describe la función de autocorrelación con la cual se estimaran los coeficientes LPC, en la Sección 3.3 se analiza la obtención de los coeficientes PARCOR utilizando el algoritmo Levinson-Durbin, en la Sección 3.4 se explica cómo a partir de lo anterior se obtienen finalmente los formantes de las vocales y finalmente en la última sección de este Capítulo se analizará el uso de los formantes obtenidos de cada vocal.

Como se mencionó anteriormente, el sonido de la voz se genera a través del tracto vocal. De acuerdo a Stanley (Stanley y Watkins, 1939), los órganos articulatorios que forman el tracto vocal (labios, mandíbula, lengua y velo del paladar) concentran la energía en frecuencias actuando como resonadores. Estas frecuencias de resonancia se conocen como formantes (F_k). Gracida y Orduña (Gracida y Orduña, 2011) señalan que los primeros dos formantes (F_1 y F_2) los que permiten la identificación de las vocales, mientras que los formantes F_3, F_4 y F_5 determinan el color de la voz, describiendo cada uno de ellos como:

- F_1: este formante depende de la forma de la cavidad de la faringe; entre más se estrecha, mayor frecuencia se produce.
- F_2: este formante depende de la posición de la lengua; si se eleva en la parte anterior de la boca la frecuencia sube, y en la parte posterior la frecuencia baja.

- F_3: este formante se encuentra relacionado con la posición de los labios, si están estirados su frecuencia es más alta, y si se encuentran redondeados su frecuencia es mas baja.

- F_4 y F_5: estos formantes varían con la anchura y longitud del tracto vocal, si es corto y estrecho el tracto vocal los formantes tienen frecuencias más agudas.

La estimación de los formantes de la producción de la voz se realiza en 4 etapas: la primera es el preprocesamiento de la señal de voz, con el objetivo de enfatizar el espectro de la señal y compensar la caída de -6 dB que presenta la señal de voz al pasar por el tracto vocal; la segunda etapa es la obtención de los coeficientes de predicción lineal obtenidos a partir de la función de autocorrelación, donde se utiliza el algoritmo Levinson-Durbin para dar solución al sistema de ecuaciones generado; la tercera etapa es la estimación de los formantes a partir de la obtención de los coeficientes LPC de la señal de voz. Por último, de acuerdo a las frecuencias obtenidas de los primeros tres formantes, se identifican las vocales pronunciadas por el individuo.

3.1. Preprocesamiento de la Señal de Voz

Con el propósito de enfatizar el espectro de la señal de voz, ésta se procesa previamente mediante técnicas de procesamiento de señales, como:

1. Pre-enfatización: la señal está sujeta a un filtro de énfasis que es realmente un filtro de respuesta de impulso finito, cuya salida y se obtiene de su entrada x como:

$$y(n) = x(n) - ax(n-1) \tag{3.1}$$

2. Enmarcado: Después del pre-énfasis y como parte del procesamiento previo, se encuadra la señal de voz utilizando en marcos de 30 milisegundos superpuestos en 20 milisegundos; el comienzo de un marco se produce 10 milisegundos después del inicio

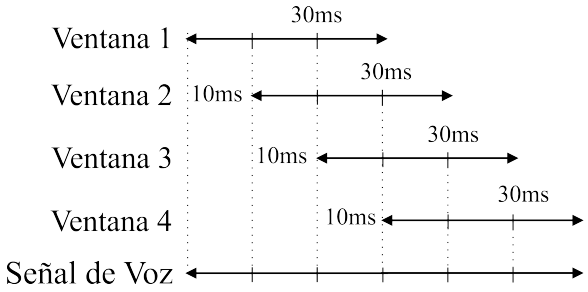


Figura 3.1: Enmarcado de una señal de audio de voz con ventanas de 30 milisegundos y traslape de 10 milisegundos.

de su anterior marco, como se muestra en la Figura 3.1.

A cada marco se aplica la ventana de Hamming, que se define como:

$$w(n) = 0.54 + 0.46\cos(2\pi n/N) \tag{3.2}$$

donde N es el tamaño de la ventana de muestras. En la Figura 3.2 se muestra el enmarcado del tamaño de la señal de voz, donde los extremos de la señal se atenúan.

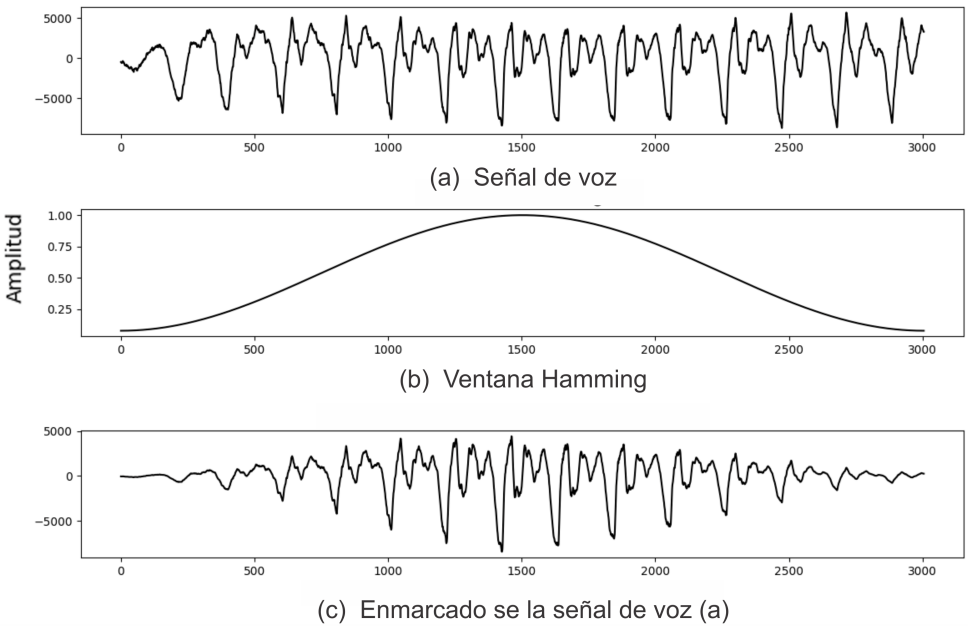


Figura 3.2: a) Señal de audio de voz; b) Ventana de Hamming del tamaño de la señal de audio de voz; c) Aplicación de la ventana de Hamming

3.2. Autocorrelación de la Señal de Voz

Una vez que se tiene la señal de voz procesada, se utiliza la función de autocorrelación, la cual contiene la energía y periodicidad de la señal. Esta periodicidad puede encontrarse en el primer máximo de la función de autocorrelación $0, \pm N_p, \pm 2N_p, \ldots$. La autocorrelación de una señal se define como:

$$\phi(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k) \tag{3.3}$$

Si la señal es periódica con periodo N_p muestras, entonces:

$$\phi(k) = \phi(k + N_p) \tag{3.4}$$

La función de autocorrelación se puede utilizar como un estimador de la periodicidad; además la energía de la señal es igual a $\phi(0)$. La función de autocorrelación de corto tiempo (SACF por sus siglas en inglés, Short-time Autocorrelation Function) $Rn(k)$ de un marco que comienza en la muestra n de una señal x se define como en la Ecuación 3.5. Se llama tiempo corto porque se determina para un solo marco de la señal de voz, donde N es la longitud del marco.

$$R_n(k) = \sum_{m=0}^{N-k} x(n+m)x(n+m+k) \tag{3.5}$$

La función de autocorrelación modificada de tiempo corto (MSACF por sus siglas en inglés, Modified Short-time Autocorrelation Function) se define como:

$$\hat{R}_n(k) = \sum_{m=0}^{N-1} x(n+m)w_1(m)x(n+m+k)w_2(m+k) \tag{3.6}$$

La diferencia entre $R_n(k)$ y $\hat{R}_n(k)$ es que la última no está limitada a su marco, sino que utiliza muestras del siguiente marco. $\hat{R}n(k)$ es una función de correlación cruzada entre dos marcos consecutivos, entre el segmento $x(n+m)w_1(m)$ y el segmento $x(n+m)w_2(m)$. La SACF sufre el hecho de que cuanto mayor sea el valor de k se suman las muestras menores de la trama y, por lo tanto, se aproxima a cero a medida que k se acerca a N. MSACF resuelve ese problema que es importante cuando la función de autocorrelación se usa para

la estimación de periodicidad. Para decidir si un marco contiene sonidos vocalizados o no vocalizados, se determina el SMACF y luego se busca el pico más grande. Si ese pico es mayor que un umbral predefinido, el marco se declara como vocalizado, por supuesto $\hat{R}_n(0)$ no se considera aquí, ya que es realmente el contenido de energía en el marco que comienza en la muestra n de la señal x y ningún otro elemento del SMACF podría ser mayor que eso.

3.3. Codificación Lineal Predictiva (LPC)

Una de las técnicas más utilizadas en el procesamiento de la señal de voz es el análisis de predicción lineal, además es una de las técnicas más precisas de los parámetros de la voz y es de bajo costo computacional. Los coeficientes de predicción lineal (LPC por sus siglas en inglés, Linear Prediction Coefficients) predicen una señal en el dominio del tiempo con base a las muestras anteriores, de α_k desde 1 hasta p, donde p es el orden de predicción. El tracto vocal esta modelado por un filtro de puros polos, un predictor lineal predice p muestras de voz en base a p muestras anteriores.

Para obtener los coeficientes de predicción lineal se debe calcular $\phi_n(i,k)$ y resolver el sistema de ecuaciones que encuentra los coeficientes α_k:

$$\sum_{k=1}^{p} = \alpha_k \phi_n(i,k) = \phi_n(i,0) \qquad\qquad i = 1,2,\ldots,p \tag{3.7}$$

Para la Ecuación de autocorrelación (3.5), el conjunto de ecuaciones se satisfacen por los coeficientes predictores tal que:

$$\sum_{k=1}^{p} \alpha_k R[|i-k|] = R[i], \qquad\qquad 1 \leq i \leq p \tag{3.8}$$

La ecuación (3.8) se puede representar por $R\alpha = r$ donde R es una matriz de Toeplitz definida por el $(i,j)^{th}$ elemento $R[|i-k|]$, α y r son vectores de columnas con elementos α_i.

El error mínimo cuadrado del p^{th} orden del predictor está dado por:

$$R[0] - \sum_{k=1}^{p} \alpha_k R[k] = \varepsilon^{(p)} \tag{3.9}$$

Para encontrar los coeficientes LPC se debe resolver el sistema de ecuaciones de la función de autocorrelacion (3.8) donde los coeficientes predictores satisfacen la ecuación:

$$R[i] - \sum_{k=1}^{p} \alpha_k R[|i-k|] = 0, \qquad i = 1, 2, \ldots, p. \tag{3.10}$$

Para estimar los valores de correlación se resuelve:

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & \ldots & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & \ldots & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & \ldots & R_n(p-3) \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ R_n(p) & R_n(p-1) & R_n(p-2) & \ldots & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ \vdots \\ R_n(p) \end{bmatrix} \tag{3.11}$$

La matriz de coeficientes de la Ecuación (3.11) es una matriz de Toeplitz, dado que es una matriz simétrica y tiene los mismos valores en cualquier diagonal, por ello se puede resolver recursivamente con el algoritmo Levinson-Durbin (Rabiner y Schafer, 2011), donde un nuevo valor de correlación para cada iteración resolviendo para el siguiente orden superior del predictor en términos del nuevo valor de correlación y el predictor encontrado anteriormente (Rabiner y Schafer, 2011), de la cual obtendremos un polinomio con la solución a la inversión (A), el error del predictor (e) y los coeficientes reflejantes (k), a estos coeficientes reflejantes se les conoce también como coeficientes PARCOR, dado que son una medida de autocorrelación parcial (de ahí su nombre, PARtial CORrelation), entre el error de predicción hacia adelante y el error de predicción hacia atrás. Los coeficientes LPC se pueden obtener a partir de los coeficientes PARCOR mediante:

$$a_i^{(i)} = k_{(i)} \tag{3.12}$$

$$a_j^{(i)} = k_i \alpha_{i-j}^{(i-1)} \tag{3.13}$$

para todo $i = 1, 2, \ldots, p$, estos coeficientes son:

$$\alpha_j = \alpha_j^{(p)} \tag{3.14}$$

Para solucionar un sistema de ecuaciones de orden $p = 2$:

$$\begin{bmatrix} R_n(0) & R_n(1) \\ R_n(1) & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1^{(2)} \\ \alpha_2^{(2)} \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \end{bmatrix} \tag{3.15}$$

Solucionando el sistema de ecuaciones para $\alpha_2(2)$, se tiene:

$$\alpha_2^{(2)} = \frac{R_n(0)R_n(2) - R_n(1)^2}{R_n^2(0) - R_n^2(1)} \tag{3.16}$$

Dividiendo el numerador y denominador entre $R_n(0)$ se tiene que:

$$\alpha_2^{(2)} = \frac{\frac{R_n(2) - R_n(1)^2}{R_n(0) - R_n^2(1)}}{R_n(0)} \tag{3.17}$$

Sustituyendo el resultado con el predictor de primer orden se tiene:

$$\alpha_2^{(2)} = \frac{R_n(2) - R_n(1)\alpha - 1^{(1)}}{R_n(0) - R_n(1)\alpha - 1^{(1)}} \tag{3.18}$$

Se tiene que $E_n^{(1)} = R_n(0) - \alpha_1^{(1)} R_n(1)$ por ello:

$$\alpha_2^{(2)} = \frac{R_n(2) - R_n(1)\alpha_1^{(1)}}{E_n^{(1)}} \tag{3.19}$$

Una vez que se conocen los coeficientes del orden inferior, se realiza lo mismo para obtener los coeficientes del siguiente orden, así sucesivamente hasta el orden deseado.

Para calcular los coeficientes predictores del orden p deseado, se calculan los coeficientes de todos los predictores de orden inferior a p. Esto permite ir calculando el error del predictor (e) y los coeficientes de autocorrelación parcial (k).

3.4. Estimación de los formantes

Para determinar los formantes se puede usar la Transformada Discreta de Fourier y luego ubicar los picos del espectro de potencia. Sin embargo, en la práctica no es tan simple, ya que hay muchas fluctuaciones en el espectro de potencia que podrían confundirse

con los formantes. Para ubicar los formantes del análisis LPC se puede utilizar el método descrito a continuación:

1. Se obtienen los coeficientes LPC de acuerdo al orden del modelo (utilizando la regla de dos veces el número de formantes deseados más 2), obteniendo las raíces del polinomio de predicción.

2. Se determinan los ángulos correspondientes de las raíces, de acuerdo a los coeficientes obtenidos.

3. Las frecuencias se convierten en radianes/muestra representada por los ángulos a Hz y se calculan los anchos de banda de los formantes.

4. Los anchos de banda de los formantes están representados por la distancia de los polos de predicción desde el círculo unitario.

Algunas raíces del polinomio de predicción $A(z)$ corresponden a las frecuencias de los formantes de los sonidos del tracto vocal. Como ejemplo en la Figura 3.3 se muestra la localización de los polos (x) del p-ésimo orden de ceros el cual se indica con un "$\circ$"para un sistema de orden $p = 12$. En estas raíces se forman los polos corresponden a los formantes que están más cerca del círculo unitario.

Los polos producen picos similares a la resonancia cuando se evalúa en el círculo unitario, tal como se muestra en la Figura 3.4, donde la frecuencia correspondiente a los polos-ceros que se encuentran cerca del círculo unitario. Éstos se encuentran etiquetados en orden del ángulo creciente en el plano z, donde los formantes se basan en la cercanía de la raíz compleja al círculo unitario.

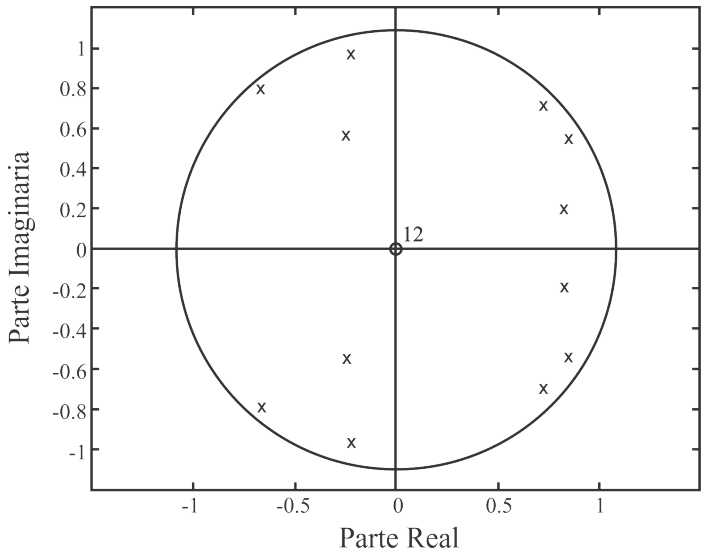


Figura 3.3: Localización de los polos y ceros en el plano z del modelo del tracto vocal para $p = 12$ (Rabiner y Schafer, 2011). Las raíces se producen en pares conjugados complejos, por lo que se ubican en un plano z de parte real y parte imaginaria, donde los anchos de banda de los formantes se representan por la distancia entre los polos y el círculo ∘ unitario.

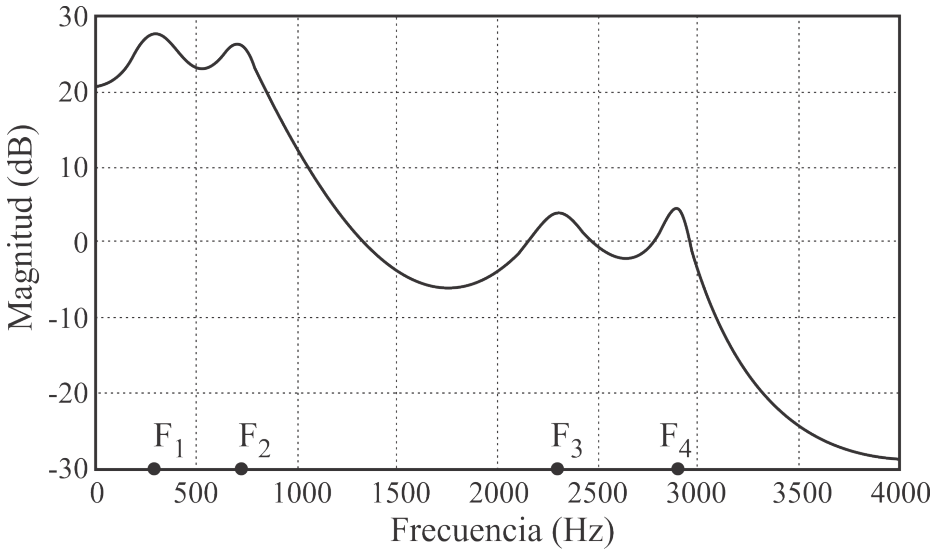


Figura 3.4: Respuesta a la frecuencia $\hat{H}(e^{jw})$ correspondiente a los polos y ceros de la Figura 3.3 (Rabiner y Schafer, 2011). Se obtiene la frecuencia correspondiente al ángulo de cada polo en el eje de la frecuencia, los cuales se asignan de acuerdo a la proximidad del polo complejo con el círculo unitario y la ubicación del polo en el intervalo de tiempo.

Tabla 3.1: Texto

Magnitud de la Raíz	θ angulo de la raíz	F angulo de la raíz en Hz	Formante
0.9308	10.36	288	F_1
0.9317	25.88	719	F_2
0.9109	82.58	2294	F_3
0.9571	104.29	2897	F_4

3.5. Análisis de formantes

Gracida y Orduña (Gracida y Orduña, 2011) señalan las regiones vocálicas del idioma español como se muestra en la Figura 3.5, donde se pueden observar las diferentes regiones para cada vocal respecto al primer y segundo formante, así como en la Tabla 3.2 se muestran los promedios de los primeros tres formantes para las cinco vocales del idioma español.

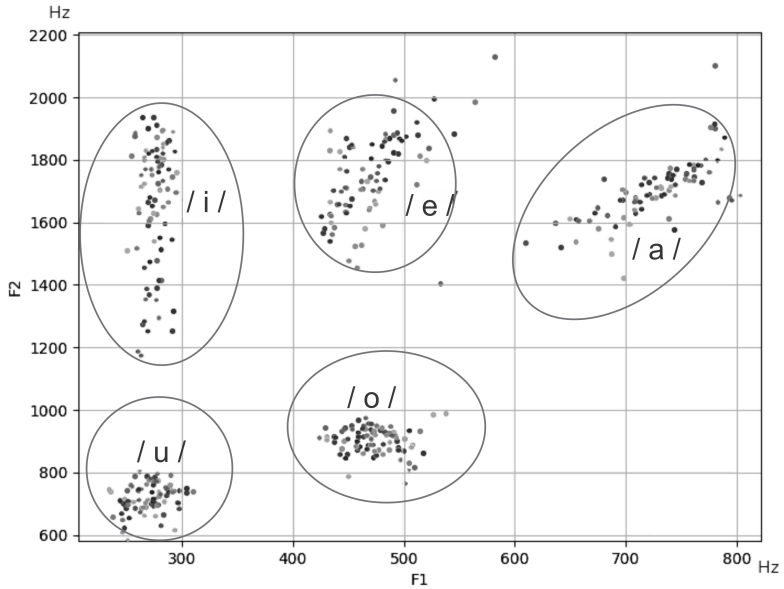


Figura 3.5: Regiones vocálicas del idioma español. Para el idioma español las cinco vocales se producen dentro de determinadas regiones, cada vocal del idioma español es identificable de acuerdo a sus dos primeros formantes.

Tabla 3.2: Promedios de fonemas vocálicos del idioma español

Vocal	F_1	F_2	F_3
/i/	300	2210	3220
/e/	520	2040	2796
/a/	720	1695	2813
/o/	470	900	2967
/u/	340	760	2250

Esto nos muestra que es posible identificar cada vocal usando los primeros 2 formantes. Los diez sonidos vocálicos del idioma inglés también cuentan con regiones que pueden ser identificadas por sus formantes como se muestra en la Figura 3.6 así mismo genera una Tabla 3.3 con los promedios de los diez sonidos vocálicos de los primeros tres formantes (Peterson y Barney, 1952).

Tabla 3.3: Promedios de fonemas vocálicos del idioma inglés (Peterson y Barney, 1952)

ARPAbet	IPA	F_1	F_2	F_3	ARPAbet	IPA	F_1	F_2	F_3
IY	i	270	2290	3010	AA	ɑ	730	1090	2440
IH	ɪ	390	1990	2550	AO	ɔ	570	840	2410
EH	ɛ	530	1840	2480	ER	ɝ	490	1350	1690
AE	æ	660	1720	2410	UH	ʊ	440	1020	2240
AH	ʌ	520	1190	2390	UW	u	300	870	2240

Las vocales son los mejores representantes de los sonidos sonoros, se han estudiado durante muchos años; en 1952, Peterson y Barney (Peterson y Barney, 1952) tomaron medidas de la frecuencia de los formantes de las vocales para varios hablantes masculinos y reportaron el promedio en la Tabla 3.3. Aún cuando existe una gran variabilidad en estos formantes, los promedios de Peterson y Barney se usan ahora para sintetizadores para producir vocales. Tomando la primera frecuencia de formantes como la coordenada horizontal y la segunda frecuencia de formantes como la coordenada vertical se pueden trazar en un plano. Las frecuencias de los formantes varían de un hablante a otro, incluso

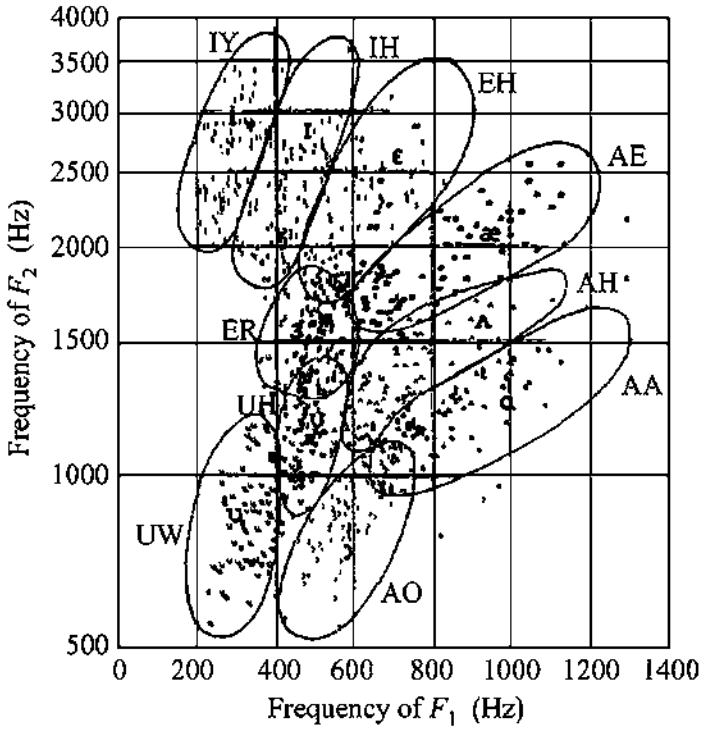


Figura 3.6: Vocales del idioma inglés en el espacio (F1,F2) (Peterson y Barney, 1952). Para el idioma inglés las diez vocales se producen dentro de determinadas regiones.

varían entre las expresiones del mismo hablante, sin embargo, la variabilidad entre los parlantes es mayor que la variabilidad entre las locuciones del mismo hablante. s

Capítulo 4

Implementación

En esta sección se explica el procedimiento utilizado de acuerdo a lo analizado en el Capítulo 3, donde para cada grupo de vocales por un mismo individuo se forman nubes de puntos en un plano que representan al individuo, dependiente del número de vocales utilizadas es el número de nubes de puntos que se obtendrá.

Primeramente se descartan los contenidos con sonidos sin voz y sonidos que no correspondían a vocales. Para ello se implementó un reconocedor de vocales utilizando una base de datos auxiliar con veintiún personas pronunciando las vocales en español. Por cada ventana que correspondía a una vocal, tanto del idioma español como del idioma inglés, se aplicó una ventana Hamming y un filtro pre-énfasis. Posteriormente se calcularon los coeficientes LPC y de ahí se obtienen los formantes. Posteriormente se agruparon los marcos que corresponden a la misma vocal y se determinaron las primeras tres frecuencias de formantes para cada marco que contiene la misma vocal. Cada uno de estos marcos puede considerarse como puntos de un grupo en un espacio tridimensional, esas tres dimensiones corresponden a las frecuencias de los tres primeros formantes. En seguida se calcula el centro del grupo y se hace lo mismo con las otras vocales. Después de este proceso, la grabación de algunos parlantes se ha convertido en un conjunto de puntos, por lo que el hablante está representado por esta nube. Dicha nube de puntos representa mejor al hablante cuando hay al menos una vocal de cada tipo en su correspondiente señal de voz. La base de datos con

una grabación de voz por individuo conocido se convierte luego en una colección de nubes de puntos por hablante. El problema de identificación del hablante independiente del texto se ha convertido en un problema de comparación de nubes de puntos que buscan la nube de puntos más similar a una nube de puntos determinada.

El algoritmo implementado para obtener los formantes es:

```
import numpy as np
import math
from scipy.signal import hamming
from scikits.talkbox import lpc

def formantes(sn, fs, p):
w = np.hamming(len(sn))
x1 = sn * w
preemphasis = 0.97
x = np.append(x1[0], x1[1:] - preemphasis * x1[:-1])
A, e, k = lpc(x, p)
rts = np.roots(A)
rts = [r for r in rts if np.imag(r) >= 0]
angz = np.arctan2(np.imag(rts), np.real(rts))
freqs = angz * (fs / (2 * math.pi))
frqs, indx = sorted(freqs), sorted(range(len(freqs)),
key=lambda k: freqs[k])
return frqs
```

La función recibe la señal *sn* de la cual se requiere obtener los formantes, la frecuencia de muestreo fs de la señal de audio, el número de coeficientes LPC p, recordando que se requiere dos veces mas dos el número de formantes que se desea obtener. Primeramente se aplica una ventana hamming y un filtro preenfasis para enfatizar la señal a analizar, una vez que se preprocesa la señal se obtienen los coeficientes de predicción lineal, los cuales regresan la solución a la inversión A, la predicción del error e y los coeficientes reflejantes k, para obtener los formantes se buscan las raices de los coeficientes A obtenidos y se calculan los ángulos, estos se convierten a Hz ya que será necesario ordenarlos del mas pequeño al más grande para poder obtener los formantes, esta función regresará $\frac{p}{2}$ número

de formantes, una vez obtenidos los formantes de un marco de la señal de voz se guarda en un diccionario de datos.

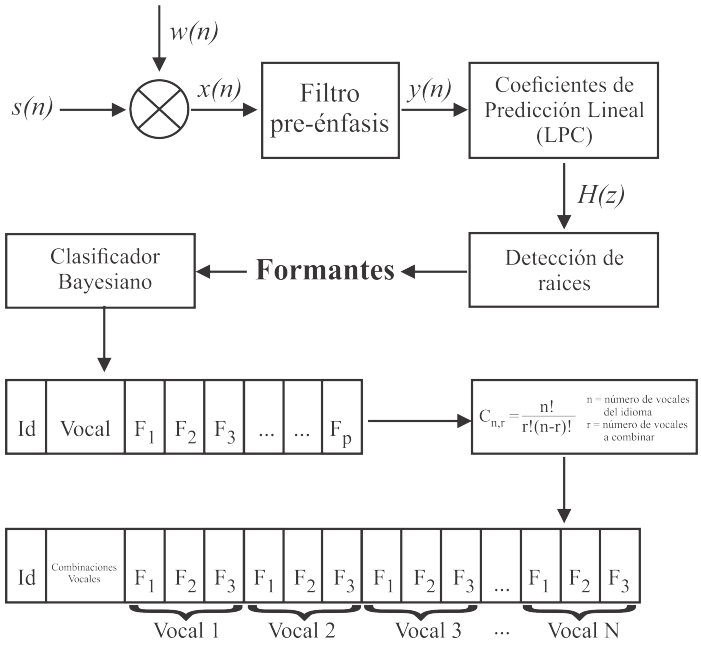


Figura 4.1: Procedimiento para la extracción de formantes. Primeramente se tiene la señal de voz a la que se le aplica un preprocesamiento que consiste en un filtro pre-énfasis y se aplica una ventana hamming para suavizar el espectro de la señal de voz, posteriormente se obtienen los coeficientes LPC con la función de autocorrelación de tiempo corto resolviendo el conjunto de ecuaciones con el algoritmo Levinson-Durbin. Una vez que se tiene el polinomio de predicción se obtienen las raíces del polinomio determinado por los ángulos correspondientes a cada raíz, obteniendo finalmente n formantes de acuerdo a p orden del predictor y se guardan en un diccionario de datos.

Una vez obtenidos todos los formantes se analizan mediante un clasificador bayesiano para clasificar qué vocal fué pronunciada, y se guarda la etiqueta de la vocal junto con sus formantes. Posteriormente se hacen todas las combinaciones de vocales tanto para 3, 4 y 5 del idioma español como para 3, 4, 5, 6, y 7 vocales del idioma inglés. Cabe mencionar que para el idioma inglés no se utilizaron combinaciones de 8, 9 o 10 ya que se detectó que en las oraciones se presentaban pocos casos donde se mencionaban mas de 8 vocales en la misma oración.

Capítulo 5

Pruebas y resultados

En este capítulo se describen las bases de datos utilizadas para la realización de las pruebas así como los resultados obtenidos utilizando diferente número de combinaciones de vocales.

5.1. Bases de datos utilizadas

Se utilizaron cuatro diferentes bases de datos para la identificación de individuos por su voz. La primera base de datos consta de 21 personas de ambos sexos, pronunciando 33 palabras en español en 4 ocasiones, siendo un total de 2,772 audios; dicha base de datos se encuentra en http://dep.fie.umich.mx/∼camarena/dsp/elocuciones21.tar.gz. La segunda base de datos conocidos como English language speech database for speaker recognition (ELSDSR), consta de 22 personas quienes pronuncian 9 diferentes frases en idioma inglés; contiene un total de 198 audios con duración entre 6 y 23 segundos; dicha base de datos se encuentra en http://www.imm.dtu.dk/lfen/elsdsr/index.php. La tercera base de datos utilizada es Diálogos Inteligentes Multimodales en Español (proyecto DIME), la cual consta de 100 personas quienes pronuncian 60 oraciones en español cada una de ellas; contiene un total de 6,000 audios, 51 mujeres y 49 hombres, en su mayoría estudiantes con una edad promedio de 24 años; esta base de datos se encuentra en http://turing.iimas.unam.mx/∼luis/DIME. La cuarta base de datos utilizada conocidos como TIMIT Acoustic-Phonetic Continuos Speech Corpus, consta de 630 personas de 8 regiones de los Estados Unidos, 438 hombres

y 192 mujeres, quienes pronuncian 10 oraciones en inglés, siendo un total de 6,300 audios; esta base de datos se encuentra en https://catalog.ldc.upenn.edu/LDC93S1. Estas características se encuentran descritas en las Tablas 5.1 y 5.2.

Tabla 5.1: Características de las bases de datos utilizadas

Base de datos	**Idioma**	**Individuos**	**Audios/individuo**	**Total audios**
Elocuciones21	Español	21	132	2772
ELSDSR	Inglés	22	9	198
DIMEx	Español	100	60	6000
TIMIT	Inglés	630	10	6300

Tabla 5.2: Características de los parlantes en las bases de datos utilizadas.

Base de datos	**fs**	**Género**	**Edad**	**Tiempo/audio**
Elocuciones21	8000 Hz	14 Hombres 7 Mujeres	ND	2 seg.
ELSDSR	16000 Hz	10 Hombres 12 Mujeres	24 a 63 años	17.6 seg
DIMEx	44100 Hz	49 Hombres 51 Mujeres	16 a 36 años	4 seg
TIMIT	16000 Hz	192 Hombres 438 Mujeres	ND	4seg

En la Tabla 5.3 se muestra que para las bases de datos en español, donde solo se tienen 5 vocales, se pueden obtener 10 diferentes triángulos de la nube de puntos por cada individuo, mientras que para las bases de datos del idioma inglés de las 10 vocales que se pronuncian se pueden obtener 120 triángulos. Se realizaron pruebas para 3, 4 y 5 vértices del idioma español, y 3, 4, 5, 6 y 7 vértices para el idioma inglés.

Para las bases de datos utilizadas, siendo Elocuciones 21 y DIMEx en idioma inglés

Tabla 5.3: Características de las bases de datos utilizadas

Idioma de las Bases de Datos	**Número de vértices**	**Número de combinaciones**
Español	3	10
	4	5
	5	1
Inglés	3	120
	4	210
	5	252
	6	210
	7	120

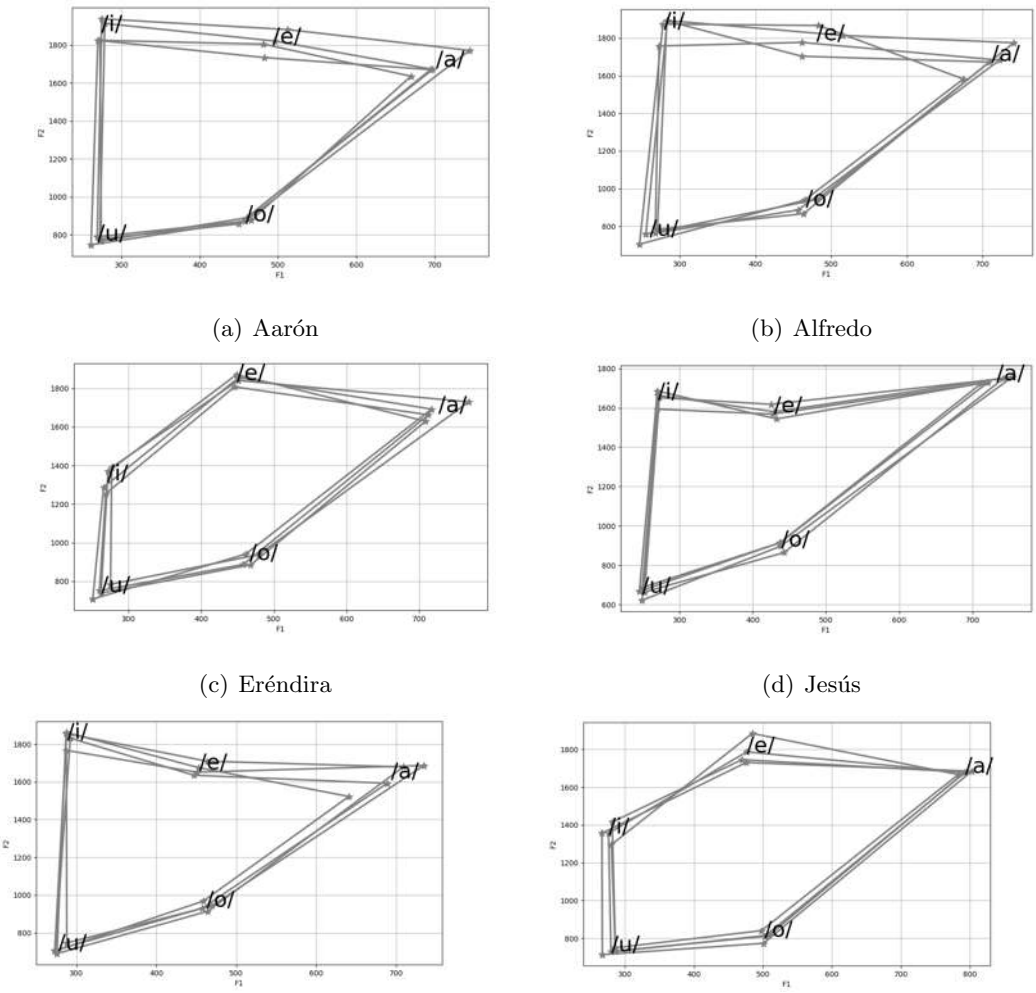
se tienen cinco vocales, dependiendo del número de vocales pronunciadas dentro de la señal de voz se tienen diferentes números de combinaciones. Se pueden pronunciar 10 diferentes combinaciones con solo 3 vocales, mientras que para las bases de datos en idioma inglés, ELSDSR y TIMIT, para 3 vocales de las 10 que tiene el lenguaje pueden tener 120 combinaciones.

La búsqueda de nubes de puntos similares donde el número de puntos varía y las nubes pueden estar sujetas a ruido (es decir, los puntos no están exactamente en la misma ubicación) se realiza buscando estructuras similares dentro de las nubes. Al seleccionar tres puntos de una nube de puntos, se determina un triángulo, cada punto de la nube está relacionado con una vocal específica, se tomaron todos los triángulos posibles de la nube de puntos, en lugar de seleccionar triángulos al azar. Como se puede observar en la Figura 5.1, para el idioma español, cada individuo forma un polígono de las cinco vocales, y estos no cambian demasiado mientras pronuncian en distintas ocasiones las vocales, lo mismo pasa para cada parlante de las bases de datos del idioma inglés, cada vértice es una vocal pronunciada, y dependiendo del número de vocales que sean pronunciadas es el número de vértices que se forman en la nube de puntos. En el caso de si el parlante pronunciara únicamente 3 de las 5 ó 10 vocales (para el idioma español e inglés, respectivamente), se

presenta como un triángulo, correspondiente a una tupla de 3 vocales ordenadas (en el diccionario). Estos vectores de características siempre se ordenan de modo que el vector de características sea el mismo independientemente del orden en que se seleccionen los vértices. De esta manera, por ejemplo, el triángulo e-a-o es el mismo que el triángulo a-e-o. Al buscar los vecinos más cercanos a K de un vector de características de un parlante de prueba en la colección de vectores de características de todos los parlantes conocidos, se determina la identidad del individuo utilizando la regla del vecino más cercano a K, que es la etiqueta de oradores más frecuente entre los vecinos más cercanos a K para determinar qué tan cerca están dos de estos vectores de características de n dimensiones, dependiendo del número de vocales que haya pronunciado.

De las pruebas realizadas se puede observar en la Figura 5.1, 6 de los 21 individuos de la base de datos utilizada, donde obteniendo los primeros dos formantes de cada vocal pronunciada por diferentes individuos, cinco vocales del idioma español, se obtienen puntos en un plano, siendo el primer formante en el eje x y segundo formante para el eje y donde cada punto de cada vocal es pronunciado en 4 ocasiones, se puede observar que existe una similitud entre cada una de las vocales pronunciadas de cada individuo, con ello se pueden formar polígonos, aduciendo que en una sola frase cada individuo hubiera pronunciado las cinco vocales. Estos polígonos que se forman son diferentes para cada individuo de la base de datos.

Siendo que al menos se pronuncien 3 vocales de las 5 para el idioma español, pueden realizarse 10 combinaciones diferentes de las vocales pronunciadas, como se puede observar en la Figura 5.2, con estas combinaciones se forman triángulos, tomando en cuenta que la combinación /i-a-o/ es la misma que /a-i-o/ o cualquier otra combinación que se forme a partir de esas 3 vocales. De cada vocal (punto formado en el plano cartesiano) se obtienen los tres primeros formantes, obtenidos a partir de los coeficientes LPC, formándose un vector de características de dimensión nueve por cada una de las combinaciones (triángulos) formadas, teniendo las cinco vocales se tienen 10 vectores (triángulos) de dimensión nueve que representaría a uno de los individuos de la base de datos. Lo mismo pasa para la base



(a) Aarón

(b) Alfredo

(c) Eréndira

(d) Jesús

(e) Mario

(f) Nayeli

Figura 5.1: Ejemplo de polígonos formados por cada individuo de la base de datos ELSDSR. Para el idioma español, las cinco vocales forman un polígono que representa a cada individuo de la base de datos, estos polígonos tienen similitud entre cada individuo pronunciando en diferentes ocasiones palabras que contienen las cinco vocales, con ello podemos observar que el polígono formado por Aarón es diferente que el polígono formado por Nayeli y del polígono formado por Eréndira, así para todos los individuos de la base de datos. De esto se asume que es posible identificar a los individuos únicamente con los formantes de las vocales pronunciadas, tanto para las bases de datos del idioma español como para las bases de datos del idioma inglés.

de datos en idioma inglés, a diferencia que al contar con 10 vocales del idioma inglés, se pueden obtener más combinaciones (de acuerdo a la Tabla 5.3), y como se muestra en la Figura 5.3 se pueden obtener diferentes combinaciones dependiendo del número de vocales

pronunciadas por cada individuo.

Un problema que se tuvo con la base de datos DIMEx fue el etiquetado de los audios, Figura 5.4, donde se puede observar a la izquierda que el etiquetado es:

"todosobrelagerakontrelterorismo"

que es la frase:

"Todo sobre la guerra contra el terrorismo"

Otra de las frases contiene el etiquetado:

"konbensionsobrelkomersinternasenaldespesesamenasadasdefauneiflorassilbestr"

que es la frase

"Convención sobre el comercio internacional de especies amenazadas de fauna y flora silvestre"

El problema radica en que entre una letra y otra en ningún momento hay silencio, entonces hay segmentos de las señales en el tiempo marcado que ó tenían parte silencio ó parte de la siguiente letra pronunciada, como se muestra en la Figura 5.5. En el caso de las etiquetadas como el ejemplo "todosobrelagerakontrelterorismo" siendo "Todo sobre la guerra contra el terrorismo". Hubo casos que como en el segmento "kontreel" que es "contra el" se registraban coeficientes que no pertenecían al conjunto de vocales del mismo tipo, y al escuchar detalladamente algunos de los segmentos existía la duda de haber escuchado otra vocal diferente a la etiquetada.

Para solucionar el problema con la base de datos DIMEx se filtraron los elementos mal etiquetados de la base de datos, obteniendo la autocorrelación de cada segmento de la señal. Si la autocorrelación obtenida presentaba un pico suficientemente grande, se tomaba como un sonido vocalizado. De lo contrario ese segmento se desechaba, con ello se eliminaron 708 muestras de las 6000 que contiene la base de datos, siendo el 11.80 % de la base de datos que se encontraban mal etiquetadas. Para el caso de los segmentos donde los coeficientes

obtenidos no se encontraban dentro del rango de su grupo, también fueron desechados. Con ello se pudo trabajar con la base de datos DIMEx ya que era importante contar con suficientes muestras que fueran en el idioma español.

5.2. Resultados

Para el análisis de sensibilidad se utilizó la distancia Manhattan 5.1 (que es la distancia entre dos lugares medida en cuadras, por ello su nombre Manhattan ó City-Block), entre cada individuo de prueba y el resto de los individuos conocidos por el sistema se almacena en un archivo de registro. La distancia Manhattan se obtiene:

$$d(a,b) = \sum_{i=1}^{d} |a_i - b_i| \tag{5.1}$$

Cuando la distancia entre un individuo de prueba (representado por un triángulo, para el caso) y un individuo del conjunto de entrenamiento (representado por otro triángulo) es menor que el umbral th y la etiqueta de un individuo es la misma que la del individuo de prueba, entonces es un verdadero positivo (TP). Si la distancia entre el individuo de prueba y un individuo candidato del conjunto de entrenamiento es mayor o igual a th pero la etiqueta de los dos parlantes es la misma, entonces es un falso negativo (FN). En caso de que la distancia entre el individuo de prueba y el individuo candidato sea mayor que th y sus etiquetas sean diferentes, ya que en realidad son individuos diferentes, entonces estamos tratando con un verdadero negativo (TN). Finalmente, cuando la distancia entre los individuos es menor que th pero sus etiquetas son diferentes, entonces este es un falso positivo (FP), la Tabla 5.4 resume estas definiciones.

Tabla 5.4: Definición para el análisis de sensibilidad

Individuo	$distancia < th$	$distancia > th$
Mismo	Verdadero Positivo (TP)	Falso Negativo (FN)
Diferente	Falso Positivo (FP)	Verdadero Negativo (TN)

La Tasa de Verdaderos Positivos (TPR) se define como el número de individuos que el sistema identifica correctamente (es decir, los verdaderos positivos) dividido por el número de individuos que el sistema debería tener (es decir, positivos). TPR se calcula con la ecuación 5.2.

$$TPR = \frac{TP}{TP + FN} \tag{5.2}$$

La Tasa de Falsos Positivos (FPR) mide la frecuencia con la que el sistema confunde a un individuo con otro, es lo que comúnmente se conoce como Tasa de Alarma Falsa. FPR se calcula con la ecuación 5.3.

$$FPR = \frac{FP}{FP + TN} \tag{5.3}$$

En el plano de Características de Operación del Receptor (ROC), los ejes vertical y horizontal corresponden a TPR y FPR respectivamente, un punto en el plano ROC evalúa el desempeño de un clasificador para un valor específico de th . Al variar el rango dinámico de distancias entre los individuos, se genera una curva ROC. Cuanto mayor sea el área bajo una curva ROC, mejor se considera que es el clasificador.

Para las pruebas con el conjunto de datos en español, en las Figuras 5.6 y 5.7 se muestra la curva ROC obtenida para el conjunto de datos de Elocuciones21 y DIMEx, que usan nubes de 3, 4 y 5 vértices para caracterizar a los hablantes. Ahora, como en esta colección de grabaciones de voz, cada vocal aparece en cada grabación del individuo, de esta manera una sola nube representa al individuo representado por un vector de características de 9, 12 y 15 dimensiones (tres formantes por cada una de las cinco vocales mencionando 3, 4 y 5 vocales).

Para obtener el área debajo de la curva ROC (AUC, Area Under the ROC) se utilizó la regla trapezoidal, que es el cálculo del área como suma de áreas de trapecios, la cual se obtiene:

$$AUC = \sum_{t=1}^{T} \frac{1}{2}(FPR_t - FPR_{t-1}) \bullet (TPR_t + TPR_{t-1}) \tag{5.4}$$

donde las fracciones de FPR_t y TPR_t calculadas para $t = 1, \ldots, T$ puntos de corte.

El resultado es mejor para los polígonos que para los triángulos como se esperaba, pero con el inconveniente de que las cinco vocales deben aparecer en el habla de los individuos de entrenamiento y de prueba, mientras que para los triángulos solo tres vocales deben aparecer sin importar cuál.

Para las pruebas con el conjunto de datos en inglés, en las Figura 5.8 y 5.9 se muestran las curvas ROC obtenidas con el método propuesto para el conjunto de datos ELSDSR y TIMIT. Una de las curvas ROC que se muestran corresponde al clasificador que usa solo triángulos, los triángulos son aquellos que pueden formarse con cualquier subconjunto formado por tres vocales (sin importar cuál) entre las que aparecen en el audio del individuo identificado. Si en lugar de usar solo 3 vocales se usan 7 de las 10 vocales que pueden pronunciar en el idioma inglés, se obtiene un mejor clasificador, al igual que en el caso del conjunto de datos en español. De esta manera se logra una precisión del 85.96 % para identificar individuos en lenguaje español, y 99.71 % para identificar individuos en el lenguaje inglés.

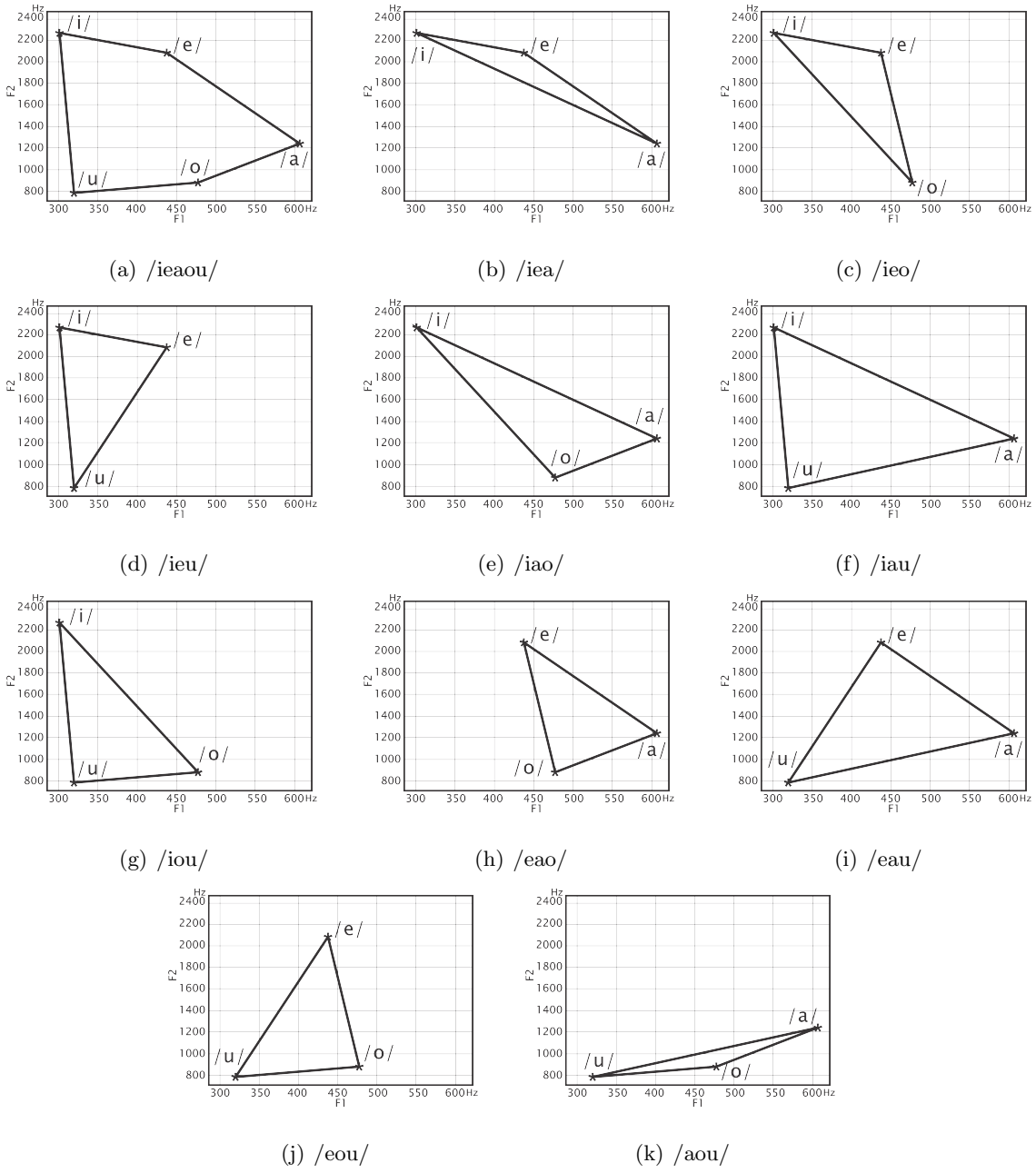
Se realizó una prueba de un audio con un discurso del actual presidente de México y un audio de un video obtenido de una persona imitando su voz, obteniendo los resultados que se muestran en la Figura 5.10, donde si solo se escucharan ambos audios no se podría distinguir las diferencias entre la voz real y la voz imitada, ambos discursos son diferentes en texto, sin embargo tienen en común las 5 vocales pronunciadas del idioma español, por lo tanto se obtuvieron formantes de cada vocal y formando los polígonos de estas vocales se puede observar que existen diferencias entre uno y otro.

Varios autores buscan la identificación de individuos de forma texto-independiente utilizando los Coeficientes Ceptrales de Mel (MFCC) junto con algún otro método para la extracción de características, como lo son el Modelo de Mezclas Gausianas (GMM), Redes Neuronales (NN), Redes Neuronales de Regresión General (GRNN), Coeficientes de Predicción Lineal (LPC), entre otros, con la finalidad de obtener características más robustas. En la Tabla 5.5 se hace una comparación con otros métodos que se utilizan para identificar

Tabla 5.5: Comparación con otros métodos

Autor	**Método propuesto**	**Resultados %**
(Chakraborty y Parekh, 2017)	MFCC	73 - 86 %
(Anand, Labati, Hanmandlu, Piuri, y Scotti, 2017)	MFCC/ISF	85.64 %
(Garcia, Arias-Vergara, Orozco-Arroyave, y Vargas-Bonilla, 2016)	MFCC/GMM	85 - 88 %
(Badran y Selim, 2000)	LPC/NN	88.94 %
(Li, Hu, Li, y Xu, 2014)	LPC/GRNN	91.9 %
Propuesto	Formantes	76.06 - 99.71 %

individuos por su voz de forma texto-independiente, donde se puede observar que el usar solo Formantes se obtienen buenos resultados, dependiendo del número de vocales pronunciadas por cada individuo.



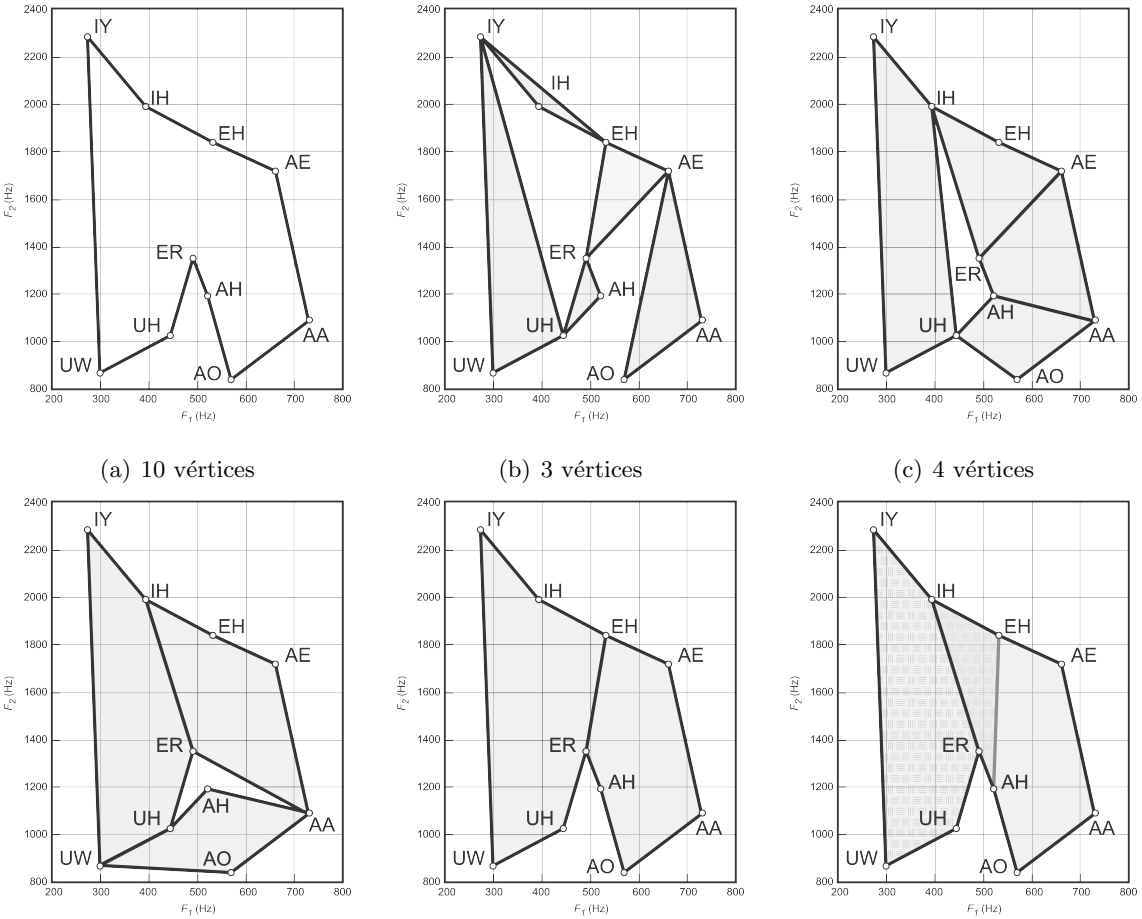
(a) /ieaou/ (b) /iea/ (c) /ieo/

(d) /ieu/ (e) /iao/ (f) /iau/

(g) /iou/ (h) /eao/ (i) /eau/

(j) /eou/ (k) /aou/

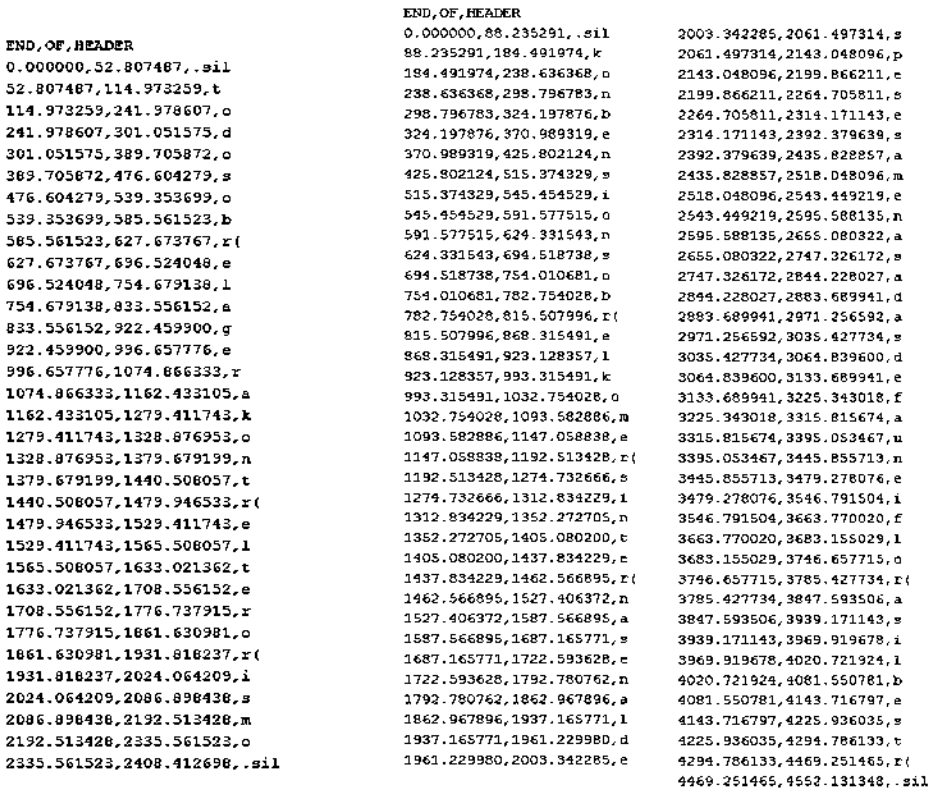
Figura 5.2: (a) Figura formada de la pronunciación de las cinco vocales del idioma español. (b)-(k) Triángulos formados por cada individuo del idioma español. Para el idioma español, si se pronunciaran al menos tres de las cinco vocales del lenguaje se formarían diez triángulos diferentes por cada individuo, por lo que se tendrían diez diferentes vectores de características de dimensión nueve (tres formantes de la vocal 1, tres formantes de la vocal 2 y tres formantes de la vocal 3))que representarían a un solo individuo, respetando que el vector del triángulo e-a-o sea el mismo que o-e-a, así para todos los vectores formados.



(a) 10 vértices (b) 3 vértices (c) 4 vértices

(d) 5 vértices (e) 6 vértices (f) 7 vértices

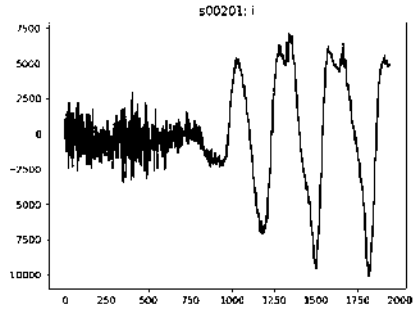
Figura 5.3: Ejemplo de figuras formadas por cada individuo del idioma inglés. Para el idioma inglés, se utilizaron combinaciones de 3, 4, 5 6 y 7 vocales pronunciadas de las diez vocales que componen el lenguaje. Como ejemplo en esta imagen se tienen (a) la figura que se obtiene si se pronunciaran las diez vocales en una sola señal de voz; en la figura (b) se observan algunos de los 120 triángulos posibles pronunciando solo 3 vocales de las 10, como se puede observar AE-AA-AO; (c) se tienen algunos de los 210 posibles cuadriláteros pronunciando 4 de las 10 vocales del idioma inglés, como ejemplo AH-AA-AO-UH; (d) algunos de las 252 posibles pentágonos pronunciando 5 de las 10 vocales del idioma, como ejemplo IH-EH-AE-AA-ER; (e) algunos de los 210 hexágonos pronunciando 6 de las 10 vocales del idioma, como ejemplo IY-IH-EH-ER-UH-UW; (f) algunos de las 120 heptágonos pronunciando 7 de las 10 vocales del idioma, como ejemplo IH-EH-AE-AA-AO-AH-ER.



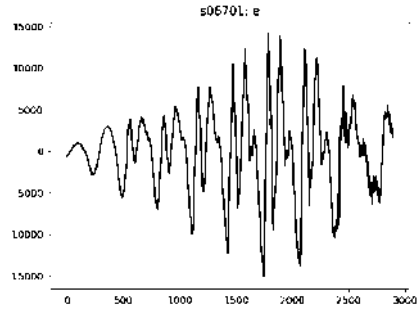
(a) Todo sobre la guerra contra el terrorismo

(b) Convención sobre el comercio internacional de especies amenazadas de fauna y flora silvestre

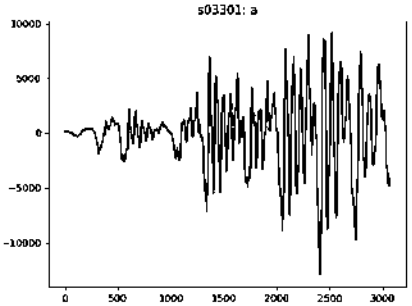
Figura 5.4: Dos ejemplos del etiquetado de la base de datos Dimex, donde se observa que entre cada letra no existen silencios, esto ocasionó que hubiera señales de audio con audio que no correspondía al etiquetado. A la izquierda el texto es “Todo sobre el terrorismo” y a la derecha el texto es “Convención sobre el comercio internacional de especies amenazadas de fauna y flora silvestre”.



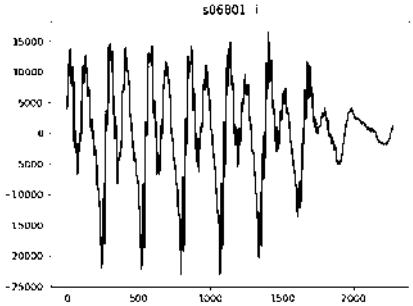
(a) Señal del audio s00201 : i



(b) Señal del audio s06701 : e



(c) Señal del audio s03301 : a



(d) Señal del audio s06801 : i

Figura 5.5: Ejemplos de segmentos de las señales de audio de la base de dato DIMEx donde se muestra que las etiquetas de un audio donde debiese ser una vocal existe información en ese segmento que podría considerarse parte de otro sonido diferente al etiquetado.

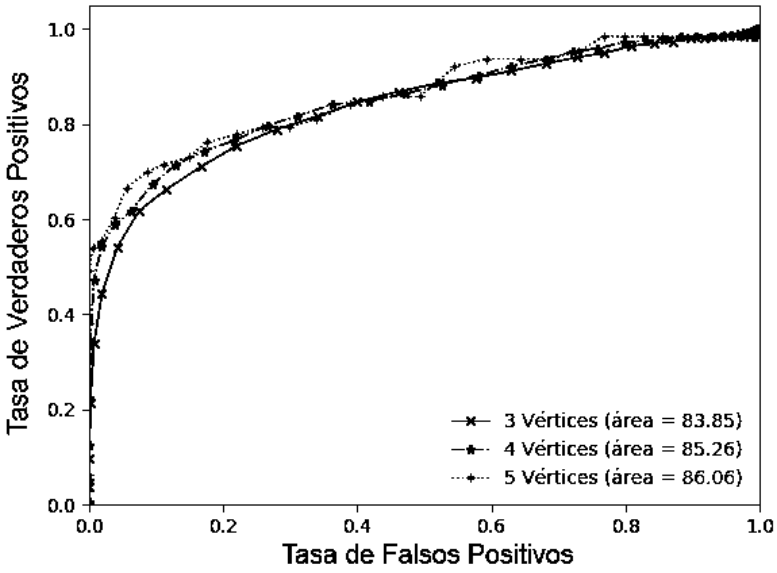


Figura 5.6: Curva Roc para la base de datos Elocuciones21 donde para 3 vocales pronunciadas en diferentes combinaciones se obtiene un 83.85 % de precisión, para 4 vocales pronunciadas en diferentes combinaciones se obtiene 85.26 % y para las 5 vocales pronunciadas en una frase se obtiene 86.06 % de precisión

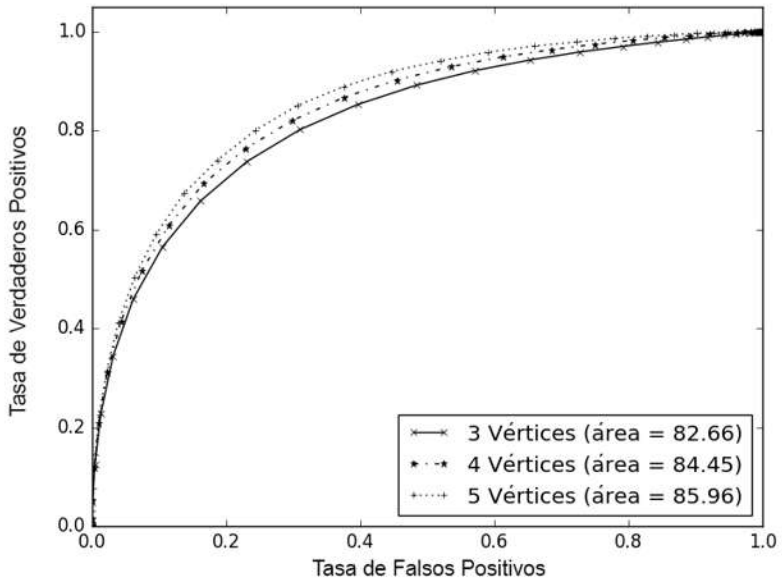


Figura 5.7: Curva Roc para la base de datos DIMEx donde para 3 vocales pronunciadas en diferentes combinaciones se obtiene un 82.66 % de precisión, para 4 vocales pronunciadas en diferentes combinaciones se obtiene 84.45 % y para las 5 vocales pronunciadas en una frase se obtiene 85.96 % de precisión

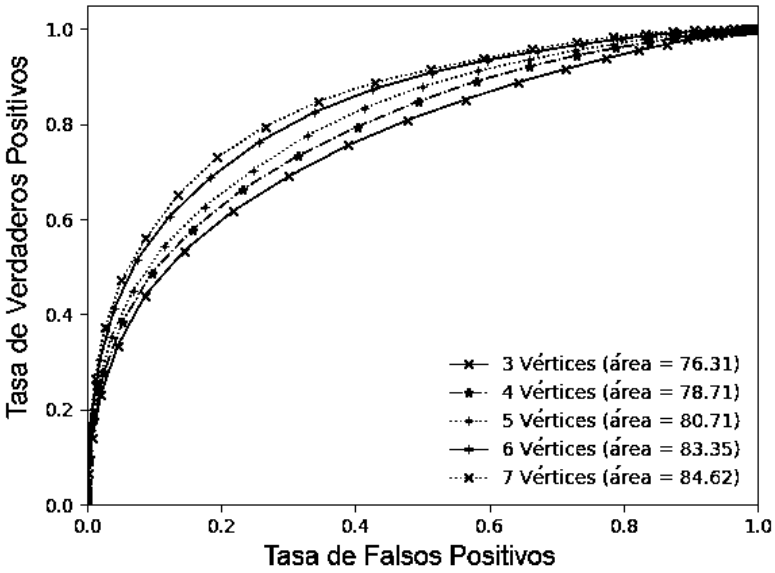


Figura 5.8: Curva Roc para base de datos ELSDSR utilizando combinaciones de 3, 4, 5, 6 y 7 vocales de las 10 vocales del idioma inglés, obteniendo desde 76.31 % hasta 84.62 % de precisión dependiendo de la cantidad de vocales pronunciadas por cada individuo

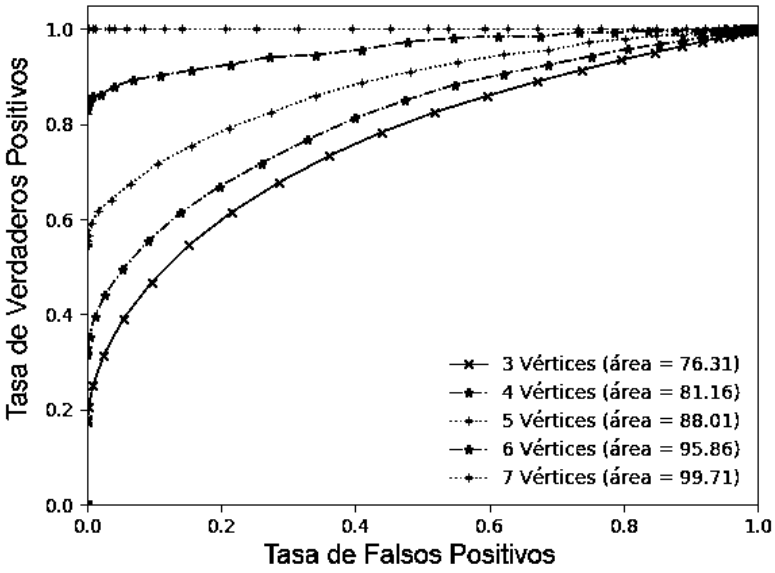
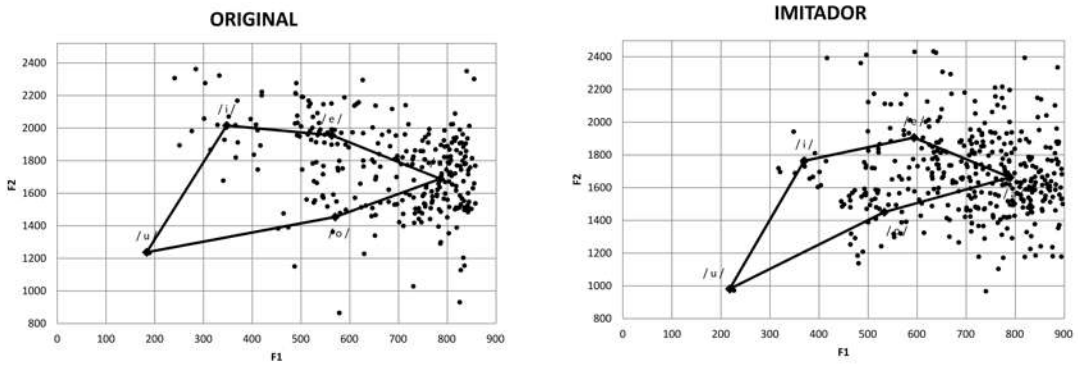


Figura 5.9: Curva Roc para base de datos TIMIT utilizando combinaciones de 3, 4, 5, 6 y 7 vocales de las 10 vocales del idioma inglés, obteniendo desde 76.31 % hasta 99.71 % de precisión dependiendo de la cantidad de vocales pronunciadas por cada individuo



(a) Voz real

(b) Voz imitada

Figura 5.10: Prueba de imitación de una voz, obteniendo los primeros dos formantes de las 5 vocales pronunciadas del idioma español, a la izquierda la voz real, a la derecha la voz imitada.

Capítulo 6

Conclusiones y Trabajo Futuro

6.1. Conclusiones Generales

Para identificar a un individuo a partir de una señal de voz independiente del texto solo se necesitan tres vocales diferentes en el discurso, sin importar que vocales específicas sean. En el caso de que se pronuncien más vocales, la TPR del sistema de identificación del hablante aumenta significativamente con el inconveniente de que se requieren más vocales en la señal de voz. También se puede aumentar la TPR del sistema utilizando más frecuencias de formantes, recordando que solo se utilizaron los primeros tres formantes. El idioma inglés es más rico en contenido por el número de vocales que se pronuncian. Cada figura formada por n vértices de una nube de puntos representa la nube completa, por lo que hay tantas entradas en el índice invertido para una nube específica, ya que las figuras pueden formarse con puntos de la nube.

6.2. Trabajos Futuros

1. Crear una base de datos del idioma español que sea utilizada para pruebas de identificación de individuos de forma texto-independiente que al menos contenga 500 individuos diferentes de distintas regiones del país.

2. Realizar un método para buscar polígonos similares, cada vértice del polígono se considera como un número complejo donde las coordenadas horizontales y verticales son las partes reales e imaginarias, respectivamente. La función $\varphi : C^n \to C$ determina una invariante de similitud para polígonos definidos como en la ecuación:

$$\varphi(z_1, z_2, \ldots, z_n) = \frac{\sum_{k=1}^{n} \lambda_n^k z_k}{\sum_{k=1}^{n} \lambda_n^{-k} z_k} \tag{6.1}$$

donde $\lambda_n = e^{2\phi i/n}$. La función φ mapea una secuencia de n números complejos en un solo número complejo por lo que transforma un polígono con n vértices en un solo punto en el plano complejo. En el caso $n = 3$, un triángulo se transforma en un punto único. La magnitud del número complejo entregado por la función φ podría usarse como un hash para buscar en un índice invertido.

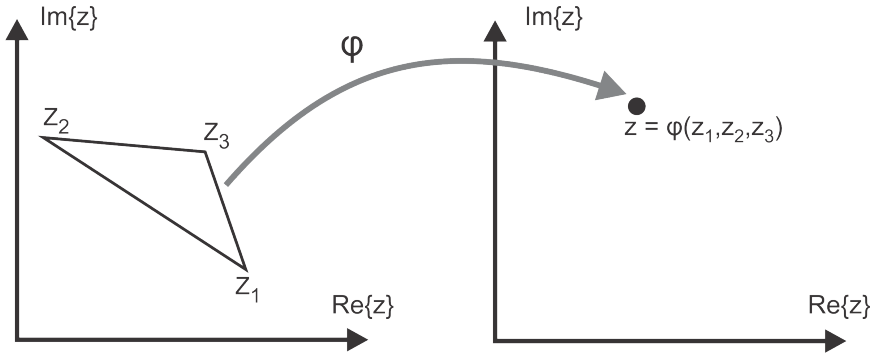


Figura 6.1: Triángulo transformado en un solo punto a través de la función φ

En cuanto al problema de escalabilidad, se obtienen muy buenos resultados ya que para una base de datos de 21 individuos y una de 630 los triángulos (o polígonos en general) deben caracterizarse por un solo número, por lo que este número se puede usar como un hash al insertar el triángulo en un índice invertido y así encontrar muy rápidamente la nube de puntos a la que pertenece.

3. Utilizar MFCC junto con los formantes para obtener resultados más robustos donde sea posible identificar un individuo que pronuncie dos o incluso sólo una vocal.

4. Implementar el funcionamiento del proceso de identificación formantes en un sistema en tiempo real.

Referencias

Anand, A., Labati, R. D., Hanmandlu, M., Piuri, V., y Scotti, F. (2017, June). Text-independent speaker recognition for ambient intelligence applications by using information set features. En *2017 ieee international conference on computational intelligence and virtual environments for measurement systems and applications (civemsa)* (p. 30-35). doi: 10.1109/CIVEMSA.2017.7995297

Badran, E. F. M. F., y Selim, H. (2000, Aug). Speaker recognition using artificial neural networks based on vowel phonemes. En *Wcc 2000 - icsp 2000. 2000 5th international conference on signal processing proceedings. 16th world computer congress 2000* (Vol. 2, p. 796-802 vol.2). doi: 10.1109/ICOSP.2000.891631

Chakraborty, S., y Parekh, R. (2017, Nov). An improved approach to open set text-independent speaker identification (osti-si). En *2017 third international conference on research in computational intelligence and communication networks (icrcicn)* (p. 51-56). doi: 10.1109/ICRCICN.2017.8234480

Dias, S. d. O., y cols. (2012). Estimation of the glottal pulse from speech or singing voice.

El-Samie, F. E. A. (2011). Information security for automatic speaker identification. En *Information security for automatic speaker identification* (pp. 1–122). Springer.

Furui, S. (2005). 50 years of progress in speech and speaker recognition research. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, *1*(2), 64–74.

Garcia, N., Arias-Vergara, T., Orozco-Arroyave, J. R., y Vargas-Bonilla, J. F. (2016, Aug). A new speech corpus in spanish for speaker verification. En *2016 xxi symposium on signal processing, images and artificial vision (stsiva)* (p. 1-7). doi: 10.1109/STSIVA.2016.7743364

Gracida, G., y Orduña, F. (2011). Evocanto: Programa de cómputo para analizar la voz cantada mediante técnicas de procesamiento digital de señales. *Computación y sistemas*, *15*(1), 39–50.

Hirano, M. (1981). Psycho-acoustic evaluation of voice. *Clinical examination of voice*, 81–84.

Itakura, F. (1975). Minimum prediction residual applied to speech recognifion. En *Ieee trans. acoustics, speech, signal proc* (pp. 67–72).

Li, P., Hu, F., Li, Y., y Xu, Y. (2014, July). Speaker identification using linear predictive cepstral coefficients and general regression neural network. En *Proceedings of the 33rd chinese control conference* (p. 4952-4956). doi: 10.1109/ChiCC.2014.6895780

Lieberman, P., y Blumstein, S. E. (1988). *Speech physiology, speech perception, and acoustic phonetics*. Cambridge University Press.

Naik, J. M., Netsch, L. P., y Doddington, G. R. (1989). Speaker verification over long distance telephone lines. En *International conference on acoustics, speech, and signal processing,* (pp. 524–527).

Peterson, G. E., y Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, *24*(2), 175–184.

Poritz, A. (1982). Linear predictive hidden markov models and the speech signal. En *Icassp'82. ieee international conference on acoustics, speech, and signal processing* (Vol. 7, pp. 1291–1294).

Quilis, A. (1980). Frecuencia de fonemas en el español hablado. *LEA: Lingüística española actual*, *2*(1), 1–25.

Rabiner, L. R., y Schafer, R. W. (1978). *Digital processing of speech signals* (Vol. 100). Prentice-hall Englewood Cliffs, NJ.

Rabiner, L. R., y Schafer, R. W. (2011). *Theory and applications of digital speech processing* (Vol. 64). Pearson Upper Saddle River, NJ.

Rose, P. (2003). *Forensic speaker identification*. CRC Press.

Sakai, T., y Doshita, S. (1962). An automatic recognition system of speech sounds.

Shaver, C. D., y Acken, J. M. (2016). A brief review of speaker recognition technology.

Shinoda, K., y Lee, C.-H. (2001). A structural bayes approach to speaker adaptation. *IEEE*

Transactions on Speech and Audio Processing, *9*(3), 276–287.

Stanley, D., y Watkins, S. S. A. (1939). *The science of voice: an application of the laws of acoustics, anatomy, physiology and psychology to the problems of vocal technic, including sections on music and interpretation, acoustics, advice to those interested in the radio and talking movies, and descriptions of original researches.* C. Fischer, inc.

Teng, G. (2016). *Language rush.* Descargado de `http://languagerush.com/`

Thévenaz, P., y Hügli, H. (1995). Usefulness of the lpc-residue in text-independent speaker verification. *Speech Communication*, *17*(1-2), 145–157.

Tola, E., Catanzaro, M., Viciano, A., y Hummel, P. (2019). *Hearing voices.* Descargado de `http://formicablu.github.io/hearingvoices` (Accedido 09-01-2019)

Tosi, O., y Tosi, O. (1979). *Voice identification: theory and legal applications.* University Park Press Baltimore.

Veiga, A. (2002). *El subsistema vocálico español* (n.º 11). Univ Santiago de Compostela.

Yarmey, A. D., Yarmey, M. J., y Todd, L. (2008). Frances mcgehee (1912–2004): The first earwitness researcher. *Perceptual and motor skills*, *106*(2), 387–394.

Zheng, T. F., y Li, L. (2017). *Robustness-related issues in speaker recognition.* Springer.