UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

ACÚSTICA FORENSE

T E S I S

QUE PARA OBTENER EL TÍTULO DE

INGENIERÍA MECÁNICA ELECTRICISTA
ÁREA ELÉCTRICA ELECTRÓNICA

P R E S E N T A

EDUARDO GARCIA LETECHIPIA

ASESOR: DR. VÍCTOR GARCIA GARDUÑO

MÉXICO, D.F. **2008**

Agradecimientos

Quisiera agradecer a todas las personas que ayudaron a lograr el presente trabajo, a mi universidad, la Universidad Nacional Autónoma de México y su personal por darme una formación académica de excelencia, fomentando en mí esa inclinación al aprendizaje constante y a la investigación, dando pauta al presente trabajo que representa varios años de dedicación y esfuerzo.

También quisiera agradecer a mi familia por el apoyo incondicional durante todo el tiempo de mi formación académica y del proceso de investigación que llevo a esta tesis, a mi esposa Gabriela, a mi madre María Teresa, a mi padre Armando y a todos mis hermanos.

Por último me gustaría terminar esta nota de agradecimiento expresando que a mi punto de vista este tipo de investigaciones son sumamente importantes para mejorar nuestro país y la seguridad de todos y cada unos de los ciudadanos que integramos el México de hoy, ya que tiene como miras prender esa mecha de la investigación y el desarrollo de aplicaciones especificas a las ciencias forenses que muchas veces se dejan relegadas, conformándonos con la adquisición de tecnologías extranjeras que en ocasiones no se adaptan por completo a los requerimientos y necesidades de nuestro país.

Capitulo I: Identificación de Locutores: Introducción

Capitulo II: Aplicación del método científico

Introducción

Para la procuración e impartición de justicia en México se hace patente la necesidad de fomentar la investigación en las áreas periciales, ya que la inmensa mayoría de la tecnología utilizada hoy en día es extranjera y en ocasiones sumamente costosa, por desgracia no invertimos en la investigación en México pero pagamos la investigación de gobiernos y empresas extranjeras que crean las técnicas y metodologías utilizadas en este ámbito al adquirir sus productos.

Con esto en mente se inicia una investigación que tiene como meta el desarrollo de una herramienta informática destinada a los estudios de acústica forense, una herramienta que permita al experto formular una hipótesis de manera rápida y objetiva, de tal suerte que logre orientar un estudio objetivo y certero que aporte una opinión calificada y bien fundamentada en un cuestionamiento concreto de alguna autoridad.

Actualmente este tipo de estudios son realizados de manera manual, involucrando el uso de aparatos especializados y el conocimiento de una o más personas, que dedica largas horas a estudiar cintas de audio o cualquier medio que pueda grabar una voz humana para determinar si se trata de la voz de uno o más probables responsables.

Estos estudios suelen ser lentos y costosos, y esto responde a razones tales como que por lo general la duración de una palabra cualquiera es de alrededor de un segundo, y los espectrógrafos utilizados suelen trabajar a una resolución tal que nos permita observar dos segundos de habla o dos palabras en promedio, por lo que imaginemos si se presenta al experto con 30 o 60 cintas de audio y cada cinta contiene alrededor de una hora de grabación, esto resultaría en 54,000 a 108,000 espectros a estudiar de manera manual y quizás algunos cientos de llamadas telefónicas.

De estos simples datos vemos que lo que parecía como algo quizás no muy complicado puede derivar en cientos de identificaciones y más aun decenas de miles de espectros a estudiar, por lo que ahora tenemos una mejor idea al porque estos estudios son lentos y costosos.

Derivado de lo anterior y teniendo en mente una máxime del derecho Mexicano que nos dice que la impartición y procuración de justicia debe ser pronta y expedita vemos que es sumamente necesario auxiliar a esta materia pericial con nuevas tecnologías informáticas como programas de cómputo que permitan al experto formular hipótesis y comprobar las mismas de una manera mucho más rápida y objetiva.

La idea de que una maquina puede reconocer nuestra voz ya sea para comprender lo que decimos o para saber quien lo dice no es nueva, esto lo podemos ver desde películas de los años 50’s o 60’s y en múltiples novelas como “Yo Robot” de Isaac Asimos, mas sin embargo hasta hace pocos años no era más que eso, una “novela”, pero hoy en día por medio de poderosas computadoras, sofisticados programas de computo y algoritmos altamente pulidos podemos por fin ver una aproximación realista a este sueño del hombre.

La presente investigación pretende sembrar esa semilla de curiosidad científica, brindando al mismo tiempo una base firme de conocimientos teóricos, todo esto

mediante la metodología científica que nos marca los pasos a seguir para lograr un trabajo valioso y completo.

En este trabajo se dará respuesta a muchas preguntas como ¿qué es la acústica forense?, o si nuestra voz es única y por lo tanto identificable, todo esto de manera gradual para que el lector pueda irse familiarizando con definiciones simples que fundamentan definiciones y temas más complejos, llevándolo de la mano hasta lograr la primera o por lo menos una de las primeras herramientas informáticas destinada a la identificación de locutores desarrollada totalmente en México.

Se dará inicio introduciendo al lector en la acústica forense, ya que esta es un ciencia ajena a una inmensa mayoría de la población, pasando después a algunos conceptos básicos como fisiología, fonética y bases matemáticas que permiten comprender la formación de la voz humana, ya que después de todo sería difícil tratar de estudiar aquello que desconocemos, pasando luego a un mayor rigor matemático en cómo se estudiara esta señal, pasando por su transducción, digitalización, y almacenaje en archivos de computadora.

Una vez cubierto esto veremos cómo podemos extraer información útil de esos archivos de computadora, por lo que veremos tres técnicas que se utilizan actualmente para tales fines, explicando en varias capas la extracción de parámetros y la interpretación de los mismos, ya que no bastara con obtenerlos, hay que comprenderlos y trabajarlos para que nos permitan ver claramente para que nos pueden ser útiles, pasando por ultimo a la explicación de la implementación informática así como la evaluación de resultados obtenidos durante las numerosas pruebas realizadas.

Capitulo I
"Identificación de Locutores: Introducción"

El objetivo del presente trabajo es el desarrollo de una herramienta informática destinada al auxilio en la identificación de locutores con fines forenses, mismo que muestra de inmediato la diferente aproximación necesaria contra otras muchas aplicaciones que existen actualmente en el mercado con fines tales como el reconocimiento del habla, la síntesis de voz y mas.

Esta herramienta informática brindara una rápida y objetiva orientación al perito en acústica forense o identificación del locutores, pero por las pretensiones de este proyecto se hace necesario el explicar algunos términos para tener una idea mas definida, así que empezaremos por definir de manera concisa algunas palabras que ya hemos ocupado en esta breve presentación:

Identificación: Acción de identificar, reconocer a una persona, objeto, animal o cosa, la cual se busca y esta relacionada con situación jurídica.

Identificar: Determinar de manera inequívoca la verdadera personalidad de un individuo un lugar o una cosa, sin existir confusión.

Identidad: Del latín *identitas*. Conjunto de rasgos y características físicas de una persona, haciéndola única a si misma y distinta de todas las demás.

Acústica: Parte de la física que trata del sonido y de todo lo que a el se refiere.

Forense: Perteneciente o relativo al foro o tribunal de justicia. Relativo al especialista designado por la ley para asistir en las actuaciones judiciales y ante los tribunales de justicia como perito en lo criminal y en lo civil.

Acústica Forense: Rama de la ciencia donde se estudian e investigan las voces y ruidos relacionados con alguna controversia presentada ante el órgano jurisdiccional, la cual sirve de base para establecer la verdad.

Perito: Proviene del latín *peritus*, que significa sabio, experimentado o hábil. El perito integra el conocimiento del juzgador o del ministerio público cuando se requiere la aportación de conocimientos especiales sobre una ciencia.

Objetivo, -va: Relativo al objeto en si, y no a nuestro modo de pensar y de sentir

Ya con los anteriores términos podemos decir que se pretende una herramienta informática que permita a un perito en acústica forense el identificar a un sujeto determinado de una manera objetiva y de acuerdo al método científico que establece la siguiente metodología:

1. Detección, delimitación y planteamiento de un problema
2. Antecedentes y estado actual del problema
3. Justificación
4. Propósitos y objetivos
5. Hipótesis

6. Diseño de la investigación
7. Resultados
8. Discusión
9. Conclusiones
10. Problemas pendientes
11. Resumen y bibliografía

Ahora, antes de poder compenetrarnos en lo que es el desarrollo de una aplicación informática que solucione nuestro problema debemos de comprender cabalmente el problema y el motivo del problema, que es en si la voz humana, lo cual lo podremos realizar contestando algunas preguntas que yo creo que muchos de nosotros ya se formulan:

¿Como podemos abordar en una solución informática la voz humana y su estudio?

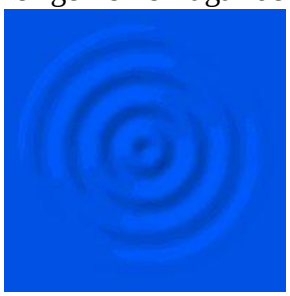
Para dar respuesta e esto requerimos profundizar en como una computadora puede manejar y procesar información correspondiente a la voz humana, pero para esto debemos de primera instancia comprender que es lo que vamos a estudiar dando las bases requeridas para comprender que es el sonido, los diferentes tipos de sonido, la voz humana, su proceso de formación y una breve introducción a la fisiología del aparato fonador humano.

¿Qué es el sonido y como se estudia?

El sonido, como todo fenómeno físico tiene diferentes maneras de aproximación para su estudio, ya que dependerá de que se desea abordar dentro de el y conque propósito por lo que encontraremos diversas definiciones y formas de trabajar con este tipo de señales, ya sea para su clasificación, trabajo o con fines de investigación.

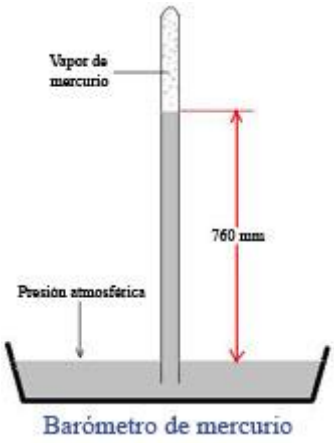
El sonido

El sonido es una perturbación espacio temporal de un medio de transmisión quasi-estático y elástico, (para una definición más amplia lea el siguiente apartado), esta perturbación consiste en que al nosotros emitir un sonido producimos una salida de aire a una determinada velocidad e intensidad o presión, esto se puede ver como una honda que se propaga en un medio, mismo que se puede ejemplificar al arrojar una piedra a un estanque de agua, veremos que se forman ondas concéntricas con origen en el lugar de impacto, mismas que se propagan de manera homogénea por el medio de transmisión (el agua), entre cresta y cresta de las ondas de agua podemos medir un nivel ligeramente inferior al nivel promedio del agua (profundidad) en el estanque, esto dado a que por física elemental, si las crestas tienen un nivel superior al estándar podemos decir que esa agua extra tiene que provenir de algún lugar, por lo que se ocasiona un nivel mas bajo en el área inmediata contigua, lo que da como resultado áreas de mayor altura al promedio y áreas de menor altura que el promedio, bueno, lo mismo ocurre con el aire, se formaran ondas que se propagan por el medio con una mayor presión, produciendo en el área contigua áreas de menor presión al promedio que a su vez viajan a la misma velocidad y dirección, esto produce áreas de menor y de mayor presión de forma alternada con una duración no determinada (dependerá de la naturaleza del sonido), esto es en si, el fenómeno del sonido y su forma de transmitirse, se puede ver por supuesto desde un punto de vista más riguroso matemáticamente o físicamente, pero no es objeto de este documento el profundizar de esa manera.



Medio de transmisión

Para la mayoría de los casos será el aire a presiones estándar, al nivel del mar esta presión será de aproximadamente 100,000 [Pa], cabe aclarar que esta presión es el resultado del peso o la presión que ejerce una columna de aire que va del punto de referencia al punto más alto de nuestra atmósfera, mismo que como se puede ver claramente varía según el lugar de referencia, así como de condiciones climatológicas, ya que el aire húmedo pesa más, además de factores como viento y otros que pueden afectar dicho resultado, como composición química de los gases que nos rodean (no todos los elementos químicos tienen el mismo peso atómico), pero en general y para los fines que se persiguen podemos olvidar dichos factores que modifican la presión, lo importante es que en cada lugar de referencia existe una presión de referencia o quasi-estática.



Barómetro de mercurio

Componentes del sonido (Cualidades Acústicas)

Los componentes fundamentales del sonido, como casi cualquier otro tema tienen varios puntos de vista o formas de estudiarlos, para el caso del sonido serán desde su punto de vista físico y perceptivo respectivamente siendo estos:

Físico	Perceptivo
Frecuencia	Tono (Pitch)
Presión Acústica	Sonia (Loudness)
Resonancia (espectro)	Timbre
Tiempo	Duración

¿Por qué se ven desde dos puntos de vista?

Esto se debe a que el sonido se puede medir o estudiar básicamente en dos etapas, esto es en su lugar de producción y desde el punto de vista del que escucha el sonido, básicamente se refieren a lo mismo definiciones como frecuencia y tono, más las que medimos de una manera real son las físicas, ya que las perceptivas dependerán también de quien las escuche, agudeza auditiva, patologías del oído y muchos otros factores ajenos al tema de este documento, por lo que a pesar de que algunos componentes son mas conocidos por su nombre o definición perceptiva, los elementos que se estudian son los físicos, ya que son los que se pueden medir directamente, por lo que daremos mayor énfasis a estos.

Frecuencia

Lo que entenderemos por frecuencia será el tono fundamental de una voz humana, misma que tiene su origen a nivel inconsciente desde la etapa primaria de la formación de la voz, esto es, son los pulsos de aire que se forman en los pulmones (es importante mencionar que estará en una sola frecuencia, sin armónicas, por lo que es una vibración simple, como la de un diapasón) para excitar posteriormente las llamadas "cuerdas vocales", y pasar a las distintas cámaras resonantes, que darán su carácter definitivo a los sonidos emitidos y que conforman lo que conocemos como voz. Esto mismo lo hace un elemento interesante, ya que es sumamente difícil modificar dicho tono fundamental y más aún mantenerlo fuera de su frecuencia natural, que resulta peculiar de cada individuo, esto lo podemos ver en las múltiples clasificaciones y tipos de voz, voces agudas, graves, etc.

Presión acústica

Por presión acústica debemos de entender que tanta fuerza o presión ejercemos sobre el aire al hablar, esta misma presión se puede traducir en la intensidad con que percibimos el sonido en el receptor, o lo que en equipos de sonido se le llama volumen, este parámetro es sumamente variante, ya que la presión que ejercemos al hablar puede variar fuertemente dentro del habla normal, por ejemplo cuando nos sentimos intimidados o incómodos suele bajar considerablemente nuestro volumen de habla, en cambio en situaciones de mucha excitación, por ejemplo en llamadas telefónicas en que se quiere infundir temor o se quiere demostrar determinación y control sobre alguna situación sube nuestro volumen del habla, como en llamadas de amenazas y algunas otras, por lo que vemos que no es un elemento que aporte fuertemente a la identificación de un individuo.

Resonancia

Este concepto, más conocido como timbre, a pesar de que esta es su contraparte perceptiva, es de suma importancia, ya que es el aspecto que más elementos aporta a la identificación de un sujeto en cuestión, esto debido a que se manifiestan un gran número de peculiaridades propias del individuo, englobando características tales como las armónicas del tono fundamental, la distribución en las distintas frecuencias de dichas armónicas, los niveles energéticos de cada armónica, mismas que son consecuencia de las múltiples cámaras de resonancia que encontramos en el aparato fonador humano, por lo que podemos decir que la resonancia nos permite el modelado del aparato fonador, pero podemos ver que este mismo modelado en si permite observar un gran número de peculiaridades, que con fines identificativos aporta un fuerte punto de referencia para la conformación de los fonemas, la transición entre fonemas, los morfemas, lexias y muchas mas características únicas a un individuo.

Para el estudio de la resonancia se utilizan los espectrogramas, dado que en forma gráfica es mucho más fácil analizar este gran número de datos y poder establecer parámetros comparativos, y de una manera más general se utilizan los sonogramas, que presentan la misma información pero permitiendo agregar el eje del tiempo, teniendo así una perspectiva mucho más amplia y útil a la vez.

Tiempo

Este es el más simple de los conceptos, ya que como su nombre lo dice y el de su contraparte perceptiva, es la duración de la porción en estudio, llámese fonema, morfema, lexia, frase, etc. Será la duración espacial de la perturbación en la estabilidad relativa del medio de transmisión.

¿Qué tipos de sonidos existen?

Sonidos simples

Un sonido simple será el resultado de una perturbación en una sola frecuencia o un sonido puro, un ejemplo de un sonido simple o puro puede ser el sonido generado por un diapasón, además de que por la naturaleza misma del sonido es difícil que se genere un sonido puro de manera natural, pero las computadoras pueden generar sonidos puros en cualquier frecuencia.

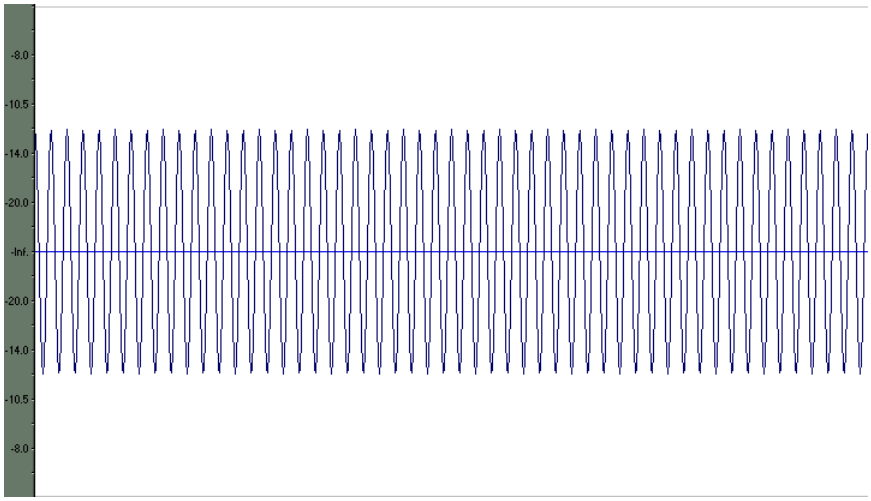


Figura 1: Sonido simple

Sonidos complejos

Un sonido complejo será la suma algebraica de un sonido simple agregando sus armónicas, que por cierto caen dentro de esto la mayoría de los sonidos. El sonido complejo tendrá como elemento primario a F_0 (tono fundamental), pero el sonido se puede expresar como $F_0 + F_1 + F_2 + ... + F_n$, donde $F_n = (n + 1) * F_0$, además de que es pertinente mencionar que a nivel glotal la amplitud de las armónicas decae 12 dB por octava (cada octava representa el doble de frecuencia), pero el sonido final tendrá una composición muy diferente a esto, ya que las cámaras resonantes de la laringe, la boca y las cámaras nasales pueden dar énfasis a cualquier grupo de armónicas, pudiendo incluso rebasar la amplitud del tono fundamental.

Más adelante se ejemplifica la formación de un sonido complejo a nivel grafico para tener un mejor entendimiento de este tipo de sonidos que son el centro de este trabajo.

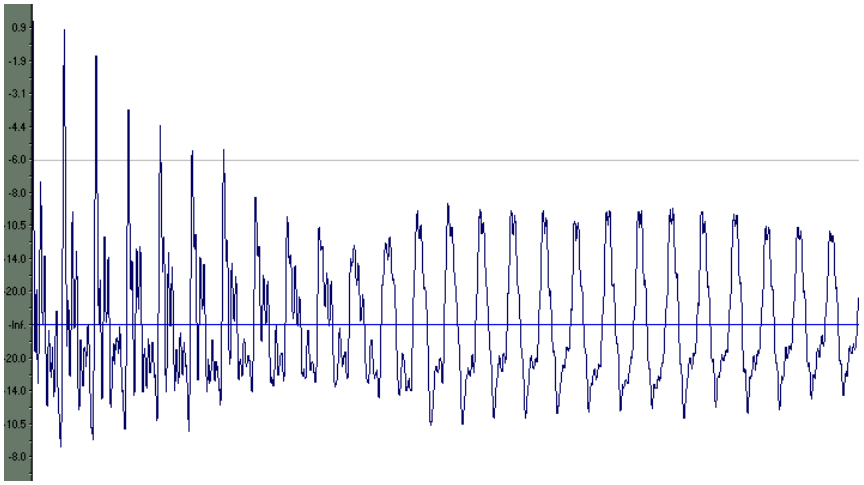


Figura 2: Sonido complejo

Ejemplo de formación de un sonido complejo

En este ejemplo se explicara como se forma una señal compleja de sonido y por que puede esta ser descompuesta en sus señales fundamentales por medio de procedimientos de filtrado o analizado por medio de herramientas matemáticas como la transformada rápida de Fourier.

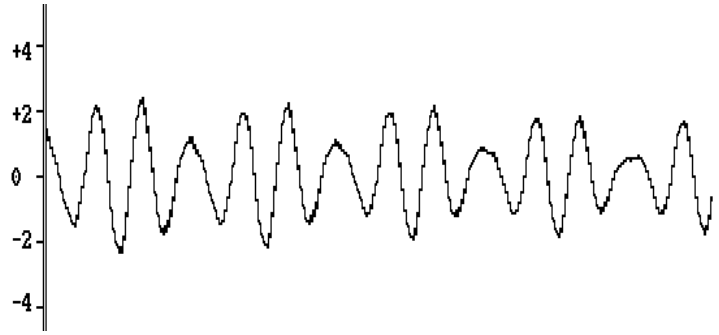


Figura 3: Parte del sonido Chimes

Para comenzar primero presentamos el grafico de la representación visual del sonido Chimes que viene como parte del paquete Windows originalmente en formato WAV, lo que se exporto a formato DAT para analizarlo con el programa Procsig (proporcionado en forma gratuita por el Dr. Sergio Suárez Guerra de Instituto Politécnico Nacional).

Ahora presentaremos gráficas de señales senoidales puras a frecuencias de 10, 50, 100 y 150 Hz, siendo la de 10 Hz la fundamental y las demás señales serán tres de sus armónicas, mismas que se aprecian a continuación.

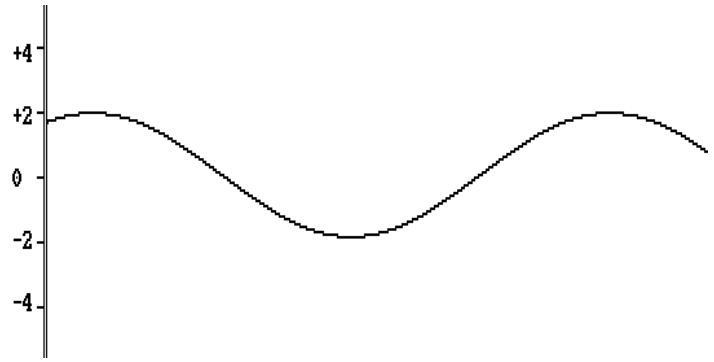


Figura 4: Señal senoidal de 10 Hz

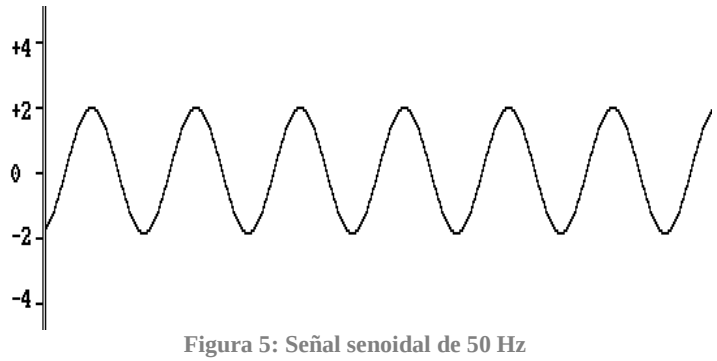


Figura 5: Señal senoidal de 50 Hz

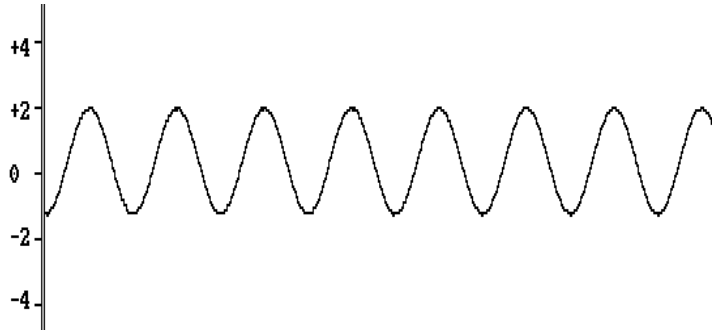


Figura 6: Señal senoidal de 100 Hz

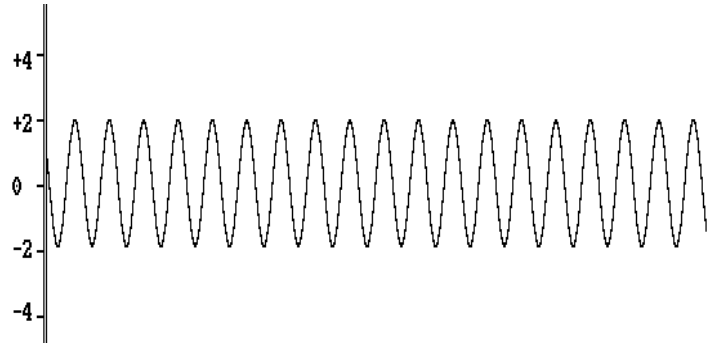


Figura 7: Señal senoidal de 150 Hz

Por excelencia se trabaja con señales senoidales, ya que son funciones simples y periódicas, lo que facilita su estudio, por función periódica queremos decir que se comportara de la misma manera en periodos de 2π radianes, ya que se da una vuelta completa a una circunferencia unitaria, lo que nos deja que para todo fin practico seria lo mismo calcular dichas funciones para valores de π que de 3π y así periódicamente.

Ahora para tener una mejor idea de cómo se formaría un sonido común sumemos las funciones senoidales que ya se presentaron, sumaremos primero las señales de 10 Hz y de 150 Hz (suma punto a punto), dándonos como resultado la figura 8.

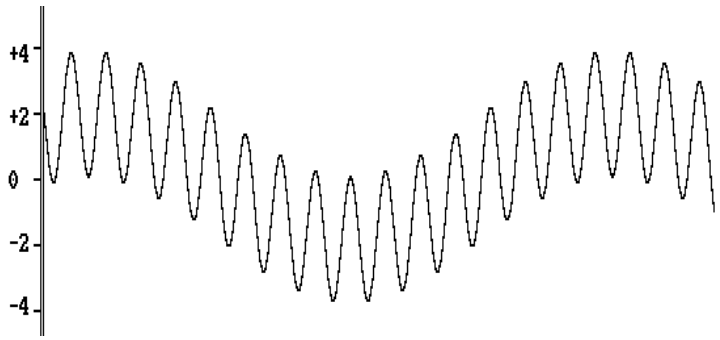


Figura 8: Suma de señales senoidales de 10 y 150 Hz

Como podemos ver pareciera que montáramos la señal de 150 Hz en la de 10 Hz, que para todo fin practico es lo que ocurre, pero ahora veamos que le sucede a la señal al sumarle las señales de 50 Hz en la figura 9 y por ultimo la de 100 Hz en la figura 10.

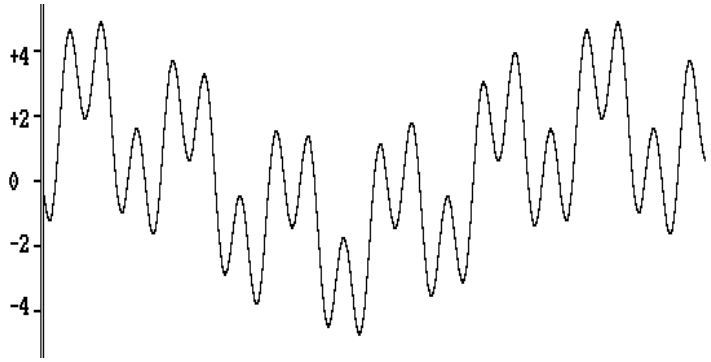


Figura 9: Suma de señales senoidales de 10, 50 y 150 Hz

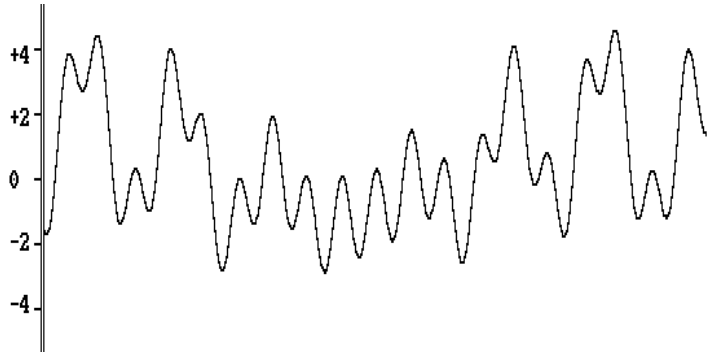


Figura 10: Suma de señales senoidales de 10, 50, 100 y 150 Hz

De las propiedades de la suma nos conviene resaltar la conmutatividad y la asociatividad, esto en el afán de ver si se nos permite matemáticamente realizar las funciones anteriores sin modificar el resultado final, cómo seria en la naturaleza en realidad, en la cual todas las señales y sus armónicas son presentadas en un solo tiempo e incluso potenciadas algunas de ellas. Estas propiedades nos dicen que en una suma cualquiera:

Conmutatividad $a + b + c + d = b + a + c + d = d + c + a + b$
Asociatividad $a + b + c = a + (b + c)$

De dichas propiedades podemos ver que se pueden sumar los términos en cualquier orden y podemos realizar sumas parciales sin temor de que esto afecte el resultado final.

Ahora analizando la figura 10 vemos que esta es la que más se parece a la figura 3 del sonido Chimes, que es una señal de sonido real, por lo que podemos entender mejor los teoremas que nos dicen que una señal puede ser descompuesta en un numero infinito de componentes simples, que es lo que realizaremos con una transformación de Fourier, donde cada componente tiene una elemento que representa el peso de la componente o que tanto afectara al resto de la señal y un segundo elemento que representa la frecuencia de dicho componente como veremos mas adelante.

Con las anteriores explicaciones ya sabemos que existen sonidos simples que prácticamente no existen en la naturaleza y sonidos complejos, entre los cuales se encuentra la voz humana, y sabemos también que para todo fin práctico podemos considerar a un sonido complejo como una sumatoria de sonidos simples, con esto en mente pasaremos ahora a ver como se forma la voz humana y la fisiología del aparato fonador para comprender mejor esta señal de audio conocida como voz humana.

¿Cómo se forma la voz humana?

Habiendo previamente visto lo que es en general el sonido, pasaremos ahora a definir con mas cuidado lo que es la voz humana o el conjunto de sonidos que puede general el ser humano, ya que después de todo este es la materia sobre la que girara el resto del trabajo, definiendo, buscando y aislando características o elementos que pueden ser de utilidad en un proceso que tiene como fin el probar la identidad de una persona por medio de su voz.

La voz

La voz o fonación, es el sonido producido en la laringe por la salida del aire (espiración) que, al atravesar las cuerdas vocales, las hace vibrar. La voz se define en cuanto a su tono, calidad e intensidad o fuerza. El tono óptimo o natural para el habla, al igual que su rango de variación, depende de cada individuo y está determinado por la longitud y masa de las cuerdas vocales. Por tanto, el tono puede alterarse, variando la presión del aire exhalado y la tensión sobre las cuerdas vocales. Esta combinación determina la frecuencia a la que vibran las cuerdas: a mayor frecuencia de vibración, más alto es el tono.

Otro aspecto de la voz es la resonancia. Una vez que ésta se origina, resuena en el pecho, garganta y cavidad bucal. La calidad de la voz depende de la resonancia y de la manera en que vibran las cuerdas vocales, mientras que la intensidad depende de la resonancia y de la fuerza de vibración de las cuerdas.

Esto suena hasta cierto punto fácil, sin embargo intervienen una serie de estructuras fisiológicas que permiten un muy amplio rango de sonidos que un ser humano puede lograr, por lo que en el afán de entender mejor este proceso y su complejidad vamos a estudiar de manera general las estructuras básicas que intervienen y su función en la generación de la voz.

Los 3 niveles del aparato vocal

Clásicamente, este puede dividirse en tres partes, que son:

a) Los fuelles
b) El vibrador
c) Los resonadores

Los fuelles

La voz puede considerarse como una espiración sonorizada. En la fonación la espiración es activa, el aire es expulsado de los pulmones por la acción los músculos espiratorios. La espiración activa necesaria para que se produzca la voz se denomina *soplo fonatorio*.

El soplo fonatorio puede producirse de varias formas, en ocasiones lo produce el descenso de la caja torácica (soplo torácico superior) en las expresiones simples, en ocasiones lo origina la acción de los músculos abdominales (soplo abdominal) en la

proyección vocal, y por ultimo la flexión torácica en un contexto de esfuerzo, como sucede con la voz de insistencia o de apremio y en el forzamiento vocal.

El diafragma, principal músculo inspirador, es una lámina muscular en forma de bóveda. Separa el tórax del abdomen. El diafragma desempeña una importante función en la proyección vocal.

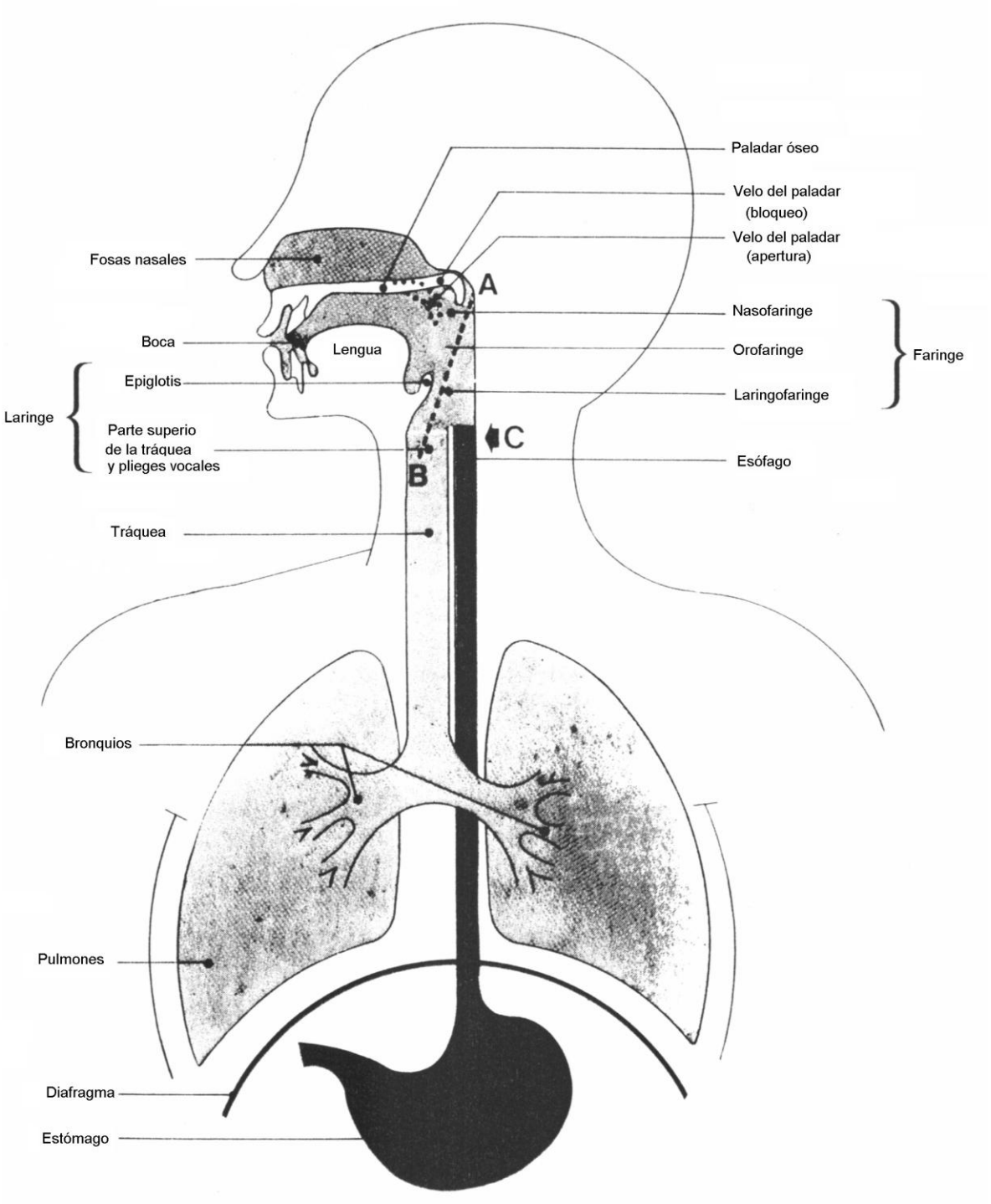


Figura 11: Los 3 niveles del aparato vocal

Durante el impulso respiratorio, el aire se introduce en los pulmones a través de la traquea y luego de los bronquios para acabar en los alvéolos pulmonares.

Durante la fonación, el aire recorre el camino inverso para acceder a la laringe con una velocidad y una presión que varia según la voz que va a producirse.

El vibrador

La laringe, extremo superior del tubo traqueal que se conecta con la faringe, es el principal órgano de la voz; aunque esta función es secundaria, su función primordial es facilitar la obstrucción de la traquea. Esta formada por cartílagos unidos entre sí por ligamentos y fascias, así como por músculos recubiertos por una mucosa. Los pliegues vocales (todavía denominados con frecuencia *cuerdas vocales*) forman parte de la laringe y lo constituyen dos de estos músculos y sus mucosas que lo recubren.

Los pliegues vocales son como dos labios horizontales situados en el extremo superior de la traquea, uno a la derecha y otro a la izquierda. Unidos por delante, pueden separarse o aproximarse entre sí por detrás; cuando se aproximan, pueden vibrar por acción del soplo pulmonar.

La glotis es el espacio comprendido entre los pliegues vocales cuando están alejados entre si. Por encima de los pliegues vocales existen dos repliegues algo parecidos, los pliegues vestibulares (o falsas cuerdas vocales) que no desempeñan función alguna en la producción de la voz normal.

Por otra parte la función de la epiglotis es la de una válvula que replegándose hacia atrás en el momento de la deglución, forma una tapadera para la laringe, de modo que los alimentos pasen hacia él estomago vía esófago. La epiglotis forma parte de la laringe aunque cuando se eleva se situé por entero en la cavidad faríngea.

Los resonadores

La laringe termina por arriba de la faringe, que no es otra cosa que la cavidad posterior de la boca o garganta, la cual sigue a la boca por detrás de la lengua.

Es una cavidad muscular capaz de estrecharse lateralmente y de atrás hacia delante. Asimismo, el volumen de la faringe puede variar verticalmente. Estas modificaciones dependen de los movimientos de elevación y de descenso de la laringe que acabamos de comentar y desempeñan una importante función en la articulación de las vocales.

Esta cavidad se divide en tres niveles superpuestos que, de abajo a arriba, son:

- Laringofaringe
- Orofaringe
- Nasofaringe

Laringofaringe

La laringofaringe corresponde a toda la zona faríngea situada por debajo de la parte libre de la epiglotis. Durante la deglución, como ya hemos comentado, la epiglotis desciende para cerrar el tubo y que el bolo alimenticio pase por encima de la epiglotis abatida y por los lados del tubo.

De este modo, en la laringofaringe desembocan dos conductos: la laringe por delante y el esófago por detrás.

El esófago es un tubo aplanado de unos dos centímetros de anchura, y que va de la faringe al estomago.
El orificio de comunicación del esófago con la laringofaringe se denomina boca del esófago, que puede cerrarse por acción de un anillo muscular o en la deglución se relaja para permitir que los alimentos pasen al esófago al tiempo que la epiglotis desciende para cubrir la laringe y cierra la traquea.

Orofaringe

Cuando se abre mucho la boca, se observan en su fondo, son dos repliegues de mucosa situados verticalmente y separados por debajo por la base de la lengua, se unen por arriba para formar la úvula (o campanilla).

Nasofaringe

Cuando el velo del paladar permanece descendido, la Orofaringe se comunica con la zona posterior de la nariz o nasofaringe.

El velo del paladar puede imaginarse como una válvula que al elevarse impide que el aire pase por la nariz. Durante el habla, el velo del paladar permanece descendido para las vocales y las consonantes nasales (m, n, ñ) y se eleva para los demás sonidos.

Con esto se cubre de manera elemental las estructuras generales que intervienen en al formación de la fonación, mas sin embargo su fisiología es un tanto mas compleja. Bien vale la pena adentrarse un tanto mas en la fisiología no solo por conocimiento general, sino porque aquello que queremos estudiar se debe de comprender lo mejor posible, y después de todo algo que se toma tan simple como hablar involucra procesos complejos y en ocasiones difíciles de comprender aun para los estudiosos del tema, razón por la cual veremos un poco mas a fondo la fisiología del aparato fonador.

Fisiología básica del aparato fonador
Como hemos visto de la división clásica de fuelles, vibradores y resonadores, la etapa primaria de formación de la voz tiene su origen en las fuelles, mismas que contribuyen con un soplo fonatorio (que es la función secundaria del aparato respiratorio, siendo su función primaria por supuesto realizar la respiración), ahora explicaremos de una forma más amplia la fisiología del soplo fonatorio, pasando después a una breve explicación de la fisiología foniatríca de la laringe, para terminar con la fisiología de la articulación del habla, misma con la que se concluirá una aproximación al funcionamiento mismo del aparato fonador y las partes que lo constituyen.

Fonación y las dos fases de la respiración
La fonación lleva aparejada la adopción de un ritmo respiratorio especial, fundamentalmente distinto del propio de la respiración tranquila. En efecto, en la respiración tranquila el ritmo respiratorio es rítmico, y la duración de cada ciclo varía poco de un ciclo a otro. Las dos fases respiratorias tienen una duración similar; la espiración es solo algo mas prolongada que la inspiración.

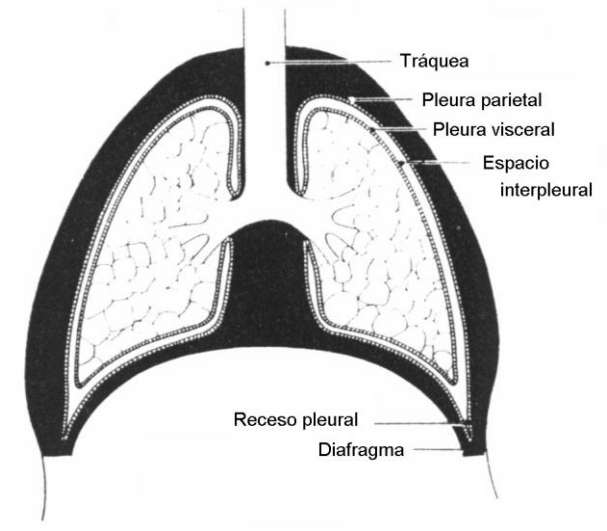


Figura 12: Pulmones

En la fonación, el ritmo respiratorio pierde esta regularidad porque:

1) La inspiración se acorta de manera considerable.
2) La espiración, convertida en soplo fonatorio, se prolonga variablemente, entrecortada por pausas con bloqueo laríngeo, que corresponden a las naturales fluctuaciones que señalan el ritmo del habla espontánea.

Este completo cambio del ritmo indica que la función del habla pasa a primer plano en el comportamiento respiratorio y que las necesidades de la respiración pasan a un segundo plano, pudiendo llegar a ocurrir que a fuerza de pasar a un segundo plano la respiración esta sea insuficiente, dando lugar a un jadeo que dificulte el habla.

Fisiología foniatríca de la laringe

Como hemos expuesto anteriormente la laringe es en principio un órgano destinado a cerrar la traquea. Durante la deglución, la glotis también se cierra para oponerse al paso de los alimentos hacia la traquea, simultáneamente con el descenso de la epiglotis.

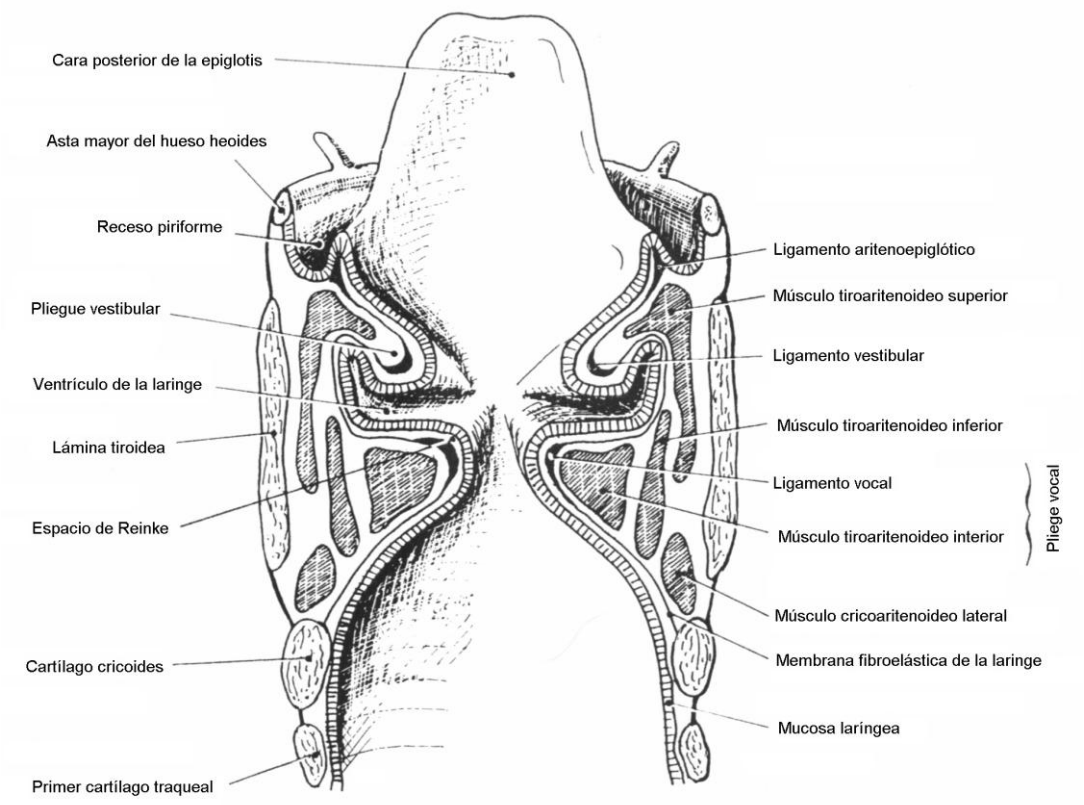


Figura 13: Corte frontal de la laringe

En el curso de la evolución animal, la laringe ha adquirido solo secundariamente una función vocal, que es lo que a continuación vamos a estudiar. Esta es realmente la parte más compleja desde el punto de vista fisiológico, razón por la cual procederemos desde la teoría más rudimentaria hasta las más actuales, viendo como se van complementando y mejorando en miras de explicar este complejo proceso

Teoría Mioelástica (Ewald, 1898)

Esta teoría se caracteriza por dos importantes conceptos:

1) La vibración de los pliegues vocales se considera pasiva
2) Las características del sonido emitido dependen exclusivamente de la presión infraglótica y de la tensión de los pliegues vocales.

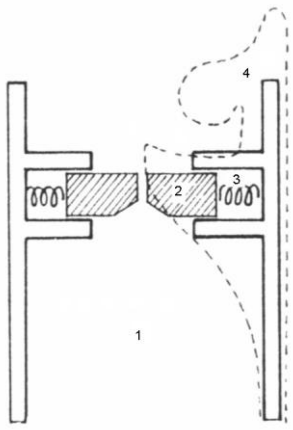


Figura 14: Esquema de Grémy

El esquema de la figura 14 (tomado de Grémy, 1968) permite con más facilidad hacerse una idea de la mecánica foniatríca según la teoría Mioelástica.

La traque esta representada por un tubo (1) y los pliegues vocales por piezas metálicas (2), introducidas en conductos horizontales (3), dentro de los que pueden deslizarse. Un muelle situado entre el fondo del conducto y la pieza metálica empuja esta hacia la otra pieza. El muelle representa la fuerza elástica de los pliegues vocales, que tiende a aproximarlos.

Al principio, el tubo permanece cerrado y se supone que esta obstrucción es hermética. Detalle importante, como se aprecia en el esquema, la arista inferior e interna de cada pieza metálica esta biselada (lo que corresponde al aspecto ojival de la infraglotis).

Si se aumenta la presión aérea en el tubo mediante un fuelle (que ejerce la función de los pulmones), esta presión aplicada sobre el bisel de las piezas metálicas crea una fuerza tal que tiende a separarlas. Sin embargo, la duración de esta separación es limitada por la presión misma.

Efectivamente, esta separación permite que se escape una pequeña cantidad de aire, lo cual origina de forma inmediata una disminución de la presión ejercida por debajo de las piezas metálicas. Los muelles intervienen entonces aplicando una fuerza que unirá de nuevo a las piezas metálicas (cierre de la glotis). Como el fuelle esta actuando constantemente, la presión infraglótica vuelve a elevarse, lo que induce el encadenamiento de los mismos fenómenos.

Teoría Neurocronáxica (Husson)

Según Husson, la laringe es una sirena de activación periódica, en la que los pliegues vocales poseen una función activa. La frecuencia de los impulsos motores procedentes del nervio laríngeo lo que condiciona la frecuencia de su vibración y, por tanto, la frecuencia del sonido.

De este modo, el mecanismo regulador de la frecuencia de los sonidos seria independiente del mecanismo que regula la intensidad de los sonidos (presión infraglótica)

La cuerda vocal según Goertier

La teoría de Husson se basa en un peculiar concepto de la anatomía microscópica del músculo del pliegue vocal, según los trabajos histológicos de Goertier (1950), quien considera que la capa interna del músculo está formada por dos músculos, el tirovocal y arivocal, que se insertan, respectivamente en la cara posterior del cartílago tiroides, y el segundo, en la apófisis vocal, y que sus fibras se entrelazan como los dientes de un peine con las del primero.

El impulso recurrencial

Según Husson, cuando un impulso nervioso accede a los pliegues vocales procedentes del nervio laríngeo, las fibrillas de los músculos arivocales y tirovocales se contraen. Entonces, el borde libre de cada pliegue sé curva hacia el exterior, separándose un breve instante un pliegue vocal del otro, con lo cual pasara entre ellos una pequeña cantidad de aire infraglótico que se escapa hacia el pabellón faringobucal, antes de que la

relajación de las fibras musculares arivocal y tirovocal produzca una fracción de segundo mas tarde el cierre de la glotis por unión de los pliegues vocales.

En otras palabras, siempre que se separen los pliegues vocales por influjo de un impulso nervioso, se escapara un poco de aire a través de la glotis, lo cual inducirá un aumento de la presión en la cavidad faringobucal. La frecuencia de las variaciones de presión, es decir, la frecuencia del sonido emitido, que corresponde a la frecuencia de las aperturas góticas, depende para Husson de la frecuencia de los impulsos nerviosos.

Teoría mucoondulatoria y mioelástica perfeccionada

Estas teorías aparecen como reacción contra la teoría Neurocronáxica, cuya exposición a principios de los años 50 reactivo las investigaciones referentes a la fisiología laríngea.

Ley de Bernouilli
La teoría mucoondulatoria y la teoría mioelástica perfeccionada explican lo que ocurre en la laringe atendiéndose a la ley de Bernouilli. Recordemos a continuación este clásico fenómeno de la física de los fluidos.

Cuando se hace circular un fluido, agua por ejemplo, por un tubo de diámetro variable, se aprecia que la velocidad de flujo en las zonas donde el diámetro del tubo es mayor disminuye, lo que es muy comprensible; ahora bien, es menos evidente que se de una disminución de la presión en las zonas del tubo que poseen un diámetro mas reducido, donde la velocidad es consecuentemente mas elevada.

Aplicación de la cinética laríngea
Si se observa un corte frontal de la laringe, se aprecia que la glotis forma un estrechamiento del conducto aéreo.

Cuando una corriente de aire suficientemente rápida circula a través de la glotis entreabierta, se produce obligatoriamente un descenso de la presión del aire a ese nivel. Esta baja de presión es susceptible de causar la aproximación de la mucosa de los pliegues vocales; ahora bien, cuando culmina esta aproximación, la glotis esta cerrada, por lo que acaba la baja de presión mencionada.

Entonces se eleva la presión infraglótica, lo cual induce la apertura de la glotis y una nueva corriente de aire, que a su vez vuelve a originar una baja presión a nivel de la glotis, con la subsiguiente cierre de los pliegues vocales, y así sucesivamente.

Además, se aprecia que la separación de los pliegues vocales, que como hemos expuesto comienza en la parte inferior de la glotis, se lleva a cabo de una manera bastante pausada, en tanto que la aproximación consecutiva se efectúa con mayor brusquedad. Evidentemente, la explicación radica en que la baja presión intraglótica se incrementa de forma paralela al progresivo estrechamiento de la hendidura glótica y de la velocidad del flujo de aire resultante.

Teoría Impulsional

Cornut y Lafon llegan, en primer lugar, a la conclusión de que el funcionamiento laríngeo se explica perfectamente a partir de tres elementos, como son la fuerza de oclusión gótica, la presión infraglótica y la fuerza de retroceso debido al efecto de Bernouilli, sin que sea necesario recurrir a un mecanismo que haga depender la frecuencia vocal de los impulsos nerviosos recurrentes. Además subrayan que el funcionamiento laríngeo debe plantearse fundamentalmente no como frecuencial sino como impulsional. En efecto, los pliegues vocales no vibran como lo hacen las cuerdas de un violín o el brazo de un diapasón. La laringe, cuyo funcionamiento se manifiesta por una alternancia de oclusiones y aperturas, puede compararse con un oscilador que produce rítmicamente impulsos.

El concepto de impulso laríngeo es importante en lo que se refiere a la fisiología del habla. En efecto, los impulsos laríngeos originan una sucesión de bruscas variaciones de presión, susceptibles de excitar las cavidades infraglóticas. Los fenómenos acústicos que corresponden al habla se explican mejor si se tiene en cuenta este aspecto impulsional del funcionamiento de la laringe.

Teoría Neurooscilatoria

Esta teoría, expuesta en 1968, afirma como la de Husson, que la vibración del pliegue vocal es un fenómeno que depende directamente de la actividad del músculo vocal.

Mac-Léod compara este músculo vocal con el de las alas de los insectos, denominado asincrónico. El músculo asincrónico se caracteriza por la posibilidad de entrar en vibración, siempre que la carga que se le aplique sea reactiva y no resistiva, como es lo más habitual en un músculo. En el caso del insecto, la reactividad depende de la elasticidad de sus estructuras torácicas.

La frecuencia de las vibraciones de un músculo asíncrono se relaciona exclusivamente con la masa y la elasticidad de las estructuras en movimiento, y es independiente de la frecuencia de los impulsos nerviosos que acceden al músculo.

Mac-Léod, a través de una serie de experimentos, demuestra que a nivel de la laringe existen este tipo de movimientos. Existen argumentos histológicos que abogan por esta analogía entre los músculos de las alas de los insectos y la laringe.

Por otra parte, existe similitud a nivel de las membranas intrafibrilares, pues la intervención es también algo similar. En efecto, tanto en el músculo vocal como en el de las alas de los insectos, cada fibra nerviosa termina en múltiples sinapsis, lo cual párase ajustarse exactamente a una actividad rítmica.

El interés de esta teoría es que responde de forma más satisfactoria a las críticas suscitadas por la teoría mioelástica concernientes a la energía necesaria para la actividad fonatoria.

No obstante, hay que hacer hincapié en que su fundamento en la arquitectura del músculo vocal y la teoría de Husson de que la laringe es una sirena de activación periódica, implican que la mucosa laríngea no cumple ninguna función importante. Las

investigaciones recientes no han permitido hasta el momento presente profundizar más en esta posible analogía entre el músculo vocal humano y el de las alas de los insectos.

Teoría osciloimpedencial

Esta teoría complementa hasta cierto punto las de Ewald y de Cornut-Lafon al afirmar que la laringe es, en realidad, un oscilador con amortiguación reducida. Recuperando los conceptos de Hirano, que diferencian la estructura del pliegue vocal y el su recubrimiento (la mucosa), Dejonckére añade que se trata de un oscilador complejo. Dado que estas dos estructuras anatómicas no presentan las mismas características mecánicas, efectivamente puede hablarse de un oscilador de múltiples componentes.

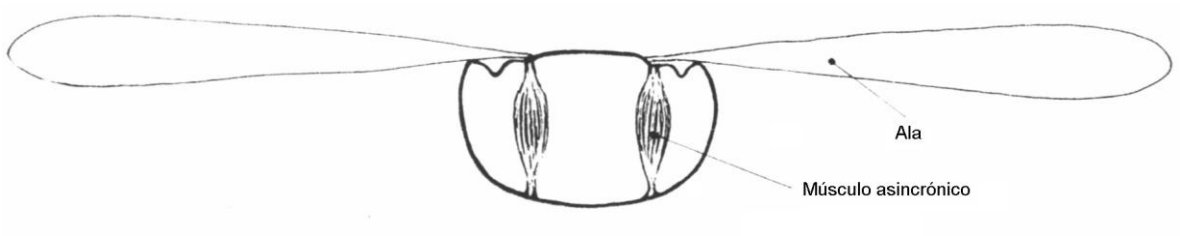


Figura 15: Tórax de un insecto

Por otra parte, Dejonckére aprecia una diferencia física en el ciclo vibratorio entre la elongación del pliegue vocal y la onda de presión infraglótica, desfase que mantiene la oscilación al proporcionar a cada ciclo la energía suficiente.

Dejonckére propone la siguiente ecuación:

Presión infraglótica – Presión supraglótica = Debito transglótico * Impedancia glótica

Y añade que la impedancia glótica depende de:

- La frecuencia y la amplitud oscilatoria de los bordes libres de los dos pliegues vocales
- La longitud de la parte vibrante de la glotis
- La orientación del eje de oscilación del borde libre de cada uno de los pliegues vocales
- La duración de la fase de adhesión de los pliegues vocales

Este concepto explica que disminuya el rendimiento vocal cuando el tejido del pliegue vocal pierde su flexibilidad por motivos patológicos, y justifica asimismo que alteraciones morfológicas, como pólipos alteran el funcionamiento vocal.

Fisiología de la articulación del habla

Solo se contemplara de manera básica los movimientos elementales de los órganos que componen el pabellón faringobucal, ya que la complejidad de los movimientos del pabellón faringobucal es enorme, como puede comprobarse mediante la revisión de las posibilidades de desplazamiento de cada uno de los órganos móviles que lo componen.

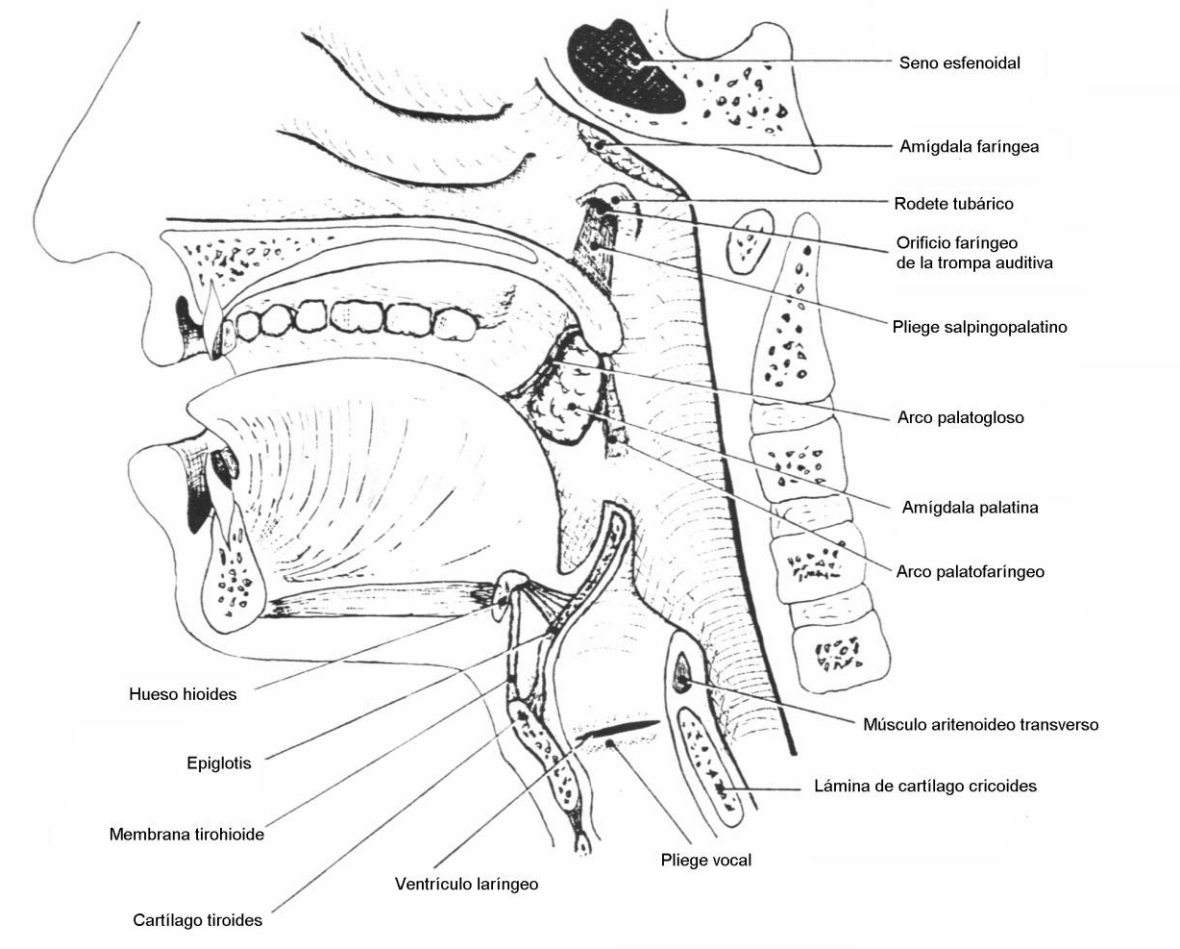


Figura 16: Resonadores

Labios

En situación de reposo, los labios mantienen en principio un simple contracto entre sí, pero este puede incrementarse si se eleva su tono muscular. Pueden separarse uno de otro, con lo cual la cavidad bucal se pone en comunicación con el exterior. Esta separación puede producirse por un descenso de la mandíbula, pero también, si esto no sucede, cuando se descubren más o menos los dientes.

Al alargarse, los labios pueden oponerse a la apertura de la boca, aunque descienda la mandíbula. Las comisuras labiales pueden alejarse una de otra, si se estiran los labios o, al contrario, acercarse. Los labios también pueden invertirse, contactando con su borde los incisivos, "recogerse", ser mordisqueados, etc.

Mandíbula

Cuando desciende la mandíbula, aumenta el volumen de la cavidad bucal. Por otra parte, sus movimientos repercuten en la posición del labio inferior y, sobre todo, en la lengua. Asimismo, como ya hemos comentado, puede realizar otros múltiples movimientos laterales, hacia atrás, adelante, etc.

Lengua

La lengua puede extenderse o extenderse lateralmente. Su cara superior puede ahuecarse más o menos hasta formar un canal mediante el enrollamiento lateral de sus bordes. Los bordes pueden contactar con las encías o con las arcadas dentales superiores para

obturas. Su punta puede ser impulsada hacia delante para salir de la boca o aplicarse contra las encías (o contra los dientes), Asimismo, puede curvarse hacia arriba y abajo, o realizar movimientos laterales. Su raíz puede impulsarse hacia atrás, hacia la pared posterior de la laringe. Su dorso puede elevarse y entrar en contacto con el paladar y con el velo palatino, a fin de ocluir la comunicación entre la faringe y la boca.

Mejillas

Pueden dejarse distender por la presión del aire bucal, o ser aspiradas por una presión negativa intrabucal. También pueden ejercer presión hacia el interior cuando se contrae su musculatura.

Velo del paladar

Puede elevarse para obturar la comunicación entre la nasofaringe y la cavidad nasal, mientras que a la vez amplia el paso entre la faringe y la cavidad bucal. A la inversa, puede descender e interrumpir así, juntamente con la elevación del dorso de la lengua, la comunicación entre la faringe y la cavidad bucal.

Faringe

Es capaz de estrecharse lateralmente en ambos ejes gracias a la acción de los músculos constrictores de la faringe y la retropulsión de la base de la lengua. Los movimientos de inclinación anterior y de retroceso de la cabeza también pueden reducir su diámetro

.

La dimensión vertical de la faringe varía con los movimientos de elevación y de descenso de la laringe. Este volumen puede aumentas asimismo con la separación de las mandíbulas y con la tensión de sus músculos, como sucede en el bostezo.

Narinas (orificios nasales externos)

Por ultimo, las narinas pueden dilatarse o estrecharse en mayor o menor grado. Si durante el habla sus movimientos son acentuados y más o menos constantes, esto indica la existencia de un proceso patológico.

Los seis puntos de articulación

En la práctica, puede considerarse que los ruidos producidos por los órganos fonatorios de una persona surgen de seis puntos, más o menos precisos, que pueden denominarse "elementos del habla" y que son:

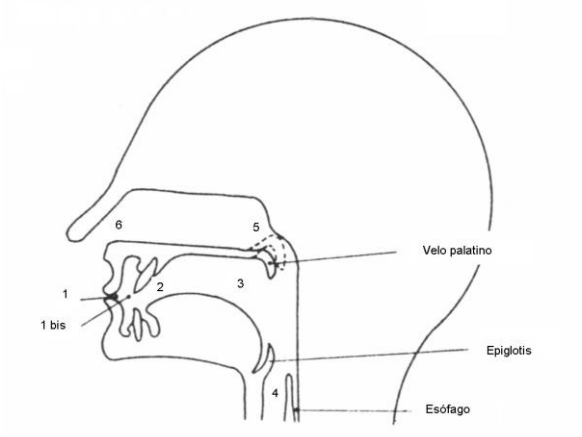


Figura 17: Seis puntos de articulación

- El 1 son los dos labios, su variante 1 Bis la constituyen el labio inferior, por una parte, y el borde inferior de los incisivos superiores, por otra.
- El 2 es la parte más anterior del dorso de la lengua y la cara posterior o de las encías superiores.
- El 3 esta formado por el dorso de la lengua y por el paladar
- El 4 esta constituido por los pliegues vocales
- El 5 lo integran el velo del paladar y el "techo" de la nasofaringe.
- El 6 lo componen los orificios de las narinas o la hilera nasal (este sexto elemento no interviene en el habla normal)

En cada uno de estos elementos pueden producirse ruidos de tres formas: *escape*, *explosión* y *vibración*. Como ya hemos comentado, los órganos articuladores del habla son susceptibles de obstaculizar el flujo de aire pulmonar, que puede ser frenado (ruido de escape), interrumpido completamente y luego desbloqueado (ruido de explosión), o inducir una vibración de uno u otro de estos órganos.

Punto de vista mecánico y punto de vista acústico

Existen diferencias entre nuestro esquema de producciones fónicas y fonéticas y aquellos otros que exponen la clasificación de los fonemas desde el punto de vista de la fonética clásica, lo cual se relaciona con el hecho de que él nuestro se refiere exclusivamente a las formas de producción fonética, basándose en el estudio del instrumento del habla, mientras que la fonética se interesa, en principio, por los ruidos producidos y se preocupa ampliamente de las características diferenciales de los fonemas, es decir, de las características acústicamente perceptibles que permiten distinguir los fonemas entre si.

Sin embargo, no es difícil establecer correspondencias entre estos dos enfoques de los acontecimientos fonéticos. Las consonantes fricativas o constrictivas corresponden a un ruido de escape; las oclusivas, a un ruido de explosión; las sonoras, a la adición de la vibración laringea; las nasales y laterales, a un "mecanismo complejo", etc.

Respecto a los puntos de articulación, cabe apreciar que corresponden a los "elementos del habla" del esquema anterior, donde:

Tipo	Elemento del habla
Bilabial	1
Labiodental	1 Bis
Apicodental	2
Predorso-prepalatal, dorsopalatal o posdorsovelar	3
Laringea	4

Es más difícil establecer las demás correspondencias, dado que las diferencias básicas continúan siendo, a pesar de todo, algo distintas y la sustitución de la designación exacta por los puntos de articulación es solo una ejemplificación, no es materia de este documento entrar en detalles excesivamente rigurosos que puedan constituir un obstáculo mas que un auxiliar de entendimiento.

Resumen: La formación de la voz

Ya vimos que el sistema fonador humano es fisiológicamente similar en todas las personas, pero tiene a la vez características muy especificas a cada individuo, dándose aspectos como diferentes tonos de voz (soprano mezzosoprano, contralto, tenor, barítono y bajo que describen el tono en el canto), vicios del habla o lenguaje como arcaísmos, barbarismos, hiato, etc. o muchas otros aspectos interesantes que individualizan la voz de cada persona, haciendo de esta un medio de identificación certero hasta en un 99.16 % de confiabilidad (fuente FBI) en condiciones favorables, como una grabación de buena duración y con una calidad de regular a buena.

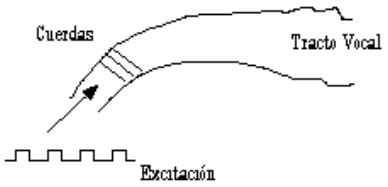


Figura 18: Modelo simple de la formación de la voz

En resumen la voz se forma mediante una señal de excitación, que proviene de nuestros pulmones (pulsos de aire a una frecuencia T_0), dándonos nuestro tono fundamental, esta a su vez excita a los pliegues vocales, los cuales vibraran aunados a esta señal de excitación, formando un sonido, de manera similar a las cuerdas de una guitarra.

Pero esta señal carece de la inteligencia necesaria para poder transmitir información, que es la función primordial de la voz, por lo que necesitamos de una serie de filtros, que selectivamente amplifican ciertas frecuencias, dejan pasar otras intactas y atenúan otras mas, generando estas lo que conocemos ya como voz, dichos filtros están conformados por las cavidades bucales, senos nasales, etc. y cada persona manipula dichos resonadores y el flujo del aire o excitación para formar su voz.

También es importante mencionar que no todos los sonidos de nuestra voz provienen o utilizan las cuerdas vocales, a los sonidos en los que intervienen las cuerdas vocales se les llama *sonidos guturales*, pero también tenemos los llamados *sonidos no guturales* o *sonidos silbantes*, y estos son producidos por el paso del aire por algunas de las cavidades antes mencionadas, mas no involucran las cuerdas vocales.

Por ultimo y aunque no es tema de interés del presente documento, recordemos que nuestra voz también es una manifestación sonora sumamente compleja que puede transmitir un mensaje inteligente, entendible y descifrable, que es el fin primordial de nuestra voz, permitiéndonos la comunicación de ideas y mensajes, y existen múltiples maneras de estudiar esta manifestación, como veremos a continuación.

¿Qué es la fonética?

Esta es una rama de la lingüística que estudia la producción, naturaleza física y percepción de los sonidos de una lengua. Sus principales ramas son: fonética experimental, fonética articulatoria y fonemática o fonética acústica.

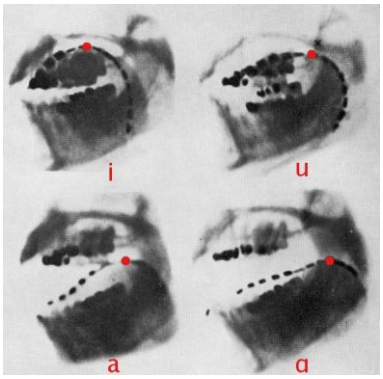
Fonética experimental

Es la que estudia los sonidos orales desde el punto de vista físico, reuniendo los datos y cuantificando los datos sobre la emisión y la producción de las ondas sonoras que configuran el sonido articulado. Utiliza instrumentos como los rayos X y el quimógrafo, que traza las curvas de intensidad. El conjunto de los datos analizados al medir los sonidos depende únicamente de la precisión del instrumental así como de otros conocimientos conexos. También se han descubierto diferencias importantes en cada sonido oral.

Fonética articulatoria

Es la que estudia los sonidos de una lengua desde el punto de vista fisiológico, es decir, describe qué órganos orales intervienen en su producción, en qué posición se encuentran y cómo esas posiciones varían los distintos caminos que puede seguir el aire cuando sale por la boca, nariz, o garganta, para que se produzcan sonidos diferentes. No se ocupa de todas las actividades que intervienen en la producción de un sonido, sino que selecciona sólo las que tienen que ver con el lugar y la forma de articulación. Los símbolos fonéticos y sus definiciones articulatorias son las descripciones abreviadas de tales actividades. Los símbolos fonéticos que se usan más frecuentemente son los adoptados por la Asociación Fonética Internacional en el alfabeto fonético internacional (A.F.I.) que se escriben entre corchetes.

Los órganos que intervienen en la articulación del sonido son móviles o fijos. Son móviles los labios, la mandíbula, la lengua y las cuerdas vocales, que a veces reciben el nombre de órganos articulatorios. Con su ayuda el hablante modifica la salida del aire que procede de los pulmones. Son fijos los dientes, los alvéolos, el paladar duro y el paladar blando. Los sonidos se producen cuando se ponen en contacto dos órganos articulatorios por ejemplo el bilabial [p], que exige el contacto entre los dos labios; también cuando se ponen en contacto un órgano fijo y otro articulatorio, y el sonido se nombra con los órganos que producen la juntura, o punto de articulación, como por ejemplo el sonido labiodental [f] que exige el contacto entre el labio inferior y los incisivos superiores. Cuando es la lengua el órgano móvil no se hace referencia a ella en la denominación del sonido, así el sonido [t] que se produce cuando la lengua toca la parte posterior de los incisivos superiores se llama dental.



El modo de articulación se determina por la disposición de los órganos móviles en la cavidad bucal y cómo impiden o dejan libre el paso del aire. Esta acción puede consistir en la interrupción instantánea y completa del paso del aire para las implosivas; en dejar abierto el paso nasal pero interrumpido el oral para las nasales; en producir un contacto con la lengua pero dejar libre el paso del aire a uno y otro lado para las laterales; en

producir una leve interrupción primero y dejar el paso libre después para las africadas; en permitir el paso del aire por un paso estrecho por el que el aire pasa rozando para las fricativas, y en permitir el paso libre del aire por el centro de la lengua sin fricción alguna para las vocales.

Se emiten diferentes clases de vocales según varíe la posición de la lengua, tanto a partir de su eje vertical (*alta, media y baja*), como a partir de su eje horizontal (*anterior, central y posterior*). Por ejemplo, en español son vocales altas las vocales de la palabra *huir,* es decir, la [i] y la [u]. Son vocales medias la [e] y la [o], es decir las vocales de la palabra *pero* y es vocal baja la [a] de la palabra *va.* Así, la lengua va de abajo arriba para pronunciar las dos vocales seguidas de la palabra *aire,* pero desciende a una posición media para pronunciar su última vocal. Hace el camino contrario de arriba abajo para pronunciar *puerta.* Son vocales anteriores del español la [i] y la [e], es decir las vocales seguidas de la palabra *piel;* las vocales posteriores son la [o] y la [u], es decir las vocales de la palabra *puro;* la [a] es la vocal central. La lengua se mueve de atrás hacia adelante para emitir las vocales de la palabra *totales,* hace el camino contrario para emitir las vocales de la palabra *piélago.* Las posiciones que mantiene la lengua para emitir las vocales [u], [i] y [a] constituyen los vértices del llamado esquema vocálico *uai.*

Fonemática

Es el estudio de los sonidos en el discurso, es decir, de los fonemas que son las unidades mínimas distintivas.

Por ejemplo, entre las palabras *las* y *los* sólo existe una diferencia de significado y de forma que es la que representa la distinción entre los fonemas [a] y [o]. Lo mismo sucede entre *pala, para, paga, pana* y *pasa,* las diferencias de significado se apoyan en los diferentes fonemas que las distinguen, esto es, [l], [r], [ɣ], [n] y [s]. Los fonemas están configurados también por unidades mínimas que los diferencian entre sí y son los llamados rasgos distintivos. La única diferencia que existe entre el fonema [p] que corresponde a una consonante bilabial, oclusiva, sorda y el fonema [b] que corresponde a una consonante bilabial, oclusiva sonora, es su modo de articulación: sorda la primera, frente a la segunda que es sonora. No siempre se mantienen como fonemas distintos las diferencias que proceden de un solo rasgo distintivo, por ejemplo la primera *d* de la palabra *dedo* corresponde a una consonante dental *oclusiva* sonora, y la segunda es dental *fricativa* sonora. En este caso no estamos ante dos fonemas sino ante dos valores del mismo fonema; a veces dos fonemas diferentes en una lengua dada son el mismo en otra, por ejemplo el español mantiene la diferencia fonética entre los sonidos [r] y [l], pero el japonés no ni el habla andaluza tampoco. De acuerdo con todo esto hay que distinguir entre fonemas y letras, aunque existen muchas coincidencias también hay desacuerdos muy importantes que apoyan esta diferencia. El fonema es un concepto ideal que está representado por unos signos escritos, las letras, aunque no todas representan un fonema. La letra *v* del español actual corresponde al fonema [b] que es una consonante bilabial, oclusiva, sonora; pero el fonema [ʙ] que corresponde a una consonante labiodental, fricativa, sonora ha

desaparecido en el sistema fonético actual, aunque estuvo presente en la historia de la lengua hasta el siglo XVIIII, y todavía hoy se usa en algunos países de América del Sur. Además hay letras que no representan fonema alguno como es el caso de la letra *h* que es muda en nuestra lengua. La escribimos como recuerdo histórico de una aspiración o de una f inicial del latín, pero no tiene valor fonético. Por otro lado, algunas letras expresan distintos fonemas, como la *c*, [ɵ] y [k] en España, y [s] y [k] en Latinoamérica y zonas de Andalucía.

Fonética acústica

Es la que estudia la onda sonora como la salida de un resonador cualquiera; esto es, equipara el sistema de fonación con cualquier otro sistema de emisión y reproducción de sonidos. En la comunicación, las ondas sonoras tienen un interés mayor que la articulación o producción de los sonidos, para un determinado auditorio recibe y descodifica la impresión a pesar de que haya sido emitida por medio de una articulación oral, o por medio de un determinado aparato emisor de sonidos o incluso por medio de un perico. Para grabar las características más significativas de las ondas sonoras y para determinar el resultado de las distintas actividades articulatorias se puede emplear el espectrógrafo. De forma experimental, para poder llegar a saber cuáles son los rasgos necesarios y suficientes que identifican los sonidos de la lengua.

Definiciones básicas

A continuación se darán una serie de definiciones que resultaran en un lenguaje técnico comprensible para futuras referencias o temas a tratar, ya que como en toda parte del saber humano, la fonética cuenta con sus propias acepciones y definiciones, mismas con las que debemos de familiarizarnos a fin de un mejor entendimiento y una fácil comunicación, ya que como se ha mencionado y sé a podido ver de los diversos temas tratados, el análisis de voz es una tarea multidisciplinar.

Comunicación: Desde el punto de vista psicológico es la respuesta discriminatoria de un organismo a un estimulo.

Código: Un de los elementos que integran el sistema de comunicación, y esta definido como el conjunto de reglas no ambiguas, previamente convenidas, por medio de las cuales los mensajes se convierten de una representación en otra.

Lengua: Es el sistema de signos que emplea una comunidad lingüística como instrumento de comunicación.

Habla: Es el uso individual que cada persona realiza del modelo general de la lengua.

Fonología: Es la disciplina que estudia los sonidos dentro de una lengua, establece las normas para su ordenamiento.

Fonema:

Es el concepto básico de la fonología. Por medio de este término designamos un conjunto de aquellas propiedades recurrentes que se usan en una lengua dada para distinguir palabras de distinto significado.

Desde el punto de vista psicológico se define como “aunque los hablantes produzcan diferentes sonidos y los oyentes los perciban como objetivamente diferentes, no son

conocedores de tal diferencia; el hablante cree producir el mismo sonido, y el oyente la impresión de oír el mismo sonido"

Como una realidad física, el fonema puede corresponder a alguna peculiaridad o particularidades, características de todos los sonidos en cuestión y solo de ellos. Cada fonema tiene una serie de características, y no se pueden repetir en conjunto para un determinado fonemas, no son divisibles pero sí substituibles.

El fonema es la unidad fonológica más pequeña. Su número es reducido. No tiene significado por si mismo, pero el significado de una palabra cambia si se intercambian dos fonemas; esto da lugar al fenómeno de la oposición.

Morfema: Unidad lingüística mínima con significado, es una colección de fonemas, es indivisible en unidades con significado.

Lexia: Unidad superior al morfema, es una unidad superior y es la palabra, unidad léxica de la lengua, ya constituida.

Frase: Es la unidad superior a la lexia, es la unidad más grande de la descripción gramatical

Variantes de los fonemas:
Los fonemas no son realizados siempre de la misma manera, por lo que estas diferencias dependerán del estilo del habla y/o del contorno fonético, mas aunque varían estos factores no se modifica ningún cambio de significado.

a) Variantes combinatorias (alófonos); cuando dos o más unidades fónicas que tienen semejanza articulatoria o acústica, no se presenta nunca en el mismo entorno.
b) Variantes libres; Cuando dos o más variantes aparecen en el mismo contorno.
c) Variantes individuales; cuando la realización de un fonema puede dar indicaciones sobre el hablante, pero no son el resultado de una elección por su parte.

Campo de dispersión: No deberán de sobrepasarse en sus realizaciones los fonemas de los limites acústicos y articulatorios (márgenes de seguridad) que estén condicionados por los campos de dispersión de los demás fonemas del sistema fonológico de la lengua.

Archí-fonema: Conjunto de particularidades distintivas que son comunes a los dos fonemas neutralizados y se representa por medio de una letra mayúscula.

Alófono: Son las diferentes realizaciones de un mismo fonema según el entorno en el que este situado. El significado de la palabra no cambia por el intercambio de alófonos.

Suprasegmental (o prosódica): En el nivel de la expresión se establece la dicotomía entre los fonemas y los otros elementos, como el acento, la entonación, el tono, etc. Los suprasegmentos del español son dos: el acento y la entonación.

*Dicotomía***:** Método de clasificación en que las divisiones y subdivisiones solo tienen dos partes. Aplicación de este método; división en dos.

*Grafía***:** Modo de escribir o representar los sonidos y, en especial, empleo de tal letra o tal signo grafico para representar un sonido dado.

Coarticulación: Fenómeno que se presenta en el habla cotidiana, es la variación en la realización espectral de un fonema por el contexto que lo rodea, ya que existen interconexiones entre los distintos fonemas. Este Fenómeno al observarse su fuerte variación con respecto del hablante hace pensar que dicho aspecto puede ser considerado como un parámetro de identificación del hablante (especialmente en la "m"), abordado de nuevo otras publicaciones en que se menciona que no es suficiente como para establecer una identificación, pero aporta mucho.

*Oclusiva***:** Consonante explosiva, la cual se produce cerrando momentáneamente la salida del aire en algún lugar de la boca, como *p*, *t*, *k*, *m*, *b*, *d*, etc.

*Fricativa***:** Producido por la fricción del aire al pasar entre dos órganos bucales que se acercan hasta formar una abertura muy estrecha, por ejemplo *f*, *j*, *s*, *z* y *c*.

*Africada***:** Consonante que resulta de combinar una oclusión con una fricción verificadas en el mismo lugar de articulación, con los mismos órganos, y con una duración aproximadamente igual a la de un sonido oclusivo, como la *ch* y *y*.

*Nasal***:** Consonante en cuya producción el aire espirado pasa totalmente o parcialmente por la nariz, como la *m*, *n* y *ñ*.

*Lateral***:** Consonante en cuya articulación la lengua impide al aire espirado su salida normal por el centro de la boca, dejándose pasar por uno o por los dos lados, como en la *l* y *ll*.

*Vibrante***:** Son los sonidos cuya pronunciación se caracteriza por un rápido contacto oclusivo, simple o múltiple, entre los órganos de la articulación, como la *r* y *rr*.

*Liquido***:** Dícese de la consonante que, precedida de una muda y seguida una vocal, forma silaba con ellas. En español, la *l* y la *r* son las únicas letras de esta clase. Ambas forman silabas con la *b*, la *c*, la *f*, la *g*, la *p* y la *t*. La *r* la forma además con la *d*.

Diferencias entre la trascripción fonológica y fonética

1) La trascripción fonológica es la reproducción grafica de la constitución fonológica de una lengua dada, abstracción hecha de la diversidad de los sonidos que realizan esta constitución del habla.
2) La trascripción fonética es la representación grafica de los medios más diversos que realizan la constitución fonológica de una lengua. Incluso la trascripción fonética más precisa solo tiene el valor de un instrumento auxiliar, ya que es incapaz de expresar toda la riqueza de matices articulatorios y acústicos que representa el habla viva y, por consiguiente, no puede sustituir a los análisis de la fonética experimental.

   Para este fin podemos dar una definición breve de que es la fonética y la fonología para apreciar mejor este criterio de distinción.

*Fonética***:** Alfabeto o escritura cuyos signos transcriben exactamente los sonidos, por ejemplo la ortografía, y que se basa en la pronunciación y no en la etimología. Rama de la lingüística que estudia los sonidos del lenguaje en su realización física.

*Fonológica***:** Rama de la lingüística, que a diferencia de la fonética, estudia los fonemas no es su realización física, sino en sus valores funcionales, esto es, diferenciadores dentro del sistema propio de cada lengua.

El análisis de la lengua se realiza a tres niveles

- Nivel fonológico; se estudian las unidades lingüísticas mínimas: fonemas. El conjunto de los fonemas se establecen por oposición, es decir, si se cambia un sonido de una palabra y la palabra cambia de significado, al sonido se le considera fonema.
- Nivel morfosintáctico; se estudian las palabras estableciendo su genero, numero y tiempo y las relaciones entre ellas.
- Nivel semántico; se estudia el significado de las frases y su coherencia.

Clasificación de los fonemas

Por su punto de articulación

Consonantes:

- *Bilabiales*: contactan los labios superiores e inferiores [p, b, m]
- *Labiodental*: contacta el labio inferior con los incisivos superiores [f]
- *Linguodentales*: contacta el ápice de la lengua con los incisivos superiores [t, d]
- *Linguointerdentales*: se sitúa el ápice de la lengua entre los incisivos superiores e inferiores [θ]
- *Linguoalveolares*: contacta el ápice o predorso de la lengua con los alvéolos [l, s, n, r, r̂]
- *Linguopalatales*: contacta el predorso de la lengua con el paladar duro [t͡ʃ, y, ɲ, ʎ]
- *Linguovelares*: se aproxima o toca el postdorso de la lengua con el velo del paladar [x, k, g]

Vocales:

- *Anterior*: la lengua se aproxima a la región delantera o zona del paladar duro [i, e]
- *Centrales*: la lengua se encuentra en la parte central del paladar [a]
- *Posteriores*: la lengua se aproxima a la zona velar [o, u]

Por el modo de articulación

Consonantes:

- *Oclusivas*: se establece un cierre completo de los órganos articulatorios y el aire sale de forma explosiva tras la interrupción [p, t, k, b, d, g]
- *Fricativas*: existe un estrechamiento de dos órganos articulatorios done pasa el aire espirado [f, ɵ, s, y, x]
- *Africadas*: se forma por combinación de una oclusiva seguida de una fricativa [t͡ʃ]
- *Nasal*: se establece un cierre de los órganos articulatorios y el aire sale por los conductos nasales [m, n, ɲ]
- *Laterales*: durante su emisión el aire se escapa por un lado o por los dos de la lengua [l, ʎ]
- *Vibrantes*: se produce una o varias vibraciones del ápice de la lengua [r]
- *Vibrante múltiple*: muy similar a la anterior pero continuada y acentuada[r̂]

Vocales:

- *Cerradas*: la lengua se encuentra muy cerca del paladar [i, u]
- *Medias*: la lengua esta en una distancia intermedia del paladar [e, o]
- *Abiertas*: la lengua se separa totalmente del paladar [a]

Por la vibración de las cuerdas vocales

- *Sordas*: no existe vibración de las cuerdas vocales [p, t, k, f, j, z, s, x, t͡ʃ]
- *Sonoras*: existe vibración de las cueras vocales [b, d, g, ʎ, m, n, y, l, ×, r, r̂] y [a, e, i, o u]

Por la acción del velo del paladar

- *Nasales*: el velo del paladar está separado de la pared faríngea [m, n, ɲ]
- *Orales*: el velo del paladar está unido a la pared faríngea y no permite el paso del aire hacia la cavidad nasal para el resto de los fonemas

AFI: Sistema internacional de trascripción

Vocales según AFI

Vocal	Grafía	Articulatorios	Acústicos
[i]	"i"	Alto, anterior	Vocálico, no consonántico, difuso, agudo
[e]	"e"	Medio, anterior	Vocálico, ni consonántico, denso, agudo
[a]	"a"	Bajo, central	Vocálico, no consonántico, denso

[o]	"o"	Medio, posterior	Vocálico, no consonántico, denso, grabe
[u]	"u"	Alto, posterior	Vocálico, no consonántico, difuso, grabe

Tabla 1: Vocales

Consonantes según AFI

Consonante	Grafía	Articulatorios	Acústicos
[p]	"p"	Oclusivo, bilabial, sordo	No vocálico, consonántico, difuso, grave, oral, interrupto, sordo, mate
[b]	"b" o "v"	Bilabial, sonoro	No vocálico, consonántico, difuso, grave, oral, sonoro
[t]	"t"	Oclusivo, dental, sordo	No vocálico, consonántico, difuso, agudo, oral, interrupto, sordo, mate
[d]	"d"	Dental, sonoro	No vocálico, consonántico, difuso, agudo, oral, sonoro
[k]	"c + a, o, u" o "qu + e, i" o "k"	Oclusivo, velar sordo	No vocálico, consonántico, denso, grave, oral interrupto
[g]	"g + a, o, u" o "gu + e, i"	Velar, sonoro	No vocálico, consonántico, denso, grave, oral sonoro
[f]	"f"	Fricativo, labiodental, sordo	No vocálico, consonántico, difuso, grave, oral, continuo, sordo, mate
[θ]	"c + e, i" o "z + a, o, u"	Fricativo, interdental, sordo	No vocálico, consonántico, difuso, agudo, oral, continua, sordo, mate
[s]	"s"	Fricativo, alveolar, sordo	No vocálico, consonántico, denso, agudo, oral, continuo, sordo, estridente
[ǰ]	"y" o "hi + vocal"	Fricativo, palatal, sonoro	No vocálico, consonántico, denso, agudo, oral, sonoro
[x]	"j + vocal" o "g + e, i"	Fricativo, velar, sordo	No vocálico, consonántico, denso, grave, oral, continuo, sordo, mate
[t͡ʃ]	"ch"	Africado, palatal, sordo	No vocálico, consonántico, denso, agudo, oral, interrupto, sordo, estridente
[m]	"m"	Nasal, bilabial, sonoro	No vocálico, consonántico, difuso, grave, nasal, continuo
[n]	"n"	Nasal, alveolar, sonoro	No vocálico, consonántico, difuso, agudo, nasal, continuo
[ɲ]	"ñ"	Nasal, palatal, sonoro	No vocálico, consonántico, denso, agudo, nasal, continuo
[r]	"r"	Vibrante simple, ápicoalveolar, sonoro	vocálico, consonántico, interrupto simple
[r̂]	"rr"	Vibrante múltiple, ápicoalveolar, sonoro	vocálico, consonántico, interrupto múltiple
[l]	"l"	Lateral, alveolar, sonoro	vocálico, consonántico, difuso, continuo
[ʎ]	"ll"	Lateral, palatal,	vocálico, consonántico, denso,

		sonoro	continuo

Tabla 2: Consonantes

Triangulo de las vocales

Derivado de múltiples trabajos de investigación, se puede ver que las vocales son sumamente parecidas, contando básicamente con el mismo tono fundamental de generación, variando exclusivamente en la articulación, por lo que se impulso un trabajo para clasificar las vocales conforme sus formantes F1 y F2, en la grafica se pueden ver también los campos de dispersión de dichas vocales, que se obtiene de la varianza de los sujetos adultos del sexo masculino con los que se obtiene la siguiente tabla para las cinco vocales tónicas en posición fonética normal.

Realización de	F_1	F_2
[i] en pípa	200 Hz	2100 Hz
[e] en pépa	324 Hz	1950 Hz
[a] en pápa	600 Hz	1300 Hz
[o] en pópa	324 Hz	729 Hz
[u] en púpa	200 Hz	620 Hz

Tabla 3: Formantes de las vocales

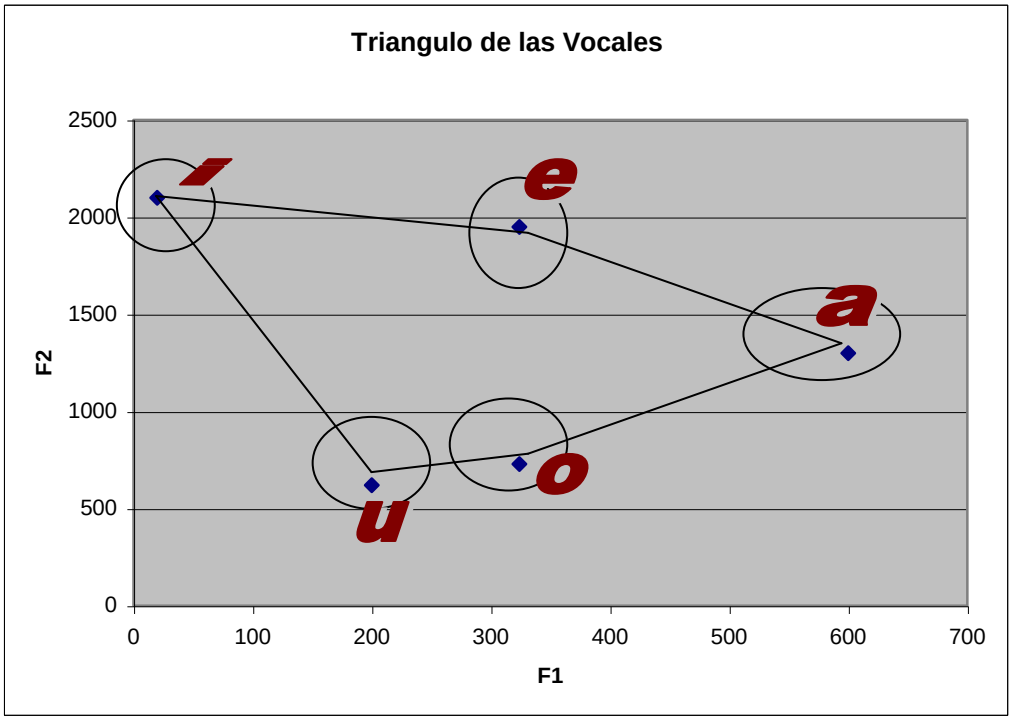


Figura 19: Triangulo de las vocales

Del triangulo de las vocales vemos que existen multitud de parámetros y comportamientos que se pueden medir y utilizar para estudiar la voz humana, que es compleja y no siempre fácil de utilizar como un elemento identificativo, sin embargo mas adelante veremos algunos métodos y técnicas por lo que abordaremos este fin por medio de una herramienta informática.

¿Cómo estudiar la voz humana en la computadora?

De lo ya expuesto se puede decir que tenemos una buena comprensión de cómo se genera nuestra voz, de que es un sonido, que tipos de sonido existen e incluso de la historia de la identificación de locutores, lo cual suena muy bien, sin embargo pretendemos lograr esto por medio de una herramienta informática, por lo que surge la pregunta básica, ¿Cómo hacemos esto?

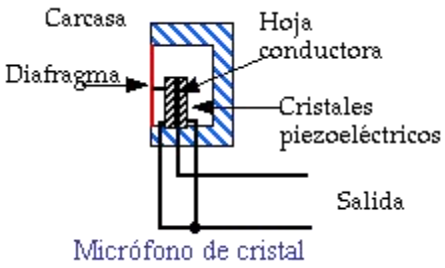
Transductores

Como primer punto tenemos que la voz humana es una señal física, es una serie de variaciones de presión del aire en un ambiente quasi estático (o de relativa estabilidad), las cuales al llegar a nuestro oído producen una serie de movimientos o vibraciones en el tímpano que posteriormente se transmitirán a la cóclea para que el cerebro los convierta de nuevo en una señal eléctrica que puede procesar y entender, sin embargo por ahora esto no es lo que nos interesa, sino mas bien como podemos transformar de primera instancia este sonido a una señal eléctrica que puedo conservar (grabar) o trabajar con una computadora.

El encargado de realizar dicho proceso es conocido como *micrófono*, que es un transductor electroacústico que convertirá las vibraciones sonoras u oscilaciones de presión del aire en una señal eléctrica, pero de hecho este cambio será realizado en dos etapas, un transductor acústico-mecánico y un transductor mecánico-eléctrico. Los micrófonos se clasifican según su directividad, según su transductor y por su utilidad y esta clasificación incluye:

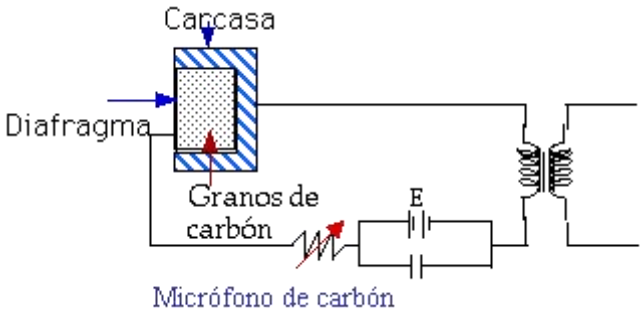
Micrófonos:

- Unidireccional
- Bidireccional o de gradiente
- Unidireccional
- Electrostáticos
- Dinámicos
- Piezoeléctricos
- De carbón
- Magnéticos
- De mano
- De estudio
- Etc.



Como el tipo de micrófonos no es el tema a tratar solo ejemplificaremos con un micrófono de carbón o piezoeléctrico el funcionamiento básico para comprender como se realiza este cambio de una señal acústica a una señal eléctrica.

En ambos casos se utilizaran las características eléctricas de un material determinado y como se comporta este a distintas presiones, en el caso de un micrófono de carbón se tendrá un deposito con carbón sellado por una membrana, y en el caso de un micrófono piezoeléctrico se tiene también un contenedor sellado con una membrana que contiene un cristal, tanto el carbón como el cristal cambiaran sus características propias con los cambios de presión a los que se someta la membrana.



Aquí es donde sucede el primer cambio, se someterá a la membrana a los ya mencionados cambios de presión o vibración sonora, y aquí tenemos el cambio acústico-mecánico, los cambios de presión ocasionaran que la membrana tenga una vibración o movimiento, ahora para el cambio mecánico-eléctrico tenemos lo siguiente.

En el caso del micrófono de carbón se cambiara la resistencia eléctrica del material contenido en la carcasa, y en el caso del cristal se generara una pequeña carga eléctrica al existir cambios de presión interna en el encapsulado producidos por la vibración mecánica de la membrana, en ambos casos este cambio será utilizado para generar una corriente eléctrica que será el resultado de la transducción.

Digitalización

Ahora ya tenemos una señal eléctrica, que es después de todo es con lo que trabajan las computadoras, mas sin embargo tenemos todavía otro contratiempo, la computadora en su inmenso poder y complejidad solo entiende un sistema binario de niveles eléctricos, es decir un CERO o apagado y un UNO o prendido, y la señal eléctrica producto de nuestro transductor esta muy lejos de ser así.

Para comprender mejor este asunto empezaremos por un par de definiciones:

Señal analógica: es aquella función matemática continua en la que es variable su amplitud y periodo en función del tiempo. Algunas magnitudes físicas comúnmente portadoras de una señal de este tipo son eléctricas como la intensidad, la tensión y la potencia.

Señal digital: es una señal donde las magnitudes de la misma se representan mediante valores discretos en lugar de variables continuas. Por ejemplo, un interruptor de luz, sólo toma dos valores: abierto o cerrado

Poniendo esto en palabras mas comunes y al alcance de todos, una señal analógica es aquella donde la señal puede tomar una infinidad de valores, por ejemplo, pensemos, si quiero saber cuantos números existen entre el 1 y el 9 muchos contestaran NUEVE, mas sin embargo otros dirán infinito, que es la respuesta adecuada, pues esta el numero 1.01, 1.001, y así puedo continuar hasta donde yo guste, y de igual manera para cada digito, que son los nueve dígitos que algunos pensaron.

De igual manera, un micrófono puede dar una infinidad de valores eléctricos, de ahí que decimos que produce una señal analógica, continua y con una infinidad de posibles valores, por lo que tenemos la necesidad de digitalizar esta para poderla trabajar en una computadora digital, por lo que procedemos con la digitalización ahora que sabemos que es necesaria.

La digitalización tiene su nombre correcto en la conversión analógica-digital y consiste en la trascripción de una señal analógica en una señal digital y esta conversión o trascripción de manera básica es realizar medidas de manera periódica sobre la amplitud de la señal analógica y traducirlas a un valor numérico, este proceso lo veremos con mas detenimiento en sus tres etapas que lo componen.

- Muestreo
- Cuantificación
- Codificación

Muestreo

El muestreo consiste en tomar muestras periódicas de la amplitud de onda. La velocidad con que se toman esta muestra o el número de muestras por segundo, es lo que se conoce como frecuencia de muestreo.

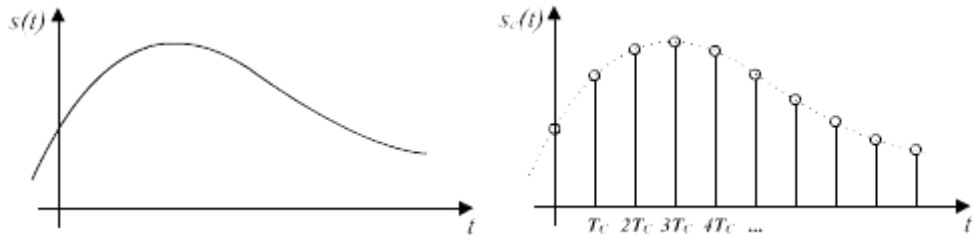


Figura 20: Señal original y señal muestreada

Como ya comentamos, la tasa o frecuencia de muestreo es el número de muestras por unidad de tiempo que se toman de una señal continua para producir una señal discreta o digital, generalmente se expresa en hercios (Hz o ciclos por segundo) y esta determinara la calidad con la que se captura el sonido apoyándonos en el teorema de Nyquist, por lo que aquí presentaremos de manera breve como seleccionar esta frecuencia de acuerdo a las necesidades.

Teorema de Nyquist-Shannon

Este teorema nos dice que para poder reconstruir con precisión la forma de onda de una señal digitalizada se requiere que la frecuencia de muestreo sea por lo menos el doble de la máxima frecuencia a muestrear, pero como en la realidad no existen los filtros paso-bajas ideales hay que dejar un margen de error entre la frecuencia máxima y la frecuencia de Nyquist. El hecho de que una señal puede ser reconstruida de manera exacta si se cumple con este teorema puede ser comprobado por medio de la transformada de Fourier, mas sin embargo no es tema de este documento entrar en el rigor matemático de este teorema.

Lo que hay que enfatizar es que no a mayor taza de muestreo se tendrá una mejor reconstrucción o mejor calidad de audio, una vez satisfecho el teorema de Nyquist no existirá diferencia, si tenemos una señal de 50 Hz, bastara con una frecuencia de muestreo de 120 Hz, por lo que una frecuencia de 250 Hz solamente utilizara mayores recursos y no presentara ganancia alguna en calidad o fidelidad de señal.

Ahora, respetando esta frecuencia mínima de muestreo y agregando el margen de seguridad ya mencionado tenemos que se manejan una serie de frecuencias ya

estandarizadas por la industria y que son como a continuación se mencionan con sus respectivas aplicaciones.

Frecuencia de muestreo	Aplicación típica
8000 Hz	Utilizada en telefonía y telecomunicaciones, es apta para trasmitir voz humana y permite reproducir señales de hasta 3.5 KHz
22050 Hz	Es la utilizada por la radio, permite la reproducción de señales de hasta 10 KHz
32000 Hz	Es utilizada por el formato DV de cámaras digitales
44100 Hz	Es el utilizado en los discos compactos musicales, permite reproducir señales de hasta 20 KHz
47250 Hz	Formato PCM, de uso digital permitiendo reproducir señales de hasta 22 KHz, es un formato escasamente utilizado
48000 Hz	Utilizado en la televisión digital, para los DVD y los DAT de audio profesional
96000 o 192400 Hz	Utilizado por el HD DVD, el Bluray y sistemas de audio de alta definición.

Tabla 4: Tabla de frecuencias comunes

Como podemos ver existen múltiples frecuencias ya estandarizadas para diferentes aplicaciones, existiendo incluso aquellas que permiten frecuencias en las señales a reconstruir mucho más altas que las que el oído humano puede captar, pero por lo que trata a este trabajo por lo general frecuencias de entre 8000 y 22050 Hz son mas que suficientes para reconstruir señales de voz, por lo que vemos que frecuencias de muestreo mayores para la aplicación en mano simplemente serian un desperdicio de recursos.

Efecto aliasing

Este efecto se presenta al utilizar una frecuencia menor a la establecida por el teorema de Nyquist, y es una distorsión conocida como *aliasing*; algunos autores traducen este término como solapamiento. El aliasing impide recuperar correctamente la señal cuando las muestras de ésta se obtienen a intervalos de tiempo demasiado largos. La forma de la onda recuperada presentaría pendientes muy abruptas.

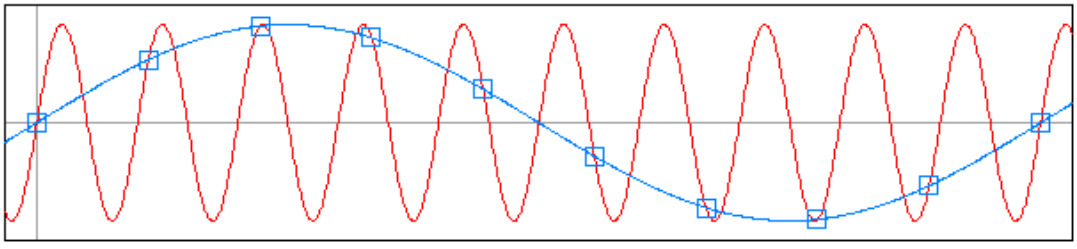


Figura 21: Distorsión o aliasing

En este ejemplo hasta cierto punto burdo la señal roja ejemplifica a la señal original a muestrear, y la señal azul seria la reconstrucción posible con una frecuencia de muestreo mucho mas baja de lo necesaria y como se generaría una reconstrucción totalmente errónea. Una pendiente abrupta genera cierta dispersión de la señal. Esta dispersión es la

responsable de que se generen desfases o desplazamientos temporales de la señal). El efecto aliasing y la dispersión que introduce quedaron demostrados por los experimentos de Lagadec y Stockham.

La cancelación del efecto aliasing es relativamente simple, la mayoría de los sistemas de digitalización incluyen filtros paso bajo, que eliminan todas las frecuencias que sobrepasan la frecuencia crítica en la señal de entrada. Es decir, todas las frecuencias que queden por encima de la frecuencia de muestreo seleccionada son eliminadas. El filtro paso bajo para este uso concreto recibe el nombre de filtro antialiasing. Sin embargo, abusar de los filtros antialiasing, puede producir errores graves, por lo que también debe de evitarse el sobre filtrar o filtrar recursivamente lo que puede degenerar la señal y provocar que la onda final presente una pendiente marcada. Por esta desventaja del filtro antialiasing se ha generalizado la técnica conocida como sobremuestreo de la señal.

Sobremuestreo

Para evitar las caídas abruptas se utiliza la técnica conocida como sobremuestreo (over sampling), que permite reconstruir, tras la conversión digital-analógica, una señal de pendiente suave.

Un sobremuestreo consiste en aplicar un filtro digital que actúa sobre el tiempo (dominio de frecuencia), cambiando de lugar las muestras, de forma que al superponerlas, se creen muestreos simultáneos virtuales. Estos muestreos simultáneos no son reales, son simulaciones generadas por el propio filtro. Estos muestreos simultáneos se obtienen utilizando el llamado coeficiente de sobremuestreo (n).

Las muestras obtenidas se superponen con los datos originales y los conversores analógico-digital los promedian, obteniendo una única muestra ponderada (por ejemplo, si se hacen tres muestreos, finalmente, la muestra tomada no es ninguna de las tres, sino su valor medio). Para evitar el aliasing, también se introduce a la entrada un filtro paso bajo digital, que elimine aquellas frecuencias por encima de la mitad de la frecuencia de muestreo. No obstante, a la salida, la frecuencia de muestreo utilizada para reproducir la señal ya no es la misma que se utilizó para tomar las muestras a la entrada, sino que es tantas veces mayor como números de muestreo se hayan hecho.

Como ejemplo podemos considerar la digitalización de música en formato CD. Imaginemos que para digitalizar el CD se hacen 3 muestreos a 44,1 kHz que se interpolan. Se introduce un filtro paso bajo, llamado "*decimator*", que elimina las frecuencias por encima de los 20 kHz, pero la frecuencia de muestreo utilizada para reconstruir la señal será tres veces mayor: 132,3 kHz. De este modo se reconstruye la señal suavizando la pendiente. A este proceso de filtrado durante la conversión digital-analógico se lo conoce como diezmado.

Es evidente que incorporar la técnica del sobremuestreo encarece considerablemente el equipo, sin embargo es utilizado en algunos equipos modernos y de ahí las muy altas frecuencias de muestreo utilizadas en formatos como el HD DVD y el Bluray, que en una forma de trabajo convencional sobrepasarían en mucho a la capacidad del oído humano, por lo que no representarían ganancia alguna en calidad o contenido.

Cuantificación

Lo que se hace básicamente es convertir una sucesión de muestras de amplitud continua en una sucesión de valores discretos preestablecidos según el código utilizado.

Por ejemplo;

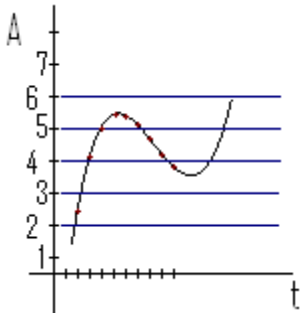


Figura 22: Niveles de cuantificación

En esta grafica 22 tenemos varios aspectos interesantes:

- Tenemos una grafica que representa a una señal continua, esta será la curva en la grafica
- Tenemos un eje que representa al tiempo con una frecuencia de muestreo *x*, a cada intervalo de tiempo se tomara la amplitud de la señal continua
- Por ultimo tenemos un eje de amplitud, que se hace discreto al dividirse en niveles representados por las líneas azules, a todo valor intermedio entre dos niveles se le asignara el inmediato inferior, por ejemplo, todo valor entre 2 y 3 será codificado como 2

Teniendo en cuenta estos puntos veremos que cada *t* segundos se tomara la amplitud de la señal (marcado como pequeñas cruces rojas), y se cuantificara de acuerdo a la regla definida, valores de 0 a 1 valen 0, valores de 1 a 2 valen 1 y así continuamos, de tal suerte que al terminar nuestro muestreo tendremos una serie de valores que nos dicen para nuestro ejemplo que la señal vale {1, 2, 4, 5, 5, 5, 5, 4…} el tiempo no se requiere mencionar, ya que para reconstruir la señal sabemos la frecuencia de muestreo *x* y podemos asumir para todo fin practico que empezamos de cero, y a cada tiempo determinado tenemos una muestra, pero como podemos ver en este rudimentario ejemplo la señal sufre perdida de resolución, por lo que es sumamente importante escoger la cuantificación adecuada para el problema.

Regresando a la explicación, durante el proceso de cuantificación se mide el nivel de tensión de cada una de las muestras, obtenidas en el proceso de muestreo, y se les atribuye a un valor finito (discreto) de amplitud, seleccionado por aproximación dentro de un margen de niveles previamente fijado.

Los valores preestablecidos para ajustar la cuantificación se eligen en función de la propia resolución que utilice el código empleado durante la codificación. Si el nivel obtenido no coincide exactamente con ninguno, se toma como valor el inferior más próximo.

En este momento, la señal analógica (que puede tomar cualquier valor) se convierte en una señal digital, ya que los valores que están preestablecidos, son finitos, pero es pertinente mencionar que todavía no se trata de valores binarios, sino tan solo de valores finitos, esto se realizara en la codificación.

Sin embargo, la señal digital que resulta tras la cuantificación es sensiblemente diferente a la señal eléctrica analógica que la originó, por lo que siempre va a existir una cierta diferencia entre ambas que es lo que se conoce como error de cuantificación que se produce cuando el valor real de la muestra no equivale a ninguno de los escalones disponibles para su aproximación y la distancia entre el valor real y el que se toma como aproximación es muy grande. Un error de cuantificación se convierte en un ruido cuando se reproduzca la señal tras el proceso de decodificación digital.

Para minimizar los efectos negativos del error de cuantificación, se utilizan distintas técnicas de cuantificación:

- *Cuantificación uniforme o lineal*. Se utiliza un bit rate constante. A cada muestra se le asigna el valor inferior más próximo, independientemente de lo que ocurra con las muestras adyacentes.
- *Cuantificación no uniforme o no lineal*. Se estudia la propia entropía de la señal analógica y se asignan niveles de cuantificación de manera no uniforme (bit rate variable) de tal modo que, se asigne un mayor número de niveles para aquellos márgenes en que la amplitud de la tensión cambia más rápidamente.
- *Cuantificación logarítmica*. Se hace pasar la señal por un compresor logarítmico antes de la cuantificación. Como en la señal resultante la amplitud del voltaje sufre variaciones menos abruptas la posibilidad de que se produzca un ruido de cuantificación grande disminuye. Antes de reproducir la señal digital, esta tendrá que pasa por un expansor.
- *Cuantificación vectorial.* En lugar de cuantificar las muestras obtenidas individualmente, se cuantifica por bloques de muestras. Cada bloque de muestras será tratado como si se tratara de un vector, de ahí, el nombre de esta tipología.

La calidad de la cuantificación para archivos digitales de audio es lo que se conoce como la profundidad de muestreo, esto es, con que precisión se cuantifica la amplitud de la señal, digamos que tenemos una señal analógica continua idéntica a la del ejemplo anterior, y con la misma frecuencia de muestreo *x*, pero ahora subimos la resolución al doble, que se ilustra con las líneas azules mas claras.

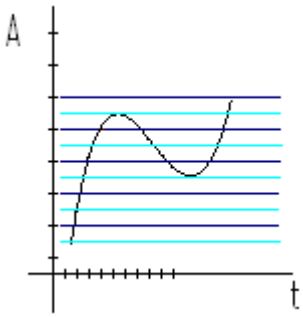


Figura 23: Cuantificación mejorada

Como podemos ver, se tiene que cada nivel de la cuantificación mide solo la mitad del ejemplo anterior, lo cual como podemos fácilmente ver al intentar encontrar de nuevo los valores de amplitud en los puntos de muestreo tendrán una mayor semejanza a la función original, ahora imaginemos subir la resolución de nuevo al doble y así sucesivamente, llegara un momento dado en que el tamaño de las zonas de cuantificación sea tan pequeñas que no tengamos perdidas importantes de señal.

La amplitud permitida de la señal para una digitalización de audio estará limitada por el rango de respuesta del digitalizador, por lo que todo nivel inferior al umbral bajo será ignorado y tomado por silencio y todo nivel superior al umbral máximo será tomado como el máximo valor del umbral mismo, lo que se conoce como saturación.

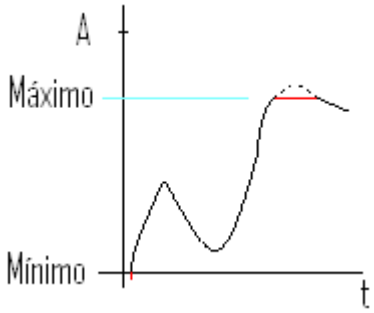


Figura 24: Pérdidas por saturación

En la grafica 24 se puede ver una señal cualquiera que se digitalizara, el mínimo valor de umbral del digitalizador cera cero y el máximo el indicado por la línea azul, por lo que las porciones de la señal marcadas con rojo serán perdidas, la primera por debajo de cero será ignorada y tomada por cero, y la porción de datos por encima del máximo se tomara como una serie de valores máximos, por lo que "recortara" la señal.

Algunos digitalizadores tienen controles para modificar su sensibilidad, pero en caso de no ser así se deberá de manejar el nivel de la señal con una especie de pre-amplificador o control de volumen en el aparato que reproduce la señal original, y una vez que tenemos una señal que respeta los valores permitidos por el digitalizador estamos preparados para ejemplificar por ultimo como nos afectara la mencionada profundidad de muestreo.

Sabemos que tenemos un nivel mínimo de señal, que normalmente se tomara como cero y un valor máximo cualquiera que llamaremos *Max*, ahora la profundidad de muestreo (que normalmente se mide en bits) nos dará es el numero de zonas o regiones en las que se dividirá este rango del digitalizador, por ejemplo, con 8 bits que es uno de los estándares mas antiguos por la formula 2^8 (que es la base elevada al numero de bits) que tenemos 256 posibles zonas o valores, pero para 16 bits tenemos 65536 niveles y para 24 bits 16777216 niveles, por lo que al no variar el rango del digitalizador (que podemos pensar como el tamaño de la señal) sino tan solo el numero de zonas, a mayor profundidad de muestreo menor será el tamaño de cada zona, y por lo tanto la señal digital se parecerá mucho mas a la original, en esta ocasión si es directamente proporcional, a mayor profundidad de muestreo mayor calidad.

Codificación

Este es un paso mas fácil de comprender, ya que primordialmente comprende el codificar los números discretos anteriormente obtenidos a lenguaje binario que las

computadoras pueden trabajar, existen otros sistemas, como octal, hexadecimal, mas sin embargo al ser de menor uso y no nativos a las computadoras digitales por ahora no los tomaremos en cuenta, pero también agregaran alguna información que permita la lectura y reconstrucción de los datos anteriormente procesados.

La codificación consiste en la traducción de los valores de tensión eléctrica analógicos que ya han sido cuantificados (ponderados) al sistema binario, mediante códigos preestablecidos. La señal analógica va a quedar transformada en un tren de impulsos digital, que es una sucesión de ceros y unos, y este proceso es llevado a cabo mediante un Códec específico.

Códecs y formatos de audio

El códec es el código específico que se utiliza para la codificación/decodificación de los datos, la palabra Códec es una abreviatura de Codificador-Decodificador.

Parámetros que definen un códec

- *Número de canales*: Indica el tipo de sonido con que se va a tratar: monoaural, binaural (estéreo) o multicanal
- *Frecuencia de muestreo*: La frecuencia es la cantidad de muestras de amplitud tomadas por unidad de tiempo en el proceso de muestreo como ya se comento.
- *Resolución*: Determina la precisión con la que se reproduce la señal original. Se suelen utilizar 8, 10, 16 o 24 bits por muestra. Esta resolución como ya se vio esta fuertemente ligada a la calidad de la misma.
- *Bit rate:* El bit rate es la velocidad o tasa de transferencia de datos. Su unidad es el bit por segundo (bps).
- *Pérdida:* Algunos códecs al hacer la compresión eliminan cierta cantidad de información, por lo que la señal resultante, no es igual a la original

Existe una gran cantidad de códecs para audio, entre ellos los de audio sin perdidas, los de audio con perdidas y los destinados a voz humana y algunos de ellos son:

- Apple Lossless (ALAC).
- Direct Stream Transfer (DST).
- FLAC (Free Lossless Audio Codec).
- LPAC (Lossless Predictive Audio Codec).
- Monkey's Audio (APE).
- OptimFROG.
- RealAudio Loseless.
- WavPack.
- MP1 (MPEG audio layer-1), MP2 (MPEG audio layer-2) y MP3 (MPEG audio layer-3)
- Ogg Vorbis
- WMA (Windows Media Audio).
- AC3 (Dolby Digital A/52).
- DTS (Digital Theather Systems).
- ADPCM.
- AMR.
- G.711 (Ley Mu y Ley A).
- G.722.

- G.723.
- G.726.
- GSM
- Perceptual Audio Coding (usado en radio digital y vía satélite).
- Speex (libre de patentes).

Como podemos ver realmente existen una gran cantidad de códecs de audio digital y por si fuera poco existen también una gran variedad de formatos o tipos de archivos, pudiendo incluso algunos de estos utilizar diferentes códecs en un mismo formato, en la investigación realizada se localizaron 280 formatos diferentes, algunos de uso muy común, otros muy especializados, y tan solo por ilustrar daremos una lista con los mas comunes de esa lista:

Extensión	Descripción
.aac	Archivo de codificación avanzada de audio, o por sus siglas en ingles Advanced Audio Coding
.aif	Formato de intercambio de audio o por sus siglas en ingles Audio Interchange File
.amr	Formato de múltiples tasas adaptable, es muy utilizado en los nuevos aparatos celulares, por sus siglas en ingles Adaptive Multi-Rate
.iff	Formato de intercambio o por sus siglas en ingles Interchange File Format
.midi	Formato de audio para MIDI que guarda información de notas e instrumentos, pero no del audio en si
.mp3	Formato de archivos MP3 con todas sus variantes como MP4 o M4A de Apple que son sumamente parecidos y tienen su principal uso en audio para Internet o reproductores de música portátiles
.mpa	Formato que usa la codificación MPEG para archivos de audio
.ra	Formato de audio comprimido utilizado principalmente en Internet y recibe su nombre por sus siglas en ingles Real Audio
.ram	Formato similar al anterior, recibe su nombre de las siglas en ingles Real Audio Media

.wav	Formato sumamente versátil, puede usar múltiples códecs y es muy utilizado a nivel profesional, es propiedad de Microsoft y recibe su nombre de la palabra en ingles WAVE
.wma	Formato propietario de Microsoft para venta de música por Internet o para audio en general, recibe su nombre de las siglas Windows Media Audio

Tabla 5: Formatos comunes de audio

Como no es tema de este trabajo ya no nos adentraremos mas en este aspecto, tan solo es conveniente mencionar que los códecs de audio con perdidas deben ser manejados con cautela, ya que las perdidas son irrecuperables, estas perdidas normalmente se basan en modelos preceptúales del oído humano y básicamente estipulan "si la mayoría de las personas no lo pueden distinguir no es necesario", lo cual también puede ejemplificarse con un par de calcetines, si estos van debajo del pantalón o dentro del zapato desde el punto de vista visual no son necesarios, podríamos solo utilizar bandas entre el zapato y el borde inferior del pantalón para tapar esta porción de la pierna y sirven al mismo propósito, sin embargo muchos de nosotros nos sentiríamos "descalzos".

¿Qué formato usaremos?

Para trabajar archivos de audio en la computadora vemos que de primera instancia debemos elegir un formato a utilizar, ya que como comentamos, hay mas de 280 para escoger, y si a esto le sumamos el numero tan grande de códecs que existen tenemos aun mas dificultades para ver cual utilizar, pero dado que esto no es lo que nos interesa atacar por el momento usaremos uno de los formatos de uso mas común y que permite trabajar audio sin compresión y por ende sin perdidas, utilizaremos el formato de Microsoft e IBM Waveform o WAV como comúnmente se le conoce, y a continuación explicaremos lo básico de dicho formato.

Formato de audio Waveform

WAVEform significa forma de onda, es un formato de audio digital normalmente sin compresión de datos desarrollado propiedad de Microsoft e IBM que se utiliza para almacenar sonidos en las computadoras personales, admite archivos mono y estéreo a diversas resoluciones y velocidades de muestreo, y la extensión de sus archivos es .wav.

Es una variante del formato RIFF (Resource Interchange File Format o formato de fichero para intercambio de recursos), es un método para almacenamiento en "paquetes". El formato toma en cuenta algunas peculiaridades de los procesadores Intel, y es el formato principal usado por Windows.

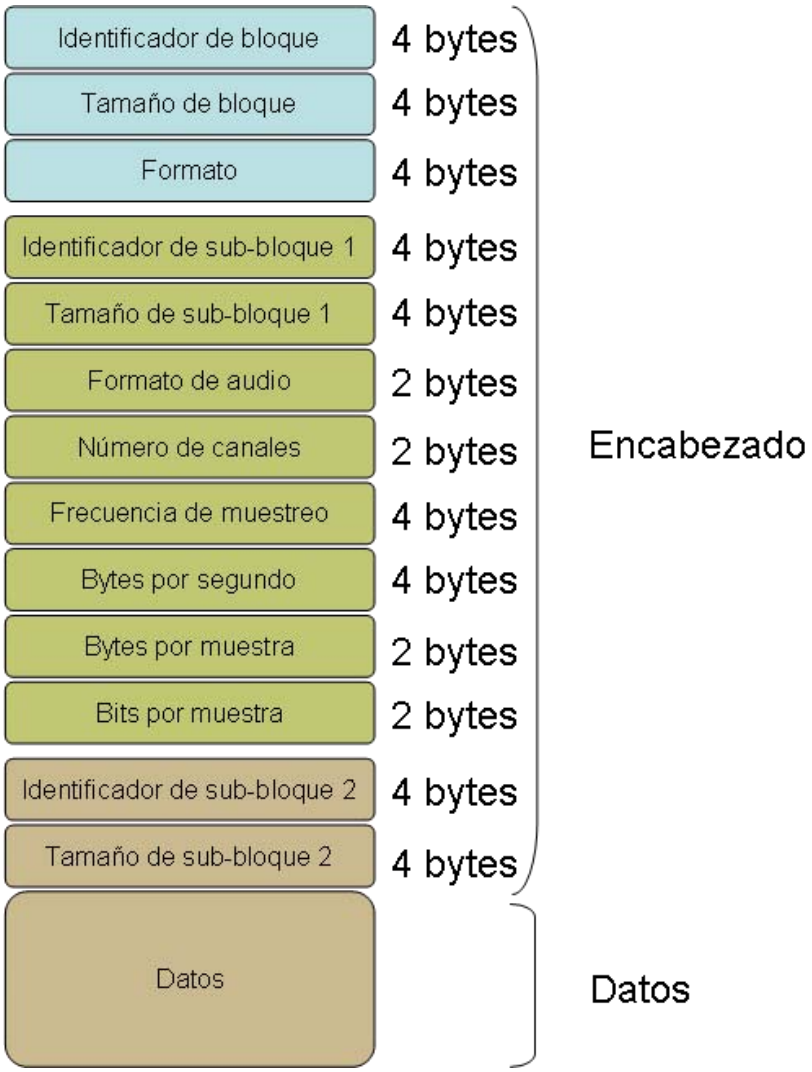
El formato WAV puede soportar casi cualquier códec de audio, pero utiliza primordialmente el formato PCM (no comprimido) y al no tener pérdidas es ampliamente utilizado por profesionales del audio. Para tener calidad disco compacto se necesita que el sonido se grabe a 44,100 Hz y a 16 bits, por cada minuto de grabación se utilizan aproximadamente 5 MB de disco duro, una de sus limitantes es que máximo puede manejar un archivo de 4 GB que equivalen aproximadamente a 6.6 horas en

calidad de disco compacto y se debe a que en la cabecera del fichero la longitud se guarda en un número entero de 32 bits, lo que limita el tamaño del archivo.

Pasando a lo que nos interesa veremos como esta integrado un archivo en formato WAV y como lo debemos de trabajar para lograr extraer la información o características que realmente nos interesan, en general veremos que estos archivos están formados por dos porciones, un encabezado y un bloque de datos de audio, y en ocasiones un bloque mas de datos adicionales que se colocan al final de los datos de audio, veamos ahora cada parte.

El encabezado

El encabezado de un archivo WAV esta conformado por 44 bytes y estos incluyen la siguiente información:



Esta información tiene el siguiente significado o uso, comienza por el encabezado RIFF que contiene:

- *Identificador de bloque*: contiene la palabra “RIFF” en código ASCII

- *Tamaño de bloque*: contiene el tamaño total del archivo WAV menos 8 bytes que corresponden a estos dos primeros campos de información, se podría decir que es del siguiente campo al final del archivo.
- *Formato*: contiene la palabra "WAVE"

A continuación se presenta el sub-bloque del formato del archivo, y este contiene:

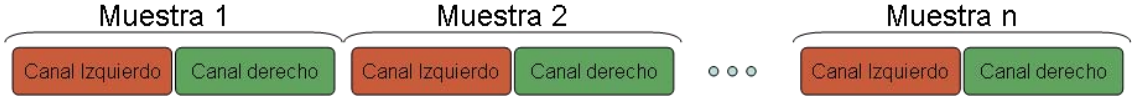
- Identificador del sub-bloque 1: contiene la palabra "fmt" que es la abreviación de *formato* en ingles.
- Tamaño de sub-bloque 1: vale 16 para PCM, es el número de bytes que le quedan al sub-bloque con información.
- Formato de audio: para PCM vale 1, si este numero es mayor de 1 significa que el archivo contiene algún tipo de compresión
- Numero de canales: Para una grabación monoaural vale 1 y una grabación estéreo vale 2, puede contener hoy en día más canales.
- Frecuencia de muestreo: contiene la frecuencia de muestreo que nos indica las máximas frecuencias registradas en el archivo, puede ser por ejemplo 8000, 44100, etc.
- Bytes por segundo: es el numero de bytes necesarios para cada segundo o lo que se conoce como "bit rate", es igual a $bytesPorMuestra * frecuenciaDeMuestreo$
- Bytes por muestra: numero de bytes por cada muestra incluyendo todos los canales, es igual a: $canales * \frac{bitsPorMuestra}{8}$
- Bits por muestra: este puede ser de 8, 16 o 24 bits hasta la fecha

A continuación se presenta el sub-bloque de datos del archivo, y este contiene:

- Identificador del sub-bloque 2: contiene la palabra "data "
- Tamaño de sub-bloque 2: contiene el tamaño de la porción de datos del archivo, con este fácilmente podemos calcular el numero de muestras en el archivo con este simple despeje $numeroDeMuestras = \frac{tamañoSubBloque2}{bytesPorMuestra}$

Los datos

A partir de este punto encontramos realmente lo que es la información de audio ordenada de la siguiente manera:



Para un archivo de 16 bits se tendrá que cada canal esta formado por dos bytes, para 24 bites estará formado por tres bytes, y por supuesto para 8 bites por un byte, cuando se trata de un solo byte no existe problema alguno, pero cuando se trata de mas de uno existe la pregunta

¿Cómo los ordenaremos?

Para responder esto tenemos que hacer un pequeño paréntesis cultural:

> *Endianness:*
> El término inglés Endianness designa el formato en el que se almacenan los datos de más de un byte en una computadora y como para casi todo en esta vida hay dos criterios. El primero se denomina little-endian y el segundo big-endian, y sus nombres nacen de la novela "Los viajes de Gulliver" de Jonathan Swift en la que los habitantes de los imperios de Lilliput y Blefuscu libran una encarnizada guerra por una disputa sobre el lado por el que debían abrir un huevo tibio, por el lado pequeño o grande y de ahí el termino big-end que da forma al termino moderno.
>
> El sistema big-endian adoptado por Motorola entre otros, consiste en representar los bytes en el orden "natural": así el valor hexadecimal 0x4A3B2C1D se almacenaría en memoria en la secuencia {4A, 3B, 2C, 1D}. En el sistema little-endian adoptado por Intel, entre otros, el mismo valor se almacenaría como {1D, 2C, 3B, 4A}.

Regresando a nuestra pregunta, y recordando que el Waveform trabaja con el formato adoptado por Intel que es little-endian, un numero como 65280 que en binario seria '1111111100000000' se almacenaría como {00, 00, FF, FF}.

Ahora que sabemos como se ordenara la información y en que orden podemos crear un programa de computo que no solo de lectura al encabezado, sino que una vez realizado esto de lectura a todos los datos de audio que es realmente lo que nos interesa.

Para este punto sabemos que es la voz humana, sabemos básicamente como se propaga, como se estudia y clasifica desde el punto de vista fonético, como se convierte en una señal eléctrica, después digital y como se almacena en por lo menos un formato de computadora, ahora veremos que hacer con estos datos.

¿Qué es el modelado?

El modelado es una técnica cognitiva que consiste en crear una representación ideal de un objeto real mediante un conjunto de simplificaciones y abstracciones, cuya validez se pretende constatar. La validación del modelo se lleva a cabo comparando las implicaciones predichas por el mismo con observaciones.

En otras palabras, se trata usar un modelo irreal o ideal, y reflejarlo sobre un objeto, como crear una escultura o representación del objeto real

Un modelo es una simplificación de la realidad, se recogen aquellos aspectos de gran importancia y se omiten los que no tienen relevancia

para el nivel de abstracción dado. Se modela para comprender mejor un sistema. Los sistemas complejos no se pueden comprender en toda su completitud (según el en foque del algoritmo Dijkstra-Scholten "divide y vencerás").

Los principios de modelado son:

- **Primero**: la elección de los modelos tiene una profunda influencia en el acometimiento del problema y en como se da forma a la solución
- **Segundo**: los modelos se pueden representar en distintos niveles de detalle, los analistas se suelen centrar en el qué, mientras que los diseñadores en el cómo
- **Tercero**: los mejores modelos se mantienen ligados a la realidad.
- **Cuarto**: un único modelo no es suficiente. Cualquier sistema no trivial se aborda mejor mediante un pequeño conjunto de modelos casi independientes, es decir, que se puedan construir y estudiar por separado pero que estén interrelacionados

Modelos de producción de voz

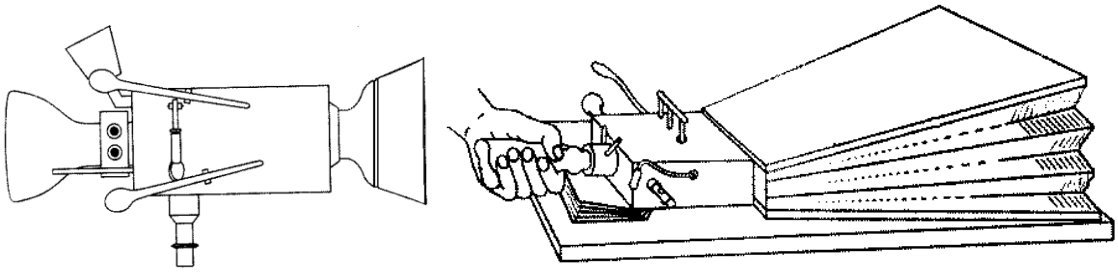
Por este motivo se presentar algunos modelos de la producción de la voz que nos serán de utilidad. Mucho del trabajo relacionado con el estudio de la voz es derivado de una extensa investigación en acústica analógica, parte de lo cual fue el modelado de la voz abordado a partir de mediados del siglo 20, por lo que una descripción completa de este tema seria demasiado extensa, pero los sistemas digitales de procesamiento de voz requieren solo de una ligera noción de los resultados de las teorías acústicas, que es lo que se pretende presentar.

Bosquejo histórico

La principal razón que motivo al ser humano a investigar sobre la formación de la voz es que esta es su primordial forma de comunicación. Por medio de estas investigaciones ahora ya son conocidos múltiples aspectos de la voz humana, mas sin embargo hasta la fecha se continua investigando otros temas a los que no se les ha encontrado una explicación completa o suficientemente convincente, y a su vez estos estudios permiten mejores técnicas de sintetizado de voces, codificación de estas y nuestro principal interés, el reconocimiento de voz.

En los primero intentos de modelar y entender la producción de la voz dio como resultado maquinas mecánicas que hablaban, con el avance de la técnica se derivo a maquinas electrónicas analógicas y en tiempos mas recientes a sistemas computarizados o digitales.

Uno de los primeros trabajos documentados es por parte de C. G. Kratzenstein en 1779 en el que intento producir artificialmente y explicar las cinco vocales inglesas (/e, i, Y, o, yu/). El construyo un resonador acústico parecido en forma al tracto vocal y lo excitaba con una lengüeta vibrante que interrumpía el paso del aire. En esa misma época Wolfgang Ritter von Kempelen en 1791 demostró una maquina mucho mas exitosa, misma maquina que posteriormente fue mejorada por Sir Charles Wheatstone para la asociación de ciencia avanzada en 1835 en Dublín.



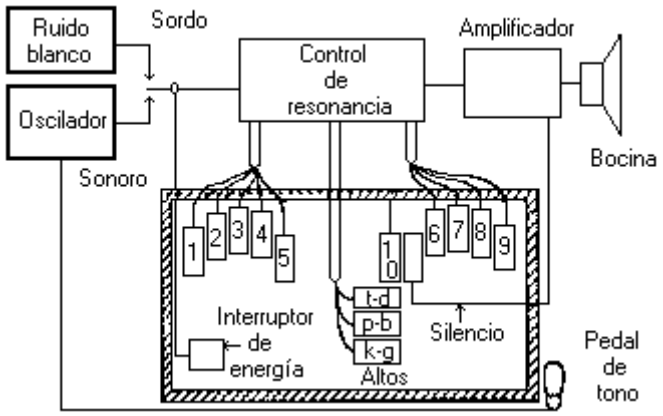
Maquinas de Von Kempelen y versión mejorada por Wheatstone

Como un dato curioso cabe destacar que un niño Escocés tuvo la oportunidad de ver la maquina de Wheatstone en Edimburgo, y al quedar seriamente impresionado por esta y con ayuda de su hermano se dedico a tratar de construir su propia maquina parlante, lo cual derivo en la patente U.S. 174465, el nombre del niño era Alexander Graham Bell y su hermano Melville Bell, y la patente se refiere al teléfono para voz.

El desarrollo continuo a través de los años, en 1846 Joseph Faber demostró su maquina llamada "Euphonia" que era algo parecido a un orégano que producía voz en lugar de notas musicales, mismo que presentaba la gran ventaja de poder cambiar el tono fundamental, en otras aproximaciones se usaron una serie de diapasones por Helmholtz en 1875, posteriormente Pager en 1930 construye modelos de polímetros plásticos y hule que podía producir prácticamente cualquier vocal y consonante inglesa.

Euphonia

Ya en 1939 Wagner construye un circuito eléctrico que podía copiar las vocales al controlar la frecuencia sobre las primeras cuatro regiones formanticas inglesas, y una de las primeras maquinas totalmente eléctricas se desarrollo en 1922 por J. Q. Stewart, y en 1939 se desarrollo el primer sintetizador totalmente eléctrico conocido como "Voder" de Voice Demonstrator, cuyo esquema se muestra en la figura siguiente, el cual requería un operador especializado para usar el Voder y producir voz, mismo que fue mejorado por la maquina eléctrica de H. K. Dunn en 1950, que podía producir voz de mucho mejor calidad con un tracto vocal eléctrico.



Futuros desarrollos en modelado y síntesis de voz continuaron con el desarrollo de las computadoras, ya que al mejorar el modelado se lograba una mejor codificación del

habla y procesos de sintetizado mas eficientes, lo cual nos lleva a estudiar en si el modelado, llevándonos al final de esta breve narración de la historia del modelado de la voz humana.

Voder

Propagación del sonido

Para poder caracterizar por completo la producción de la voz humana se requeriría una serie de ecuaciones diferenciales que describen los principios físicos de la propagación del aire en el sistema vocal. La generación y propagación del sonido requiere la caracterización de temas tales como:

a) La naturaleza cambiante a través del tiempo de la forma del tracto vocal
b) El acoplamiento de las cavidades nasales
c) El efecto de los tejidos blandos a través de las paredes del tracto vocal
d) El efecto del acoplamiento sub-glótico (pulmones y traquea) con la estructura resonante del tracto vocal
e) Perdidas producidas por la fricción viscosa en las paredes del tracto vocal y las condiciones de temperatura a través del tracto vocal

Una descripción completa requeriría de un análisis y modelado matemático detallado basado en teorías acústicas y mecánica de fluidos de baja viscosidad (aire), pero a pesar de que se han realizado una gran cantidad de estudios no existe a la fecha una teoría de aceptación universal, por lo que nos preguntamos ¿Como es que se llevan a cabo los trabajos actuales sobre este tema?, y la respuesta es debido a que estos trabajos se centran alrededor de sonidos estacionarios como las vocales, que son sumamente similares a la producción del sonido de un órganos de tubos, por lo que gran parte de este apartado se basara en el modelado a través tubos acústicos y su analogía eléctrica, estos modelos posteriormente se asociaran con filtros discretos para su uso en procesamiento de voz computarizado.

De lo ya estudiado podemos ver que se puede dividir la producción de voz en tres partes para su modelado, estas partes son la excitación, el tracto vocal y la radiación de la voz, por ejemplo una vocal en estudio en el tiempo finito puede ser representada por las siguientes tres transformadas de Fourier.

$$S(\Omega) = U(\Omega)H(\Omega)R(\Omega) \qquad (1)$$

Donde U (Ω) representa la excitación, H (Ω) representa la dinámica del tracto vocal y R (Ω) representa los efectos de radiación, cabe destacar que algunos autores incluyen el efecto de los labios en la dinámica del tracto vocal, pero otros lo sitúan en los efectos de radiación por su efecto de transformación flujo-presión, además este modelado asume que las tres componentes son lineales y separables, por lo que no se incorporara ningún acoplamiento entre estos subsistemas. Otra cosa que se debe tomar en cuenta es que se asume una propagación uniforme, esto es que cuando se crea la vocal una presión constante es producida y se expande para llenar el tracto vocal, propagándose uniformemente hasta los labios.

Las asunciones de propagación uniforme y que no existe la necesidad de acoplamiento entre subsistemas es una necesidad practica para producir modelos viables para su implementación informática, y por este medio se a logrado la inmensa mayoría de los modelados que apoyan los algoritmos de procesamiento de voz actuales, la exclusión de ciertos detalles en estos modelos es factible debido a que lo que deseamos es caracterizar a grosso modo características temporales y propiedades en el dominio de la frecuencia, los detalles mas finos de estos modelos son únicamente necesarios para analizar por ejemplo el movimiento de las cuerdas vocales, efectos de las patologías sobre la voz humana y otras áreas similares que no son de interés al presente trabajo.

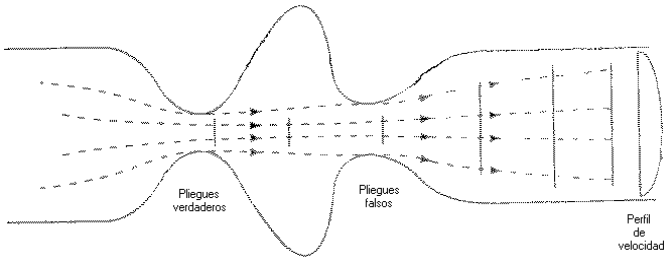


Figura 25: Modelo clásico de propagación uniforme

Modelo de la fuente de excitación

Existen dos formas básicas de excitación y estas son:

a) Excitación sonora; un movimiento periódico de las cuerdas vocales resultando en un flujo de resoplidos de aire quasi periódicos.
b) Excitación sorda; una turbulencia insonora causada por el flujo de aire a través de una constricción estrecha

Es importante también mencionar las siguientes formas de excitación que representan importantes combinaciones o variantes de la excitación sonora y sorda y que para propósitos de modelado usualmente se toman como distintas categorías.

c) Excitación plosiva; causada por el incremento de presión del aire por una porción del tracto vocal totalmente cerrada, seguida por una súbita liberación de la presión, el flujo liberado produce un sonido sonoro o sordo dependiendo del fonema

d) Susurro; el susurro es una pronunciación creada al forzar aire a través de la glotis parcialmente abierta para excitar una pronunciación que seria normalmente articulada.

e) Silencio; se debe de incluir como una forma de excitación para modelado, a pesar de que en esta no existen regiones quasi estáticas en el habla sin sonido, este tipo de excitación ocurre por ejemplo en la pausa que precede un sonido plosivo.

Excitación sonora
Para la producción sonora un flujo de aire de los pulmones es interrumpida por una vibración quasi periódica de las cuerdas vocales como se ve en la figura 26, que como se explico anteriormente en la fisiología del aparato fonador entra en una oscilación sostenida, la taza a la cual las cuerdas vocales cierran y abren es determinada por la presión sub-glótica, la tensión y tono muscular de las cuerdas vocales y el área de apertura glótica.

El tono fundamental es uno de los parámetros más básicos y fáciles de medir, en adición otros aspectos importantes como son la duración de cada pulso de la laringe y la forma de cada pulso. Un análisis de estos parámetros requiere la reconstrucción o estimación desde una señal de voz, estos métodos son llamados algoritmos de filtrado inverso glótico.

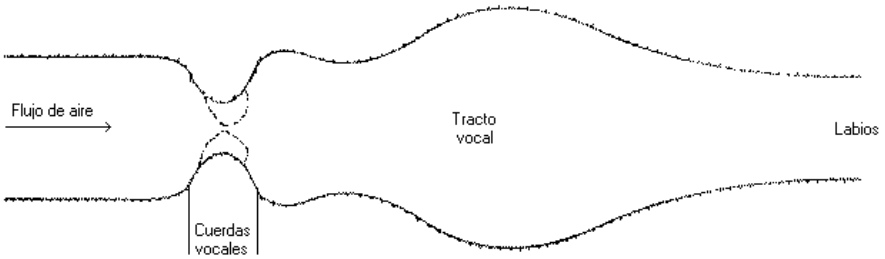


Figura 26: Esquemático del sistema vocal

En las diversas fuentes bibliográficas existen desacuerdos sobre esta aproximación, ya que algunos autores dicen que existe una segunda fuente de excitación durante la apertura glótica o incluso cuando esta se encuentra cerrada, por lo que se han desarrollado modelos con múltiples excitaciones, existen a su vez otros modelos, como los que usan una pulso de señales senoidal, pero dichos modelos son utilizados en técnicas de sintetizado, mas no en codificación o reconocimiento por su baja calidad.

Excitación sorda
Este tipo de excitación se presentan en los sonidos fricativos o plosivos, y es modelado frecuentemente como ruido blanco, este tipo de excitación teóricamente no tiene efecto en la forma del espectro ya que la densidad de poder es uniforme en todas las frecuencias, por lo que se da poca importancia a esta excitación, ya que se pueden obtener muy buenos resultados sin importar la fase, ya que la amplitud espectral es mas importante que la fase para la percepción de la voz.

Modelado del tracto vocal

Una onda sonora es producida cuando las cuerdas vocales vibran o por movimiento aleatorio de partículas de aire. La propagación sigue las reglas de la física, incluyendo la conservación de la masa, momento y energía. El aire puede ser considerado como un fluido de baja viscosidad comprimible, por lo que aplican las leyes de mecánica de fluidos y la termodinámica y sabiendo que las ondas sonoras tendrán un comportamiento no radial debido a los tejidos blandos del tracto vocal, estas se

propagaran en una sola dirección, por lo que para simplificar el modelado vamos a asumir que el sonido sigue una propagación plana a lo largo del eje del tracto vocal, pero teniendo en consideración que esto solo se cumple siempre y cuando la longitud de onda sea grande comparada con el diámetro del tracto vocal (aproximadamente frecuencias menores a los 4000 Hz).

$$\lambda = \frac{c}{F} = \frac{340\ m/seg}{4000\ 1/seg} = 8.5\ [m] \qquad (2)$$

Donde 'c' es la velocidad de propagación del sonido en el aire, y vemos que los 8.5 centímetros es mucho mayor que el diámetro promedio del tracto vocal de 2 centímetros, asimismo se asumirá también que el tracto vocal puede ser modelado como un cuerpo de paredes rígidas sin perdidas como se ve en la figura.

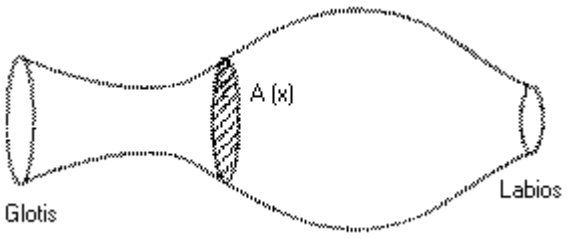


Figura 27: Tracto vocal ideal

La propagación plana de una dimensión asume que todas las partículas en un desplazamiento 'x' tendrán la misma velocidad independientemente de su posición a través de la sección transversal 'A', por lo tanto el análisis es más conveniente si consideramos la velocidad de un volumen de aire y no de una simple partícula.

Este modelo caracteriza la velocidad del volumen y la presión del sonido a través del tracto desde la glotis (x = 0) hasta los labios (x = n o 17.5 cm. en un varón típico), pero solo se puede solucionar para algunas configuraciones simples, incluso tomando algunas asunciones como son que en la formación de vocales simples el área A (x, t) no variara con el tiempo, la solución es muy compleja, por lo que se emplearan mas simplificaciones a fin de llegar a una solución mas razonable.

Modelo de tubo sin pérdidas

Uno de los problemas para resolver el modelo anteriores es la parametrización del área A (x, t) a lo largo del tracto vocal, para simplificar un poco este problema se va a asumir que el tracto vocal se puede tratar como un tubo de área transversal igual, así remplazamos A (x, t) por A como se puede ver en la figura siguiente, algunos trabajos se auxiliaron de los rayos X para observar y extraer los parámetros durante la pronunciación de las vocales, posteriormente se aplicaron al tubo sin perdidas con buenos resultados en la producción de vocales sintetizadas, lo cual valida esta aproximación.

También se puede ver de la figura que el modelo del tubo no toma en consideración la curvatura del tracto vocal, pero en los trabajos por los trabajos de Sondhi M.M. se pudo ver que el efecto de la curvatura no afecta mucho el resultado, así que es valido el tomar este modelo para una primera aproximación.

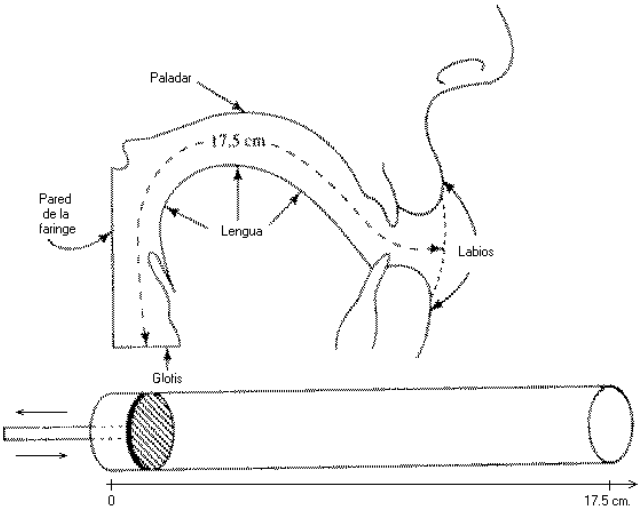


Figura 28: Diagrama de rayos X y modelo de tubo uniforme

Aun considerando ya una simplificación considerable como el tubo sin perdidas debemos tener en consideración aspectos tales como si este tendrá una terminación abierta o cerrada, que en este caso representa la apertura de los labios.

Para una terminación abierta la variación de la presión ambiental y los labios es cero, aunado a que nos interesa un análisis estático donde la fuente glótica se modela como una excitación exponencial simple, por lo que con estas condiciones de frontera y considerando A (x, T) como una constante las ecuaciones que definen el comportamiento de la propagación del sonido se simplificarían considerablemente.

Para una terminación cerrada que sucede cuando el tubo esta totalmente cerrado, como cuando los labios están cerrados, se ve claro que la velocidad del aire en los labios es cero, ya que no sale aire por los labios por lo tanto resulta en una función de transferencia igual a cero.

Modelo de múltiples tubos sin pérdidas

Dado que la producción de voz es caracterizada por los cambios de forma del tracto vocal se ve que un modelo mas realista consistiría en una serie de tubos que pueden variar como una función del tiempo y del desplazamiento a través de los ejes de propagación del sonido, esta formulación puede resultar muy compleja. Una forma de simplificar esto seria el considerar el tracto vocal como una serie de tubos acústicos conectados sin perdidas, como se puede ver en la figura 29, esto es una serie de tubos con un área transversal A_k y un largo l_k.

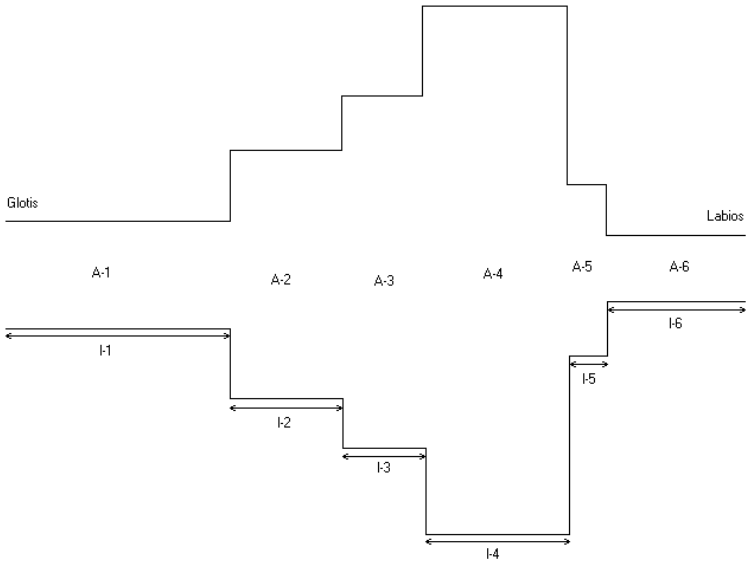


Figura 29: Modelo de seis tubos

Las áreas transversales y longitudes se escogen de manera que se aproximen al área de la función A (x) del tracto vocal, mientras mas tubos de menor longitud sean utilizados las formantes resultantes se aproximaran cada vez mas a un modelo se área transversal con cambios continuos.

Para determinar la interacción de las ondas viajeras entre los tubos podemos considerar la unión entre dos tubos cualesquiera que sean, ya que las ondas cumplen con la ley de la continuidad y de fuerza de Newton la presión y velocidad del volumen son continuos tanto en el tiempo como en el espacio, la ley de la continuidad requiere que se cumpla que la presión y velocidad de volumen del final del tubo 'k' deben de ser iguales a la presión y velocidad de volumen del inicio del tubo 'k+1'.

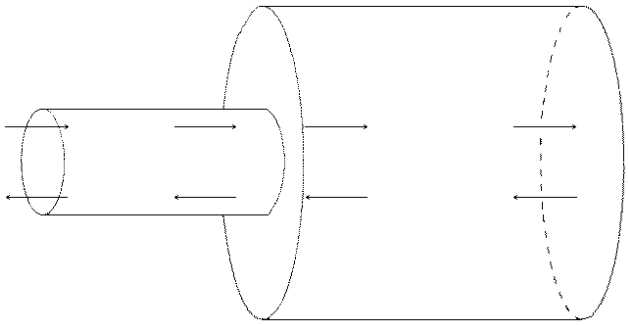


Figura 30: Unión de dos tubos

En la figura se puede apreciar que una porción de la señal viajera positiva (de izquierda a derecha) es transmitida del tubo 'k' al 'k+1' y otra parte es reflejada de nuevo al tubo 'k', al igual la señal viajera negativa (de derecha a izquierda) del tubo 'k+1' una porción se transmite al tubo 'k' y otra porción es reflejada al tubo 'k+1'. La relación que ilustra la transmisión y reflexión de la propagación de las ondas en una unión se resuelve mediante ecuaciones diferenciales.

De estas ecuaciones veríamos que la velocidad del volumen positivo en el tubo 'k+1' esta compuesta de una parte transmitida desde el tubo 'k' y una porción reflejada del mismo tubo 'k+1' y que la velocidad del volumen negativo del tubo 'k' consiste de una parte que se transmite del tubo 'k+1' y otra parte que se refleja desde el mismo tubo 'k',

por lo que si asumimos que la señal viajera negativa en el tubo 'k+1' es cero las ecuaciones se reducirían considerablemente.

Por este hecho y dado que los coeficientes están relacionados, es de uso común el usar un solo coeficiente de reflexión en estudios analíticos y de modelado, por lo que apegándonos a esta notación se usa el coeficiente ρ_k omitiendo así el signo y dado que el área transversal de los tubos debe de ser positiva, el coeficiente de reflexión en cualquier unión de tubos estará limitado por la unidad.

$$-1 \leq \rho_k \leq 1$$

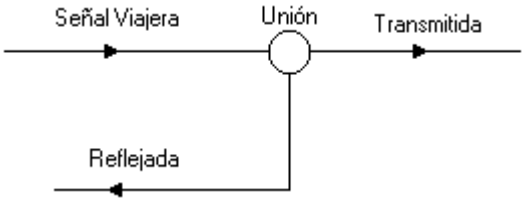


Figura 31: Diagrama de flujo de una unión

En la figura vemos un diagrama de flujo en la juntura de los tubos 'k' y 'k+1' para una onda que viaja en sentido positivo.

Las ecuaciones que resuelven este modelo son conocidas como Kelly-Lochbaum y fueron desarrolladas en 1962, de esta ecuaciones se llego al diagrama de la figura 32 para la unión en un punto cualquiera, esta estructura fue utilizada por primera vez por Kelly y Lochbaum en 1962 para la síntesis de voz. Nótese que este diagrama de flujo contiene información equivalente a la figura de los tubos, pero en lugar de caracterizar las secciones por su área y largo aquí se caracteriza por coeficientes de reflexión y retardos en la señal.

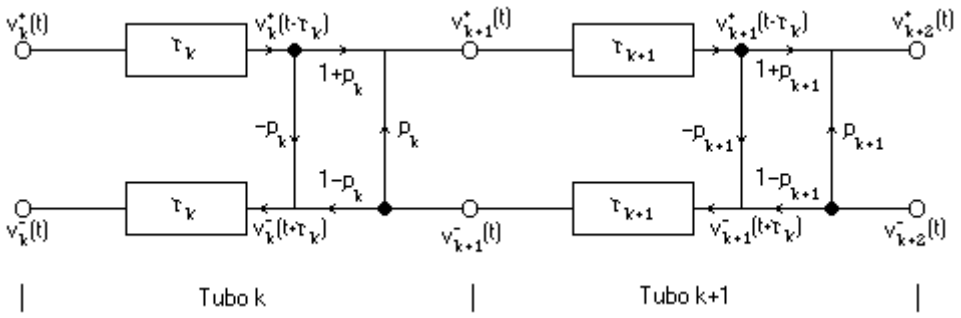


Figura 32: Diagrama de flujo de dos tubos sin pérdidas

La figura de la unión aísla una unión cualquiera e ilustra la reflexión y transmisión de una onda viajera. La propagación de una onda sonora experimenta un retraso al viajar de un lado a otro de un tubo, que se representa como un retraso de la señal τ_k en la figura del modelo. Dado que no existen pérdidas en los tubos no existe atenuación alguna en la propagación, pero al llegar la señal a la unión de dos tubos una porción de la señal será transmitida $(1+\rho_k)$ y el remanente es reflejado $(-\rho_k)$.

El porcentaje transmitido y reflejado dependerá de la diferencia en área transversal entre los tubos, si las áreas son similares $(A_k \approx A_{k+1})$ la mayoría de la señal será transmitida,

llegando al caso ideal cuando las áreas son idénticas (transmisión perfecta), pero si las áreas son sumamente diferentes ($A_k >> A_{k+1}$) se reflejara mas señal de la que se transmite.

A partir de este modelo se puede generar un modelo de 'n' tubos, en cada frontera las condiciones de flujo y dinámica de presión están descritas en términos de coeficientes de reflexión y retrasos y se muestra a continuación.

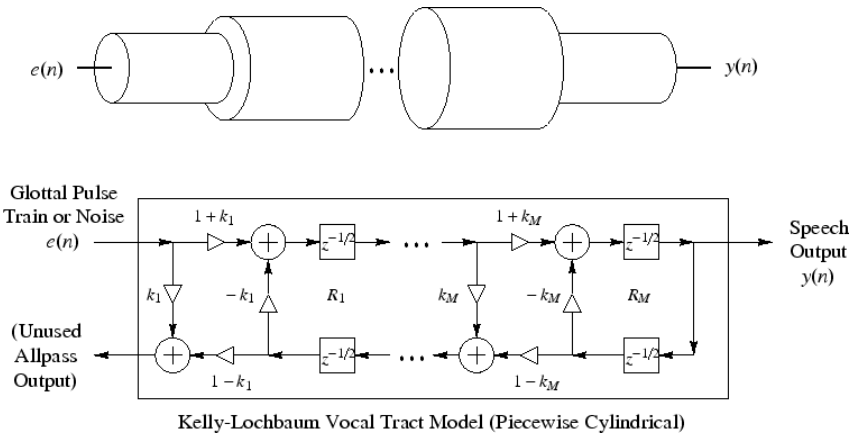


Figura 33: Modelo de 'n' tubos

Modelos de tubo para la glotis y los labios
Para completar el modelo sin perdidas de producción de voz es necesario tomar en consideración los efectos de frontera en los labios y la glotis, mismo que se abordara en este momento.

La formulación de un modelo de tubo para la glotis o los labios requiere la consideración de un tubo acústico infinito en uno de sus extremos, esto es, parte de un punto determinado pero no tiene fin en el otro extremo, por lo que este quedaría caracterizado por una área transversal A y un largo infinito, por lo tanto si se inyecta una señal de voz en uno de sus extremos esta se propagara hacia la derecha indefinidamente, y dado que no existen mas fronteras en dicho tubo no se presentara reflexión alguna de la señal. Por ultimo y debido a que es un tubo sin perdidas, al hacerlo tender a infinito la impedancia en un lugar 'x' del tubo será simplemente la impedancia característica del tubo.

$$\lim_{l\to\infty} Z_{\substack{Tubo \\ Abierto}}(x,\Omega) = Z_0 \qquad (3)$$

Dado que se asume que dicha impedancia es un número real, un tubo infinito puede ser usado para modelar una carga resistiva, por lo que la impedancia puede ser expresada simplemente como una resistencia R_0.

Ahora hay que considerar la unión entre el ultimo tubo del tracto vocal y los labios, por lo que si consideramos un tracto vocal de N tubos, el tubo N + 1 serán los labios, y dado que una vez que la voz abandona los labios ya no existirá reflexión, por lo que la onda viajera negativa cera cero.

También se puede utilizado un tubo infinito para modelar el comportamiento en la glotis, como se puede suponer una parte de la señal que llega por la glotis será

transmitida al primer tubo del tracto vocal y otra parte será reflejada a la glotis, pero como esta porción reflejada pasa a la traque y los pulmones, no contribuye de forma significativa a la producción del sonido, dado que en general se considera que los tejidos blandos de los pulmones absorben la mayoría de esta energía, por lo que en términos de modelado esta señal viajera negativa es ignorada, por lo que la terminación glótica se modelara como una fuente con una impedancia $Z_{glótal}$ en paralelo.

$$Z_{glotis}(\Omega, t) = R_{Glotis}(t) + j\Omega L_{Glotis}(t) \qquad (4)$$

Esta es una impedancia acústica variable con el tiempo que es función del inverso del área transversal de la glotis, por lo que por ejemplo cuando la glotis se cierra la impedancia es infinita, por lo que decimos que la impedancia variara entre infinito y un valor finito, mismo que lleva a que la velocidad del volumen varara entre cero a un valor finito, dado una señal pulsante, como ya lo habíamos mencionado.

Es conveniente mencionar que ambos coeficientes de reflexión son dependiente de la frecuencia debido a la participación de la impedancia, así mismo la impedancia de la glotis también suele ser considerada como un numero real para facilitar su modelado, por lo que los efectos de la glotis y los labios puede ser representada en términos de impedancia y una velocidad de volumen para analogías con circuitos eléctricos.

También es importante mencionar que ya tenemos todos los elementos necesarios para crear un modelo completo de producción de voz, y que normalmente estos modelos emplearan un número de aproximadamente 14 tubos o secciones para el modelado del tracto vocal con el fin de tener una mejor aproximación y mejores resultados.

Efectos de pérdidas en modelos de tubos
Para este efecto deberíamos de regresar al inicio del tema, ya que se tomaron una serie de presunciones poco realistas con objeto de simplificar el análisis, pero principalmente la suposición de que se puede modelar como una serie de tubos rígidos sin pérdidas es poco certera.

De hecho existen perdidas de energía debido a la fricción viscosa entre el flujo de aire y las paredes del tracto, la vibración de las paredes y condiciones de temperatura del tracto vocal, pero la introducción de tales factores desde el inicio nos llevaría a una solución mucho más compleja y difícil de comprender.

Debido a esto es que la aproximación que se a tomado por los ingenieros dedicados a este tema es anticipar ciertos efectos en rangos de frecuencias definidos que son resultado de dichas perdidas, y es así que auxiliándonos de estos elementos podemos basarnos en los modelos sin perdidas obteniendo buenos resultados.

El efecto de perdidas mas importante es debido a la vibración de las paredes del tracto vocal, ya que estas responden principalmente a las frecuencias bajas con un efecto vibratorio, y para las frecuencias altas las mayores perdidas se verán por efectos térmicos y de fricción viscosa, afectándose así ligeramente las formantes obtenidas con estos modelos, por lo que podemos observar ligeros incrementos o decrementos en frecuencia de las formantes en estos modelos.

Ya conocemos cuales son los efectos y las causas, mas como no es tema de este trabajo la síntesis de voz sino el reconocimiento de locutores por el momento se pasaran por alto dichas perdidas, ya que al no afectar significativamente los resultados se podrán crear modelos flexibles que puedan a futuro tomar en consideración algunos efectos de perdida y compensarlos.

Modelado de sonidos nasales y fricativos

La producción de sonidos nasales requiere que el velo este abierto y los labios cerrados, lo que crea un modelado muy difícil con una configuración de tubos como los antes descritos. En el caso de los sonidos nasales la cavidad oral crea una cavidad lateral derivada de la cavidad principal. Analíticamente esto resultaría en un juego de tres ecuaciones diferenciales, las cuales para muchos fines no es practica, por lo que usualmente es suficiente el tomar en cuenta ciertos efectos en frecuencia causados por la nasalización, ya que las cavidades nasales tienden a atrapar energía llevándonos a anti-resonadores o ceros en adición de los polos en el espectro de magnitud.

Polos y ceros
Como un paréntesis cultural, en las matemáticas y procesamiento de señales, la Transformada Z convierte una señal que esté definida en el dominio del tiempo discreto en una representación en el dominio de la frecuencia compleja.

El nombre de Transformada Z procede de la variable del dominio, al igual que la S para la Transformada de Laplace. Esta transformada esta dada por la ecuación:

$$X(z) = Z\{x[n]\} = \sum_{-\infty}^{\infty} x[n] z^{-n} \qquad (5)$$

Donde *n* es un numero entero y *z* es un numero complejo y $z = Ae^{j\omega}$ donde *A* es el modulo y ω la frecuencia.

De aquí una ecuación cualquiera representada por H podría expresarse como sigue:

$$H(z) = \frac{(1-q_1 z^{-1})(1-q_2 z^{-1})...(1-q_m z^{-1})}{(1-p_1 z^{-1})(1-p_2 z^{-1})...(1-p_n z^{-1})} \qquad (6)$$

En esta representación en el numerador tenemos *m* raíces llamadas *ceros* y en el denominador *n* raíces llamadas *polos*, por lo que una ecuación que presenta únicamente polos como las que de manera general veremos, únicamente tienen raíces en el denominador.

El tema de la transformada Z es sumamente complejo y amplio, pero realmente no es de suma relevancia para este trabajo, por este motivo únicamente se presenta este pequeño paréntesis para comprender un poco mejor de lo que se habla aquí.

En general es un tanto complejo el modelado de las consonantes y las fricativas son un caso particularmente complejo, pero en general se ha observado que un modelo de tres tubos para el tracto vocal trabaja bien para el modelado de las cavidades principales anterior y posterior conectadas por una estrecha constricción. Las fricativas al igual que

las nasales tienden a crear anti-resonadores o ceros en el espectro, y se a encontrado que las fricativas presentan poca energía debajo de su primer cero en su espectro.

Derivado de todo esto, al igual que en el caso de las perdidas se usa un modelo sin pérdidas ni cavidades auxiliares para analizar sonidos nasales y fricativos, ya que el modelo puede sufrir ligeras modificaciones posteriores para compensar los ceros extra en el modelo, ya que es sumamente deseable que el modelo presente únicamente polos.

Modelado en tiempo discreto

En la figura 34 podemos observar un diagrama completo desde la glotis hasta los labios, de este podemos notar la gran similitud que existe entre este y un filtro digital, este modelo corresponde a un tracto vocal conformado por cuatro tubos sin perdidas.

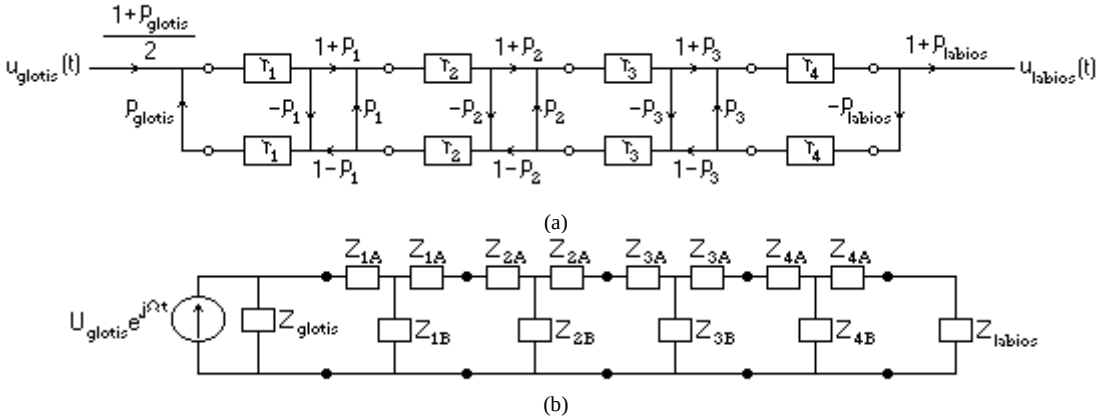


Figura 34: Diagrama de flujo de un modelo de cuatro tubos y su equivalente en una línea de transmisión

Este diagrama consiste únicamente de sumas, restas, multiplicaciones y retrasos y estas operaciones pueden fácilmente ser implementadas en un modelo de tiempo discreto, la única restricción a satisfacer es que cada retraso sea un múltiplo de T, siendo este el periodo de muestreo del sistema en tiempo discreto. Para logra esto considérese el retraso total experimentado por una señal que entra al sistema.

$$\tau_{total} = \sum_{j=1}^{N} \tau_j = \frac{1}{c} \sum_{j=1}^{N} l_j \qquad (7)$$

Donde τ_{total} representa el tiempo necesario para que la señal de voz viaje a través del modelo del tracto vocal $L = \sum_{j=1}^{N} l_j$. Para asegurar una transición suave al dominio del tiempo discreto consideremos N tubos, cada uno digamos de un largo $\Delta_x = L/N$, en este caso se trabajara del mismo modo que en el antes expuesto, a excepción de que el retraso en cada tubo será igual al de los otros.

$$\tau = \frac{L}{cN} = \frac{\Delta_x}{N} \qquad (8)$$

Esto limitara el numero de variables necesarias para simular la función A (x, t), con esta simplificación se logra simplificar el diagrama de la figura 34 al de la figura 35, donde todos los retrasos son sustituidos por un retraso igual.

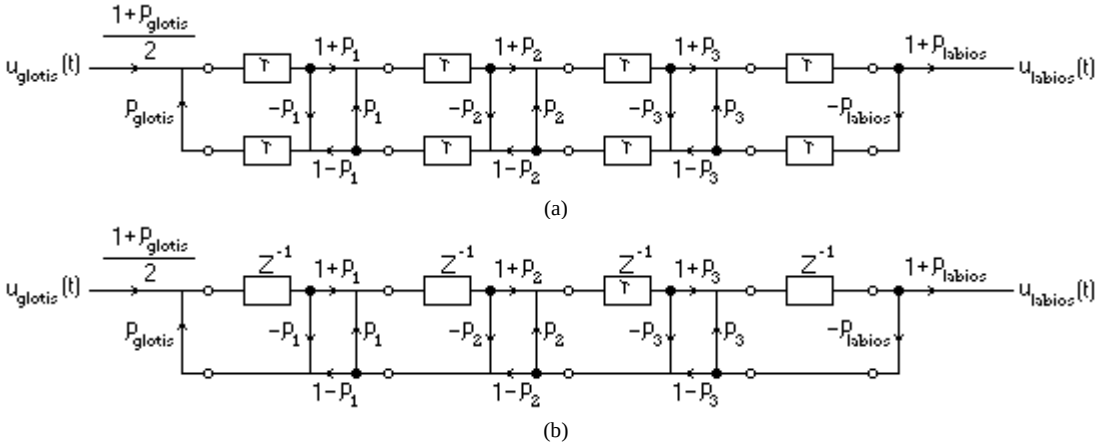


Figura 35: Modelo simplificado y modelo en tiempo discreto

Si se inyecta una señal discreta en el modelo, lo mas rápido que una salida puede ocurrir es en $N\tau$ unidades de tiempo (N = 4 en este ejemplo), y si tenemos un periodo de muestreo de $T = 2\tau$ entonces el tiempo $N\tau$ corresponde a un corrimiento de N/2 muestras.

Como una nota importante cabe mencionar que este modelo implica el tracto vocal desde la glotis a los labios, pero es importante mencionar que los tubos '0' y 'N+1' no representan en si a los labios o la laringe, sino simplemente se usan para modelar los efectos de frontera de estos elementos y el tracto vocal, pero no corresponde de ninguna manera con G (z) y R (z).

Modelado con filtros discretos para producción de voz

Meramente como agregado de interés general en la figura mostramos un modelo lineal discreto para producción de voz, este modelo es llamado "modelo de terminal analógica", significando que las señales y el sistema involucrados en el modelo son solamente superficialmente analógicas. Este modelo intenta representar el proceso de producción de voz basado en sus características de salida, y este modelo produce voz de razonablemente buena calidad para propósitos de codificación.

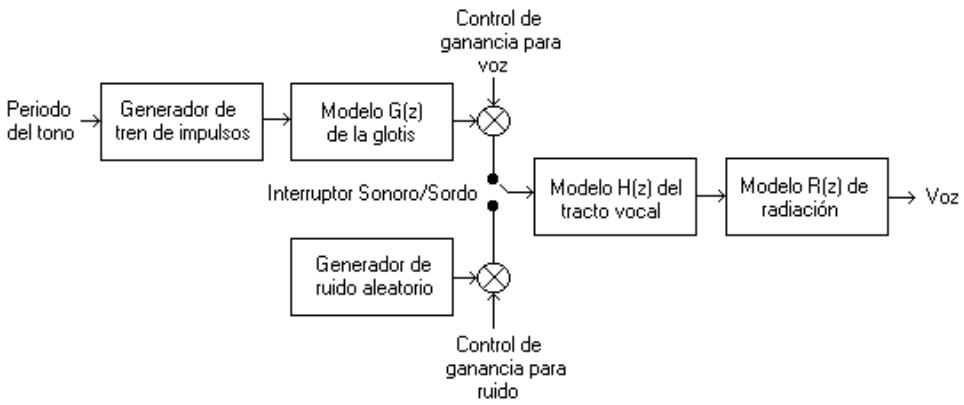


Figura 36: Modelo general discreto para producción de voz

En este modelo general el modelo del tracto vocal y el modelo de radiación son excitados desde una fuente de señales discretas, en los sonidos sordos, la fuente de excitación será un generador de ruido aleatorio, durante periodos de sonidos sonoros la excitación proviene de un estimado del tono fundamental que impulsa un generador de tren de impulsos que excita al modelo de la glotis, pero cabe hacer mención que este modelo no puede manejar señales que tienen mas de una fuente de excitación como sonidos fricativos.

La síntesis o producción de voz artificial no es tema de interés del presente trabajo, sin embargo cuando se trabaja con modelos de este tipo que tienen únicamente polos y es que en técnicas como LPC, que se vera posteriormente, se obtiene un modelo de polos únicamente por métodos informáticas, que es sumamente útil en aplicaciones de reconocimiento de voz, codificación e incluso síntesis con algunas adaptaciones.

Vocoders

Por vocoders queremos decir VOice CODERS (codificadores de voz) mismo nombre que es común para los métodos de síntesis por análisis, estos métodos usan la información espectral contenida en las señales de voz para una codificación mas eficiente.

Estos métodos tratan de identificar parámetros que representan mejor la voz, y luego usar dichos parámetros en el receptor para reconstruir la señal (síntesis), estas técnicas se han hecho populares por el hecho de que pueden representar la voz en frecuencias de muestreo muy bajas (de 2400 a 9600 bits por segundo) con buena calidad, por lo que resulta de gran utilidad en aplicaciones donde el ancho de banda o el espacio de almacenamiento es crucial, también es importante mencionar que este tipo de codificación tiene un limite para su calidad, mismo que no mejorara a pesar de aumentar la frecuencia de muestreo.

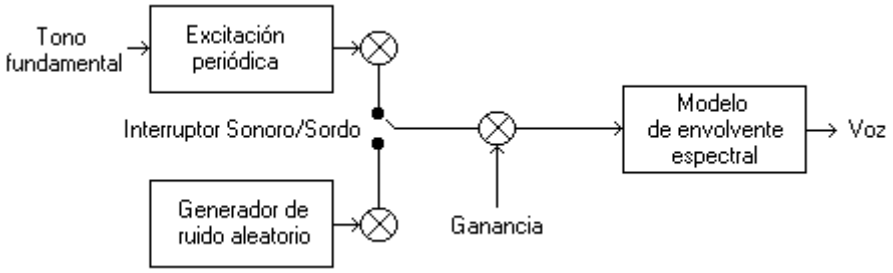


Figura 37: Modelo de un vocoder

La razón por la que reciben el nombre de vocoders es que se utiliza un modelo explicito para la producción de voz, como se ve en la figura, en este modelo la excitación esta separada del modelo espectral, y cada uno se codifica por separado, aumentando así el ahorro en bits, en la figura 38 se muestra la transformada de Fourier de una onda de audio de 20 mili-segundos de voz, esta a su vez se codifica mediante su envolvente que se apega al espectro.

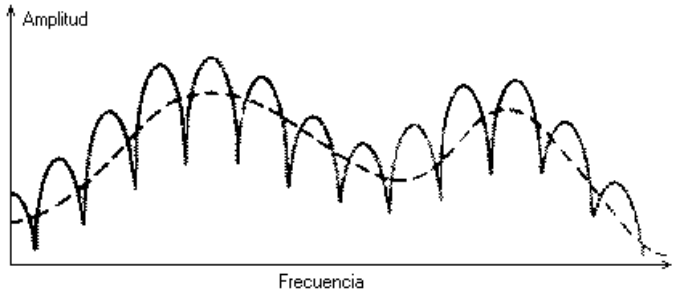


Figura 38: Especto con envolvente

El vocoder mas utilizado es el de código de predicción lineal o LPC, mismo que se discutirá mas a fondo después, pero también existen otro tipo de vocoder comunes, teniendo en primer termino el vocoder de canal pues es la metodología mas antigua y estudiada, misma que se basa en el análisis de Fourier a corto plazo, otro método es el vocoder mono-forma que utilizan una parametrización distinta, basándose en un espectro logarítmico de la voz, otros vocoder son los de fase que es una técnica en el dominio de la frecuencia donde se codifica la amplitud espectral y por ultimo los vocoders por formantes que usan modelos que usan la representación del tracto vocal por sus formantes, así mismo existen mas métodos como los vocoder fonéticos, los vocoder excitados por voz y la cuantización de vectores, pero no serán tratados a mayor profundidad pues la idea es dar una semblanza general de los vocoders, mas no es tema del presente trabajo.

¿Qué son las Técnicas de reconocimiento y cuales usaremos?

Como nos lo marca la Real Academia Española, una técnica es un "Conjunto de procedimientos y recursos de que se sirve una ciencia o un arte", y el reconocimiento es "Acción y efecto de reconocer o reconocerse", por lo tanto las 'técnicas de reconocimiento' son el conjunto de procedimientos y recursos destinados a efecto de reconocer o reconocerse.

Ahora que tenemos el concepto claro de a que nos referimos podemos decir que aquí se pretende explicar las diversas metodologías o procedimientos de reconocimiento que se pueden aplicar a la voz humana, debemos ver que esto puede ser un tema muy extenso, ya que comprende desde luego la extracción de parámetros, que puede ser realizada tanto en el dominio del tiempo como en el de la frecuencia, facilitándose en cada una la extracción de ciertos parámetros, pasando de aquí a establecer patrones con dichos parámetros, con el propósito de clasificarlos por sus diferencia y similitudes según el caso, y por ultimo un algoritmo de toma de decisiones si es que este se considera necesario, por lo que a lo largo de este tema trataremos de ir tocando todos estos temas de manera breve y dejar un marco general de trabajo para esta aplicación y su correspondiente investigación.

Extracción de parámetros

La extracción de parámetros es en si la acción de obtener algunos datos representativos de un mayor numero de datos, de esta manera es mas fácil el poder trabajar algunos parámetros que una gran cantidad de datos.

Tomemos por ejemplo un grupo de niños de escuela, digamos 80 alumnos del segundo grado de primaria, es en general mas representativo dar algunos parámetros de este grupo que todos los datos de cada alumno, pensemos que si queremos la edad, promedio de calificaciones de todo el año de clases por materia, el promedio general, y si son niños o niñas, etc. Será mas fácil tener una serie de datos representativos que el querer estudiar cada alumno individualmente para poder comparar contra otro grupo y ver su aprovechamiento.

Podríamos tener por ejemplo una edad promedio de 8 años, un promedio en matemáticas de 8, un promedio en español de 9, un promedio general de 8.5, etc. Datos que son más fáciles de evaluar y comparar contra los mismos de otro grupo.

De esta misma manera vemos que la extracción de parámetros para muestras de voz es el proceso de buscar datos representativos de dicha muestra, como son su tono fundamental, estructuras formanticas, niveles de energía, cruces por cero y otros muchos parámetros que nos pueden ser de utilidad, pero con el debido cuidado de que estos datos nos proporcionen información fidedigna sobre la muestra en si y el individuo que la produce, ya que es fácil caer en parámetros no representativos, como podría ser para el ejemplo anterior el color de ojos del alumno, mismo que a pesar de ser un dato verdadero y legitimo no muestra relación alguna con el aprovechamiento, por lo que no es representativo o útil.

Desafortunadamente no es tan simple ver que parámetros pueden o no ser representativos o útiles en muestras de voz, ya que además de que sea propio a la muestra en si tenemos efectos secundarios no deseados de elementos tales como ruidos

en la grabación, la calidad de los equipos con los que se realiza la grabación y digitalización de las muestras o simplemente datos que muestren muy poca diferencia entre distintos sujetos o que no sean representativas debido a la inter variabilidad intra personal del sujeto, ya que recordemos que de la definición misma de formante será una estructura que el hablante cree es la misma y el receptor percibe como la misma, siendo estos ajenos a las diferencias existentes.

Por lo expuesto se ve claro que no es fácil encontrar cuales parámetros pueden ser los adecuados para la tarea que nos proponemos, por lo que resulta fácil apreciar el por que se implementaran distintas técnicas que trabajan con diferentes parámetros, esto en busca de encontrar los que sean mas útiles y confiables, ya que además podemos tener métodos que obtengan parámetros que en algunas ocasiones funcionen bien y en otras caigan en contradicción, indicándonos que su porcentaje de confiabilidad es bajo y en algunas ocasiones inaceptable debido a constantes falsas identificaciones.

Pasaremos posteriormente a explicar algunos de los juegos de parámetros que se proponen para solucionar nuestro problema, así como su respectiva implementación y prueba para ver claramente lo expuesto, pero antes veremos algunos aspectos importantes de los patrones.

Reconocimiento de patrones

Ya hemos tocado el tema de la adquisición de las señales que nos interesan, su digitalización y su correspondiente extracción de parámetros que nos dará una representación más aceptable en términos generales.

Cabe hacer mención que las señales en el dominio de la frecuencia y no del tiempo son en general más aceptables, ya que al transformar la señal original eliminaremos o por lo menos disminuiremos ruidos o señales no deseadas por un método en el cual no se pierde información, sino tan solo se modifica la forma en que esta se manifiesta o interfiere con nuestra señal primaria, pero esta transformación la veremos mas adelante.

Recordemos que el propósito general del procesamiento de bajo nivel que se lleva a cabo en una primera etapa tan solo nos provee de una información más susceptible de un procesamiento de alto nivel como el que involucra el reconocimiento de patrones.

La interpretación de la información de manera general esta basada en el reconocimiento de estructuras conocidas o en las similitudes que presenten con una estructura conocida, de esta manera se podrá clasificar la información en grupos llamados comúnmente patrones.

Por supuesto, como es fácil notar, el clasificar la similitud de un conjunto de datos con otro dependerá de que información extraemos del conjunto de datos que se nos presenta, por lo que se ve la importancia que tiene la correcta extracción de parámetros así como la necesidad de cerciorarse que dichos parámetros sean representativos de cada uno de los “n” patrones que conforman nuestro universo de datos.

Reconocimiento de patrones y la voz humana

Para explicar este tema y basándonos en lo ya expuesto plantearemos un ejemplo usando nuestros conocimientos de foniatría, donde veremos en forma rápida cuales son los datos que nos pueden ser de mayor utilidad al ser sumamente representativos de un

conjunto de patrones y mas aun de un sujeto, razona por la cual son de interés para nosotros, recordemos de manera rápida.

Un fonema es : "es la unidad fonológica más pequeña, su número es reducido, no tiene significado por si mismo, pero el significado de una palabra cambia si se intercambian dos fonemas", además "por medio de este termino designamos un conjunto de aquellas propiedades recurrentes que se usan en una lengua dada para distinguir palabras de distinto significado" y por ultimo "cada fonema tiene una serie de características, y no se pueden repetir en conjunto para un determinado fonemas", por lo que vemos que los fonemas son un número reducido de patrones, que sus características no se pueden repetir sin variar el fonema y que sumando dichos fonemas formamos una lengua, lo cual lo hace optimo para los fines que perseguimos.

Aunado a esto podemos agregar que "aunque los hablantes produzcan diferentes sonidos y los oyentes los perciban como objetivamente diferentes, no son conocedores de tal diferencia; el hablante cree producir el mismo sonido, y el oyente la impresión de oír el mismo sonido", lo cual se conoce como la inter-variabilidad intra-personal de un sujeto, razón por la cual se confirma que cada patrón será un campo de dispersión de un juego de características que conforman un solo fonema único dentro de una lengua y característico a un sujeto.

Por supuesto que existen otros métodos, pero estos son básicamente variaciones de este mismo, como es el tono fundamental (T_0), el cual es la base de los formantes mismos (la frecuencia fundamental de cada formante) o parámetros tales como LPC, que por métodos numéricos extrae una serie de características basados en los formantes mas no detecta ni trabaja los formantes en si.

Una vez que sabemos que es lo que trabajaremos y que hemos verificado que esta información es representativa y que forma un numero finito de patrones podemos ver que para los fines que persigue el presente trabajo formaríamos una serie de patrones para tener los más elementos posibles representativos de una lengua, y estos se podrán englobar en un patrón mas grande que representa a cada individuo que estudiamos, llevándonos como es adecuado de lo general a lo particular para obtener una respuesta de nuestro sistema.

De esto mismo podemos ver que será un proceso complejo y difícil de implementar en un sistema computacional, ya que comparar vectores de características obtenidos de una serie de patrones, que es el próximo paso representa ya un problema bastante complejo, por tal motivo las redes neuronales pueden ser un paso muy interesante a implementar, ya que brindan un procesamiento alterno para resolver este problema, pero que por su complejidad no se aborda en el presente trabajo, pero seria el siguiente paso lógico, ya que continua la mímica con los seres biológicos al imitar el procesamiento de un ser humano, que después de todo es la mejor maquina de reconocimiento de voz e identificación de locutores.

Formación del espacio vectorial

Continuando con el ejemplo anterior y en lo que respecta al reconocimiento de un patrón pensemos que tenemos una serie de tomas de voz o grabaciones ideales, las variaciones de velocidad de dicción son mínimas, la entonación es la misma, las

variaciones anímicas de la voz son mínimas y no hay ruidos externos, por lo que obtendríamos un vector de datos de cada muestra, al cual llamaremos z_i.

Al ubicar estos vectores en un espacio vectorial vemos que se pueden agrupar a los que son sumamente parecidos entre si en patrones, donde cada patrón representa un fonema, como se observa en la figura 39, por supuesto, para poder clasificarlos por su similitud será necesaria una comparación de distancia entre los distintos vectores y determinar un error o distancia máxima para poder considerarlo dentro de un solo patrón o pertenecientes al mismo fonema, lo cual se conoce como "Campo de Dispersión del Fonema" pero teniendo cuidado de no caer en las zonas de transición de un fonema al inmediatamente continuo, ya que recordemos que en el habla coloquial no existe separación entre los diversos fonemas que conforman una palabra o incluso una frase.

Una vez que tenemos un espacio vectorial con los 24 patrones (24 fonemas en el español) óptimamente o su mayor numero posible esto podrá englobarse en un patrón general para cada individuo, ya que el objetivo de este trabajo es el reconocimiento del locutor y no de que se dice, por lo que formaremos ahora un numero finito o infinito de patrones según el numero de individuos en estudio o si se desea formar una base de datos de voces de sujetos con fines forenses, por ejemplo todos los individuos sujetos a un proceso legal (ingreso a reclusorio).

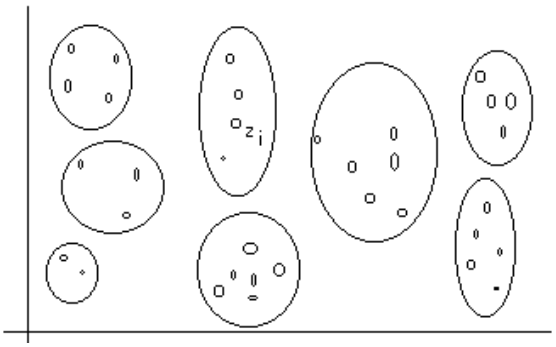


Figura 39: Ejemplo de patrones en un espacio vectorial

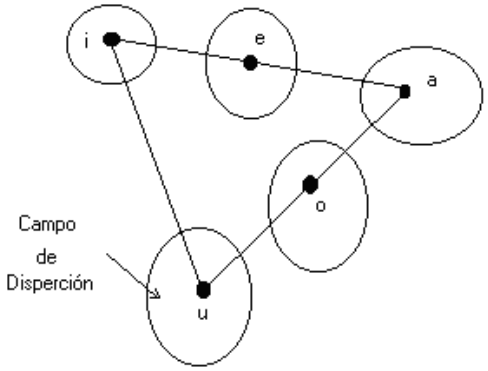


Figura 40: Campos de dispersión o alófonos de las vocales

Ahora que ya tenemos un espacio vectorial llamaremos los elementos S_i a cada patrón de un individuo, y dentro de estos llamaremos Z_i a los patrones de cada fonema, por lo que tendremos un espacio compuesto por *n* numero de patrones S y cada uno de ellos compuesto a su vez por un máximo de 24 patrones Z para este espacio vectorial propuesto, quedando claro que no es la única solución ni la mejor quizás, sino tan solo una propuesta inicial.

Reconocimiento de un individuo

De igual manera y continuando con la propuesta vamos a suponer que se captura una voz que recibirá el mencionado procesamiento en donde la grabación será dividida en un *m* numero de vectores Z^*, donde cada vector representa a un fonema, por ejemplo, la palabra "patata" estará conformada por 6 fonemas, pero observemos que se repiten algunos de ellos, por lo que solo tenemos 3 fonemas únicos.

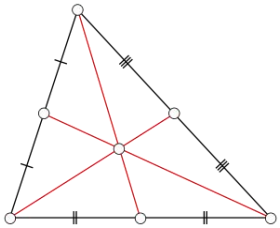
Uno de los principales problemas como veremos mas adelante será que la computadora pueda distinguir los fonemas que conforman una palabra, pero suponiendo que ya logramos esto, tenemos 3 vectores, Z_1^*, Z_2^*, Z_3^*, los cuales se pudran agrupar en un único locutor, el cual a su vez podrá compararse contra el espacio vectorial de un sujeto desconocido para tratar de identificarlo.

Diremos que S^* pertenece a S_i sí se cumple que:

$$S^* \cong S_i \text{, donde } Z_1^* \cong Z_1^i,\ Z_2^* \cong Z_2^i \text{ y } Z_3^* \cong Z_3^i \qquad (9)$$

Ahora los patrones Z_i de cada patrón S_i debe representarse de alguna manera, por lo que usaremos el concepto de baricentro, mas sin embargo por facilidad de comprensión usaremos el concepto de centro de masa que en los casos donde coincide con el centro de gravedad y centroide puede usarse de manera indistinta, los primeros dos se refieren a cualidades físicas, y el tercero a un aspecto geometrico, mas sin embargo en nuestro caso los usaremos como una analogía para entender mejor el concepto abstracto.

Un centroide tiene la siguiente definición:
En geometría, el centroide o baricentro de un objeto X perteneciente a un espacio n-dimensional es la intersección de todos los hiperplanos que dividen a X en dos partes de igual n-volumen con respecto al hiperplano.



Para entender mejor esto, el centro de gravedad (recordemos que para esta explicación son intercambiables y coinciden en el mismo lugar) será aquel punto que nos permite balancear un objeto y sostenerlo de manera estable con un solo punto de apoyo, y desde el punto de vista del centro de masa será el punto geométrico que dinámicamente se comporta como si estuviese sometido a la resultante de las fuerzas externas al sistema, por lo que vemos que en general es el punto que *representa* al cuerpo ya sea desde el punto de vista físico o geométrico, y en nuestro caso especifico será el punto que representa a todo el vector o conjunto de vectores, que es en si el baricentro del patrón en cuestión.

Teniendo esto en cuenta sabemos que se obtendrá en base a los vectores originales que conforman el fonema en cuestión un único vector que representa de forma optima a todos los vectores con los que se construye el patrón, aunado a la distancia máxima del baricentro al campo de dispersión para saber que fonemas pertenecen al mismo patrón y cuales caen fuera o están en el margen critico de pertenencia.

Es fácil vislumbrar que nunca tendremos un vector idéntico al otro, por lo que por métodos estadísticos se puede obtener la distancia de un vector al baricentro y con un margen de error se determina la pertenencia a un grupo definido, la distancia más simple es la euclidiana o Euclidea, representada por la siguiente ecuación:

$$d(Z_{testigo}, Z_{problema}) = (\sum_{k=1}^{m} (Z_{testigo}(k) - Z_{problema}(k))^2)^{1/2} \qquad (10)$$

Esta ecuación es muy simple, nos dice que para cada vector Z de *m* elementos se restara el elemento *k* (que va de 1 a *m*) del testigo contra el mismo *k* elemento del problema, elevándose al cuadrado dicho resultado parcial y obteniéndose la raíz cuadrada de toda la sumatoria de m elementos.

Apoyado en esta ecuación podemos cambiar la regla de decisión de manera que al grupo que presente la menor variación se le establece un porcentaje de proximidad, lo cual es lo mismo que un porcentaje de certidumbre de que se trata del mismo individuo.

$$S^{problema} \approx S_i \text{ si } d(S^{problema}, S^i) < d(S^{problema}, S^j) \text{ para todo } j \neq i \qquad (11)$$

Esta ecuación nos dice sencillamente que el sujeto planteado como problema tiene una mejor probabilidad de ser el elemento *i* del conjunto S si las distancias Z en general son más pequeñas que para el resto de los *j* elementos.

Normalización

En algunas ocasiones será conveniente llevar a cabo una normalización de los datos antes de proceder con los cálculos ya descritos o por describir, si es el caso podemos utilizar el siguiente procedimiento para lograr esto.

$$Z_k' = \frac{Z_k}{D}, \text{ donde } D = \max(Z_k) - \min(Z_k) \qquad (12)$$

En esta ecuación se nos dice que se dividirán todos los elementos del vector Z_k por un escalar D que será obtenido para cada vector Z_k al restar al máximo valor localizado dentro de este mismo el mas pequeño.

Ventanas

Las ventanas son funciones matemáticas usadas con frecuencia en el análisis y el procesamiento de señales para evitar las discontinuidades al principio y al final de los bloques analizados.

En procesamiento de señales, una ventana se utiliza cuando nos interesa una señal de longitud voluntariamente limitada. En efecto, una señal real tiene que ser de tiempo finito; además, un cálculo sólo es posible a partir de un número finito de puntos.

La utilización de ventanas cambia el espectro en frecuencia de la señal. Existen distintos tipos de ventana que permiten obtener distintos resultados en el dominio de las frecuencias como veremos más adelante.

Su concepto es muy simple, de un universo de datos con "*n*" elementos, estudiaremos un subconjunto de estos que pertenece en su totalidad al universo y que se caracterizara por poderse mover dentro del universo sin cambiar nunca su numero "*m*" de elementos, por lo que vemos que simplemente representa una porción del total como se ve en la ilustración.

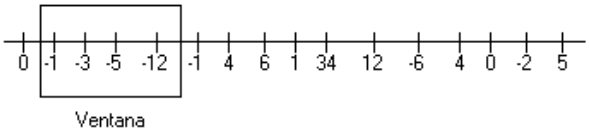


Figura 41: Ventana de datos

Aquí observamos una simple ilustración que ejemplifica una recta con números aleatorios, dentro de esta recta se tomaran solo cuatro números a la vez para su estudio, pudiendo moverse la ventana a cualesquiera cuatro números consecutivos que se pretendan estudiar.

Tipos de ventanas y solapamiento

Una vez visto la idea básica podemos adentrarnos un poco más, la ventana que vimos con anterioridad es una ventana rectangular, debido a que todo lo que esta dentro de ella permanece con su valor original, y todo lo que este fuera de ella se tomara como cero, pero también tenemos otro tipos de ventanas, como son ventanas triangulares, Hamming, Hann y muchas otras, en las que se modifican los valores originales por medio de una función matemática que rige dicho tipo de ventana, esto a fin de evitar errores de calculo, efectos de borde o discontinuidades y en ocasiones incluso por su efectos en la obtención de señales en el dominio de la frecuencia.

Efecto de borde o discontinuidades

Si los valores que están en la orilla de la ventana son muy grandes, podría presentarse el efecto de borde o discontinuidades, esto es que se marcara de manera muy acentuada la diferencia entre una ventana y la inmediatamente continua, para evitar esto se utiliza un tipo de ventana que reduce los valores en sus extremos o incluso se pueden solapara las ventanas, esto es, una ventana toma los elementos 1, 2, 3 y 4, la siguiente puede tomar los elementos 3, 4, 5 y 6, solapándose con la primera en un 50 por ciento, y si se aplican ambas técnicas simultáneamente se conseguirá el mejor efecto posible para señales de voz, evitando en gran medida el efecto de borde.

Ejemplo grafico de una ventana

Para entender mejor cada uno de estos conceptos veamos un rápido ejemplo de que hace una ventana a un conjunto de datos, para estos ejemplos usaremos la ventana de Hamming, ya que de manera empírica se ha visto que es la que da mejores resultados para señales de voz, que es el enfoque que nos interesa, para esto tendremos una serie de valores de lo que podría ser una señal de audio y su respectiva aplicación de ventanas.

De primera instancia mostraremos la grafica de una señal compuesta por un 67 datos y con un relleno de 13 ceros al final para completar las cuatro ventanas que proponemos, cada ventana tomara una subconjunto de 20 datos, por lo que tendremos una ventana del dato 1 al 20, del 21 al 40, del 41 al 60 y del 61 al 80 consecutivamente.

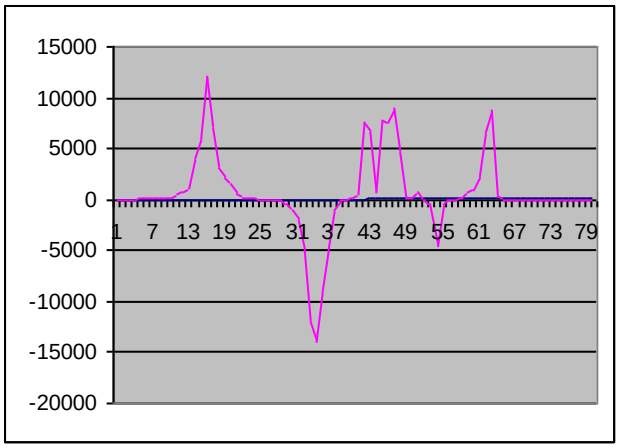


Figura 42: Serie de 80 datos aleatorios

Una vez que hemos visto la graficación de los datos sin ventana alguna agregaremos la grafica de los datos posteriormente a la aplicación de la mencionada ventana de Hamming con los siguientes resultados.

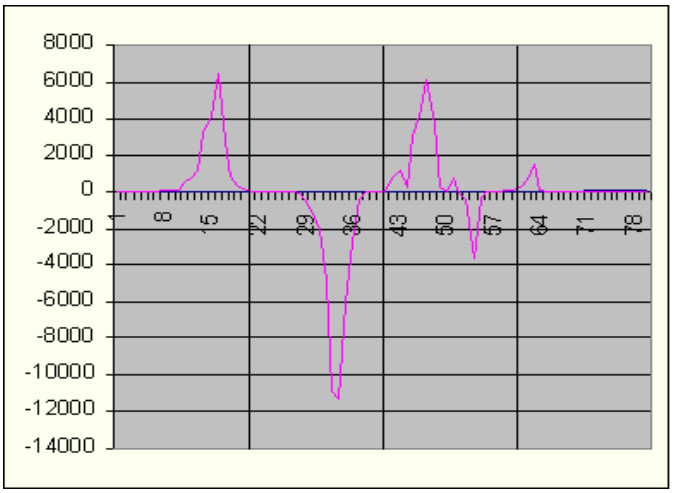


Figura 43: Aplicación de cuatro ventanas de 20 datos

Como podemos apreciar en la figura 43 la ventana de Hamming nos da una disminución de los valores grandes que se encuentran cercanos a las orillas de las ventanas y en general una adaptación de los valores a una señal coseno, la cual se ejemplifica en la figura 44, que muestra la grafica de los valores de la ventana mencionada.

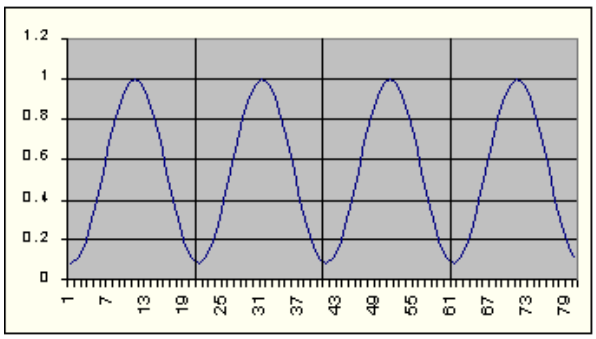


Figura 44: Graficación de los valores de una ventana de Hamming

Como se puede observar, todos los valores de la función de Hamming son positivos y mayores de cero, razón por la cual tan solo se generara una atenuación en la función a la

que apliquemos la ventana, más no se generan cambios de signo o eliminación de datos, ya que esto alteraría la señal y por ende el resultado final.

Solapamiento

Con respecto al solapamiento de ventanas, en los estudios de señales de voz, que es en si una señal de naturaleza cambiante, pero a su vez quasi-estática en periodos cortos de tiempo, se ve que el estudio por medio de ventanas optimiza los resultados y nos permite obtener parámetros representativos de dicha porción de audio, pero sumado el solapamiento de ventanas que suavizara los resultados es un modo ideal de realizar este tipo de estudios.

Para permitirnos observar la evolución de los distintos parámetros que se estudien como son formantes o tono fundamental podemos solapar ventanas en un determinado porcentaje, por ejemplo 10%, lo cual significa que si tomamos ventanas de 20 mseg. (Ventana temporal en que la voz se comporta de manera quasi-estática) se solapara una ventana sobre la anterior en 2 mseg., lo cual aumenta él numero de ventanas y de cálculos, pero permitirá un resultado más suave y homogéneo entre regiones, así como minimizar la perdida de información por la aplicación de dicha ventana.

De manera grafica esto se podría ejemplificar así, pero cabe mencionar que carece de significado físico, ya que se repiten porciones de información, pero para fines de cálculo nos da mejores resultados como veremos mas adelante.

Como se puede apreciar fácilmente, entre la grafica de la función con la ventana de Hamming aplicada anteriormente y la figura 45 existen zonas en que se aprecia una menor atenuación en porciones que anteriormente se atenuaban fuertemente, por lo que resultaría obvio que al calcular algunos parámetros a partir de dichos datos, serán más representativos los de las ventanas solapadas que los de las ventanas consecutivas.

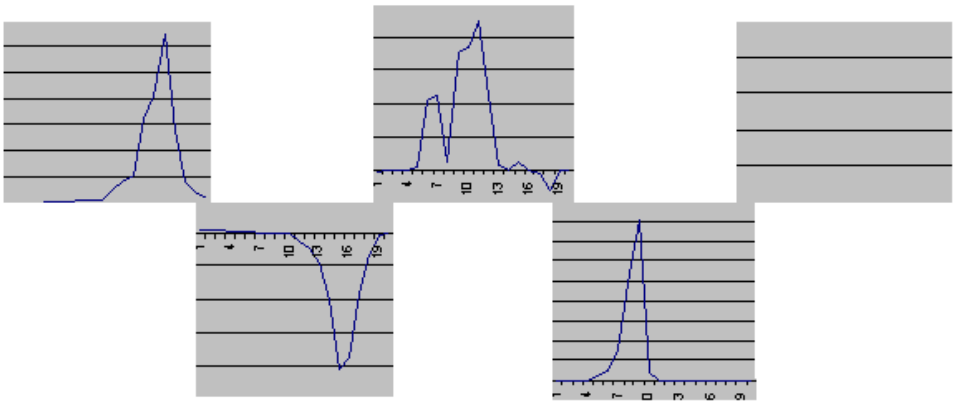


Figura 45: Ventanas solapadas al 10%

Es un poco difícil comprender de entrada el beneficio de la aplicación de estas ventanas matemáticas, pero veremos en los resultados finales del presente trabajo los diferentes resultados obtenidos con y sin la aplicación de este pre-procesamiento.

Como un ultimo agregado con respecto de la aplicación de ventanas y sus beneficios, mostraremos de manera grafica sus efectos, ya que no hemos cubierto que es un espectrograma, veremos el efecto que se obtiene al aplicar una ventana matemática y por que nos es útil aplicar este algoritmo antes de transformar nuestra señal,

consideremos la onda resultante del ejemplo de *sonidos complejos*, el resultado de aplicar una ventana de Hamming a dicha señal se aprecia en la figura 46.

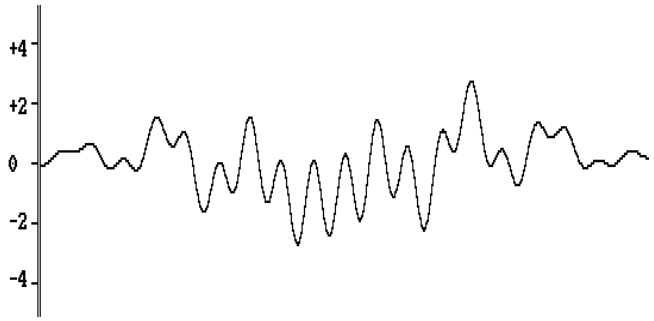


Figura 46: Suma de las señales con una ventana de Hamming

Aquí vemos que esta señal compleja después de aplicarle una ventana de Hamming se ve en general amoldada a una señal coseno y suavizada en sus extremos, lo cual ya habíamos visto, mas sin embargo si posteriormente aplicamos una transformada de Fourier y graficamos el resultado antes y después veremos lo siguiente:

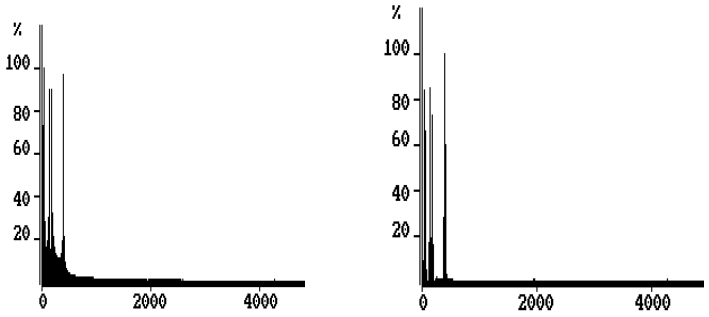


Figura 47: Espectro de la señal antes y después de la ventana

De estos espectrogramas podemos apreciar que se percibe mucho mas ruido en la grafica antes de aplicar la ventana de Hamming, esto se debe a que al aplicar la ventana ajustamos las amplitudes de la señal a una función coseno con amplitud igual a la sección a la que se aplica la ventana, por lo que vemos disminuida la amplitud de la señal en los extremos, de lo contrario las discontinuidades generadas por el procesamiento al realizar la transformación se ven convertidos en ruido y no nos dejan apreciar claramente las componentes en frecuencia de nuestra señal.

Ventanas más comunes

Como algunas de las ventanas más comunes tenemos la ventana rectangular, la ventana de Hamming, la ventana de Hann y algunas otras, la diferencia entre distintas ventanas será la función matemática por la cual se rigen, mismo por lo que se modificaran los datos cuestión de estudio.

A continuación se presentan los algoritmos de algunas de las mas ventanas comunes y su graficas correspondientes para tener un mejor entendimiento del efecto que causa cada una de estas, recuérdese que dichas funciones tan solo nos darán los coeficientes de la ventana, los nuevos coeficientes de la señal de audio se calcularan a partir de la función $\hat{c_m} = f_m * c_m$ donde los coeficientes primos son el resultado de la aplicación de la ventana.

También se agrega la grafica del espectro de una señal senoidal simple a la que se le aplico la ventana para que se pueda apreciar su efecto en el dominio de la frecuencia.

Ventana rectangular

$$f(n)=\begin{cases} w(n)=1 \rightarrow 0 \le n \le N-1 \\ w(n)=0 \rightarrow resto \quad de \quad la \quad señal \end{cases}$$

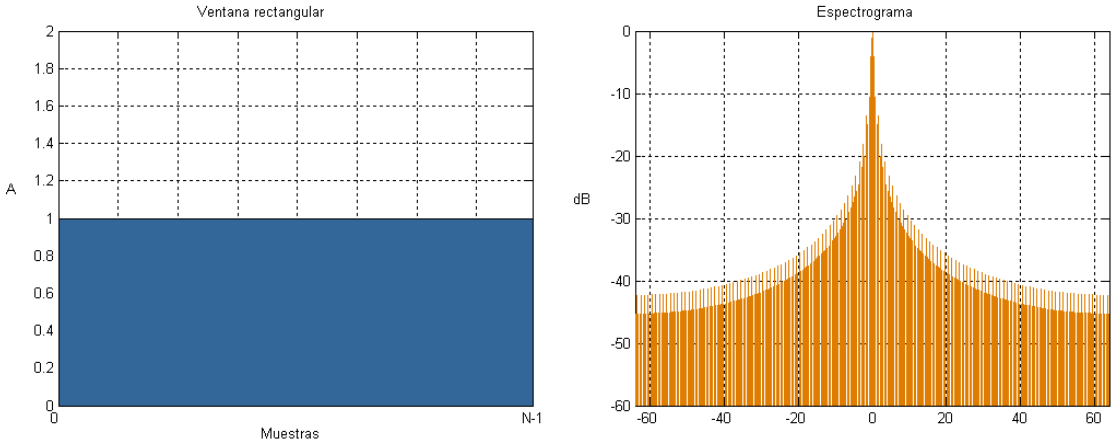


Figura 48: Ventana rectangular

Ventana Hamming

$$f(n)=\begin{cases} w(n)=0.54-0.46*\cos(2\pi n/N) \rightarrow 0 \le n \le N-1 \\ w(n)=0 \rightarrow resto \quad de \quad la \quad señal \end{cases}$$

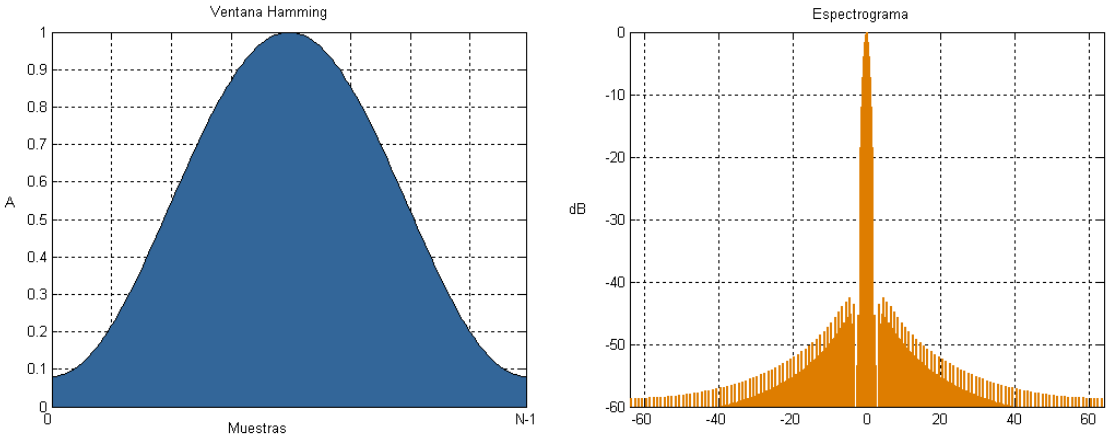


Figura 49: Ventana Hamming

Ventana Hann

$$f(n)=\begin{cases} w(n)=0.5-0.5*\cos(2\pi n/N) \rightarrow 0 \le n \le N-1 \\ w(n)=0 \rightarrow resto \quad de \quad la \quad señal \end{cases}$$

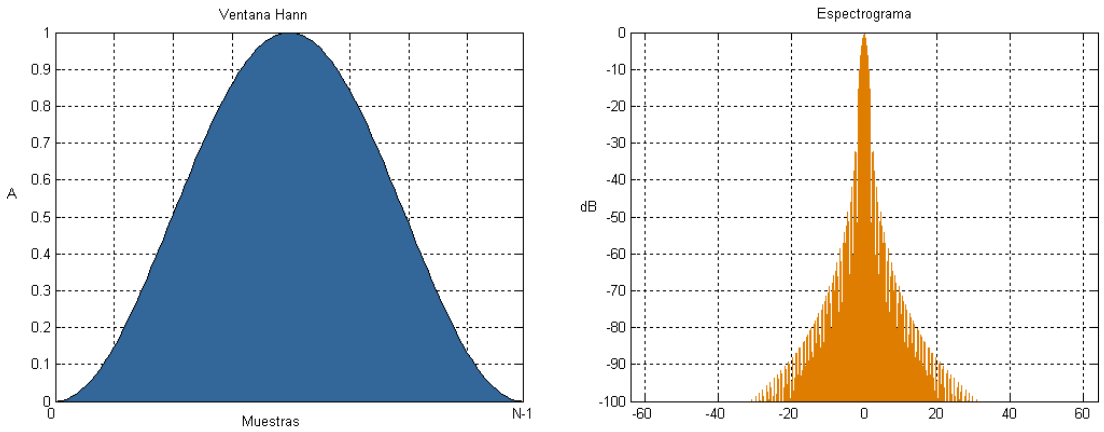


Figura 50: Ventana Hann

Ventana Blackman

$$f(n)=\begin{cases} w(n)=0.42-0.5*\cos(2\pi n/N-1)+0.08*\cos(4\pi n/N-1)\rightarrow 0\le n\le N-1 \\ w(n)=0\rightarrow resto \quad de \quad la \quad señal \end{cases}$$

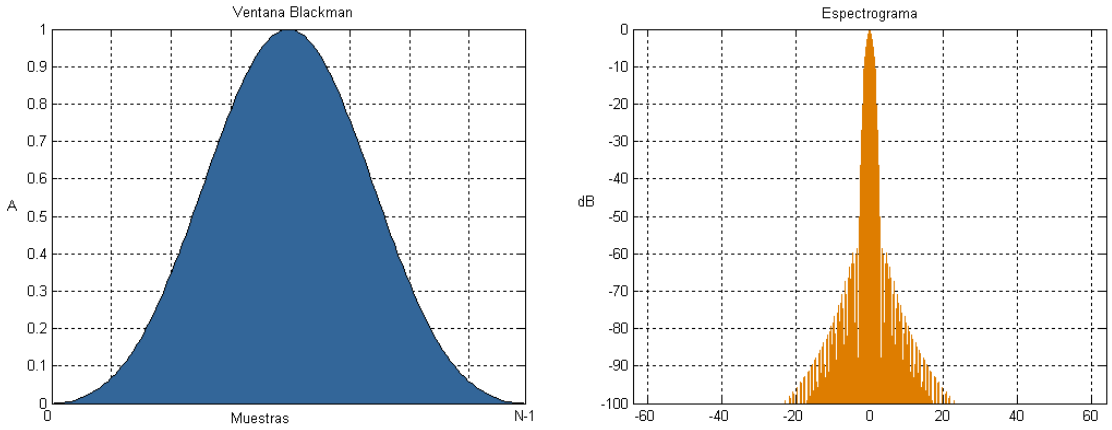


Figura 51: Ventana Blackman

Ventana Gauss

$$f(n)=\begin{cases} w(n)=e^{-\frac{1}{2}\left(\frac{n-(N-1)/2}{\sigma(N-1)/2}\right)^2} \\ w(n)=0\rightarrow resto \quad de \quad la \quad señal \end{cases} \qquad \text{Donde } \sigma\le 0.5$$

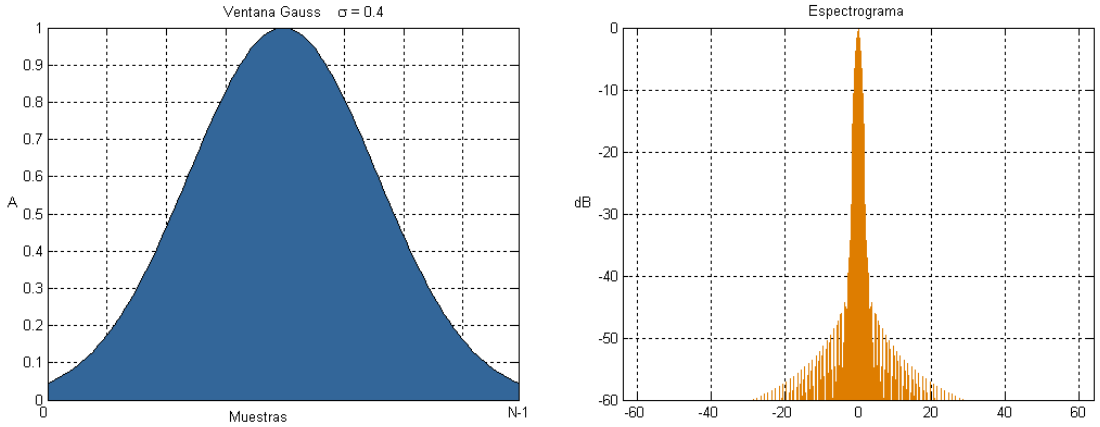


Figura 52: Ventana Gauss

Ventana Triangular

$$f(n)=\begin{cases} w(n)=\dfrac{N}{2}-\left|n-\dfrac{N-1}{2}\right| \\ w(n)=0 \rightarrow resto \quad de \quad la \quad señal \end{cases}$$

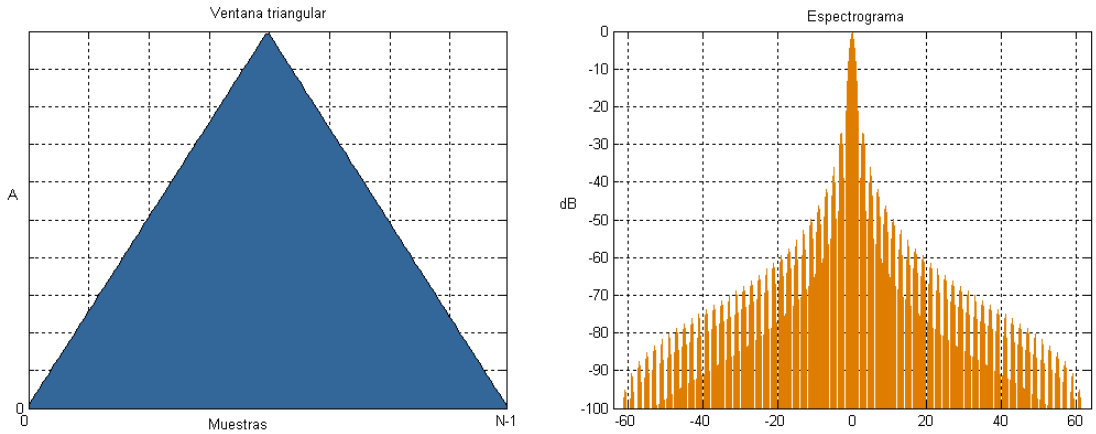


Figura 53: Ventana Triangular

Parámetros en el tiempo

Los parámetros en el tiempo son los de mas fácil extracción, ya que el archivo no requiere de ningún procesamiento previo para su análisis, en este tema se pretende explicar la metodología que se implementara, así como una breve explicación de que representa cada uno de los parámetros que aquí se trabajan y cual puede ser su aportación a la identificación de un sujeto.

Como es natural se tiene la sospecha de que el camino mas fácil es normalmente el equivocado, mas sin embargo se puede apreciar también que tiene cierta lógica el que puede ser un camino normalmente olvidado y dado por erróneo por su misma simpleza, por lo que se considera adecuado utilizar esta aproximación y estudiar sus resultados contra técnicas mas complejas.

En esta aproximación se buscaran parámetros dentro de una señal de audio, en especifico de voz, que además de que sean representativas de la señal en si sean de carácter distintivo entre distintos sujetos en estudio, ya que se piensa que los parámetros

que a continuación se explican sean lo suficientemente propios de cada voz como para representarlo como una clase distinta de las restantes e identificable.

Es importante destacar que el proceso de obtención de parámetros en el tiempo, a pesar de su simpleza deberá de realizarse con cierto cuidado, ya que si obtenemos un parámetro para toda una palabra por ejemplo veremos que este parámetro es poco representativo, por lo que se hace necesario el calculo de dichos parámetros para porciones de audio menores, llamadas ventanas, dichas ventanas pueden presentar a su vez algún tratamiento como son las ventanas de Hamming, de Hann y demás tipos conocidos, así como solapamientos como ya se explico en el apartado correspondiente a ventanas.

En cuanto al tamaño de la ventana, es conveniente observar cautela en la selección de dicho parámetro, una ventana muy corta producirá un exceso de datos repetitivos, y una ventana muy grande pasara por alto información importante. Recordando de diversos estudios sobre la voz humana sabemos que esta tiene un comportamiento quasi estacionario en periodos de aproximadamente 10 mseg, periodo en el cual sus parámetros son sumamente parecidos en todas las muestras que tomamos, por lo que se considera el numero ideal de muestras para parámetros mas representativos, y en cuanto al solapamiento de las ventanas sabemos que no deberá de ser muy grande para no repetir información ni muy pequeño para que siga teniendo algún efecto en la señal estudiada.

En cuanto al tipo de ventana, se ha observado durante la historia del procesamiento digital de la voz que da mejores resultados una ventana Hamming, mas esto no descarta que pueda ser de interés otro tipo de ventana, en el presente trabajo se tomara únicamente la ventana de Hamming por su ya reconocida trayectoria, dejándose abierta la posibilidad de otras ventanas en futuras revisiones.

Energía

Ahora que ya tenemos un tamaño de ventana definido de 10 mseg y una sugerencia de solapar las ventanas en un 10%, se dejara la decisión final al usuario, pudiendo variarse desde el 2% hasta el 15%, y usando o no una ventana de Hamming, pero iniciaremos nuestro estudio por obtener el parámetro de la energía de la señal, la energía estará representada por la sumatoria del cuadrado de los elementos de la ventana dividido entre el numero de muestras de dicha ventana, como se ve en la siguiente expresión.

$$E(n) = \frac{1}{N}\sum_{k=0}^{N-1} x^2(k) \qquad (13)$$

Energía; virtud para obrar o producir un efecto.

De lo anterior y basándonos en al ecuación que define la energía de una señal vemos que su extracción es simple, tomaremos los datos que debe de tener una ventana, mismos que serán elevados al cuadrado y sumados, dividiéndose por ultimo por el numero de elementos de una ventana, aunque cabe hacer mención que no se da un numero fijo de muestras para un periodo de 10 mseg ya que esto dependerá de la frecuencia de muestreo en que se digitalizo el archivo en estudio.

Esta acción será repetida con un desplazamiento de tantas muestras como el tamaño de la ventana menos el porcentaje de solapamiento hasta procesar el total de los datos que contiene el archivo en cuestión, información que después podrá ser graficada o analizada estadísticamente para su comparación contra otras muestras.

Como resultado tendremos una curva de la distribución de energía y su evolución en el tiempo conforme el sujeto habla, mismo patrón que podrá ser comparado contra la curva de energía de otra muestra de voz en busca de similitudes y diferencias para establecer un porcentaje de certidumbre de pertenencia, esto es, que probabilidad existe de que se trate del mismo sujeto, mismo que puede ser logrado mediante la medición de la distancia de una curva contra otra que se presuma del mismo locutor.

Media

Por lo que respecta a la media su obtención es sumamente parecida a la de la energía, ya que como se ve incluso son sumamente parecidas sus formulas, a excepción de que para la media no se elevan al cuadrado las muestras, sino simplemente se sumaran las muestras de una ventana y se dividirán entre el numero de muestras de dicha ventana.

$$\overline{M}(n) = \frac{1}{N}\sum_{k=0}^{N-1} x(k) \qquad (14)$$

Media; cantidad que representa el promedio de varias otras.

La media como lo dice su definición es el valor promedio o medio del numero de muestras contenidas en una ventana, valores que también pueden ser graficados, obteniendo una representación alterna de la graficación de un archivo “wav”, pero en lugar de graficar la amplitud de cada muestra con respecto al tiempo se graficara su media con respecto al numero de ventanas, lo que es una representación compacta del comportamiento de la amplitud de la señal con respecto al tiempo.

Magnitud

La magnitud se obtiene también de una manera muy similar a la media, pero en esta ocasión será la sumatoria de los valores absolutos de las muestras dividido por el número de muestras en una ventana, tomándose en cuenta solamente su magnitud o tamaño contra las muestras vecinas.

$$M(n) = \frac{1}{N}\sum_{k=0}^{N-1} \left|x(k)\right| \qquad (15)$$

Magnitud; cualidad de un cuerpo o fenómeno que, referida a una unidad de la misma especie, puede ser medida.

Como se puede suponer este parámetro también es de fácil interpretación en su forma grafica, y es sumamente parecida a la grafica de la media, pero en esta ocasión no se tomaran en cuenta los cambios de signo de la señal, que recordemos que estos cambios de signo representan presiones superiores al promedio ambiental (valores positivos), y valores inferiores al promedio ambiental (valores negativos), que en toda señal de voz se encuentran, ya que se presenta una onda de presión superior seguida de una de presión inferior a la estándar como en las ondas que genera una piedra al se arrojada a

un estanque con agua. Al analizar estos resultados veremos que es mas simple en ocasiones ver como se comporta la magnitud de la señal sin tomar en cuenta los cambios de signo que se graficarían como súbitas bajadas y subidas bruscas en la grafica.

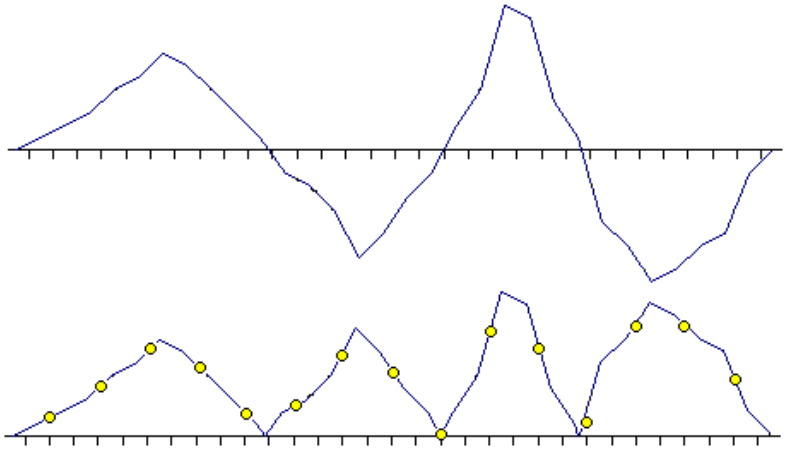


Figura 54: Ejemplo de magnitud

Como se ve en la figura 54 tenemos una señal que podría ser audio, con valores positivos y negativos, en la parte inferior vemos que se han obtenido el valor absoluto de todos los valores, por lo que ahora todos los valores son positivos, y se divide la señal en partes, siendo cada punto amarillo el representante de la función en dicha ventana, por lo que vemos que se forma una nueva señal con muchos menos datos y que representa de manera general a la anterior, siendo esta mas fácil de analizar y de proceso mas rápido.

Cruces por cero

Los cruces por cero son una medida simple que nos dice cuantas veces la señal atraviesa el valor cero o línea de referencia, esto es, de un valor positivo pasara a uno negativo o vise versa, incluso tomando el valor de cero para después asumir un valor de signo contrario al valor anterior al cero, y esta es una medida que nos puede ayudar tanto a determinar la naturaleza de la señal que estamos viendo y algunas características de la misma voz humana como se explicara después.

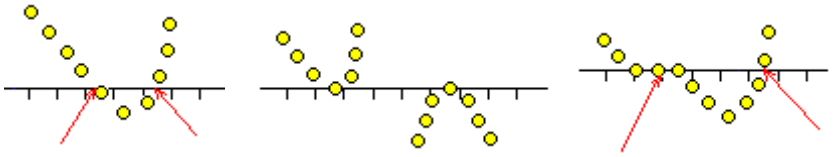


Figura 55: Cruces por cero

En la primera grafica de la figura 55 vemos que se presentan dos cruces por cero señalados por las flechas rojas, ya que tenemos que aunque nunca se toma el valor de cero en la señal se tiene un cambio de signo de un punto al siguiente, en la segunda grafica, vemos que a pesar de que se toma el valor de cero en ambos casos nunca se cruza la línea cero, por lo que no pueden ser considerados cruces por cero, y en la tercera grafica observamos que a pesar de que se tienen múltiples valores sobre la línea de cero solo se consideran dos cruces por cero cuando la señal tiene valores positivos y negativos de cada lado de los valores cero, este algoritmo es el que se implementara en el sistema para identificar cruces por cero.

Ahora para entender mejor la utilidad de este parámetro pensemos en los sonidos y como se conforman, sabemos que el ruido presenta un gran numero de cruces por cero por su naturaleza sumamente cambiante, así mismo sabemos por lo ya estudiado que ciertos sonidos como los fricativos o los vibrantes presentan un elevado numero de cruces por cero por la naturaleza vibrante de dichos sonidos, mientras que los nasales presentan un numero bastante pequeño de estos, por lo que podemos suponer que esta característica puede ayudarnos a ver que tipo de sonido tenemos y que elementos propios de cada sujeto presenta esta señal.

Máximos y Mínimos

En cuanto a los máximos relativos, estos se presentaran cuando se llega a un valor alto y después se produce un descenso de la señal, a esto se le conoce como una cima de la grafica, deberá de presentarse un valor máximo o alto para inmediatamente después seguirle valores menores a la cima en ambas direcciones.

Para los mínimos relativos, deberá de presentarse un valor mas bajo que todos los valores aledaños, esto es conocido como una sima de la grafica, debe presentarse un valor menor a la generalidad con valores mayores de ambos lados de la sima, y al igual que los máximos relativos pueden presentarse cantidad de ellos en una porción de audio común, ya que la señal constantemente sube y baja, dando un numero variable de estos elementos, así como otras características ya discutidas.

Esta medida es similar a los cruces por cero, nos dará una medida de que tanto y tan violentamente evoluciona la señal, valores que pueden englobar características propias de la voz de un sujeto y mas aun características propias del tipo de sonido que se a realizado, por lo que se considera interesante observar su comportamiento en una señal de audio.

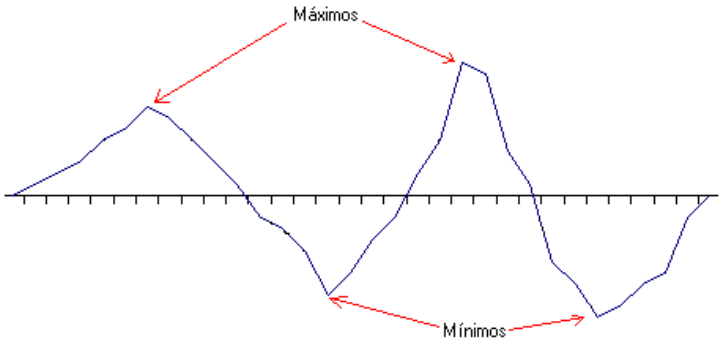


Figura 56: Máximos y mínimos

Como se ve en la figura 56, se tienen dos máximos relativos, dos mínimos relativos y tres cruces por cero para esta ventana de datos, elementos que al ser cuantificados podrán ser elementos que aporten a la identificación de un sujeto así como al aislamiento y clasificación de los diversos fonemas que se encuentran plasmados en la muestra de voz.

Sumario

Como un breve resumen de los elementos que nos pueden ser de interés se presenta una tabla con algunas características obtenidas de forma empírica al estudiar diversas señales de voz con los diferentes grupos de fonemas del español.

Como ya hemos visto, los parámetros de energía, magnitud y media son sumamente parecidos, y aunque representan diversas características de la señal pueden ser englobados para una mas fácil distinción por su miembro mas representativa, en este caso la energía, y lo mismo ocurre para los cruces por cero, máximos y mínimos, estos pueden también ser representados en general por su miembro mas notable, los cruces por cero, razón por la cual solo se habla en la tabla de energía y cruces por cero, pero en análisis mas detallados es conveniente manejar todos los parámetros sin importar su similitud, ya que cada parámetro es aportativo y distinto del anterior.

Esta tabla como ya se menciono se formo de la simple observación, y solo tiene como finalidad orientar con respecto de la utilidad de este estudio, mas no es de carácter definitivo y mucho menos puede tomarse como una referencia, ya que esto requeriría de mayor estudio y comprobación para poderse exponer de manera mas precisa.

Es también importante mencionar que estos parámetros pueden ser usados en la implementación de un detector de silencio, el cual buscara un elevado número de cruces por cero así como un muy bajo nivel de energía, pudiendo aislar así el silencio de las palabras útiles para su estudio.

	Energía	Cruces por Cero
Oclusivas	Reducida	Escasos a normal
Nasales	Reducida	Escasos
Fricativas	Muy reducida	Muy abundantes
Africadas	Menor al promedio	Muy abundantes
Laterales	Promedio	Promedio
Vibrantes	Promedio	Muy abundantes
Vocálicas	Muy estable	Muy estable

Tabla 6: Aportaciones en el tiempo

Parámetros en el dominio de la frecuencia

El matemático francés Joseph Fourier nacido en 1768, es conocido por sus importantes contribuciones en los campos de la matemática y la física matemática, aporto principalmente las series de Fourier, que permiten expresar funciones discontinuas como la suma de una serie infinita de senos y cósenos, y la transformada de Fourier, que nos permite el paso de una función en el dominio del tiempo al dominio de la frecuencia.

Joseph Fourier

La transformada de Fourier es una herramienta de análisis muy utilizada en diversos campos científicos, entre los cuales nos interesa en especifico el procesamiento digital de señales de audio, esta solo transformara una señal representada en el dominio del

tiempo al dominio de la frecuencia sin alterar su contenido de información, sólo es una forma diferente de representarla.

Recordemos que después de todo, este tipo de procesamiento es ideal para aplicaciones que involucran sonido, ya que el oído humano realiza una descomposición similar a nivel del caracol, el cual, por medio de una variación de diámetro en su conducto ira separando en manera natural las diferentes componentes en frecuencia de un sonido, mismas que serán recogidas por diversos nervios a través del caracol.

También resulta conveniente recordar que las series de Fourier nos dicen que podemos descomponer una función discontinua como una sumatoria infinita de senos y cósenos, o de funciones periódicas simples que son lo mismo, que después de todo es lo que realiza desde un punto de vista matemático la transformada de Fourier y el oído humano mismo.

Dejaremos de lado las transformaciones de una función continua, debido a la naturaleza de las señales que trataremos y que estas serán procesadas mediante programas de computo que digitalizaran y manejaran siempre de forma discreta las señales de audio, por lo que usaremos una aproximación o implementación a dicho algoritmo de una manera computacionalmente eficiente llamada transformada rápida de Fourier o la transformación de Fourier para señales discretas.

A su vez, también dejaremos de lado las anti-transformaciones o el regreso de las señales al dominio del tiempo, ya que para los fines que perseguimos no es útil dicho algoritmo, puesto que partimos de una señal de voz digitalizada en el dominio del tiempo y llegaremos a su transformación al dominio de la frecuencia, su extracción de parámetros y por ultimo la interpretación de estos resultados.

La transformada de Fourier aplicada al procesamiento digital de señales

La transformada de Fourier, y la ya mencionada transformada rápida de Fourier tienen un gran campo de aplicación en el procesamiento digital de señales, pudiendo encontrar entre otras aplicaciones las telecomunicaciones, procesamiento de video, procesamiento de voz y muchas otras.

Estos algoritmos se implementan ya sea en computadoras de uso general o de uso especifico, y pueden tener campos de aplicación tan bastos como la calibración y corroboración de correcto funcionamiento en equipos de transmisión de datos, como satélites y microondas, en algoritmos de visión robótica, en aplicaciones de reconocimiento de voz, y muchas otras en las que es útil una representación alterna de la información original.

En lo que respecta a señales de audio, que es lo que nos interesa en el presente trabajo, el proceso de transformación es la descomposición de una señal compleja en un sumario de señales simples. Esta misma acción es la que se lleva a cabo en el caracol

del oído humano, se presenta una descomposición del sonido en sus frecuencias fundamentales y así se pasa la información para su posterior proceso en el cerebro.

Anteriormente se ejemplifica la formación de un sonido complejo basado en la suma de señales simples, por lo que no tiene caso repetir este proceso, pero la transformada de Fourier tiene la función inversa, a partir de una señal compleja nos entregara una serie se señales simples que son equivalentes, esto es, sus componentes en frecuencia.

De esta explicación se ve de forma clara que la transformada de Fourier es una poderosa herramienta en el análisis de una voz, ya que haciendo una mímica de los procesos mas simples que se llevan a cabo en un ser humano podemos abordar el problema de distinguir a un hablante de un grupo de estos.

Para entender un poco mejor lo que estamos haciendo nos basaremos en la ecuación básica de Fourier que nos dice:

$$F\left(\frac{n}{NT}\right) = F\left(\frac{n}{N} f\right) = \sum_{k=0}^{N-1} m_k e^{-j\frac{2\pi nk}{N}} \qquad (15)$$

Donde '*N*' es el numero de muestras por ventana, '*T*' es el periodo de muestreo, '*f*' es la frecuencia de muestreo y m_k es la muestra késima de la ventana.

De esta formulación podemos ver que cuando n = 0 se trata de la frecuencia 0 Hz, por lo que representa la componente de continua de la señal si es que existe, y el resto de los valores de 'n' hasta N-1 son las distintas frecuencias que se pueden obtener o estudiar.

Según el criterio de Nyquist, el ancho de banda de la señal coincide con la mitad de la frecuencia de muestreo '*f*', por lo que '*n*' recorrerá desde cero hasta la frecuencia de muestreo y las señales de frecuencia mas alta estarán representadas en nuestra señal digital por la mitad de esta frecuencia o por '*N/2*' o sea los valores n = 0, 1, 2, 3,…, $N/2-1$, por lo que los valores que se obtienen para $0 \le n < N/2$ coinciden con los de el intervalo $N/2 \le n < N-1$ (cuando *n* es par) por lo que es suficiente realizar los cálculos en una de las dos mitades.

Es conocido también y fácil de apreciar que al aumentar el valor de '*N*' conseguiremos un análisis con un mayor numero de frecuencias o lo que es lo mismo una mayor resolución en frecuencia pero a costo de mayor tiempo de proceso por el mayor número de cálculos

Cuando tenemos ya una frecuencia de muestreo y un tamaño de ventana definido vemos que esta función dependerá únicamente del parámetro 'k', por lo que la operación primordial es la multiplicación de cada muestra m_k del bloque de datos por el valor de la exponencial, dicha multiplicación hace la función de comparador entre ambos miembros.

Dicha exponencial representa un número complejo en coordenadas polares y sabemos que $x * e^{j\theta} = x(\cos\theta + jsen\theta)$, donde x es el modulo, por lo que regresando a la función de Fourier vemos que la exponencial es un complejo girando en sentido

antihorario con una velocidad angular de $\theta = 2\pi n$ con saltos discretos de k/N radianes, por lo que la expresión anterior quedaría como:

$$\sum_{k=0}^{N-1} m_k * \left(\cos\left(-2\pi n \frac{k}{N} \right) + j * sen\left(-2\pi n \frac{k}{N} \right) \right) \qquad (16)$$

La razón por la que la sumatoria actúa como comparador es que el bloque de la señal analizada es parecida al seno (o coseno) por el que se multiplican las muestras, y el resultado será un valor diferente de cero, esto es porque los valores positivos del seno o coseno se multiplican por valore positivos de la señal y los negativos por los negativos de la señal, en el caso de que la señal no se parezca al seno o coseno, los valores positivos y negativos se irán contrarrestando y el resultado de la sumatoria se aproximara a cero.

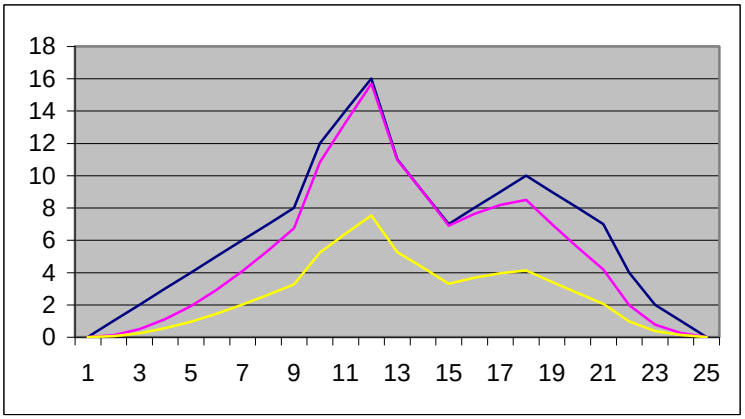


Figura 57: Señal multiplicada por una senoidal

Como ejemplo tenemos la figura 57, la cual presenta una función cualquiera en su línea azul (la grafica mas externa), y esta se multiplica por dos senoidales de distintas frecuencias, la primera, línea amarilla (línea mas interna) vemos que no es de una frecuencia similar, por lo que los resultados son de mucho menor magnitud, y en cambio la segunda senoidal es de una frecuencia muy similar a la se la función original y vemos que atenúa muy poco la función en la línea magenta, que es lo que mencionábamos, compara que tanto se parece la señal en estudio con las funciones generadas por la exponencial que es una función compleja de senos y cósenos.

De manera similar la transformada de Fourier es un proceso en el cual se comparara sucesivamente la señal compleja con una gama de funciones simples, dándonos los pesos de que tanto contribuye dicha frecuencia a la conformación de la señal original, o sus componentes en frecuencia como mejor se les conoce.

En este caso la sumatoria será un numero complejo que indica la similitud de la señal analizada con el seno y coseno de la frecuencia '*n*', la parte real se refiere al coseno y la imaginaria al seno, pero para ponderar la importancia de ambos valores normalmente se usara la distancia Euclidea o lo que es mejor conocido como el modulo de dicho numero complejo.

$$Modulo = \sqrt{(\mathrm{Re}\,al)^2 + (\mathrm{Im}\,aginario)^2} \qquad (17)$$

La transformada rápida de Fourier

La transformada rápida de Fourier (FFT) es un algoritmo que resuelve de una manera más eficiente la transformada discreta de Fourier, recordemos que la transformada discreta de Fourier se define por:

$$F\left(\frac{n}{NT}\right) = F\left(\frac{n}{N}f\right) = \sum_{k=0}^{N-1} m_k e^{-j\frac{2\pi mk}{N}} \text{ para h = 0, 1, 2,…, N-1} \qquad (19)$$

Donde '*T*' es el periodo de muestreo, '*N*' el número de puntos por ventana, '*n*' la variable índice de frecuencias y '*k*' la variable índice de la muestra.

Como una nota cabe recordar el hecho que si trabajamos con una señal a una frecuencia máxima de f_{max} y esta se muestrea al doble de dicha frecuencia (teorema de Nyquist) tendremos que $T = \frac{1}{2f_{\max}}$, por lo que al obtener su espectro calculando todas las frecuencias (h = 0, 1, 2,…, N-1) se obtiene el espectro que buscamos y su simétrico, por lo que solo es necesario calcular hasta $n = \frac{N}{2} - 1$ si N es par y hasta el entero de $n = \frac{N}{2}$ si N es impar, ya que el resto de los cálculos son redundantes.

Ahora si consideramos los siguientes cambios de nomenclatura con fines de simplificación de notación tendremos que $W^{nk} = e^{\frac{-j2\pi mk}{N}}$, g (k) = m_k y $G(n) = G\left(\frac{n}{NT}\right)$, por lo que la ecuación anterior se puede escribir como sigue:

$$G(n) = \sum_{k=0}^{N-1} g(k) W^{nk} \qquad (20)$$

Para continuar con el tema presente es necesario recordar o aprender un poco sobre matrices, motivo por el cual realizaremos un pequeño apartado dedicado a este tema, así que sin más:

Matrices matemáticas

Una matriz es una tabla o arreglo rectangular de números. Los números en el arreglo se denominan elementos de la matriz.

Las líneas horizontales en una matriz se denominan filas y las líneas verticales se denominan columnas. A una matriz con m filas y n columnas se le denomina matriz m-por-n (escrito m×n), donde m y n son sus dimensiones. Las dimensiones de una matriz siempre se dan con el número de filas primero y el número de columnas después.

Un elemento de la matriz A que se encuentra en la fila i-ésima y la columna j-ésima se le llama entrada i,j o entrada (i,j)-iésima de A, que se escribe como $A_{i,j}$ o A[i,j].

Una matriz con una sola columna o una sola fila se denomina a menudo vector, y se interpreta como un elemento del espacio euclídeo. Una matriz 1 × n (una fila y n columnas) se denomina vector fila, y una matriz m × 1 (una columna y m filas) se denomina vector columna.

Un ejemplo de una matriz de 3 x 4 seria la siguiente:

$$A = \begin{bmatrix} 1 & 5 & 8 & 3 \\ 5 & 1 & 0 & 7 \\ 9 & 3 & 2 & 5 \end{bmatrix}$$

Con estos elementos matemáticos llamados matrices podemos llevar a cabo operaciones matemáticas, como por ejemplo:

Suma de Matrices

La suma de matrices únicamente se puede dar cuando ambas matrices tienen el mismo tamaño, es decir, mismo numero de filas y columnas en ambas matrices, sin importar que el número de columnas y filas no sea el mismo, ya que esta se realiza elemento a elemento y seria como se explica por medio de un ejemplo a continuación:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1+0 & 0+1 \\ 0+1 & 1+0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Multiplicación de una matriz por un escalar

Esta es sumamente simple, solo multiplicaremos cada elemento de la matriz por el escalar, como se muestra en el ejemplo siguiente:

$$2\begin{bmatrix} 3 & 2 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2*3 & 2*2 \\ 2*1 & 2*0 \end{bmatrix} = \begin{bmatrix} 6 & 4 \\ 2 & 0 \end{bmatrix}$$

Multiplicación de matrices

El producto de dos matrices se puede definir sólo si el número de columnas de la matriz izquierda es el mismo que el número de filas de la matriz derecha. A continuación veremos con un ejemplo esta operación:

$$\begin{bmatrix} 2 & 1 \\ 0 & 1 \\ 3 & 4 \end{bmatrix} x \begin{bmatrix} 0 & 1 & 2 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} (2*0)+(1*3) & (2*1)+(1*2) & (2*2)+(1*1) \\ (0*0)+(1*3) & (0*1)+(1*2) & (0*2)+(1*1) \\ (3*0)+(4*3) & (3*1)+(4*2) & (3*2)+(4*1) \end{bmatrix} = \begin{bmatrix} 3 & 4 & 5 \\ 3 & 2 & 1 \\ 12 & 11 & 10 \end{bmatrix}$$

Descomposición de matrices

Para no enviciar el tema del presente trabajo que no requiere ahondar de tal manera en el tema se presenta una breve presentación de la descomposición de matrices, ya que por desgracia y a pesar de que lo ocuparemos brevemente para la explicación siguiente, este tema es muy amplio y complejo por si solo, por tal motivo distraería del tema principal en comento.

En la disciplina del algebra lineal perteneciente a las matemáticas una descomposición de matrices es la factorización, que es en si la descomposición de un objeto como puede ser un numero, un polinomio o una matriz en un producto de otros objetos o factores mas simples, que al multiplicarse dan por resultado el original, por ejemplo, el numero 15 puede factorizarse o descomponerse en 3x5 y esto es igual a 15 de nuevo.

Regresando a las matrices, esta factorización de matrices dará como resultado dos o más matrices de alguna forma canónica que es una forma normal o estándar que sigue un teorema de clasificación, como podrían ser matrices triangulares inferiores o superiores, matrices ortogonales, diagonales, etc.

Existen una gran variedad de descomposiciones matriciales, y cada una de estas tiene aplicación en un tipo de problema en particular, por mencionar algunas existen las descomposición LU que es una descomposición en una matriz triangular inferior (Lower) y una matriz triangular superior (Upper), la descomposición por bloques, la Cholesky (que a su ves se divide en múltiples opciones), la QR, espectral, la Jordan y muchas mas que existen actualmente.

Al ser cada técnica muy diferente de las demás y relativamente complejas ya no nos ocuparemos mas en explicarlas a detalle, ya que esto nos distraería del tema de este trabajo, en el presente tan solo se mencionara que se llevo a cabo una descomposición, pero no entraremos en el rigor matemático al no ser necesario para comprender el tema.

Antes de este paréntesis habíamos llegado a la siguiente ecuación después de algunos cambios de nomenclatura:

$$G(n) = \sum_{k=0}^{N-1} g(k) W^{nk} \quad (21)$$

Esta expresión puede ser desarrollada y representada de forma matricial, para este ejemplo tomaremos un valor de N = 4 esta quedaría como:

$$\begin{bmatrix} G(0) \\ G(1) \\ G(2) \\ G(3) \end{bmatrix} = \begin{bmatrix} W^0 & W^0 & W^0 & W^0 \\ W^0 & W^1 & W^2 & W^3 \\ W^0 & W^2 & W^4 & W^6 \\ W^0 & W^3 & W^6 & W^9 \end{bmatrix} \begin{bmatrix} g(0) \\ g(1) \\ g(2) \\ g(3) \end{bmatrix} \quad (22)$$

Esta ecuación en forma matricial podemos ver que para solucionarla necesitaríamos N^2 multiplicaciones (16 para este ejemplo) y N(N-1) sumas (12 para este ejemplo), lo cual compararemos mas adelante contra la transformada rápida de Fourier.

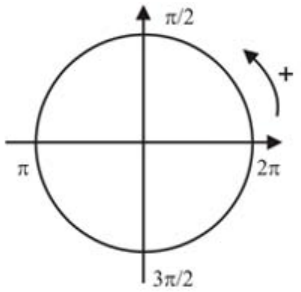
Ahora por definición sabemos que $W^0 = e^{\frac{-j2\pi 0}{4}} = e^0 = 1$, lo que nos permite reescribir la matriz de la siguiente manera:

$$\begin{bmatrix} G(0) \\ G(1) \\ G(2) \\ G(3) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & W^1 & W^2 & W^3 \\ 1 & W^2 & W^4 & W^6 \\ 1 & W^3 & W^6 & W^9 \end{bmatrix} \begin{bmatrix} g(0) \\ g(1) \\ g(2) \\ g(3) \end{bmatrix} \quad (23)$$

Desarrollando esta última ecuación como un producto de dos matrices por medio de una descomposición y tomando en cuenta que la función '*W*' es periódica y se repite cada 2π radianes, pero aquí de nuevo realizaremos un pequeño paréntesis para recordar o aprender que es un radian:

Radianes

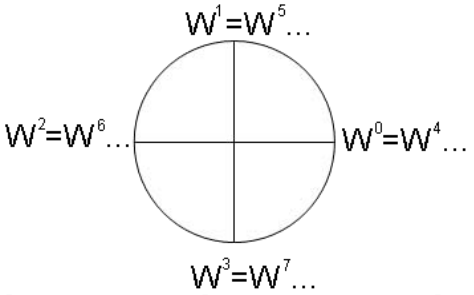
El radián se define como el ángulo que limita un arco de circunferencia cuya longitud es igual al radio de la circunferencia. Una definición más general, indica que el ángulo formado por dos radios de una circunferencia, medido en radianes, es igual a la longitud del arco formado sobre el radio, es decir, $\theta = s/r$, donde θ es el ángulo, *s* es la longitud del arco y *r* es el radio. Por tanto, el ángulo α de un círculo completo en radianes es:



$$\alpha = \frac{LongitudCircunferencia}{r} = \frac{2\pi r}{r} = 2\pi$$

De esta definición vemos que cuando se llega a 2π regresamos de nuevo al inicio de la circunferencia, por lo que también es fácil ver que las funciones senoidales y cosenoidales son igualmente periódicas, variando sus valores de cero a 1 y a -1, pero regresando a cero (para el seno), por lo que al poder expresar W^{nk} como una función de senos y cósenos veremos que es periódica, en la tabla siguiente podemos observar la matriz con los valores de $\frac{-j2\pi nk}{N}$ evaluada para la ecuación anterior:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \frac{1}{2}\pi & \pi & \frac{3}{2}\pi \\ 1 & \pi & 2\pi & 3\pi \\ 1 & \frac{3}{2}\pi & 3\pi & \frac{9}{2}\pi \end{bmatrix}$$



En esta tabla y con la explicación de los radianes vemos claramente que cada cuatro valores se repetirán los valores, por lo que $W^0=W^4$ y $W^2=-W^0$ como se aprecia en la anterior grafica, y teniendo esto en cuenta podemos ahora expresar la matriz descompuesta como:

$$\begin{bmatrix} G(0) \\ G(1) \\ G(2) \\ G(3) \end{bmatrix} = \begin{bmatrix} 1 & W^0 & 0 & 0 \\ 1 & W^2 & 0 & 0 \\ 0 & 0 & 1 & W^1 \\ 0 & 0 & 1 & W^3 \end{bmatrix} \begin{bmatrix} 1 & 0 & W^0 & 0 \\ 0 & 1 & 0 & W^0 \\ 1 & 0 & W^2 & 0 \\ 0 & 1 & 0 & W^2 \end{bmatrix} \begin{bmatrix} g(0) \\ g(1) \\ g(2) \\ g(3) \end{bmatrix} \qquad (24)$$

Ahora si resolvemos esto por medio de un resultado intermedio al aplicar la propiedad asociativa de la multiplicación de matrices podemos decir que:

$$\begin{bmatrix} g_1(0) \\ g_1(1) \\ g_1(2) \\ g_1(3) \end{bmatrix} = \begin{bmatrix} 1 & 0 & W^0 & 0 \\ 0 & 1 & 0 & W^0 \\ 1 & 0 & W^2 & 0 \\ 0 & 1 & 0 & W^2 \end{bmatrix} \begin{bmatrix} g(0) \\ g(1) \\ g(2) \\ g(3) \end{bmatrix} \qquad (25)$$

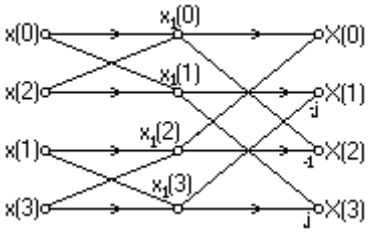
y por ultimo

$$\begin{bmatrix} g_2(0) \\ g_2(1) \\ g_2(2) \\ g_2(3) \end{bmatrix} = \begin{bmatrix} 1 & W^0 & 0 & 0 \\ 1 & W^2 & 0 & 0 \\ 0 & 0 & 1 & W^1 \\ 0 & 0 & 1 & W^3 \end{bmatrix} \begin{bmatrix} g_1(0) \\ g_1(1) \\ g_1(2) \\ g_1(3) \end{bmatrix} \qquad (26)$$

Cabe señalar que los resultados aparentan estar ligeramente desordenados, pero esto es por las operaciones matriciales llevadas a cabo, y el resultado final es el siguiente:

$$\begin{bmatrix} G(0) \\ G(1) \\ G(2) \\ G(3) \end{bmatrix} = \begin{bmatrix} g_2(0) \\ g_2(2) \\ g_2(1) \\ g_2(3) \end{bmatrix} \qquad (27)$$

De este resultado podemos obtener lo que se conoce como la 'mariposa' que se aprecia en la figura 58, en esta vemos que son necesarias solo ocho sumas (contra 12 anteriormente) y 4 multiplicaciones (contra 16 anteriormente), pero incluso las multiplicaciones pueden ser sustituidas fácilmente debido a que la función periódica '*W*' tiene valores muy simples que quedarían como a continuación se expresa.



$$W^0 = W^4 = W^8 = \ldots = 1$$
$$W^1 = W^5 = W^9 = \ldots = -j$$
$$W^2 = W^6 = W^{10} = \ldots = -1$$
$$W^3 = W^7 = W^{11} = \ldots = j$$

Figura 58: Mariposa para FFT con N = 4

Y a manera de un resumen grafico de la forma de trabajo de la transformada rápida de Fourier en la grafica vemos un análisis comparativo del número de operaciones de la transformada de Fourier contra la transformada rápida de Fourier.

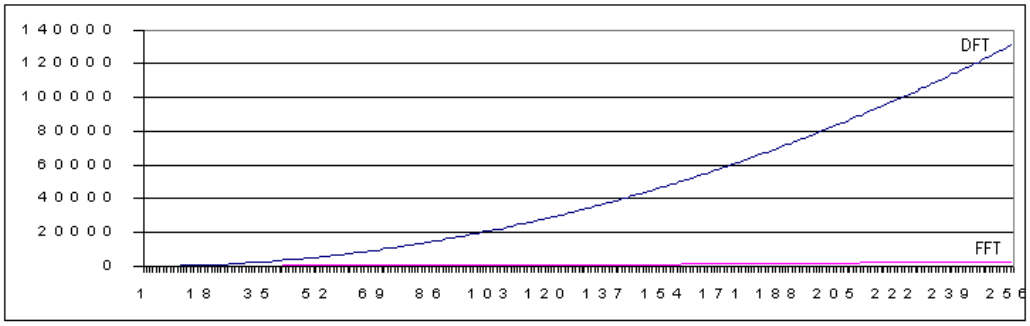


Figura 59: Numero de operaciones en la DFT y FFT cuando varía 'N'

Como podemos ver cuando 'N' es pequeño la diferencia en el numero de operaciones no es muy significativa, pero como es común en procesamiento de muestras de voz que se tome un valor de 'N' de 256, 512 o incluso 1024 se puede ver claramente que el numero de operaciones es mucho menor, resultando en un proceso mucho mas rápido y eficiente, la grafica va de N = 1 a N = 256 para ilustrar esto mismo.

Obtención del tono fundamental y las formantes

En el presente tema se pretende cubrir una forma de obtención de los mencionados parámetros desde un punto de vista simple y rápido para su implementación en un sistema informático, ya que esta técnica se implementara en el siguiente capitulo como una de las formas que se emplean en la identificación de locutores.

Esta técnica se basa en la forma de trabajo de los distintos laboratorios de acústica forense de habla hispana alrededor del mundo, por ejemplo en Colombia, España, México, y otros más. Este método se basa en los conocimientos del perito para establecer la identidad de un individuo por medio de su voz con la ayuda de espectrógrafos computarizados de alta calidad, mismos que son prácticamente únicos en el mercado y de un considerable costo, perteneciendo casi todos a la marca Kay Elemetrics, y en algunos casos Medav (en uso actualmente en Alemania).

Espectrógrafo Kay Elemetrics modelo 5500-1 en las oficinas de la Policía Científica Española

Dichos espectrógrafos tienen primordialmente dos usos, aplicaciones medicas por parte de foniatras en la detección y cura de padecimientos relacionados con la voz y aplicaciones forenses, primordialmente identificación de locutores.

La técnica es simple, pero al mismo tiempo de considerable dificultad para la obtención de un buen resultado, ya que requiere extensos conocimientos de diversas áreas del conocimiento humano, como son la foniatría, la ingeniería, la medicina, la psicología, cirujanos dentistas y algunas otras mas que también pueden aportar elementos de interés a la técnica, todo esto debido a que la base es un estudio fónico aplicado con la ayuda de complejos aparatos, pero teniéndose que tomar en cuenta factores tales como trastornos mentales (algún trastorno de personalidad o enfermedad mental), estados anímicos, patologías temporales o permanentes (enfermedades), variaciones en las cavidades resonantes (perdida de dientes u otros similares) y demás factores que pueden dar pauta a una variación en los patrones de la voz.

De lo ya expuesto se hace obvia la pregunta ¿puede un programa de computadora hacer todo esto?, y la respuesta también es obvia, “no lo puede hacer”, lo que pretenden los diversos programas de computadora actualmente en desarrollo alrededor del mundo es una herramienta mas que auxilie al experto en la metería a llevar a cabo sus funciones de una manera mas rápida y certera, ya que recordemos que de un resultado de esta índole puede depender la situación jurídica de uno o mas sujetos.

La técnica que pretende implementarse con el uso de estos programas complementa a la llamada técnica mixta, esto es que la computadora nos da una lista de candidatos o simplemente nos da un porcentaje de probabilidades de que se trate de un mismo sujeto y basándose en estos resultados el perito realizara el resto del estudio de forma tradicional.

Pero por ejemplo si tomamos un caso hipotético de que se pida la comparativa de voz de diez probables responsables contra una cinta problema (recibe este nombre la cinta que contiene la grabación de la voz del delincuente, por ejemplo chantajista, secuestrador, etc.), si tomamos en cuenta que cada sujeto estudiado puede llevar de tres a siete días seria un periodo de respuesta de hasta mas de dos meses, y con la ayuda de programas de computadora como el que se pretende desarrollar esto se puede agilizar tremendamente, ya que la computadora estudia los diez sujetos y da una lista con el mas probable primero, de esta manera el perito analizara la lista en el orden sugerido pudiendo encontrar al culpable al primer intento, respondiendo en un lapso de tres a siete días.

El estudio en si es la búsqueda, identificación y comparación de estructuras formanticas dentro de diversas muestras de voz, pero en la actualidad con los equipos mencionados se hace esto dividiendo la grabación en pociones de aproximadamente dos segundos de audio por cada pantalla de resultados, y si tomamos en cuenta el hecho de que de manera general una sola palabra tiene una duración de un segundo aproximadamente tenemos el hecho de que se realizara el estudio de dos palabras a la vez, por lo que dependiendo de la duración de las cintas en estudio puede llevar días o meses incluso, ya que no se escucha una sola vez, sino en múltiples ocasiones para tratar de identificar peculiaridades del habla, vicios del habla, acentos regionales, modismos o cualquier otro elemento que nos pueda auxiliar.

Esta misma acción se repetirá para la cinta testigo (recibe este nombre la cinta con la grabación de los sujetos presuntamente responsables y en los que un Ministerio Publico o autoridad competente da Fe de la identidad de dicho sujeto), pero además se deberán de buscar estructuras lo mas similares posibles a las que pretendemos comparar, ya que requiere de mucha mas experiencia y conocimientos el poder hacer estudios comparativos entre palabras distintas (estudio independiente del texto), por lo que se ve el porque del elevado tiempo de respuesta de estos laboratorios.

Se considera pertinente también mencionar que las cintas problema y testigo pueden a la vez recibir los nombres de cinta dubitada y cinta indubitada respectivamente, refiriéndose a que en la primera existe la duda de la identidad de los sujetos que participan en dicha grabación y en la segunda no existe duda de la identidad del sujeto, ya que esta grabación se obtiene en presencia de una autoridad pertinente que da Fe de la veracidad de la grabación y de la identidad del sujeto.

A continuación se incluye una tabla de los tiempos aproximados de respuesta, el numero de peritos, el numero de casos y el porcentaje de casos que no se puede determinar la identidad por la mala calidad de las grabaciones o la escasa cantidad de material en algunos de los países Europeos que pertenecen a la ENFSI, esta taba se deriva de la reunión llevada a cabo en el año de 1999.

	Numero de Peritos	Numero de Casos al Año	Porcentaje de Rechazo	Tiempo de Respuesta
Bélgica	2	87	25 %	3 a 5 días
Finlandia	1	20	10 %	2 a 3 días
Francia (Gendarmes)	2	15	50 %	3 días
Francia (PTYC)	2	50	10 %	21 días
Alemania	5	108	50 %	3 a 5 días
Italia (Carabinieri)	5	35	6 %	7 días
Italia (policía)	5	50	No se sabe	7 días
Lituania	9	180	15 %	8 horas
Holanda	2	30	10 %	40 horas
Polonia	4	50	20 %	7 días
Rusia	112	+2000	7 %	5 días
Eslovaquia	2	30	40 %	8 horas
Eslovenia	1	10	1 %	14 días
España	10	85	50 %	16 horas
Suiza	2	35	40 %	25 horas
Suecia	1	5	80 %	No se sabe
Turquía	2	60	84 %	10 días

Tabla 7: Tiempos de intervencion

Lo que se pretende en el presente trabajo es que la computadora pueda estudiar grabaciones de audio, establecer el tono fundamental y buscar las primeras dos formantes, que son las mas estables y de mas fácil obtención, pero como tenemos mas de veinte formantes se ve que este estudio no es lo suficientemente confiable como para relegar la decisión por completo a la computadora, sin mencionar que la computadora no puede analizar estados anímicos o alguna condición física que dificulte el estudio. Sin embargo puede analizar fríamente y de manera muy rápida las estructuras mencionadas dando una buena aproximación o guía para posteriores estudios.

Para entender mejor la acción a tomar estudiemos la figura 60 que es un espectrograma de una porción de audio, en especifico de un sonido gutural como una vocal, en esta se ve que la señal presenta máximos locales en las regiones que deben de contener cada uno de los elementos que se desean estudiar, mismos que están delimitados en esta misma figura.

Recordemos además que las formantes por definición misma son armónicas del tono fundamental, esto es, múltiplos del tono fundamental o la multiplicación de este por un numero entero, y por ultimo vemos también que en lo que seria el máximo de la tercera formante se a colocado una línea punteada horizontal la cual tiene la leyenda “3 dB” la cual simboliza de forma cualitativa que la porción comprendida de la cima del máximo a la línea punteada se encuentra la porción que nos interesa de dicha formante, ya que debajo de tres decibeles ya no será de interés para su estudio por su poca aportación.

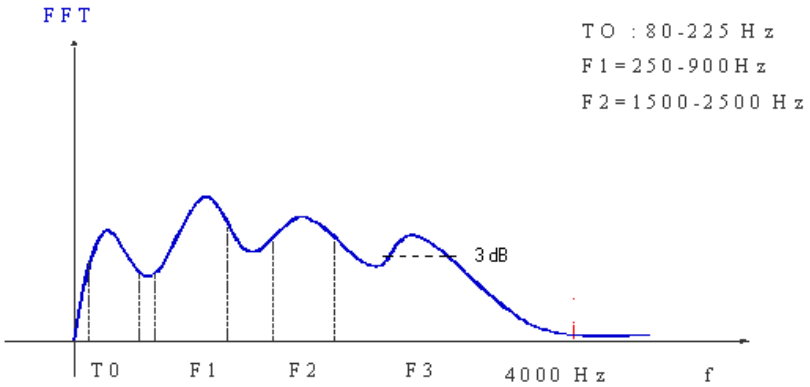


Figura 60: Espectro de sonido gutural

Viendo esta figura y apoyándonos en que tenemos ya la información en el dominio de la frecuencia podemos imaginar fácilmente un algoritmo que ira recorriendo el espectro en los rangos deseados para buscar dichos máximos y así poder identificar la estructura que estamos buscando, siendo el primer parámetro a obtener el tono fundamental, el cual simplemente será el máximo local comprendido en el rango especificado para T_0.

Una vez que tenemos localizado el tono fundamental deberemos de crear una lista de los posibles armónicos que pueden existir de este, ya que como una medida extra implementaremos que para las formantes además de que sean el máximo relativo en su rango cumplan con la definición misma de formante y esto es que sean un armónico de T_0, ya que recordemos que por cuestiones de ruido y malas grabaciones se podrían presentar estructuras que la computadora pueda confundir con una formante.

Por ultimo para la obtención de las formantes también pondremos la condicionante de que una vez localizada la formante en cuestión solo se tomara esta en consideración mientras se encuentre en el rango de amplitud comprendido de la cima de la formante y hasta 3 dB por debajo de esta, esto en el afán de tener solo la estructura formantica en su mas estricta definición y no considerar datos poco relevantes.

Como hemos planteado en si la obtención del tono fundamental y de las dos primeras formantes no es en si muy compleja, pero vale la pena repetir que la computadora no tomara en cuenta factores ajenos a la muestra de voz, como enfermedades u otros factores, pero al mismo tiempo esto hace que el estudio sea sumamente objetivo, por lo que se puede apreciar el valor de esta herramienta al perito en acústica forense.

Para cerrar el tema se presenta un resumen de cómo se localiza cada uno de los parámetros deseados y su posible utilidad en el campo de la acústica forense.

Obtención del tono fundamental

Una vez que se tiene la señal en el dominio de la frecuencia, ya sea por medio de la transformada de Fourier o de la transformada rápida de Fourier se establecerá un rango en el cual se deberá de encontrar el tono fundamental, en este caso tomaremos un rango de 60 a 250 Hz para su ubicación.

La señal de audio será dividida en ventanas de duración definida por el usuario, pero tomando en cuenta de que a partir de este tamaño de ventana variara la resolución en frecuencia del estudio, se aplicara una ventana de Hamming y se solaparan las ventanas

en el porcentaje seleccionado por el usuario, dejándonos ya con el camino listo para localizar he identificar el tono fundamental.

Se tomara cada una de las mencionadas ventanas y se buscara el máximo absoluto en el rango de frecuencia seleccionado, este será conservado para cada una de las ventanas que conforma la muestra de voz en estudio y promediado para obtener el tono fundamental de la persona, ya que este debe de presentar muy poca variación a lo largo de una o mas palabras, esto por su naturaleza misma, que recordando de la ficología humana, el tono fundamental estará dado por el sistema de fuelle que proporciona los pulsos y la presión necesaria para la producción del habla.

Obtención de las primeras dos formantes

Basándonos en los resultados previos del tono fundamental se construye una tabla de los posibles armónicos de este, ya que como hemos mencionado cada formante debe ser por definición un armónico de T_0.

Una vez construida esta tabla se localizara el máximo absoluto por cada ventana en los rangos de 250 a 900 Hz para F_1 y de 1500 a 2500 para F_2, cumpliéndose además que este máximo sea un armónico de T_0, dichos valores serán guardados como estructuras formanticas, en una posterior revisión deberá de cumplir además con la condicionante de que a partir del máximo de la formante misma y solo mientras no rebase una caída superior a 3 dB será considerada una formante, por debajo de los 3 dB no será de interés por su poca aportación o porque puede confundirse con otra estructura, recordemos que las primeras dos formantes solo estarán presentes en los sonidos vocálicos, por lo que seria un error si estas se localizaran en toda la muestra de voz.

Sumario

A forma de auxiliar al lector con la localización de las estructuras formanticas y de un mejor entendimiento de los resultados obtenidos en el presente tema se presenta una tabla que contiene a grandes rasgos las características que pueden presentar las distintas vocales en sus dos primeras formantes, además de explicar a grandes rasgos como se deberán de interpretar de manera manual los resultados de un espectrograma.

Debido a que el tono fundamental se genera de manera involuntaria y es de difícil variación por periodos prolongados se ve que este será en general sumamente estable y de muy poca variación cuando se trata de diversas tomas de voz de un mismo individuo, por esto mismo se maneja de manera generalizada un valor máximo de diferenciación de aproximadamente 30 Hz para poder considerar que se trata de un mismo sujeto, siempre y cuando no exista razón alguna para justificar dicha variación.

De manera similar existe un estándar para la diferenciación de las estructuras formanticas y este es de 50 Hz, esto se debe a que las formantes pueden variar fuertemente dependiendo de estados anímicos, del tipo de entonación y muchos factores mas, por lo que es obligado tomar un mayor rango de variación.

La siguiente tabla muestra los valores de las dos primeras formantes para las cinco vocales tónicas en posición fonética normal para un sujeto masculino:

Realización de	F_1	F_2
[i] en pípa	200 Hz	2100 Hz
[e] en pépa	324 Hz	1950 Hz
[a] en pápa	600 Hz	1300 Hz
[o] en pópa	324 Hz	729 Hz
[u] en púpa	200 Hz	620 Hz

Tabla 8: Valores comunes en frecuencia

Parámetros LPC

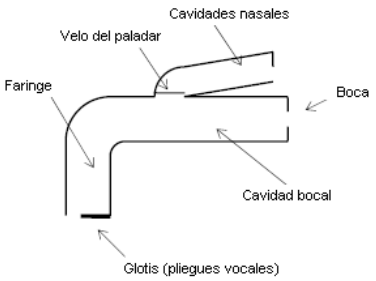
En el presente tema se pretende cubrir de manera básica una forma de obtención de los parámetros LPC de una señal de audio desde un punto de vista simple y rápido para su implementación en un sistema informático, ya que esta técnica se implementara en el siguiente capitulo como una de las formas que se emplean en la identificación de locutores.

Una de las técnicas más poderosas para el procesamiento de voz es el análisis de predicción lineal (LPC Linear Predictive Coding), esta se ha convertido en la técnica predominante para la estimación de los parámetros básicos como tono fundamental, formantes, espectro, etc.

Las ventajas de este método radican tanto en su habilidad de dar estimados sumamente precisos de los parámetros de voz como de su relativa alta velocidad de procesamiento en sistemas computacionales.

Para comenzar LPC empieza con la presunción de que una señal de voz es producida por un zumbador al final de un tubo para sonidos vocálicos, con la ocasional adición de sonidos silbantes y plosivos, esto puede parecer muy burdo, pero en realidad es un modelo cercano a la realidad de la producción vocálica.

La glotis produce los zumbidos y esta caracterizado por su intensidad y frecuencia, el tracto vocal conformado básicamente por la garganta y la boca, forman el tubo y se caracteriza por su resonancia, el efecto de estas resonancias es lo que conocemos como formantes, y los silbantes y plosivos son generados por la acción de la lengua, labios y garganta.

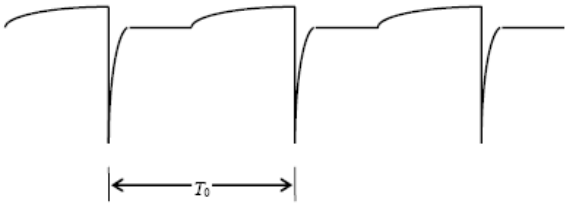


La técnica LPC analiza la señal de voz al estimar las formantes, removiendo su efecto de la señal de voz, y estimando la intensidad y frecuencia del zumbido restante, el proceso durante el cual se remueven las formantes es llamado filtrado inverso y la señal resultante después de la extracción es llamado residuo.

Los números que describen la intensidad y frecuencia del zumbido, las formantes y la señal residuo pueden ser almacenados para su transmisión y al invertir este proceso se puede reconstruir la señal de voz original.

Ahora, por la naturaleza cambiante del sonido este proceso se repite en periodos cortos de señales de voz, los cuales reciben el nombre de cuadros o ventanas, y por lo regular de 30 a 50 ventanas resultan en una señal de buena calidad.

Recordatorio;
Recordemos que las vocales y las consonantes sonoras tienen intervención de los pliegues vocales, en estas los pliegues vocales producen un tren de impulsos como se muestra a continuación.



En los sonidos sordos que no hay intervención de los pliegues vocales se puede decir que la excitación es únicamente un 'hiss' o ruido blanco, que se apreciaría como algo similar a esto.

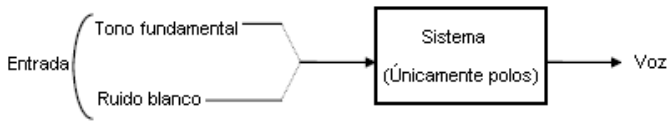
La idea básica de este método es que una muestra de voz puede ser aproximada como una combinación lineal de muestras anteriores, esto es logrado al minimizar la suma del cuadrado de las diferencias (dentro de un periodo finito) entre la muestra actual de voz y las que se obtuvieron por la predicción lineal, por lo que se puede obtener un juego único de coeficientes de predicción.

Cuando se aplica la técnica de LPC al procesamiento de voz el término de predicción lineal se refiere a una variedad de formulaciones equivalentes con respecto al problema básico de modelado de la forma de onda de voz (digitalización de la voz), entre las distintas aproximaciones o formulaciones equivalentes para resolver este problema tenemos:

- El método de la covarianza
- La formulación por auto-correlación
- El método de la celosía
- La formulación de filtro inverso
- La formulación por estimación de espectro
- La formulación por máximo parecido
- La formulación por producto interno

El modelo usado en general por la técnica LPC en su forma más básica se puede apreciar en la siguiente ilustración y su representación matemática en la forma más común por el algoritmo Levinson-Durbin esta dado por la ecuación:

$$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i) \qquad (28)$$



Donde $\hat{x}(n)$ es el valor de la señal predicha, $x(n-i)$ son los valores anteriores medidos y a_i son los coeficientes de predicción y el error generado por la predicción esta dado por:

$$e(n) = x(n) - \hat{x}(n) \qquad (29)$$

Y por ultimo la suma del error cuadrático que deseamos minimizar esta dada por:

$$E = \sum_n e^2(n) = \sum_n \left(x(n) - \sum_{i=1}^{p} a_i x(n-i) \right)^2 \qquad (30)$$

Estas ecuaciones como ya se menciono pueden ser resueltas de múltiples formas, mas sin embargo por su complejidad no es tema de este trabajo extenderse en este tema, sino mas bien comprender su importancia y aportación a la identificación de locutores, por este motivo no entraremos en el rigor matemático de los métodos para obtener los ya mencionados coeficientes.

Obtención de los coeficientes

Para los que desean profundizar más sus conocimientos sobre las metodologías para resolver las ecuaciones anteriores les recomendamos la lectura de algún libro sobre codificación de señales de audio, pero para aquellos que ya estamos mas interesados en la identificación de un locutor que en la solución matemática vamos a ofrecer una solución rápida de estas ecuaciones que puede ser resuelta de manera rápida por una computadora por medio del algoritmo Levinson-Durbin, y este nos plantea lo siguiente:

De primera instancia vamos a requerir los coeficientes de auto-correlación R[l], y una vez que tenemos estos procedemos de la siguiente manera:

0) Iniciamos con l=0, lo cual nos lleva a que J_0=R[0]

Ahora para l=1, 2, 3…M donde M será el orden LPC deseado

1) Calcular el RC correspondiente (RC es un resultado intermedio, y finalmente será una constante que nos permite resolver las ecuaciones que se formulan de forma matricial)

$$k_l = \frac{1}{J_{l-1}} \left(R[l] + \sum_{i=1}^{l-1} a_i^{(l-1)} R[l-i] \right) \qquad (31)$$

Una vez que tenemos esto calculamos los coeficientes LPC

$$a_l^{(l)} = -k_l \qquad (32)$$
$$a_i^{(l)} = a_i^{(l-1)} - k_l a_{l-i}^{(l-1)} \text{ para i=1, 2…l-1}$$

Si l=M se acaba el proceso

2) Calcular el error mínimo cuadrático de predicción

$$J_l = J_{l-1}(1-k_l^2) \qquad (33)$$

Sumar 1 a l y regresar al paso 1

De estos pasos podemos ver que aunque es algo complicada la obtención de los coeficientes LPC realmente ya con las soluciones que actualmente existen no es muy complicado implementar una de estas en un sistema informático, y que comparativamente con otros métodos como la transformada rápida de Fourier es computacionalmente económica.

Ahora la pregunta de todos me supongo que serán dos, como obtengo los valores R[l] de la auto-correlación y donde intervienen los datos de la voz en estudio, y la respuesta es la misma, estos valores provienen de los datos, por lo que incluso vemos que si el orden M no cambia pues tampoco cambiaran las ecuaciones que las determinan, solo será necesario sustituir los correspondientes valores de R[l] para obtener los resultados finales, por lo que seguimos viendo que esta puede ser una buena metodología a implementar.

Obtención de los coeficientes de auto-correlación

Esto es un dato realmente fácil de obtener, ya que la auto-correlación esta dada por la siguiente ecuación:

$$R[l] = \sum_n x_n x_{n-l} \qquad (34)$$

En esta ecuación x es el conjugado complejo, pero para una función real como lo es el sonido $x = x$, y x es una función discreta que representa los datos en una ventana de audio.

Interpretación de los coeficientes LPC

Los coeficientes LPC obtenidos anteriormente describen el comportamiento de las formantes, como ya lo vimos anteriormente, este modelo es útil para que con ayuda del residuo que se obtiene al procesar una señal de audio nos permite reconstruir la señal de voz original, mas sin embargo también nos permite observar de una manera mucho mas clara la estructura formantica, como lo podemos apreciar en la grafica 61.

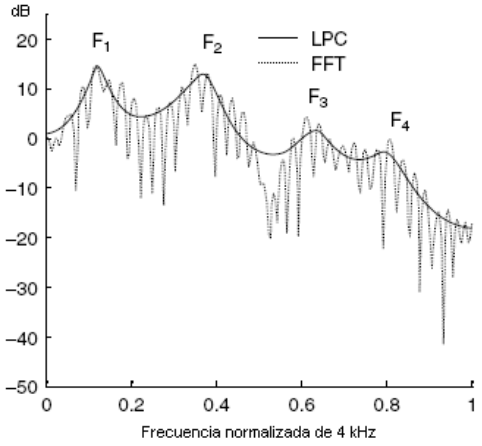


Figura 61: Envolvente

Aquí podemos ver la grafica de los coeficientes LPC y un espectro obtenido por una FFT o transformada rápida de Fourier, como podemos apreciar ambas muestran el mismo comportamiento en periodos cortos de tiempo, y aun mas podemos decir que la grafica de LPC "cubre" al espectro, razón por la cual en múltiples ocasiones recibe el nombre de *envolvente*, pero con mucha menos información y de una forma por decir "mas simple".

Este comportamiento es precisamente el que nos permite decir que cada fonema puede ser en un momento dado caracterizado por un modelo LPC, que recordemos desde un inicio fue uno de nuestros planteamientos básicos, es un grupo de objetos bien definidos, con características propias y que si llegaran a variar serian un fonema distinto.

De esto podemos ver fácilmente que para un programa de reconocimiento de habla ya nos lleva a mas de la mitas, ya que al tener modelados todos los fonemas y sus alófonos podemos de una manera relativamente simple y rápida reconocer un fonema determinado, y por ende a la suma de un grupo de fonemas, que después de todo es una palabra.

Regresando a nuestro tema de interés vemos que esto también puede ser aplicado al reconocimiento de un locutor, ya que al modelar todos o los mas fonemas posibles podemos de alguna manera caracterizar o modelar a un sujeto determinado, que es el objetivo que perseguimos, ya que recordando un poco como funcionan los programas de dictado vemos que estos requieren un periodo de entrenamiento, durante el cual el programa refinara sus modelos a los del usuario que usa la maquina, y si un usuario ajeno a la misma intentara usar el programa este tendría malos o nulos resultados por la falta de entrenamiento o modelado de su usuario actual, lo cual nos lleva a confirmar la idea de que cada locutor es distinto.

Distancia vectorial

En el presente tema se pretende dar al lector una idea general de las metodologías que se utilizan en la determinación de la distancia entre dos vectores, cualesquiera que estos sean o lo que es lo mismo y para fines del presente trabajo determinar la similitud o coincidencia entre dos o mas grupos de datos, de manera que se pueda discriminar si se trata del mismo sujeto o no por medio de muestras de su voz.

Medición de distancias

Dados dos vectores '*x*' y '*y*' en un espacio multidimencional muchas veces nos interesara ver que tan "cercano" es uno del otro, por ahora solo diremos que estos vectores pertenecen al espacio cartesiano de 'N' dimisiones con elementos reales que se denotara $\Re^N$ y una métrica $d(\cdot,\cdot)$ de $\Re^N$ será una función con un valor real y que cumple con las siguientes tres propiedades para todo '*x*' y '*y*' $\in \Re^N$.

1.- $d(x, y) \geq 0$
2.- $d(x, y) = 0$ si y solo si x = y
3.- $d(x, y) \leq d(x, z) + d(z, y)$

La primera nos dice que no existen distancias negativas, la segunda nos dice que la distancia nula solo existe al medir dos vectores idénticos, y la tercera nos dice que la

distancia mas corta será siempre la línea recta y si el vector 'z' es colineal (coincide con la línea recta entre '*x*' y '*y*') con los puntos '*x*' y '*y*' será igual la suma de las distancias que la original pero nunca menor.

Cualquier función que cumpla las tres condiciones anteriores es una métrica legitima de un espacio vectorial, ahora, existen muchas clases de medidas, pero en aplicaciones de análisis de voz se utilizan casos particulares de la métrica Minkowski o similares, esta métrica se define de la siguiente manera, si tomamos x_k como el elemento kaésimo del vector 'x' la métrica de orden 's' o l_s entre los vectores 'x' y 'y' es:

$$d_s(x,y) = \sqrt[s]{\sum_{k=1}^{N} |x_k - y_k|^s} \qquad (35)$$

Algunos casos particulares son:

1.- La l_1 o métrica de bloques de ciudad

$$d_1(x,y) = \sum_{k=1}^{N} |x_k - y_k| \qquad (36)$$

2.- La l_2 o métrica Euclidiana

$$d_s(x,y) = \sqrt{\sum_{k=1}^{N} |x_k - y_k|^2} \qquad (37)$$

3.- La l_∞ o métrica Chebyshev ($s \to \infty$)

$$d_\infty(x,y) = \max_k |x_k - y_k| \qquad (38)$$

La métrica Euclidea

De las mediciones de distancia en $\Re^N$ la Euclidea es probablemente la más utilizada en problemas de ingeniería, la razón para esta popularidad reside en que encaja perfectamente en nuestra noción física de distancia.

Uno de los puntos que deberemos de cuidar es que los ejes de referencia por lo que representamos a nuestros vectores cumplan con ser orto-normales (que la información de cada eje no tenga ninguna componente en otro eje distinto a si mismo, por ejemplo para dos dimensiones que formen un ángulo recto entre ambos ejes), ya que si los vectores que queremos comparar no cumplen con esta condición deberán de ser primero transformados a ejes de referencia orto-normales para que esta distancia tenga un significado coherente.

En lo que respecta al presente trabajo vemos que se cumple la condición de ejes de referencia orto-normales debido a que por ejemplo uno de los ejes representa al tiempo y otro puede representar una magnitud determinada como energía, cruces por cero, etc. o en su defecto en el dominio de Fourier un eje representa frecuencia y otro representa magnitud o tiempo, no presentando ninguna dependencia entre si.

Al mismo tiempo también es importante mencionar que a pesar de lo ya expuesto la relación de distancias que exista entre dos vectores dados con respecto a su métrica no variara a pesar de no cumplir con la condición de orto-normalidad o incluso después de la transformación, este es solo un comentario para que dicha distancia tenga un

significado físico mas fácil de comprender. Una transformación de esta naturaleza solo remueve la información redundante que se pudiera presentar.

Esta transformación puede tener un mayor significado cuando habamos ya de un ejemplo en concreto, ya que además de importarnos la correlación entre los parámetros también tenemos que ver el efecto de la escala, por ejemplo, si de una señal de audio se extraen los parámetros de cruces por cero y energía podremos ver claramente que la energía será numéricamente mucho mayor que los cruces por cero, por lo que al juntar estas características en un vector de parámetros este factor de escala puede afectar nuestra percepción de los resultados.

Del mismo ejemplo anterior podríamos ver que además de la alta diferencia de escala podría presentarse alguna correlación, como que al bajar los cruces por cero sube la energía, por lo que puede ser recomendable una transformación, pero sobre todo un proceso de escalamiento, debido a que por la naturaleza pequeña de los cruces por cero estos perderán importancia o peso contra un parámetro como la energía, recordemos además que al escalar a valores pequeños se disminuye la varianza, por lo que es recomendable aplicar un factor de escalamiento para disminuir la energía y obtener mejores resultados.

Una vez cubiertos todos estos puntos que nos dan la base para comprender la puesta en marcha de un sistema que auxilie a un perito en la materia a evaluar muestras de voz para determinar si se trata o no de un mismo locutor pasaremos en si al sistema, así que sin mas preámbulo pasaremos al segundo capitulo.

Capitulo II
"Aplicación del método científico"

Recordemos rápidamente los pasos que nos marca el método científico de investigación, ya que son los que a continuación desarrollaremos en miras de lograr una solución a nuestro problema de identificar a uno o mas sujetos por medio de sus registro vocales.

1. Detección, delimitación y planteamiento de un problema
2. Antecedentes y estado actual del problema
3. Justificación
4. Propósitos y objetivos
5. Hipótesis
6. Diseño de la investigación
7. Resultados
8. Discusión
9. Conclusiones
10. Problemas pendientes
11. Resumen y bibliografía

De esta metodología vemos de manera clara que de primera instancia debemos de detectar, delimitar y plantear el problema en mano, así que sin mas empecemos con el primer punto para tener una mejor comprensión del problema que tenemos ante nosotros.

Detección, delimitación y planteamiento del problema

Como fácilmente podemos deducir en delitos tales como la privación ilegal de la libertad, el secuestro, el secuestro express, el secuestro virtual, el chantaje, el fraude telefónico y otros delitos, la principal evidencia puede ser una grabación con una o mas conversaciones telefónicas, por lo que se ve la importancia de estudios de esta naturaleza para no dejar impunes este tipo de delitos tan lacerantes a la sociedad.

De este hecho ahora sabemos que es de suma importancia social es estar en la posibilidad de poder lograr la identificación de uno o mas sujetos por medio de sus voces bajo la hipótesis que veremos mas a detalle en el desarrollo de esta tesis de que nuestra voz es "*única e identificable*".

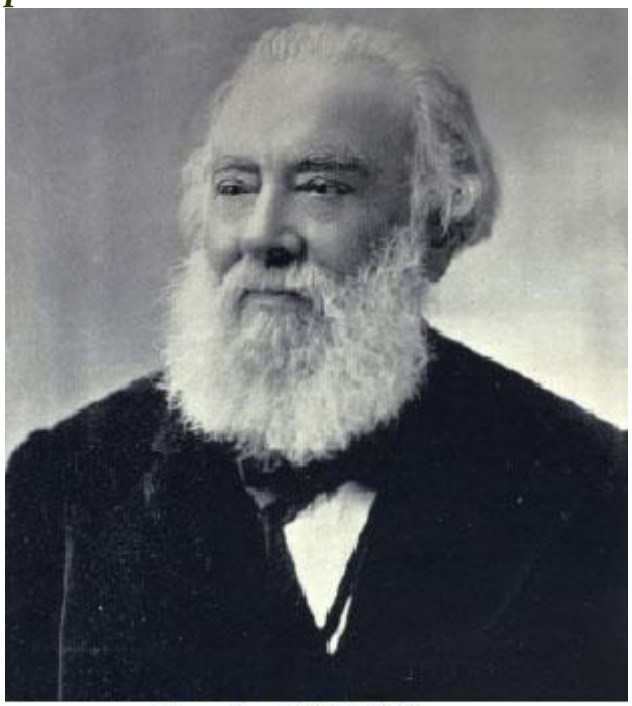
Es importante delimitar de primera instancia metas bien marcadas para esta investigación, ya que querer saltar de lleno al problema nos puede llevar a grandes problemas donde no podremos fácilmente identificar el motivo de la falla o comportamiento errático, pudiendo ser este el canal de comunicación utilizado para las grabaciones, el tono de voz usado, la calidad de grabación o simplemente el sistema, por lo que el presente trabajo utilizara un *corpora* (plural de corpus que en Latín significa cuerpo) de pruebas recabadas en ambiente controlado, con un canal de grabación directo a equipo de computo con un micrófono de buena calidad y con un tono de voz plano tipo lectura.

Es también importante mencionar que este tipo de sistemas pueden ir aumentando su utilidad a costo de complejidad al hacerse independientes del texto y en algunas ocasiones incluso independientes del idioma, mas sin embargo este tipo de sistemas se encuentran actualmente en desarrollo y prueba, no teniéndose hasta la fecha ninguno de estos funcionando a nivel nacional, por lo que seria pretencioso tratar de saltar estos dos importantes pasos, por lo que esta tesis aborda un sistema dependiente del texto y por consecuencia del idioma.

De lo ya expuesto ahora sabemos cual es el problema y hemos fijado delimitaciones bien marcadas para el alcance de la presente investigación, por lo que ahora si podemos plantear el problema a atacar y que será identificar a uno o mas sujetos por medio de sus registros vocales, dichos registros o corpora será obtenido en condiciones de laboratorio con un micrófono de buena calidad como se describe mas adelante y en base a repeticiones de palabras trisilábicas o mayores, y este trabajo sentara precedente para continuar a manera futura con la investigación y desarrollo, ya que por lo general en el ámbito forense trabajaremos con características adversas y con señales de pobre calidad, por lo que se requiere tener una buena base para mejorar las probabilidades de una correcta identificación o descarte.

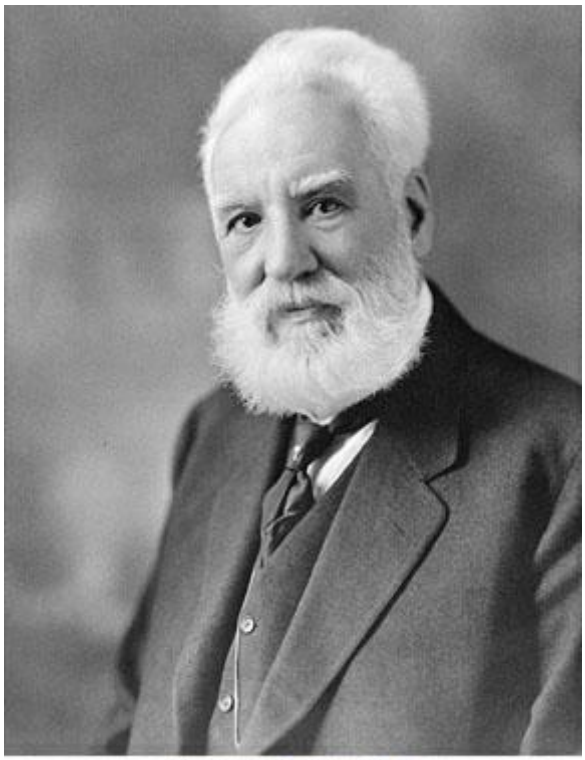
Antecedentes y estado actual del problema

La historia de la identificación de locutores por medio de su voz tiene sus orígenes hace poco mas de 100 años y apunta hacia Alexander Melville Bell, el desarrollo una representación visual de las palabras habladas. Esta representación visual mostraba mucha mas información sobre la pronunciación que lo que jamás podría mostrar un diccionario de pronunciación sobre la misma.



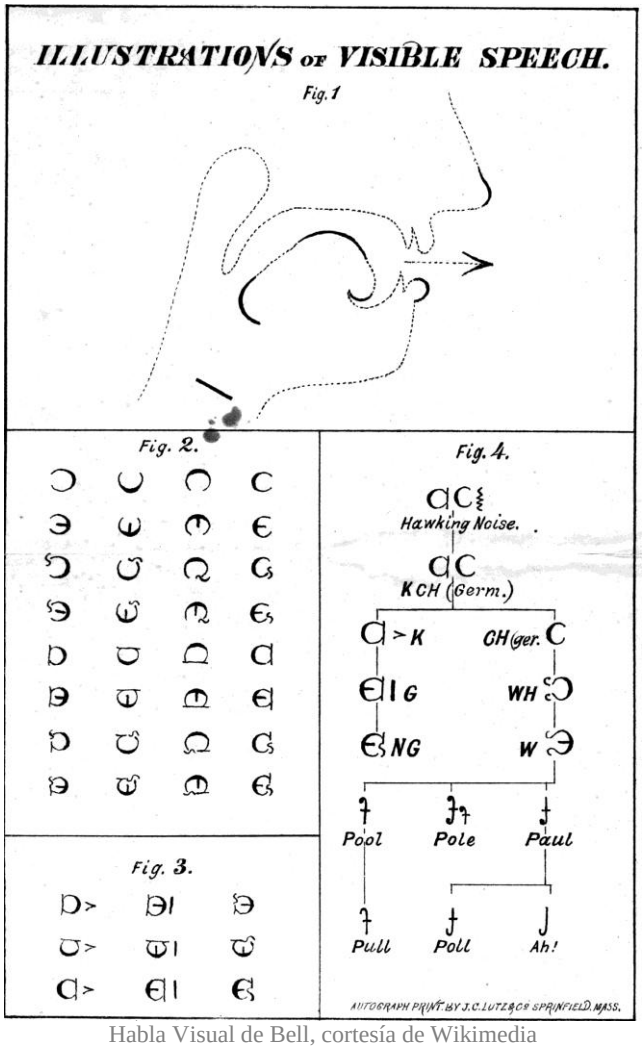
Alexander Melville Bell

Esta representación visual mostraba sutiles diferencias en como diferentes personas pronunciaban una misma palabra. Este tipo de representación visual para análisis desarrollada por Bell recibió el nombre de "Habla visual". Su método de codificación para la gran variedad de sonidos era por símbolos escritos a mano y era independiente del idioma, esta codificación mostraba de manera rápida las diferencias en como una palabra podía ser pronunciada por dos o mas personas diferentes, este sistema fue utilizado por Bell y por su hijo Alexander Graham Bell para ayudar a las personas sordas a que aprendieran a hablar.



Alejander Graham Bell

A continuación se puede apreciar una ilustración que muestra el "Habla Visual" de Bell, cabe destacar que esta imagen al igual que algunas otras utilizadas en esta tesis son cortesía de Wikimedia Common, que es un archivo en Internet de imágenes libres de derechos, ya que por su antigüedad o derechos de autor se encuentran ya libres para su uso sin permisos o licencias.



Habla Visual de Bell, cortesía de Wikimedia

Fue a principios de los cuarentas nace una nueva metodología para el análisis de los sonidos de la voz, Potter, Kopp y Green trabajando para Laboratorios Bell en Murria Hill en Nueva Jersey comienzan a trabajar en el desarrollo de un espectrógrafo para poder observar las características de los sonidos. Esta maquina de manera automática analizaba las ondas del sonido produciendo un esquema de frecuencia, intensidad y tiempo, esta investigación se intensifico durante la Segunda Guerra Mundial cuando se sugirió el uso de esta tecnología para identificar la voz de los enemigos por medio de la radio, pero la guerra termino sin que la técnica pudiese ser perfeccionada.

En 1947 Potter, Kopp y Green publicaron su trabajo en un libro llamado “Visible Speech” y contiene un buen estudio de espectrogramas de voz y estaba destinado a interpretar visiblemente patrones lingüísticos, de manera similar al “Habla Visual” de Bell y su código estos científicos codificaron los sonidos del habla con el uso de patrones producidos por un espectrógrafo en lugar de símbolos escritos a mano.

La investigación en la identificación de locutores bajo dramáticamente de ritmo al término de la guerra, y no es sino hasta finales de los cincuentas o principios de los sesentas que la investigación inicia de nuevo, esta vez a petición de la policía de Nueva York que estaba recibiendo un gran numero de amenazas de bomba hacia los aeropuertos.

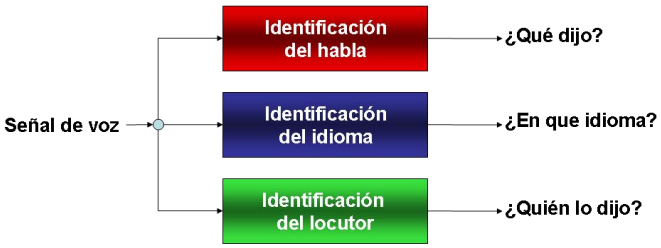
En esta ocasión la investigación fue encargada a Lawrence G. Kersta de laboratorios Bell, en dos años Kersta desarrollo un método para identificar a una persona por medio de su voz en los que reportaba resultados con un 99.65% de certeza.

A partir de esto y a partir de 1966 la policía de Michigan comienza como pionero la aplicación de los métodos de identificación de locutores para la resolución de casos criminales, recibiendo los especialistas capacitación por el mismo Kersta.

Este es el inicio de la identificación de locutores por medio de su voz, mas sin embargo lo que nos atañe es la identificación de locutores auxiliada por herramientas informáticas, esto por la necesidad tanto de certeza en este tipo de estudios como de acelerar el proceso de identificación que con las técnicas empleadas por los expertos actualmente y que a cambiado poco en los últimos 50 años, conlleva una serie de procedimientos tardados que orillan a que un estudio de esta índole pueda tardar meses o años en algunas ocasiones.

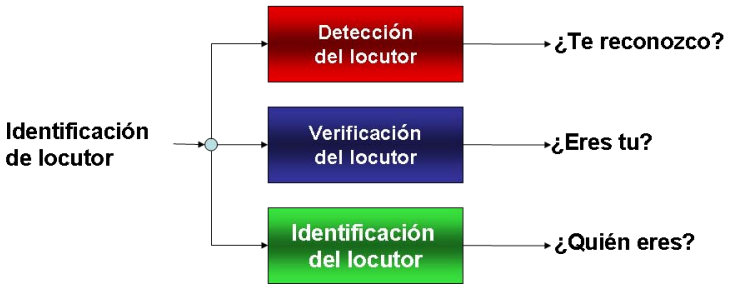
Ahora, también es importante mencionar que a pesar de que se usan herramientas computarizadas en el desempeño diario de las funciones de cualquier perito en acústica forense, estas herramientas tan solo sustituyen al los antiguos espectrógrafo, llevando a cabo todos los cálculos y presentación de resultados en pantallas de computadora por medio de programas informáticos.

Pero más importante aun que distinguir simplemente entre herramientas informáticas y los programas de identificación automática o semi-automática hay que adentrarse un poco mas en todo lo que se puede trabajar en cuanto a la voz humana se trata, por lo que de inicio dividiremos de esta manera el estudio de la voz, que recordemos, después de todo siempre será la extracción de información o características, en nuestro caso de forma automática.



Aquí vemos una primera división, al estudiar la voz podemos estar interesados en *que se dice*, como la inmensa mayoría de la investigación y programas actualmente disponibles, *en que idioma se dijo*, que no suele ser muy común, ya que por lo general tiene aplicaciones muy especificas, y por ultimo *quien lo dijo*, que es la pregunta que nos interesa, por lo que las otras dos ramas ya no serán tratadas en este documento.

Pero como veremos todavía hay mas sobre este tema, pudiendo de nuevo dividirse el estudio como a continuación se muestra de nuevo.



De esta subdivisión vemos que no será igual detectar a un locutor dentro de un grupo cerrado de voces, como podría ser el caso de un banco con seguridad biométrica o un sistema de acceso a un edificio de oficinas, a un sencillo verificador como puede ser una chapa digital que después de introducir nuestro *pin* o numero de seguridad nos pedirá decir una frase previamente grabada para verificar nuestra identidad y por ultimo el mas difícil de todos los campos, la identificación de un locutor contra un numero sumamente grande de posibles candidatos (población de una ciudad o país, etc.) contra un material determinado del cual se desea conocer a su autor o locutor. Esta terna en que dividimos la identificación en general es la que comúnmente se conoce como *verificación*, *autentificación* e *identificación*.

Ahora sabemos que aun descartando ya por de facto temas tales como la codificación de la voz o síntesis de voz existen múltiples ramas de estudio para la voz humana, siendo uno de los mas difíciles la identificación de locutores, ya que existen múltiples factores que afectan o dificultan el resultado, como pueden ser que el sistema no puede tener entrenamiento o calibración, que el numero de posibles locutores es para todo fin practico infinito, que las grabaciones problema y testigo tendrán diferentes circunstancias de grabación, que normalmente se contara con poco material y de mala calidad, el uso de filtros como pañuelos, modificadores de voz, etc.

Como ya vimos en las explicaciones anteriores nos incumbe una muy particular porción dentro del estudio de la voz humana para fines informáticos, y como podemos también vislumbrar la mas difícil y menos trabajada, y esto no es de sorprenderse, después de todo, que pensamos que tenga un mayor mercado potencial, ¿un programa de dictado automatizado o un sistema de identificación de voz?, y la respuesta es obvia, el mercado es poco, pero no de menor importancia, después de todo las aplicaciones policiales y forenses son de suma importancia para nosotros como sociedad.

Existen a la fecha algunas tesis y trabajos de investigación atacando este particular tema, mas sin embargo estos están centrados no en el desarrollo y prueba de un sistema de esta índole, sino mas bien en algún aspecto teórico especifico, siendo innovador este trabajo en su punto de vista que no se adentra en un solo aspecto, sino mas bien de da un punto de vista general que nos permite apreciar el problema claramente y evaluar por medio de un sistema algunas metodologías que actualmente se conocen.

En cuanto a antecedentes históricos en el tema en cuestión realmente no encontraremos nada importante, ya que los pocos desarrollos informáticos que actualmente existen, teniéndose conocimiento de cuatro de ellos, un sistema Ruso desarrollado por ex agentes de la KGB, un sistema Americano desarrollado por la CIA, un sistema Israelita desarrollado por la Agencia de Inteligencia y un sistema Español desarrollado de manera conjunta entre la universidad Complutense de Madrid y la Policía Científica, mismos que son por supuesto altamente secretos.

Resumiendo, podríamos llenar muchas hojas dando la historia de cómo se ha logrado mejorar los programas de dictado automático, o como las chapas digitales de alta seguridad han evolucionado con el tiempo para incluir y mejorar sus mediciones biométricas, mas sin embargo de poco impactarían al actual trabajo, ya que aunque parientes cercanos no son el tema de nuestro interés, por lo que mejor daremos paso al siguiente punto y continuaremos sabiendo en general la historia de la identificación de locutores y que es un tema sumamente fértil donde seguir evolucionando e investigando y de suma importancia para la sociedad en general.

En cuanto al estado actual, existe poca información al publico en general, pero podemos comentar que de los cuatro sistemas que se mencionaron anteriormente solo dos de ellos están a la venta, siendo restringida la misma a instituciones gubernamentales de impartición y procuración de justicia y que cada sistema oscila entre los $100,000 a $750,000 dólares por sus modelos básicos, por lo que entendemos bien su secreto en cuanto a algoritmos y formas de trabajo.

Justificación

Como por desgracia estamos concientes estamos viviendo épocas de gran violencia, donde delitos tales como la privación ilegal de la libertad, el secuestro, el chantaje, el fraude telefónico, las amenazas y muchos otros se están dando cada día mas, seguramente entre nuestros compañeros de trabajo, amistades o familiares podremos fácilmente localizar a una o mas victimas de delitos como los ya mencionados, por lo que vemos que hoy en día no es cuestión de ser rico o famoso para sufrir este tipo de ataques a nuestras personas o familias, siendo estos delitos además de sumamente violentos, lacerantes a la victima y a sus familiares y amigos, causando un gran impacto social, por lo que vemos que contar con herramientas de primer nivel no es un capricho tecnológico, sino una necesidad social.

Desde sus inicios hace poco mas de 50 años con las amenazas de bomba a aeropuertos de la ciudad de Nueva York al hoy en día común secuestro virtual que se lleva a cabo todos los día en esta ciudad de México, vemos como una necesidad imperativa el poder identificar a una persona por medio de sus registros vocales, aunado a muchos otros métodos de identificación como son huellas dactilares, genética, química y demás materias que ofrecen una batería de estudios para responder a cuestionamientos concretos de los Ministerios Públicos y de los Jueces para la adecuada procuración e impartición de justicia en nuestro país.

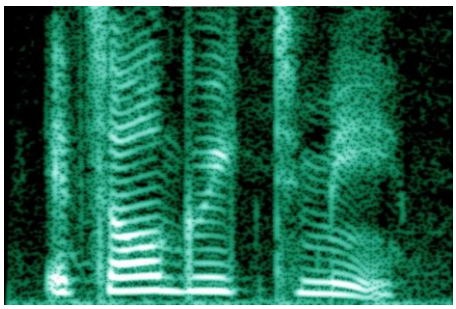
Y por si fuera poco el auge delictivo que vivimos cada vez nos enfrentamos con un criminal más preparado y tecnificado, por lo que cada vez debemos de tener mas y mejores herramientas para luchar contra el criminal, y poder ofrecer a la sociedad un lugar mas seguro donde vivir, así como garantizar a la sociedad que nadie quedara impune cuando incurre en un hecho delictivo, por esta razón y atendiendo a esta demanda de la sociedad se implementan cada vez mas áreas especializadas, tratando de abarcar todo rubro que pueda ser necesario.

De esta manera y por las nuevas necesidades de servicio surge el laboratorio de acústica forense, utilizando tecnologías de vanguardia para cumplir su función primordial de poder determinar si una voz en una o mas grabaciones corresponden contra un o mas determinados sujetos, así como la posible localización de un individuo entre un universo de voces registradas para tratar de encontrar a un presunto responsable.

Por ultimo podemos ver que se justifica plenamente en nuestro país la investigación en la materia tanto por el elevado precio de los pocos sistemas actualmente disponibles como por el aspecto de desarrollo de tecnologías nacionales destinadas a este poco contemplado aspecto del procesamiento de señales de voz con fines identificativos, pudiéndose en un futuro no muy distante desarrollarse programas orgullosamente hechos en México que se puedan incluso vender a otros países o simplemente satisfacer las necesidades internas con un costo mucho menos, e incluso incluyendo aspectos regionales como acentos, dialectos indígenas y muchas mas variantes propias de nuestro país y cultura.

Propósitos y objetivos

Como se ha venido comentando, el propósito del presente trabajo es ofrecer al lector una síntesis general de la identificación de locutores con fines forense, desde sus orígenes, justificaciones y por supuesto análisis llevándolo de la mano hasta el desarrollo e implementación de un sistema informático destinado a la asistencia en la identificación de locutores.



Espectrograma de una voz humana, cortesía de Wikimedia

Como también ya se comento no es difícil localizar trabajos de investigación que traten sobre la identificación de locutores, mas sin embargo estos por lo general trataran muy profundamente un tema en especifico sobre este amplio campo, siendo en general de muy difícil comprensión para una persona común que no este ya inmersa en el tema.

Como podremos ver mas ampliamente en el resto del documento la investigación y trabajo con señales biológicas como la voz presentan retos que no son tan fáciles de vislumbrar de entrada, como la generación misma de la voz, el porque nuestra voz es única e identificable, aspectos sicológicos del habla, posibles diversificaciones de investigación, modelado matemático, procesamiento digital y muchos otros temas, por lo que vemos que en general el conocimiento sobre una sola materia o carrera universitaria será poco útil, requiriéndose conocimientos de medicina, psicología, foniatría, lingüística, ingeniería y quizás algunas otras que por el momento no me vienen a la mente.

Durante el presente se pretende dar al lector toda la base teórica, y si no es tema de este documento el enseñar a una persona a llevar a cabo un estudio comparativo de tipo forense se trataran en general todas sus bases y fundamentos, preparándonos así para continuar con futuras investigaciones o para poder comprender mejor los documentos altamente especializados que se pueden localizar en Internet o en otras fuentes.

Teniendo ya en cuenta este propósito general y la idea primaria que es el desarrollo de una herramienta informática que auxiliara a los peritos en los estudios de identificación de locutores podemos sentar una serie de objetivos que se irán cubriendo durante el presente, siendo estos:

- Antecedentes generales
- Justificación
- Base teórica
- Planteamiento del sistema
- Implementación del sistema
- Pruebas del sistema
- Conclusiones

Como podemos incluso ver ya se han cubierto los dos primeros objetivos, por lo que sin mas continuaremos cubriendo estos respondiendo una simple pregunta ¿Es esto posible?

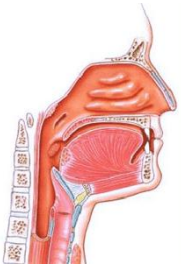
Hipótesis

Como creo que ya algunos de ustedes se habrán preguntado y si no al comenzar a leer esto se preguntaran:

¿Es posible identificar a una persona por medio de su voz?

La respuesta corta es: Por supuesto

Ahora vamos a ver esto con un poco mas de detenimiento, de primera instancia tenemos que ver que es la voz humana, esto de entrada ya presenta un problema, ya que desde el punto de vista lingüístico la voz es una expresión semántica de los referentes que participan en una oración, desde el punto de vista de la comunicación y el lenguaje es un transporte de información, y desde el punto de vista fisiológico es el sonido producido por el aparato fonador humano, que como vemos es el que nos interesa por ahora.

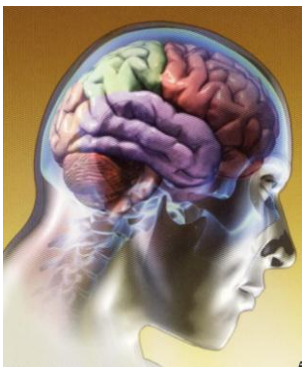


Por ahora y para no desviarnos de la pregunta planteada no veremos de nuevo como es que se produce la voz humana desde el punto de vista fisiológico, esto ya lo vimos anteriormente, por ahora solo pensemos que es el resultado de un proceso sumamente complejo en el cual intervienen muchos músculos y una buena parte de la función cerebral, razón por la que no debemos de hablar por celular al conducir un automóvil, y no la absurda idea de que con un manos libres ya estamos a salvo pues no utilizamos las manos.

De este hecho de que son los sonidos producidos por el aparato fonador del ser humano ya tenemos nuestra primera pista al porque nuestra voz es única e identificable, al intervenir no una sino varias partes de nuestro cuerpo para formar la voz podemos pensar de una manera casi inmediata, estas estructuras son únicas a cada ser humano, seguro, todos tenemos pulmones, lengua, mejillas y labios, pero ¿Son iguales los míos a los de la persona de al lado?

La respuesta es que no, son similares, mas no iguales, cumplen las mismas funciones, tienen en general las mismas características, pero por simple lógica no tendrán el mismo tamaño ni desplazaran la misma cantidad de aire los pulmones en un niño de un metro con veinte centímetros que en un individuo que mida mas de dos metros. Con esto esta dada la primera parte de la respuesta, al intervenir no una sino múltiples estructuras fisiológicas esto hace que nuestra producción de voz sea particular a cada individuo, única e identificable, pero esto no es todo.

Para el segundo punto pensemos por un instante en como es que hablamos, ya mencionamos que intervienen muchos músculos y varias estructuras fisiológicas, pero ¿Quien comanda todo esto?, la respuesta es el cerebro, este será el encargado de dirigir y coordinar toda la función del habla, de recordar las palabras, su pronunciación, su significado, la forma de ensamblar las frases, y en un contexto mas amplio de entender la comunicación y regresar una respuesta coherente para que no se interrumpa el ciclo de comunicación que es la función primordial del habla en nuestra sociedad.

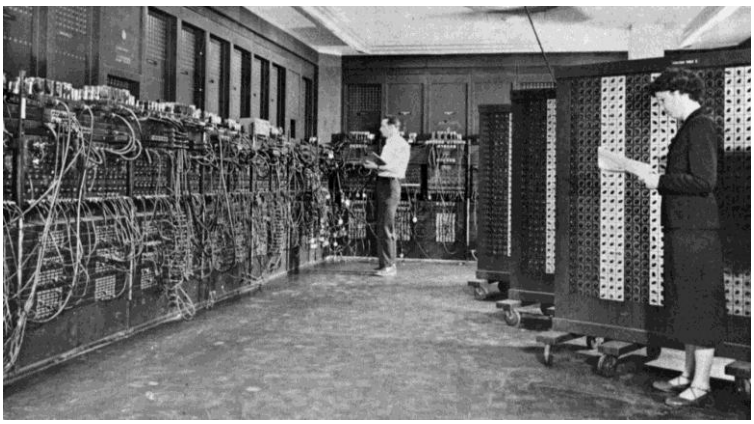


De esto podemos ver ya de donde vendrá la segunda justificación, y sin ir a algo tan burdo como que el cerebro de dos personas es distinto vamos a algo un poco mas allá, la forma de aprendizaje, como incluso se a extrapolado a la ingeniería y a la ciencia que trata de imitar la forma en que nuestro cerebro funciona, después de todo, a la fecha sigue siendo la computadora mas impresionante del mundo, siendo superada quizás por las computadoras digitales modernas en

velocidad de reacción, pero no acercándose siquiera a su capacidad y potencial, y a este método de funcionamiento se le conoce como redes neuronales.

Para entender esto mejor tendremos que ver desde el punto de vista de la psicología conductista, esta nos dice que el aprendizaje tiene como proceso fundamental la imitación o la repetición de un comportamiento observado, este proceso de imitación toma por supuesto tiempo para observar a detalle aquello que se desea imitar.

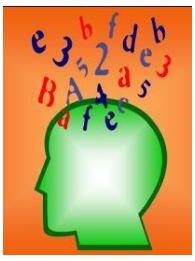
Desde el punto de vista molecular y como lo comprobó el trabajo ganador del premio Nóbel en el año 2000 por Erick Richard Kandel, el proceso de aprendizaje cambia físicamente la forma en que nuestras neuronas realizan sus conexiones sinápticas y el RNA (ácido ribonucleico) interno de las células, y todo esto es en si el resultado de asociaciones entre estímulos y respuestas mediante la práctica en un nivel elemental, esto es decir, la imitación con una retroalimentación.



Computadora ENIAC, cortesía de Wikimedia

Ahora sabemos algo nuevo, como aprendemos, de este hecho hay algo que se hace patente, aprendemos extremadamente diferente de cómo aprenden las computadoras digitales, donde basta con cargar un disco de cualquier tipo para que la computadora aprenda una tarea totalmente nueva sin jamás haberla hecho antes, la realizara rápido, a la perfección (casi siempre) y de una manera muy eficiente, por repetitiva que sea esta tarea, mas sin embargo enfrentaremos un trago amargo si queremos que una computadora aprenda a pintar la próxima Mona Lisa, ya que carecerá de imaginación y genialidad, aunque podría imprimir quizás unas cuantas miles de copias por hora a diferencia de Leonardo da Vinci.

Del hecho de que el ser humano aprende por la imitación podemos ver el segundo punto importante de porque nuestra voz es única e identificable, este proceso tratara de imitar un comportamiento, por lo que es imposible que se adquiera la forma de hablar "idéntica" a la de la persona que estamos imitando, aunado a que este proceso de aprendizaje se llevara a cabo durante la infancia por lo general, durante la cual tendremos la influencia de un padre, una madre, hermanos, tíos, etc. Por lo que este proceso de aprendizaje se vera plagado de diferentes modelos o comportamientos.



Otro ejemplo rápido con el que podemos ejemplificar este hecho de las patentes diferencias entre diversos sujetos expuestos a un mismo comportamiento que se desea imitar o aprender serian las clases de ingles o de pintura, aquí podremos ver como con un único comportamiento a imitar veremos alumnos con una facilidad asombrosa para adquirir un segundo idioma o para aprender a pintar, y por el otro extremos tendremos aquellos individuos que difícilmente lograran estos fines, ya que como coloquialmente se maneja "no tienen esa habilidad", aquí podemos fácilmente plantearnos que un grupo de individuos "imitaran" en diferentes

grados y con toques característicos cualquier rama del aprendizaje, incluida el habla.

De estas explicaciones ya tenemos claro que la voz de cada persona es única e identificable, ya que tanto sus características fisiológicas como neuronales determinaran sus forma de hablar, siendo esta totalmente característica a cada individuo, mas sin embargo todavía hay un factor mas que es digno de mencionarse, quizás no es tan importante como para definirse como una justificación mas o como un elemento que nos podrá individualizar a un sujeto, pero si es importante en las ciencias forenses, ya que gracias a este factor podemos de manera rápida ubicar a un sujeto, aportando incluso información sobre el mismo, y este factor es *el medio de aprendizaje*.

El medio de aprendizaje o en otras palabras el lugar y circulo social donde crecimos nos dotara de una serie de influencias que caracterizara nuestra forma de hablar, existiendo factores como el social, el económico, el nivel educativo y el lugar geográfico.

Estos factores formativos son muchas veces de fácil utilización y otras no tanto, pero como explicaremos determinaran muchos aspectos de la forma de comunicarse de un individuo, empezaremos por la ubicación geográfica, de esta podemos ver aspectos tales como el acento o entonación del habla coloquial, en esta podemos pensar quizás en el acento del "Norteño", del "Veracruzano" y del "Chilango", estamos de acuerdo, no todos los individuos que nacieron en el Distrito Federal tendrán ese característico tono "cantadito" como algunos lo clasifican, pero si puede presentarse, estudiarse e incluso clasificarse.

Para mostrar la importancia de los tonos regionales en manos de un experto bien entrenado en la acústica forense este podría llegar a descubrir el probable lugar de nacimiento de una persona, el lugar donde creció y el lugar donde vive tan solo por sus inclinaciones idiomáticas y su entonación característica, este tipo de estudios están fuera del alcance del presente documento, pero forman parte del perfil vocálico.

Grupo de gente rural en el México de 1900, cortesía de Wikimedia

En cuanto al nivel de educación y nivel socio económico estos factores suelen ir de la mano, los niveles socio económicos bajos suelen tener un nivel educativo menor, adquiriendo por estos factores algunos vicios del habla como pueden ser la sustitución de fonemas y la omisión de fonemas. Otro factor importante en los jóvenes es el uso de palabras conocidas como "slang" o de uso vulgar (o lenguaje urbano) y los extranjerismos.

Por ultimo tenemos como una importante influencia el entorno educativo, el resultado de un bajo nivel educativo será un bajo dominio del idioma en grupos o sectores poblacionales, llegando a encontrase grupos sociales que según los estudiosos de la materia tienen vocabularios

de alrededor de 250 palabras, elevándose a 1,000 en promedio en grupos con educación de buena calidad y alrededor de 5,000 para personas sumamente cultas.

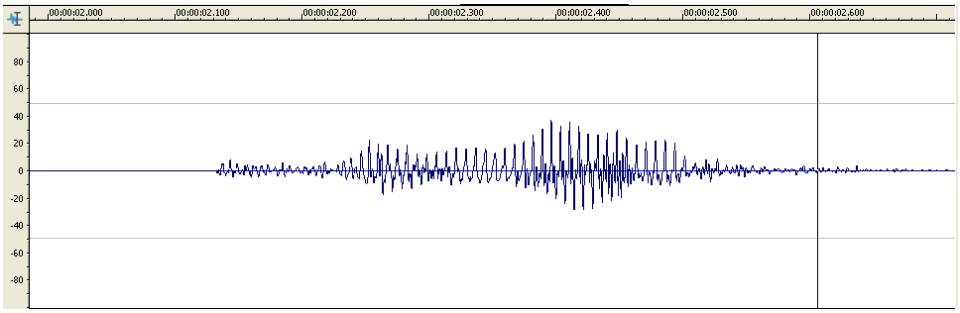
De estos tres factores podemos concluir que de manera definitiva, la voz de una persona es única e identificable, podremos encontrar individuos con voces similares por compartir entornos y modelos a imitar, como son los hermanos, y por supuesto los imitadores de voces, mas sin embargo podemos garantizar que no serán idénticos, sino tan solo parecidos, con esta respuesta ahora si sabemos que esta biométrica tiene una sólida base para afirmar que podemos identificar a una persona por medio de su voz.

Diseño de la investigación

En este respecto sabemos ya que trabajaremos sobre tres juegos de parámetros o tres metodologías, siendo la primera el estudio de las señales en el dominio del tiempo, realmente creo que no tenemos mucho más que explicar en este tema, ya que hemos cubierto de manera adecuada los antecedentes, mas sin embargo podemos quizás facilitar un poco mas la comprensión de lo que se pretende con algunas simples explicaciones:

¿Qué veo en el dominio del tiempo?

En el dominio del tiempo y como ya vimos se puede observar una representación de la señal de audio, en este caso en concreto de voz humana donde apreciaremos de manera general una representación de las variaciones de presión del medio quasi estático de transmisión de la onda de voz.



Voz humana en el dominio del tiempo

Como podemos ver en la grafica anterior podemos observar en esta como la señal sube y baja de un cero o presión estándar y va variando de presiones altas a bajas y viceversa, de esta grafica podemos ver que existen los ya mencionados cruces por cero, máximos y mínimos relativos, energía y demás parámetros que vimos anteriormente.

Este tipo de graficas como se sabe en el medio de los que nos dedicamos a la acústica forense realmente no aportan elementos identificativos, ya que es sumamente difícil poder caracterizar o detectar parámetros en este tipo de señales, salvo los muy obvios como silencios, duración, etc.

Por este motivo mas adelante veremos que tanto puede aportar un estudio mas objetivo y matemático de esta señal, que aunque es el origen y motivo de estudio no facilita el trabajo que pretendemos llevar a cabo.

¿Qué son los espectrogramas y sonogramas?

Como casi todos hemos experimentado existen múltiples soluciones y puntos de vista a un problema, casi siempre encontrándose caminos alternos, no siempre mejores, simplemente distintos, por lo que resulta conveniente saber que existen estas distintas aproximaciones antes de tomar la decisión de ir por un camino, por lo que sin mas pasemos a ver que son y para que nos pueden ser útiles.

Espectrogramas y sonogramas

Como un error común, lo que normalmente conocemos como espectrograma debe llamarse sonograma, ya que si recurrimos a definiciones más precisas o más próximas a

su significado exacto tenemos que un espectrograma es la representación gráfica de una transformación del dominio del tiempo al dominio de la frecuencia, para este tipo de aplicaciones usaremos lo que se conoce como una FFT (Fast Fourier Transform o Transformada Rápida de Fourier), este tipo de gráficas son instantáneas, lo que significa que representan tan solo un instante (porción o segmento designado por el analista como número de muestras a trabajar) determinado de un sonido, sin importar la duración del sonido en estudio, como se puede ver en la figura siguiente.

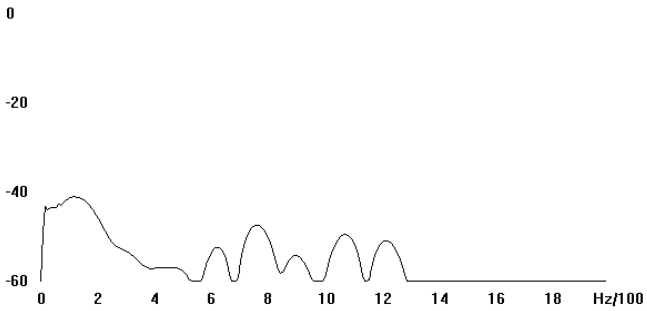


Figura 62: Espectrograma

Existe lo que se llama un espectrograma en cascada, que es el despliegue sucesivo (simulando una gráfica en tres dimensiones), en la que se agrega el eje del tiempo, y se podrá ver como varía el espectrograma en cada segmento de la señal estudiada, este tipo de representación es de difícil lectura, ya que las cimas de una gráfica pueden tapar las simas de otra gráfica anterior en el tiempo, lo cual dificulta la apreciación de factores importantes, como las fluctuaciones de las formantes, como se aprecia en la figura.

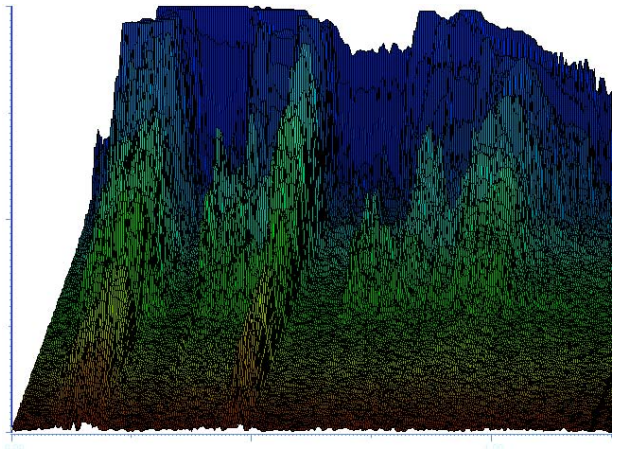


Figura 63: Espectrograma en cascada

De dicha problemática se buscaron nuevas formas de representar las señales que permitieran una visualización más rápida y fácil de leer, de esto nace el representar los espectrogramas como se vieron anteriormente girando los ejes de la frecuencia y la amplitud, ahora de este tipo de representación formamos un código de colores o tonos de gris que represente la intensidad, siendo por citar un ejemplo, los tonos de gris donde el negro es el máximo valor de frecuencia y el blanco la mínima o ausencia de amplitud, con tantos tonos intermedios como resolución se desee, esto formaría una barra de un instante como la de la figura siguiente, y si por último tomamos prestado el concepto de agregar el eje del tiempo, como en un espectrograma en cascada, obteniendo un sonograma, como se muestra finalmente.

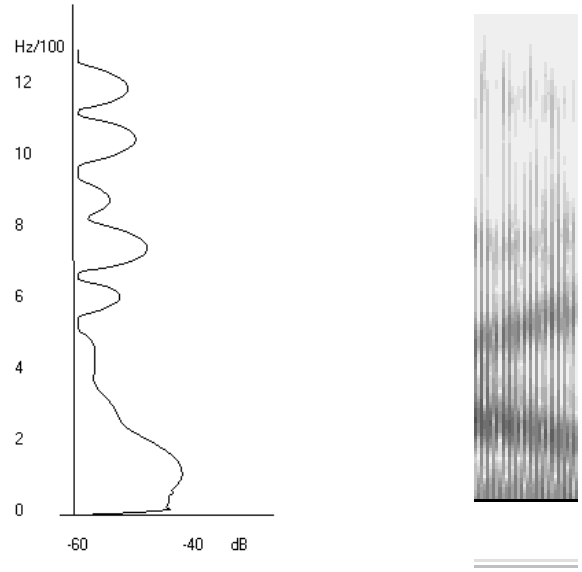


Figura 64: Espectrograma y porción de sonograma

De esta representación podemos ver que el eje de las abscisas será el tiempo, y el eje de las ordenadas representa la frecuencia, pudiendo tener sonogramas de banda ancha o de banda estrecha, como ya se explico antes, y el color o tono de gris representa la amplitud o nivel energético de la señal.

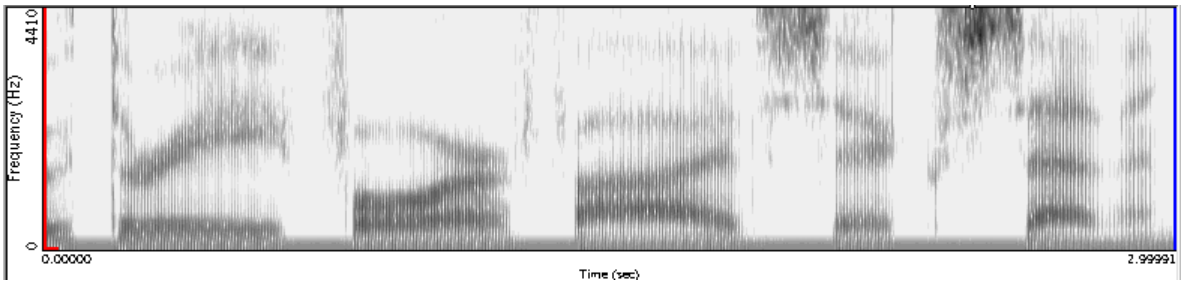


Figura 65: Sonograma

¿Que me aporta un análisis LPC?

En realidad un análisis LPC nace con la compresión y transmisión de señales de audio en mente. Es la metodología utilizada hoy en día para tecnologías tales como VoIP o Voice over IP, que es la telefonía sobre protocolos de comunicación IP o Internet, la telefonía celular GSM que es la que actualmente tiene mayor auge, vocoders para la generación de programas de computo parlantes y muchas mas tecnologías que utilizamos de manera diaria.

La adaptación de esta tecnología a los fines que perseguimos viene de un hecho bastante simple, se genera un modelo de las estructuras formanticas o de los resonadores de manera indirecta, esto con la finalidad por supuesto de agregarle alguna información mas como el residuo, el tono fundamental y los datos adicionales que sean convenientes, transmitir esta información por un medio de comunicación y poder reconstruir la señal de voz original en el otro extremo.

Pero que tal si en lugar de transmitir esta información le damos un procesamiento o tratamiento distinto, aquí como ya se comento aprovechamos el hecho de que el modelo LPC nos permite

obtener la envolvente, y como ya también vimos la envolvente coincide con el comportamiento espectral que se puede obtener por una transformación de Fourier, pudiendo obtenerse una forma de representación alternativa a los ya conocidos espectrogramas.

Ahora recordemos que los coeficientes LPC son los polos de un filtro, ya que el método de LPC lo que hace es construir un filtro que tratara de aplanar el espectro, o como se le conoce en ocasiones, "blanquear el espectro", por lo que obtendrá una serie de valores tales que eliminen las componentes en frecuencia de acuerdo a la ecuación que norma la función de transferencia del modelo, que es la siguiente:

$$H(z) = \frac{1}{1 + \sum_{j=1}^{p} a_j z^{-j}}$$

De esto podemos pensar que en el lado contrario se requiere amplificar o potenciar aquellas frecuencias que en el lado del codificador fueron aplanadas, por lo que si graficamos el filtro inverso de los coeficientes LPC esto nos dará la envolvente, que marca el comportamiento en frecuencia de la voz modelada, o sus formantes.

Esta envolvente como ya lo mencionamos tiene en general el mismo comportamiento que el espectro producido por una transformada de Fourier, razón por la cual se le conoce también como espectro LPC, pudiendo este graficarse en cascada o por colores como su contraparte del análisis en frecuencia y en ocasiones con mejores resultados, a continuación se muestra una grafica de un espectro LPC en cascada, lo cual nos hace ver por simple analogía con lo ya tratado en espectros y transformadas de Fourier su utilidad y uso.

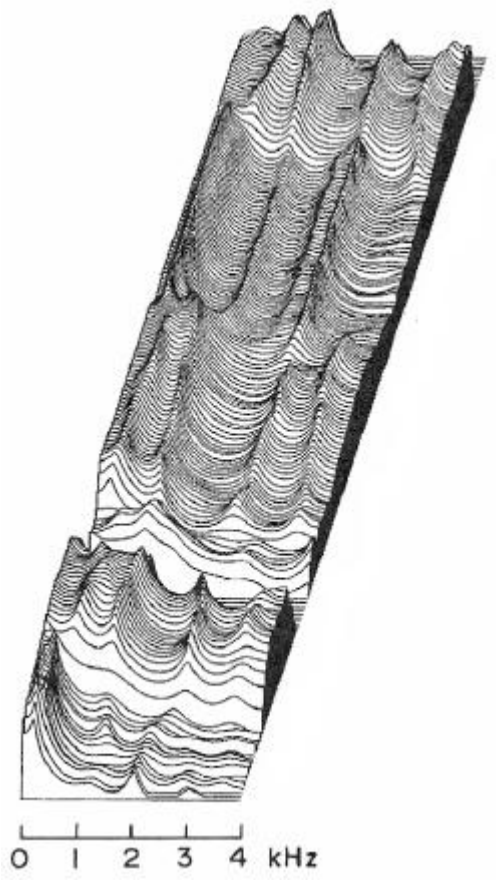


Figura 66: Espectro LPC de cascada

El sistema

Ahora que sabemos a grandes rasgos como se trabaja con los tres métodos o aproximaciones y sus fundamentos pues realmente lo que nos queda es explicar como trabaja el sistema que se desarrollo en base a estas tecnologías y las partes que lo componen.

Dentro de este apartado se detalla el desarrollo un sistema informático que pone en uso algunas de las técnicas descritas anteriormente, siendo el fin de este sistema el de brindar una herramienta lo mas automatizada posible que de una gama de parámetros comparativos entre diversas muestras de voz previamente digitalizadas para su estudio.

Se explicaran los principales módulos de este sistema informático, la estructura de los archivos de salida, ya sea intermedios o de resultados finales y se explicara como se pueden analizar las distintas pantallas de resultados que ofrece el sistema, así como la finalidad y funcionamiento de las diversas herramientas informáticas que se implementan.

Como ya hemos discutido y analizado la voz es un aspecto único a cada individuo, lo cual lo hace de carácter aportativo para una posible identificación, esto es debido a que la voz presenta rasgos distintivos desde muchos puntos de vista como pueden ser:

1) Diferencias fisiológicas en el tracto vocal, lo cual lo hace único a cada individuo, ya que como toda estructura orgánica nunca será idéntica en dos seres vivos, dando variantes como velocidad de dicción, tono de voz o timbre, fisiologías o patologías individuales, como malformaciones o perdidas de partes del aparato fonador por accidentes, etc.
2) Influencias ambientales, de lo cual podemos decir que no será la misma forma de hablar de una persona de un estrato socio-cultural y económico alto que el de uno de nivel bajo, donde veremos diferencias en el modo de la inflexión o conjugación, modismos distintivos del habla de una región o estrato social, etc.
3) Grupo étnico de pertenencia, donde podemos ver de forma clara que podremos fácilmente distinguir entre una persona que nació y creció en México, España o Argentina, por mencionar unos cuantos, todos hablamos Castellano, mas desde el punto de vista estricto de lenguaje, se podría decir que hablamos respectivamente Mexicano, Español y Argentino, ya que el significado de algunas palabras varia de una etnia a otra, así como modismos del lenguaje y hábitos de habla.
4) Diferencias a nivel neuronal, que comprenden el hecho de que nuestra forma de aprender y hacer las cosas no es por un aprendizaje exacto como el de las computadoras que siempre harán una misma función de manera idéntica, sino por el contrario realizamos una mímica de lo aprendido, parecido, pero nunca igual, esto se puede ver claramente en que la caligrafía, la voz, y otras funciones aprendidas nunca serán iguales en dos individuos.
5) Condiciones de ambiente y variantes personales, de donde se desprende que no será igual nuestra voz a nivel del mar que en una montaña a 5000 metros sobre el nivel del mar, presentándose también pequeñas variantes por aspectos como humedad relativa del ambiente, composición química de los gases atmosféricos (contaminantes, porcentaje de oxigeno, etc.), estado emocional del individuo, tipo de voz que se usa (voz de mando, susurro, voz de lectura, etc.), estado de salud, estados de agitación, ínter variación intra personal y otras muchas razones.

Como podemos ver el número de parámetros a tomar en consideración nos da una idea de lo difícil que es el poder establecer una identidad por medio de la voz, pero a la vez nos da un amplio rango de características únicas e identificables para grupos de personas o individuos, que es el objetivo de este sistema.

Por todo lo expuesto vemos que las funciones del perito en acústica forense no pretenden ser sustituidas por un sistema automatizado, sino por el contrario es una mas de las herramientas con que se contara para el trabajo forense, este sistema tiene como función el analizar y dar un resultado preliminar de que tanto parece ser o no una misma persona en dos o mas tomas de voz, mismas que posteriormente podrán ser examinadas por el perito para el establecimiento de una identidad o descarte de la misma.

El nombre de la aplicación, como a continuación observamos en la pantalla de arranque es el que se considero adecuado a su función, Acústica Forense 1.0, con lo que damos

inicio a la explicación ya del primera parte del sistema, pero no sin mencionar que el desarrollo se llevo a cabo en Visual Basic 6.0 y que el sistema busca sus archivos de trabajo en una carpeta llamada C:\Sonidos de manera fija.

Figura 67: Pantalla de inicio

Las Herramientas

Dentro de lo que se considero como herramientas útiles a incorporar al sistema se encuentran en la pantalla principal dos de ellas, que comprenden la extracción de la información del encabezado de un archivo WAV de Microsoft que consta de 44 bytes, los cuales contienen la siguiente información:

	Bytes	Descripción
Campo 1	0..3	Palabra "RIFF" en código ASCII
Campo 2	4..7	Tamaño total del archivo menos 8 bytes
Campo 3	8..15	Palabra "WAVEfmt " en código ASCII (hay un espacio después de la t)
Campo 4	16..19	Formato, para PCM vale 16
Campo 5	20..21	Formato, para PCM vale 1
Campo 6	22..23	Indica modo mono (1) o estéreo (2)
Campo 7	24..27	Frecuencia de muestreo (por ejemplo 44,100, 8000, etc.)
Campo 8	28..31	Numero de bytes por segundo en grabación o reproducción
Campo 9	32..33	Numero de bytes por captura, que puede ser 1, 2 o 4
Campo 10	34..35	Bits por muestra que pueden ser 8 o 16
Campo 11	36..39	Palabra "data" en código ASCII
Campo 12	40..43	Total de bytes ocupados por las muestras

Tabla 9: Datos del encabezado Wav

Examinar

El comando o botón "Examinar" nos da esta información, y con respecto al campo 2 compensa el tamaño al sumarle los 8 bytes que le hacen falta para dar el mismo valor que veríamos en un programa como Explorador de Windows, en la siguiente pantalla observamos un ejemplo de información de un archivo de 16 bytes monoaural muestreado a 8000 Hz con una duración de 0.5515 segundos que es el resultado de dividir el campo 12 (bytes totales de las muestras) entre el campo 8 (bytes en un segundo de reproducción).

Figura 68: Pantalla principal

También es importante mencionar que por razones del compromiso velocidad de proceso y calidad de resultados se tomo la decisión de que todas las funciones de procesamiento de este sistema se llevaran a cabo con archivos de tipo monoaural (no es recomendado el sonido estereofónico, ya que no aporta ningún elemento útil o ventaja sobre el sonido monoaural, pero haría mas lento el proceso de reconocimiento) y en 16 bytes (mayor calidad de muestreo, ya que el rango de las muestras va de -32,768 a 32,767 en lugar del rango de -128 a 127 de los 8 bytes), dejando la frecuencia de muestreo abierta al gusto del usuario, mas sin embargo recomendando fuertemente una frecuencia de 8000 Hz. ya que recordemos que por la naturaleza de una gran parte de los trabajos en Acústica Forense por lo menos una de las grabaciones en estudio será proveniente de una grabación telefónica, y recordemos que las líneas telefónicas tienen integrado un filtro paso bajas de 8000 Hz, y por el teorema de Nyquist la mayor frecuencia que no se perderá estará alrededor de los 4000 Hz, frecuencia que coincide con los tonos mas altos de la voz humana normal, encontrando poca o ninguna información útil por arriba de dicha frecuencia.

Cambios de formato

Por lo que respecta a archivos de formato estéreo o de 8 bytes, al examinarlos se dará un aviso de que el archivo no es mono de 16 bytes y se ofrece la opción de convertir el archivo a formato mono de 16 bytes, por lo que como podemos ver existen 3 tipos de conversión posible, que son:

Tipo	Procedimiento
Archivo estéreo de 16 bytes	Suma los dos canales en uno solo (no pierde calidad)
Archivo estéreo de 8 bytes	Suma los dos canales y multiplica por 128 (no recomendable por su baja calidad)
Archivo mono de 8 bytes	Multiplica por 256 (no recomendable por su baja calidad)

Tabla 10: Formatos de trabajo

Graficación

En cuanto a la opción de “Graficar” archivos se tiene que el sistema puede graficar los cuatro tipos de archivos WAV mencionados, haciendo una división por colores en espacios de 10 milisegundos, que es el periodo en el cual se considera que la voz presenta un comportamiento quasi-estático y por lo tanto se pueden estudiar sus características como si se tratara de una unidad invariante en el tiempo, pero también presenta la opción de graficar archivos HAM, que como se explicara a detalle mas tarde son archivos WAV a los que se les ha aplicado una ventana de Hamming y si se desea un traslape de ventanas variable de acuerdo al usuario, pero dicho archivo conserva un encabezado idéntico al del archivo WAV, modificando el tamaño si es necesario, pero de hecho se puede renombrar el archivo a WAV y lo podrá leer cualquier programa que pueda reproducir archivos WAV de Microsoft para observar el efecto del traslape y la aplicación de la ventana de Hamming.

La graficación de un archivo activara un control de deslizamiento con el cual se puede avanzar o retroceder en el tiempo para observar todo el archivo en caso de que ocupe mas de una pantalla (8000 puntos por pantalla) y un botón para cerrar la grafica actual, también bastara con hacer clic con el ratón sobre la grafica para ocultar esta misma pero sin cerrar la graficación actual.

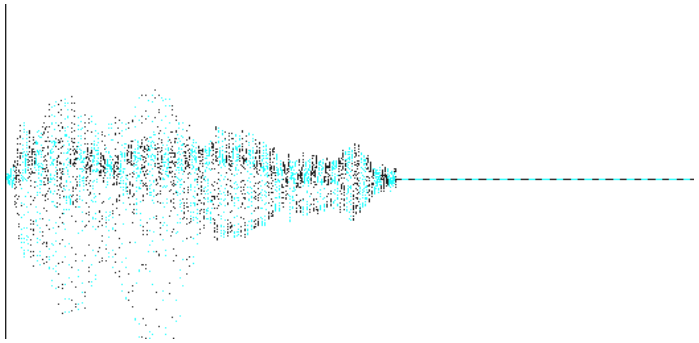


Figura 69: Grafica de archivo WAV

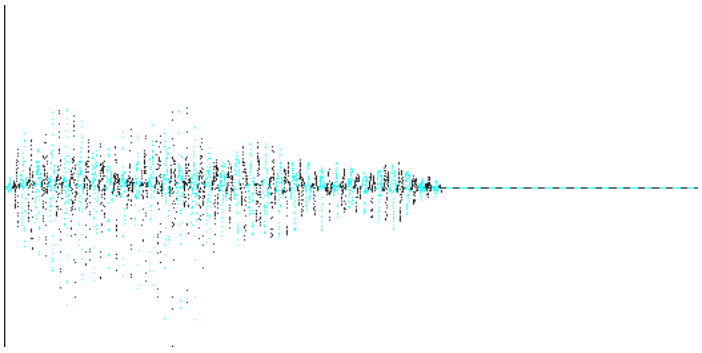


Figura 70: Grafica de archivo HAM

Como se puede apreciar en las figuras anteriores se ve el efecto de aplicar un traslape de ventanas del 10% (1 milisegundo) y de aplicar una ventana de Hamming y como se modifica y alarga ligeramente el resultado graficado.

Interpolación

Como una cuarta y ultima herramienta se incorporo al sistema la interpolación en el tiempo de un archivo, esto es, se toma de los dos archivos seleccionados, como se ve en la figura siguiente, el mayor en tamaño (duración de grabación) y se interpolara el archivo menor con respecto a su mismo contenido para prolongar su duración hasta

igualar la del archivo mayor, que fue una idea que se implemento para el procesamiento en el dominio del tiempo con resultados favorables, como se comentara mas adelante en dicho apartado, pero se dejo dicha herramienta abierta para que el usuario pueda interpolar cualquier par de archivos, pero cabe mencionar que este proceso se creo teniendo en mente grabaciones del tipo repetitivo, como mencionar dos veces una misma palabra con la misma entonación, tratando de compensar un poco el elemento de la ínter variabilidad intra personal.

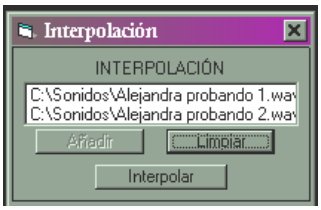


Figura 71: Interpolación de archivos

La ventana de interpolación consta de tres botones, "Añadir" para agregar un archivo a la lista, "Limpiar" para borrar toda la lista e "Interpolar" para llevar a cabo el proceso, agregando al nombre del archivo el postfijo "Intra", por lo que no modificara ninguno de los dos archivos originales.

Identificación por Medio de Parámetros en el Dominio de Tiempo

Por esta aproximación se pretende poner a prueba la hipótesis de que los parámetros básicos en el dominio del tiempo sirven para lograr la identificación de un sujeto, sin más proceso que estudiar los archivos en cuestión con respecto a seis parámetros que se consideraron de interés y de aportación.

Esta función se encuentra en la pantalla principal del sistema bajo el menú "Tiempo", que ofrece a su vez tres opciones, que son "Comparar Dos Archivos", "Manipulación Múltiple" y "Calcula Parámetros", los cuales iremos explicando en el orden que se considera mas oportuno para un optimo entendimiento de su funcionalidad.

Calculo de Parámetros

Para este menú se tiene opción de seleccionar si se desea graficar el resultado de la operación o no, esto se logra por medio de las casillas de selección que tienen el nombre de cada parámetro a un lado, además explicaremos de que se trata cada uno de los seis parámetros que se tienen disponibles, y su posible utilidad:

Magnitud; Se define como la sumatoria del valor absoluto de las muestras que se encuentran en una ventana, se divide entre el numero de muestras para normalizar el resultado y se refiere a una medida cuantitativa que al compararse contra otras de la misma clase pueden ser comparadas.

$$M(n) = \frac{1}{N}\sum_{k=0}^{N-1}\left|x(k)\right| \qquad (39)$$

Media; Se define como la sumatoria de todas las muestras de una ventana y divididas por el numero de muestras y se refiere a la cantidad que representa al promedio de varias muestras.

$$\overline{M}(n) = \frac{1}{N}\sum_{k=0}^{N-1} x(k) \qquad (40)$$

Energía; Se define como la sumatoria de las muestras al cuadrado y se divide por el numero de muestras para normalizar el resultado y se refiere a la capacidad de producir trabajo de la onda.

$$E(n) = \frac{1}{N}\sum_{k=0}^{N-1} x^2(k) \qquad (41)$$

Cruces por Cero; Estos nos indican el numero de veces que la señal atraviesa el nivel cero en cualquiera de los dos sentidos, para que se cumpla dicha condición se deberá de dar un cambio de signo en la señal, no se considerara un cruce por cero el que la señal tome el valor de cero si no es seguido de un valor de signo contrario al anterior al valor cero, así mismo si la señal permanece en cero tampoco se tomara como cruces por cero.

Máximos; Este será el calculo del numero de máximos locales que ocurren en una ventana, se considera un máximo local cuando una muestra tiene un valor superior a la anterior y posterior, si se diera el caso de que varias muestras iguales sean superiores a las anteriores y posteriores se considerara la muestra que se encuentra en medio.

Mínimos; Este será el calculo del numero de mínimos locales que ocurren en una ventana, se considera un mínimo local cuando una muestra tiene un valor inferior a la anterior y posterior, si se diera el caso de que varias muestras iguales sean inferiores a las anteriores y posteriores se considerara la muestra que se encuentra en medio.

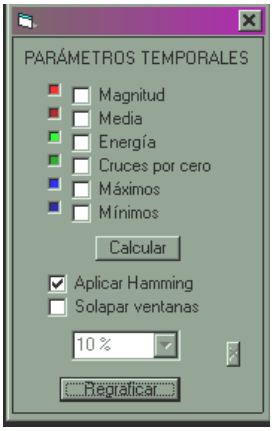


Figura 72: Calculo de Parámetros

La utilidad de algunos de estos parámetros puede ser por ejemplo, que los cruces por cero nos ayudan a detectar la calidad del sonido, ya que las secciones de mucho ruido tendrán un numero más elevado de cruces, así mismo se pueden observar secciones del habla con sonidos fricativos, ya que estos también presentan un aumento de cruces por cero, así mismo la energía y la magnitud por ejemplo nos pueden ayudar a detectar secciones de sonidos sordos o sonoros, ya que los sonoros presentaran mayor energía y magnitud, y en un momento dado se puede emplear una técnica similar para implementar un detector de silencio si ese considerara necesario.

Como se puede observar en la figura anterior, se tiene un color al lado de cada casilla de selección, esto es debido a que en dicho color se graficara cada parámetro seleccionado, facilitándose así la lectura de la grafica obtenida en caso de desear graficar mas de un parámetro simultáneamente.

La graficación de los resultados puede ser obtenida de dos maneras, como una etapa final de la opción "Calcular", ya que la computadora verificara si alguna casilla esta seleccionada y graficara dicho parámetro, activándose un botón para cerrar dicha grafica, y por otra parte el botón de "Regraficar" que graficara de nuevo cualquiera de los seis parámetros de un archivo de resultados previos.

Además de esto observamos que tenemos dos casillas mas de selección, y un menú de caída, estos son "Aplicar Hamming" que aplicara una ventana de Hamming a los datos antes de procesarlos y la opción de "Solapar Ventanas" que en manera conjunta con el menú drop-down nos permite seleccionar entre solapar al 2%, 5%, 10% (default), 12% y 15%, esto se maneja en porcentajes debido a que tenemos abierta la frecuencia de muestreo y se aplicara el mencionado porcentaje del numero de muestras contenidas en 10 milisegundos de grabación a dicha frecuencia.

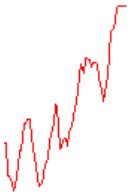


Figura 73: Grafica de magnitud

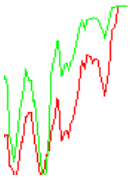


Figura 74: Grafica de magnitud y energía

Resultados parciales y finales

Como podemos observar en el directorio de trabajo "C:\Sonidos" (que previamente se menciono es el directorio fijo de trabajo), se observan varios archivos nuevos, que son los resultados parciales y finales de las distintas opciones de análisis que se implementaron en este modulo del sistema, a continuación se explica el formato del encabezado de dichos archivos, su función y el formato de los datos guardados en cada uno de ellos.

Archivos "win"; estos archivos contienen el resultado de la aplicación de un solapamiento de ventanas, esto es, si tenemos por ejemplo los datos x = {0, 1, 3, 5, 12, 6, 3, -1, -5, -6, 1, 8, 6, 2, 0}, tenemos un total de 15 muestras y tomaremos una ventana de 5 muestras, por lo que si solapamos al 20% significa un solapamiento de 1 muestra por ventana, por lo que nuestro vector aplicando el solapamiento quedaría como x' = { (0, 1, 3, 5, 12) , (12, 6, 3, -1, -5) , (-5, -6, 1, 8, 6) , (6, 2, 0, 0, 0) }.

Como se ve en la figura siguiente el solapamiento de ventanas es un proceso en el cual cada ventana se recorre el número de posiciones que representa el porcentaje seleccionado del tamaño de una ventana, el tamaño de la ventana será en este caso variable, dependiendo como ya se menciono de la frecuencia de muestreo, pero será el numero de muestras que tengan 10 milisegundos de audio, y el porcentaje lo escoge el usuario, así que ahora nuestro vector del ejemplo tiene 20 muestras, ya que se agregaron 2 ceros a la ultima ventana para rellenarla y que este completa esta ultima.

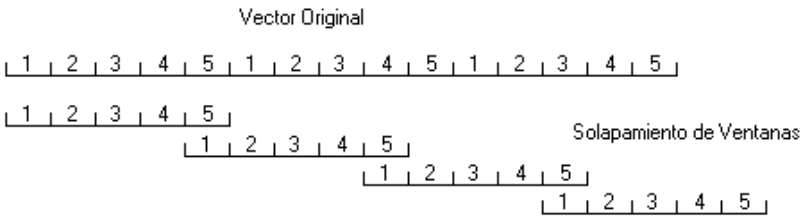


Figura 75: Solapamiento de ventanas

Es pertinente mencionar que los archivos "win" tienen un encabezado idéntico al de un archivo "wav", pudiendo incluso renombrar el archivo y escucharlo con un programa común que reproduzca archivos "wav" compatible con el estándar de Microsoft para apreciar el efecto del solapamiento de ventanas en la señal.

Archivos "ham"; estos archivos contienen el resultado de la aplicación de una ventana de Hamming a partir de un archivo "win" si se aplico solapamiento de ventanas o de un archivo "wav" en su defecto, dicha ventana es sugiere como la mas conveniente para el procesamiento de señales de voz con fines forenses por la WGFSAA (Working Group for Forensic Speech & Audio Analysis) perteneciente a la ENFSI (European Network of Forensic Sciences Institutes).

Logotipos

La ventana de Hamming esta definida de la siguiente manera:

$$w(t) = 0.54 - 0.46 * \cos(2\pi t / L) \qquad \text{en} \qquad 0 \le t \le L-1$$

$$w(t) = 0 \qquad \text{Para el resto de la señal}$$

Dicha ventana si se grafica tiene la siguiente forma:

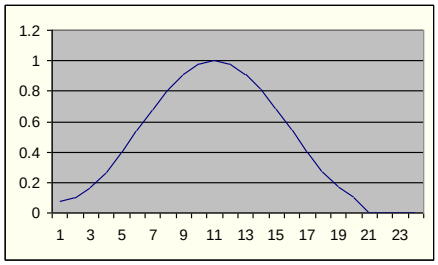


Figura 76: Ventana de Hamming

Y en general podemos decir que lo que hace es adaptar los valores de una señal para que tengan la forma de la figura anterior, esto es, la señal será atenuada en ambos extremos adaptándola o la curva de un coseno, dejando intactos los valores al centro de la ventana, donde la ventana tiene un valor unitario y no modificara los valores originales de la señal, esto con el fin de evitar el efecto de borde o de frontera como se le conoce, esto es, se tendería a falsear un poco los resultados obtenidos por el efecto de los cortes de la ventana rectangular común (no modifica la señal), dando valores poco confiables, por lo que el aplicar una ventana de Hamming y aun mas un solapamiento de ventanas se elimina en buena medida el efecto de frontera, ofreciendo un análisis mas confiable.

Esto se puede apreciar como una ligera diferencia en la forma general de la señal como se pudo ver en las graficas de la herramienta de graficación, que siendo la misma señal se aprecia como se alarga un poco la señal (por el solapamiento de ventanas) contenida en el archivo "ham" y a su vez se tiende a observar las curvas tipo coseno que realiza la señal cuando cambia de color (para diferenciar una ventana de la anterior).

Es también pertinente mencionar que los archivos "ham" tienen un encabezado idéntico al de un archivo "wav", pudiendo incluso renombrar el archivo y escucharlo con un programa común que reproduzca archivos "wav" compatible con el estándar de Microsoft para apreciar el efecto de la ventana en la señal.

Archivos "tie"; estos archivos contienen en si el resultado del procedimiento, esto es, basándose en el archivo "ham" si se aplico Hamming o un archivo "win" o "wav" en su defecto, se procesa cada ventana del archivo para obtener su Magnitud, Media, Energía, Cruces por Cero, Máximos y Mínimos relativos, como ya se explico antes, formándose un archivo como a continuación se detalla.

	Bytes	Descripción
Campo 1	0..3	Numero de ventanas procesadas
Campo 2	4	Numero de archivos procesados
Campo 3	5..8	Posición de inicio de los datos mas uno
Campo 4	9..	Nombres de los archivos procesados

Tabla 11: Encabezado del archivo tie

Como se puede ver el cuarto campo del encabezado tiene una longitud variable, misma que depende de dos factores, el numero de archivos procesados (como se explicara mas adelante) y el largo del nombre del archivo procesado, pero hace falta mencionar que los nombres de los diferentes archivos que conforman un solo lote de procesamiento estarán separados entre si por el código ASSCII 42 (símbolo "*"), que se toma como un carácter invalido para un nombre, así que es perfectamente seguro usarlo como un separador.

Así mismo vemos que los datos variaran también su posición de inicio, por lo que esta se guarda en el campo tres para su fácil acceso, y los datos estarán ordenados de la siguiente manera y en variables del tipo indicado.

Magnitud	4 bytes	Long
Media	4 bytes	Long
Energía	4 bytes	Long

Cruces por Cero	2 bytes	Integer
Máximos	2 bytes	Integer
Mínimos	2 bytes	Integer

Tabla 12: Tipos de las variables

Cabe aclarar que esta designación de variables es valida para Visual Basic 6.0 de Microsoft, pero puede variar en otros lenguajes, como por ejemplo C++ en el cual variables Integer y Long tienen 4 bytes ambas, pero distintas características.

Archivos "nor"; estos archivos son únicamente una normalización de un archivo "tie" para su graficación, pero de hecho estos valores no tienen otro uso, un archivo "nor" tomara los valores máximos de cada uno de los seis parámetros ya calculados y los escalara para que puedan ser graficados en una resolución vertical de 4096 puntos o twips, que es una unidad de medida de Microsoft que es igual a 1/20 parte de un punto de impresión, y aproximadamente 1/567 de un centímetro de la pantalla, esta unidad de medida se usa para que sin importar la resolución de la computadora que se use el resultado sea el mismo en la visualización.

Cabe mencionar que se agrego un 10% por seguridad a la normalización, por lo que ningún grafica llegara al limite de la ventana de graficación, dado que dificultaría su lectura, además otra ventaja es que podemos graficar diferentes parámetros al mismo tiempo sin que importe que tan diferente sea la escala de uno contra el otro, todos se graficaran en la misma escala de puntos verticales para mejor visualización.

La estructura de los datos y el encabezado es idéntica al de un archivo "tie", por lo que no se profundizara en su explicación.

Comparar dos archivos

En este menú se como se aprecia en la figura siguiente, se tienen dos opciones de comparación, "Comparación Analítica" y "Comparación Grafica", además de los respectivos botones "Añadir", para agregar un archivo "tie" de resultados a la lista y "Limpiar" para borrar la lista de archivos.

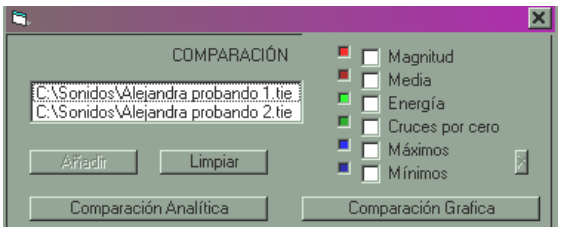


Figura 77: Comparar Dos Archivos

Como una medida de seguridad el sistema verifica que no se de de alta dos veces el mismo archivo, en cuyo caso da una señal de alerta y borra la lista de archivos.

En el caso de la comparación analítica, como se muestra en la figura siguiente, se ve que se da un porcentaje de proximidad por cada parámetro, así como un porcentaje total, siendo importante mencionar que en la forma en que trabaja el sistema, al contrario de cómo solemos interpretar los porcentajes comunes, el 0% representa una extremada similitud en el parámetro y el 100% expresa la mayor distancia posible (la cual nunca se

logra), de la misma manera es necesario explicar que existen tres porcentajes por cada parámetro, variando solamente la forma de obtener dicho porcentaje.

Para entender mejor esto daremos un ejemplo simple de que representa un porcentaje y posteriormente proceder a explicar como se calculan los tres porcentajes ya mencionados. Un porcentaje se define de diccionario como "un tanto por ciento" y representa una proporción, por ejemplo, el numero 40 representa el 40 % de 100, pero el 80 % de 50, por lo que se ve que es una proporción contra otro numero o referencia.

En este caso se tienen tres porcentajes, los que están bajo la columna "Totales" representan la distancia Euclidea contra el máximo rango que las respectivas variables pueden contener, como se explica en la tabla sobre los archivos "tie" previamente presentada, siendo en este caso 65,536 para las variables integer y 4,294,967,296 para las variables long, lo cual es poco representativo, ya que no tiene nada que ver con las voces comparadas, sino con la capacidad numérica de las variables que lo contienen, más sin embargo se conservo solo como un porcentaje mas de comparación a pesar de su poca aportación a la identificación.

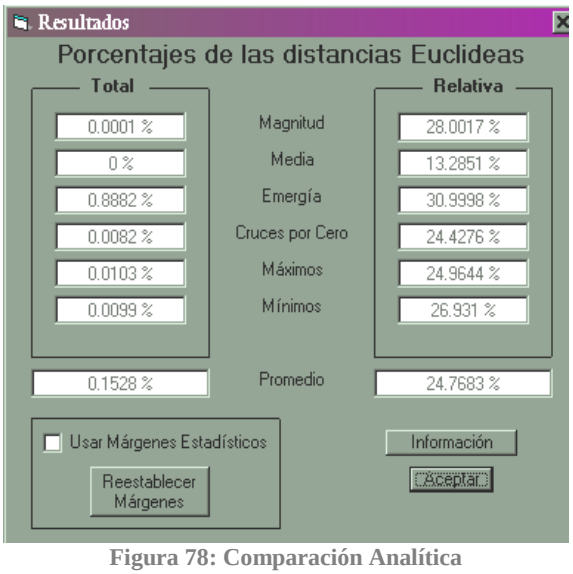


Figura 78: Comparación Analítica

Solo como una nota interesante, el rango máximo de una variable esta dado por el numero de valores distintos que esta puede tomar, por ejemplo, si tomamos una variable de 2 bites sus posibles valores son "00", "01", "10" y "11", por lo que decimos que su rango máximo es 4, o lo que es lo mismo para toda variable, si elevamos la base numérica que es en este caso 2 (binario, que es el lenguaje interno de las computadoras digitales, cero para apagado y uno para prendido), a la potencia del numero de dígitos de la misma variable, en este caso también 2, esto es $2^2 = 4$, mismo que se respeta para todo valor, como $2^{16} = 65{,}536$ en el caso de variables integer, rango que la computadora puede distribuir en que los posibles valores que acepta dicha variable es de -32,768 a 32,767, nótese que existe una diferencia unitaria entre el rango negativo y el positivo, este valor esta reservado para el cero, o en variables sin signo de 0 a 65,536 (referencia en C++).

Para los segundos porcentajes notamos que debajo del cuadro de "Totales" se puede ver un cuadro de selección que dice "Usar márgenes estadísticos", esto es que para los rangos máximos de comparación se usaran los valores contenidos en un archivo que se

encuentra en el directorio fijo de trabajo “C:\Sonidos” bajo el nombre “ACET.cfg”, el cual contiene únicamente los valores máximos que ha alcanzado cualquier archivo procesado en la misma computadora, y como se ve en la pantalla se tiene un botón debajo del cuadro de selección que dice “Restablecer Márgenes”, el cual tiene como función borrar el contenido de dicho archivo y empezar a formarlo de nuevo, que seria como comenzaría a trabajar el sistema recién instalado, por lo que podríamos ver que si es la primera vez que se corre el sistema al seleccionar dicha casilla se modificaran los porcentajes quedando idénticos a los de la columna “Relativa”, pero en caso de que no estén recién borrados cambiaran los porcentajes a algo parecido al de la columna “Relativa” pero nunca igual.

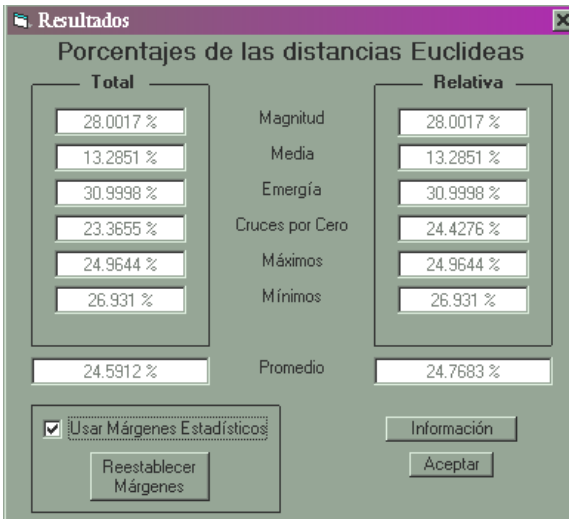


Figura 79: Usar Márgenes Estadísticos

Estos márgenes de obtienen al comparar la distancia Euclidea entre las dos muestras contra la distancia Euclidea máxima que a procesado el sistema, como se explico anteriormente, y de manera muy similar bajo la columna “Relativa” se tiene la distancia Euclidea tomando como referencia únicamente los dos archivos que se estén comparando, que aunque muy parecida a los porcentajes obtenidos usando los márgenes estadísticos se vera que con el uso del sistema los márgenes estadísticos se irán estrechando, obteniendo resultados mas confiables conforme se procesen una mayor cantidad de archivos, por lo que estos terceros márgenes serán de poca utilidad cuando el sistema tenga relativamente pocos archivos procesados.

Por lo ya expuesto se hace claro el hecho de que este estudio es dependiente del texto, ya que comparar dos palabras distintas daría resultados poco representativos, mas sin embargo la valides de este estudio se basa en la suposición de que los parámetros como magnitud, media y energía serán fuertemente dependientes del locutor, por parte de los cruces por cero, máximos y mínimos, se ve que estos son aportativos con cierto tipo de sonidos como los fricativos, por lo que también dependerán del locutor, pero también dependen fuertemente de las condiciones y calidad de la grabación en cuestión, por lo que se ve que este estudio es un tanto exigente en cuanto a que depende fuertemente de la grabación en si y de las palabras que se estén comparando.

También es importante mencionar que si los dos vectores de datos obtenidos tienen un numero de muestras desigual, lo cual es lo mas factible, ya que es difícil que de manera natural se tengan dos muestras de exactamente la misma duración debido a la inter variabilidad intra personal de cada sujeto, por lo que se implemento en una primera

aproximación la acción de completar el vector con el menor numero de muestras al tamaño del mayor rellenando de ceros, para así poder realizar la distancia Euclidea, que como se vio anteriormente es una distancia punto a punto, por lo que se generara un nuevo archivo "tie" al que se le agrega "Com" al final de su nombre, el cual es idéntico al original pero se rellena con ceros al final para igualar su tamaño al del otro archivo en estudio.

Por ultimo, en las dos figuras anteriores vemos un botón de "Información", el cual despliega la siguiente pantalla, que únicamente contiene los archivos que conforman cada archivo "tie", que como ya se menciono puede estar conformado por mas de un archivo, información misma que se obtiene de su encabezado, el cual ya se explico previamente.

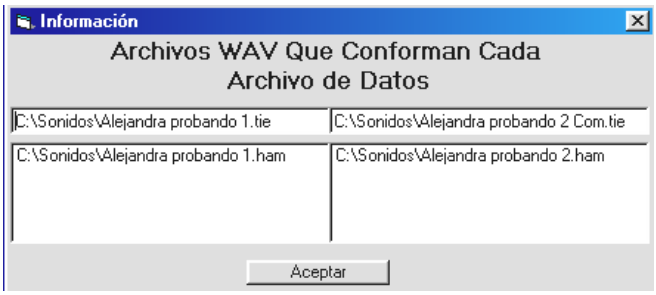


Figura 80: Información del encabezado

Ahora, para la comparación grafica vemos en la figura del comparador de archivos que se tienen las seis casillas de selección para cada parámetro, y que si queremos obtener una grafica comparativa y no hemos seleccionado ningún parámetro recibiremos un aviso de que se debe de seleccionar al menos un parámetro, una vez que seleccionamos un parámetro obtendremos una grafica similar a la siguiente figura, en la que se compara la magnitud de dos muestras.

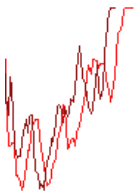


Figura 81: Comparación de medias

De esta figura vemos la extremada semejanza entre dos muestras de voz en uno de sus parámetros, mismo que dio la idea de que estos pueden ser de carácter identificativo, y también es importante mencionar que se tomaran dos colores distintos, uno para cada señal para poder distinguirlas de manera mas clara, pero por ejemplo si tomamos la magnitud y la media y las graficamos al mismo tiempo podríamos obtener dos graficas por cada tono, prestándose a confusión, aunado a que cuando graficamos dos o mas parámetros en una sola pantalla y a pesar de los distintos colores es poco legibles, como se ve en la figura siguiente.

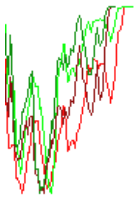


Figura 82: Dos comparaciones simultáneas

Manipulación múltiple

Esta es una solución que nace del uso del sistema así como el estudio de los resultados obtenidos en la comparación de dos archivos, y como su nombre lo dice tiene como función primordial manipular varios archivos simultáneamente de una manera más rápida y fácil.

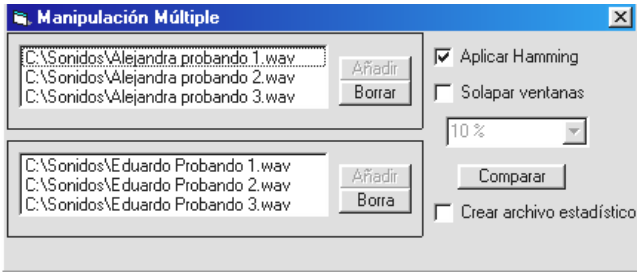


Figura 83: Manipulación Múltiple

En este primer menú que se muestra en la figura vemos que se solicita dar de alta tres archivos en cada una de las listas, esto por supuesto con el botón de "Añadir", que a diferencia de cómo trabajamos comúnmente, al seleccionarlo un primer archivo automáticamente abre de nuevo para seleccionar un segundo y un tercer archivo a procesar, esto para hacer mas rápida la selección, y de manera similar para la segunda lista, contando por supuesto con sus botones de "Borrar" para cada lista para borrar los archivos seleccionados. También contamos con las casillas de selección para aplicar una ventana de Hamming y un solapamiento de ventanas variable, mismo que se toma de la ventana drop down con los distintos porcentajes disponibles.

Una vez que hemos seleccionado tres archivos para cada sujeto, los cuales deberán de ser la misma palabra y con la pronunciación mas similar posible, se puede seleccionar la casilla de comparar, lo cual ocasionara que las dos listas se ordenen de manera que el archivo de mayor duración sea el primero, y así consecutivamente en orden descenderte, para posteriormente aplicarle una interpolación a los archivos menores, esto con la idea de igualar la duración temporal de las tres grabaciones. Esta idea nace del hecho de que las características de engría, media, mediana, cruces por cero, máximos y mínimos no sufrirán cambios al agregar algunos puntos extra siguiendo los patrones de la onda original, esto es, si tomamos dos puntos y agregamos un tercero intermedio con un valor que sea la media de ambos no se compromete la confiabilidad del estudio, sino al contrario se incrementa su certeza, esto debido a que ahora tenemos una mayor similitud en el espaciamiento de las características propias del individuo y esto mejora la medición de distancias Euclideas al tener mejor concordancia de elementos y características, y como ya se dijo anteriormente se agregara al nombre la palabra "Intra" a los archivos interpolados, como se ve en la figura 84.

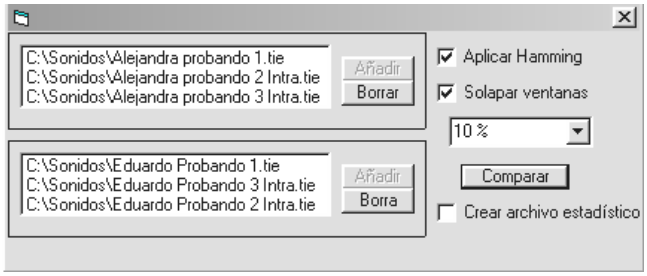


Figura 84: Archivos procesados

Como resultado final de este proceso notaremos que se genera un nuevo archivo "tie", el cual tiene el nombre del primer archivo mas la palabra "Multi" y su correspondiente archivo "nor" para graficación, el cual es la media aritmética de los tres archivos "tie" generados a partir de su respectivo archivo WAV, el cual ya es un archivo que se puede estudiar por las herramientas ya explicadas en "Comparar Dos Archivos".

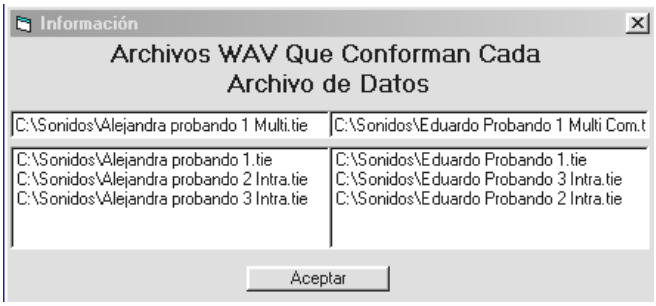


Figura 85: Pantalla de información de una comparación múltiple

Por ultimo y como un pasó mas en la búsqueda de resultados más confiables se agrego el multi procesamiento de diez archivos simultáneos, esto con el objetivo de formar un modelo mas acertado de un sujeto. Este proceso trabaja de manera idéntica a lo explicado anteriormente, pero tomara diez muestras de un mismo sujeto para formar un archivo estadístico, por lo que vemos que este modo se activara al seleccionar la casilla de "Crear archivo estadístico", la cual no se había explicado, modificándose este menú a como lo muestra la figura.

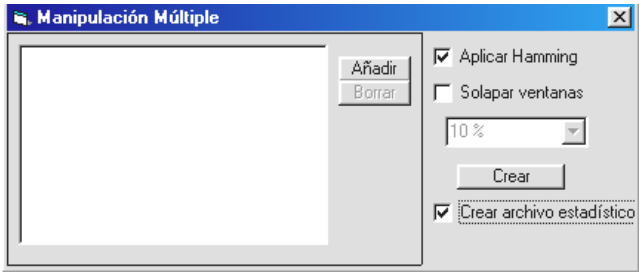


Figura 86: Crear Archivo Estadístico

El resultado final de este proceso será de manera similar a lo ya expuesto un archivo "tie" y otro "nor" con el nombre del primer archivo de la lista mas la palabra "Esta" al final del nombre, mismo archivo que puede ser procesado en el modulo de "Comparar Dos Archivos", como ya se discutió anteriormente, pero con mejores resultados, ya que ahora tomamos una media aritmética de diez archivos, ofreciendo una base firme y confiable para este tipo de comparación.

Identificación por Medio de Formantes

Para esta segunda aproximación se pretende poner a prueba la hipótesis de que las primeras dos formantes y el tono fundamental, que son mejor visibles en el dominio de la frecuencia sirven para lograr la identificación de un sujeto, esto en el entendimiento de que dichas formantes estarán presentes en todos los sonidos vocálicos, y presentan una evolución y estabilización uniforme y constante para un mismo individuo.

Esta función se encuentra en la pantalla principal del sistema bajo el menú "Fourier", que ofrece a su vez dos opciones, que son "Fourier" y "Análisis Estadístico", los cuales iremos explicando en el orden que se considera mas oportuno para un optimo entendimiento de su funcionalidad.

Fourier

Para este menú tendremos en si la transformación al dominio de la frecuencia o transformada de Fourier, por dominio de la frecuencia se entiende que en lugar de tener como variable libre o independiente al tiempo tendremos como variable independiente a la frecuencia, por lo que podremos fácilmente obtener información con respecto a la frecuencia y a sus componentes, como ya se a explicado en la transformada de Fourier y los espectrogramas anteriormente.

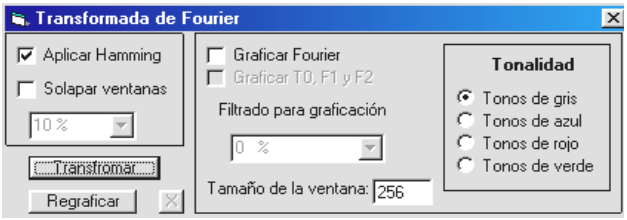


Figura 87: Transformada de Fourier

Para entender bien como trabaja este menú explicaremos cada una de sus opciones y su funcionalidad, comenzando por que se tienen las casillas de selección para aplicar una ventana de Hamming y un solapamiento de ventanas variables con su menú drop down, como ya se ha visto en múltiples ocasiones, pero también tenemos la casilla de "Graficar Fourier" la cual tiene como función llamar a la visualización del espectrograma correspondiente una vez que se a procesado el archivo, al activar esta casilla se activara a su vez la casilla "Graficar T_0, F_1 y F_2, lo cual tiene como función el que sobre el espectrograma ya obtenido se graficaran en distintos colores lo que la computadora a identificado como el tono fundamental (T_0), el cual además se informa su promedio numérico en Hertz al final de la operación, y las primeras dos formantes.

En este mismo apartado vemos también que están dos opciones mas, que son "Filtrado para graficación" el cual establecerá un filtro en porcentajes que van del 0 % al 5 % en intervalos de 0.5, esto es 0, 0.5, 1…, lo cual únicamente tiene como fin tratar de reducir la cantidad de ruido visual que se genera al transformar en Fourier, esto es, sonidos con muy poco peso espectral y por lo tanto muy poca o ninguna aportación para fines identificativos, lo cual se ve como tonos bajos de color en el espectrograma (grises tenues por ejemplo), y que dificultan su lectura, pero no modifica en si la transformación numérica.

Por ultimo tenemos también lo que es "Tamaño de la ventana" que nos da la opción de cambiar el tamaño de las ventanas que se procesaran al pasar al dominio de la frecuencia, lo cual también se sugiere en la mayoría de las bibliografías en un tamaño de

256 muestras (default) o 512 muestras por ventana, ya que esto eleva la resolución espectral, permitiéndonos tomar rangos de frecuencia mas pequeños y por lo tanto cálculos de mejor calidad, valores mas grandes que esto ya no ayudan, sino al contrario pierden información, ya que son demasiadas muestras en muy pocas ventanas, tendiéndose a confundir o mezclar fonemas, dando resultados poco confiables, y valores menores tampoco son recomendados, ya que en general se tiene una pobre resolución espectral, no permitiendo una adecuada identificación de los fonemas o el tono fundamental.

Para la localización de las formantes y del tono fundamental se utilizan los valores internacionalmente reconocidos para su ubicación en el espectro, siendo estos de 60 a 250 Hz para el tono fundamental, de 250 a 900 Hz para la primera formante y de 1500 a 2500 Hz para la segunda formante, pero para la localización de los formantes se toman en cuenta dos aspectos mas que valen la pena de mencionar, el primero es que el concepto de formante nos dice que esta será el máximo local dentro del rango en que esta se presenta y estará conformada desde la cúspide y hasta 3 dB por debajo de la cima, lo cual representa el 70 por ciento de la señal en cuestión, y además tenemos otro concepto que nos ayuda a localizarlas, y esto es que las formantes serán armónicas del tono fundamental o lo que es lo mismo el tono fundamental multiplicado por un numero entero.

Así que como vemos la metodología es simple, se localiza el tono fundamental que es el máximo local en el intervalo de 60 a 250 Hz, dependiendo de la voz en estudio, en base a esto se obtienen sus armónicos y se conservan temporalmente, para después buscar los máximos en los rangos de las formantes y que sean armónicos del tono fundamental, para a partir de esto buscar la señal hasta donde se pierde por debajo de los 3 dB, por esta razón el tono fundamental se graficara como una línea continua y las formantes como líneas discontinuas, ya que recordemos que estas formantes solo estarán presentes en sonidos vocálicos, por lo que seria un error que fueran continuas.

Adicionalmente hay que explicar las opciones que tenemos para las tonalidades, opción que solo estará disponible si no graficamos T_0, F_1 y F_2, ya que si se selecciona esta opción de graficación se desactivan las tonalidades dejando únicamente tonos de gris para una mejor visualización del tono fundamental y las formantes que se grafican en colores, pero de no ser así se nos da la opción de ver el espectrograma en tonos de azul, rojo y verde además de tonos de gris, lo cual es con la finalidad de brindar una mejor visualización y localización de estructuras de forma manual.

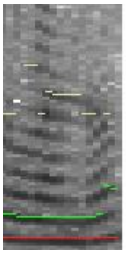


Figura 88: Espectrograma con T0, F1 y F2

Resultados parciales y finales

Como podemos observar en el directorio de trabajo “C:\Sonidos” (que previamente se menciono es el directorio fijo de trabajo), se observan varios archivos nuevos, que son

los resultados parciales y finales de las distintas opciones de análisis que se implementaron en este modulo del sistema, a continuación se explica el formato del encabezado de dichos archivos, su función y el formato de los datos guardados en cada uno de ellos.

Archivos "win"; estos archivos contienen el resultado de la aplicación de un solapamiento de ventanas como ya se explico anteriormente.

Archivos "ham"; estos archivos contienen el resultado de la aplicación de una ventana de Hamming a partir de un archivo "win" si se aplico solapamiento de ventanas o de un archivo "wav" en su defecto como ya se explico anteriormente.

Archivos "tmp"; estos archivos contienen el resultado de la transformada de Fourier de un archivo "win", "ham" o "wav" respectivamente según la selección deseada, o sea un recuento de las componentes de la señal de acuerdo a las frecuencias o lo que es lo mismo una descomposición de la señal en sus componentes principales.

Aunado a lo anterior es conveniente mencionar que los resultados de una transformación de Fourier da resultados extremadamente dispares o con una gran desviación estándar (distancia promedio de los datos con respecto a la media), esto es que tenemos números extremadamente pequeños y números extremadamente grandes, por lo que es poco representativo el comparar un numero alto con uno pequeño, por lo que es común que se aplique un normalización, en este caso por medio de una escala logarítmica para que disminuya la desviación estándar de los datos sin afectar la calidad de su contenido.

Este tipo de archivos son como su nombre lo indica temporales, por lo que solo veremos su aparición por un corto periodo de tiempo, mismo que será procesado y borrado posteriormente, ya que no tiene utilidad alguna por su elevada desviación estándar como ya se menciono.

El encabezado de estos archivos es como a continuación se detalla.

	Bytes	Descripción
Campo 1	0..3	Numero de ventanas procesadas
Campo 2	4	Numero de archivos procesados
Campo 3	5..6	La mitad del tamaño de la ventana procesada
Campo 4	7..10	Posición de inicio de los datos mas uno
Campo 5	11..	Nombres de los archivos procesados

Tabla 13: Encabezado del archivo tmp

Y los datos estarán almacenados de manera continua en variables de tipo double (8 bytes por dato), esto es, se tienen el numero de datos del campo 3 por ventana y tantas ventanas como se tienen en el campo 1, lo que nos deja ver que el resultado de esta transformación tiene una resolución igual a la mitad del tamaño de la ventana, por lo que se ve mas claro el por que es importante tener una ventana de un tamaño mas grande al utilizado en el procesamiento en el tiempo, pero no tan grande como para que se tengan varios fonemas en una sola ventana.

Para entender mejor el concepto de resolución tomaremos por ejemplo un archivo muestreado a 8000 Hz (filtro de la línea telefónica) y una ventana de 256 muestras (default del sistema) tendremos que cada dato representa aproximadamente 31 Hz, por lo que si buscamos un máximo en el rango del tono fundamental (60 a 250 Hz) tendremos 8 datos para buscarlo, en cambio si comparamos contra el tamaño de ventana de 80 (la que se usaría en análisis en el tiempo) en que tenemos que cada dato representa 100 Hz solo tendremos 2 datos para procesar, lo cual seria muy poco representativo.

Archivos "fou"; estos archivos contienen el resultado de la aplicación de una escala logarítmica a un archivo "tmp", el cual tiene el mismo encabezado y la misma estructura de datos, pero este archivo es apto para su estudio y procesamiento ya que tiene una desviación estándar mucho mas aceptable y por lo tanto mas propicio para su estudio.

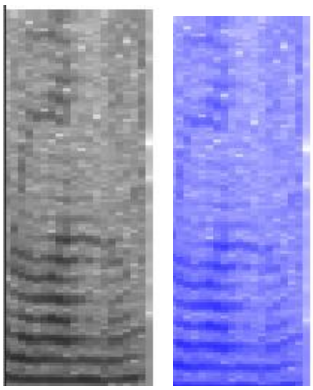


Figura 89: Espectro de una señal

Recordemos que como se ve en la figura (el mismo espectro en tonos de gris y de azul) las componentes principales son los tonos de gris (o de color) de mayor intensidad, así como sus respectivas armónicas , esto se ve como tonos de menor intensidad, lo cual es mas bien ruido y no aporta a la identificación de un locutor, razón por la cual se implementa un filtro para graficar, el cual filtra o disminuye estas componentes extra para ver de manera mas clara las formantes y el tono fundamental.

Archivos "nof"; estos archivos contienen también el resultado de la aplicación de una escala logarítmica a un archivo "tmp", el cual tiene el mismo encabezado y la misma estructura de datos, pero este archivo es apto para su graficación, ya que se puede aplicar además de la escala logarítmica un filtro, el cual tiene como finalidad la reducción del ruido visual que presenta un espectrograma, como ya se comento, además de que se fija el rango de variación máximo de 0 a 255 (256 valores o tonos), esto debido a que este sistema maneja 256 tonalidades de gris, azul, rojo o verde según se seleccione la graficación.

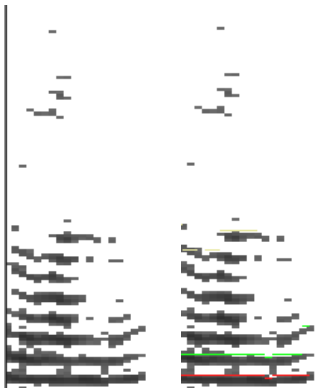


Figura 90: Espectrograma filtrado al 3%

Como se puede ver en la figura, que es el mismo espectrograma de la figura anterior pero en esta ocasión se remueve el ruido espectral por medio de la aplicación de un filtro del 3% lo cual nos deja apreciar de manera mas clara las estructuras formanticas, aunque existe el riesgo de que un filtrado excesivo pierda parte de la información importante de una muestra de voz y dificulte o impida la identificación, pero también es importante mencionar que este filtrado es únicamente para fines de visualización del espectrograma, no afectando los cálculos del tono fundamental y formantes que el sistema hace de manera automática.

Archivos "tff"; estos archivos contienen el resultado final de este modulo del sistema, contienen una lista de los valores del tono fundamental y de las formantes en cada una de las ventanas que contiene el archivo "fou" que se explico anteriormente, recordemos que debido a que en algunas muestras puede ser que no este presente el tono fundamental o que no tenga alguna de las primeras dos formantes el sistema asignara un valor de 0 Hz a los puntos en los que no se puede determinar el parámetro, esto para que se tenga una fácil manera de localizar en donde están presentes dichos parámetros para su posterior comparación contra otra muestra de voz.

El encabezado de estos archivos es como a continuación se detalla.

	Bytes	Descripción
Campo 1	0..3	Numero de ventanas procesadas
Campo 2	4..7	Posición de inicio de los datos mas uno
Campo 3	8..	Nombres de los archivos procesados

Tabla 14: Encabezado del archivo tff

Por lo que respecta a los datos estos estarán ordenados grabándose primero el tono fundamental, la primara formante y la segunda formante, todos en variables de tipo long (4 bytes por dato), y existirán tantos juegos de datos como se especifica en el campo 1.

Recordemos que para la localización de las formantes se toman en cuenta tres parámetros, y esto es que sean el máximo relativo dentro del rango en que comúnmente se encuentran, el que sean un armónico del tono fundamental previamente obtenido y por ultimo que no caigan por debajo de 3 dB del máximo identificado como la cima de la estructura formantica.

Análisis estadístico

En esta opción se puede llevar a cabo la comparación de dos archivos previamente procesados en Fourier, y como se puede ver en la figura siguiente tenemos el botón de "Agregar" para seleccionar los archivos a comparar, que de manera automática pedirá un segundo archivo para su comparación, verificando que no se de de alta dos veces el mismo archivo, cambiando después su función a "Calcular", función que nos dará el resultado de los cálculos de comparación entre el primer y el segundo archivo.

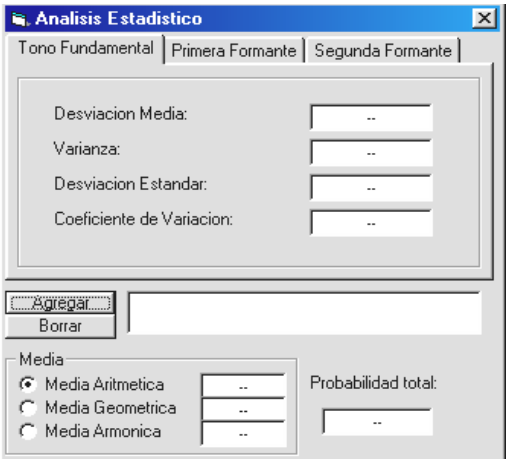


Figura 91: Análisis Estadístico

Es importante hacer notar que se tienen tres opciones de selección para el parámetro de la media, esto es debido a que para los cuatro parámetros que se obtienen de cada formante y tono fundamental, que son, la desviación media, la varianza, la desviación estándar y el coeficiente de variación, es necesaria la media para su calculo, y al tener disponibles tres tipos de media variaran todos los parámetros de acuerdo a la media seleccionada, pero se calcularan y mostraran las tres medias, para que el usuario pueda decidir en una futura comparación entre la que resulte mas adecuada.

En cuanto a las distintas medias estas se definen de la siguiente manera, y así mismo se dará un ejemplo de su cálculo para que el usuario pueda ver cual considera mas adecuada.

Media aritmética: (la más común)

$$\bar{x} = \frac{\sum x_j}{n} \qquad (42)$$

Media geométrica:

$$x = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n} \qquad (43)$$

Media armónica:

$$\bar{x} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}} \qquad (44)$$

A manera de ejemplo tomemos la siguiente secuencia numérica, x = {2, 4, 6, 8, 3, 7, 6, 2, 4, 9, 3, 1}, entonces tendremos que su media aritmética es $\bar{x} = \frac{55}{12} = 4.5833$, su media

geométrica es $x = \sqrt[12]{10450944} = 3.8453$ y por ultimo su media armónica es $x = \frac{12}{3.8789} = 3.0936$, por lo que se puede ver que es significativamente distinto un parámetro del otro, pudiendo dar también variantes considerables en los cálculos que involucran a la media,

Una vez que hemos presionado el botón de "Calcular" se desplegaran los resultados como se aprecia en la figura siguiente, en donde vemos que tenemos tres pestañas en la parte superior, las cuales están etiquetadas "Tono Fundamental", "Primera Formante" y "Segunda Formante", y en cada una de ellas podemos observar los cuatro parámetros antes mencionados y que a continuación se describen a detalle, también podemos observar que el botón "Calcular" a cambia de nuevo a "Recalcular" lo cual nos permite seleccionar una media distinta y ver como modifica los resultados.

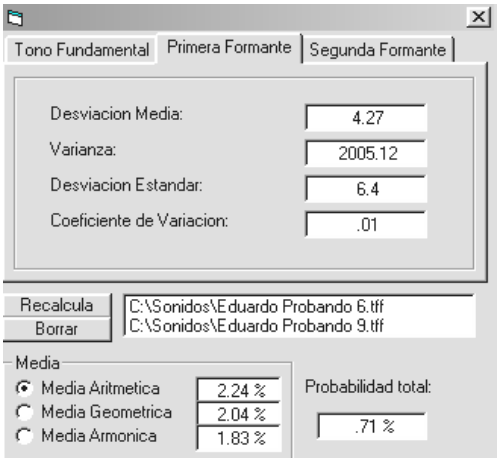


Figura 92: Resultados

Por otra parte, en la parte baja de la pantalla vemos que están las tres medias ya mencionadas expresadas como el porcentaje de la distancia Euclidea, las cuales también cambiaran al seleccionar las distintas pestañas y vemos una "Probabilidad total", la cual es una media aritmética de las distintas medias disponibles, siendo en si el parámetro de mas fácil interpretación para la identificación de un locutor, el resto de los resultados obtenidos para cada pestaña son para un análisis manual por parte del usuario en caso de desearlo así, datos que nos dan la distancia Euclidea de cada parámetro entre las dos muestras de voz, y donde se ve que mientras mas pequeña sea la diferencia mas probabilidad existe de que se trate del mismo sujeto en estudio.

Por ultimo mencionemos que el botón "Borrar" sirve a una doble función, si se han seleccionado dos archivos y se quiere modificar la selección al presionar este botón se borran los archivos previamente seleccionados, regresando el anterior botón a su función de "Agregar" y por otra parte también nos sirve para que una vez que hemos analizado dos archivos podemos borrar toda la forma y seleccionar otros archivos que deseamos comparar.

A continuación y para cerrar ya este modulo del sistema se explicaran los cuatro parámetros que se calculan en este modulo y se explicara su aportación a la posible identificación de un sujeto en estudio.

Medidas de dispersión

Estos son parámetros estadísticos que miden como se distribuyen una serie de datos, los más utilizados se refieren al grado de lejanía de los datos con respecto a la media.

Desviación Media: es el promedio de los valores absolutos de las desviaciones $|x_j - \bar{x}|$ de cada elemento con respecto a su media.

$$DM = \sum \frac{|x_j - \bar{x}|}{n} \qquad (45)$$

Varianza: es el promedio del cuadrado de las desviaciones $(x_j - \bar{x})^2$ de cada elemento con respecto a su media.

$$v = \sum \frac{(x_j - \bar{x})^2}{n} \qquad (46)$$

Desviación Estándar: o desviación típica es la raíz cuadrada de la varianza.

$$\sigma = \sqrt{v} \qquad (47)$$

Coeficiente de Variación: es el cociente entre la desviación típica o estándar y la media de la distribución.

$$CV = \frac{\sigma}{\bar{x}} \qquad (48)$$

Este ultimo parámetro sirve para relativizar el valor de la desviación estándar y así poder comparar la dispersión de dos poblaciones estadísticas con gamas de valores muy discretos, y a pesar de que en este caso no se cumple o por lo menos no se debe de cumplir o necesitar este parámetro se calcula para tener una mejor idea de que tan discreta es la gama de las poblaciones estudiadas.

Como se ve de las definiciones mismas, a excepción del coeficiente de variación se obtiene con respecto a la media de las formantes la desviación propia de cada sujeto, que debe ser muy pequeña para el tono fundamental y relativamente pequeña para las dos formantes calculadas, para después obtener la distancia entre las dos muestras de voz y así determinar una probabilidad de pertenencia o de certidumbre de que se trata de un mismo sujeto o no.

Identificación por Parámetros LPC

Para esta tercera y ultima aproximación se pretende poner a prueba la hipótesis de que por medio de un juego de parámetros LPC (linear prediction code o código de predicción lineal), que establecen un modelo del aparato fonador sirven para lograr la identificación de un sujeto o el reconocimiento del habla pudiendo incluso ser independiente del texto si se usa la aproximación adecuada, misma que no esta dentro del alcance del presente trabajo tanto por su complejidad como porque no seria representativo contra los dos métodos anteriores que son dependientes del texto.

Esta función se encuentra en la pantalla principal del sistema bajo el menú “LPC”, que ofrece a su vez dos opciones, que son “Modelo LPC” y “Comparación de Modelos”, los cuales iremos explicando en el orden que se considera mas oportuno para un optimo entendimiento de su funcionalidad.

Modelo LPC

Para este menú tenemos en si la obtención del modelo del aparato fonador de un sujeto basado en un archivo previamente digitalizado, el sistema implementa el método de la celosía, ya que este método no requiere del calculo de la matriz de correlación, mismo que resulta en un sistema mas eficiente y rápido para la obtención de resultados, pero también cabe mencionar que dichos resultados son coincidentes con los del método de auto-correlación, pero a diferencia de este se garantiza que se obtiene siempre un modelo estable.

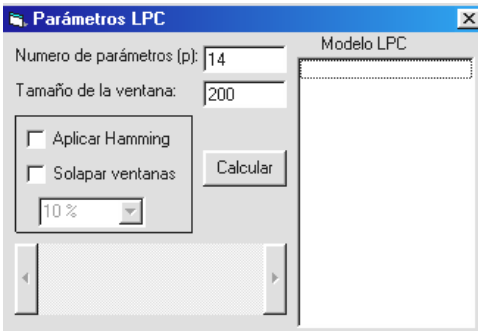


Figura 93: Obtención del modelo LPC

Es conveniente comentar que por su formulación el método de la celosía no requiere de la aplicación de una ventana a los datos, aun así se ofrece la opción de aplicar una ventana de Hamming, así como el solapamiento de ventanas como se ha venido ofreciendo en el resto de las metodologías anteriormente empleadas, mas sin embargo no es recomendado su uso, por lo que por default no se utilizan, por otra parte tenemos dos parámetros que suelen ser establecidos por el usuario para este análisis, siendo estos el parámetro ‘p’ que especifica el numero de coeficientes del modelo, donde se recomienda tener un coeficiente por kilo Hertz de audio mas 3 o 4 para modelar la fuente de excitación, por lo que recomendamos un total 14 para audio de 8000 Hz, pudiendo variarse hasta un máximo de 19 (por cuestión de trabajo interno del sistema y de la capacidad de direccionamiento de memoria de Visual Basic 6), así mismo tenemos el tamaño de la ventana que se recomienda de 80 a 320 para una señal de 8000 Hz, utilizándose el punto medio de 200 como recomendación general.

Como podemos ver de la simplicidad de la pantalla misma este tipo de modelado es sumamente eficiente para usos automatizados, pues además de los pocos parámetros que se pueden variar, los cambios en los resultados no son muy importantes al variar los parámetros, por lo que al presionar el botón de “Calcular” solo se pedirá el nombre del archivo y el resto del proceso es automático.

Al culminar este proceso de obtención del modelo, se desactivaran las opciones de los parámetros y el botón de “Calcular”, pero se activara la barra deslizante de la parte baja y aparecerá en el recuadro “Modelo LPC” los coeficientes obtenidos, que recordemos será un juego de ‘p’ coeficientes por cada ventana de ‘x’ muestras (200 por default) y con la barra deslizante podremos ir recorriendo todos los parámetros obtenidos para la

porción de audio en estudio, quedándonos solo la opción de cerrar este menú para regresar al menú principal del sistema.

Resultados parciales y finales

Como podemos observar en el directorio de trabajo "C:\Sonidos" (que previamente se menciono es el directorio fijo de trabajo), se observan varios archivos nuevos, que son los resultados parciales y finales de las distintas opciones de análisis que se implementaron en este modulo del sistema, a continuación se explica el formato del encabezado de dichos archivos, su función y el formato de los datos guardados en cada uno de ellos.

Archivos "win"; estos archivos contienen el resultado de la aplicación de un solapamiento de ventanas como ya se explico anteriormente.

Archivos "ham"; estos archivos contienen el resultado de la aplicación de una ventana de Hamming a partir de un archivo "win" si se aplico solapamiento de ventanas o de un archivo "wav" en su defecto como ya se explico anteriormente.

Archivos "lpc"; estos archivos contienen el resultado del modelo LPC de un archivo "win", "ham" o "wav" respectivamente según la selección deseada, esto como se explico son los coeficientes que pueden generar un modelo alternativo para análisis espectral, síntesis de voz o para su estudio y comparación contra otros modelos.

El encabezado de estos archivos es como a continuación se detalla.

	Bytes	Descripción
Campo 1	0..3	Numero de ventanas procesadas
Campo 2	4	Numero de archivos procesados
Campo 3	5	El numero de coeficientes 'p'
Campo 4	6..9	Posición de inicio de los datos mas uno
Campo 5	10..	Nombres de los archivos procesados

Tabla 15: Encabezado del archivo lpc

Y los datos estarán almacenados de manera continua en variables de tipo double (8 bytes por dato), esto es, se tienen el numero de datos del campo 3 por ventana y tantas ventanas como se tienen en el campo 1, el tamaño de la ventana utilizada no tiene importancia una vez que se tienen los resultados, y no existe regla alguna que diga que se invalida el resultado si comparamos dos archivos con diferentes tamaños de ventana respectivamente, pero por sentido común esto no es conveniente, pues estamos forzando los resultados y eso no es representativo, pudiendo incrementar el error o falsear un resultado.

Comparación de modelos

En lo que respecta a esta opción vemos que este menú es también muy simple, en la parte superior vemos los botones de "Agregar" y "Borrar" seguidos de una ventana que muestra los dos archivos a comparar, al presionar el botón de "Agregar" se nos pedirá seleccionar dos archivos con extensión "lpc" mismos que serán comparados, el sistema verifica que no se intente comparar un archivo consigo mismo y que el numero de

coeficientes ‘p’ sea el mismo en los dos archivos, si no se cumple alguna de estas condiciones el sistema dará un error y borrar la selección de manera automática, por otra parte al presionar el botón “Borrar” este limpiara la lista de archivos y dejara la ventana lista para seleccionar dos archivos de nuevo.

Figura 94: Comparación de dos modelos

La ventana de número de parámetros es solo informativa y nos dice el numero de coeficientes que se compararan, una vez que tenemos seleccionados dos archivos podemos presionar el botón de “Comparar” que realizara el calculo de la distancia Euclidea entre los dos modelos en cuestión, ofreciéndonos en la ventana grande la distancia normalizada entre cada uno de los ‘p’ coeficientes y una distancia promedio de todos los coeficientes, cabe hacer mención que el análisis de dichas distancias se establecerá después en el trabajo cuando estudiemos de forma comparativa los resultados entre la población de este trabajo, por ahora nos concretamos a mostrar los resultados para dos archivos de prueba para tener una mejor idea de que es lo que nos ofrece esta comparación.

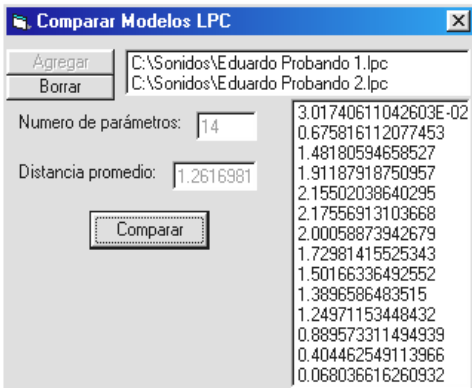


Figura 95: Resultados de una comparación

Resultados

En este apartado se probara el sistema informático que se propuso anteriormente, el cual detalla los procedimientos y técnicas utilizadas, junto con la descripción de su funcionamiento y opciones de trabajo.

Como hemos podido notar hasta este punto no se ha prestado gran atención a los resultados, esto se debe en gran medida a que es necesario una serie de pruebas de funcionamiento aunado a una evaluación estadística de los resultados en miras de obtener los rangos en los que se puede medir cada uno de los parámetros estudiados por el sistema, ya que recordemos que al estar evaluando una señal biológica esta presenta variaciones importantes entre una muestra y la otra, aun que se trate del mismo sujeto, por lo tanto se deberán de obtener los campos de dispersión y obtener los rangos en los que es normal que varié, pudiéndose así finalmente evaluar la eficiencia o utilidad de cada uno de los tres métodos propuestos y desarrollados anteriormente.

Recordemos que un diagrama o campo de dispersión se obtiene al graficar sobre dos ejes coordenados una señal bidimensional y que a cada elemento de esta distribución bidimensional (conjunto de datos en la que intervienen dos variables ‘x’ y ‘y’) le corresponde su ordenada y su abscisa, esta nube que se formara es lo que se conoce como diagrama o campo de dispersión, existiendo varios métodos estadísticos que nos permiten evaluar su comportamiento.

Las medidas de dispersión son los parámetros estadísticos que miden como están diseminados los datos de una distribución, siendo los mas comunes los que se refieren al grado de lejanía de los datos con respecto de su media siendo estos la desviación media, la varianza y la desviación estándar, estos parámetros ya fueron explicados con anterioridad, por lo que no se detalla de nuevo la forma de su obtención.

Condiciones de trabajo y equipos

En este apartado se pretende cubrir aspectos tales como las metodologías de prueba para la evaluación de sistemas de reconocimiento de locutor y reconocimiento de habla, así como aspectos relevantes que intervienen en la evaluación del presente sistema, como es el equipo utilizado para las grabaciones, las condiciones de grabación y demás aspectos importantes que se irán desarrollando.

Cuerpo de evaluación para sistemas de reconocimiento del locutor

Uno de los aspectos relevantes cuando se evalúa un sistema de este tipo es el poder controlar el numero de variables de un experimento al otro, y sobre todo es de suma importancia el poder contar con un juego estándar de grabaciones de evaluación, ya que de esta manera en diferentes partes del mundo o diferentes personas podrán comparar resultados y medir la eficiencia de una aproximación contra otra metodología distinta.

Para los sistemas de habla inglesa o de otras lenguas extranjeras (varios idiomas del continente Europeo) existen una variedad de cuerpos de prueba estándar que incorporan diferentes características tanto de grabación, ancho de banda y numero de sujetos de prueba y que los hacen adecuados para evaluar y compartir resultados entre investigadores de estos países, entre ellos podemos mencionar cuerpos tales como:

TIMIT; Uno de los primeros cuerpos de prueba que se creo con un gran número de sujetos de prueba, pero hoy en día poco utilizado debido a sus condiciones de grabación poco realistas, ya que se realizan las grabaciones en una cabina acústica y con un gran ancho de banda, pero es base para otra serie de cuerpos como FFMTIMIT (igual a TIMIT pero con un micrófono lejano), NTIMIT (mismas grabaciones transmitidas a través de líneas telefónicas reales locales y de larga distancia), CTIMIT (mismas grabaciones transmitidas a través de telefonía celular), etc.

Numero de locutores	630 (438 H/192 M)
Numero de sesiones por locutor	1
Intervalo entre sesiones	Ninguno
Tipo de habla	Lectura de sentencias
Micrófonos	De diadema con ancho de banda amplio
Canales	De gran ancho de banda
Ambiente acústico	Cabina de grabación
Cuenta con proceso de evaluación	Si

Cuerpo TIMIT

SIVA; Cuerpo Italiano obtenido a través de diferentes aparatos telefónicos, se implemento una contestadora telefónica con un numero gratuito al que la gente podía llamar y seguir unas simples instrucciones para realizar una grabación de 28 palabras, algunas preguntas y lectura de texto.

Numero de locutores	671 (335 H/336 M)
Numero de sesiones por locutor	1 a 26
Intervalo entre sesiones	Días o meses
Tipo de habla	Lectura de sentencias, palabras y preguntas cortas
Micrófonos	Diversos aparatos telefónicos
Canales	PSNT
Ambiente acústico	Caso o oficinas
Cuenta con proceso de evaluación	Si

Cuerpo SIVA

De esta manera podríamos seguir enumerando diversos cuerpos como PolyVar, POLYCOST, KING, YOHO, Switchboard I-II, SPIDRE, NIST, OGI y otros más, pero no es muy relevante ya que al no localizar con un cuerpo en el idioma Español de libre acceso no nos son de utilidad, sino simplemente nos sirven como antecedente para el que se usara en este trabajo.

Como se puede ver no es simple el tener un cuerpo adecuado para la evaluación de un sistema de identificación de locutores, además de que podría considerarse un trabajo digno de una tesis el implementar un cuerpo de prueba tan completo como los que se han mencionado, sin mencionar el costo y tiempo para llevar a cabo algo como esto, mas sin embargo se implementara un cuerpo de prueba mas simple, pero tratando de conservar las características necesarias para una honesta evaluación del presente sistema.

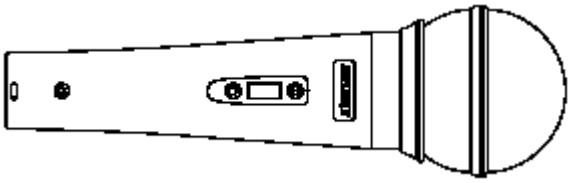
Formación del cuerpo de evaluación

Como un primer aspecto que salta a la mente es el numero de locutores que se usaran para las presentes pruebas, que como es fácil ver, mientras mas amplio mas tiempo requiere para su formación representando también un mayor gasto, por lo que a pesar de que seria mas útil un numero elevado (comparativamente 51 el mas pequeño, KING y 671 SIVA, el mas grande) en el presente trabajo se utilizara un total de 20 sujetos, 10 hombres y 10 mujeres para formar nuestro cuerpo de evaluación.

También es importante mencionar que al no tener nuestro sistema una etapa que tome una decisión final de la identificación, sino tan solo una serie de medidas que orientaran el perito en la materia, no cuenta con una etapa de entrenamiento, lo cual hace que el número de 20 locutores sea escaso pero suficiente para las pruebas.

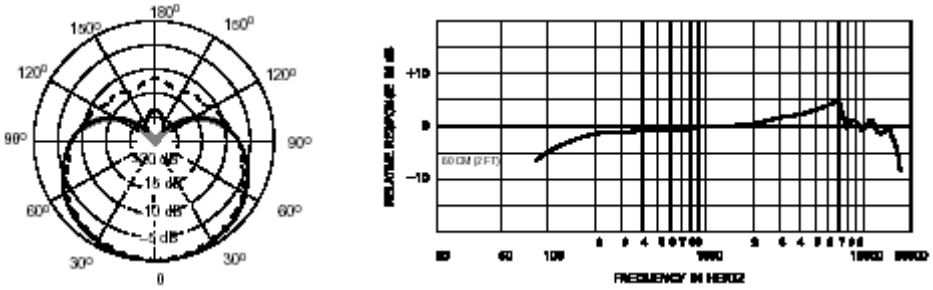
El número de sesiones por locutor será de una sola, ya que de lo contrario elevaría el tiempo de formación, se tomaran 10 muestras de tres palabras que posteriormente se seleccionaran para cada locutor, dando un total de 30 palabras aisladas, ya que el presente sistema no contempla habla continua.

El micrófono utilizado para llevar a cabo las grabaciones es de la marca Shure modelo 12A que es un micrófono dinámico con un patrón de captación cardioide como el que se ilustra a continuación y con las siguientes características:



Micrófono Shure 12A

Tipo:	Dinámico
Patrón de captación:	Cardioide (ver figura 109)
Respuesta en frecuencia:	80-14000 Hz (ver figura 110)
Impedancia nominal:	150Ω
Nivel de salida:	56 dBV/Pa (a 1 Khz.)
Conector:	XLR



Patrón de captación y respuesta en frecuencia

En cuento al canal de transmisión o grabación hay que tener en cuenta que el presente sistema esta orientado a aplicaciones forenses, mismas que en su mayoría por no decir en su totalidad son grabaciones realizadas en condiciones muy distintas a las ideales de

laboratorio, estas serán grabaciones de conversaciones telefónicas en su mayoría, y con ruidos de oficina o casa comúnmente, por lo que se elige para el presente cuerpo un ancho de banda de 8000 Hz que es suficiente para estas aplicaciones, así mismo se llevaran a cabo las grabaciones de manera directa en una computadora personal IBM modelo 300PL multimedia con procesador Pentium III a 500 MHz, 64 MB de memoria RAM y una tarjeta de sonido Audio Crystal WDM.

La tarjeta Audio Crystal WDM con chip set CS4235-KQ cuenta con las siguientes características; canales de 18 bits para conversión analógica digital y digital analógica, rango de frecuencias de 8,000 a 44,000 Hz en modo estéreo o monoaural, para este cuerpo se uso un canal de 16 bits con un ancho de banda de 8000 Hz en modo monoaural.

En lo que respecta al ambiente acústico este fue seleccionado en un nivel bajo de ruido de oficina, ya que como se ha mencionado es común tener un cierto ruido ambiental en las grabaciones de índole forense, como puede ser ruido de oficinas, casas o de teléfonos públicos, mas sin embargo no es incumbencia del presente trabajo el trabajar con ambientes sumamente ruidosos o en grabaciones en que intervengan múltiples voces, por lo que se trabaja con una sola voz y en un ruido ligero de oficina.

Numero de locutores	20 (10 H/10 M)
Numero de sesiones por locutor	1
Intervalo entre sesiones	Ninguna
Tipo de habla	Palabras aisladas
Micrófonos	Shure 12A
Canales	16 bits a 8000 Hz
Ambiente acústico	Oficina
Cuenta con proceso de evaluación	No

Cuerpo RVIL (Reconocimiento de Voz: Identificación de Locutores)

Para la digitalización correspondiente de las palabras usadas en este cuerpo se utilizo el software comercial Sound Forge 8.0 de la compañía Sony, esto en razón de múltiples ventajas que presenta este programa con respecto a sus similares, sobre todo en la formación de filtros sumamente poderosos con fines de limpieza de señales de audio, aislamiento de pociones útiles y demás características que lo hacen un favorito dentro de los diversos laboratorios de Acústica Forense a nivel mundial.

Selección de las palabras para el cuerpo RVIL

En lo que respecta al criterio de selección de las tres palabras que formaran el presente cuerpo se eligieron palabras de uso común y que presentan alguna característica significativa que facilita o dificulta su identificación, las palabras son “probando”, “patata” y “ratón”, y a continuación se presenta cada palabra y su análisis fonético para entender mejor su utilidad o aportación dentro del presente cuerpo.

Para el mejor entendimiento de la presente explicación se utilizo el software comercial Sound Probe 2.51 escrito por David O’Reilly y Kris Delaney que presenta un poderoso modulo de graficación de espectrogramas, razón principal por la que se escogió, ya que

sus similares de otras compañías presentan mucho menos interés en los espectrogramas si acaso los incorporan.

Palabra "patata": La trascripción fonética de la palabra "patata" seria "[p+a+t+a+t+a]" que se aprecia rápidamente como una palabra muy simple, ya que incluso en alfabeto AFI no presenta uniones de fonemas, su tono es totalmente discontinuo como se aprecia en la figura siguiente, en dicha grafica observamos el sonido oclusivo del la "p" sonorizado por la vocálica "a", seguida de la sonorización de la oclusiva "t" por la vocálica "a" terminado de nuevo con la sonorización de la "t" por la "a".

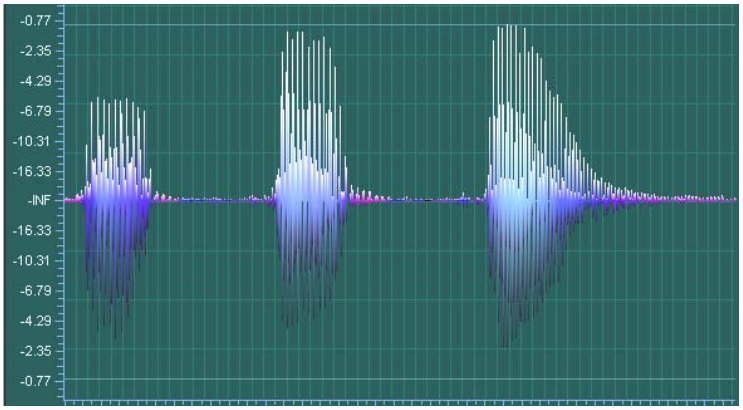


Figura 96: Sonógrama de la palabra "patata"

La sonorización de una consonante es debido a que algunas constantes son imposibles de producir sin una vocal, por ejemplo la consonante "b", intentemos pronunciar una "b" sin una vocal que nos auxilie, veremos que es imposible, por lo que al unirla con una vocal, por ejemplo "be" se sonoriza dicha consonante pasando a formar una unión con dicha consonante y contagiándose de algunas características de dicha vocal.

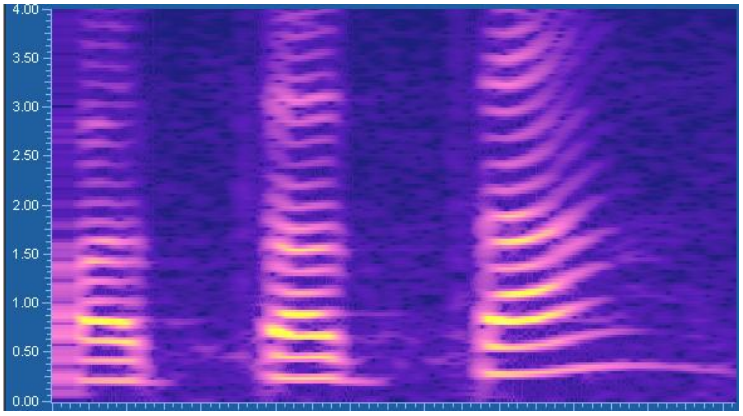


Figura 97: Espectrograma de la palabra "patata"

Del espectrograma anterior podemos ver de manera aun mas clara que se tienen tres unidades que son sumamente parecidas entre si, esto se debe a que la sonorización de todas estas es con la vocal "a" por lo que fuera de factores de entonación, estado anímico y demás agentes, deben de ser prácticamente idénticos, pero observemos en el inicio del espectro la mancha difusa de regular duración, esta corresponde a la realización de la consonante "p" y las "t" son mas angostas y con mayor amplitud espectral, así también vale la pena mencionar la diferencia que existe en la evolución de las estructuras formanticas de la vocal "a" sobre todo en la ultima parte del espectro.

En conclusión esta es una palabra de fácil identificación, ya que se muestra de manera clara la evolución de las formantes, no se tiene mas que una sola vocal a identificar y tenemos pausas claras y prolongadas entre las distintas secciones de la palabra, por eso es nuestra primera elección como una de las palabras para el presente cuerpo.

Palabra "probando": La trascripción fonética de la palabra "probando" seria "[p+r+o+b+a+n+d+o]" que incluso a simple vista no presenta mayores irregularidades o dificultades, ya que a pesar de estar en alfabeto AFI no presenta uniones de fonemas, su tono es bastante sostenido como se aprecia en la figura, en esta observamos el sonido oclusivo del la "p" seguido por la sonorización de la vibrante "r" por la vocálica "o", seguida a su vez de la sonorización de la bilabial "b" por la vocálica "a" continuada por la nasal "n" que presenta mucha estabilidad y terminado con la dental "d" sonorizada por la vocálica "o", esto por su modo de articulación.

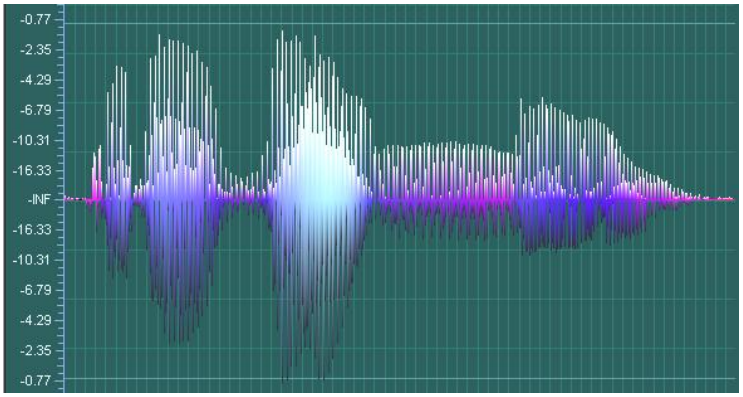


Figura 98: Sonógrama de la palabra "probando"

Del espectrograma podemos observar que se distinguen claramente las mismas estructuras que en el sonógrama, pero incluso se hace mas patente la separación "/p/+/ro/+/ba/+/n/+/do/" en donde de manera clara se ve que no existe separación en "ro", "ba" o "do" entre sus fonemas que lo conforman por su sonorización, aquí también se diferencian fácilmente las formantes en cada realización, así como la ausencia de formantes en sonidos como la "n", también podemos prestar atención a la evolución de cada formante, que tiene un ataque, estabilidad y caída, pudiendo ser similar, igual o sumamente distintos entre si, dependerá de la persona y la palabra, así como estados anémicos y demás factores que contribuyen al resultado final.

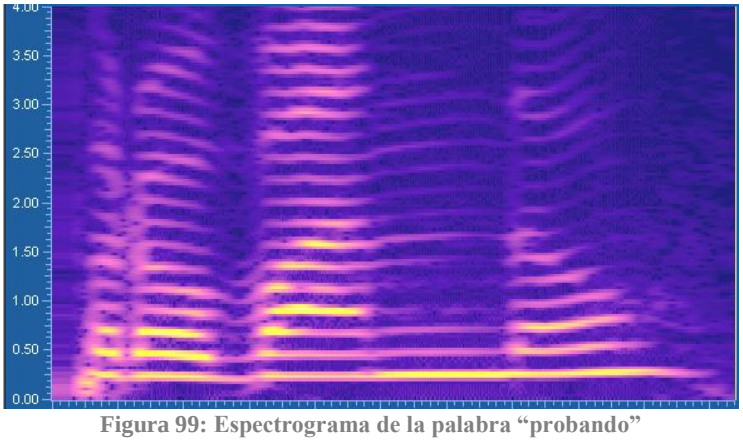


Figura 99: Espectrograma de la palabra "probando"

Del presente análisis podemos concluir que la palabra presenta un nivel de dificultad medio en su identificación, puesto que tiende a formar bloques poco separados entre si en los cuales un sistema automatizado puede tender a unir dichos bloques llevando a una falsa interpretación de resultados, así mismo vemos que los niveles de las formantes son sumamente similares entre las diversas silabas a pesar de no ser iguales, por lo que resulta interesante ver como se desempeña el sistema con esta palabra de uso común.

Palabra "ratón": La trascripción fonética de la palabra "ratón" seria "[r̄+a+t+õ+n]" lo primero que salta a la vista en alfabeto AFI es la presencia de una "rr" al inicio, pero recordemos que todas las "r" a inicio de palabra recibirán la trascripción de una "rr" por ser de pronunciación fuerte en la figura observamos el sonido vibrante múltiple del la "rr" sonorizado por la vocálica "a", seguida a su vez por la sonorización de la oclusiva "t" por la vocálica "o" que remata con la nasal "n" que decrece paulatinamente.

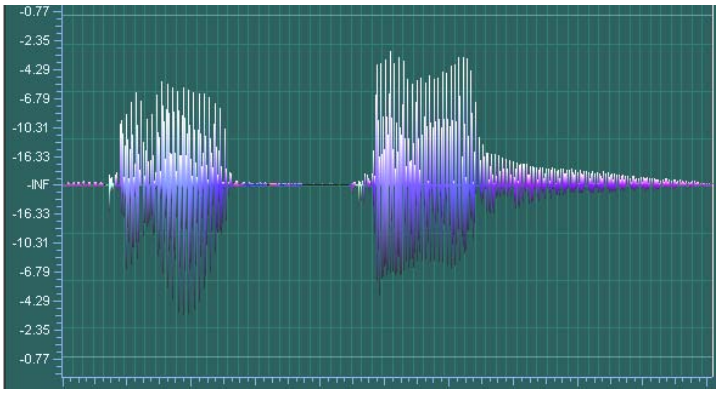


Figura 100: Sonógrama de la palabra "ratón"

Como se puede ver del espectro, esta palabra muestra una vibrante múltiple inicial que se nota va fusionándose con su vocal, seguida del espacio de presurización para la oclusiva, pero de lo que resulta mas interesante es la terminación del sonido nasal "n", que decrece en intensidad, pero su espectro es continuo e incluso aumenta de frecuencia, modificando el tono fundamental de la palabra, lo cual puede propiciar una falla en el algoritmo de detección de tono fundamental de este sistema.

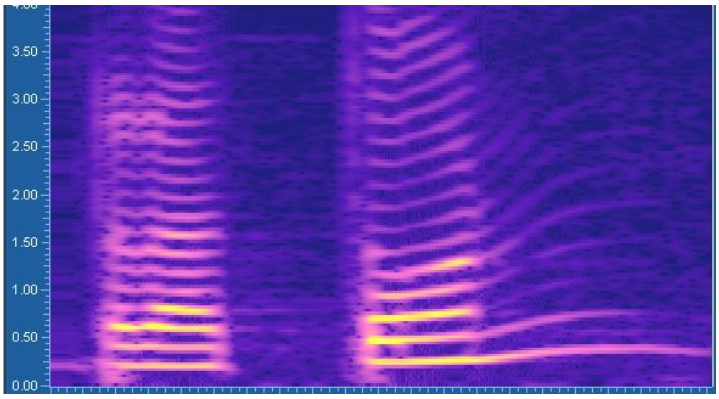


Figura 101: Espectrograma de la palabra "ratón"

En conclusión, esta palabra presenta mas que nada un reto a los métodos de trabajo en frecuencia, ya que el tono fundamental a pesar de ser sumamente estable aquí presenta una fuerte variación, por lo que resulta interesante ver como se comportara el sistema

con este problema, por lo demás el caso es similar a la palabra "patata" es una separación de silabas por la oclusiva, lo cual la hace fácil de detectar hasta cierto punto.

Evaluación del sistema

Una vez establecido y delimitado el cuerpo de prueba sabemos bien con que se probara el sistema y como se obtienen las muestras que estaremos trabajando, ahora hay que establecer una metodología de trabajo, de tal suerte que si se quisiera repetir dichas pruebas se obtenga el mismo resultado, de esta manera se puede seguir trabajando y mejorando el presente sistema o incluso compararlo certeramente contra otra aproximación que se desarrolle a futuro.

Como un primer punto es importante mencionar que el sistema no emite una decisión de si se trata del mismo sujeto o no, tan solo proporciona una serie de medidas de distancia entre una muestra y otra, por tal motivo es importante establecer parámetros que ayuden al perito en la materia a interpretar dichos resultados para que este pueda tomar una decisión o recibir una rápida orientación para basar futuros estudios.

En base a los ya mencionados campos de dispersión se pretende poner a prueba la teoría de que al establecer los campos de dispersión de cada sujeto, así como las medidas de dispersión de dichos campos, podremos sentar una base de que tanta distancia es "normal" que se aleje o aproxime una muestra de voz de otra distinta, pudiendo entonces llegar a una decisión de si se trata o no de la misma persona, lo cual se explica de manera grafica a continuación.

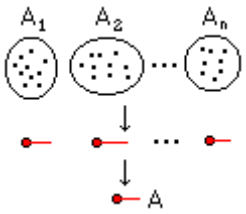


Figura 102: Campos de dispersión

Se tiene un campo de dispersión A_x, de este se obtiene un baricentro y un vector con sus medidas de dispersión, una vez obtenido esto se procede a obtener un baricentro general de todos los campos A_x para obtener un solo vector A que nos dice que tanta distancia puede existir en promedio dentro de un mismo sujeto.

Por lo tanto la idea general es tomar en este caso 10 muestras de hombres y 10 de mujeres, por lo que de inicio por la diferencia en tonalidad entre las voces masculinas y femeninas se propone la formación de dos clases a partir de los 10 sujetos, "Hombres adultos entre 20 y 40 años" y "Mujeres adultas entre 20 y 40 años", una vez evaluados se vera la posibilidad de unir estas dos clases en una sola clase de "Seres humanos adultos entre 20 y 40 años", pero regresando a las dos primeras clases se procederá como a continuación se propone:

1) Evaluación de las distancias entre las muestras de voz de una solo sujeto y de la misma palabra contra sus similares para formar los campos A_1, $A_2 \ldots A_{10}$.
2) Representación de cada campo por su media y sus medidas de dispersión, por lo que se podrá representar un sujeto como una media y un vector de distancias.
3) Obtención del baricentro de las 10 representaciones anteriores para obtener un vector único a cada sujeto por cada palabra.

4) Evaluación de los resultados para ver la factibilidad de unir las tres palabras y los 10 sujetos de cada clase.
5) Obtención de una media de las medidas dispersión permisibles para que un sujeto sea el mismo.
6) Comparación contra las medidas de dispersión entre dos sujetos distintos.
7) Evaluación de resultados.

Una vez completados los estudios antes mencionados se podrá estudiar que tan factible es la formación de una sola clase, ya que las distancias en principio no serán susceptibles de grandes cambios entre sujetos masculinos y femeninos, mas sin embargo pueden llegar a presentar variaciones significativas que no hacen viable la formación de una sola clase.

También es conveniente recordar que la voz humana en sus primeras etapas es sumamente cambiante, estabilizándose alrededor de los 20 años para su periodo de mayor estabilidad, la vida adulta, siendo en general de los 20 años a los 40, que es cuando se puede presentar otro cambio a la voz senil, variando la edad limite de acuerdo a cada individuo, y esta se conservara prácticamente igual por el resto de la vida natural, por esta razón para fines forenses la mayor utilidad es en la etapa adulta de 20 a 40 años que es el promedio de la edad en delincuentes activos.

Por esta misma metodología se pretende evaluar las tres aproximaciones propuestas en el presente trabajo, para que una vez que se tengan los resultados se pueda realizar una comparación certera de la veracidad y certidumbre de cada uno de los métodos propuestos, de tal suerte que este trabajo sirva como base a futuras investigaciones y metodologías en el ámbito del reconocimiento de locutores en el área de Acústica Forense.

Identificación por medio de parámetros en el dominio de tiempo

En este apartado se tomara la primera aproximación que es el análisis de seis parámetros en el dominio del tiempo, siendo estos la magnitud, media, energía, cruces por cero, máximos y mínimos, que ya se explicaron con anterioridad, pero aquí y de primera instancia se comenzara por probar todas las variantes posibles que nos ofrece el sistema con respecto a un solo sujeto de prueba, teniendo registradas un total de tres palabras (“Patata”, “Probando” y “Ratón”) pronunciadas en 10 ocasiones cada una, de las cuales se estudiara el efecto que produce las distintas configuraciones posibles y cual combinación nos ofrecen los mejores resultados.

Es importante mencionar que el sistema nos ofrece únicamente medidas de distancia entre un sujeto y otro, mas no nos da una decisión con respecto a si se trata de un mismo sujeto o no, por lo que esta etapa de evaluación de resultados pretende sentar una base de conocimientos con los que el usuario pueda discernir la mejor forma de uso del programa y la utilidad y aportación de cada método y parámetro con la intención de dar como ya se a mencionado en múltiples ocasiones una poderosa y rápida orientación al experto en la materia de Acústica Forense cuando enfrenta un problema de identificación de locutores.

Opciones de configuración

Como una primera prueba se pretende evaluar los resultados utilizando la palabra "Patata" con cinco muestras seleccionadas de manera aleatoria entre las 10 muestras disponibles del sujeto masculino numero uno (SM01 en lo posterior), estas muestras se compararan entre si, teniéndose un total de 20 comparaciones que se manejaran como un campo de dispersión, pero como lo que se pretende es evaluar las diferentes opciones de configuración para procesar estas muestras tendremos que evaluar estas 20 comparaciones con cada una de las posibilidades de configuración.
Las distintas configuraciones comprenden el uso opcional de una ventana de Hamming, el solapamiento de ventanas con un porcentaje de 2, 5, 10, 12 o 15 por ciento y por ultimo la opción de usar la herramienta de interpolación de archivos para igualar los archivos en su duración antes de extraer los parámetros, por lo que tenemos 24 combinaciones posibles, lo que nos da un gran total de 480 pruebas que se evaluaran en 24 grupos de resultados o 24 campos de dispersión para buscar la mejor configuración posible.

Ahora también es fácil notar que no es necesario ir pasando por cada uno de los puntos posibles de configuración, sino en general solo observar si alguna de las opciones de configuración nos da una mejor medida que la otra, por lo que se tomara únicamente el solapamiento de ventanas en 2% y 15% para ver si contribuye o no, por lo que se reduce el numero de configuraciones a 12 opciones, cabe hacer mención que aquí solo se pondrán las tablas de resultados con respecto a las medias, los resultados de las desviaciones se pueden consultar en las tablas de resultados completos que se anexan, pero si se presentaran sus graficas en el resumen de resultados, las medias se comportaron como sigue:

Configuración #1; Sin aplicar interpolación, sin Hamming y sin solapamiento

	Desviación Media	**Desviación Estándar**
Magnitud	18.0949 ± 1.3639	18.0949 ± 1.5494
Media	21.0320 ± 1.1029	21.0320 ± 1.1730
Energía	18.0883 ± 0.8720	18.0883 ± 1.0266
Cruces por Cero	21.4733 ± 0.7080	21.4733 ± 0.8594
Máximos	19.2812 ± 1.4651	19.2812 ± 1.5296
Mínimos	17.3582 ± 1.0580	17.3582 ± 1.2256
Total	19.2213 ± 0.6659	19.2213 ± 0.7989

Configuración #2; Sin aplicar interpolación, con Hamming y sin solapamiento

	Desviación Media	**Desviación Estándar**
Magnitud	19.5177 ± 2.0637	19.5177 ± 2.3999
Media	31.0700 ± 2.9939	31.0700 ± 3.4717
Energía	20.1485 ± 2.2969	20.1485 ± 2.5218
Cruces por Cero	21.5153 ± 1.5000	21.5153 ± 1.7779
Máximos	19.5474 ± 2.6484	19.5474 ± 3.0259
Mínimos	18.0936 ± 2.2743	18.0936 ± 2.6684
Total	21.6487 ± 1.1110	21.6487 ± 1.2611

Configuración #3; Sin aplicar interpolación, con Hamming y con solapamiento del 2%

	Desviación Media	Desviación Estándar
Magnitud	18.3482 ± 1.3676	18.3482 ± 1.5913
Media	14.5982 ± 1.0810	14.5982 ± 1.2303
Energía	15.1477 ± 1.2586	15.1477 ± 1.4398
Cruces por Cero	24.7832 ± 2.3586	24.7832 ± 2.7165
Máximos	21.2464 ± 2.1926	21.2464 ± 2.7443
Mínimos	17.7887 ± 3.0306	17.7887 ± 3.2986
Total	18.6555 ± 1.3792	18.6555 ± 1.6719

Configuración #4; Sin aplicar interpolación, con Hamming y con solapamiento del 15%

	Desviación Media	Desviación Estándar
Magnitud	18.8658 ± 2.4165	18.8658 ± 2.6575
Media	13.8336 ± 0.7828	13.8336 ± 0.8934
Energía	15.2437 ± 2.3746	15.2437 ± 2.5052
Cruces por Cero	22.0827 ± 1.9148	22.0827 ± 2.2102
Máximos	18.6247 ± 2.0873	18.6247 ± 2.3698
Mínimos	16.8221 ± 2.6244	16.8221 ± 3.0450
Total	17.5788 ± 1.4267	17.5788 ± 1.6265

Configuración #5; Sin aplicar interpolación, sin Hamming y con solapamiento del 2%

	Desviación Media	Desviación Estándar
Magnitud	20.7022 ± 1.5661	20.7022 ± 1.5720
Media	14.7817 ± 1.2469	14.7817 ± 0.9274
Energía	16.9059 ± 2.0457	16.9059 ± 1.5725
Cruces por Cero	24.7809 ± 2.3804	24.7809 ± 2.3349
Máximos	20.8323 ± 1.8739	20.8323 ± 2.1511
Mínimos	17.1359 ± 3.1344	17.1359 ± 3.1446
Total	19.1898 ± 1.4875	19.1898 ± 1.3633

Configuración #6; Sin aplicar interpolación, sin Hamming y con solapamiento del 15%

	Desviación Media	Desviación Estándar
Magnitud	19.8117 ± 1.3112	19.8117 ± 1.5559
Media	14.4158 ± 0.9310	14.4158 ± 1.1206
Energía	16.6983 ± 1.4367	16.6983 ± 1.7275
Cruces por Cero	22.2708 ± 1.8862	22.2708 ± 2.1791
Máximos	18.4876 ± 2.2247	18.4876 ± 2.5600
Mínimos	16.3765 ± 3.3590	16.3765 ± 3.7182
Total	18.0101 ± 1.6061	18.0101 ± 1.9223

Configuración #7; Con interpolación, sin Hamming y sin solapamiento

	Desviación Media	Desviación Estándar
Magnitud	16.0397 ± 2.3854	16.0397 ± 2.8082
Media	17.6428 ± 1.4771	17.6428 ± 1.7158
Energía	19.2014 ± 3.2346	19.2014 ± 3.8845
Cruces por Cero	19.6165 ± 1.7480	19.6165 ± 1.9487
Máximos	17.1681 ± 2.1988	17.1681 ± 2.7022
Mínimos	8.9360 ± 0.3869	8.9360 ± 0.4501
Total	16.4104 ± 1.1218	16.4104 ± 1.3314

Configuración #8; Con interpolación, con Hamming y sin solapamiento

	Desviación Media	Desviación Estándar
Magnitud	15.8670 ± 2.2661	15.8670 ± 2.5331
Media	23.8416 ± 2.6701	23.8416 ± 3.2715
Energía	17.3248 ± 2.6792	17.3248 ± 3.0102
Cruces por Cero	19.4944 ± 1.5978	19.4944 ± 1.8042
Máximos	16.2920 ± 1.8974	16.2920 ± 2.2166
Mínimos	12.1671 ± 0.7243	12.1671 ± 0.8800
Total	17.4978 ± 1.1928	17.4978 ± 1.3121

Configuración #9; Con interpolación, con Hamming y con solapamiento del 2%

	Desviación Media	Desviación Estándar
Magnitud	17.7268 ± 2.2349	17.7268 ± 2.4277
Media	18.1685 ± 3.9525	18.1685 ± 4.3487
Energía	16.5877 ± 1.5705	16.5877 ± 1.8503
Cruces por Cero	20.5537 ± 1.5934	20.5537 ± 1.9193
Máximos	17.1430 ± 1.9943	17.1430 ± 2.3763
Mínimos	12.5505 ± 1.5281	12.5505 ± 1.7174
Total	17.1217 ± 1.2233	17.1217 ± 1.3785

Configuración #10; Con interpolación, con Hamming y con solapamiento del 15%

	Desviación Media	Desviación Estándar
Magnitud	16.7263 ± 2.1794	16.7263 ± 2.5317
Media	18.6533 ± 4.6712	18.6533 ± 5.1361
Energía	168261 ± 2.8909	168261 ± 3.4908
Cruces por Cero	17.5267 ± 1.8261	17.5267 ± 2.1133
Máximos	16.2892 ± 1.9624	16.2892 ± 2.2633
Mínimos	11.7172 ± 0.7429	11.7172 ± 0.8425
Total	16.2906 ± 1.5541	16.2906 ± 1.6425

Configuración #11; Con interpolación, sin Hamming y con solapamiento del 2%

	Desviación Media	Desviación Estándar
Magnitud	16.7142 ± 2.0319	16.7142 ± 2.2997
Media	14.3479 ± 1.8852	14.3479 ± 2.2438
Energía	16.3892 ± 2.3065	16.3892 ± 2.6695
Cruces por Cero	20.4718 ± 1.3476	20.4718 ± 1.6445
Máximos	17.8887 ± 2.7745	17.8887 ± 3.3252
Mínimos	9.5829 ± 0.7022	9.5829 ± 0.7846
Total	15.8946 ± 0.7090	15.8946 ± 0.8210

Configuración #12; Con interpolación, sin Hamming y con solapamiento del 15%

	Desviación Media	Desviación Estándar
Magnitud	15.3768 ± 1.6768	15.3768 ± 1.9170
Media	15.7723 ± 2.6992	15.7723 ± 3.1502
Energía	14.4997 ± 1.9407	14.4997 ± 2.2495
Cruces por Cero	17.6670 ± 1.7874	17.6670 ± 2.0956
Máximos	17.3643 ± 2.6473	17.3643 ± 3.1378
Mínimos	8.5267 ± 0.5111	8.5267 ± 0.5753
Total	14.8701 ± 0.7522	14.8701 ± 0.8382

En conclusión con respecto a la configuración para trabajo en el tiempo podemos ver de manera grafica el comportamiento de las distancias totales y sus correspondientes desviaciones medias en la grafica siguiente, tomando en cuenta que se busca que la distancia sea la menor posible y con el mejor compromiso de desviación, ya que si la distancia es poca y su desviación también tendremos un campo de dispersión mas compacto, esto debido a que todos los cálculos son con una misma voz como se comento anteriormente.

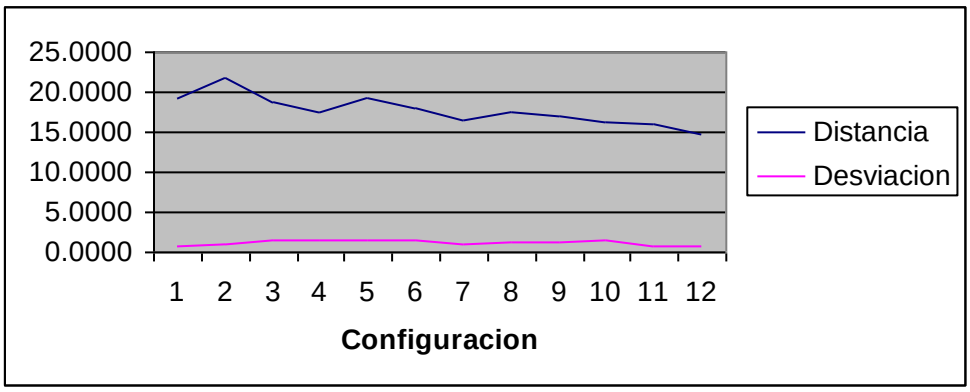


Figura 103: Grafica de las diversas configuraciones

De manera analítica se puede ver que la configuración con la menor distancia total es la opción numero doce con una distancia promedio de 14.8701 unidades, y la menor desviación se presenta con la configuración numero uno con 0.6660 unidades, pero para seleccionar mejor la configuración que se trabajara observemos la siguiente tabla que muestra las tres mejores configuraciones por distancia y desviación correspondientemente.

	Distancia	Desviación
Primera	12	1
Segunda	11	11
Tercera	10	12

De esta tabla nos es fácil ver que la configuración que presenta las mayores ventajas es la numero once con una distancia de 15.8946 y una desviación media de 0.7090 unidades, por lo que en lo sucesivo se trabajara el resto de las pruebas con interpolación de archivos, sin aplicar una ventana de Hamming y con un solapamiento del 2%, así mismo cabe mencionar que como se aprecia en las tablas de resultados anteriores aparentemente no existe una gran correlación entre las distintas características medidas, de hecho el comportamiento es un tanto aleatorio, mismo que puede dar cabida a una sombra de duda sobre la eficacia de la presente metodología, mas sin embargo se realizaran las pruebas de identificación y se analizaran los resultados contra las restantes metodologías propuestas.

Pruebas de identificación

Ahora que tenemos la mejor configuración del sistema se comenzaran a hacer una serie de pruebas de identificación, teniendo entre estas la comparación de dos individuos distintos entre si y en ocasiones dos muestras de un mismo sujeto, esto en miras de ver cuando el sistema sugiere que se trata de la misma persona y cuando se trata de una persona distinta.

Por ultimo, y viendo el comportamiento un tanto irregular de esta aproximación en la etapa de configuración se propone el usar una muestra comparada contra si misma pero con un ligero corrimiento en el tiempo para ver como se desempeña el sistema con esta pequeña trampa y así evaluar su confiabilidad.

Márgenes de variación personal
Aquí se abordara como primera instancia la comparación entre tres muestras de voz de cada uno de los diez locutores del sexo masculino contra su propia voz en la palabra "patata" para obtener así lo que deberían de ser los márgenes aceptables de distancia para un mismo sujeto, esto en el afán de tener un margen en el que debe de tratarse de la misma persona y fuera del cual deberá de tratarse de una persona distinta.

De acuerdo a lo expuesto se evaluaron las distancias de tres muestras por cada sujeto masculino contra si mismas usando la palabra "patata" para obtener el margen en el que debe de ser normal que se trate del mismo sujeto, las distancias que se encuentren fuera de estos promedios serán en teoría identificaciones negativas, y las que estén dentro identificaciones positivas respectivamente, a continuación se presenta la tabla con los resultados y los márgenes que se usaran para evaluar la eficacia de esta metodología propuesta.

Medida	Desviación Media	Desviación Estándar
Magnitud	17.1940 ± 2.4719	17.1940 ± 2.7253
Media	15.4970 ± 2.7949	15.4970 ± 2.9746
Energía	15.8497 ± 2.5035	15.8497 ± 2.7074
Cruces por Cero	18.3434 ± 1.6020	18.3434 ± 1.7879
Máximos	17.4249 ± 2.0925	17.4249 ± 2.3425
Mínimos	9.5485 ± 0.6003	9.5485 ± 0.6491
Total	15.7447 ± 1.2599	15.7447 ± 1.4014

Tabla 16: Configuración en el tiempo para hombres

Siguiendo la misma metodología anterior se precedió a analizar un grupo de sujetos del sexo femenino para obtener los márgenes de variación que se utilizaran en las pruebas de identificación, así como estudiar la posibilidad de englobar en un solo grupo a sujetos masculinos y femeninos con los siguientes resultados.

Medida	Desviación Media	Desviación Estándar
Magnitud	19.7227 ± 3.3159	19.7227 ± 3.5097
Media	15.4577 ± 3.8060	15.4577 ± 3.7820
Energía	17.7444 ± 2.6065	17.7444 ± 3.0780
Cruces por Cero	20.2067 ± 1.4582	20.2067 ± 1.5686
Máximos	18.3396 ± 1.7024	18.3396 ± 1.3345
Mínimos	9.8566 ± 0.8102	9.8566 ± 0.7857
Total	16.8540 ± 1.5462	16.8540 ± 1.5810

Tabla 17: Configuración en el tiempo para mujeres

De estos resultados podemos ver que coincidimos con la mayoría de los investigadores que consideran mejor separar en dos grupos a sujetos del sexo masculino y femenino, ya que al presentarse fuertes variaciones en el tono y otras características de las voces de

ambos géneros es más conveniente tratarlo por separado para evitar errores en los resultados, como se ve incluso en programas de dictado que separan en genero.

Pruebas de identificación

A continuación se tomaran dos muestras de cada sujeto masculino con la palabra "patata" y se probaran una contra la otra así como con las restantes 18 muestras de los 9 sujetos distintos y se evaluara el numero de aciertos y de equivocaciones con las que el sistema supuestamente identifica al sujeto de prueba, el numero total de pruebas a realizar es de 380 comparaciones, siendo el 100% de eficiencia 380 aciertos, si el total de aciertos no es mayor al 50% el método se clasificara como ineficiente y no útil.

Nota: las muestras empleadas en esta prueba y la anterior son distintas entre si con fines de evitar vicios en el proceso de evaluación.

Los resultados completos de estas pruebas se pueden apreciar en la tabla anexa "Pruebas de Identificación Para Sujetos Masculinos", pero en resumidas cuentas lo que nos interesa es en si el análisis de los resultados obtenidos y que se pueden expresar como sigue.

La prueba consta de la comparación de veinte muestras de voz de diez locutores, dos por cada locutor, estas son comparadas cada una contra las restantes 19, obteniéndose 19 resultados por muestra a un total de 20 muestras son 380 pruebas a realizar. Dichos resultados se evaluó la distancia total de los seis parámetros estudiados contra los máximos permitidos, estos máximos son el resultado de los márgenes de variación del punto "*Márgenes de variación personal*" que nos da una distancia máxima en promedio de 17.1461 unidades, de aquí se obtiene la siguiente tabla de posibles resultados:

1	Resultado correcto
0	Resultado erróneo
1	Resultado correcto en margen critico

El resultado correcto en margen critico se refiere a que el sistema obtuvo el resultado esperado pero la distancia se encontró a menos de una unidad de los márgenes utilizados de 17.1461 unidades, el resultado correcto es que se logro el resultado esperado, recordemos que hay dos clases de resultados, cuando debe de ser positivo (dos muestras del mismo locutor) y cuando debe de ser negativo (dos muestras de distintos locutores), pero en general "1" significa que el sistema dio el resultado esperado, sea que se esperaba identificación o descarte, y por ultimo "0" significa que el sistema dio un resultado incorrecto.

Los resultados de esta prueba se pueden apreciar en la siguiente tabla, en esta se da de primera instancia el número de errores y aciertos del sistema en las 380 pruebas efectuadas, para después dar un desglose de dichos resultados conforme lo que se esperaba de resultado.

354	Aciertos	**4**	**Falsa eliminación**
26	Equivocaciones	**16**	**Identificación positiva**
93.16%	Eficiencia	**22**	**Falsa identificación**
		338	**Diferente locutor**

Aquí se ve que de 380 pruebas el sistema obtiene 354 con resultado favorable y tan solo 26 errores, por lo que la eficiencia es del 93.16% de eficiencia, también se ve que el sistema obtiene 4 falsas eliminaciones (dos muestras del mismo locutor y lo dio como negativo), 16 identificaciones certeras (dos muestras del mismo locutor con resultado positivo), 22 falsas identificaciones (dos muestras de distintos locutores con resultado positivo) y 338 descartes acertados (dos muestras de distintos locutores con resultado negativo).

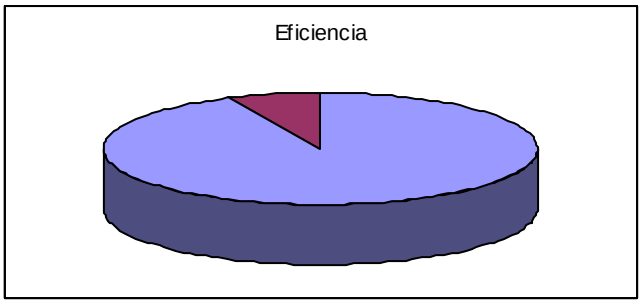


Figura 104: Eficiencia para sujetos masculinos 93.16%

De la figura podemos ver que el sistema obtiene una muy buena eficiencia del 93.16 % que en comparación contra el estándar internacional de 99.65 % (fuente FBI) no es despreciable, recordemos que esta es una primera aproximación y mas que nada una investigación para fundamentar futuros trabajos y mayores eficiencias, además de que vale la pena mencionar que falta evaluar el sistema con sujetos femeninos.

De manera similar para el conjunto de sujetos femeninos se obtienen los siguientes resultados:

332	Aciertos	**6**	**Falsa eliminación**
48	Equivocaciones	**14**	**Identificación positiva**
87.37%	Eficiencia	**42**	**Falsa identificación**
		318	**Diferente locutor**

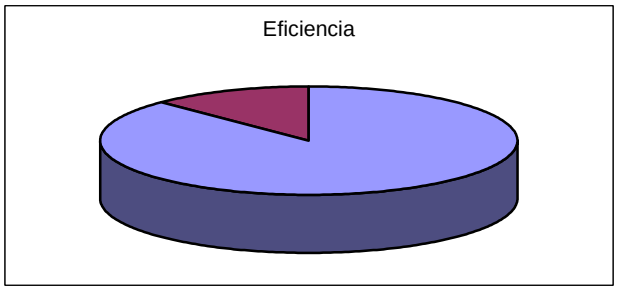


Figura 105: Eficiencia para sujetos femeninos 87.37%

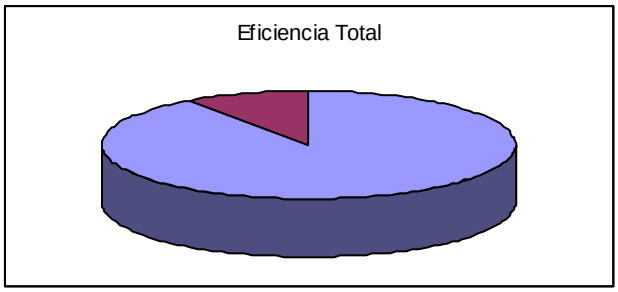


Figura 106: Eficiencia combinada 90.26%

El sistema presenta una eficiencia combinada de 90.26% para sujetos de ambos géneros, misma que será comparada contra las otras dos metodologías implementadas y así mismo vale la pena observar especial interés al hecho de que la eficiencia en sujetos femeninos fue considerablemente menos a la de los sujetos masculinos, hecho que se espera confirmar con las otras dos metodologías.

Por ultimo solo resta la prueba que se comento al principio, se comparara una muestra de voz contra si misma pero con un desplazamiento de tiempo y veremos como se

comporta el sistema, se presentara una grafica de cómo aumenta la distancia conforme se aumenta el desplazamiento para ver que tan critico es la correcta selección de las muestras cuando se va a realizar un estudio.

Desplazamiento (milisegundos)	Distancia
10	11.0487
20	10.9911
30	14.4508
40	15.3008
50	15.7487
60	17.3676
70	17.9762
80	18.6498
90	20.3328
100	21.3015

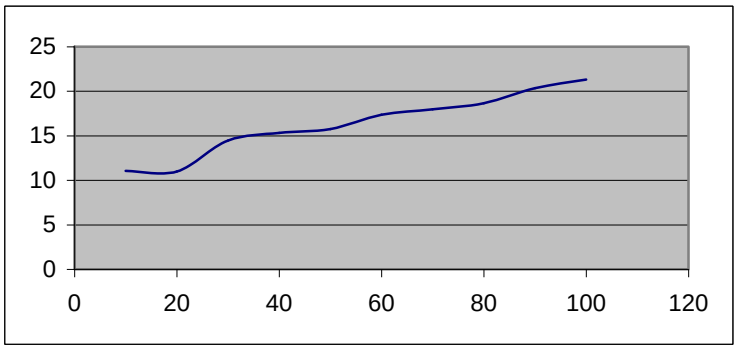


Figura 107: Aumento de la distancia por desplazamiento

Como se puede ver de la figura anterior y de la tabla analítica, arriba de los 60 milisegundos de desplazamiento en el inicio del archivo lleva a que el sistema de una distancia mas allá de los aceptable según los márgenes establecidos en este mismo apartado, por lo tanto no es tan fácil que una ligera equivocación a la hora de seleccionar el audio que se estudiara con el sistema produzca un error de identificación, ya que un desplazamiento de mas de 20 milisegundos en una palabra que en promedio tiene una duración menor a un segundo es difícil de pasar por alto.

Identificación por medio de formantes en el dominio de la frecuencia

En este apartado se pretende poner a prueba el planteamiento de que se puede llevar a cabo la identificación de un sujeto mediante las estructuras formanticas que se pueden identificar después de llevar a cabo la transformación de la señal al dominio de la frecuencia con la transformada de Fourier.

Como un breve recordatorio, se identificara el tono fundamental de la persona y en base a este se intentara localizar de manera automatizada las dos primeras formantes respetando las diversas reglas que estas deberán de cumplir, como ser un múltiplo del tono fundamental, el hecho de que solo se considerara como una parte de la estructura formantica hasta los 3 dB por debajo de la cima y demás reglas antes mencionadas e implementadas en el presente sistema.

Ahora bien recordemos que también esta opción nos da como parámetros las diversas medidas estadísticas de las distancias entre las diversas estructuras formanticas detectadas, mas no un resultado definitivo de si se trata del mismo locutor o no, por lo que es importante ver cual de las diversas opciones de configuración nos ofrece las mejores condiciones de trabajo, lo cual nos lleva al siguiente apartado.

Opciones de configuración

En el dominio de la frecuencia tenemos como variantes la aplicación de interpolación, ventana de Hamming, solapamiento de ventanas del 2%, 5%, 10%, 12% y 15%, y el tamaño de ventana variable (valores aceptables van de 128 a 1024), por lo que vemos que se pueden tener un gran total de 21,504 combinaciones de configuración, por lo que se ve que no es viable el probar todas estas combinaciones.

Siguiendo la metodología antes empleada, lo que nos interesa es ver en general como se comporta la distancia reportada por el sistema con las diferentes combinaciones, por lo que no es necesario el probar cada una de las posibilidades, lo que se pretende hacer es probar la tendencia que se manifestara con cada parámetro y sus combinaciones, por lo que se llevaran a cabo un total de 36 pruebas de funcionamiento que quedaran como a continuación se presenta.

Nota; para asegurar una honesta comparación se usaran las mismas muestras que en las pruebas de tiempo para toda la evaluación de la presente metodología.

Como otra prueba de configuración es necesario ver si se utiliza la media aritmética, media geométrica o media armónica por una simple prueba de cuatro comparaciones y se vera si se disminuye o no la distancia total, ya que no tiene caso agregar otras 72 pruebas para esta opción de visualización, ya que esto no afecta los cálalos, sino tan solo la presentación de resultados, en esta prueba el sistema se comporto como sigue:

	Desviación Media	Varianza	Desviación Estándar	Coeficiente de Variación
	Primera Comparación			
Media Aritmética	2.4	877.22	9.23	0.03
Media Geométrica	4.33	921.93	9.52	0.04
Media Armónica	6.83	1119.98	10.64	0.05
	Segunda Comparación			
Media Aritmética	0.91	539.6	5.88	0.02
Media Geométrica	1.67	555.58	5.95	0.02
Media Armónica	2.46	611.73	6.07	0.02
	Tercera Comparación			
Media Aritmética	3.02	423.92	4.68	0.02
Media Geométrica	3.43	445.72	4.83	0.03
Media Armónica	3.74	529.09	5.29	0.04
	Cuarta Comparación			
Media Aritmética	0.26	620.26	6.7	0.02
Media Geométrica	1.72	652.22	6.92	0.03
Media Armónica	4.53	808.05	7.88	0.04

De esta tabla que expresa la variación del tono fundamental de cuatro comparaciones se ve fácilmente que la mejor opción es la media aritmética, ya que las demás tienden a aumentar las distancias, por lo que a futuro solo se usara media aritmética.

Para las presentes pruebas de configuración se dejara fuera la varianza, ya que es directamente proporcional con la desviación estándar y el coeficiente de variación, ya que no es muy significativo, pero se agregara la media aritmética y la probabilidad total que proporciona el sistema, quedando las pruebas como sigue.

Configuración #1; Sin interpolación, sin solapamiento, sin Hamming, ventana de 128

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	1.79	4.093	3.986	26.041	19.738	6.936
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		71.129	88.157	3.362	**5.038 ± 1.42**	

Configuración #2; Con interpolación, sin solapamiento, sin Hamming, ventana de 128

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	6.929	7.755	4.47	14.925	9.216	8.324
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		51.052	73.758	7.129	**7.127 ± 1.874**	

Configuración #3; Sin interpolación, con solapamiento del 2%, sin Hamming, ventana de 128

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	11.46	7.152	5.782	21.466	15.478	7.883
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		37.125	32.462	2.359	**6.364 ± 1.449**	

Configuración #4; Sin interpolación, con solapamiento del 15%, sin Hamming, ventana de 128

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	4.57	2.659	4.806	19.625	17.703	4.627
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		57.936	65.144	3.236	**4.625 ± 1.31**	

Configuración #5; Sin interpolación, con solapamiento del 2%, con Hamming, ventana de 128

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	7.153	7.097	3.121	24.166	12.639	8.302
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		147.912	167.989	7.399	**6.551 ± 1.908**	

Configuración #6; Sin interpolación, con solapamiento del 15%, con Hamming, ventana de 128

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	6.59	5.756	4.077	31.611	23.864	7.452
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		125.552	119.001	5.673	**6.506 ± 1.6155**	

Configuración #7; Con interpolación, con solapamiento del 2%, sin Hamming, ventana de 128

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	8.261	3.948	5.557	15.145	16.849	6.175
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		43.884	69.2	7.872	**7 ± 2.268**	

Configuración #8; Con interpolación, con solapamiento del 15%, sin Hamming, ventana de 128

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	4.492	3.52	4.909	13.959	15.797	5.04
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		71.948	99.726	7.606	**5.96 ± 0.963**	

Configuración #9; Con interpolación, con solapamiento del 2%, con Hamming, ventana de 128

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	5.281	5.783	3.795	18.743	22.338	4.5
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		78.635	105.718	6.241	**4.624 ± 1.237**	

Configuración #10; Con interpolación, con solapamiento del 15%, con Hamming, ventana de 128

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	2.908	3.432	2.83	24.704	25.967	6.989
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		127.5	143.166	8.211	$\mathbf{6.082 \pm 1.351}$	

Configuración #11; Sin interpolación, sin solapamiento, con Hamming, ventana de 128

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	3.638	2.452	3.337	22.52	16.52	8.35
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		131.874	160.117	5.556	$\mathbf{6.079 \pm 1.505}$	

Configuración #12; Con interpolación, sin solapamiento, con Hamming, ventana de 128

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	6.604	5.226	4.336	12.761	18.846	6.297
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		102.52	133.839	9.013	$\mathbf{6.761 \pm 2.605}$	

Configuración #13; Sin interpolación, sin solapamiento, sin Hamming, ventana de 256

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	2.652	5.512	9.14	28.733	30.268	4.414
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		133.058	170.441	6.184	$\mathbf{7.145 \pm 20.099}$	

Configuración #14; Con interpolación, sin solapamiento, sin Hamming, ventana de 256

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	5.805	4.576	8.224	27.268	29.66	8.33
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		93.321	109.908	8.801	$\mathbf{8.441 \pm 2.861}$	

Configuración #15; Sin interpolación, con solapamiento del 2%, sin Hamming, ventana de 256

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	6.273	9.628	7.477	18.327	11.996	3.977
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		115.349	119.842	4.34	$\mathbf{5.024 \pm 1.177}$	

Configuración #16; Sin interpolación, con solapamiento del 15%, sin Hamming, ventana de 256

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	5.206	5.968	6.613	17.538	15.399	3.104
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		91.945	102.116	4.155	$\mathbf{4.751 \pm 1.0975}$	

Configuración #17; Sin interpolación, con solapamiento del 2%, con Hamming, ventana de 256

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	4.302	6.424	7.056	20.898	18.258	7.246
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		162.35	150.66	7.639	$\mathbf{6.787 \pm 1.6625}$	

Configuración #18; Sin interpolación, con solapamiento del 15%, con Hamming, ventana de 256

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	3.441	5.685	6.053	13.574	9.873	4.589
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		153.436	148.192	6.765	$\mathbf{5.454 \pm 1.072}$	

Configuración #19; Con interpolación, con solapamiento del 2%, sin Hamming, ventana de 256

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	5.848	5.459	7.58	23.027	19.488	9.117
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		146.044	169.508	7.534	**8.51 ± 2.9775**	

Configuración #20; Con interpolación, con solapamiento del 15%, sin Hamming, ventana de 256

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	3.424	3.057	2.915	34.469	34.136	7.837
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		90.288	122.876	6.048	**5.836 ± 1.6433**	

Configuración #21; Con interpolación, con solapamiento del 2%, con Hamming, ventana de 256

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	4.105	4.842	8.522	37.364	33.212	6.495
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		177.926	188.448	9.287	**7.894 ± 2.094**	

Configuración #22; Con interpolación, con solapamiento del 15%, con Hamming, ventana de 256

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	5.597	6.978	1.568	41.167	33.117	5.483
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		147.271	161.591	7.55	**5.422 ± 1.578**	

Configuración #23; Sin interpolación, sin solapamiento, con Hamming, ventana de 256

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	2.982	4.571	8.69	36.126	36.402	5.393
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		165.309	185.205	7.24	$\mathbf{7.314 \pm 1.2795}$	

Configuración #24; Con interpolación, sin solapamiento, con Hamming, ventana de 256

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	5.564	4.906	5.555	44.804	34.061	2.641
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		105.228	122.299	6.949	$\mathbf{5.27 \pm 0.7223}$	

Configuración #25; Sin interpolación, sin solapamiento, sin Hamming, ventana de 512

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	2.294	4.333	5.703	18.252	19.634	8.468
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		126.044	164.154	4.307	$\mathbf{6.046 \pm 1.2567}$	

Configuración #26; Con interpolación, sin solapamiento, sin Hamming, ventana de 512

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	5.296	6.231	5.385	34.946	42.189	12.226
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		115.403	145.587	6.03	$\mathbf{7.511 \pm 1.531}$	

Configuración #27; Sin interpolación, con solapamiento del 2%, sin Hamming, ventana de 512

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	6.676	9.122	3.546	42.444	41.323	9.407
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		98.197	90.022	3.165	$\mathbf{5.536 \pm 1.8}$	

Configuración #28; Sin interpolación, con solapamiento del 15%, sin Hamming, ventana de 512

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	7.959	11.122	10.037	40.475	33.67	9.715
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		78.406	78.406	3.172	$\mathbf{7.413 \pm 2.037}$	

Configuración #29; Sin interpolación, con solapamiento del 2%, con Hamming, ventana de 512

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	4.058	7.224	6.8	29.936	29.62	11.246
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		105.988	83.98	7.813	$\mathbf{8.911 \pm 3.129}$	

Configuración #30; Sin interpolación, con solapamiento del 15%, con Hamming, ventana de 512

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	8.79	11.342	10.357	22.874	20.328	13.894
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		80.286	72.74	8.136	$\mathbf{10.616 \pm 3.357}$	

Configuración #31; Con interpolación, con solapamiento del 2%, sin Hamming, ventana de 512

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	6.981	9.971	11.434	37.828	37.2	14.309
		F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**	
		113.83	135.504	8.376	$\mathbf{11.279 \pm 3.99}$	

Configuración #32; Con interpolación, con solapamiento del 15%, sin Hamming, ventana de 512

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	13.69	14.37	11.277	28.666	32.365	8.63

F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**
117.312	147.82	7.403	$\mathbf{8.658 \pm 2.595}$

Configuración #33; Con interpolación, con solapamiento del 2%, con Hamming, ventana de 512

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	6.084	9.381	10.861	38.728	43.168	9.762

F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**
151.324	162.471	9.188	$\mathbf{8.874 \pm 2.4835}$

Configuración #34; Con interpolación, con solapamiento del 15%, con Hamming, ventana de 512

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	7.9761	11.791	7.693	31.599	33.662	13.783

F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**
134.31	148.281	8.229	$\mathbf{9.265 \pm 2.453}$

Configuración #35; Sin interpolación, sin solapamiento, con Hamming, ventana de 512

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	3.479	7.413	4.642	14.025	19.919	7.833

F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**
122.565	123.211	10.81	$\mathbf{8.404 \pm 2.897}$

Configuración #36; Con interpolación, sin solapamiento, con Hamming, ventana de 512

	T_0 Desviación Media	T_0 Desviación Estándar	T_0 Media	F_1 Desviación Media	F_1 Desviación Estándar	F_1 Media
SM01	7.508	9.302	7.012	55.29	54.057	15.272

F_2 Desviación Media	F_2 Desviación Estándar	F_2 Media	**Probabilidad Total**
113.72	144.322	8.145	$\mathbf{9.412 \pm 2.467}$

En conclusión con respecto a la configuración para trabajo en el dominio de la frecuencia podemos ver de manera grafica el comportamiento de la probabilidad total y su correspondientes desviaciones medias en la grafica siguiente, tomando en cuenta que se busca que la distancia sea la menor posible y con el mejor compromiso de desviación, ya que si la distancia es poca y su desviación también tendremos un campo de dispersión mas compacto, esto debido a que todos los cálculos son con una misma voz como se comento anteriormente.

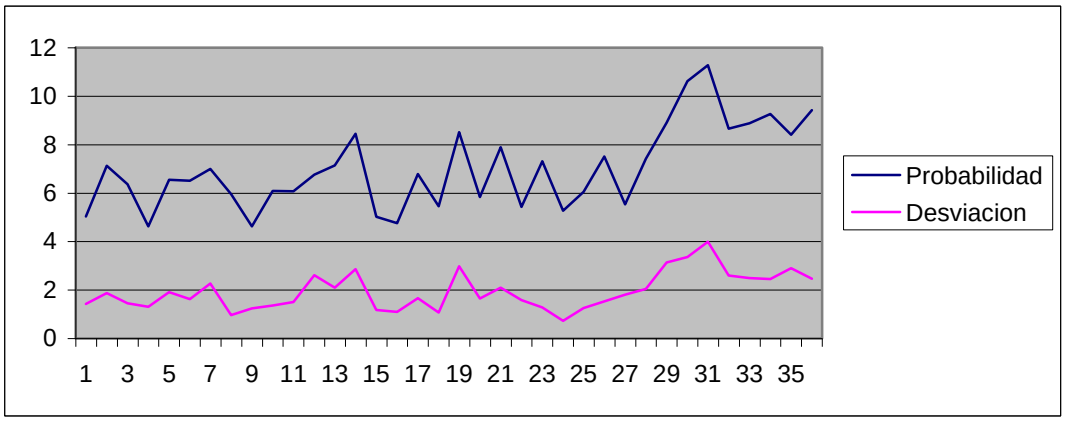


Figura 108: Grafica de las diversas configuraciones

De manera analítica se puede ver que la configuración con la menor distancia total es la opción numero doce con una distancia promedio de 14.8701 unidades, y la menor desviación se presenta con la configuración numero uno con 0.6660 unidades, pero para seleccionar mejor la configuración que se trabajara observemos la siguiente tabla que muestra las seis mejores configuraciones por distancia y desviación correspondientemente.

	Probabilidad	Desviación	Suma
Primera	9	24	16
Segunda	4	8	9
Tercera	16	18	4
Cuarta	15	16	24
Quinta	1	15	15
Sexta	24	9	1

De esta tabla si asignamos el lugar como puntuación se elige la configuración que presenta el menor numero en cuanto a posición y consecuentemente las mayores ventajas, siendo esta configuración la numero dieciséis con una probabilidad de 4.751 y una desviación media de 1.0975 unidades, por lo que en lo sucesivo se trabajara el resto

de las pruebas sin interpolación de archivos, sin aplicar una ventana de Hamming, con un solapamiento del 15% y una ventana de 256 muestras.

Pruebas de identificación

Ahora que tenemos la mejor configuración del sistema se comenzaran a hacer una serie de pruebas de identificación, teniendo entre estas la comparación de dos individuos distintos entre si y en ocasiones dos muestras de un mismo sujeto, esto en miras de ver cuando el sistema sugiere que se trata de la misma persona y cuando se trata de una persona distinta.

Por ultimo y con miras de evaluar el sistema de una forma imparcial se aplicara la misma prueba de corrimiento de tiempo en una muestra para observar y comparar si esta presente metodología mejora este aspecto o no con respecto al anterior.

Márgenes de variación personal

Aquí se abordara como primera instancia la comparación entre tres muestras de voz de cada uno de los diez locutores del sexo masculino contra su propia voz en la palabra "patata" para obtener así lo que deberían de ser los márgenes aceptables de distancia para un mismo sujeto, esto en el afán de tener un margen en el que debe de tratarse de la misma persona y fuera del cual deberá de tratarse de una persona distinta.

De acuerdo a lo expuesto se evaluaron las distancias de tres muestras por cada sujeto masculino contra si mismas usando la palabra "patata" para obtener el margen en el que debe de ser normal que se trate del mismo sujeto, las distancias que se encuentren fuera de estos promedios serán en teoría identificaciones negativas, y las que estén dentro identificaciones positivas respectivamente, a continuación se presenta la tabla con los resultados y los márgenes que se usaran para evaluar la eficacia de esta metodología propuesta.

Medida	Desviación Media	Desviación Estándar
Desviación Media T_0	6.119 ± 2.457	6.119 ± 2.700
Desviación Estándar T_0	5.323 ± 2.238	5.323 ± 2.446
Desviación Media F_1	16.848 ± 6.930	16.848 ± 7.527
Desviación Estándar F_1	14.169 ± 6.774	14.169 ± 7.419
Desviación Media F_2	66.629 ± 25.209	66.629 ± 28.858
Desviación Estándar F_2	64.326 ± 32.289	64.326 ± 34.744
Probabilidad Total	4.036 ± 0.940	4.036 ± 1.034

Tabla 18: Valores en frecuencia para hombres:

Siguiendo la misma metodología anterior se precedió a analizar un grupo de sujetos del sexo femenino para obtener los márgenes de variación que se utilizaran en las pruebas de identificación, así como estudiar la posibilidad de englobar en un solo grupo a sujetos masculinos y femeninos con los siguientes resultados.

Medida	Desviación Media	Desviación Estándar
Desviación Media T_0	12.268 ± 4.794	12.268 ± 5.369
Desviación Estándar T_0	9.643 ± 3.661	9.643 ± 4.128
Desviación Media F_1	28.097 ± 11.362	28.097 ± 12.518

Desviación Estándar F_1	21.708 ± 9.921	21.708 ± 10.618
Desviación Media F_2	48.887 ± 20.316	48.887 ± 22.466
Desviación Estándar F_2	44.767 ± 17.322	44.767 ± 19.664
Probabilidad Total	6.154 ± 1.206	6.154 ± 1.326

Tabla 19: Valores en frecuencia para mujeres

De estos resultados podemos ver que de nuevo coincidimos con la mayoría de los investigadores que consideran mejor separar en dos grupos a sujetos del sexo masculino y femenino, ya que al presentarse fuertes variaciones en el tono y otras características de las voces de ambos géneros es más conveniente tratarlo por separado para evitar errores en los resultados.

Pruebas de identificación

A continuación se tomaran dos muestras de cada sujeto masculino con la palabra "patata" y se probaran una contra la otra así como con las restantes 18 muestras de los 9 sujetos distintos y se evaluara el numero de aciertos y de equivocaciones con las que el sistema supuestamente identifica al sujeto de prueba, el numero total de pruebas a realizar es de 380 comparaciones, siendo el 100% de eficiencia 380 aciertos, si el total de aciertos no es mayor al 50% el método se clasificara como ineficiente y no útil.

Nota: las muestras empleadas en esta prueba y la anterior son distintas entre si con fines de evitar vicios en el proceso de evaluación.

Los resultados completos de estas pruebas se pueden apreciar en la tabla anexa "Pruebas de Identificación Para Sujetos Masculinos", pero en resumidas cuentas lo que nos interesa es en si el análisis de los resultados obtenidos y que se pueden expresar como sigue.

La prueba consta de la comparación de veinte muestras de voz de diez locutores, dos por cada locutor, estas son comparadas cada una contra las restantes 19, obteniéndose 19 resultados por muestra a un total de 20 muestras son 380 pruebas a realizar. Dichos resultados se evaluó la distancia total de los seis parámetros estudiados contra los máximos permitidos, estos máximos son el resultado de los márgenes de variación del punto "*Márgenes de variación personal"* que nos da una distancia máxima en promedio de 5.07 unidades, de aquí se obtiene la siguiente tabla de posibles resultados:

1	Resultado correcto
0	Resultado erróneo
1	Resultado correcto en margen critico

El resultado correcto en margen critico se refiere a que el sistema obtuvo el resultado esperado pero la distancia se encontró a menos de una unidad de los márgenes utilizados de 5.07 unidades, el resultado correcto es que se logro el resultado esperado, recordemos que hay dos clases de resultados, cuando debe de ser positivo (dos muestras del mismo locutor) y cuando debe de ser negativo (dos muestras de distintos locutores), pero en general "1" significa que el sistema dio el resultado esperado, sea que se esperaba identificación o descarte, y por ultimo "0" significa que el sistema dio un resultado incorrecto.

Los resultados de esta prueba se pueden apreciar en la siguiente tabla, en esta se da de primera instancia el número de errores y aciertos del sistema en las 380 pruebas efectuadas, para después dar un desglose de dichos resultados conforme lo que se esperaba de resultado.

330	Aciertos
50	Equivocaciones
86.84%	Eficiencia

16	**Falsa eliminación**
4	**Identificación positiva**
34	**Falsa identificación**
326	**Diferente locutor**

Aquí se ve que de 380 pruebas el sistema obtiene 330 con resultado favorable y tan solo 50 errores, por lo que la eficiencia es del 86.84% de eficiencia, también se ve que el sistema obtiene 16 falsas eliminaciones (dos muestras del mismo locutor y lo dio como negativo), 4 identificaciones certeras (dos muestras del mismo locutor con resultado positivo), 34 falsas identificaciones (dos muestras de distintos locutores con resultado positivo) y 326 descartes acertados (dos muestras de distintos locutores con resultado negativo).

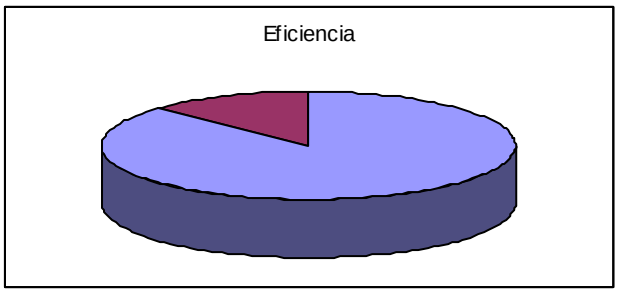


Figura 109: Eficiencia para sujetos masculinos 86.84%

De la figura podemos ver que el sistema obtiene una buena eficiencia del 86.84 % que en comparación contra el mencionado estándar internacional de 99.65 % (fuente FBI) no es despreciable, mas sin embargo deja bastante que desear, este método podría ser considerablemente mejorado mediante la implementación de diversas metodologías, como medidas de distancia mas avanzadas o mejoras en los algoritmos de identificación de tono fundamental e identificación de estructuras formanticas.

De manera similar para el conjunto de sujetos femeninos se obtienen los siguientes resultados:

244	Aciertos
136	Equivocaciones
64.21%	Eficiencia

6	**Falsa eliminación**
14	**Identificación positiva**
130	**Falsa identificación**
230	**Diferente locutor**

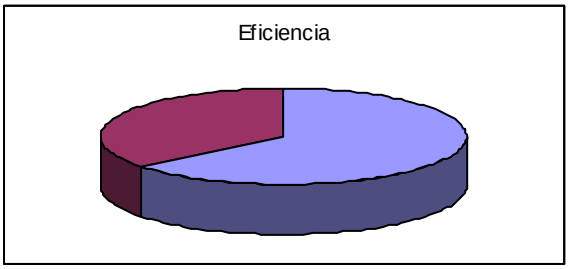


Figura 110: Eficiencia para sujetos femeninos 64.21%

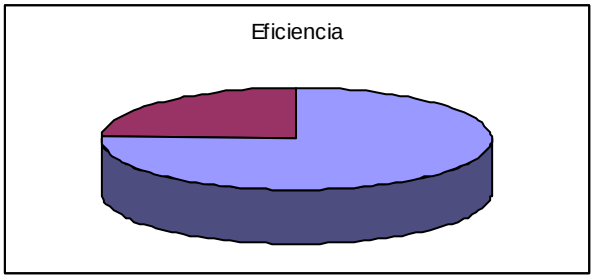
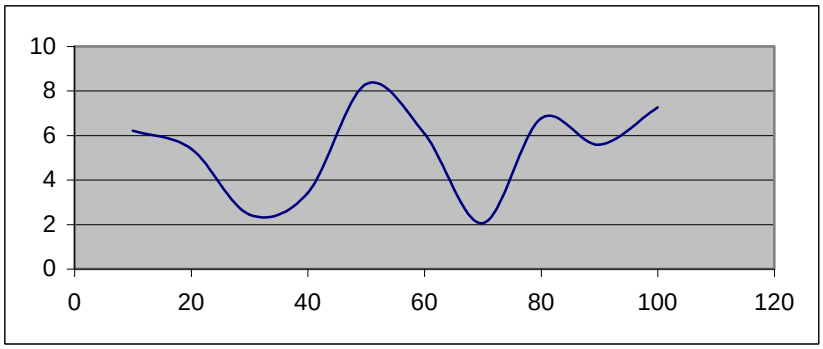


Figura 111: Eficiencia combinada 75.53%

El sistema presenta una eficiencia combinada de 75.53% para sujetos de ambos géneros, misma que será comparada contra las otras dos metodologías implementadas y así mismo vale la pena observar especial interés al hecho de que la eficiencia en sujetos femeninos fue considerablemente menos a la de los sujetos masculinos de manera consistente con la pasada prueba de identificación con parámetros en el tiempo.

Por ultimo solo resta la prueba que se comento al principio, se comparara una muestra de voz contra si misma pero con un desplazamiento de tiempo y veremos como se comporta el sistema, se presentara una grafica de cómo aumenta la distancia conforme se aumenta el desplazamiento para ver que tan critico es la correcta selección de las muestras cuando se va a realizar un estudio.

Desplazamiento (milisegundos)	Distancia
10	6.21
20	5.39
30	2.44
40	3.42
50	8.31
60	6.08
70	2.05
80	6.77
90	5.58
100	7.26



Aumento de la distancia por desplazamiento

De la figura anterior y la tabla analítica vemos que el comportamiento es oscilatorio, siendo bastante critica la selección de tiempo de las muestras, ya que un desplazamiento de 10 milisegundos provoca una falsa eliminación del sujeto, pero vemos que a los 30, 40 y 70 milisegundos el sistema da una identificación acertada, mismo que nos hace pensar que un modulo de pre-procesamiento o acondicionamiento de la señal es sumamente necesario para el trabajo en frecuencia, modulo que seleccionaría de manera automática el silencio y las porciones con voz así como tratar de filtrar y homogeneizar las muestras en aspectos tales como volumen de grabación y ruidos ajenos a la voz en estudio.

Identificación por medio de parámetros LPC

Como una última aproximación se tiene el trabajar con modelos obtenidos mediante parámetros LPC (Linear Prediction Code) o códigos de predicción lineal, por este método se obtiene un modelo del tracto vocal y de la fuente de excitación, mismo modelo que será comparado contra el modelo obtenido de otra grabación.

De esta aproximación también existen un buen numero de variantes que se pueden modificar, como el numero de parámetros que se trabajaran, el tamaño mismo de la ventana, la aplicación de ventanas de Hamming y demás, por lo que vemos que se hace necesaria la misma metodología antes empleada para poder ver cual de estas opciones nos ofrece la mejor calidad de resultados, ya que esta metodología también nos entregara una serie de distancias y no una identificación positiva o negativa.

Opciones de configuración

Cuando utilizamos un juego de parámetros LPC tenemos como variantes la aplicación de interpolación, ventana de Hamming, solapamiento de ventanas del 2%, 5%, 10%, 12% y 15%, el tamaño de la ventana (valores aceptables van de 80 a 320) y el numero de parámetros (valores aceptables de 8 a 19) por lo que vemos que se pueden tener un gran total de 63,360 combinaciones de configuración, por lo que se ve que no es viable el probar todas estas combinaciones.

Siguiendo la metodología antes empleada, lo que nos interesa es ver en general como se comporta la distancia reportada por el sistema con las diferentes combinaciones, por lo que no es necesario el probar cada una de las posibilidades, lo que se pretende hacer es probar la tendencia que se manifestara con cada parámetro y sus combinaciones, por lo que se llevaran a cabo un total de 108 pruebas de funcionamiento que quedaran como a continuación se presenta.

Nota; para asegurar una honesta comparación se usaran las mismas muestras que en las pruebas de tiempo para toda la evaluación de la presente metodología.

Para las presentes pruebas de configuración se dejara fuera el modelo LPC obtenido en si, ya que este presenta un gran numero de parámetros que en general si no se trabajan de manera estadística no nos aporta una medida de comparación de fácil comprensión, por tal motivo solo se tomara en cuenta la distancia promedio entre los dos modelos comparados.

Configuración #1; Sin interpolación, sin solapamiento, sin Hamming, ventana de 80, 10 parámetros

	Distancia Promedio
SM01	0.3790 ± 0.0560

Configuración #2; Con interpolación, sin solapamiento, sin Hamming, ventana de 80, 10 parámetros

	Distancia Promedio
SM01	0.3043 ± 0.0294

Configuración #3; Sin interpolación, sin solapamiento, con Hamming, ventana de 80, 10 parámetros

	Distancia Promedio
SM01	0.3643 ± 0.0289

Configuración #4; Con interpolación, sin solapamiento, con Hamming, ventana de 80, 10 parámetros

	Distancia Promedio
SM01	0.3098 ± 0.0231

Configuración #5; Sin interpolación, con solapamiento del 2%, sin Hamming, ventana de 80, 10 parámetros

	Distancia Promedio
SM01	0.3362 ± 0.0343

Configuración #6; Sin interpolación, con solapamiento del 15%, sin Hamming, ventana de 80, 10 parámetros

	Distancia Promedio
SM01	0.2650 ± 0.0181

Configuración #7; Con interpolación, con solapamiento del 2%, sin Hamming, ventana de 80, 10 parámetros

	Distancia Promedio
SM01	0.2719 ± 0.0377

Configuración #8; Con interpolación, con solapamiento del 15%, sin Hamming, ventana de 80, 10 parámetros

	Distancia Promedio
SM01	0.2351 ± 0.0148

Configuración #9; Sin interpolación, con solapamiento del 2%, con Hamming, ventana de 80, 10 parámetros

	Distancia Promedio
SM01	0.3862 ± 0.0446

Configuración #10; Sin interpolación, con solapamiento del 15%, con Hamming, ventana de 80, 10 parámetros

	Distancia Promedio
SM01	0.3381 ± 0.0416

Configuración #11; Con interpolación, con solapamiento del 2%, con Hamming, ventana de 80, 10 parámetros

	Distancia Promedio
SM01	0.3543 ± 0.0431

Configuración #12; Con interpolación, con solapamiento del 15%, con Hamming, ventana de 80, 10 parámetros

	Distancia Promedio
SM01	0.2950 ± 0.0333

Configuración #13; Sin interpolación, sin solapamiento, sin Hamming, ventana de 200, 10 parámetros

	Distancia Promedio
SM01	0.4780 ± 0.1188

Configuración #14; Con interpolación, sin solapamiento, sin Hamming, ventana de 200, 10 parámetros

	Distancia Promedio
SM01	0.3519 ± 0.0588

Configuración #15; Sin interpolación, sin solapamiento, con Hamming, ventana de 200, 10 parámetros

	Distancia Promedio
SM01	0.4492 ± 0.0592

Configuración #16; Con interpolación, sin solapamiento, con Hamming, ventana de 200, 10 parámetros

	Distancia Promedio
SM01	0.3513 ± 0.0505

Configuración #17; Sin interpolación, con solapamiento del 2%, sin Hamming, ventana de 200, 10 parámetros

	Distancia Promedio
SM01	0.3877 ± 0.0600

Configuración #18; Sin interpolación, con solapamiento del 15%, sin Hamming, ventana de 200, 10 parámetros

	Distancia Promedio
SM01	0.3243 ± 0.0357

Configuración #19; Con interpolación, con solapamiento del 2%, sin Hamming, ventana de 200, 10 parámetros

	Distancia Promedio
SM01	0.3075 ± 0.0459

Configuración #20; Con interpolación, con solapamiento del 15%, sin Hamming, ventana de 200, 10 parámetros

	Distancia Promedio
SM01	0.2772 ± 0.0356

Configuración #21; Sin interpolación, con solapamiento del 2%, con Hamming, ventana de 200, 10 parámetros

	Distancia Promedio
SM01	0.4850 ± 0.0800

Configuración #22; Sin interpolación, con solapamiento del 15%, con Hamming, ventana de 200, 10 parámetros

	Distancia Promedio
SM01	0.4234 ± 0.0625

Configuración #23; Con interpolación, con solapamiento del 2%, con Hamming, ventana de 200, 10 parámetros

	Distancia Promedio
SM01	0.3295 ± 0.0525

Configuración #24; Con interpolación, con solapamiento del 15%, con Hamming, ventana de 200, 10 parámetros

	Distancia Promedio
SM01	0.3559 ± 0.0412

Configuración #25; Sin interpolación, sin solapamiento, sin Hamming, ventana de 320, 10 parámetros

	Distancia Promedio
SM01	0.4453 ± 0.0730

Configuración #26; Con interpolación, sin solapamiento, sin Hamming, ventana de 320, 10 parámetros

	Distancia Promedio
SM01	0.3452 ± 0.0629

Configuración #27; Sin interpolación, sin solapamiento, con Hamming, ventana de 320, 10 parámetros

	Distancia Promedio
SM01	0.3993 ± 0.0444

Configuración #28; Con interpolación, sin solapamiento, con Hamming, ventana de 320, 10 parámetros

	Distancia Promedio
SM01	0.3243 ± 0.0375

Configuración #29; Sin interpolación, con solapamiento del 2%, sin Hamming, ventana de 320, 10 parámetros

	Distancia Promedio
SM01	0.4228 ± 0.0561

Configuración #30; Sin interpolación, con solapamiento del 15%, sin Hamming, ventana de 320, 10 parámetros

	Distancia Promedio
SM01	0.3536 ± 0.0564

Configuración #31; Con interpolación, con solapamiento del 2%, sin Hamming, ventana de 320, 10 parámetros

	Distancia Promedio
SM01	0.4076 ± 0.0531

Configuración #32; Con interpolación, con solapamiento del 15%, sin Hamming, ventana de 320, 10 parámetros

	Distancia Promedio
SM01	0.2963 ± 0.0374

Configuración #33; Sin interpolación, con solapamiento del 2%, con Hamming, ventana de 320, 10 parámetros

	Distancia Promedio
SM01	0.4318 ± 0.0648

Configuración #34; Sin interpolación, con solapamiento del 15%, con Hamming, ventana de 320, 10 parámetros

	Distancia Promedio
SM01	0.3883 ± 0.0724

Configuración #35; Con interpolación, con solapamiento del 2%, con Hamming, ventana de 320, 10 parámetros

	Distancia Promedio
SM01	0.3746 ± 0.0482

Configuración #36; Con interpolación, con solapamiento del 15%, con Hamming, ventana de 320, 10 parámetros

	Distancia Promedio
SM01	0.3475 ± 0.0416

Configuración #37; Sin interpolación, sin solapamiento, sin Hamming, ventana de 80, 14 parámetros

	Distancia Promedio
SM01	0.7187 ± 0.0687

Configuración #38; Con interpolación, sin solapamiento, sin Hamming, ventana de 80, 14 parámetros

	Distancia Promedio
SM01	0.6901 ± 0.0363

Configuración #39; Sin interpolación, sin solapamiento, con Hamming, ventana de 80, 14 parámetros

	Distancia Promedio
SM01	0.7798 ± 0.0525

Configuración #40; Con interpolación, sin solapamiento, con Hamming, ventana de 80, 14 parámetros

	Distancia Promedio
SM01	0.6503 ± 0.0538

Configuración #41; Sin interpolación, con solapamiento del 2%, sin Hamming, ventana de 80, 14 parámetros

	Distancia Promedio
SM01	0.6959 ± 0.0359

Configuración #42; Sin interpolación, con solapamiento del 15%, sin Hamming, ventana de 80, 14 parámetros

	Distancia Promedio
SM01	0.6869 ± 0.0582

Configuración #43; Con interpolación, con solapamiento del 2%, sin Hamming, ventana de 80, 14 parámetros

	Distancia Promedio
SM01	0.6166 ± 0.0321

Configuración #44; Con interpolación, con solapamiento del 15%, sin Hamming, ventana de 80, 14 parámetros

	Distancia Promedio
SM01	0.6642 ± 0.0424

Configuración #45; Sin interpolación, con solapamiento del 2%, con Hamming, ventana de 80, 14 parámetros

	Distancia Promedio
SM01	0.8039 ± 0.0596

Configuración #46; Sin interpolación, con solapamiento del 15%, con Hamming, ventana de 80, 14 parámetros

	Distancia Promedio
SM01	0.6783 ± 0.0651

Configuración #47; Con interpolación, con solapamiento del 2%, con Hamming, ventana de 80, 14 parámetros

	Distancia Promedio
SM01	0.7275 ± 0.0503

Configuración #48; Con interpolación, con solapamiento del 15%, con Hamming, ventana de 80, 14 parámetros

	Distancia Promedio
SM01	0.6651 ± 0.0549

Configuración #49; Sin interpolación, sin solapamiento, sin Hamming, ventana de 200, 14 parámetros

	Distancia Promedio
SM01	0.9246 ± 0.1725

Configuración #50; Con interpolación, sin solapamiento, sin Hamming, ventana de 200, 14 parámetros

	Distancia Promedio
SM01	0.8427 ± 0.1295

Configuración #51; Sin interpolación, sin solapamiento, con Hamming, ventana de 200, 14 parámetros

	Distancia Promedio
SM01	0.9964 ± 0.1260

Configuración #52; Con interpolación, sin solapamiento, con Hamming, ventana de 200, 14 parámetros

	Distancia Promedio
SM01	0.9484 ± 0.1008

Configuración #53; Sin interpolación, con solapamiento del 2%, sin Hamming, ventana de 200, 14 parámetros

	Distancia Promedio
SM01	0.9033 ± 0.0881

Configuración #54; Sin interpolación, con solapamiento del 15%, sin Hamming, ventana de 200, 14 parámetros

	Distancia Promedio
SM01	0.9112 ± 0.0804

Configuración #55; Con interpolación, con solapamiento del 2%, sin Hamming, ventana de 200, 14 parámetros

	Distancia Promedio
SM01	0.8240 ± 0.1105

Configuración #56; Con interpolación, con solapamiento del 15%, sin Hamming, ventana de 200, 14 parámetros

	Distancia Promedio
SM01	0.7879 ± 0.0819

Configuración #57; Sin interpolación, con solapamiento del 2%, con Hamming, ventana de 200, 14 parámetros

	Distancia Promedio
SM01	0.9841 ± 0.1807

Configuración #58; Sin interpolación, con solapamiento del 15%, con Hamming, ventana de 200, 14 parámetros

	Distancia Promedio
SM01	0.8718 ± 0.0882

Configuración #59; Con interpolación, con solapamiento del 2%, con Hamming, ventana de 200, 14 parámetros

	Distancia Promedio
SM01	0.8845 ± 0.1057

Configuración #60; Con interpolación, con solapamiento del 15%, con Hamming, ventana de 200, 14 parámetros

	Distancia Promedio
SM01	0.8092 ± 0.1058

Configuración #61; Sin interpolación, sin solapamiento, sin Hamming, ventana de 320, 14 parámetros

	Distancia Promedio
SM01	1.1008 ± 0.1655

Configuración #62; Con interpolación, sin solapamiento, sin Hamming, ventana de 320, 14 parámetros

	Distancia Promedio
SM01	0.8804 ± 0.1497

Configuración #63; Sin interpolación, sin solapamiento, con Hamming, ventana de 320, 14 parámetros

	Distancia Promedio
SM01	1.0474 ± 0.1003

Configuración #64; Con interpolación, sin solapamiento, con Hamming, ventana de 320, 14 parámetros

	Distancia Promedio
SM01	0.8922 ± 0.1739

Configuración #65; Sin interpolación, con solapamiento del 2%, sin Hamming, ventana de 320, 14 parámetros

	Distancia Promedio
SM01	1.0881 ± 0.1690

Configuración #66; Sin interpolación, con solapamiento del 15%, sin Hamming, ventana de 320, 14 parámetros

	Distancia Promedio
SM01	1.0708 ± 0.1289

Configuración #67; Con interpolación, con solapamiento del 2%, sin Hamming, ventana de 320, 14 parámetros

	Distancia Promedio
SM01	1.0103 ± 0.1613

Configuración #68; Con interpolación, con solapamiento del 15%, sin Hamming, ventana de 320, 14 parámetros

	Distancia Promedio
SM01	1.0146 ± 0.1328

Configuración #69; Sin interpolación, con solapamiento del 2%, con Hamming, ventana de 320, 14 parámetros

	Distancia Promedio
SM01	1.0042 ± 0.1196

Configuración #70; Sin interpolación, con solapamiento del 15%, con Hamming, ventana de 320, 14 parámetros

	Distancia Promedio
SM01	0.9601 ± 0.1793

Configuración #71; Con interpolación, con solapamiento del 2%, con Hamming, ventana de 320, 14 parámetros

	Distancia Promedio
SM01	0.8651 ± 0.1267

Configuración #72; Con interpolación, con solapamiento del 15%, con Hamming, ventana de 320, 14 parámetros

	Distancia Promedio
SM01	1.0296 ± 0.1270

Configuración #73; Sin interpolación, sin solapamiento, sin Hamming, ventana de 80, 19 parámetros

	Distancia Promedio
SM01	1.7598 ± 0.0853

Configuración #74; Con interpolación, sin solapamiento, sin Hamming, ventana de 80, 19 parámetros

	Distancia Promedio
SM01	1.6780 ± 0.0651

Configuración #75; Sin interpolación, sin solapamiento, con Hamming, ventana de 80, 19 parámetros

	Distancia Promedio
SM01	1.7021 ± 0.1227

Configuración #76; Con interpolación, sin solapamiento, con Hamming, ventana de 80, 19 parámetros

	Distancia Promedio
SM01	1.5204 ± 0.1229

Configuración #77; Sin interpolación, con solapamiento del 2%, sin Hamming, ventana de 80, 19 parámetros

	Distancia Promedio
SM01	1.7223 ± 0.0857

Configuración #78; Sin interpolación, con solapamiento del 15%, sin Hamming, ventana de 80, 19 parámetros

	Distancia Promedio
SM01	1.4615 ± 0.1127

Configuración #79; Con interpolación, con solapamiento del 2%, sin Hamming, ventana de 80, 19 parámetros

	Distancia Promedio
SM01	1.4741 ± 0.0986

Configuración #80; Con interpolación, con solapamiento del 15%, sin Hamming, ventana de 80, 19 parámetros

	Distancia Promedio
SM01	1.4025 ± 0.0965

Configuración #81; Sin interpolación, con solapamiento del 2%, con Hamming, ventana de 80, 19 parámetros

	Distancia Promedio
SM01	1.7540 ± 0.1058

Configuración #82; Sin interpolación, con solapamiento del 15%, con Hamming, ventana de 80, 19 parámetros

	Distancia Promedio
SM01	1.4993 ± 0.1037

Configuración #83; Con interpolación, con solapamiento del 2%, con Hamming, ventana de 80, 19 parámetros

	Distancia Promedio
SM01	1.6168 ± 0.0967

Configuración #84; Con interpolación, con solapamiento del 15%, con Hamming, ventana de 80, 19 parámetros

	Distancia Promedio
SM01	1.3724 ± 0.0875

Configuración #85; Sin interpolación, sin solapamiento, sin Hamming, ventana de 200, 19 parámetros

	Distancia Promedio
SM01	2.3299 ± 0.1430

Configuración #86; Con interpolación, sin solapamiento, sin Hamming, ventana de 200, 19 parámetros

	Distancia Promedio
SM01	2.1881 ± 0.2455

Configuración #87; Sin interpolación, sin solapamiento, con Hamming, ventana de 200, 19 parámetros

	Distancia Promedio
SM01	2.1715 ± 0.2148

Configuración #88; Con interpolación, sin solapamiento, con Hamming, ventana de 200, 19 parámetros

	Distancia Promedio
SM01	1.9779 ± 0.1300

Configuración #89; Sin interpolación, con solapamiento del 2%, sin Hamming, ventana de 200, 19 parámetros

	Distancia Promedio
SM01	5.7029 ± 3.1530

Configuración #90; Sin interpolación, con solapamiento del 15%, sin Hamming, ventana de 200, 19 parámetros

	Distancia Promedio
SM01	1.7713 ± 0.1474

Configuración #91; Con interpolación, con solapamiento del 2%, sin Hamming, ventana de 200, 19 parámetros

	Distancia Promedio
SM01	2.0598 ± 0.2116

Configuración #92; Con interpolación, con solapamiento del 15%, sin Hamming, ventana de 200, 19 parámetros

	Distancia Promedio
SM01	1.5980 ± 0.1799

Configuración #93; Sin interpolación, con solapamiento del 2%, con Hamming, ventana de 200, 19 parámetros

	Distancia Promedio
SM01	5.7019 ± 3.1393

Configuración #94; Sin interpolación, con solapamiento del 15%, con Hamming, ventana de 200, 19 parámetros

	Distancia Promedio
SM01	1.9915 ± 0.1798

Configuración #95; Con interpolación, con solapamiento del 2%, con Hamming, ventana de 200, 19 parámetros

	Distancia Promedio
SM01	2.0193 ± 0.1955

Configuración #96; Con interpolación, con solapamiento del 15%, con Hamming, ventana de 200, 19 parámetros

	Distancia Promedio
SM01	1.9214 ± 0.1902

Configuración #97; Sin interpolación, sin solapamiento, sin Hamming, ventana de 320, 19 parámetros

	Distancia Promedio
SM01	2.7999 ± 0.4017

Configuración #98; Con interpolación, sin solapamiento, sin Hamming, ventana de 320, 19 parámetros

	Distancia Promedio
SM01	2.3645 ± 0.2930

Configuración #99; Sin interpolación, sin solapamiento, con Hamming, ventana de 320, 19 parámetros

	Distancia Promedio
SM01	2.3832 ± 0.3281

Configuración #100; Con interpolación, sin solapamiento, con Hamming, ventana de 320, 19 parámetros

	Distancia Promedio
SM01	2.2967 ± 0.3531

Configuración #101; Sin interpolación, con solapamiento del 2%, sin Hamming, ventana de 320, 19 parámetros

	Distancia Promedio
SM01	2.5254 ± 0.2255

Configuración #102; Sin interpolación, con solapamiento del 15%, sin Hamming, ventana de 320, 19 parámetros

	Distancia Promedio
SM01	1.9591 ± 0.2594

Configuración #103; Con interpolación, con solapamiento del 2%, sin Hamming, ventana de 320, 19 parámetros

	Distancia Promedio
SM01	2.3777 ± 0.2786

Configuración #104; Con interpolación, con solapamiento del 15%, sin Hamming, ventana de 320, 19 parámetros

	Distancia Promedio
SM01	2.3175 ± 0.2646

Configuración #105; Sin interpolación, con solapamiento del 2%, con Hamming, ventana de 320, 19 parámetros

	Distancia Promedio
SM01	2.3676 ± 0.2657

Configuración #106; Sin interpolación, con solapamiento del 15%, con Hamming, ventana de 320, 19 parámetros

	Distancia Promedio
SM01	2.1571 ± 0.2424

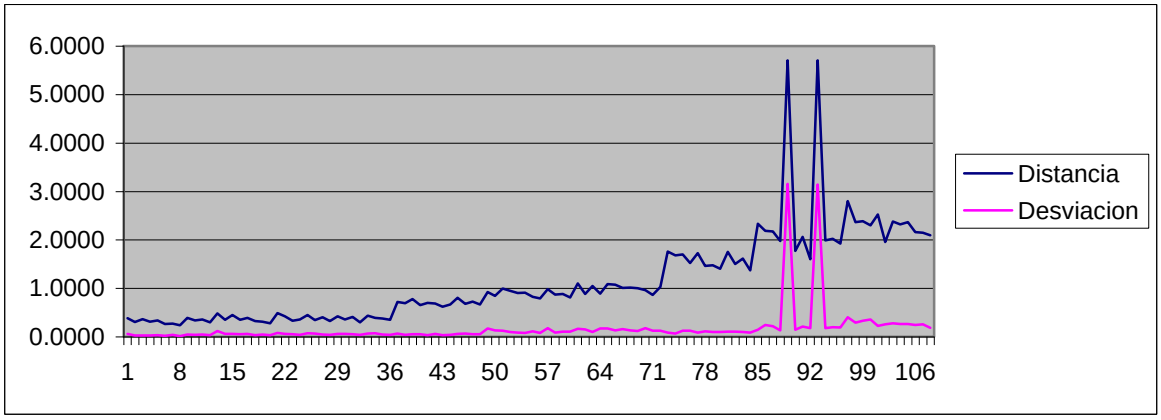
Configuración #107; Con interpolación, con solapamiento del 2%, con Hamming, ventana de 320, 19 parámetros

	Distancia Promedio
SM01	2.1491 ± 0.2545

Configuración #108; Con interpolación, con solapamiento del 15%, con Hamming, ventana de 320, 19 parámetros

	Distancia Promedio
SM01	2.0948 ± 0.1818

En conclusión con respecto a la configuración para trabajo con un modelo LPC podemos ver de manera grafica el comportamiento de la distancia promedio y su correspondiente desviación estándar en la grafica siguiente, tomando en cuenta que se busca que la distancia sea la menor posible y con el mejor compromiso de desviación, ya que si la distancia es poca y su desviación también tendremos un campo de dispersión mas compacto, esto debido a que todos los cálculos son con una misma voz como se comento anteriormente.



Grafica de las diversas configuraciones

De manera analítica se puede ver que la configuración con la menor distancia promedio es la opción numero ocho con una distancia promedio de 0.2351 unidades, y la menor desviación se presenta también con la configuración numero ocho con 0.0148 unidades, por lo que es obvia la selección de la configuración que se trabajara.

	Distancia Promedio	Desviación Estándar	Suma
Primera	8	8	8
Segunda	6	6	6
Tercera	7	4	7
Cuarta	20	3	20

De esta tabla vemos que la configuración numero ocho con una distancia de 0.2351 y una desviación estándar de 0.0148 unidades nos da la mejor opción de trabajo, mas sin embargo recordemos que la diversa bibliografía consultada a través de esta investigación sugiere que no se use pre-procesado de la señal en cuestión, por lo que se utilizara de primera instancia la segunda mejor opción que es la configuración numero seis con una distancia de 0.2650 unidades y una desviación de 0.0181 unidades, por lo que en lo sucesivo se trabajara el resto de las pruebas sin interpolación de archivos, sin aplicar una ventana de Hamming, con un solapamiento del 15%, una ventana de 80 muestras y 10 parámetros LPC.

Pruebas de identificación

Ahora que tenemos la mejor configuración del sistema se comenzaran a hacer una serie de pruebas de identificación, teniendo entre estas la comparación de dos individuos distintos entre si y en ocasiones dos muestras de un mismo sujeto, esto en miras de ver cuando el sistema sugiere que se trata de la misma persona y cuando se trata de una persona distinta.

Por ultimo y con miras de evaluar el sistema de una forma imparcial se aplicara la misma prueba de corrimiento de tiempo en una muestra para observar y comparar si esta presente metodología mejora este aspecto o no con respecto a las anteriores.

Márgenes de variación personal

Aquí se abordara como primera instancia la comparación entre tres muestras de voz de cada uno de los diez locutores del sexo masculino contra su propia voz en la palabra "patata" para obtener así lo que deberían de ser los márgenes aceptables de distancia para un mismo sujeto, esto en el afán de tener un margen en el que debe de tratarse de la misma persona y fuera del cual deberá de tratarse de una persona distinta.

De acuerdo a lo expuesto se evaluaron las distancias de tres muestras por cada sujeto masculino contra si mismas usando la palabra "patata" para obtener el margen en el que debe de ser normal que se trate del mismo sujeto, las distancias que se encuentren fuera de estos promedios serán en teoría identificaciones negativas, y las que estén dentro identificaciones positivas respectivamente, a continuación se presenta la tabla con los resultados y los márgenes que se usaran para evaluar la eficacia de esta metodología propuesta.

Medida	Desviación Media	Desviación Estándar
Distancia Promedio	0.3163 ± 0.0182	0.3163 ± 0.0207

Siguiendo la misma metodología anterior se precedió a analizar un grupo de sujetos del sexo femenino para obtener los márgenes de variación que se utilizaran en las pruebas de identificación, así como estudiar la posibilidad de englobar en un solo grupo a sujetos masculinos y femeninos con los siguientes resultados.

Medida	Desviación Media	Desviación Estándar
Distancia Promedio	0.3264 ± 0.0291	0.3264 ± 0.0320

De estos resultados podemos ver que seguimos coincidiendo en la idea de que no es conveniente la agrupación de los dos sexos, ya que aunque la diferencia es pequeña en su propia escala es sumamente significativa, por lo que no es conveniente tratar de poner ambos grupos en un solo conjunto del género humano.

Pruebas de identificación

A continuación se tomaran dos muestras de cada sujeto masculino con la palabra "patata" y se probaran una contra la otra así como con las restantes 18 muestras de los 9 sujetos distintos y se evaluara el numero de aciertos y de equivocaciones con las que el sistema supuestamente identifica al sujeto de prueba, el numero total de pruebas a realizar es de 380 comparaciones, siendo el 100% de eficiencia 380 aciertos, si el total de aciertos no es mayor al 50% el método se clasificara como ineficiente y no útil.

Nota: las muestras empleadas en esta prueba y la anterior son distintas entre si con fines de evitar vicios en el proceso de evaluación.

Los resultados completos de estas pruebas se pueden apreciar en la tabla anexa "Pruebas de Identificación Para Sujetos Masculinos", pero en resumidas cuentas lo que nos interesa es en si el análisis de los resultados obtenidos y que se pueden expresar como sigue.

La prueba consta de la comparación de veinte muestras de voz de diez locutores, dos por cada locutor, estas son comparadas cada una contra las restantes 19, obteniéndose 19 resultados por muestra a un total de 20 muestras son 380 pruebas a realizar. En dichos resultados se evalúa la distancia de los diez parámetros estudiados representados por su distancia promedio contra los máximos permitidos, estos máximos son el resultado de los márgenes de variación del punto

"*Márgenes de variación personal*" que nos da una distancia máxima en promedio de 0.3163 unidades, de aquí se obtiene la siguiente tabla de posibles resultados:

1	Resultado correcto
0	Resultado erróneo

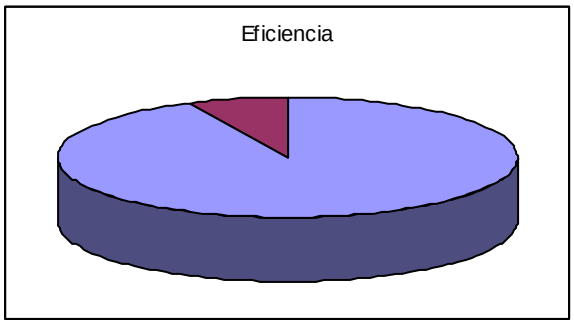
El resultado correcto en margen critico se elimino debido a que las distancias son muy pequeñas y no es muy representativo que tan cercano esta del margen dicho resultado, por otra parte recordemos que hay dos clases de resultados, cuando debe de ser positivo (dos muestras del mismo locutor) y cuando debe de ser negativo (dos muestras de distintos locutores), pero en general "1" significa que el sistema dio el resultado esperado, sea que se esperaba identificación o descarte, y por ultimo "0" significa que el sistema dio un resultado incorrecto.

Dentro de la presente prueba no se llevo a cabo la suma la desviación estándar o desviación media a la media obtenida, esto en el mismo tenor de que la distancia es muy pequeña y al sumar la desviación se reduce la eficiencia del método en 2.63%, ya que se probo de ambas maneras.

Los resultados de esta prueba se pueden apreciar en la siguiente tabla, en esta se da de primera instancia el número de errores y aciertos del sistema en las 380 pruebas efectuadas, para después dar un desglose de dichos resultados conforme lo que se esperaba de resultado.

354	Aciertos	**12**	**Falsa eliminación**
26	Equivocaciones	**8**	**Identificación positiva**
93.16%	Eficiencia	**14**	**Falsa identificación**
		346	**Diferente locutor**

Aquí se ve que de 380 pruebas el sistema obtiene 354 con resultado favorable y tan solo 26 errores, por lo que la eficiencia es del 93.16% de eficiencia, también se ve que el sistema obtiene 12 falsas eliminaciones (dos muestras del mismo locutor y lo dio como negativo), 8 identificaciones certeras (dos muestras del mismo locutor con resultado positivo), 14 falsas identificaciones (dos muestras de distintos locutores con resultado positivo) y 346 descartes acertados (dos muestras de distintos locutores con resultado negativo).



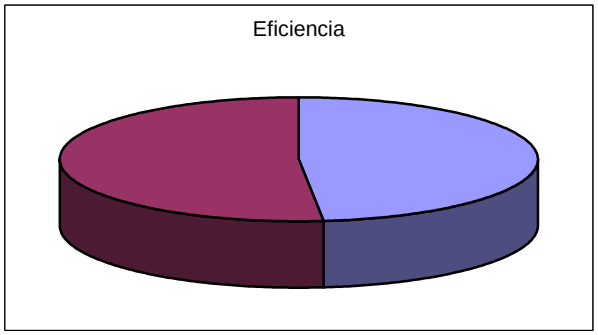
Eficiencia para sujetos masculinos 93.16%

De la figura podemos ver que el sistema obtiene una buena eficiencia del 93.16 % que en comparación contra el mencionado estándar internacional de 99.65 % (fuente FBI) no es nada despreciable, puede mejorarse, ya que al ser reducido el margen de distancias que se emplean resulta bastante critica la medición de distancias en si.

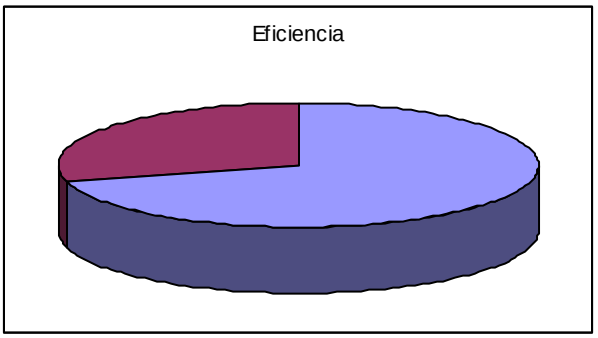
De manera similar para el conjunto de sujetos femeninos se obtienen los siguientes resultados:

184	Aciertos
196	Equivocaciones
48.42%	Eficiencia

8	**Falsa eliminación**
12	**Identificación positiva**
188	**Falsa identificación**
172	**Diferente locutor**



Eficiencia para sujetos femeninos 48.42%

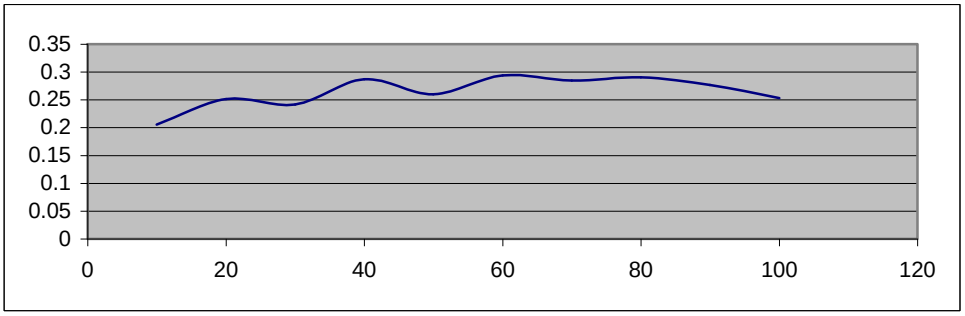


Eficiencia combinada 74.72%

El sistema presenta una eficiencia combinada de 74.72% para sujetos de ambos géneros, misma que será comparada contra las otras dos metodologías implementadas y así mismo vale la pena observar especial interés al hecho de que la eficiencia en sujetos femeninos fue sumamente inferior a la de los sujetos masculinos de manera consistente con la pasada prueba de identificación con parámetros en el tiempo y en la frecuencia.

Por ultimo solo resta la prueba que se comento al principio, se comparara una muestra de voz contra si misma pero con un desplazamiento de tiempo y veremos como se comporta el sistema, se presentara una grafica de cómo aumenta la distancia conforme se aumenta el desplazamiento para ver que tan critico es la correcta selección de las muestras cuando se va a realizar un estudio.

Desplazamiento (milisegundos)	Distancia
10	0.2056
20	0.2511
30	0.2417
40	0.2866
50	0.2598
60	0.2938
70	0.2845
80	0.2904
90	0.2766
100	0.2532



Aumento de la distancia por desplazamiento

De la figura anterior y la tabla analítica vemos que el comportamiento es bastante estable, se maximizo el error aproximadamente a los 80 milisegundos, pero entonces comenzó a bajar de nuevo, en general la prueba nos dice que el método es bastante estable y soporta corrimientos incluso de 100 milisegundos sin que el resultado se saliera de los márgenes obtenidos y trabajados anteriormente.

Discusión

Ahora bien, de lo que hemos visto en el presente trabajo no queda mas que cerrar diciendo que hay mas por hacer de lo que se ha logrado hacer a la fecha, los sistemas de reconocimiento de habla, de identificación de locutor y de verificación de locutor nos ofrecen todavía una ciencia en sus inicios, ya tenemos sistemas altamente eficientes, pero aun lejos de lo que uno podría desear o de lo que la sociedad nos demanda.

Este trabajo pretende dar una base a futuras investigaciones así como despertar el interés en sistemas de esta índole, despertando esa curiosidad científica a aquellas personas que les guste la investigación y los retos difíciles, pues como podemos ver cada vez que investigamos de nuevo encontramos una gran variedad de ideas, opiniones, nuevos algoritmos y sobre todo muchas preguntas, ¿Cómo mejorar la eficiencia?, ¿Por qué tengo esta variación de eficiencia entre sexos?, ¿Cuál método es el mejor?, ¿Puede mejorarse? Y mil preguntas más destinadas a lograr cada día mejores resultados y contestar algunas de estas incógnitas, o talvez despertar nuevas.

El procesamiento de voz con fines de identificación es una ciencia joven, tiene sus inicios en la segunda guerra mundial en proyectos que pretendían poder identificar a una persona en las transmisiones de radio, y que tuvo logros importantes en la década de los 60's cuando se lograron los primeros sistemas de identificación, por lo que es todavía una rama que ofrece muchas opciones y que esta abierta a nuevas propuestas, por lo que espero que el presente trabajo logre despertar esa curiosidad que nos permita llevar estos estudios a aplicaciones practicas, ya que en México es una ciencia aun mas joven, teniendo sus inicios escasamente a 10 años de distancia, por lo que es necesaria la investigación e implementación de sistemas y técnicas que nos permitan poner a nuestro país a un nivel competitivo y de excelencia, ya que por desgracia o por fortuna, como el lector lo desee ver, los sistemas desarrollados en lenguas extranjeras tienen poca utilidad en nuestro idioma por las diferencias lingüísticas, dando cabida a investigaciones como esta.

Separación por géneros

Como un tema interesante a discutir y como se pudo observar la separación de individuos del sexo masculino y femenino es altamente recomendable, esto en virtud de que los resultados por ejemplo en los márgenes para descartar o aceptar una identidad son significativamente distintos.

Esta idea no es de asombrarse, ya que incluso programas comerciales de dictado llevan a cabo una separación, siendo el primer paso para su configuración el determinar si el usuario es de sexo masculino o femenino para posteriormente pasar a la configuración y medición de niveles del hablante en si.

En esto coincidimos con la opinión general de que no es una buena idea el querer englobar en un solo grupo a ambos sexos, y después de todo no es tan extraño, ya que la diferencia en el tono fundamental es significativa, tendiendo a ser la voz femenina mas aguda y armónica que la masculina, por lo que merece darle un poco de atención a este aspecto con el fin de evitar errores en los resultados.

Diferencia entre sexos

Como se observo de los resultados, el sistema tiende a bajar su eficiencia de manera significativa con sujetos del sexo femenino, teniéndose una diferencia de 5.79% en el

análisis en el tiempo, 22.63% en el análisis en frecuencia y un alarmante 44.74% en modelos LPC, hecho que hace que valga la pena que en futuras investigaciones se de énfasis a localizar los causales de dicha baja en eficiencia.

Aquí como una teoría se plantea el hecho de que se trabajaron voces filtradas a 8,000 Hz que es la frecuencia de corte de una línea telefónica estándar, esto en virtud de que la inmensa mayoría de los asuntos que se estudian en las ciencias forenses son grabaciones ya sea de aparatos conectados por el usuario a sus teléfonos u obtenidas mediante la intervención de líneas telefónicas de sujetos presuntamente responsables.

De esto nace la idea de que al ser las voces femeninas mas agudas (valores cercanos al doble de frecuencia de un varón) la línea telefónica puede estar filtrando una considerable cantidad de información que seria necesaria para el correcto estudio y comparación de las mismas, factor que lleva a que la computadora "vea" mucho mas similares la voz de dos mujeres distintas que las de dos hombres.

Esta teoría puede ser probada de manera no muy complicada al subir la frecuencia de adquisición de las voces para observar si disminuye dicho error, pero aunado a esto bien valdría la pena observar especial cuidado a que en las mujeres puede ser mas estrecho el margen de identificación, hecho que podría ser observado en futuras investigaciones sobre el tema, pero por ahora se deja esta incógnita abierta al no contar con elementos para determinar un causal para esta baja de eficiencia.

Conclusiones

Como hemos podido apreciar en los resultados obtenidos durante la aplicación de este método científico podemos ver que este es solo el primer paso hacia el objetivo de lograr una herramienta que pueda asistir al perito en materia de Acústica Forense y que logre de hecho acelerar su trabajo con un elevado porcentaje de efectividad.

Recordemos que de los resultados del presente sistema puede depender el que un perito emita un dictamen que pueda modificar la situación jurídica de un individuo presuntamente responsable de un acto delictivo como puede ser el secuestro, privación ilegal de la libertad, chantajes, fraudes y demás delitos graves que afectan la seguridad de todos los que vivimos en este país, pero también puede depender la libertad de un individuo falsamente acusado de dichos delitos, por lo que se hace patente la responsabilidad con la que se debe de aproximar este problema.

Comentarios por metodología

Como una primera conclusión se dará una opinión sobre cada una de las tres metodologías implementadas y probadas en el presente trabajo, sus puntos fuertes y sus puntos débiles, así como los comentarios que se crean pertinentes para mejorar cada método con fines de que esta investigación sirva como base a futuros trabajos y mejores sistemas que lleven un paso mas allá lo aquí logrado.

Parámetros en el tiempo

Esta metodología nos dio buenos resultados con un muy aceptable nivel de error, mas sin embargo se observo durante las pruebas que los seis distintos parámetros estudiados en este apartado presentan un comportamiento en ocasiones inverso, ya que unos parámetros mejoraban su calidad con una configuración pero otros empeoraban con la misma, pero en general al poner los seis parámetros en una balanza general se obtiene un buen resultado.

Como una posibilidad de mejorar esta metodología y su eficiencia y sobre todo teniendo en cuenta que no nos interesa en gran medida el tiempo empleado en una prueba sino la confiabilidad de los resultados, podríamos implementar distintas configuraciones para cada parámetro de así ser necesario, de tal manera que logremos minimizar el error y la distancia medida en cada parámetro, para de esta manera poder reducir significativamente el error global al utilizar este tipo de aproximaciones.

En general vemos que es un método de simple implementación y que aparentemente no valía la pena ni siquiera de probarlo, logro buenos resultados donde demuestra que es una buena opción y que podría valer la pena conservar dicha metodología para futuras versiones con las mencionadas mejoras así como las que también se incluirán en las recomendaciones generales mas adelante.

Parámetros en el dominio de la frecuencia

En este apartado cabe la pena mencionar que la presente implementación aparentemente fue pobre y no muy bien lograda, ya que se esperaban mucho mejores resultados, de hecho esta metodología es la utilizada a nivel mundial por los peritos en la materia, pero también es importante mencionar que de lo experimentado por los mismos expertos es bien sabido que la calidad del hardware con el que se lleva a cabo la adquisición de la señal es de suma importancia, ya que una vez en el dominio de la frecuencia es fácil

perder el significado de las diversas componentes, así como el ruido en las grabaciones dificulta gravemente esta labor.

Como algunas de las recomendaciones que vale la pena mencionar por lo observado en la presente experimentación es que este método puede fácilmente pasar a un comportamiento mas aleatorio que predecible, por lo que se hace patente la necesidad de un mucho mejor algoritmo de localización del tono fundamental, que después de todo es uno de los parámetros básicos para poder localizar he identificar las estructuras formanticas, pudiendo incluso utilizarse métodos como LPC para la determinación de este tono fundamental.

También es importante mencionar que los márgenes permitidos de variación intra-personal no es igual para el tono fundamental que para las formantes, por lo que un trato especial a este parámetro es sumamente aconsejable, ya que de los resultados aquí observados vemos que es sumamente fácil exceder este margen permitido si el procesamiento y algoritmo de detección no funciona óptimamente.

En general la técnica es muy basta y sumamente útil, pero aparentemente de difícil implementación, ya que hay que cuidar multitud de aspectos que llevan a mecanismos complejos de detección y procesado, de esto nace la idea de que seria apropiado procesar la señal en múltiples ocasiones o en diferentes capas, buscando en cada proceso un parámetro basado en los resultados del proceso anterior y verificando la valides de todos los resultados obtenidos.

Otro aspecto que seria importante es el tratar de implementar un algoritmo tal que pueda detectar si se trata de un sonido sonoro o sordo, ya que recordemos que los sonidos sordos no tienen las estructuras formanticas básicas y que se pueden identificar por un sistema de esta índole, por lo que al aplicar el presente algoritmo a sonidos sordos el sistema puede tender a identificar estructuras que nos son validas o que no existen en lo absoluto.

Parámetros de modelos LPC

Esta técnica muestra muy buenos resultados a pesar de que su origen es para un uso totalmente distinto, utilizándose principalmente para algoritmos de sintetizado de voz o en algoritmos de reconocimiento de habla (no del hablante) por lo que tiende a ser fácil que el modelo de una persona se pueda adaptar a otra persona distinta, tendiendo al error, por lo que se debe tener especial cuidado a los márgenes donde será considerada como identificación positiva o negativa.

En general el algoritmo es bueno y muy rápido, pero podría ser una buena idea el tratar de probar otras metodologías mas avanzadas de obtención de los modelos LPC para ver si mejora su eficiencia, ya que recordemos que a partir del método empleado en el presente sistema se han basado numerosos métodos mas modernos y mejorados, pero de los cuales hay poca información disponible, ya sea porque el realizador lo esta empleando comercialmente y no desea liberarlo o por lo nuevo del algoritmo en si, pero seria una línea de investigación que puede valer la pena.

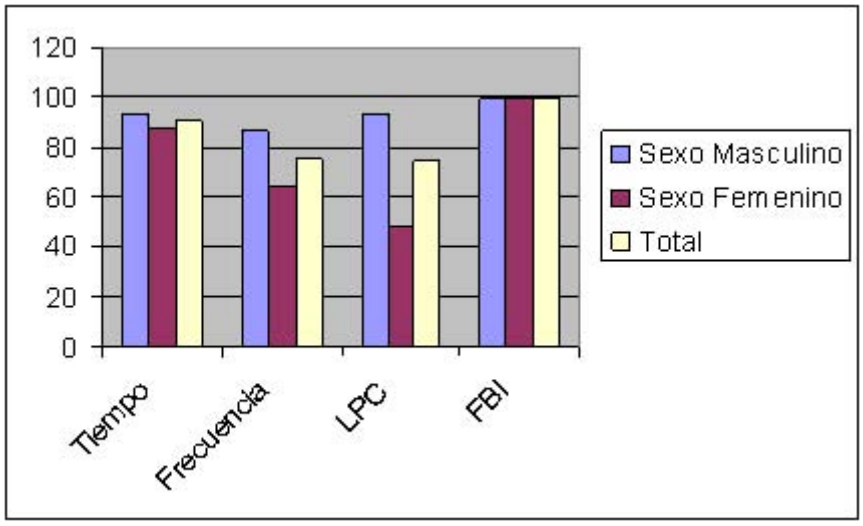
Sumario de resultados

Aquí se da un breve resumen de los resultados obtenidos por el presente sistema, ya que una vez cubiertos los detalles observados en cada metodología y las sugerencias

generales para mejorar la eficiencia que se verán mas adelante solo queda dar un comentario final para cerrar el presente trabajo.

Recordemos que el punto de comparación que se esta utilizando del 99.65% de eficiencia no es para un sistema automático de identificación como el que se desarrollo aquí, sino para un sistema de trabajo mixto que utiliza un sistema automatizado como orientación rápida para sucesivamente dejar la decisión final a un perito en la materia con el uso de espectrógrafos como los ya descritos, por lo que no se esperaba lograr de un primer intento el alcanzar o rebasar dicho porcentaje de eficiencia.

En general las pretensiones del sistema son mas que nada el establecer una base teórica para que el interesado en la materia tenga una buena idea sobre los temas que son necesarios para un mejor entendimiento de los resultados, ya que un sistema de esta índole en manos de una persona sin ningún conocimiento de causa del porque se comporta de una manera una señal de voz se le dificultara el entender los resultados y métodos aquí empleados y desarrollados.



Resumen de eficiencias

En la figura vemos un resumen grafico que compara los resultados obtenidos, la altura de las columnas representa el porcentaje y se tienen las tres metodologías implementadas y la eficiencia de los laboratorios del FBI que tienen prácticamente el 100% que seria el nivel óptimo de identificación en esta pericial.

De esta podemos ver que si tomamos únicamente sujetos del sexo masculino (que son la mayoría de los casos tratados en las ciencias forenses) vemos que logramos una eficiencia del 93.16% en las metodologías de LPC y parámetros en el dominio del tiempo, que son muy aceptables, pero para ambos sexos en el dominio del tiempo tenemos por ahora la mejor opción de análisis.

Estos resultados son alentadores, ya que a pesar de que la presente investigación no fue breve se logro un resultado aceptable y si sienta una firme base para futuras mejoras en la implementación así como dejar abiertas varias líneas de investigación con las que se presume se puede obtener un resultado bastante mejor, ya que como se ha mencionado ya en varias ocasiones, en las ciencias forenses no hay mucho lugar para el error si

queremos poder conciliar el sueño por las noches, ya que de nuestro trabajo puede depender la situación jurídica de una o mas personas.

A este respecto me gustaría tomar prestada una frase que es bien conocida en el ámbito de las ciencias forenses y que dice así:

> Si la ley te pide que opines como perito, nunca dejes de ser hombre de ciencia; tu misión no es de vengar a nadie, no es salvar a un inocente o aniquilar a un culpable, es solo encontrar la verdad científica.
>
> Georges Burgess

Georges Burgess es uno de los criminalistas mas conocidos y autor de varios libros en su país natal, Estados Unidos de Norte América, y esta es una de sus frases mas celebres que no puede expresar de manera mas clara lo que persigue un perito en cualquier materia, se persigue únicamente la verdad científica.

Es importante también mencionar que las condiciones de trabajo normales para esta ciencia pericial serán de manera común mucho mas difíciles que las aquí trabajadas, ya que las grabaciones suelen tener mala calidad, bajos niveles de grabación, muchos ruidos ajenos a la voz en estudio, otras voces solapadas con la voz en estudio, poco material de estudio y demás condiciones que dificultan aun mas un buen resultado, pero por esto mismo es necesario de primera instancia lograr un sistema que pueda trabajar eficientemente con señales de "buena" calidad para después adaptarlo y mejorarlo para su uso real, ya que no seria muy útil un sistema que trabaje correctamente únicamente en condiciones ideales, como son grabación en cabina, ancho de banda amplio y una sola voz.

Existen también otros métodos que aquí ya no se abordaron, como pueden ser las cadenas ocultas de Markov, diversas variaciones de extracción de parámetros LPC o incluso la combinación de varias metodologías, existen actualmente métodos que emplean la detección del tono fundamental por medio de parámetros LPC y la energía de la señal como parámetros de discriminación con buenos resultados, por lo que vemos que existen múltiples caminos que nos pueden llevar a mejorar los resultados aquí obtenidos y que vale la pena adentrarse pues prometen buenos resultados.

Como una ultima sugerencia por lo ya vivido durante las pruebas del presente sistema es que es sumamente recomendable la implementación de módulos que permitan el automatizar lo mas posible los procesos que ofrece el sistema, ya que uno de los aspectos mas largos de las presentes pruebas fue el proceso manual de cada uno de los archivos con las diversas opciones para obtener los márgenes de variación en los que el sistema establecerá el umbral para dar un resultado positivo o negativo, así como la obtención de la configuración optima con la que se debe de trabajar las diversas pruebas.

Como se observo el proceso de detección de la configuración óptima es largo y engorroso, ya que hay que probar una multitud de archivos con una gran cantidad de posibilidades. Estas pruebas nos permiten en general observar si un proceso

determinado ayuda o perjudica a los resultados, pero en ocasiones la combinación y permutación de las diversas opciones es la que nos ofrece el mejor resultado, razón por la cual no es fácil determinar cual es la mejor opción de configuración.

Seria conveniente para la determinación de la configuración ideal, un modulo de calibración que nos permita hacer el proceso completo de establecer los márgenes y probar los resultados completos con cada configuración que nos pueda parecer de interés o incluso con todas las opciones posibles para ver cual nos da el mayor porcentaje de aciertos, pero este proceso puede llevar varios miles de proceso y comparaciones, por lo que un modulo que pueda realizar esto sin intervención del usuario seria muy útil y ahorraría mucho tiempo, además de que dicho modulo puede llevar un control estadístico que sea útil al investigador para discernir si se requiere mayor profundización en un parámetro, corregir algún algoritmo o demás aportaciones que tienen un gran valor cuando se quiere mejorar el rendimiento del sistema.

Por otra parte se puede observar que el establecer los márgenes para fijar un umbral de decisión es también un tanto engorroso, ya que mientras mayor sea el numero de pruebas para establecer dicho umbral mejor será el resultado, por lo que se ve que es sumamente aconsejable el implementar que el sistema de manera automática pueda tomar una serie de muestras y analizarlas en busca de un umbral de buena calidad, pero al mismo tiempo que pueda seguir enriqueciéndose con cada nueva voz que se alimente al sistema, ya que en general al tratarse de un proceso estadístico para ver que márgenes deben ser aceptables, mientras mayor sea la población mejores resultados podremos obtener, además este modulo puede ser de gran utilidad para implementar un modulo final de decisión automático.

Como un ultimo paso de estos módulos de automatización, es sumamente aconsejable el implementar un modulo de decisión que nos de un resultado mas simple y de fácil comprensión para un usuario final, ya que del presente sistema se nota la dificultad de toma de decisiones ya que el programa únicamente expresa una serie de distancias, mas nunca una decisión u opinión de si se trata o no de la misma persona, en este tenor se podría utilizar quizás un sistemas de redes neuronales, ya que este tipo de algoritmos de inteligencia artificial son sumamente comunes para toma de decisiones en sistemas de proceso de voces como los programas comerciales de dictado.

Problemas pendientes

Como un punto fundamental y teniendo en cuenta lo complejo del tema en comento se hace patente la necesidad de una mejor continua, así como de la implementación de nuevas tecnologías y algoritmos, sin dejar a un lado la mejora en los algoritmos ya implementados desde el punto de vista informático, lo cual nos lleva a una serie de sugerencias generales

Medición de distancias

Una de las cosas que salta a la vista es que a pesar de que los resultados son buenos se podría mejorar significativamente los resultados mediante la implementación de mediciones de distancia mas avanzadas como podría ser alguna de las distancias de Itakura, que incluso contempla la comparación de vectores de 'n' dimensiones pero de distinto tamaño entre si, hecho que aquí fue compensado en parte por la interpolación de archivos en el dominio del tiempo.

El principal problema es que la distancia Euclidea utilizada en el presente trabajo no es optima para la comparación de vectores de 'n' dimensiones, da muy buenos resultados por ejemplo en la medición de distancias entre puntos en un plano cartesiano o incluso en la distancia entre dos puntos en el espacio, pero seria una buena idea el implementar mediciones mas avanzadas para mejorar los presentes resultados.

$$D(\hat{a},a)=(\hat{a}-a)\left[N\frac{R}{\hat{a}R\hat{a}^{t}}\right](\hat{a}-a)^{t} \quad (3.1)$$

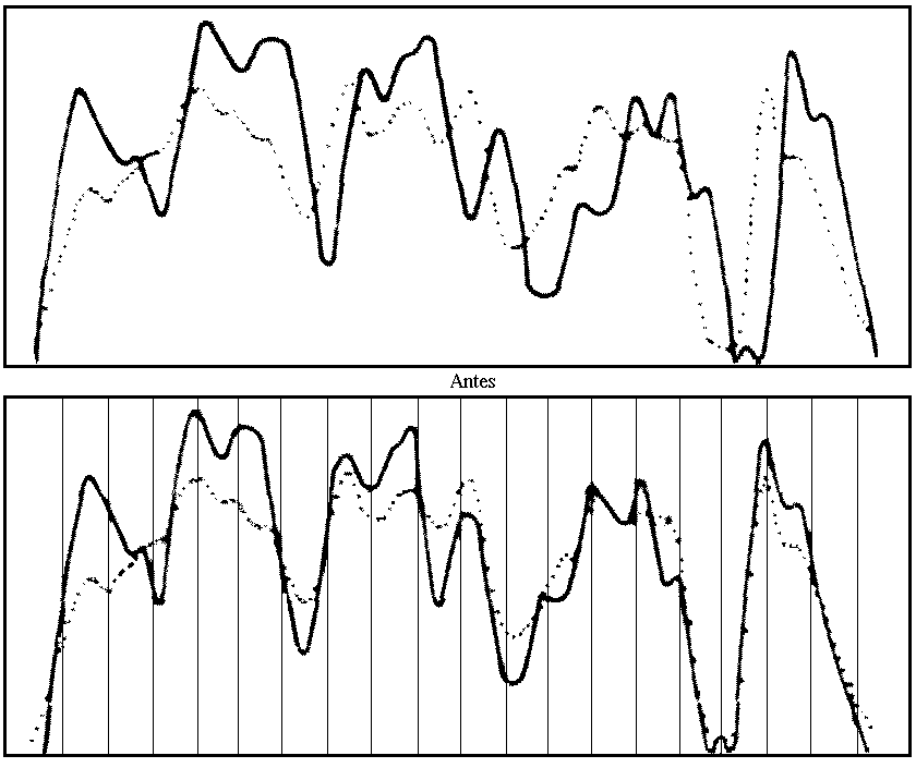
La ecuación anterior es una de las distancias propuestas por Itakura para la medición de distancia entre dos modelos LPC, en esta 'R' es la matriz de correlación de dimensiones 'p+1' donde 'p' es el numero de parámetros calculados en el modelo, los vectores 'a' representan dos muestras de voz previamente procesadas y modeladas y de las que se desea saber que tanta distancia existe entre ambas, por supuesto, mientras menor sea la distancia mayor semejanza existe entre ambas muestras, siendo la distancia cero si las muestras son idénticas.

Esto medida se propone debido a que para la medición de distancia entre parámetros LPC se sugiere una metodología totalmente distinta debido a las características propias de dichos modelos, pero Itakura propone una serie de medidas que se pueden adaptar prácticamente a cualquier situación y tipo de cantidades a comparar, pues además es importante comprender también que dependiendo de que se mide puede variar el peso que se deberá de asignar a dicha medida, y no debido a la importancia ponderada por uno mismo a cada parámetro, por ejemplo en el dominio del tiempo el parámetro de la energía dará números muy grandes en general y en contraparte los cruces por cero dará una cantidad bastante pequeña en comparación, por lo que si se le asigna el mismo peso podemos mermar o ocultar por completo la contribución de dicho parámetro.

Algoritmo "Time Warp"

Otro de los principales problemas en el reconocimiento del locutor es que una persona aunque lo quiera difícilmente podrá producir dos veces la misma palabra sin variaciones, por lo que se hace patente la necesidad de un algoritmo que pueda hacer una deformación no lineal de la escala de tiempo (time warping), estos algoritmos tienen como función el tratar de hacer que los eventos significativos de una señal de audio coincidan de la mejor manera con los eventos significativos de otra señal de

audio, esto tendera a alinear los contornos de un parámetro en especifico, por ejemplo en la figura siguiente podemos apreciar el efecto de un algoritmo de deformación no lineal aplicado al parámetro de la energía de una porción de audio.



Ejemplo de algoritmo "Time Warping"

De la figura anterior vemos que tenemos dos envolventes de la energía de dos señales, vemos que son parecidas, pero al mismo tiempo presentan desplazamientos en el tiempo que hacen no coincidir a los diversos elementos significativos de cada grafica, pero vemos que después de aplicado el algoritmo de deformación no lineal la mayor parte de los elementos significativos coinciden de mucho mejor manera, pero no se modificaron en si las características de ninguna de las señales, tan solo se deforma la escala de tiempo para hacer coincidir una señal con otra, reduciendo significativamente la distancia a lo hora de hacer una medición de distancias para determinar si se tratara de la misma persona.

Detector de silencio

Como un extra al buen funcionamiento del sistema es conveniente la implementación de un algoritmo de detección de silencio, de esta manera al sistema se le puede alimentar con porciones de sonido que contengan pausas y palabras y el programa podría separar las porciones que contengan habla para su estudio y desechar las porciones que no contengan información, de esta manera se puede evitar el que se tenga un error de medición debido a que se presente una porción de grabación que no sea parte de la palabra o palabras en estudio, aspecto que se midió en el presente trabajo con las pruebas de corrimiento en tiempo en que se probaba una misma muestra de voz con ligeros corrimientos de 10 milisegundos para ver como se comportaba el sistema y que tanto variaba sus resultados.

Dicho algoritmo podría ser implementado al hacer un estudio general del audio alimentado al sistema para observar su energía y así poder determinar que las porciones que tengan un nivel de energía por debajo de un umbral sean consideradas como pausas o silencio, pero hay que tener cuidado de que dicha pausa no sea por ejemplo el momento de aspiración para un sonido plosivo, razón que complica un poco dicho algoritmo, ya que deberá de contemplar tanto la baja de niveles de energía como la duración de dicha pausa.

Como un paso final a esta sugerencia se podría implementar el mismo detector de silencio que pueda separar varias palabras de una misma grabación separadas por pausas, pero el aspecto complicado de esta idea es que en el habla continua no existe prácticamente separación alguna entre las diversas palabras, por lo que no es fácil la separación de habla y silencios y podrían unirse varias palabras en un solo bloque al no detectar pausa alguna entre ellas, por lo que aunque seria cómodo y conveniente no es tan simple como podría pensarse de inicio.

Modulo de pre-procesado

Prácticamente para terminar se sugiere en futuras investigaciones la implementación de un modulo de procesado previo, esto es, antes de alimentar la señal incluso al detector de silencio se considera conveniente aplicar un ligero proceso a la señal, en este procesado de la señal se podría implementar una serie de filtros que “limpien” la señal a procesar de ruidos ajenos a la voz en estudio.

Otra buena acción a realizar antes de procesar una señal es la homogeneización de los niveles de grabación, esto debido a que algunos de los parámetros en estudio pueden variar fuertemente debido a factores tales como la distancia entre la boca y el micrófono o simplemente por los niveles de grabación empleados, por lo que seria bastante bueno tener un mismo nivel de grabación para todas las porciones de audio que se pretenden estudiar para disminuir al mínimo las diferencias debido a niveles o equipos de grabación y no a la voz en estudio misma.

A su vez y en general las diversas bibliografías recomiendan la implementación de un filtro de pre-énfasis antes de pasar la señal a la extracción de parámetros, esto puede ser logrado simplemente por medio de un filtro paso altas según la mayoría de los autores, pero pudiendo implementarse también un algoritmo mas complejo con mejores resultados.

Cadenas ocultas de Markov

Las cadenas ocultas de Markov o modelos ocultos de Markov son métodos muy utilizados hoy en día, estos los podemos encontrar prácticamente en todos los programas de dictado y reconocimiento de habla, pero también en los sistemas de reconocimiento de locutores comerciales, pero por su complejidad y las limitantes de tiempo no se incorporo al presente trabajo, así que damos una pequeña explicación de que son dejando al lector y a futuras investigaciones su implementación.

Una cadena de Markov recibe su nombre del matemático ruso Andrei Markov, es una serie de eventos, en la cual la probabilidad de que ocurra un evento depende del evento inmediato anterior, las cadenas de este tipo tienen memoria. "Recuerdan" el último evento y esto condiciona las posibilidades de los eventos futuros. Esta dependencia del evento anterior distingue a las cadenas de Markov de las series de eventos independientes, como tirar una moneda al aire o un dado.

Una vez que tenemos una ligera noción de que es una cadena de Markov o modelo oculto de Markov (HMM) como también se le conoce podemos entrar un poco mas a detalle en el porque es útil.

Se puede decir que las HMM son modelos estadísticos que se extraen de la voz humana, de manera muy similar al LPC aprovechando la característica de que la voz se comporta de una forma quasi estacionaria en periodos cortos de tiempo

Algunas otras razones por la que se emplea de manera muy común es que puede entrenarse o calibrarse de manera automatizada y que es un algoritmo eficiente en términos computacionales, entregando una serie de vectores de *n* dimensiones de números reales como resultado, siendo estos números coeficientes cepstrales que son empleadas para la obtención de la envolvente de una señal de voz, dando un modelo característico para cada fonema, que fue una de las bases teóricas de nuestro planteamiento

Esta técnica se podría a grandes rasgos considerar como una mejora a las técnicas de LPC, existiendo actualmente múltiples implementaciones como VTLN o la normalización por el largo del tracto vocal, MLLR o regresión lineal por máxima semejanza, etc.

Migración a Visual Studio 2008

Por ultimo como puntos propuestos a la mejora se encuentra la migración a lenguajes de programación mas poderosos, versátiles y actualizados, teniéndose como la mejor opción al gusto de un servidor el lenguaje de programación C# (pronunciado C sharp) de Microsoft.

Esta sugerencia obedece a varios puntos que se localizaron a favor, entre estos los siguientes:

a) Este lenguaje es de última generación, siendo liberado apenas hace pocos meses para su uso en general, por lo que evitamos caer en obsolescencia o en lenguajes en franco desuso.
b) Es de uso gratuito en su versión Express sin limitante alguna, únicamente tenemos como limitante que algunas de las opciones de sus hermanos mayores no están presentes, pero se tiene un lenguaje totalmente profesional de manera gratuita.
c) Permite la implementación de aplicaciones en sistemas operativos de última generación como Windows Vista y Windows Server 2008, utilizando además las nuevas tecnologías que ofrecen estos ambientes como Windows Presentation Foundation, Windows Communication Foundation que respectivamente separan la implementación de la interfase grafica y facilitan la comunicación entre aplicaciones.
d) Por medio del CLR Common Lenguaje Runtime permite separar el desarrollo del sistema operativo, pudiendo fácilmente migrarse el sistema a sistemas operativos tales como OS X, Linux y otros más, ya que el CLR es de libre uso y se ha implementado ya en múltiples sistemas operativos, a diferencia de la implementación actual del tipo COM que depende fuertemente del sistema operativo.

e) El desempeño de este nuevo lenguaje es muy superior al de Visual Basic 6, encontrándose únicamente por debajo de C++ pero con un grado de flexibilidad y simplicidad mucho más accesible.
f) La capacitación para el uso de estos nuevos lenguajes de desarrollo es totalmente gratuita y se encuentra en Internet para su difusión en los programas Desarrollador Cinco estrellas y Microsoft Virtual Academy.

Tomando en consideración estos puntos se ve como una muy Buena opción para la migración a un sistema mas moderno y poderoso, aunado a que pueden incorporarse múltiples herramientas mas que permitirán al experto en la material ahorrar mucho tiempo, así como ofrecer incluso a nivel institucional módulos de acceso a controles de avance y gestión, como lo que actualmente me encuentro desarrollando.

Soporte a mayor numero de formatos de audio

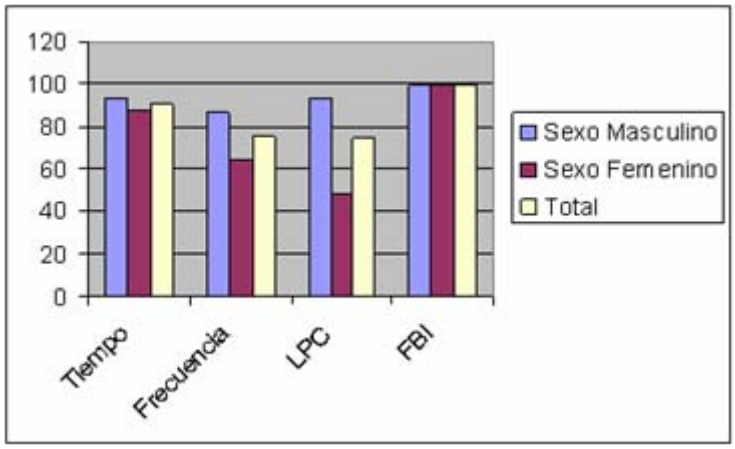
Como un ultimo punto seria sumamente conveniente agregar al sistema la posibilidad de trabajar con archivos Wav con otro tipo de codificaciones a la de Microsoft, ya que a pesar de que esta es la mas utilizada a nivel de análisis forense por sus características ya mencionadas, existen muchos otros modos de codificación y formatos distintos, que para agregar versatilidad y poder al sistema podrían ser contemplados, pero siempre teniendo en cuenta los tipos de archivos que presentan compresión con perdidas, ya que estos deben tratarse con las respectivas reservas.

Conclusiones

Sumario de resultados

A manera de cierre presento un sumario de los resultados obtenidos, donde vale la pena mencionar que se tomo como punto de comparación el 99.65% de eficiencia de los laboratorios de identificación de locutores del FBI, no un sistema automático de identificación como el que se desarrollo aquí, por lo que no se esperaba lograr de un primer intento el alcanzar o rebasar dicho porcentaje de eficiencia.

En general las pretensiones del sistema son más que nada el establecer una base para que el interesado en la materia forme una buena idea de los temas que son necesarios para una buena comprensión de los resultados, ya que un sistema de esta índole en manos de una persona sin conocimiento alguno de cómo se comporta una señal de voz y otros temas ya tratados difícilmente le sacara provecho.



Resumen de eficiencias

En la figura vemos un resumen grafico donde comparamos los resultados obtenidos, la altura de las columnas representa el porcentaje y se grafican las tres metodologías implementadas y la eficiencia de los laboratorios del FBI que tienen prácticamente el 100% que sería el nivel óptimo de identificación en esta pericial.

De esta podemos ver que si tomamos únicamente sujetos del sexo masculino (que son la mayoría de los casos tratados en las ciencias forenses) vemos que logramos una eficiencia del 93.16% en las metodologías de LPC y parámetros en el dominio del tiempo, que son muy aceptables, pero para ambos sexos en el dominio del tiempo tenemos por ahora la mejor opción de análisis.

Estos resultados son alentadores, ya que la presente investigación no fue corta, pero logra un resultado aceptable, asentando una firme base para futuras mejoras en la implementación y nuevas líneas de investigación con las que se puede obtener un mejor resultado, ya que como se ha mencionado ya, en las ciencias forenses no hay mucho lugar para el error si queremos poder conciliar el sueño por las noches, ya que de nuestro trabajo puede depender la situación jurídica de una o más personas.

A este respecto me gustaría tomar prestada una frase que es bien conocida en el ámbito de las ciencias forenses y que dice así:

> Si la ley te pide que opines como perito, nunca dejes de ser hombre de ciencia; tu misión no es de vengar a nadie, no es salvar a un inocente o aniquilar a un culpable, es solo encontrar la verdad científica.
>
> Georges Burgess

Georges Burgess es uno de los criminalistas más conocidos y autor de varios libros en su país natal, Estados Unidos de Norte América, y esta es una de sus frases más celebres que no puede expresar de manera más clara lo que persigue un perito en cualquier materia, **se persigue únicamente la verdad científica**.

Existen también diferentes métodos que no se abordaron, como pueden ser las cadenas ocultas de Markov, diversas variaciones de extracción de parámetros LPC o incluso la combinación de varias metodologías, existen actualmente métodos que emplean la detección del tono fundamental por medio de parámetros LPC y la energía de la señal como parámetros de discriminación con buenos resultados, por lo que vemos que existen múltiples caminos que nos pueden llevar a mejorar los resultados aquí obtenidos.

Sugerencias generales

En miras de una mejora continua, así como de la implementación de nuevas tecnologías y algoritmos, sin dejar a un lado la mejora en los algoritmos ya implementados desde este primer intento, se tienen las siguientes sugerencias generales.

1 - Separación por género

Como una primera sugerencia y como se vio en los resultados, la separación de individuos del sexo masculino y femenino es altamente recomendable, esto en virtud de que los resultados en los márgenes personales para descartar o aceptar una identidad son significativamente distintos.

Esta idea no es de asombrarse, incluso programas comerciales de dictado llevan a cabo esta separación, siendo el primer paso para su configuración el determinar si el usuario es de sexo masculino o femenino para posteriormente pasar al entrenamiento o calibración del sistema.

Aunado a esto encontramos la opinión general de que no es una buena idea el querer englobar en un solo grupo a ambos sexos, y después de todo no es tan extraño, ya que la diferencia en el tono fundamental es significativa, tendiendo a ser la voz femenina más aguda y armónica que la masculina, pero con mayores fluctuaciones de frecuencia, por lo que merece darle una atención especial con el fin de evitar errores en los resultados.
Esta diferencia entre sexos se hace necesaria al observar que el sistema tiende a bajar su eficiencia de manera significativa con sujetos del sexo femenino, teniéndose una diferencia de 5.79% en el análisis en el tiempo, 22.63% en el análisis en frecuencia y un alarmante 44.74% en modelos LPC, hecho que hace que valga la pena que en futuras investigaciones se dé especial énfasis a localizar las causas de esto.

Ahora bien, del hecho de que se trabajaron voces filtradas a 8,000 Hz que es la frecuencia de corte de una línea telefónica estándar, podemos formular la hipótesis de que al ser las voces femeninas más agudas, la línea telefónica puede estar filtrando una

considerable cantidad de información que sería necesaria para el correcto estudio y comparación, factor que lleva a que la computadora "vea" mucho mas similares la voz de dos mujeres distintas que las de dos hombres.

2 - Medición de distancias

Una de las cosas que salta a la vista es que a pesar de que los resultados son buenos se podría mejorar significativamente los resultados mediante la implementación de mediciones de distancia más avanzadas como podría ser alguna de las distancias de Itakura, que incluso contempla la comparación de vectores de 'n' dimensiones pero de distinto tamaño entre sí, hecho que aquí fue compensado en parte por la interpolación de archivos en el dominio del tiempo.

El principal problema es que la distancia Euclidea utilizada en el presente trabajo no es optima para la comparación de vectores de 'n' dimensiones, da muy buenos resultados por ejemplo en la medición de distancias entre puntos en un plano cartesiano o incluso en la distancia entre dos puntos en el espacio, pero sería una buena idea el implementar mediciones más avanzadas para mejorar los presentes resultados.

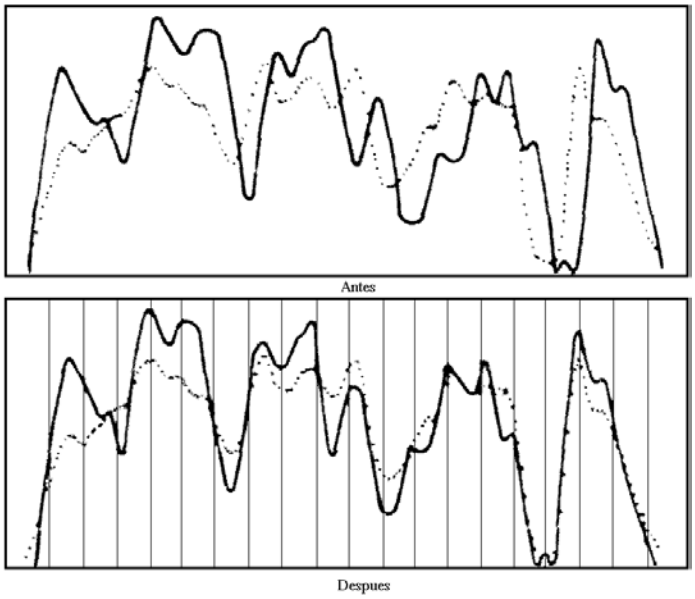
$$D(\hat{a}, a) = (\hat{a} - a)\left[N \frac{R}{\hat{a} R \hat{a}^t} \right](\hat{a} - a)^t \quad (3.1)$$

La ecuación anterior es una de las distancias propuestas por Itakura para la medición de distancia entre dos modelos LPC, en esta 'R' es la matriz de correlación de dimensiones 'p+1' donde 'p' es el numero de parámetros calculados en el modelo, los vectores 'a' representan dos muestras de voz previamente procesadas y modeladas y de las que se desea saber que tanta distancia existe entre ambas, por supuesto, mientras menor sea la distancia mayor semejanza existe entre ambas muestras, siendo la distancia cero si las muestras son idénticas.

Esto medida se propone debido a que para la medición de distancia entre parámetros LPC se sugiere una metodología totalmente distinta debido a las características propias de dichos modelos, pero Itakura propone una serie de medidas que se pueden adaptar prácticamente a cualquier situación y tipo de cantidades a comparar, pues además es importante comprender también que dependiendo de que se mide puede variar el peso que se deberá de asignar a dicha medida, y no debido a la importancia ponderada por uno mismo a cada parámetro, por ejemplo en el dominio del tiempo el parámetro de la energía dará números muy grandes en general y en contraparte los cruces por cero dará una cantidad bastante pequeña en comparación, por lo que si se le asigna el mismo peso podemos mermar o ocultar por completo la contribución de dicho parámetro.

3 - Algoritmo "Time Warp"

Otro de los principales problemas en el reconocimiento del locutor es que una persona aunque lo quiera difícilmente podrá producir dos veces la misma palabra sin variaciones, por lo que se hace patente la necesidad de un algoritmo que pueda hacer una deformación no lineal de la escala de tiempo (time warping), estos algoritmos tienen como función el tratar de hacer que los eventos significativos de una señal de audio coincidan de la mejor manera con los eventos significativos de otra señal de audio, esto tendera a alinear los contornos de un parámetro en especifico, por ejemplo en la figura siguiente podemos apreciar el efecto de un algoritmo de deformación no lineal aplicado al parámetro de la energía de una porción de audio.



Ejemplo de algoritmo "Time Warping"

De esta vemos que tenemos dos envolventes de la energía de dos señales, vemos que son parecidas, pero al mismo tiempo presentan desplazamientos en el tiempo que hacen no coincidir a los diversos elementos significativos de cada grafica, pero vemos que después de aplicado el algoritmo de deformación no lineal la mayor parte de los elementos significativos coinciden de mucho mejor manera, pero no se modificaron en si las características de ninguna de las señales, tan solo se deforma la escala de tiempo para hacer coincidir una señal con otra, reduciendo significativamente la distancia a lo hora de hacer una medición para determinar si se tratara de la misma persona.

4 - Detector de silencio

Como un extra al buen funcionamiento del sistema es conveniente la implementación de un algoritmo de detección de silencio, de esta manera al sistema se le puede alimentar con porciones de sonido que contengan pausas y palabras y el programa podría separar las porciones que contengan habla para su estudio y desechar las porciones que no, de esta manera se puede evitar el que se tenga un error de medición debido a que se presente una porción de grabación que no sea parte de la palabra o palabras en estudio, aspecto que se midió en el presente trabajo con las pruebas de corrimiento en tiempo en que se probaba una misma muestra de voz con ligeros corrimientos de 10 milisegundos para ver como se comportaba el sistema y que tanto variaba sus resultados.

Dicho algoritmo podría ser implementado al hacer un estudio general del audio alimentado al sistema para observar su energía y así poder determinar que las porciones que tengan un nivel de energía por debajo de un umbral sean consideradas como pausas o silencio, pero hay que tener cuidado de que dicha pausa no sea por ejemplo el momento de aspiración para un sonido plosivo, esto complica dicho algoritmo, ya que deberá de contemplar tanto la baja de niveles de energía como la duración de dicha pausa.

Como un paso final a esta sugerencia se podría implementar el mismo detector de silencio que pueda separar varias palabras de una misma grabación separadas por pausas, pero el aspecto complicado de esta idea es que en el habla continua no existe prácticamente separación alguna entre las diversas palabras, por lo que no es fácil la

separación de habla y silencios, pudiendo unirse varias palabras en un solo bloque al no detectar pausa alguna entre ellas, por lo que aunque seria cómodo y conveniente no es tan simple como podría pensarse de inicio.

5 - Modulo de pre-procesado

Se sugiere también para futuras investigaciones la implementación de un modulo de procesado previo, esto es, antes de alimentar la señal incluso al detector de silencio se considera conveniente aplicar un ligero proceso a la señal, en este procesado de la señal se podría implementar una serie de filtros que "limpien" la señal a procesar de ruidos ajenos a la voz en estudio.

Otra buena acción a realizar antes de procesar una señal es la homogeneización de los niveles de grabación, esto debido a que algunos de los parámetros en estudio pueden variar fuertemente debido a factores tales como la distancia entre la boca y el micrófono o simplemente por los niveles de grabación empleados, por lo que sería bueno tener un mismo nivel de grabación para todas las porciones de audio que se pretenden estudiar, disminuyendo así las diferencias debido a niveles o equipos de grabación.

A su vez y en general las diversas bibliografías recomiendan la implementación de un filtro de pre-énfasis antes de pasar la señal a la extracción de parámetros, esto puede ser logrado simplemente por medio de un filtro paso altas según la mayoría de los autores, pero pudiendo implementarse también un algoritmo más complejo con mejores resultados.

6 - Cadenas ocultas de Markov

Las cadenas ocultas de Markov o modelos ocultos de Markov son métodos muy utilizados hoy en día, estos los podemos encontrar prácticamente en todos los programas de dictado y reconocimiento de habla, pero también en los sistemas de reconocimiento de locutores comerciales, pero por su complejidad y las limitantes de tiempo no se incorporo al presente trabajo, así que damos una pequeña explicación de que son dejando al lector su implementación.

Una cadena de Markov recibe su nombre del matemático ruso Andrei Markov, es una serie de eventos, en la cual la probabilidad de que ocurra un evento depende del evento inmediato anterior, las cadenas de este tipo tienen memoria. "Recuerdan" el último evento y esto condiciona las posibilidades de los eventos futuros. Esta dependencia del evento anterior distingue a las cadenas de Markov de las series de eventos independientes, como tirar una moneda al aire o un dado.

Una vez que tenemos una ligera noción de que es una cadena de Markov o modelo oculto de Markov (HMM) como también se le conoce podemos entrar un poco más a detalle en el porqué es útil.

Se puede decir que las HMM son modelos estadísticos que se extraen de la voz humana, de manera muy similar al LPC aprovechando la característica de que la voz se comporta de una forma quasi estacionaria en periodos cortos de tiempo

Algunas otras razones por la que se emplea de manera muy común es que puede entrenarse o calibrarse de manera automatizada y que es un algoritmo eficiente en términos computacionales, entregando una serie de vectores de *n* dimensiones de

números reales como resultado, siendo estos números coeficientes cepstrales que son empleadas para la obtención de la envolvente de una señal de voz, dando un modelo característico para cada fonema, que fue una de las bases teóricas de nuestro planteamiento

Esta técnica se podría a grandes rasgos considerar como una mejora a las técnicas de LPC, existiendo actualmente múltiples implementaciones como VTLN o la normalización por el largo del tracto vocal, MLLR o regresión lineal por máxima semejanza, etc.

Comentario final

Como un último comentario se invita a los que dieron lectura a esta tesis a continuar la investigación y desarrollo en esta materia, ya que además de ser sumamente interesante recordemos que tiene como finalidad proteger a nuestros seres queridos y salvaguardar la procuración e impartición de justicia en México, existen muchas mejoras y posibles soluciones, por lo que solo la renovación e investigación constante nos llevaran a esa meta de conseguir un desarrollo de excelencia y profesionales altamente capacitados y profesionales.

Bibliografía

[1] Corpora for the Evaluation of Speaker Recognition Systems
Joseph P. Campbell Jr. y Douglas A. Reynolds
IEEE 0-7803-5041-3/99

[2] Sitio Internet de Crystal, parte del corporativo Cirrus Logic
www.crystal.com

[3] Manual del usuario Computadora Personal 300PL 37L4-470
IBM Corporation

[4] Manual Micrófono Shure 12A
Shure Brothers Incorporated

[5] Fonética Acústica de la Lengua Española
Antonio Quilis
Ed. Gredos

[6] Speaker Recognition: A Tutorial
Joseph P. Campbell, Jr.
IEEE, Vol. 85 No. 9, septiembre 1997

[7] Probabilidad y Estadística
George C. Canavos
Ed. McGraw Hill, 1986

[8] Probabilidad y Estadística para Ingenieros
Irwin Millar, John E. Freund
Ed. Prentice Hall, 1986

[9] Reconocimiento de Comandos Verbales Utilizando Cuantización Vectorial y Redes Neuronales
Ricardo Barron Fernández, Sergio Suárez Guerra
Centro de Investigación en Computación, IPN, 1999

[10] Speaker Recognition: A Tutorial
Joseph P. Campbell, Jr.
IEEE, Vol. 85 No. 9, septiembre 1997

[11] Discrete-Time Processing of Speech Signals
John R. Deller Jr., John G. Proakis
Ed. Prentice Hall

[12] Speech Coding
Thomas P. Barnasell
Ed. Georgia Tech

[13] Practical Aproches to Speech Coding
Panus E. Papamichalis
Ed. Prentice Hall

[14] Laboratorio de Criminalística
Zonderman
Ed. Limusa, 1993