UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA
INGENIERIA ELÉCTRICA – TELECOMUNICACIONES

SISTEMA DE IDENTIFICACIÓN DE LENGUAS INDÍGENAS PARA LA INDEXACIÓN DE AUDIO

TESIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRO EN INGENIERÍA

PRESENTA:
BENJAMÍN JOSÉ LÓPEZ GARCÍA

TUTOR

DR. VICTOR GARCIA GARDUÑO

MÉXICO, D. F. FEBRERO 2014

JURADO ASIGNADO:

Presidente:	Dr. García Ugalde Francisco
Secretario:	Dr. Matías Maruri José María
Vocal:	Dr. García Garduño Víctor
1 er. Suplente:	Dr. Psenicka Bohumil
2 d o. Suplente:	Dr. Gómez Castellanos Javier

Lugar o lugares donde se realizó la tesis: Ciudad Universitaria, México, D.F.

TUTOR DE TESIS:

Dr. García Garduño Víctor

_____

FIRMA

DEDICATORIA

Mis más sinceros agradecimientos al
Dr Víctor García Garduño
por su orientación, dedicación y apoyo en la
realización de este trabajo.

Agradezco a Dios por permitirme realizar este trabajo y por todas las bendiciones que me da.

Agradezco a mi familia y amigos por estar a mi lado y el apoyo que me brindan día con día.

Benjamín López García

Índice.

Capítulo 1

Estado del arte.

1.1 Introducción

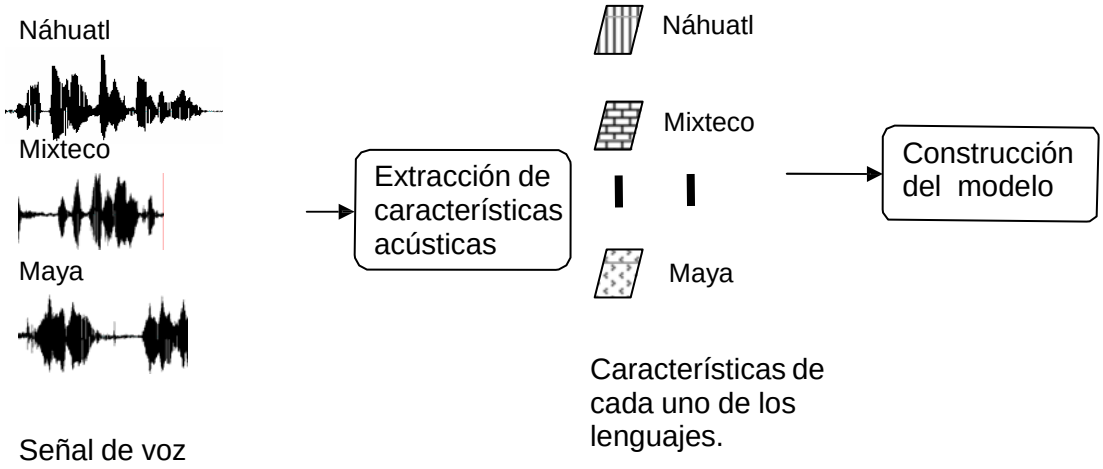
Nuestra investigación está orientada a no depender de la estructura gramatical y fonética propia de cada lenguaje para su identificación, por lo tanto nuestro objetivo consiste en explotar métodos de extracción de características de bajo y mediano nivel como MFCC (Mel-frequency cepstral coefficients), energía en bandas críticas, Pitch, etc, obteniendo información directamente de la señal de voz que nos permita obtener iguales o mejores porcentajes de identificación del lenguaje hablado que los métodos propuestos hasta ahora, sin utilizar las características fonéticas propias de cada dialecto, así como reduciendo el tiempo de ejecución y costos, ya que la identificación por fonemas es costosa en el sentido de tiempo y recursos que necesita una máquina para hacer el procesamiento.[4]

Nuestro trabajo aportará:

- Un método para extracción de características acústicas para la identificación del lenguaje hablado.
- Un método de identificación del lenguaje hablado que no utilice reconocimiento fonético.

Con base en el diagrama de componentes básicos para la identificación del lenguaje hablado sin representación fonética -ver figura I, se realizó el trabajo en dos pasos principales: el primero dedicado al procesamiento acústico de la señal de voz (extracción de características) y el segundo, enfocado a la identificación del lenguaje.[3]

ENTRENAMIENTO



CLASIFICADOR

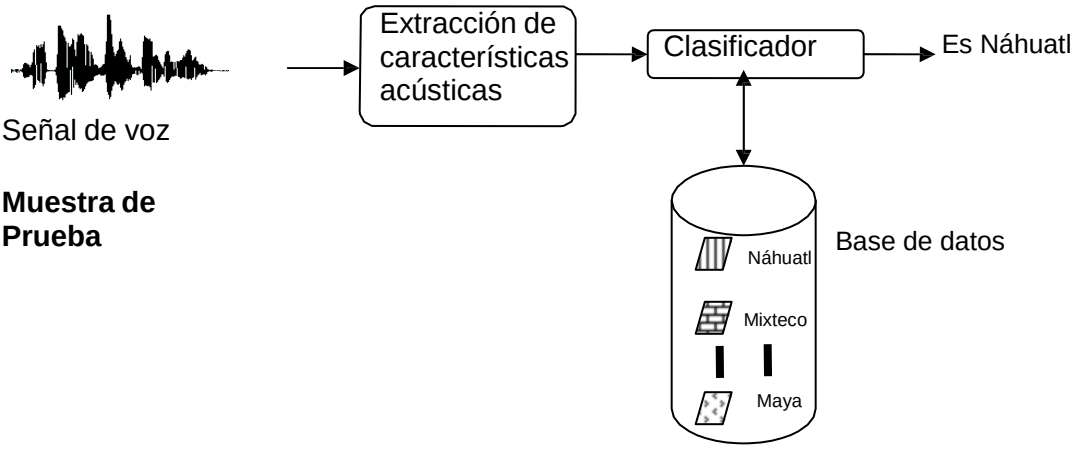


Figura I. Componentes básicos para la identificación del lenguaje hablado sin representación fonética.

1.2 Panorama general

Varias aplicaciones audiovisuales y servicios nuevos fueron posibles basados en el análisis de contenido de audio y la descripción. Los motores de búsqueda o filtros específicos pueden utilizar la descripción extraída y ayudar a los usuarios a navegar a través de grandes colecciones de datos. El análisis digital puede discriminar si un archivo de sonido contiene el habla, la música u otras entidades de audio, cuántos altavoces están contenidos en un segmento de voz, qué género son y hasta sobre de que están hablando las personas.

El contenido hablado puede ser identificado y convertido en texto. La música puede ser clasificada en categorías, tales como jazz, rock, clásica, etc. [14] A menudo es posible identificar una pieza de música, incluso cuando es realizada por diferentes artistas. Puede ser posible identificar sonidos particulares, tales como explosiones, disparos, etc. Por último, y el problema en el que nos vamos a enfocar es la identificación del lenguaje hablado, que consiste en determinar el idioma de quien está hablando, basándose solamente en muestras de voz sin considerar al hablante.

Usamos el término **"audio"** para todo tipo de señales de sonido, tales como el habla, la música, así como señales de sonido más generales y sus combinaciones. Nuestro objetivo principal es entender cómo la información significativa se puede extraer de las formas de onda del audio digital, con el fin de comparar y clasificar los datos de manera eficiente. Cuando esta información se extrae, también puede a menudo ser almacenada como descripción del contenido de una manera compacta. [1]

Estos descriptores compactos son de gran utilidad no sólo en el almacenamiento de audio y aplicaciones de recuperación, sino también para una eficiente clasificación basada en el contenido, reconocimiento, navegación o filtrado de datos. Un descriptor de datos a menudo se llama "vector de características" o "huella digital" y el proceso para la extracción de estos vectores de características o las huellas digitales de audio se denomina "extracción de características de audio" o de "reconocimiento de audio". Por lo general, una variedad de descripciones más o menos complejas pueden ser extraídas y guardadas en una huella digital de datos de audio. La eficiencia de una huella particular utilizada para la comparación y clasificación depende en gran medida de la aplicación, el proceso de extracción y la riqueza de la propia descripción. [1]

Las técnicas actuales de búsqueda de información han logrado gran éxito en su aplicación a documentos de texto, que es atestiguado por los enormes beneficios comerciales generados por las compañías de motores de búsqueda como Google y Yahoo!. En comparación, la recuperación de multimedia se encuentra en una fase inicial y los productos existentes o herramientas no han ofrecido satisfacción a los usuarios y por lo tanto no son populares en comparación con la de los motores de búsqueda basados en texto. En particular, la recuperación de clips de audio que no tienen anotaciones de texto, una función importante en muchas aplicaciones potenciales, es todavía un problema no resuelto desde el punto de vista comercial.

Los actuales sistemas de recuperación de audio dependen en gran medida de las anotaciones de texto en el tratamiento de los datos de audio. Estas anotaciones incluyen metadatos estructurados y no estructurados, por ejemplo, el título, el cantante y a veces la letra en el caso de una canción. La recuperación de archivos de audio basado en sus textos asociados esencialmente puede ser manejada de la misma manera como la recuperación de documentos de texto. A diferencia de las páginas web, a partir de las palabras clave que puede ser extraída de forma automática por medio de algoritmos, la extracción de anotaciones de texto dentro de archivos de audio puede ser difícil y propensa a errores. Por otra parte, en realidad, sólo una fracción de todos los archivos de audio, se han anotado manualmente por los usuarios, y estas anotaciones podrían estar mal para ser útiles. Así, la recuperación de audio basado en estos metadatos tiene una aplicabilidad limitada y poca fiabilidad. Y en el peor de los casos, la búsqueda de un clip sin metadatos de texto necesita de una gran paciencia y determinación por parte del usuario, incluso cuando la colección de audios en la cual se debe buscar es razonablemente pequeña, digamos 200 clips.

Además de la recuperación basada en texto, una alternativa es “content-based”, recuperación de audio basado en métricas de similitud de contenido. Por ejemplo, ha habido un hilo activo de trabajo en la consulta de música por tarareo [14] o el reconocimiento de voz en el que la consulta de entrada es un breve pasaje de la música tarareada por el usuario. El motor de búsqueda realiza una búsqueda por similitud del audio. Si bien estos trabajos lograron algunos avances notables, que trata clips de audio sin anotaciones de texto, en general, sigue siendo una tarea difícil debido a la alta dimensionalidad de características del audio, así como la falta de claridad y subjetividad

en la similitud de contenidos, que depende en gran medida del usuario y la consulta en cuestión.[1][2]

Aunque el problema de la búsqueda de cadenas de audio es relativamente nuevo, está relacionado con una serie de problemas de investigación anteriores. Los sistemas desarrollados para búsqueda de vídeo basada en audio o subtítulos pueden ser eficaces pero a menudo asumen una secuencia de texto asociado, o un flujo de audio limpio. La recuperación de la información a través de audio ha producido recientemente varios enfoques comerciales, sin embargo, estos métodos se centran generalmente en condiciones de grabación relativamente limpias y de un solo orador.

Métodos alternativos han considerado formas de comprimir el tiempo o modificar el habla con el fin de dar a oyentes humanos la capacidad de pasar con mayor rapidez a través de los datos grabados de audio. En general, los sistemas de detección de palabras clave se pueden utilizar para aplicaciones de tema o en lo general. Sin embargo, para la búsqueda de una frase, el sistema debe ser capaz de recuperarse de errores, tanto del usuario que solicita la secuencia de texto como de la jerarquización de los sitios detectados con la frase dentro de la cadena.[1]

Algunas formas de clasificación se muestran en el siguiente diagrama:

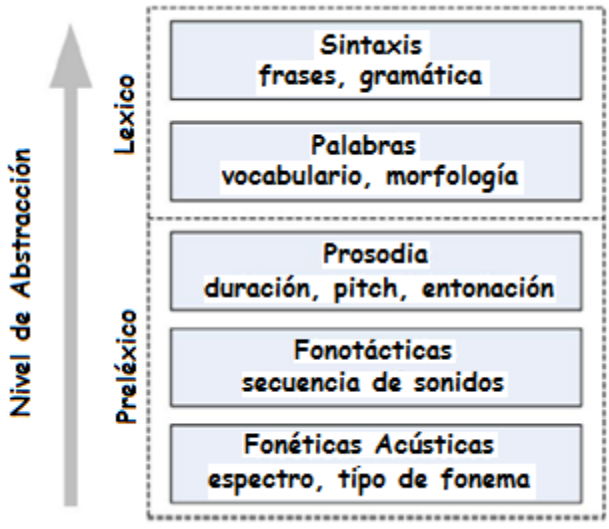


Figura 1.1 formas de clasificar audios

1.3 Lenguas indígenas en México

México es el país con mayor cantidad de personas hablantes de lenguas amerindias en América, con un total de 65 lenguas vivas registradas en el año 2010. Sin embargo, en números relativos, la proporción de estas comunidades lingüísticas es menor en comparación con países como Guatemala (52,8%) y Perú (35%) e incluso con Ecuador (9,4%) y Panamá (8,3%). A excepción hecha del náhuatl, ninguna de las lenguas indígenas de México posee más de un millón de hablantes. El náhuatl es la cuarta lengua indígena de América por el tamaño de su comunidad lingüística, detrás del quechua, el aymara y el guaraní.[6][7]

1.3.1 Clasificación de las lenguas indígenas

El estudio de las lenguas indígenas comenzó desde la llegada misma de los españoles al territorio que actualmente ocupa México. Algunos de los misioneros, por encontrarse más cercanos a los nativos, advirtieron las semejanzas que existían entre algunas de las lenguas, por ejemplo, el zapoteco y el mixteco. En el siglo XIX, las lenguas nativas fueron objeto de una clasificación semejante a la que se realizaba en Europa para las lenguas indoeuropeas. Esta tarea fue emprendida por Manuel Orozco y Berra, intelectual mexicano de la segunda mitad del siglo XIX. Algunas de sus hipótesis clasificatorias fueron retomadas por Morris Swadesh a principios del siglo XX. Las lenguas de México pertenecen a *ocho familias de lenguas* (además de algunas lenguas de filiación dudosa y otras lenguas aisladas), de las cuales las tres más importantes tanto en número de hablantes como en número de lenguas son las lenguas uto-aztecas, las lenguas mayenses y las lenguas otomangues.

Uno de los grandes problemas que presenta el establecimiento de relaciones genéticas entre las lenguas de México, es la falta de documentos escritos antiguos que permitan conocer la evolución de las familias lingüísticas. En muchos casos, la información disponible consiste en unas cuantas palabras registradas antes de la desaparición de un idioma. Tal es el caso, por ejemplo, del idioma coca, cuyos últimos vestigios lo constituyen algunas palabras de las que se sospecha pertenecen más bien a alguna variedad del náhuatl hablado en Jalisco. Swadesh calculaba que el número de idiomas hablados en el territorio mexicano llegaba a los ciento cuarenta. Actualmente sólo sobreviven sesenta y cinco. [6][7][8][9]

Las 20 lenguas indígenas más habladas en México[8]				
#	Idioma	2010	2000	Cambio %
1	náhuatl	1,544,968	1,659,029	▼ 6.8%
2	maya	786,113	892,723	▼ 11.9%
3	mixteco	471,710	510,801	▼ 7.6%
4	tzeltal	445,856	336,448	▲ 32.5%
5	zapoteco	425,123	505,992	▼ 15.9%
6	tzotzil	404,704	356,349	▲ 13.5%
7	otomí	284,992	327,319	▼ 12.9%
8	totonaco	244,033	271,847	▼ 10.2%
9	mazateco	223,073	246,198	▼ 9.3%
10	chol	212,117	189,599	▲ 11.8%
11	huasteco	161,120	173,233	▼ 6.9%
12	mazahua	135,897	151,897	▼ 10.5%
13	mixe	132,759	135,316	▼ 1.8%
14	chinanteco	131,382	152,711	▼ 13.9%
15	purépecha	124,494	136,388	▼ 8.7%
16	tlapaneco	120,072	119,497	▲ 0.4%
17	tarahumara	85,018	87,721	▼ 3.0%
18	zoque	63,022	60,093	▲ 4.8%
19	tojolabal	51,733	44,531	▲ 16.1%
20	amuzgo	49,635	48,843	▲ 1.6%

Tabla 1.1 Lenguas indígenas más habladas en México

1.3.2 Base de audios

Debido a la marginación que existe en México y al poco interés que hay hacia la preservación de las lenguas indígenas, así como a su estudio, poseemos muy poca información grabada de estas lenguas. Parte importante de nuestra investigación es tratar de identificar estas lenguas en grabaciones que se desconoce su contenido.

Poseemos 17 audios de lengua Náhuatl, 10 de Mixteco, 9 de Maya y 10 de Ñañu que solo contienen voz y son realmente útiles para nuestra investigación. También tenemos otras grabaciones pero debido a que poseen música de fondo o mucho ruido de grabación no nos son útiles en estos momentos.[8][9]

Cada audio dura 1 minuto, lo cual es bastante bueno para hacer nuestro análisis y poder obtener una base de datos para las pruebas realizadas en el presente trabajo.

1.4 Naturaleza del problema

La identificación del lenguaje hablado consiste en determinar el idioma de quien habla basándose sólo en una muestra de voz sin considerar al hablante y lo que está diciendo. De acuerdo a esto, la identificación automática del lenguaje hablado es el proceso por el cual el lenguaje (idioma/dialecto) de una muestra de señal de voz digitalizada es reconocido por una computadora.

De acuerdo a los lingüistas las diferencias entre los idiomas son múltiples y enormes. A pesar de que esas diferencias son evidentes a diferentes niveles (léxico, sintáctico, de articulación, ritmo, prosodia, etc.) la identificación del lenguaje hablado es aún un reto, debido a lo difícil que es caracterizar un idioma.[1][2]

De acuerdo al estado del arte, podemos decir, que existen dos grandes áreas para la identificación automática del lenguaje hablado, una que se basa en la representación fonética de la señal de voz, es decir en la segmentación de fonemas y sus subsecuentes procesos, y otra en donde sólo se utilizan las características acústicas de la señal de voz para la identificación de los idiomas. Este último, hasta nuestros días, no ha tenido resultados comparables a los del primer enfoque.

Los sonidos que se generan cuando hablamos pueden ser descritos en términos de un conjunto de unidades lingüísticas abstractas llamadas “fonemas”. Cada fonema corresponde a una única configuración del tracto vocal. Diferentes combinaciones de fonemas constituyen diferentes palabras. Por lo que, diferentes palabras están formadas de diferentes secuencias de fonemas que corresponden a diferentes movimientos del tracto vocal. Y más aún, diferentes combinaciones de palabras producen un gran número de oraciones que contienen toda la información que uno quiere transmitir.[2]

La fonética analiza los fonemas en términos de las características lingüísticas de esos sonidos y los relaciona con la posición y movimientos de las articulaciones. Los fonemas pueden ser clasificados por:

- *Modo de articulación*, el cual describe diferentes fonemas de acuerdo a la forma en que el tracto vocal restringe el aire que sale de los pulmones. Los

idiomas tienen diferentes categorías de fonemas: nasal, vocal, fricativa, entre otros.

- *Características de losformantes*, las consonantes pueden ser formadas dependiendo si las cuerdas vocales están o no están involucradas en su producción.
- *Lugar donde se hace la articulación,* es decir, el lugar donde se estrecha el tracto vocal durante la pronunciación.

Diferentes combinaciones de formantes, lugar y modo de articulación resultan en diferentes fonemas.

Generalmente un lenguaje no usa todas las posibles combinaciones de formantes, de articulación y de lugar de articulación. Es decir, un lenguaje usa sólo un subconjunto de todos los posibles fonemas que el ser humano puede producir. Así diferentes lenguajes tienen diferentes fonemas, por ejemplo el francés tiene 15 vocales mientras que el español sólo tiene 5, el alemán tiene vocales unidas mientras que en el inglés no están permitidas, etc. Por otro lado, la manera en que dichos fonemas se unen para formar una palabra debe respetar ciertas reglas propias de cada lenguaje, de la misma forma que cada lenguaje tiene su propia gramática.[2] [3]

Los métodos tradicionales de identificación automática del lenguaje aprovechan estas características (los fonemas de un lenguaje y sus combinaciones permitidas) para reconocer un lenguaje. Sin embargo, cuando hablamos no sólo generamos fonemas, también existen otros aportes de información dentro de la señal acústica tal como la entonación, la duración, el acento y el ritmo. Estos elementos comúnmente son agrupados por los lingüistas bajo el concepto de prosodia. Este tipo de información también es distintiva en los lenguajes humanos. Por ejemplo, el ritmo, que es la pauta de tensión formada en el mismo por la combinación de sílabas tónicas y átonas, largas y breves; es un elemento distintivo entre los idiomas, porque como no todas las lenguas hacen el mismo uso de las sílabas largas y breves, y de las tónicas y átonas, habrá distintos tipos de ritmos; lo que nos lleva a tener un elemento distintivo entre los idiomas; ya que el ritmo es uno de los prosodemas o fonemas prosódicos (o suprasegmentales) más característicos de una lengua. Los ritmos más importantes son el acentual y el silábico. Para el oído inglés el “ritmo” español resulta marcial, porque le produce el efecto subjetivo de una ametralladora, ya que da timbre pleno a todas las

vocales de las sílabas. En cambio, al oído español, el “ritmo” inglés le produce un efecto entrecortado y sujeto a tirones. El “ritmo” es probablemente el rasgo de la *base articulatoria* de una lengua cuya adquisición o dominio resulta más difícil al estudiante adulto de un idioma extranjero y, aunque la inteligibilidad depende en gran parte de su correcta emisión, a éste no se le presta la atención debida en la enseñanza de idiomas extranjeros [2].

El problema con este tipo de información (la entonación, la duración, el acento y el ritmo) es su extracción, ya que actualmente no existen métodos que extraigan la información suprasegmental del habla, simplemente se ha ligado la frecuencia fundamental F0 como un elemento de la prosodia. Así que realizar la identificación del lenguaje hablado sin utilizar la representación de los fonemas es un campo poco explorado; en el cual se desarrolla este trabajo de investigación.

1.5 Limitantes en la identificación del lenguaje hablado

La identificación del lenguaje hablado por medios automáticos es una tarea difícil que inevitablemente debe limitarse en diversos aspectos. Por ejemplo, el tipo de locutores esperados (i.e. niños, adultos, hombres, mujeres); el tipo de conversación (palabra aislada, frases claves, habla espontánea); el canal de transmisión de la señal de voz (micrófono, teléfono, entre otras); el nivel de ruido en la señal; el número de idiomas a identificar, etc. En particular, nuestro trabajo aborda la problemática de la identificación del lenguaje hablado cuándo:

(i) El canal de transmisión es el teléfono y en general audios con poca información de frecuencias altas. El canal del teléfono está limitado a anchos de banda bajos, aproximadamente de 3.2 Khz., con una frecuencia de muestreo de 8kHz, por lo que, la información en las altas frecuencias de la señal de voz se pierde, dando como resultado menos información para la discriminación;

(ii) el tipo de conversación es espontánea (introduce co-articulación y pausas);

(iii) se tiene una situación independiente del locutor (el tracto vocal en cada persona es diferente, entonces las variaciones de los hablantes en la realización de fonemas puede ser substancial).

Estas tres condiciones son los mínimos requeridos para alcanzar un sistema útil, lo cual hace el trabajo más difícil. Porque al trabajar con 8Khz de señal de voz perdemos frecuencias, las cuales podrían ser importantes en la discriminación de los lenguajes. Además, cuando el habla es espontánea existen silencios o risas que dificultan la extracción de características. [1]

Por último un sistema independiente del locutor es más difícil que uno dependiente del locutor, ya que se tendría que entrenar al sistema con un gran número de hablantes.

Los humanos no tenemos problemas en identificar un lenguaje cuando los entendemos. Similarmente, no hay duda que un sistema de identificación de lenguaje parecido al humano debería conseguir resultados impecables, si pudiera tener un gran vocabulario almacenado con el cual es preciso reconocer y adquirir el conocimiento de las reglas sintácticas y semánticas para cada lenguaje. Con las recientes técnicas y recursos computacionales, el desarrollo de un sistema de este tipo es imposible. Las razones son las siguientes:

- La ejecución de sistemas de reconocimiento del habla está aún muy lejos de los niveles de ejecución humanos. Los actuales sistemas de reconocimiento del habla trabajan mejor con grandes restricciones en el tamaño del vocabulario. Pero para los sistemas de identificación del lenguaje hablado, tales restricciones no pueden ser hechas.

- Recolectar y seleccionar el suficiente conocimiento de los múltiples lenguajes no es una tarea trivial. Para obtener una representación robusta de esta información, se requiere de una cantidad grande de datos de entrenamiento.[8][9]

Por todo esto, tomar el camino de no utilizar la representación de los fonemas es una posibilidad más viable. Aún con todos los problemas que ello implica, esta solución evita el reconocimiento de fonemas de cada lenguaje así como la creación de modelos de lenguaje respectivos. Además facilita la introducción de nuevos lenguajes en caso de ser necesario. Por otro lado, esta es la única solución posible para lenguas marginadas sin recursos lingüísticos suficientes. Tal como es el caso de muchas de las lenguas indígenas de México.[6]

1.6 Objetivos

Al no depender de la representación fonética de la señal de voz, el peso del método recae en el procesamiento acústico. Se necesita de un proceso acústico que extraiga las características más representativas para una mejor discriminación de los lenguajes. Por otro lado, tenemos que los lingüistas han estudiado el habla no solo en términos de fonemas[2], sino que además ellos definen características extralingüisticas al hablar, tales como la prosodia, el ritmo, el tono y la duración, así como las frecuencias que se usan al pronunciar las diferentes palabras.[11] Dichas características las definen como fenómenos fonético-fonológicos, los cuales no pueden segmentarse como los fonemas, porque actúan simultáneamente sobre más de un segmento, es por ello que podemos hablar de fonemas segmentales y suprasegmentales. Entonces tenemos los siguientes objetivos particulares para el procesamiento acústico de la señal de voz:

1. Extracción de características suprasegmentales del habla para la tarea de identificación del habla.

Para ello, se propone aumentar el número de coeficientes cepstrales de frecuencia Mel (MFCC) a 20 buscando obtener más detalle de las frecuencias. Comúnmente se han utilizado 12 coeficientes con muy buenos resultados para la segmentación de fonemas, principalmente en la tarea de reconocimiento del habla.

2. Extracción de características asociadas a la frecuencia.

Las frecuencias que producimos al entonar una palabra son diferentes para cada palabra y para cada idioma, estudios anteriores, muestran como cada idioma tiene frecuencias que se repiten constantemente lo cual diferencia uno de otro. Recordemos que la prosodia ha sido vinculada a la frecuencia fundamental F0 (pitch) y dicha frecuencia es la más baja, además estudios recientes muestran que el humano no procesa frecuencias individuales independientemente como lo sugiere el análisis acústico[12], a excepción del ruido blanco que es producido arficialmente; en su lugar escuchamos grupos de frecuencias y aunado a que la prosodia está en la frecuencias bajas, estamos interesados en distinguir el grupo de frecuencias bajas de la señal de voz con una muy buena resolución y en lo general distinguir la cantidad de energía utilizada en cada banda de frecuencias críticas, según el idioma analizado.[13]

3. Aplicación y evaluación de estos métodos de caracterización propuestos.

- Para poder evaluar los resultados compararemos la cantidad de esfuerzo y recursos necesarios para lograr nuestro objetivo con los esfuerzos realizados con las técnicas que utilizan fonemas.
- Para determinar el alcance de estos métodos en lenguas marginadas, realizaremos pruebas con muestras de lenguas indígenas de México, los lenguajes a utilizar son Náhuatl, Mixteco, Maya y Ñañú.[6]
- Para demostrar la eficacia de la caracterización propuesta, realizaremos experimentos con diferentes clasificadores: Algoritmo de alineamiento temporal dinámico DTW (Dynamic Time Warping) que nos permite medir la similitud entre dos secuencias que pueden variar en el tiempo o en el espacio, y por algoritmo de distancia Euclidiana para medir la diferencia entre dos vectores, siendo la distancia más corta, el vector mejor parecido.[16]

De esta manera dejaremos una base para poder seguir con esta línea de investigación en trabajos futuros, así como continuar realizando esfuerzos para poder mejorar el sistema realizado en el presente trabajo.

CAPÍTULO 2

ANTECEDENTES LINGÜÍSTICOS EN LA IDENTIFICACIÓN DEL LENGUAJE HABLADO

Existen importantes estudios desde el punto de vista lingüístico relacionados con la identificación del lenguaje hablado. Los cuales se enfocan en cómo el humano realiza la discriminación entre las lenguas. En esta sección abordaremos el tema de la discriminación de los lenguajes desde este punto de vista.

Los lingüistas, desde un punto de vista diferente al nuestro, han intentado realizar la clasificación de los lenguajes humanos basados en características prosódicas. La prosodia es un término usado típicamente para describir aspectos extralingüísticos del discurso. Ella incluye la entonación, patrones de acentuación, ritmo, melodía, etc.

Los lingüistas parten de fenómenos fonético-fonológicos, los cuales no pueden segmentarse como los fonemas, porque actúan simultáneamente sobre más de un segmento (al menos sobre la sílaba). Estos fenómenos reciben el nombre de suprasegmentales y son tres: el acento, el tono (o la sucesión de ellos, es decir, la entonación) y la duración (o cantidad). El conjunto de estos tres elementos suprasegmentales se denomina prosodia.[17]

La fonología realiza una división entre los fonemas (o fonemas segmentales) y los prosodemas (o suprasegmentos), como el acento, la duración y la entonación. Entre segmentos y suprasegmentos hay una diferencia de clase que resulta evidente: los

fonemas son segmentales (o segmentables), uno a uno, mientras que los prosodemas afectan o pueden afectar conjuntamente a varios. Sin embargo, en la realización de los suprasegmentos intervienen índices acústicos y articulatorios que también están presenten en la realización de los segmentos, como:

1. La vibración de las cuerdas vocales, que es la fuente de sonoridad de los segmentos sonoros, y también del movimiento del tono fundamental que puede utilizarse en la distinción de las palabras (tono) o de oraciones (entonación).
2. Todo segmento tiene una dimensión temporal, es decir, una duración. Ésta, además, puede desempeñar, en determinadas lenguas, una función distintiva.
3. Todo segmento, al realizarse, ha de tener alguna intensidad. Esta, además, puede desempeñar en algunas lenguas una función distintiva (acento).

Lo anterior muestra las semejanzas entre segmentos y suprasegmentos. Pero entre esos dos elementos hay también una diferencia de grado, que hace que haya que considerarlas como unidades distintas. La diferencia entre dos fonemas no es gradual. Por ejemplo, /p/ se diferencia de /t/ en que una es labial y otra dental. De igual manera, /p/ se diferencia de /b/ por el rasgo de sonoridad. Y un sonido es sonoro o no lo es. Por su parte, el acento, por ejemplo, es gradual: una vocal tona tiene más "fuerza" que una átona, pero no posee ninguna cualidad distinta [17].

Por último, existe una tercera razón para distinguir los segmentos y los suprasegmentos como pertenecientes a dos clases separadas: la función lingüística.

1. La función de los fonemas es distintiva: son unidades que en un contexto dado se excluyen mutuamente (/'pipa/ - /'pepa/ - /'papa/ - /'popa/ - /'pupa/).
2. La función de los suprasegmentos es contrastiva, ya que no pueden alternar en el mismo contexto. En la oposición "amo-amó" lo distintivo es el esquema acentual /'_ /, frente a /_ '/, pero no el acento en sí. El suprasegmento necesita la presencia contrastante de su opuesto en la misma secuencia.

2.1 Definiciones generales

Fonotáctica, trata de la normas y reglas que regulan la combinación de los fonemas de una lengua, por ejemplo, las reglas fonotácticas del español impiden plurales como “clubs” o “films”, siendo las formas correctas, de acuerdo con la fonotáctica de esta lengua “clubes” y “filmes”.

Átono, en fonética articulatoria, el adjetivo átono se aplica a las sílabas y vocales que carecen de acento, en las lenguas como el español o el francés las vocales de las sílabas átonas conservan prácticamente el mismo timbre que el de las acentuadas. Pero en otras, como el ruso o el inglés, las átonas tienden a una centralización; en esta última lengua la centralización se materializa en /e/, /i/ o /u/, como se puede comprobar en las sílabas átonas como *language, necessary o plentiful.* Esta característica hace que en la conversación normal se pueda confundir la pronunciación de palabras como *vacation* o *vocation*, aunque en una emisión oral cuidada se puedan diferenciar sin mayor problema.

Tono, existe varias definiciones, una de ellas define al “tono” como la *altura musical* de cada sílaba. Tradicionalmente al tono se le ha llamado acento melódico. Vistos desde la fonética articulatoria, los “tonos” constitutivos de la entonación, se forman por la vibración de las cuerdas vocales y se mueven en la escala de mayor a menor vibración (agudo- grave), mientras que los acentos, componentes de las pautas rítmicas, se mueven en la de mayor a menor intensidad (tónico-átono); cuando mayor sea la vibración, tanto más agudo será el tono. Acústicamente los “tonos” están relacionados con la frecuencia. Cada persona tiene su *tono normal* de voz, es decir, la nota que dentro de su registro individual se produce con más naturalidad y menor fatiga. En torno a ella se suceden los movimientos ascendentes y descendentes. Se comprueba que, descartando las diferencias individuales, las gentes de determinadas regiones o países suelen expresarse en un tono normal medio más agudo o más grave.

También se llama “tono” a la función distintiva que cumple la *frecuencia fundamental* en el nivel de la palabra [17]. De la misma manera que en el español el acento tiene una función distintiva, como recurso de diferenciación léxica, por ejemplo, “pérdida” y “perdida”, “ingles” e “inglés”, entre otras. Existen lenguas

como el Chino o el Vietnamita, que se sirven del tono para estos fines. Por ejemplo, /ma/ puede significar varias cosas distintas, desde *madre* hasta *caballo*. Con un “tono” estático alto significa *madre*, con un “tono” dinámico ascendente significa *cáñamo*, con un “tono” dinámico descendente-ascendente significa *caballo*, y con un tono descendente significa *riña*. A las lenguas que usan el tono como recurso en la formación de las palabras se las llama lenguas tonales, por ejemplo: las lenguas de la familia congo-nigeriana, sino-tibetanas y algunas de las lenguas indígenas de México (otimí, mazahua, pame y chichimeca entre otras).

El tono fundamental depende, básicamente, de las vibraciones de las cuerdas vocales; pero, además, hay una serie de factores fonéticos que la condicionan:

1. Existe una relación entre la cualidad o el timbre de la vocal y la altura relativa de su frecuencia fundamental, de modo que las vocales más altas / [i], [e]/ tienen un tono fundamental más elevado.
2. Las frecuencias fundamentales más altas aparecen después de las consonantes sordas, y las más bajas, tras las consonantes sonoras.
3. Además del tono fundamental, la duración y la intensidad también intervienen en la producción y la percepción de la entonación.[18]

2.2 El acento

El acento es un rasgo suprasegmental que recae sobre una sílaba de la cadena hablada y la destaca o realza frente a otras no acentuadas (o átonas) [17].

Esta prominencia silábica suele interpretarse tradicionalmente como reflejo de intensidad; por eso, se le ha llamado "acento de intensidad". La realidad, sin embargo, es más compleja: la prominencia resulta de la conjunción de varios factores articulatorios:

1. Una mayor fuerza respiratoria, que genera una mayor intensidad.
2. Una mayor tensión de las cuerdas vocales, que genera una elevación del tono fundamental.
3. Una mayor prolongación en la articulación de los sonidos, que supone un aumento de la duración silábica.

Así pues, la sílaba tónica, habitualmente, es más intensa, más alta y más larga que las sílabas átonas adyacentes. En español, el índice acústico primario del acento es el tono, aunque los otros dos índices (intensidad y duración) también colaboran en la acentuación, en proporciones variables.

La mayoría de las palabras poseen una sílaba tónica y otras átonas. Sólo algunos monosílabos pueden considerarse palabras átonas. Cuando las palabras son más largas, una sílaba posee el acento principal y otra el acento secundario. Dentro de una frase, el último acento principal se denomina acento de frase.

En las distintas lenguas del mundo, el acento puede tener las siguientes funciones lingüísticas:

1. Contrastiva: distingue sílaba tónicas/átonas en el eje sintagmático. Por ejemplo: "El libro es de él".
2. Distintiva: distingue unidades en el eje paradigmático (en lenguas con acento libre). Por ejemplo: "amo"/"amó".
3. Demarcativo: en lenguas de acento fijo, señala los límites de las unidades en la secuencia. Por ejemplo: el final de una palabra en turco.
4. Culminativa: en las lenguas de acento libre, señala la presencia de una unidad acentual, sin indicar sus límites.

2.3 La entonación.

La entonación es uno de los componentes más complejos de una lengua. Se ha definido de muchas maneras, dependiendo básicamente del interés de cada autor: por el tono fundamental, por una conjunción de parámetros acústicos (tono, acento y duración, primordialmente), por su función lingüística, etc.

Quilis [17] define la entonación como "la función lingüísticamente significativa, socialmente representativa e individualmente expresiva de la frecuencia fundamental en el nivel de la oración".

Desde el punto de vista articulatorio, el tono depende básicamente de las cuerdas vocales: de su longitud, su grosor su tensión.

Según la utilización lingüística del tono, las lenguas se dividen en tonales y entonadas:

1. Las lenguas tonales utilizan los tonos para distinguir significados. Cumple, entonces una función distintiva en el léxico. Por ejemplo, el chino, el tailandés.

2. Las lenguas entonadas utilizan la sucesión de tonos, es decir, la curva melódica de la entonación, no ya para distinguir significados léxicos, sino para modificar significaciones secundarias (expresividad, intencionalidad, etc.). Cumple, entonces una función expresiva en la frase. A este tipo de lenguas pertenecen todas las románicas.

2.4 La duración.

La duración es también un fenómeno segmental, puesto que cada sonido posee una duración propia. Así por ejemplo, es sabido que las consonantes fricativas son más largas que las oclusivas, que las sordas son las más largas que las sonoras, etc.

Algunas lenguas poseen pares de fonemas en función de la duración. Por ejemplo, el italiano distingue entre ciertas consonantes breves y largas o "dobles". El latín clásico distinguía entre vocales breves y largas.

De acuerdo a la articulación, la duración se basa en el mantenimiento por más o menos tiempo de una determinada configuración articulatoria. Por el fenómeno de la coarticulación (la cual describimos con un ejemplo dado anteriormente, es diferente la pronunciación de “to” en las palabras: *to*do y en es*to*rnudo), dicha configuración (y, consiguientemente, la duración) se ve alterada en función del contexto.

Dentro de la cadena hablada, los segmentos se agrupan en unidades cada vez mayores: sílabas, palabras y enunciados. La fonosintaxis es el estudio de las modificaciones que sufren los fonemas al agruparse en la cadena hablada. El concepto básico aquí es el de coarticulación: los sonidos no se pronuncian aislados, y la proximidad articulatoria de unos con otros hace que se influyan mutuamente.

Los sonidos se agrupan, como hemos visto, en unidades cada vez mayores: la sílaba -que no suele considerarse objeto específico de la fonosintaxis-, la palabra y la oración. Sin embargo, la fonosintaxis distingue otra unidad, intermedia entre las dos últimas: el sirrema.

2.5 El sirrema

El sirrema es "la agrupación de dos o más palabras que constituyen una unidad gramatical, unidad tonal, unidad de sentido y que, además, forman la unidad sintáctica intermedia entre la palabra y la frase" [17].

Las palabras que constituyen el sirrema permanecen siempre unidas: entre ellas no puede haber pausa. La razón de ser de dicha unidad es acentual: el sirrema aglutina a una serie de elementos silábicos átonos que no pueden producirse aislados, sino en torno a alguna otra sílaba acentuada, para formar con ella una unidad indisoluble.

En general, cada lengua tiene su propio inventario de las partes de la oración que forman sirrema. Fuera de esas combinaciones, las demás agrupaciones están sujetas a una gran variabilidad en lo referente a pausas y entonación. En español, forman sirrema las siguientes partes de la oración:

1. El pronombre átono y el elemento gramatical que le antecede. Por ejemplo: dile que venga (/dile ke 'beNga/).
2. El artículo y el sustantivo. Por ejemplo: el carro (/el'kaʁo/).
3. El adjetivo y el sustantivo, o viceversa. Por ejemplo: perro blanco (/'peʁo'blaNko/).
4. El sustantivo y el complemento determinativo. Por ejemplo: el perro de Javier (/el 'peʁode'javier/).
5. Los tiempos compuestos de los verbos. Por ejemplo: he comido (/'eko'mido/).
6. Los elementos de una perífrasis o una frase verbal. Por ejemplo: hemos dejado de ser (/'emosde'xadode'ser/).
7. El adverbio y verbo, adjetivo o adverbio. Por ejemplo: los más destacados alumnos (/los'masdesta'kadosaluNnos/).
8. La conjunción y la parte del discurso que la introduce. Por ejemplo: Juan y Pedro (/'xuan i'pedRo/).
9. La preposición con su término. Por ejemplo: voy con Juan (/'boi koN'xuan/).

2.6 El ritmo.

El término ritmo puede tener en lingüística, al menos, dos acepciones:

1. En un sentido amplio se llama ritmo a las sensaciones auditivas que se perciben a los intervalos regulares de tiempo, producidas por repeticiones isofónicas de cualquier recurso prosódico del lenguaje, como puede ser la rima, la censura, etc.

2. En un sentido estricto, el ritmo es un prosodema básico de la cadena hablada, junto con la entonación y el acento. Aún siendo conscientes, por una parte, de que lo que realmente se percibe auditivamente es una prominencia, conviene separar, en lo posible, los rasgos de tensión y los de melodía que se manifiestan en la cadena hablada; los rasgos de melodía corresponden a la entonación y los rasgos de tensión corresponden al ritmo (también llamado ritmo verbal para diferenciarlo del ritmo musical).

El ritmo de un grupo fónico es la pauta de tensión formada en el mismo por la combinación de sílabas tónicas y átonas, y largas y breves. El ritmo es uno de los prosodemas o fonemas prosódicos (o suprasegmentales) más característicos de una lengua. Como no todas las lenguas hacen el mismo uso de las sílabas largas y breves, y de las tónicas y átonas, habrá distintos tipos de ritmos; los más importantes son el acentual y el silábico.

Ritmo acentual (o *stress-timed*) quiere decir que las pautas que se forman en el grupo fónico tienen un *tempo* marcado por el acento, o sea, están acompasadas por el acento, mientras que en el **ritmo silábico** (o *syllable timed*) es la sílaba la que sella el *tempo*, es decir, el ritmo está acompasado por la sílaba; el del inglés es de tipo acentual, mientras que el del español es silábico.

Las vocales del inglés poseen dos rasgos peculiares que no existen en el español, y que contribuyen a que las diferentes pautas rítmicas sean distintas a las del español:

a) pueden ser largas y breves, con lo que unas sílabas tendrán mayor duración que otras; y
b) en las sílabas átonas pierden su timbre pleno.

En cambio, en español los fonemas vocálicos poseen la misma duración aproximadamente, y la diferencia entre los timbres de las vocales tónicas y átonas apenas es perceptible. Como ya mencionamos anteriormente, para el oído inglés el "ritmo" español resulta marcial, ya que da timbre pleno a todas las vocales de las sílabas. En cambio, al español, el "ritmo" inglés le produce un efecto "entrecortado" y sujeto a "tirones", el de esta lengua se caracteriza, además, por la *isocronía*, es decir, por la tendencia a dejar el mismo tiempo entre dos sílabas tónicas, con independencia del número de sílabas átonas que hay entre las dos tónicas.

2.7 Clasificando los lenguajes humanos a través de su ritmo.

Los lingüistas, desde un punto de vista diferente al nuestro, han intentado realizar la clasificación de los lenguajes humanos basados en las características suprasegmentales, específicamente el ritmo. El ritmo lo definimos como la organización sistemática de unidades prominentes y no prominentes del habla en unidades de tiempo [19]. Una unidad, dependiendo del lenguaje, puede ser una sílaba o un intervalo vocálico; y la prominencia está dada por su duración, su intensidad o una frecuencia fundamental alta. Es posible hablar de patrones rítmicos en el habla específicos o no a un lenguaje. Nuestro interés recae en los patrones rítmicos distintivos de un lenguaje. Los primeros intentos por clasificar a los idiomas con base en el ritmo recaen en la habilidad documentada en infantes para distinguir lenguajes [20].

Estos primeros intentos distinguían dos clases :

- Los lenguajes mostrando patrones de igual duración entre sílabas prominentes (sílabas acentuadas) a esta clase se le conoce como "stress timed", (por ejemplo, el inglés o el alemán) y

- Los lenguajes con sílabas de igual duración o "syllable timed"(por ejemplo, el francés o el español)

Sin embargo, investigaciones posteriores encontraron que esta clasificación era demasiado restrictiva. Lo que es más se llegó afirmar que el ritmo era sólo un fenómeno perceptual imposible de extraer de la señal acústica. Trabajos recientes han buscado redefinir el concepto de ritmo para clasificar los lenguajes. Ramus et al [21] introdujo una nueva forma de medición del ritmo. Esta se basa en la duración de los intervalos vocálicos y consonánticos. Un intervalo vocálico es un fragmento de la señal de habla

asociado a una o varias vocales (diptongos) del idioma en cuestión. De manera similar, un intervalo consonántico es un fragmento donde se realizan únicamente consonantes. En particular los trabajos basados en la proporción relativa de las vocales han obtenido resultados interesantes. En estos trabajos las medidas para describir el ritmo se basan en tres conceptos principales:

- La proporción de los intervalos vocálicos en una oración, que es la suma de los intervalos vocálicos entre la duración total de la oración (%V).

- La desviación estándar de la duración de los intervalos vocálicos en cada oración (ΔV).oración (ΔC).

A partir de ellas se pudieron determinar el ritmo de algunos lenguajes. La gráfica 2.1 muestra la distinción de los lenguajes en función de la proporción relativa de los intervalos vocálicos (%V) y su desviación estándar (ΔV). A partir de las ideas de este trabajo nuevos esquemas de clasificación han sido planteados.

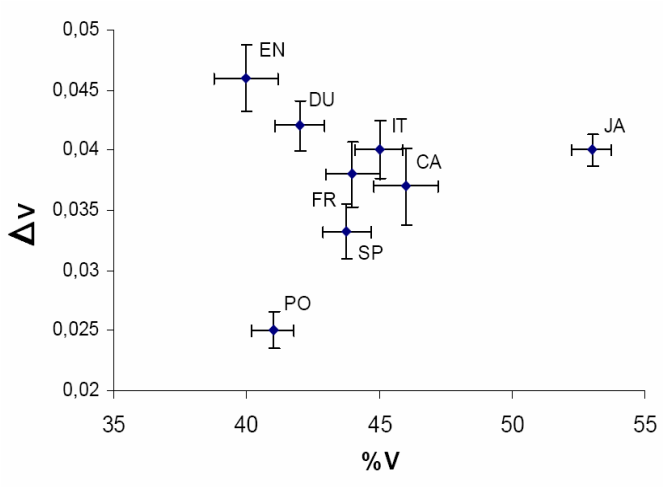


Figura 2.1 Distinción rítmica de los lenguajes en función de sus intervalos vocálicos

Es a partir de estos estudios que podemos orientar nuestra investigación para la integración del ritmo en la caracterización de la señal acústica. Por supuesto, aun falta trabajo por realizar. En todos estos trabajos el interés no es crear un sistema automático, de ahí que la segmentación e identificación de fonemas (más aun la definición de un fonema) se realizó de manera manual sobre un conjunto reducido de

grabaciones. Sin embargo, para nuestro trabajo se pretende realizar una extracción automática de estas características y por lo cual se exploran métodos para lograr hacerlo y poder tener un sistema autónomo.

2.8 La identificación de idiomas por los seres humanos.

Otro estudio relacionado con la discriminación de idiomas es el llevado por Muthusamy. En él se comprueba la capacidad de los seres humanos para la identificación del lenguaje. Con escuchar la voz unos segundos, la gente es capaz de determinar de qué lenguaje se trata, siempre y cuando conozcan el lenguaje en particular; y en el caso de que sea un lenguaje que ellos no están familiarizados, pueden realizar un juicio subjetivo de acuerdo a los lenguajes similares que ellos conocen, por ejemplo, suelen decir: "suena parecido al alemán". De acuerdo a esto, Muthusamy en 1994 [22] realizó un estudio para obtener las mejores marcas que tienen los humanos en la identificación del lenguaje hablado. Sus pruebas consistieron en dos casos: la identificación del lenguaje hablado por personas monolingües, es decir, que sólo conocen su lengua materna; y el segundo caso por personas que conocen varios lenguajes. Para el experimento se analizaron 28 personas (14 mujeres y 14 hombres); de los cuales fueron 10 hablantes nativos del lenguaje inglés y 2 personas para cada uno de los 9 lenguajes restantes. Todas las personas conocían el lenguaje inglés. Las personas podían escuchar 10 segundos de señal de voz espontánea tantas veces como ellos desearan.

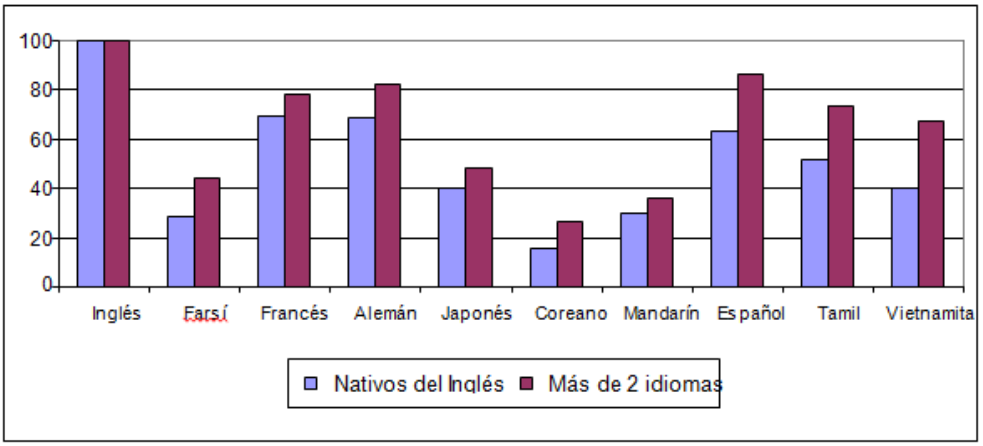


Figura 2.2 Porcentaje de identificación del lenguaje hablado por hablantes nativos del inglés y los que hablan más de dos idiomas en 6 segundos de señal de voz.

El porcentaje de reconocimiento para personas que conocían 4 idiomas fue de 66.7%, para personas que conocían 3 idiomas 57.9%, para las personas que conocían 2 fue de 51.1% y las que sólo conocían uno fue de 44.1%. Ver figura 2.3. Con esto Muthusamy [22] concluyó que el porcentaje de discriminación de los lenguajes aumenta cuando se tiene conocimiento de más lenguas. *Lo importante de esto es notar que incluso la identificación del lenguaje hablado hecho por los humanos no tiene grandes porcentajes de discriminación, aún en el caso de personas que dominan cuatro idiomas.*

CAPÍTULO 3

FUNDAMENTOS PARA EL ANAÑISIS DE LA SEÑAL DE AUDIO

3.1 Panorama general

El habla es una señal continua, la cual varía en el tiempo. Esta señal es el producto de las variaciones en la presión del aire cuando hablamos. Un micrófono convierte esas variaciones de presión del aire a variaciones en voltaje, lo que comúnmente llamamos señal analógica. La señal analógica, se puede transmitir a través de un circuito telefónico o puede ser almacenado en una cinta magnética. Sin embargo, para la computadora es necesario digitalizar la señal, convirtiendo la señal analógica a una serie de valores numéricos con una frecuencia regular (frecuencia de muestreo). El número de valores está limitado por el número de bits seleccionados para representar a cada muestra.

La resolución es la cantidad de bits utilizados para almacenar la voz; es decir, cada muestra se representa con un valor digital, limitando el rango de valores discretos correspondiente al original. Por ejemplo: utilizando 4 bits se pueden representar 16 valores diferentes y con 8 bits se pueden representar 256 valores.

La resolución del teléfono es de 8bits/muestra, es decir, si muestreamos a 8kHz, tenemos 8000 muestras por segundo y así 8000 * 8 = 64000 bits por segundo. En cambio en un CD la resolución es de 16 bits/muestra, por lo tanto, 44100 muestras por segundo *16 bits = 705600 bits por segundo en modo mono aural, si es estéreo, se duplica.[5]

3.2 Percepción auditiva

Los fonemas aparentemente tienen parámetros acústicos claramente definidos, pero más bien, los fonemas tienden a ser abstracciones implícitamente definidas por la pronunciación de las palabras en un lenguaje.

La forma acústica de un fonema depende fuertemente del contexto acústico en el que sucede. A este efecto se le llama coarticulación, por ejemplo, es diferente la pronunciación de “to” en las palabras: *to*do y en es*to*rnudo. Este concepto se utiliza para distinguir entre la característica conceptual de un sonido del habla (fonema) y una instancia o pronunciación específica de ese fonema (fono).

La variabilidad del habla producida por la coarticulación y otros factores, tales como el tipo de locutores, el tipo de conversación, si es espontánea, el canal de transmisión de la señal de voz y el nivel de ruido en la señal; hacen el análisis de la voz extremadamente difícil. La facilidad del humano en superar estas dificultades sugiere que un sistema basado en la percepción auditiva podría ser un buen enfoque. Desafortunadamente nuestro conocimiento de la percepción humana es incompleto. Lo que sabemos es que el sistema auditivo está adaptado a la percepción de la voz.

El oído humano detecta frecuencias de 20Hz a 20,000 Hz, pero es más sensible entre 100 y 6000 Hz. También es más sensible a cambios pequeños en la frecuencia en el ancho de banda crítico para el habla. Además el patrón de sensibilidad a cambios en el tono (pitch) no corresponde a la escala lineal de frecuencia de ciclos por segundo de la acústica.

Para representar mejor el patrón de percepción del oído humano se desarrolló una escala llamada Mel, la cual es una escala logarítmica. Y a partir de dicha escala se crearon los coeficientes cepstrales de Frecuencia Mel. Lo cual se detalla en el siguiente apartado.

3.3 Frecuencia y amplitud

El sonido puede definirse como la decodificación que efectúa nuestro cerebro de las vibraciones percibidas a través de los órganos de la audición. Estas vibraciones se transmiten en forma de ondas sonoras. Todos los sonidos causan movimientos entre las moléculas del aire. Algunos sonidos, tales como los que produce una cuerda de guitarra, producen patrones regulares y prolongados de movimiento del aire. Los patrones de sonidos más simples son los sonidos puros, y se pueden representar gráficamente por una onda senoidal.

La amplitud de una onda sonora corresponde fisiológicamente al movimiento del tímpano de oído. La distancia desde la posición de reposo hasta la de máxima presión alcanzada por una partícula de aire se llama amplitud; la cual es una medida de fuerza de la onda; entonces el volumen de un sonido refleja la cantidad de aire que es forzado a moverse y su unidad es el decibel (dB).

La frecuencia es el número de vibraciones (ciclos) del tono por segundo, por ejemplo, si tenemos 100 ciclos por segundo esto equivale a 100Hz. Los tonos altos se representan por frecuencias altas y los tonos bajos con frecuencias bajas [5].

El humano produce señales de voz desde los 100 Hz. (hombre) ó 200 Hz. (mujer) hasta los 15000 Hz. El teléfono muestrea a 8000 Hz, perceptible pero con baja calidad, comparado con un CD, que muestrea a 44100 Hz. En el caso de los instrumentos musicales el ancho de banda es mayor que para la voz, por lo que la diferencia es audible; esto requiere un mayor espacio para almacenar y transmitir.

El habla no es un tono puro, es una serie de múltiples frecuencias y se representa como una señal compuesta, es decir, el resultado de la adición de un número determinado de ondas sinusoidales simples. El teorema de Fourier (1822) demostró que toda señal compuesta periódica (es decir, que repite periódicamente su perfil) puede descomponerse en un número limitado de ondas sinusoidales simples. En la figura 3.1 tenemos tres ondas periódicas simples de 100,200 y 300 Hz, y en la parte de abajo podemos ver la onda periódica compuesta (la línea de trazo grueso), que es el resultado de la suma algebraica de las ondas simples (las líneas discontinuas). La frecuencia de cada una de esas ondas es múltiplo de la frecuencia fundamental F0, que es la más

baja. Las vocales se componen de dos o más ondas simples, son ricas en frecuencias secundarias y tienen estructuras internas que incluyen ondas cíclicas y no cíclicas.

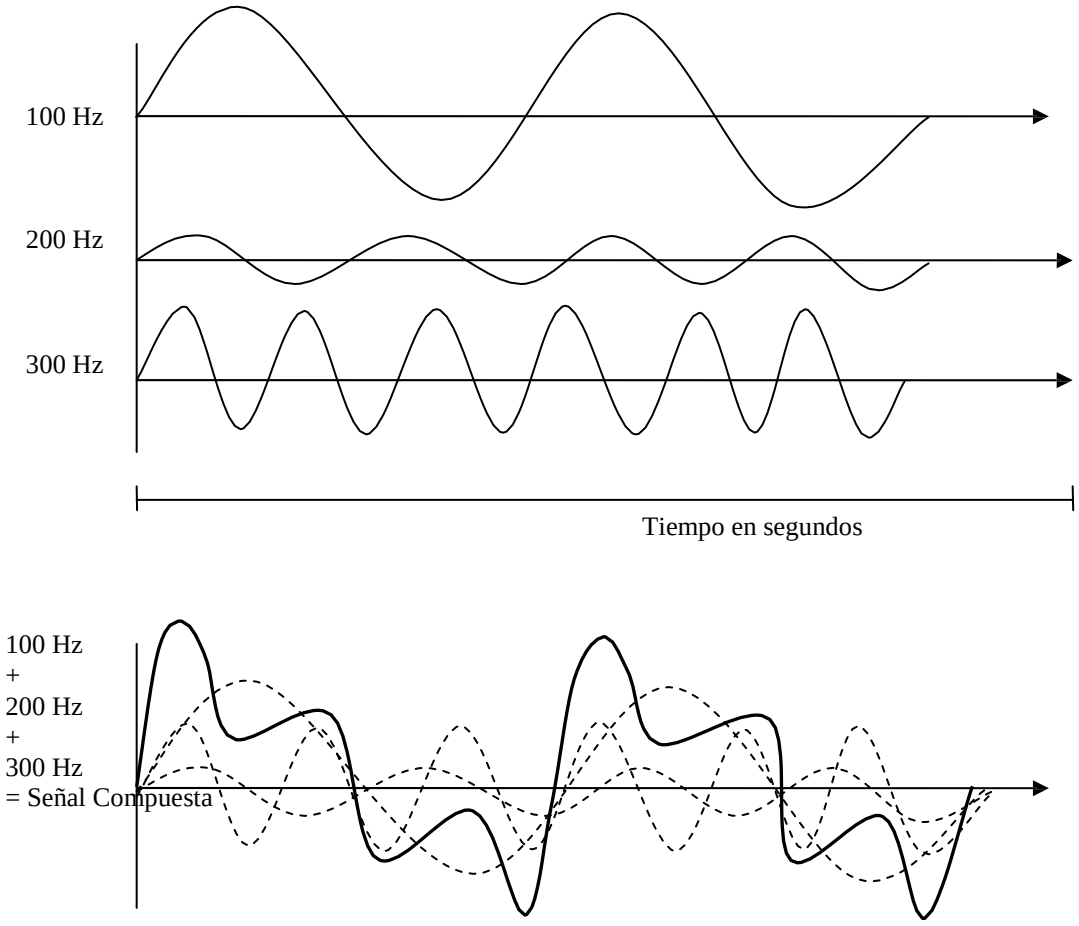


Figura 3.1 Ejemplo de señal compuesta, resultado de la suma algebraica de ondas simples.

Hay básicamente dos tipos de ondas compuestas: las cíclicas (periódicas) y las no cíclicas (aperiódicas). Las ondas cíclicas repiten periódicamente su perfil, debido a cambios regulares en la presión del aire, sus componentes son múltiplos de la frecuencia fundamental, además generan un espectro en línea. Las ondas no cíclicas no repiten periódicamente su perfil, porque hay cambios irregulares en la presión del aire, tienen componentes de todas las frecuencias, y generan un espectro continuo.

3.4 Descripción de contenidos de audios.

Las descripciones del contenido de audio pueden incluir:

- La información que describe los procesos de creación y producción del contenido (director, autor, título, etc.)
- La información relacionada con el uso del contenido (indicadores de derechos de autor, historial de uso, el horario de transmisión).
- Información sobre las características de almacenamiento de los contenidos (formato de almacenamiento, codificación).
- La información estructural de los componentes temporales de los contenidos.
- Información sobre las características de bajo nivel en el contenido (de distribución de energía espectral, timbres de sonido, descripción de la melodía, etc.
- Información conceptual sobre la realidad captada por el contenido (los objetos y acontecimientos, las interacciones entre objetos).
- Información sobre cómo navegar por el contenido de una manera eficiente.
- Información sobre colecciones de objetos.
- Información acerca de la interacción del usuario con el contenido (las preferencias del usuario, historial de uso).

La figura 3.2 ilustra un posible escenario de aplicación. Las características de audio se extraen en línea o fuera de línea, de forma manual o automática, y se almacenan en un formato reconocible por el sistema (por ejemplo MPEG-7) junto a las descripciones extraídas en una base de datos. Estas descripciones pueden ser los descriptores de bajo nivel de audio, los descriptores de alto nivel, o el texto hablado.[1]

Considere un usuario, o un agente, que sólo le interese para escuchar contenidos de audio específico, como noticias. Un filtro específico procesará las descripciones MPEG-7 de los diversos canales de audio y sólo proporcionan al usuario el contenido que

coincida con su preferencia. Tenga en cuenta que el procesamiento se realiza en los descriptores ya extraídos, y no en el contenido de audio en sí. La solicitud se somete a un motor de búsqueda, que a su vez consulta las descripciones MPEG-7 almacenadas en la base de datos. La eficiencia y la precisión de filtrado, la navegación y la consulta, dependerá en gran medida de la riqueza de las descripciones. En el escenario de la aplicación anterior, es de gran ayuda si los descriptores MPEG-7 contienen información sobre la categoría de los archivos de audio (es decir, si los archivos de difusión son noticias, música, etc.) Y, como en nuestro caso, a menudo es posible clasificar los archivos de audio basados en los descriptores de bajo nivel almacenados en la base de datos.

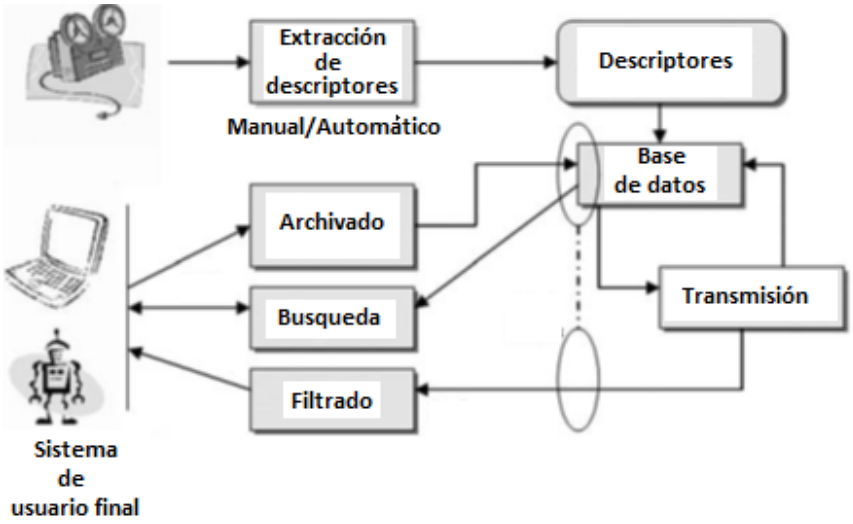


Figura 3.2 MPEG-7 escenario de aplicación.

3.4.1 Descriptores básicos

La figura 3.3 representa ejemplificaciones de los descriptores de MPEG-7 de audio básicos para fines ilustrativos, los cuales son el descriptor de forma de onda de audio y el descriptor de potencia de audio. Estas son descripciones en el dominio del tiempo. La variación temporal de los valores de los descriptores nos proporciona mucha información sobre las características de la señal de audio original [1].

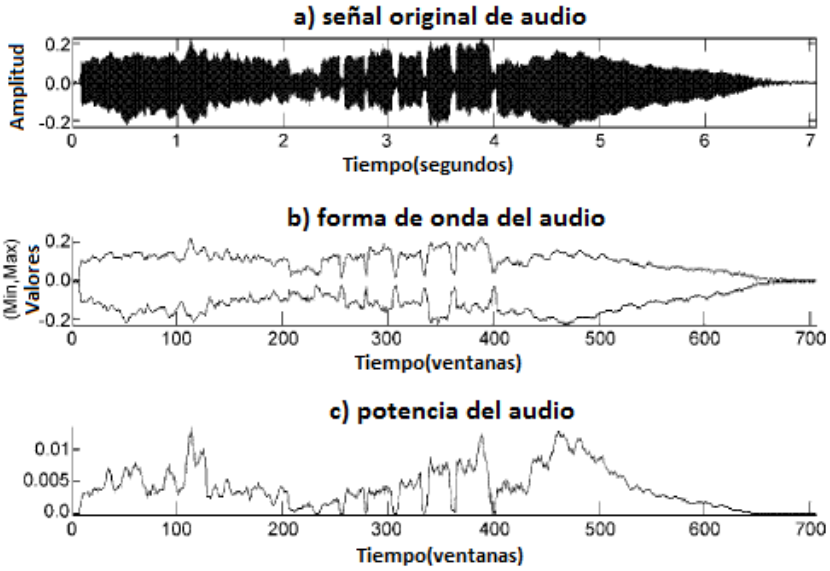


Figure 3.3 MPEG-7 Descriptores básico extraídos de una señal de audio.

- Descriptores básicos espectrales

  Los cuatro descriptores básicos de audio espectrales se derivan de un único análisis tiempo-frecuencia de una señal de audio. Ellos describen el espectro de audio en función de su dotación, centro de gravedad, el alcance y la plenitud.

- Descriptores de parámetros de la señal

  Los dos descriptores de parámetros de la señal sólo se aplican a las señales periódicas o cuasi-periódicas. Ellos describen la frecuencia fundamental de una señal de audio así como la armonicidad de una señal.

- Descriptores temporales de timbre

  Descriptores temporales de timbre pueden ser utilizados para describir las características temporales de los segmentos de sonidos. Son especialmente útiles para la descripción del timbre musical (calidad de sonido característica independiente de tono y volumen).

3.4.2 Descriptores de bajo nivel

Los descriptores de bajo nivel (LLDS) se componen de una colección de características de baja complejidad de audio que pueden ser utilizadas para caracterizar cualquier tipo de sonido.
Los LLDS ofrecen flexibilidad a la norma, lo que permite crear nuevas aplicaciones además de que se pueden diseñar basándose en herramientas de alto nivel MPEG-7[1].
La capa de base comprende una serie de 18 LLDS genéricas que consisten en una parte normativa (la sintaxis y la semántica del descriptor) y una opcional, no normativa. Los LLDS temporales y espectrales se pueden clasificar en las siguientes grupos:

- Descriptores básicos: forma de onda de audio (AWF), potencia de la señal de audio (AP).
- Descriptores espectrales básicos: envolvente espectral de audio (ASE), los centroides espectrales de audio (ASC), propagación de espectro de audio (ASS), amplitud de espectro de audio (ASF).
- Parámetros de la señal básicos: armonicidad de audio (AH), la frecuencia fundamental de audio (AFF).
- Descriptores temporales tímbricos: Tiempo de duración (LAT) y centroides temporales (TC).
- Descriptores tímbricos espectrales: centroide espectral armónico (HSC), desviación armónica espectral (HSD), propagación del espectro armónico (HSS), variación espectral armónico (HSV) y el centroide espectral (SC).
- Representaciones espectrales básicas: Fundamentos espectrales de audio (ASB) y la proyección espectral de audio (ASP).

3.5 Bandas críticas

El ancho de banda crítico es un concepto desarrollado por Fletcher, que puede interpretarse como una medida de la selectividad frecuencial del oído.

El ancho de banda crítico permite explicar por qué, dado un tono de una cierta frecuencia, una banda de ruido estrecha centrada en dicha frecuencia produce la misma cantidad de enmascaramiento sobre el tono que una banda ancha de ruido, aun cuando el

nivel de densidad espectral de ambos ruidos sea igual y, por ende, la energía del ruido de banda estrecha sea menor [23].

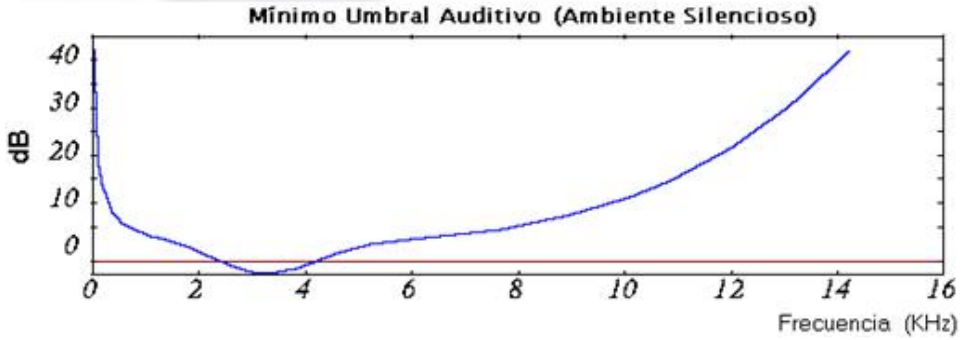


Figura.3.4 mínimo umbral auditivo.

Bajo la suposición de que un tono "sumergido" en una banda de ruido es apenas audible cuando la intensidad del tono es igual a la intensidad total del ruido enmascarante, Fletcher determinó que, cuando el ancho de la banda de ruido cae por debajo de cierto valor crítico, la densidad espectral del ruido debe ser inversamente proporcional al ancho de dicha banda para que el tono permanezca enmascarado; cuando el ancho de la banda de ruido supera dicho valor crítico, la densidad espectral del ruido enmascarante debe permanecer constante para que el tono sea apenas audible.[12]

En otras palabras, si el ancho de la banda de ruido varía, para enmascarar al tono es necesario que la energía del ruido contenida en un intervalo de frecuencias alrededor del tono sea constante. (Figura 3.5)

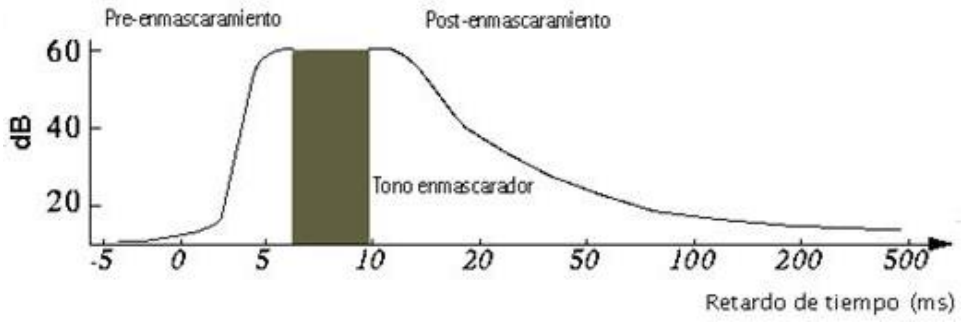


Figura.3.5 Enmascaramiento de un tono.

La energía efectiva de la señal enmascarante es aquella confinada en tal intervalo, mientras que el resto no contribuye al enmascaramiento del tono [12]. El ancho de este intervalo crítico ha sido denominado ancho de banda crítico.

A pesar de los errores implícitos en la definición de Fletcher, el concepto de un ancho crítico sigue siendo válido, puesto que numerosos experimentos psicoacústicos indican que las respuestas de los sujetos ante distintos fenómenos perceptuales cambian abruptamente cuando los estímulos sobrepasan un cierto ancho de banda [13].
Así pues, se define una banda crítica (BC) como un intervalo de frecuencia que representa la máxima resolución frecuencial del sistema auditivo en diversos experimentos psicoacústicos. Adicionalmente, puede decirse que una banda crítica constituye el intervalo de frecuencia en el cual el oído interno efectúa una integración espacial (es decir, espectral) de la intensidad de la señal sonora: la banda crítica es el intervalo en el cual se "suma" la energía de las distintas componentes espectrales de la señal.

número	frecuencia central (Hz)	banda crítica (Hz)	frecuencia de corte inferior (Hz)	frecuencia de corte superior (Hz)
1	50	-	-	100
2	150	100	100	200
3	250	100	200	300
4	350	100	300	400
5	450	110	400	510
6	570	120	510	630
7	700	140	630	770
8	840	150	770	920
9	1000	160	920	1080
10	1170	190	1080	1270
11	1370	210	1270	1480
12	1600	240	1480	1720

número	frecuencia central (Hz)	banda crítica (Hz)	frecuencia de corte inferior (Hz)	frecuencia de corte superior (Hz)
13	1850	280	1720	2000
14	2150	320	2000	2320
15	2500	380	2320	2700
16	2900	450	2700	3150
17	3400	550	3150	3700
18	4000	700	3700	4400
19	4800	900	4400	5300
20	5800	1100	5300	6400
21	7000	1300	6400	7700
22	8500	1800	7700	9500
23	10500	2500	9500	12000
24	13500	3500	12000	15500

Tabla 2.1. Bandas Críticas

En la Tabla 1 se muestran los valores que definen las primeras 24 BCs, según Zwicker, los cuales se han convertido en un estándar "de facto" para describir la distribución de las bandas críticas en función de la frecuencia.

3.6 Transformada de Fourier (TF)

La transformada de Fourier es una herramienta matemática que se podría considerar la más importante para el procesamiento de audio. Ya que asigna una función dependiente del tiempo a una función dependiente de la frecuencia, y esta revela el espectro de frecuencia que compone la función original; es decir, la transformada de Fourier nos muestra dos lados de la misma información.

Cuando una función *f* depende del tiempo esta solo muestra la información en el tiempo ocultando la información que podemos obtener en frecuencia. Un ejemplo claro de información que se puede obtener es cuando en una grabación de un instrumento, por ejemplo en una guitarra una nota es producida, se ve reflejada la información en el tiempo, pero lo que no se muestra es que notas han sido ejecutadas.

La transformada de Fourier de una función $\hat{f}$ esconda toda aquella información que se mostraba en el tiempo y en cambio muestra toda la información acerca de las frecuencias. En este caso haciendo referencia al ejemplo de la guitarra solo observaríamos aquellas notas que han sido ejecutadas pero no tendríamos idea de en qué tiempo estas fueron tocadas.

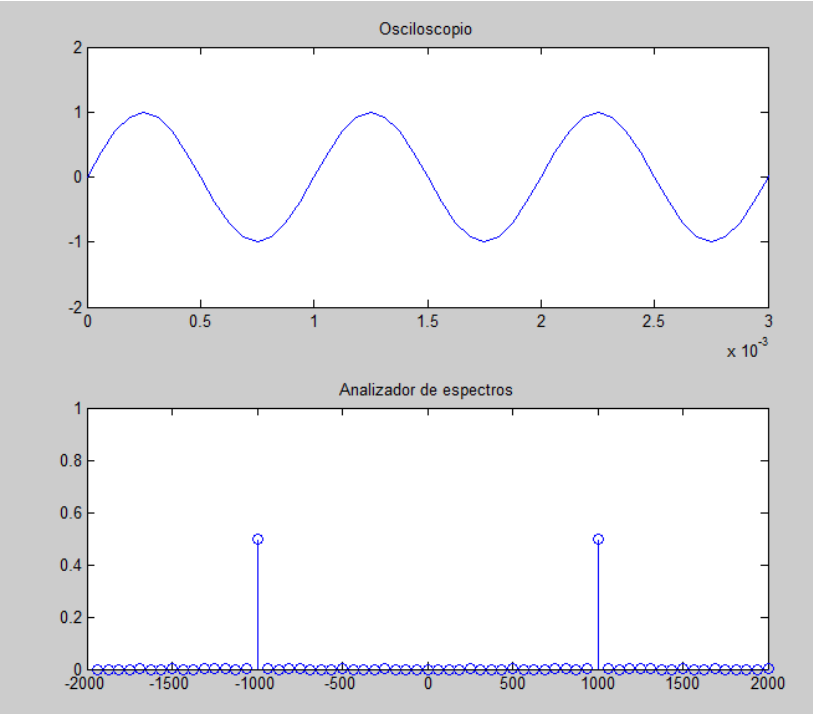


Figura 3.6 Representación de una señal en tiempo y en frecuencia.

En forma continua es como hemos tratado la transformada de Fourier. Como es sabido, ésta tiene su similar en forma discreta. Tales señales poseen una representación Fourier que puede ser considerada como una superposición ponderada de señales senoidales de frecuencias diferentes, el cálculo de le transformada de Fourier consiste en evaluación de integrales o sumas infinitas.
Para realiza el cálculo en la práctica se debe de realizar aproximaciones, para obtener la transformada por medio de sumas infinitas, esto puede realizarse eficientemente por la conocida transformada rápida de Fourier (FFT).

La transformada de Fourier en su forma continua puede ser definida como en (2.2),

$$x(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x(t).e^{-j\omega t} dt$$

En la práctica en el procesamiento de señales y otras áreas se utiliza *t* y ω, que están designadas al tiempo y frecuencia respectivamente. La expresión que nos lleva del dominio de la frecuencia al dominio del tiempo se encuentra descrita en (2.3).

$$x(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} x(\omega).e^{j\omega t} d\omega$$

La idea base de la representación de Fourier, es la de representar una señal como una superposición ponderada de las funciones de frecuencia elementales independientes. Cada una de las ponderaciones expresa la extensión que le corresponde, su función elemental que contribuye con la señal original, para revelar ciertos aspectos de la señal.

3.7 Transformada discreta de Fourier

El análisis de frecuencias de señales discretas en el tiempo, es el más utilizado en procesamiento digital de una señal; para trabajar en el análisis de frecuencias en tiempo discreto de una señal x(n) tenemos que transformar la secuencia del dominio del tiempo al domino de la frecuencia en una representación equivalente como hemos visto anteriormente.[24]

Al tratar de realizar una transformada de Fourier de una señal se requiere realizar la evaluación de integrales o de sumatorias infinitas, que es en general una tarea no viable. Por lo cual la forma en que se emplea esta transformación es considerando la representación de una secuencia *x(n)* de su espectro *X(ω)*, tal representación en el dominio de la frecuencia nos lleva a la transformada discreta de Fourier(DFT) que es una herramienta muy útil para el análisis de frecuencia de señales en el tiempo discreto.

Con el muestreo en el dominio de la frecuencia de una secuencia finita aperiódica x(n), en general las muestras separadas en frecuencia.

$$\mathrm{X}\left(\frac{2\pi k}{N}\right), k = 1, 2, 3, ..., (N-1)$$

No representa únicamente la secuencia original x(n), cuando x(n) tiene duración infinita. Pero en cambio la frecuencia de las muestras en (2.4), corresponden a una secuencia periódica $x_p(n)$ que tiene periodo N, donde $x_p(n)$ es un alias de x(n) como se indica en (2.5)

$$x_p = \sum_{l=-\infty}^{\infty} x(n - lN)$$

Como la frecuencia de las muestras es obtenida evaluando la transformada de Fourier *X(ω)*, de un set de *N* (igualmente espaciadas) de frecuencias discretas, en relación es llamada transformada discreta de Fourier(DFT) de x(n).

$$X(k) = \sum_{k=0}^{N-1} X(k) e^{-j2\pi kn/N} \qquad \text{k=1,2,3,…,N-1}$$

Y para recobrar la secuencia de x(n) de las nuestras de frecuencia, tenemos la expresión en (2.7)

$$X(k) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j2\pi kn/N}$$

En la práctica es utilizada la llamada transformada rápida de Fourier, matemáticamente la DFT es una aproximación de la transformada de Fourier, donde DFT de tamaño *N* es una asignación lineal de $\mathbb{C}^N \to \mathbb{C}^N$ dado por una matriz de *NxN*

$$DFT_N = \left(\Omega_N^{kj}\right)_{0\le k,j\le N} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & \Omega_N & \cdots & \Omega_N^{(N-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \Omega_N^{(N-1)} & \cdots & \Omega_N^{(N-1)(N-1)} \end{pmatrix}$$

Donde $\Omega_N = e^{-2\pi i/N}$, para un vector de entrada $v = \left(v(o), v(1), ..., v(N-1)\right)^T \in \mathbb{C}^N$ la evaluación de la DFT está dado por el productor matriz-vector $\hat{v} = DFT_N \bullet v$, donde $\hat{v} = \left(\hat{v}(o), \hat{v}(1), ..., \hat{v}(N-1)\right)^T \in \mathbb{C}^N$ que denota la salida del vector. Se puede notar el cálculo directo del producto de la matriz-vector requiere $O\left(N^2\right)$ sumas y multiplicaciones. Para la mayoría de las aplicaciones esto es demasiado lento en muchos casos se tiene que lidiar con largos de $N >> 10^5$. El punto importante es que existe un algoritmo eficiente, que es llamado transformada rápida de Fourier (FFT) el cual solo requiere *O(NlogN)* sumas y multiplicaciones. La idea de la FFT fue originalmente ideada por Gauss, pero fue redescubierta por Coley and Tukey basándose en la factorización de la matriz DFT consistiendo de *O(logN)* escasas matrices cada una de las cuales se puede evaluar con O(N) operaciones.

3.8 Filtros digitales

Un filtro es un sistema cuya finalidad es modificar el espectro de la señal de entrada ya sea este continuo o discreto, en el audio estos filtros son utilizados para modificar la forma de onda en modo específico para a alterar las propiedades de su espectro. Siendo unos de los más referidos para este trabajo aquellos que atenúan ciertas partes a un determinado ancho de banda; todos los filtros satisfacen cierta linealidad, estabilidad y varianza, la cual puede ser expresada mediante una convolución, una transformada discreta de esta señal *s*. La máxima tasa de muestreo de un filtro digital es igual a la mitad de la frecuencia de muestreo del teorema de Shannon.

En general, los filtros digitales se pueden clasificar en dos categorías que son: filtros con respuesta al impulso infinita (IIR) y los filtros con respuesta al impulso finita (FIR) también conocidos como filtros transversales. En el caso de los filtros IIR se aprovechan métodos para diseñar filtros analógicos, y estos se transforman en filtros digitales usando diferentes transformaciones (lineal, bilineal, etc.). Estos filtros son inestables ya que la transformada rápida tiene tanto polos como ceros, y en el caso de los filtros FIR estos siempre son estables. Hay distintos caminos para diseñar este tipo de filtros pero este va a depender de la circunstancia en la que sea empleado.

Algunas ventajas de un filtro digital sobre su símil analógico son: la respuesta en frecuencia es más cercana a la ideal; no requieren de sintonización, pueden multiplexarse para el procesamiento de más señales; su rediseño es sencillo ya que solo implica cambiar los coeficientes de este, sus componentes son independientes de la frecuencia de operación del filtro; tienen un alto grado de integración por lo que se tiene mayor confiabilidad.

Para el diseño de filtros IIR, se requiere de la transformación de un filtro analógico a un digital que satisfaga las necesidades requeridas. Este tipo de diseño es directo, ya que se aplican métodos de diseño ya derivados de los filtros analógicos y así en ocasiones se requerirán de un filtro analógico usando un digital.

Hay una variedad de filtros analógicos que pueden ser usados para el diseño de filtros digitales, como son el filtro Butterworth, Chevyshev, Elípticos, Bessel, entre otros.

La transformada de Fourier de la señal se conoce como la respuesta en frecuencia de un filtro que puede referirse a importantes características sobre la selectividad de un filtro subyacente. La respuesta en frecuencia puede ser utilizada para especificar ciertas características de un filtro a como lo requiera la aplicación.

3.9 Diseño y especificaciones de los filtros.

En el proceso de diseño de un filtro se empieza por las especificaciones del mismo las cuales nos darán las limitaciones en magnitud y/o fase de la frecuencia de respuesta, limitaciones sobre las muestras o unidad de respuesta del paso del filtro, así como el tipo

de filtro (FIR o IIR), y el orden del filtro. Una vez que estas especificaciones del filtro se han definido el siguiente paso es encontrar un conjunto de coeficientes en el filtro que produzcan un filtro adecuado. Cuando esté terminado el diseño de un filtro por último surge su implementación en el sistema ya sea por medio del hardware o software. De ser necesario se deben cuantizar los coeficientes del filtro y escoger la estructura adecuada para el filtro.

Si suponemos que queremos diseñar un ideal filtro pasa-bajas con una frecuencia de corte en ω_0. En la teoría podríamos obtener dicho filtro simplemente con invertir la frecuencia de respuesta como se observa en la figura 2.5.

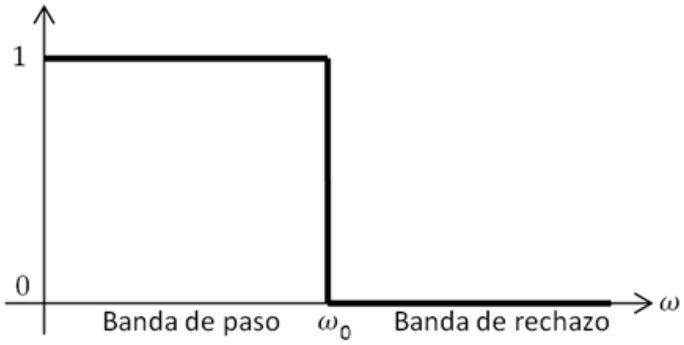


Figura 3.7 Filtro ideal.

Si denotamos una función sic como sigue:

$$\text{sicn(t)} = \begin{cases} \dfrac{\sin \pi t}{\pi t} & para \quad t \neq 0 \\ 1 \quad para \quad t = 0 \end{cases}$$

y luego mediante un cálculo encontramos que los coeficientes para un filtro ideal está dado por $h(n) = 2\omega_0$ sinc $(2\omega_0 n)$ para $n \in \mathbb{Z}$. Sin embargo este filtro tiene muchos inconvenientes dado que tiene un número infinito de coeficientes del filtro diferentes de cero no es casual ni estable.

En la actualizad realizar un filtro ideal con estas características no es posible, por lo que se tiene que trabajar con aproximaciones que pueden tener fenómenos como los que

siguen: la frecuencia de respuesta *H* de un filtro que es realizable presenta rizos en la banda de paso y la banda de rechazo.

Además de que *H* no tiene ningún punto de corte brusco en la banda de paso y en la banda de no paso, véase la figura 2.6, tampoco *H* puede caer de la unidad a cero abruptamente.

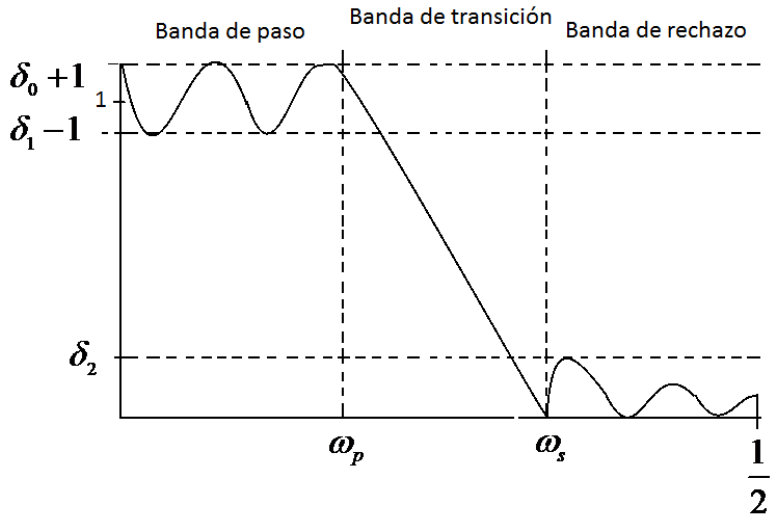


Figura 3.8 Filtro Pasa Bajas (FPB).

En las aplicaciones algunas cosas pueden ser toleradas como son los rizos en la banda de paso y en la banda de no paso. La transición de la respuesta en frecuencia de la banda de paso a la banda de no paso figura (2.6), es llamada banda de transición de filtro. El límite de la banda de frecuencia ω_p define el límite de la banda de paso, mientras que el parámetro ω_s va a definir el comienzo de la banda de rechazo. La diferencia que existe entre estas dos $\omega_s - \omega_p$ es conocida como ancho de transición, de manera similar en la parte de la banda de paso se conoce como ancho de banda de filtro.

Si hay rizos en la banda de paso del filtro, la máxima desviación de los rizos por arriba y por debajo de 1 son denotados por δ_0 y δ_1 respectivamente. Si la magnitud de $|H|$ varia dentro del intervalo de $[1-\delta_1, 1+\delta_0]$. La máxima magnitud de los rizos en la

banda de rechazo del filtro esta denotada por δ_2 , y de la misma manera estas características pueden ser definidas para un filtro pasa-altas.

En el caso de un filtro pasa-banda tiene dos bandas de rechazo, así como también tiene dos bandas de transición a la izquierda y a la derecha. En el diseño habitual de un filtro, un filtro *h* es construido de forma que se encuentra dentro de sus especificaciones y clase dentro de lo que es requerido. Y del grado en que *H* se aproxime a las especificaciones del filtro depende del orden del filtro.

Como se había mencionado anteriormente existen otras especificaciones referidas a la respuesta en fase del filtro. Si un filtro *h* es la función de muestra de frecuencia elemental $n \to e_\omega(n) = e^{2\pi i \omega n}$:

$$\left(h * e_\omega\right)(n) = \sum_{k \in \mathbb{Z}} h(k) e^{2\pi i \omega (n-k)} = H(\omega) e_\omega(n) = \left|H(\omega)\right| e^{2\pi i (\omega n + \phi_h(\omega))}$$

Se puede notar que la fase induce a un desplazamiento en el tiempo en función de la frecuencia elemental, que generalmente depende de ω. Para una señal de entrada en general tal retraso en todos los componentes de frecuencia conduce a un retardo global en el proceso de filtrado que no es considerado como una distorsión lo cual lleva a ciertas definiciones.

Si $\phi_h = c\omega$ módulo 1 para algún $c \in \mathbb{R}$, el filtro h se dice que es de fase lineal, la función $\tau_h : [1, 0] \to \mathbb{R}$ está definida por:

$$\tau_h(\omega) = \frac{d\phi_h}{d\omega}(\omega)$$

Es llamado retraso de grupo de un filtro h donde las discontinuidades en la fase son a consecuencia de las ambigüedades que no se consideran. El valor de τ_h puede ser interpretado como el tiempo de retraso que un componente de la señal de frecuencia ω sufre en el proceso del filtrado.

3.10 Los coeficientes cepstrales de frecuencia Mel

Para representar mejor el patrón de percepción del oído humano se desarrolló una escala llamada Mel, que es una escala logarítmica. Entonces, los coeficientes cepstrales de frecuencia Mel (MFCC por sus siglas en inglés Mel Frequency Cepstral Coefficients), son coeficientes para la representación del habla basados en la percepción auditiva humana. Los MFCC se derivan de la transformada de Fourier, la diferencia es que en MFCC las bandas de frecuencia están situadas logarítmicamente, según la escala de Mel. Los MFCC modelan la respuesta auditiva humana más apropiadamente que las bandas espaciadas linealmente de FT. Esto permite un procesado de datos más eficiente, por ejemplo, en compresión de audio.

La escala Mel, fue propuesta por Stevens, Volkman y Newman en 1937, el punto de referencia entre esta escala y la frecuencia normal se define equiparando un tono de 1000 Hz, 40 dBs por encima del umbral de audición del oyente, con un tono de 1000 mels. Por encima de 500 Hz, los intervalos de frecuencia espaciados exponencialmente son percibidos como si estuvieran espaciados linealmente. En consecuencia, cuatro octavas en la escala de hercios por encima de 500 Hz se comprimen a alrededor de dos octavas en la escala mel.

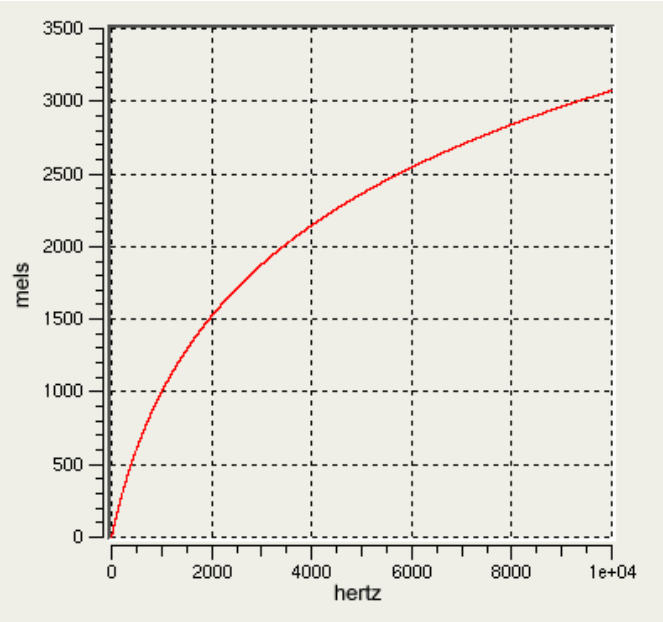


Figura 3.9. Escala Mel.

3.11 MFCC (Mel Frequency Cepstrum Coefficients)

Un método más eficiente para extraer características y que es el más utilizado actualmente en reconocedores comerciales son los Coeficientes Cepstrales en Escala de Mel (MFCC, por sus siglas en inglés), este método es *robusto*, además hace uso de la Transformada de Fourier para obtener las frecuencias de la señal. El objetivo es desarrollar un conjunto de características basadas en criterios perceptuales, diversos experimentos muestran que la percepción de los tonos en los humanos no está dada en una escala lineal, esto hace que se trate de aproximar el comportamiento del sistema auditivo.

Los coeficientes Cepstrales en Frecuencia en Escala de Mel (MFCC) son una representación definida como el cepstrum de una señal ventaneada en el tiempo que ha sido derivada de la aplicación de una Transformada Rápida de Fourier, pero en una escala de frecuencias no lineal, las cuales se aproximan al comportamiento del sistema auditivo humano [16].

El cálculo de los coeficientes mels utiliza dos de las herramientas más conocidas en el análisis de señales:

- La transformada de Fourier para la representación del contenido espectral de una señal.
- ·El diseño de un banco de filtros para permitir la selección de bandas de frecuencia de la señal bajo análisis.

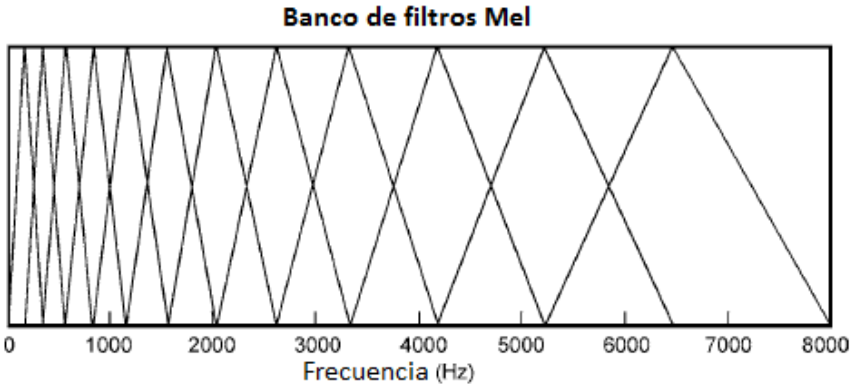


Figura. 3.10 Banco de filtros Mel.

Con la transformada de Fourier se conoce el contenido en frecuencia (espectro) de la señal y con los filtros diseñados (sintetizados), se logra obtener las componentes de frecuencia que a cada banda les aporta la señal analizada.

El principio de ponderar la energía que aporta a cada banda de frecuencias la señal bajo análisis y luego calcular en términos de un coeficiente para cada valor de energía en banda de frecuencia, es a lo que llamamos coeficientes cepstral.

El algoritmo o método para el cálculo de estos coeficientes, usando las dos herramientas mencionadas es lo que se describe a continuación, figura 3.11.

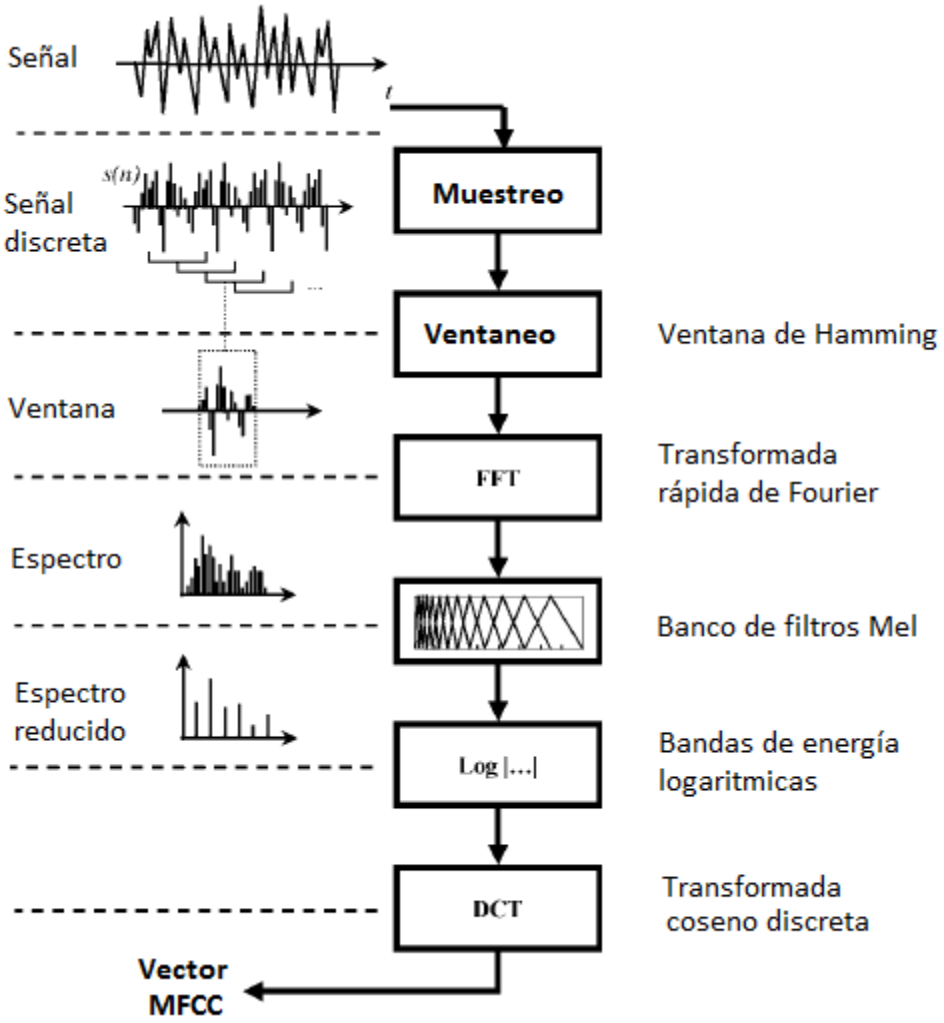


Figura 3.11. Diagrama de flujo para el cálculo de los coeficientes MFCC

Davis y Mermelstein en 1980 demostraron que los MFCC son beneficiosos para el Reconocimiento Automático del Habla.
Dada una Transformada Discreta de Fourier de una señal de entrada:

Se distribuye el comportamiento espectral en bandas de frecuencia mediante un banco de filtros con los que se calcula el promedio del espectro alrededor de cada frecuencia central. Se puede definir a fl como la frecuencia más baja y a fh como la frecuencia más alta del banco de filtros en Hz, Fs es la frecuencia de muestreo en Hz, M el número de filtros y N el tamaño de la Transformada Rápida de Fourier.

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-\frac{j2\pi nk}{N}} \quad k = 0,1,\ldots..,N-1$$

Normalmente Normalmente los filtros triangulares son repartidos en el rango de frecuencias completas, desde cero hasta la frecuencia de Nyquist. Sin embargo, un criterio de limitación en banda es comúnmente útil para rechazar frecuencias no deseadas o evitar la construcción de filtros en regiones de frecuencia en las cuales no existe energía de la señal útil. Para el análisis de filtros solamente, las frecuencias de corte más bajas y las más altas se pueden establecer utilizando parámetros de configuración LOFREQ y HIFREQ, por ejemplo:

LOFREQ=300
HIFREQ=3400

Pueden ser utilizadas para procesar señales telefónicas. Cuando se especifican las frecuencias de corte baja y alta de esta forma, el número especificado de canales de los bancos de filtro son distribuidos de forma igual a lo largo de toda la escala de Mel resultando en un conjunto de filtros pasa-banda, en donde la frecuencia de corte más baja del primer filtro es igual a LFREQ y la frecuencia de corte del último filtro está en HIFREQ.

3.12 El pitch

El análisis del pitch involucra diferentes tópicos dentro del estudio de señales sonoras aún no explorados completamente. En virtud de la imprecisión en su propia definición se pueden implementar una gran variedad de algoritmos para su adquisición.
Históricamente se ha definido al pitch como la frecuencia fundamental de espectro de frecuencias del habla y se lo ha asociado al movimiento que realiza la glotis en la generación del sonido. Desafortunadamente cualquiera sea la forma en la que se lo defina no se ajustará a la realidad, porque la oscilación glotal es una función cuasi-periódica.
Además, esta frecuencia no es fácilmente identificable debido a que en algunas situaciones prácticamente desaparece de la onda sonora. Esto ocurre cuando las articulaciones del tracto vocal hacen que la energía del sonido se concentre en algunos de sus armónicos. No obstante no se lo pierde completamente y se puede utilizar dichos armónicos para su rastreo.
Se ha observado que esta vibración no es constante a lo largo del discurso, detectándose variaciones a lo largo de la frase y también dentro mismo de una palabra.
Estas variaciones se deben tanto a la entonación de la frase, como a la acentuación de los fonemas así como al estado emocional del orador.

El pitch, es una propiedad del sonido que puede ser percibida en una escala de frecuencia determinada y puede ser utilizado para interpretar una melodía. El pitch de un sonido cualquiera puede ser medido en Hertz y a manera de ejemplo es posible representarlo con una señal senoidal.

El pitch está ligado a otra característica del sonido llamada timbre. Lo cual para la percepción del ser humano es muy importante; ejemplo, supongamos que percibimos el canto del ave, de cierta manera podemos percibir la tonalidad del canto del ave. ¿Pero como sabemos que es un ave la cual está emitiendo el sonido?, el timbre es lo que nos va a brindar esto, dado que el timbre es la mezcla de frecuencias acompañadas con la frecuencia fundamental (pitch) que otorgan las características al sonido que oímos, como es: la posición. La intensidad y otra serie de atributos que puede no se hayan definido aún. Y es de esta manera como el ser humano percibe la música. Hay notas musicales

que podemos percibir, y con el timbre podemos definir si es una guitarra, un piano, un violín, etc.

El pitch es una propiedad subjetiva, la cual implica problemas involucrados en psicoacústica. En el cual se realizan los estudios de procesamiento y de la percepción auditiva. Para el tema de estudio expuesto en este trabajo de tesis le veremos como la frecuencia fundamental de una tonalidad.

Al percibir un sonido, existe de por medio como es conocido, una frecuencia fundamental. Pero esto implica que esta va a tener presentes armónicos los cuales van a influenciar en la representación musical. Este evento puede ser percibido pde una mejor forma en la representación de sonidos polifónicos que veremos más adelante,

En la forma de onda de una señal de audio, no puede percibirse ninguna periodicidad aparente, como en la representación del pitch con una onda senoidal. Hay una gran variedad de instrumentos y timbres de voz. Que pueden ser analizados, los cuales tendrán formas de onda muy complejas y en adición en ciertos intervalos de tiempo habrá similitudes. Es por esto que existen diversos procesos para obtener el pitch de una onda de audio. En nuestro caso lo obtenemos filtrando la señal y luego obteniendo la FFT.

CAPÍTULO 4

DESARROLLO DEL SISTEMA

4.1 Extracción de características.

Se tienen grabaciones de 1 minuto aproximadamente cada una y se procesan de tal manera que sean casi iguales en cuanto a amplitud y frecuencias.

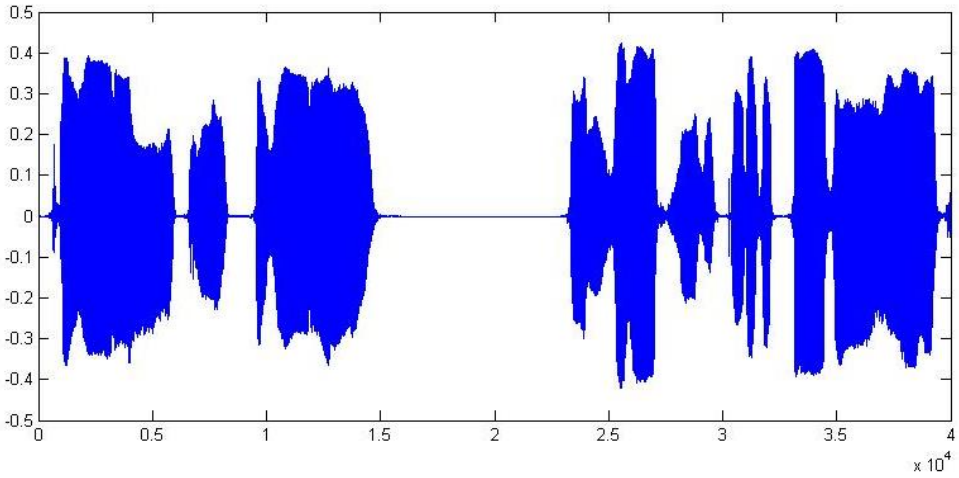


Figura 4.1 Grabación de voz

4.2 El pre-procesamiento.

La etapa de pre-procesamiento es común para el posterior trabajo del sistema en las etapas de entrenamiento, verificación y trabajo cotidiano del mismo. Su importancia radica en una serie de pasos que normalizan las características en el dominio del tiempo de las señales de voz grabadas o captadas directamente desde un micrófono.

Estos pasos son:

4.3 Preénfasis y normalizar.

El preénfasis es para restaurar la pérdida que sufre la señal de voz en sus componentes de alta frecuencia, por efecto de propagación y radiación al salir de la cavidad bucal a través de los labios al ambiente exterior. Se realiza pasando toda la señal a través de un filtro pasa alta de un solo coeficiente, figura 4.

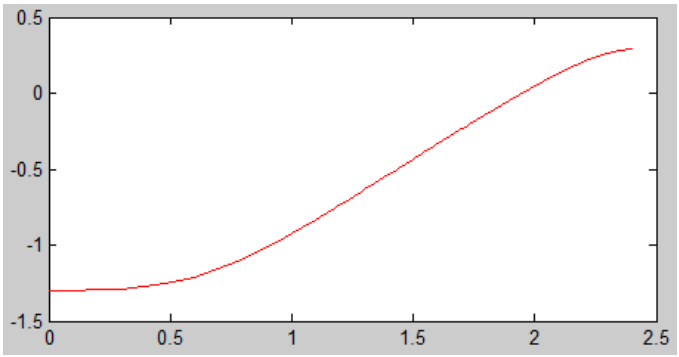


Figura 4.2. Respuesta en frecuencia del filtro de preénfasis. El filtro queda de la siguiente manera:

El filtro queda de la siguiente manera:

$$y(n) = x(n) - a * x(n-1)$$

Donde:

y es la señal preénfatizada. x es la señal de entrada.

a el coeficiente del filtro, 0.90 ≤ a ≤ 0.97.

n es la secuencia n-esima de muestras

Normalizar un vector consiste en obtener otro vector unitario, de la misma dirección y sentido que el vector dado. Para normalizar un vector se divide éste por su módulo.

audioNorm= audioPre / |audioPre|

Donde:

audioNorm va a ser el resultado al normalizar el audio.

audioPre es la señal preénfatizada.

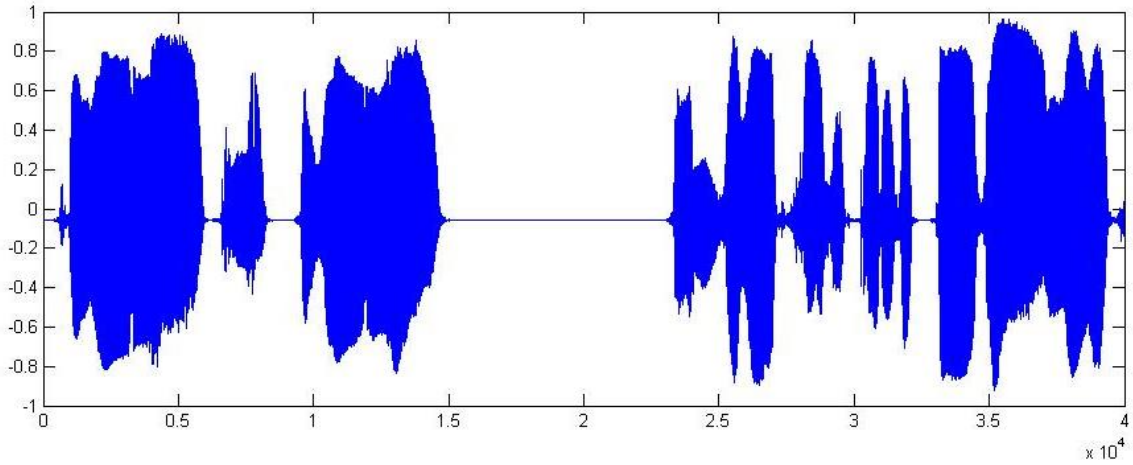


Figura 4.3. Audio con preénfasis y normalizado.

Se identifica donde está la mayor cantidad de energía en el audio lo cual esperamos que equivalga a voz y se eliminan los silencios

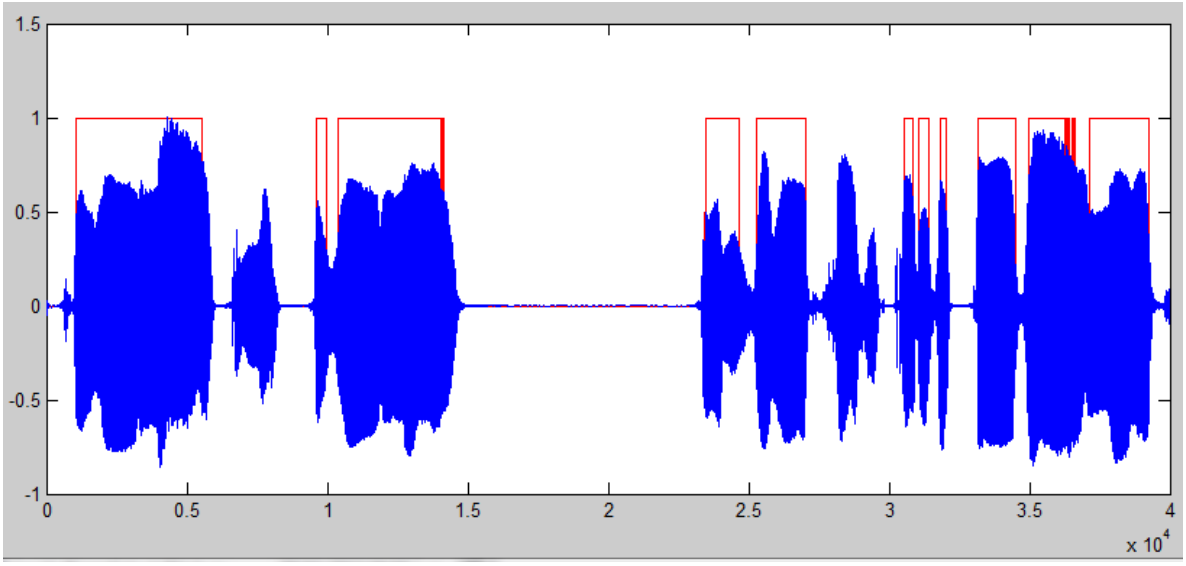


Figura 4.4. Audio con identificación de energía.

Generamos un archivo con aproximadamente 10 segundos de audio exclusivamente sin silencios y con el cual podemos trabajar a gran velocidad de procesamiento para extraer otras caracteristicas como MFCC, Pitch, Energia en Bandas Críticas, etc

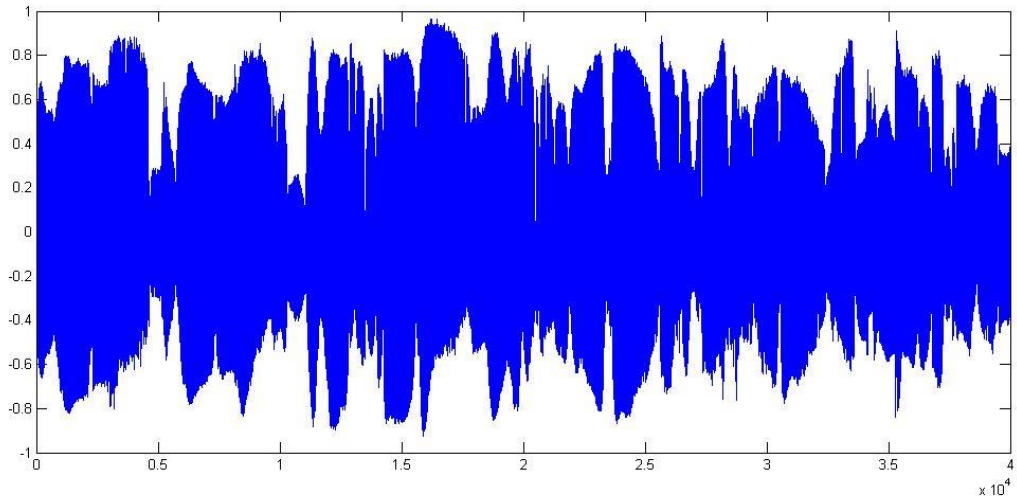


Figura 4.5. Audio recortado, quitando los silencios.

4.4 Filtrado

EI filtro es un sistema diseñado para obtener una característica de transferencia deseada. Esto es, opera sobre una señal (o señales) de entrada en una forma predeterminada. Los filtros lineales pasivos por lo general se consideran parte del estudio de circuitos, redes o sistemas lineales. Están compuestos de una combinación de resistencias, inductores y condensadores. Aunque es posible obtener una amplia variedad de características de transferencia utilizando estos elementos, a menudo se requieren gran cantidad de componentes. Esto conduce a buscar alternativas a filtros activos.

Un filtro es un circuito que se ha diseñado para pasar una banda de frecuencias especificada, mientras atenúa todas las señales fuera de esa banda. Los circuitos de filtrado pueden ser activos o pasivos. Los circuitos de filtrado pasivo contienen solo resistencias, inductores y condensadores. Los filtros activos, emplean transistores o amp op más resistencias, inductores y condensadores. Los inductores a menudo no se utilizan en los filtros activos, debido a que son voluminosos y costosos y tienen grandes componentes resistivos internos.

4.4.1 Clasificaciones de los filtros:

Los filtros se pueden clasificar por:

- Función de Transferencia
- Orden
- Respuesta en Frecuencia
- Activos o Pasivos
- Analógicos, Digitales o Mecánicos
- Piezoeléctricos, etc.

4.4.2 Función de Transferencia:

La Función de Red determina la forma en que la señal aplicada cambia en amplitud y fase al atravesar el filtro, algunos filtros habituales son:

- o Filtro Butterworth
- o Filtro Chebyshev.
- o Filtro Bessel.
- o Filtro Cauer o elípticos.
- o Filtro Sallen Key.

En el presente trabajo utilizamos el Filtro Butterworth: en honor al ingeniero británico Stephen Butterworth; es un filtro básico, con respuesta más plana en la banda de paso y caída aguda en la frecuencia de corte a razón de 20n [dB/dec], donde *n* es el orden.

La función de transferencia del filtro en función de la ganancia K_{pb} a w=0, la frecuencia de corte y el orden del filtro *n* es:

$$\left|H(j\omega)\right| = \frac{K_{Pb}}{\sqrt{1+\left(\omega/\omega_c\right)^{2n}}}$$

$$n = 1,2,3,\ldots$$

El orden del filtro tiene que ver con el número de polos de la función de transferencia o con el número de redes presentes en la estructura. Mientras mayor sea el orden del filtro más aproximada será su respuesta a la respuesta ideal del filtro.

Si la frecuencia *w* es mucho mayor que la frecuencia de corte, puede demostrarse que la atenuación del filtro viene dada por:

$$Atenuación = -20 \cdot n \cdot \log\left(\frac{\omega}{\omega_c}\right)$$

$$n = 1,2,3,...$$

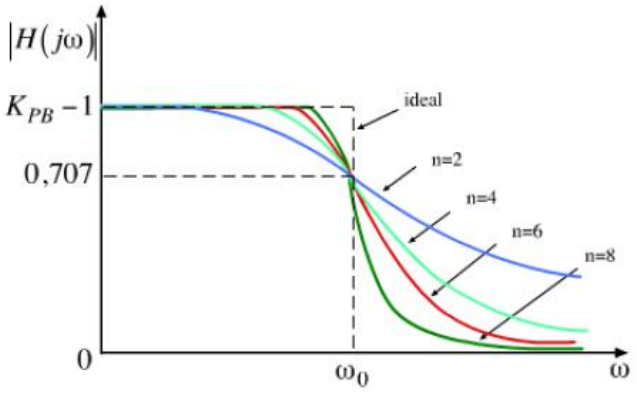


Figura 4.6. Coeficientes para un filtro Butterworth.

Es decir, un filtro Butterworth de primer orden tiene una atenuación de 20 dB/década, el de segundo orden 40 dB/década y el tercer orden 60 dB/década. Valores con respecto a la ganancia máxima 20 log K_{pb}.

Tenemos nuestro archivo de audio, previamente recortado, y con solo la señal de voz, y sobre este es el que vamos a trabajar.

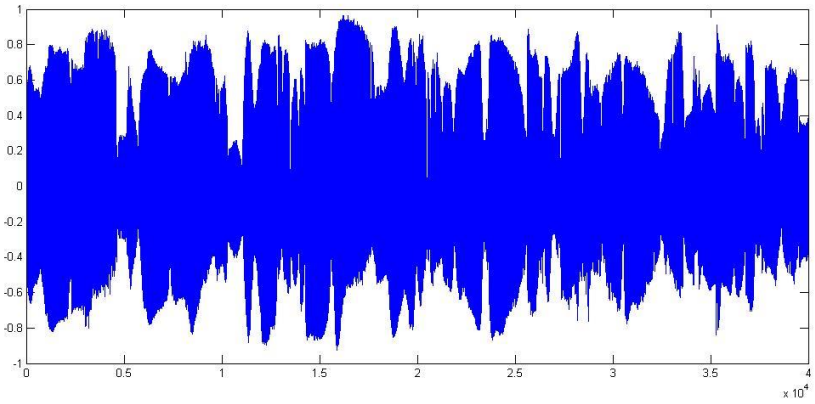


Figura 4.7. Audio con preénfasis y normalizado.

Si obtenemos la transformada de fourier para ver las componentes en frecuencia, obtenemos graficas como las que se muestran a continuación:

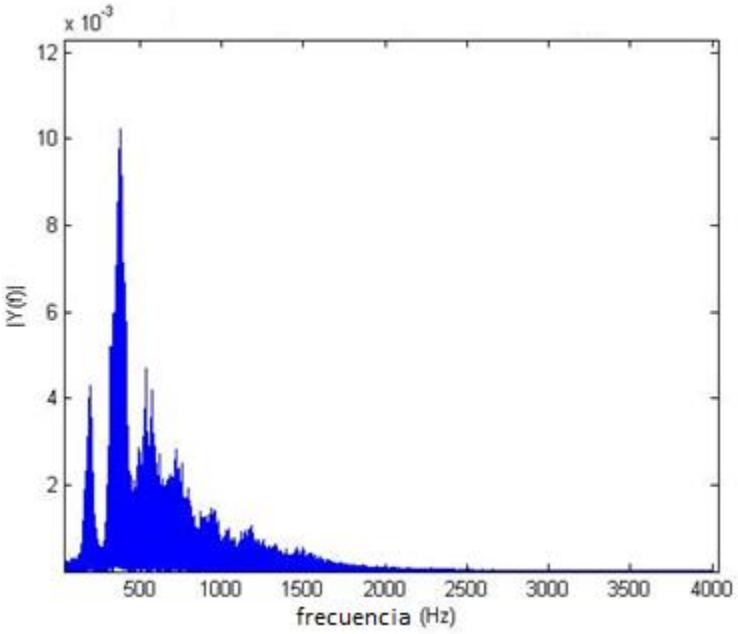


Figura 4.8. Transformada de Fourier de un audio.

4.5 Ventaneo

El mecanismo de ventaneo muestra como a cada porción de la señal de voz (de un tamaño predefinido por el usuario), se le asigna una ventana, de tal forma que las muestras queden ponderadas con los valores de la función escogida. En este caso, las muestras que se encuentran en los extremos de la ventana tienen un peso mucho menor

que las que se hallan en el medio, lo cual es muy adecuado para evitar que características de los extremos del bloque varíen la interpolación de lo que ocurre en la parte central, la cual es la más significativa, de las muestras del segmento seleccionado.

La colocación de las ventanas puede realizarse de tal forma que existan solapamientos y, aunque ello repercutirá en los tiempos de respuesta del sistema reconocedor, proporcionará una mejor calidad en los resultados obtenidos.

Las funciones de ventaneo más comunes se muestran en la tabla 2.

Ventana Rectangular	$w(t)=1$ en el intervalo temporal del bloque $(0 \le t \le L-1)$ $w(t)=0$ en el resto de la señal
Ventana Hanning	$w(t)=0.5-0.5*\cos\left(\frac{2\pi t}{L}\right)$ en $(0 \le t \le L-1)$ $w(t)=0$ en el resto de la señal
Ventana Hamming	$w(t)=0.54-0.46*\cos\left(\frac{2\pi t}{L}\right)$ en $(0 \le t \le L-1)$ $w(t)=0$ en el resto de la señal

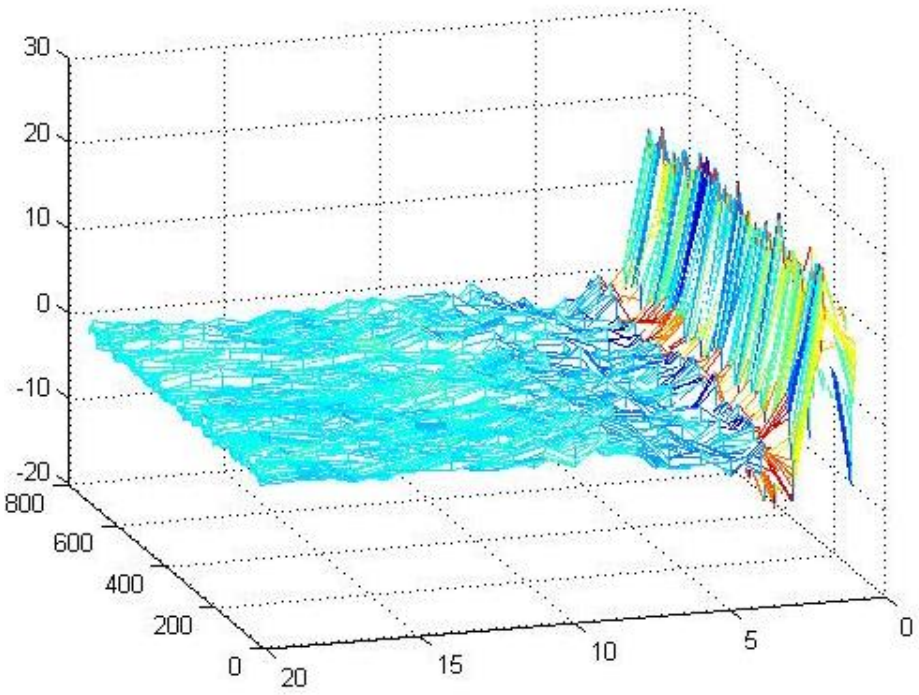
Tabla 4.1. Tipos de ventanas

La ventana más utilizada es la de Hamming, dado que es la que mejor respuesta en el dominio de la frecuencia aporta para eliminar los ruidos que introduce el truncamiento de la señal en segmentos. Para cualquier ventana, su duración determina la cantidad de cambios que se podrán obtener; con una duración temporal larga se omiten los cambios locales producidos en la señal, mientras que con una duración demasiado corta se reflejan demasiado los cambios puntuales y se reduce la resolución espectral.

4.6 Obtención de coeficientes MFCC

Después de obtener el archivo recortado y filtrado, procedemos a tomar ventanas de 512 muestras cada una para su análisis. Aquí mostramos las gráficas obtenidas de los MFCC para un ejemplo de cada uno de los 4 dialectos que tenemos.

a)



b)

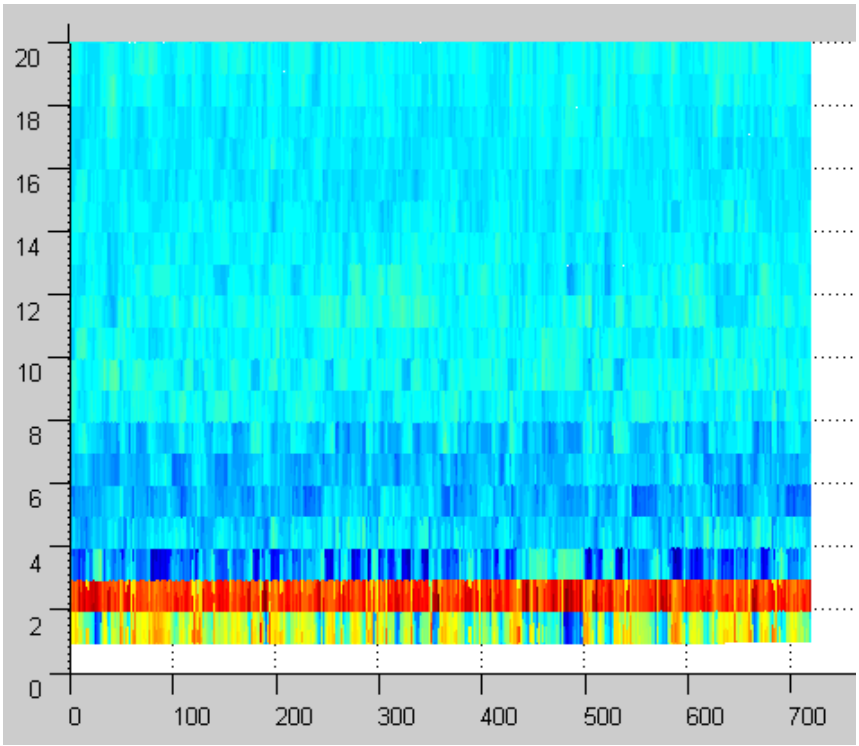
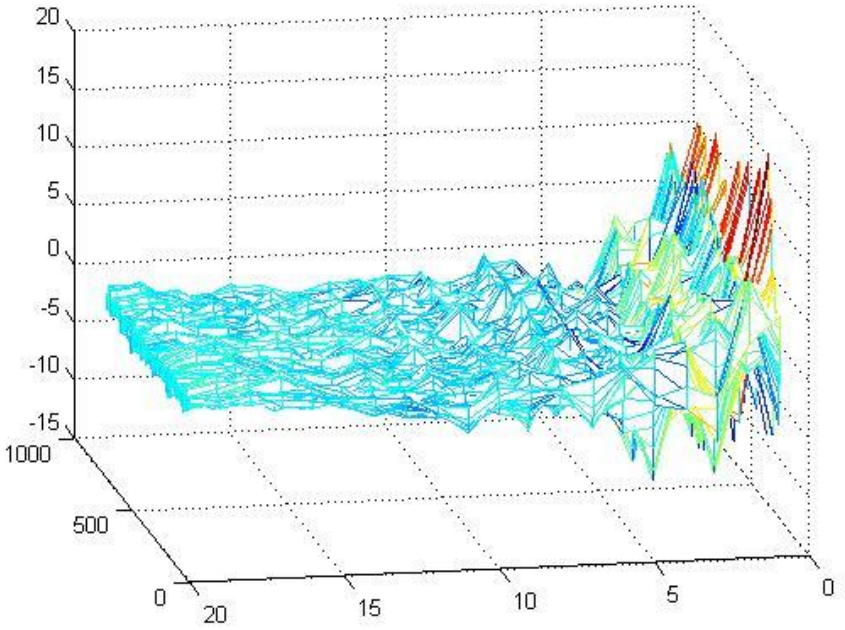
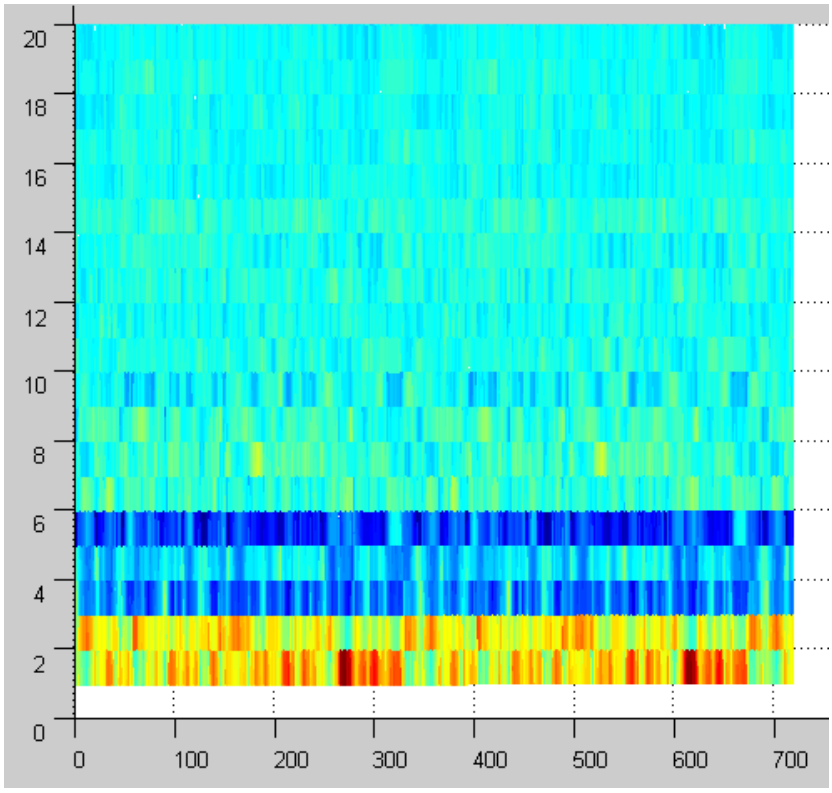


Figura. 4.9 Coeficientes MFCC para el dialecto Náhuatl. a) Vista 3D. b) Vista superior.

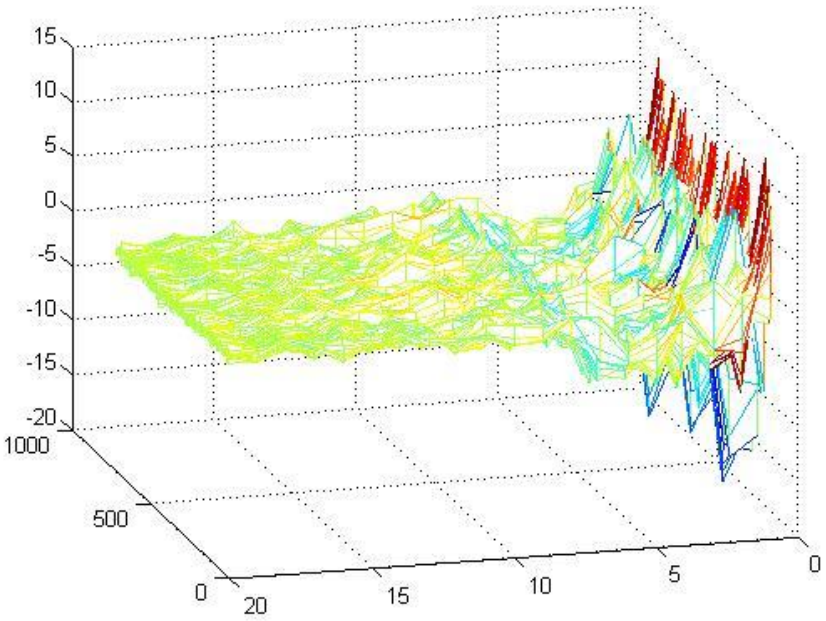


a)

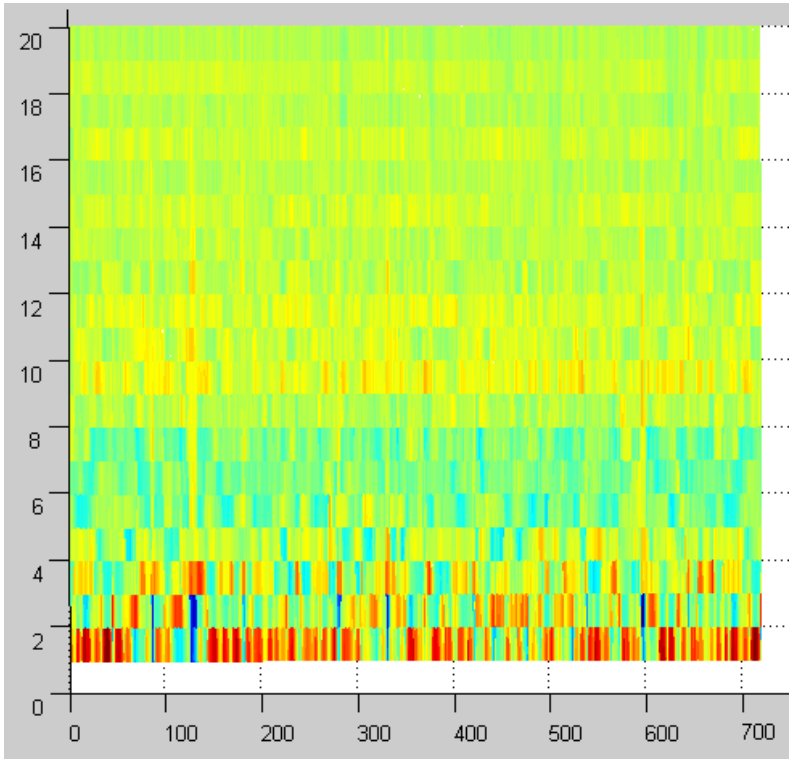


b)

Figura 4.10 Coeficientes MFCC para el dialecto Mixteco. a) Vista 3D. b) Vista superior.

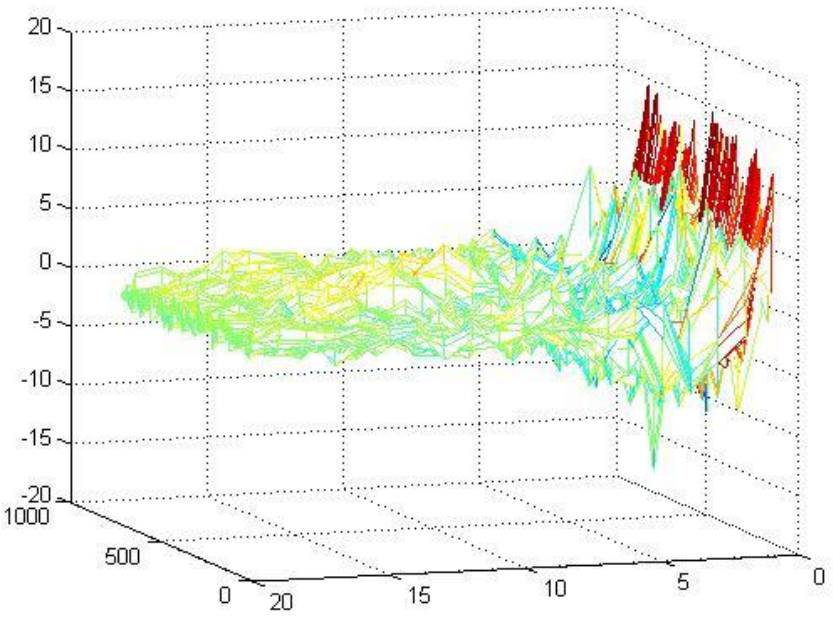


a)

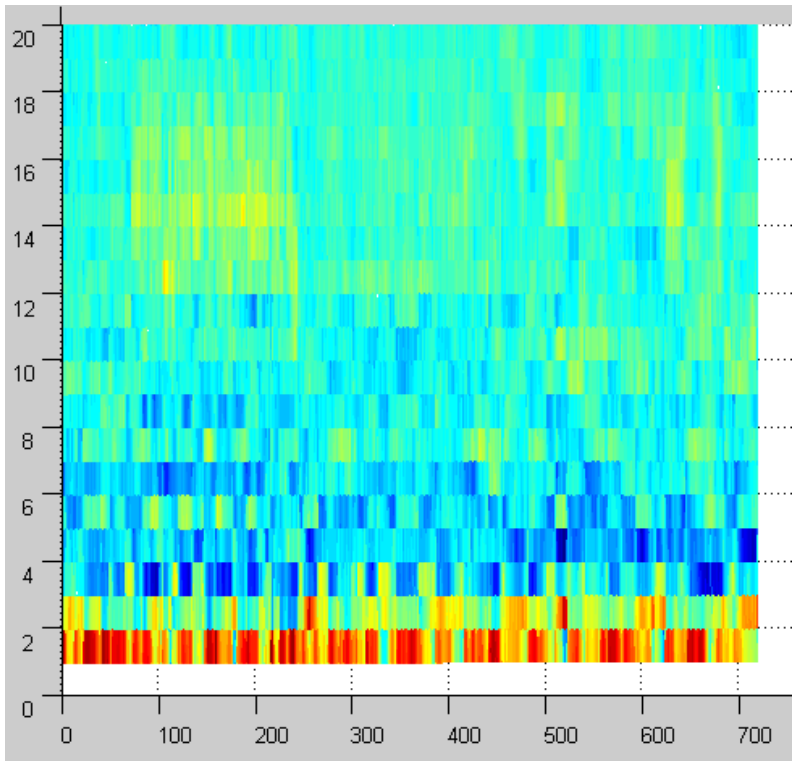


b)

Figura. 4.11 Coeficientes MFCC para el dialecto Maya. a) Vista 3D. b) Vista superior.



a)



b)

Figura 4.12. Coeficientes MFCC para el dialecto Ñañu. a) Vista 3D. b) Vista superior.

4.7 Obtención de Pitch

Con el mismo archivo recortado de 10 segundos, filtramos de 50 Hz a 500 Hz, que es el rango de frecuencias del pitch y procedemos a tomar ventanas de 512 muestras para hacer el cálculo (grafica azul) y posteriormente aplicamos un filtro de medianas para suavizar picos y obtener una gráfica mejor definida (gráfica roja). A continuación mostramos las gráficas obtenidas del pitch para cada uno de los 4 dialectos.

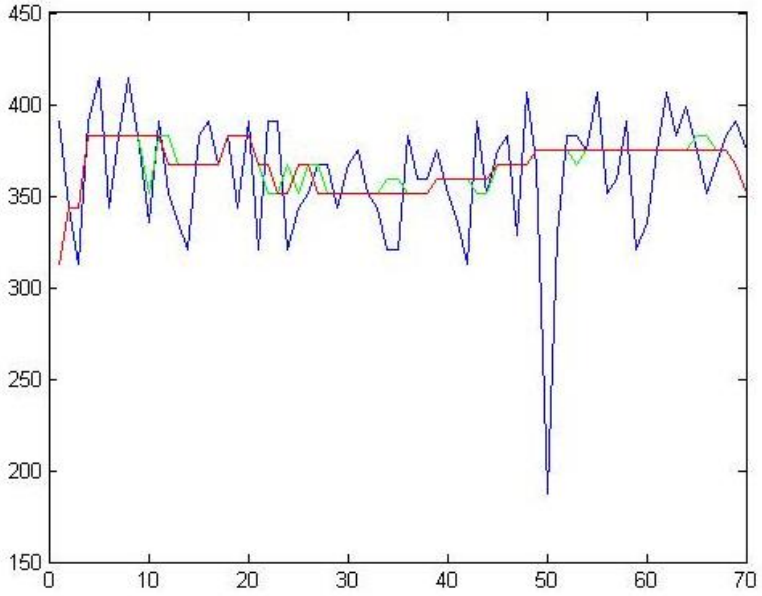


Figura 4.13. Gráfica del pitch para el dialecto Náhuatl.

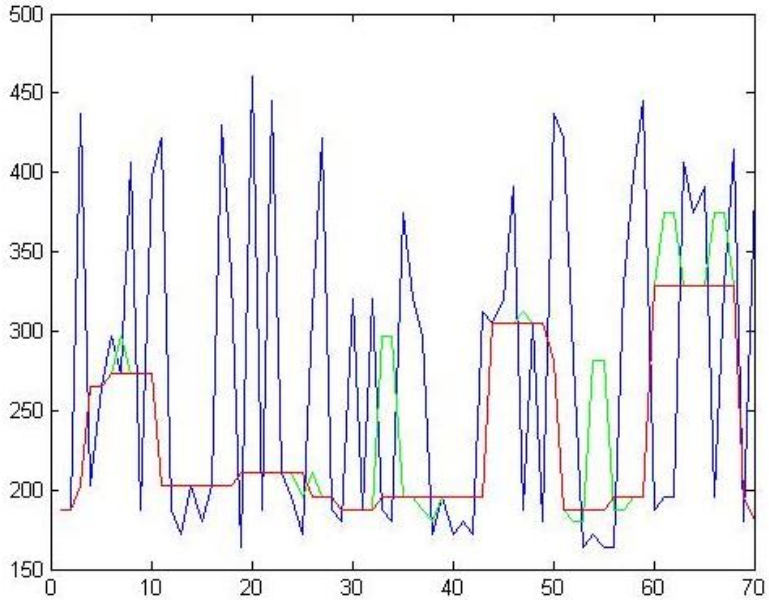


Figura 4.14. Gráfica del pitch para el dialecto Mixteco

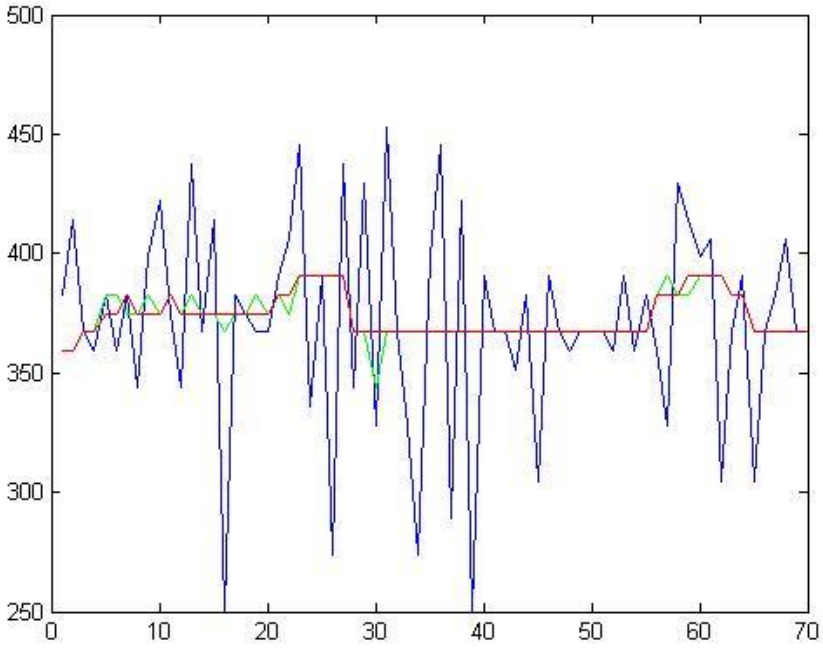


Figura 4.15. Gráfica del pitch para el dialecto Maya

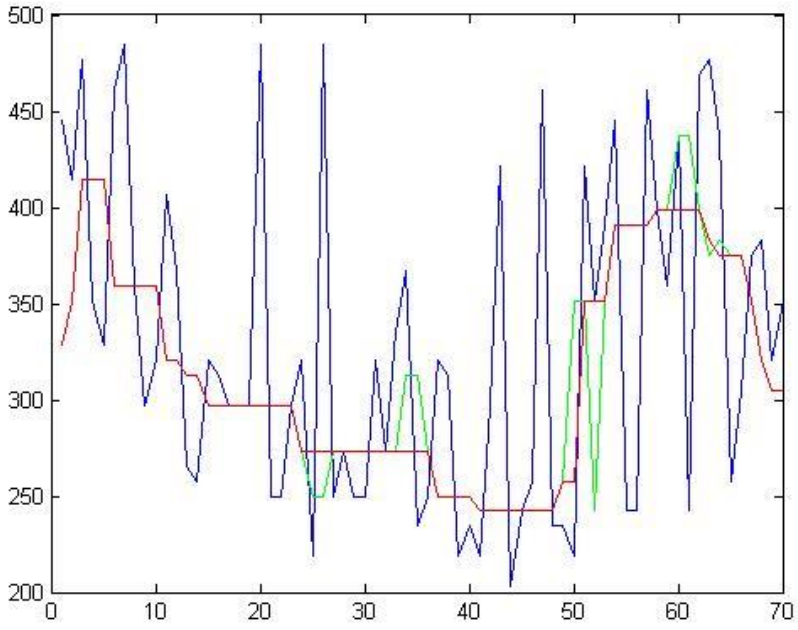


Figura 4.16. Gráfica del pitch para el dialecto Ñañu

4.8 Obtención de energía en bandas críticas

Con el mismo archivo recortado de 10 segundos, procedemos a tomar ventanas de 512 muestras y filtramos en bandas críticas para después obtenemos la energía que hay en cada una de ellas. A continuación mostramos las gráficas para cada uno de los 4 dialectos con los que contamos.

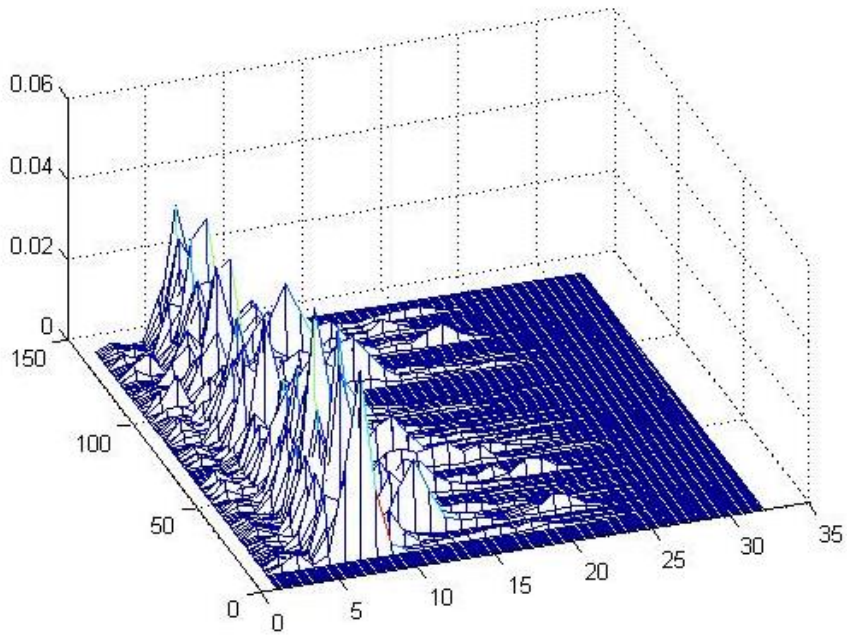


Figura 4.17. Gráfica de energía en bandas críticas para el dialecto Náhuatl

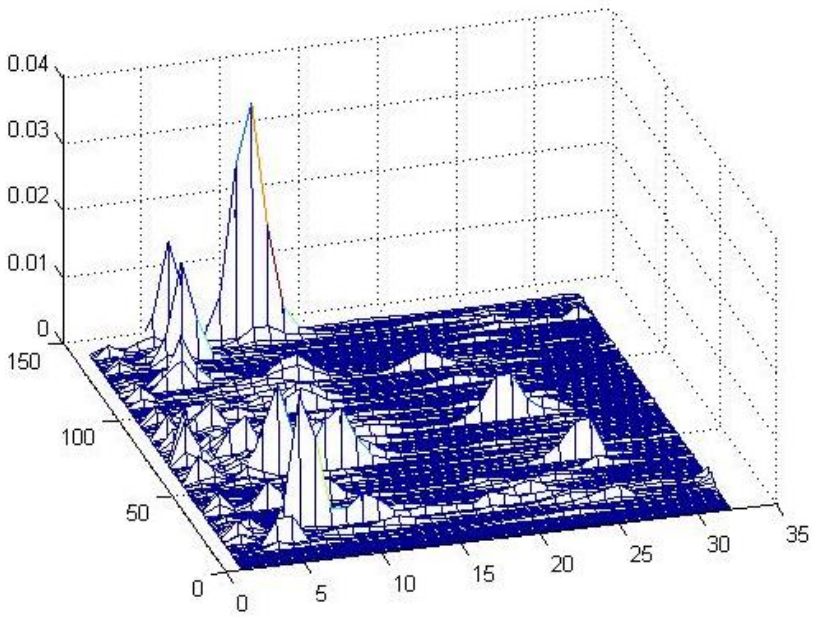


Figura 4.18. Gráfica de energía en bandas críticas para el dialecto Mixteco

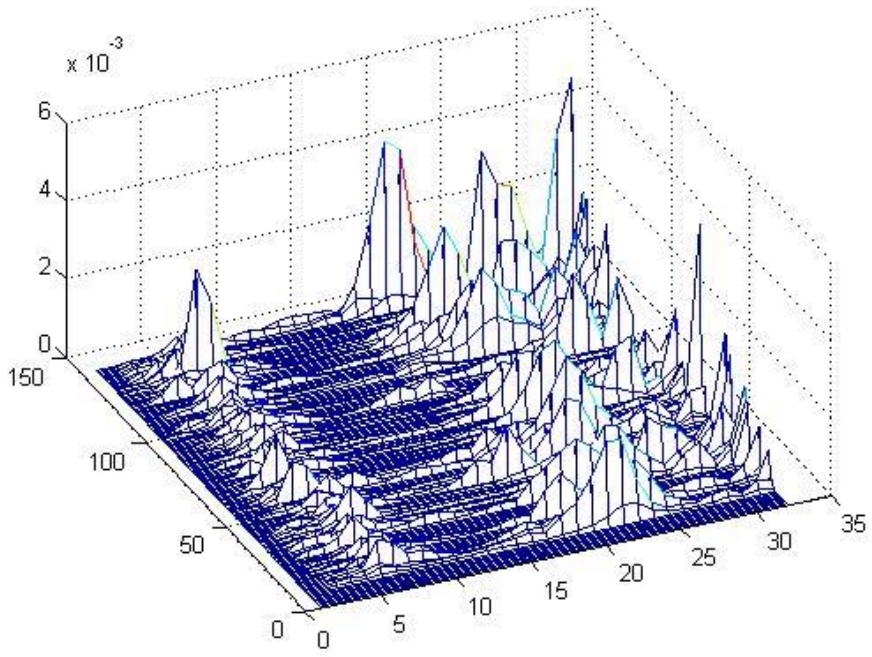


Figura 4.19. Gráfica de energía en bandas críticas para el dialecto Maya

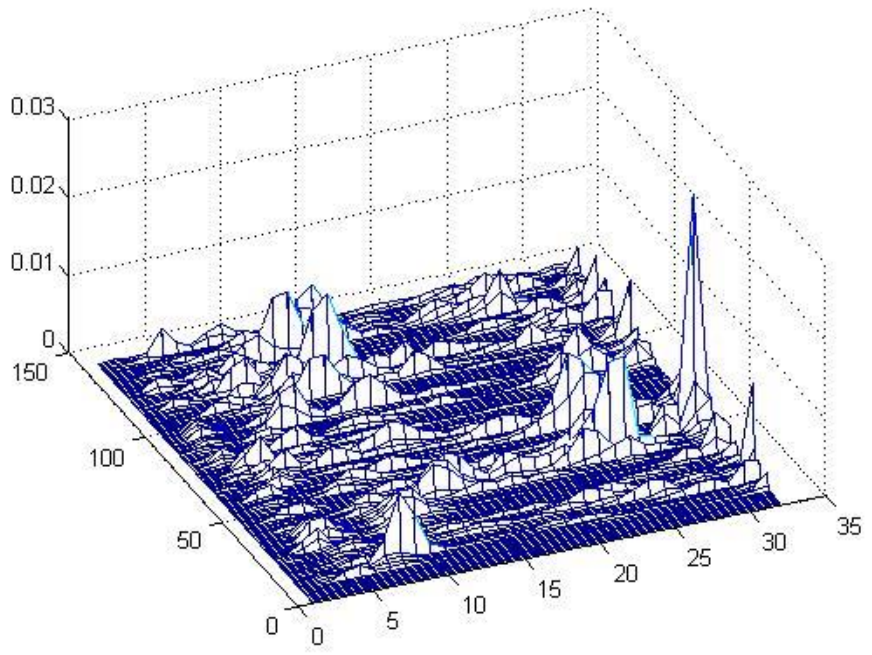


Figura 4.20. Gráfica de energía en bandas críticas para el dialecto Ñañu

4.9 Entrenamiento del sistema y creación de base de datos

Con las características anteriores (Mfcc, energía en bandas críticas y pitch) obtenidas de cada uno de los audios de muestra que tenemos, procedemos a armar una base de datos, la cual servirá como marco de referencia para comparar con la misma información extraída de un audio desconocido y de esta manera poder determinar el dialecto desconocido.

Esto se analizará con más detalle en el siguiente capítulo.

CAPÍTULO 5

EVALUACIÓN DEL SISTEMA

5.1 Extracción de características del audio a evaluar

Una vez que tenemos la base de datos con las características de los audios de cada dialecto que si conocemos, procedemos a extraer las mismas características en el audio a evaluar (coeficientes MFCC, pitch y energia en bandas críticas).

5.2 Cálculo de las distancias para el reconocimiento del dialecto

Para hacer la comparación de los vectores obtenidos de nuestro audio desconocido con los audios conocidos, nos valdremos de dos métodos conocidos:

-Distancia euclidiana

-Distancia Dynamic Time Warping (DTW).

5.2.1 Cálculo de distancia euclidiana

El significado más común en la vida cotidiana de la palabra distancia es el de lejanía. Por ejemplo, la distancia de México a Acapulco es de 400 km. La lejanía es determinada por un número que tiene unidades, kilómetros en el ejemplo anterior.

La palabra ha sido utiliza coloquialmente para indicar la diferencia notable entre unas cosas y otras. Desde el punto de vista matemático, la distancia euclidiana o euclídea es la distancia "ordinaria" (que se mediría con una regla) entre dos puntos de un espacio euclideo, la cual se deduce a partir del teorema de Pitágoras.

Por ejemplo, en un espacio bidimensional, la distancia euclidiana entre dos puntos P1 y P2, de coordenadas cartesianas (x1, y1) y (x2, y2) respectivamente, es:

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

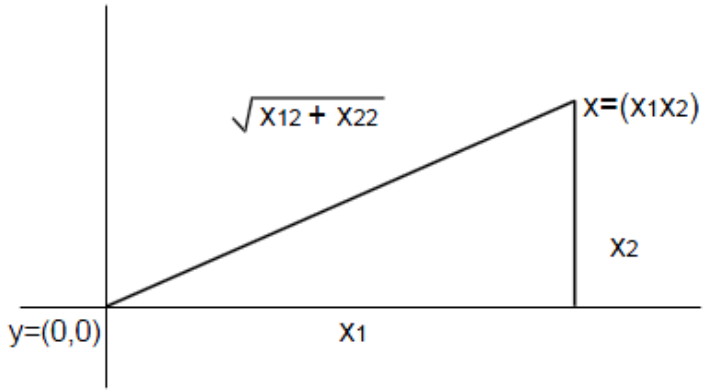


Figura 5.1.Modelo distancia Euclidiana

5.2.2 Cálculo de distancia DTW

El Alineamiento Temporal Dinámico (Dynamic Time Warping, DTW), es una técnica surgida de la problemática inherente a diferentes realizaciones de una misma locución, en las que se observa una variabilidad interna en la duración de los grupos fónicos que la forman, de modo que no existe una sincronización temporal (alineamiento temporal). Además, esta falta de alineamiento no obedece a una ley fija (p. e., un retardo constante), sino que se da de forma heterogénea, produciéndose así variaciones localizadas que aumentan o disminuyen la duración del tramo de análisis.

La problemática asociada hace referencia a la dificultad añadida en el proceso de medida de distancia entre patrones, puesto que se estarán comparando tramos que pueden corresponder a unidades fónicas distintas. Será necesario alinear temporalmente la locución para proceder a realizar una medida de distancia entre patrones cuyo nuevo eje temporal haya homogeneizado las variaciones iniciales. La figura siguiente muestra dos realizaciones de la misma locución (contorno de energía localizada) antes y después de ser alineadas:

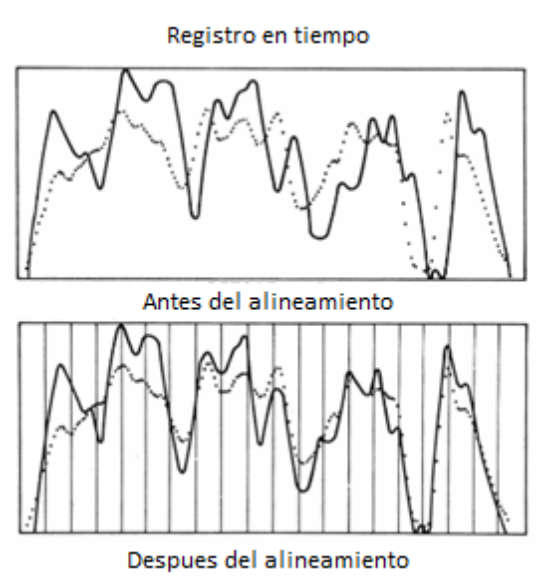


Figura 5.2. DTW

5.3 Comparación de distancias y clasificación.

Para el cálculo de la distancia entre las caracteristicas MFCC y energía en bandas criticas, se utilizo la distancia euclidiana y para el pitch se utilizó la distancia DTW

Para el primer caso se evaluaron los coeficientes MFCC, como podemos notar en las gráficas a) es nuestro archivo desconocido, b) nuestro archivo que más se parece y c) nuestro archivo que menos se parece.

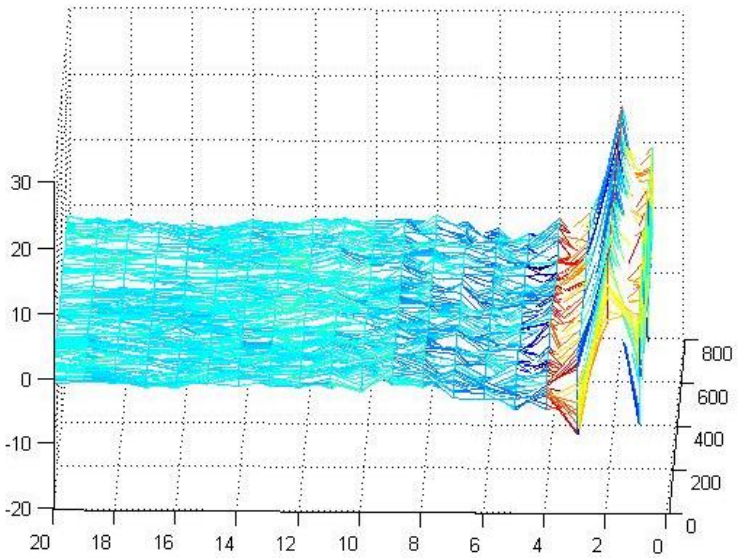


Figura 5.3 a) Archivo desconocido

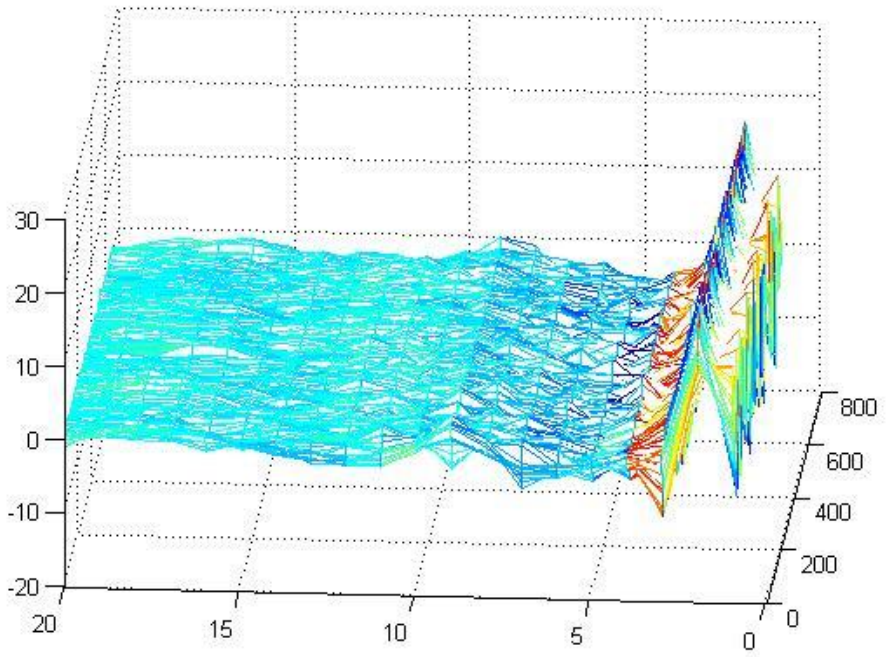


Figura 5.4. b) Archivo que más se parece al desconocido (nosotros sabemos que este archivo es el número 9 y es Náhuatl)

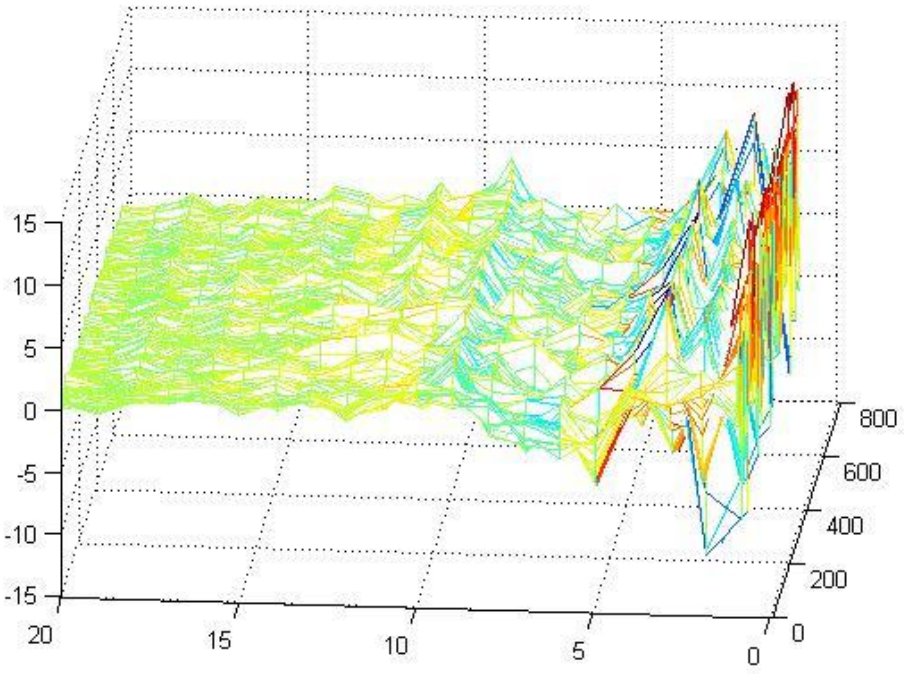


Figura 5.5 c) Archivo que menos se parece al desconocido (nosotros sabemos que este archivo es el 32 y es Maya)

Por lo tanto, para este caso con MFCCs, podemos concluir que nuestro archivo desconocido muy probablemente sea del dialecto Náhuatl.

Ahora para el caso de Pitch, comparamos el archivo desconocido (azul) con el archivo conocido (rojo) y como podemos ver, para el caso a) es donde más se parecen las gráficas (distancia más chica) y para el caso b) es donde menos se parecen las gráficas (distancia más grande)

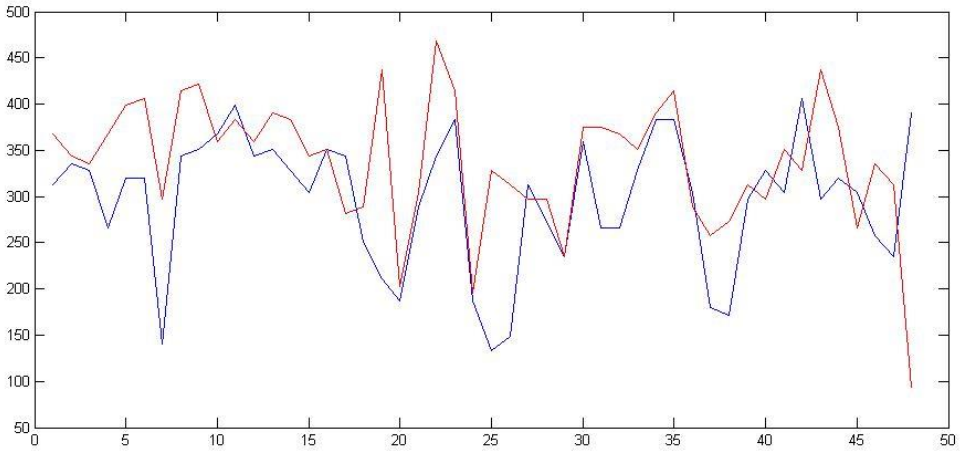


Figura 5.6 a) Archivo que más se parece al desconocido (nosotros sabemos que este archivo es el 9 y es Náhuatl)

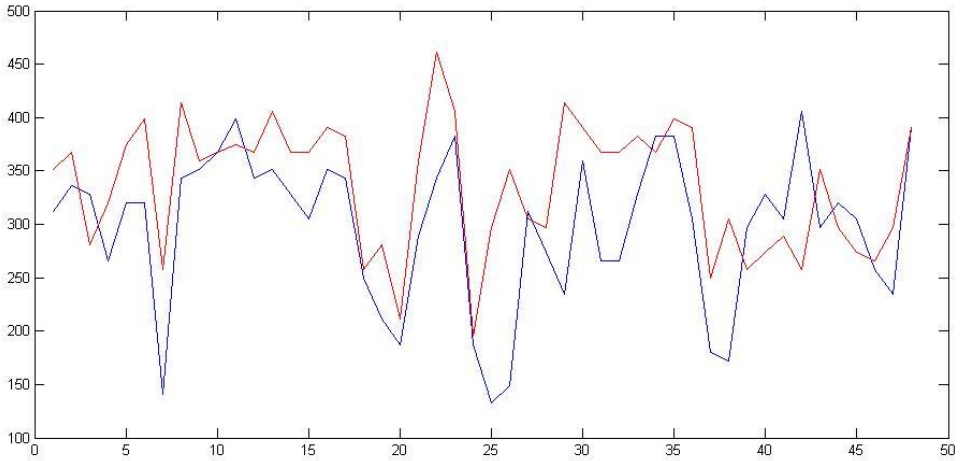


Figura 5.7 b) Archivo que menos se parece al desconocido (nosotros sabemos que este archivo es el 31 y es Maya)

Por lo tanto, para este caso con Pitch, podemos concluir que nuestro archivo desconocido muy probablemente sea del dialecto Náhuatl.

Para el caso de energía en bandas críticas podemos notar en las gráficas a) es nuestro archivo desconocido, b) nuestro archivo que más se parece y c) nuestro archivo que menos se parece.

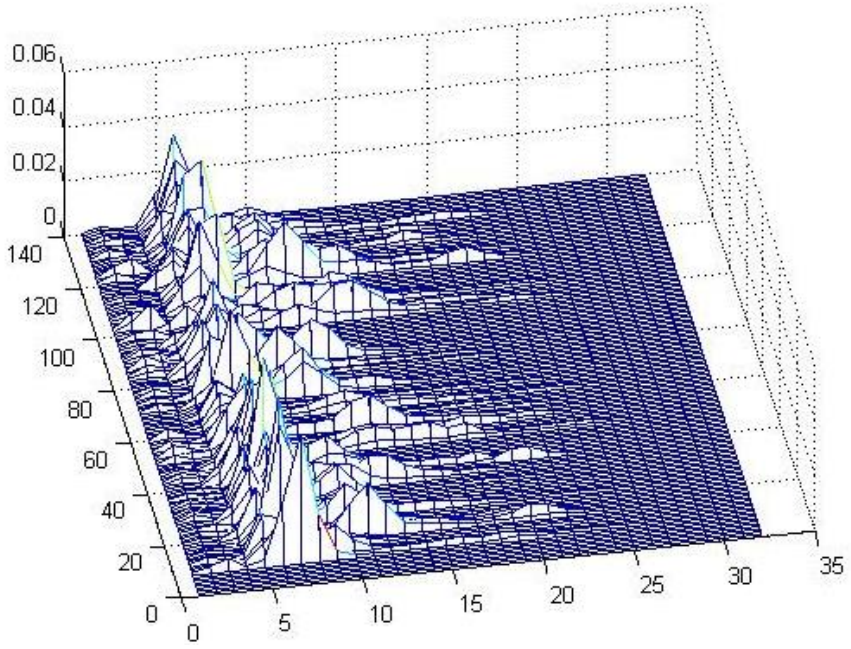


Figura 5.8 a) Archivo desconocido.

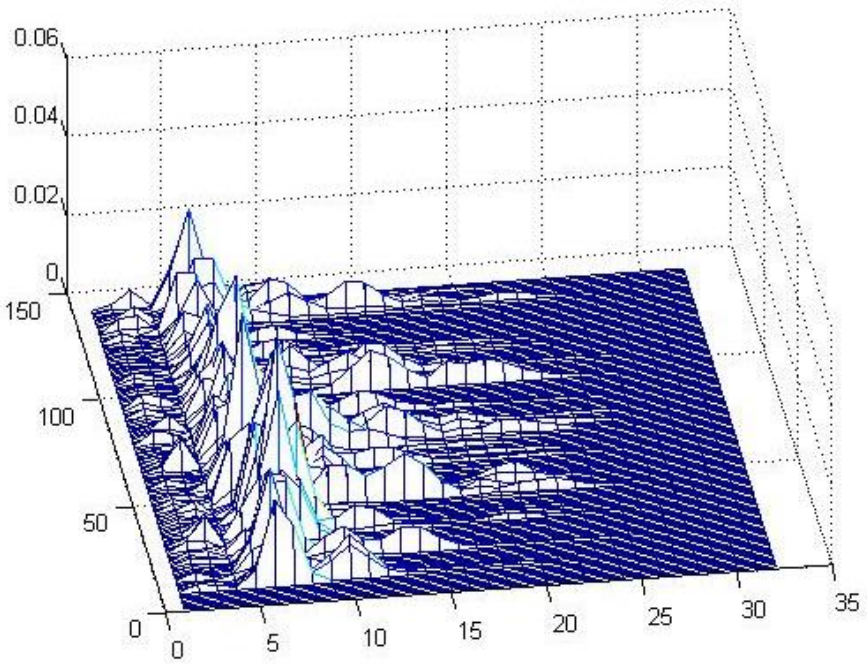


Figura 5.9 b) Archivo que más se parece al desconocido (nosotros sabemos que este archivo es el 16 y es Náhuatl)

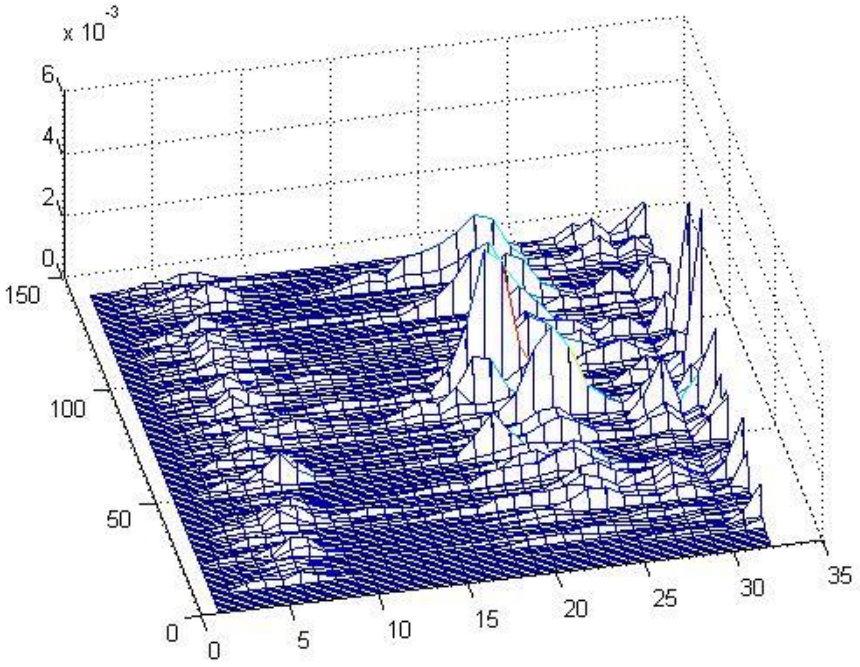


Figura 5.10 c) Archivo que menos se parece al desconocido (nosotros sabemos que este archivo es el 31 y es Maya)

Por lo tanto, para este caso con energía en bandas críticas, podemos concluir que nuestro archivo desconocido muy probablemente sea del dialecto Náhuatl.

Para este caso en especial, las 3 características del dialecto nos arrojan que el archivo desconocido es Náhuatl.

5.4 Resultados generales

Haciendo diversas pruebas y con archivos desconocidos comparándolos con la base de datos pudimos obtener, en promedio, el 81% de certeza en la clasificación de los dialectos.

MFCC	Nahuatl	Mixteco	Maya	Ñañu	total audios por dialecto	**% aciertos**
Náhuatl	**15**	0	0	2	17	**88.24**
Mixteco	2	**7**	1	0	10	**70.00**
Maya	1	1	**7**	0	9	**77.78**
Ñañu	0	0	0	**11**	11	**100.00**

Tabla 5.1 Resultado obtenidos para los coeficientes MFCC

Energia	Nahuatl	Mixteco	Maya	Ñañu	total audios por dialecto	% aciertos
Náhuatl	**15**	0	0	2	17	**88.24**
Mixteco	1	**7**	0	2	10	**70.00**
Maya	1	1	**6**	1	9	**66.67**
Ñañu	0	0	0	**11**	11	**100.00**

Tabla 5.2 Resultado obtenidos para los coeficientes MFCC

Pitch	Nahuatl	Mixteco	Maya	Ñañu	total audios por dialecto	% aciertos
Náhuatl	**13**	0	3	1	17	**76.47**
Mixteco	0	**7**	0	3	10	**70.00**
Maya	2	1	**4**	2	9	**44.44**
Ñañu	0	1	0	**10**	11	**90.91**

Tabla 5.3 Resultado obtenidos para los coeficientes MFCC

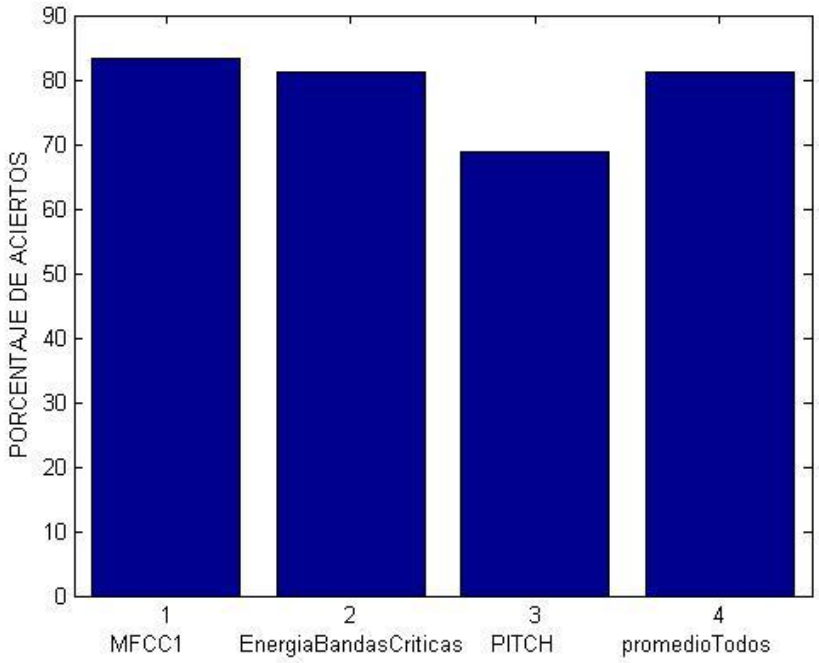


Figura 5.11. Gráfica que muestra los porcentajes totales de acierto para cada característica del dialecto

A manera de resumen, las siguientes tablas muestran un concentrado de los resultados obtenidos por los trabajos más significativos usando los diferentes enfoques: con reconocimiento fonético y sin él.

Autor /año	**Tipo de procesamiento acústico**	**Tiempo de muestra (segundo)**	**Cantidad de idiomas a verificar/ identificar**	**Recursos lingüísticos**		**Método de clasificación**	**Corpus**	**Resultados**	
				Reconocimiento fonético	Modelos de lenguaje			Exactitud en reconocimiento	Error en verificación (EER)
Casseiro 2000	12 DMFCC	10	6 lenguajes	Sólo uno (para el portugués)	Uno para cada lenguaje a reconocer	Interpolación de probabilidades de los bi-gramas	SpeechDat corpus	79.6%	
Torres/ Singer 2002	SDC Delta-cepstral coeficientes	45	12 lenguajes	tokens en lugar de fonemas	Uno para cada lenguaje a reconocer	GMM-Tokenizacion	OGI_TS		6.7%

Tabla 5.4 Comparativo de métodos para la identificación del lenguaje hablado con reconocimiento fonético.

Autor /año	**Tipo de procesamiento acústico**	**Tiempo de muestra (segundos)**	**Cantidad de idiomas a reconocer**	**Tipo de Clasificación**	**Método de clasificación**	**Tamaño de Corpus**	**Porcentaje de identificación**
Samouelian 1998	12 MFCC 12 DMFCC 1 DEnergía	45 seg entrenamiento	3 lenguajes	Multiclase	árbol de decisión C4.5	OGI_TS 50 hablantes c/idioma	45s: 53% 10s: 48.6%
Cummins 1999	Delta F0 DEnergía	50	5 lenguajes	Binaria	Red neuronal back-propagation LSTM	OGI_TS 50 hablantes c/idioma	*Ver tabla 5.6*
Rouas 2003 y 2005	Intervalos de vocales y consonantes	45 seg 10 seg	10 lenguajes	Binaria	GMM-modelo prosódico	OGI_TS	*Ver tabla 5.7*

Tabla 5.5 Comparativo de métodos para la identificación del lenguaje hablado sin reconocimiento fonético.

	Alemán	Español	Japonés	Mandarín
Inglés	52	62	57	58
Alemán	-	51	58	65
Español	-	-	66	47
Japonés	-	-	-	60

Tabla 5.6 Porcentajes de discriminación obtenido por Cummins et al [25].

	Alemán	Español	Mandarín	Vietnamita	Japonés	Coreano	Tamil	Farsi
Inglés	60	68	75	68	68	79	77	76
Alemán	_	59	62	66	66	71	70	72
Español	_	_	81	62	63	76	65	67
Mandarín	_	_	_	50	51	74	74	76
Vietnamita	_	_	_	_	69	56	71	67
Japonés	_	_	_	_	_	66	59	67
Coreano	_	_	_	_	_	_	62	75
Tamil	_	_	_	_	_	_	_	70

Tabla 5.7 Porcentajes de discriminación obtenido por Rouas et al [26].

Comparando nuestros mejores resultados (Figura 5.11, porcentajes arriba del 80% de identificación) con los resultados obtenidos por Cummins [25](tabla 5.6) y Rouas [26] (tabla 5.7) podemos decir que superan en gran medida por los obtenidos por nosotros usando los MFCC y la energía en bandas críticas.

Por otro lado, comparado con los resultados obtenidos por Torres-Carrasquillo, evaluado en el NIST (tabla 5.4) con un margen de error del 6.7%, estamos todavía un poco lejos de tal porcentaje de aciertos, sin embargo, si comparamos el costo que tiene el identificador fonético, nosotros estamos muy por debajo y nuestro sistema es mucho más sencillo de implementar así como de agregar algún nuevo dialecto/idioma que queramos en cualquier momento.

CAPÍTULO 6

CONCLUSIONES GENERALES Y TRABAJO FUTURO

6.1 Conclusiones

En este trabajo de investigación, se propusieron 3 métodos de extracción de características específicos para la identificación del lenguaje hablado, sin utilizar la representación fonética de la señal de voz, los dos están basados en las características suprasegmentales, distintivas entre los idiomas.

La idea central del *primer método* fue la *inclusión de información suprasegmental*, obteniendo la caracterización basándonos en la existencia de los elementos suprasegmentales de los fonemas, como la prosodia, la entonación y la duración. Para ello nos basamos en el procesamiento de la señal de voz por medio de la transformada de Fourier, específicamente los coeficientes cepstrales de frecuencia Mel (MFCC). Todos los trabajos anteriores han utilizado sólo la frecuencia fundamental F0. Nosotros propusimos el uso de los cepstrales MFCC, con 20 coeficientes, capturando además de la frecuencia fundamental F0, frecuencias secundarias que pueden ser importantes en la tarea de discriminar lenguajes. Los resultados son comparables con el estado del arte en sistemas que no utilizan representación fonética.

Otra observación que tenemos es que con muestras pequeñas de señal de voz (1 minuto aproximadamente y reducido a 10 segundos) se obtuvieron buenos resultados. Por lo tanto, este es un punto a discutirse, ya que con una muestra de voz más grande, pensamos que se deben de obtener mejores resultados, pero implicaría

mayor costo computacional debido a que se requiere más capacidad para correr el algoritmo.

Por otro lado, la frecuencia fundamental es la más baja de todas y fue el parámetro más utilizado por los métodos del estado del arte para representar la prosodia. Por lo tanto podemos asumir que en las frecuencias bajas hay información relevante para la identificación del lenguaje hablado; las cuales podrían representar al ritmo, la entonación y la duración –en general las características suprasegmentales que usamos al hablar –. Por lo que propusimos el uso de la energía en bandas críticas para el procesamiento de la señal de voz. Haciendo una separación entre las altas y bajas frecuencias. Este método es completamente diferente a los usados anteriormente basados en la transformada de Fourier. Y con el cual se obtuvieron grandes resultados ya que cada dialecto mostró tener una concentración de energía diferente en cada rango de frecuencias.

En el caso el caso del pitch, los resultados fueron satisfactorios, sin embargo, falta mucho trabajo por realizar en este esquema, debido a que la voz humana es bastante compleja y cada individuo tiene un pitch diferente lo que hace complejo poder tener una caracterización propia para cada dialecto basándonos en esta característica.

Los métodos anteriormente mencionados permiten el tratamiento de cualquier lenguaje, incluyendo los lenguajes marginados, ya que dichos métodos no dependen de ningún tipo de información lingüística. Lo que nos abre las puertas hacia un método de identificación automática para las lenguas indígenas de México.

Con estos resultados (arriba del 80%) nos estamos acercando a los porcentajes de discriminación que obtienen los sistemas que utilizan representación fonética, que como hemos dicho anteriormente son los que mejores resultados han obtenido hasta ahora (arriba del 90%), pero desgraciadamente, con altos costos todavía y no son tan fáciles de implementar y mucho menos agregar, cuando se desee, una nueva lengua.

6.2 Trabajo futuro

Como trabajo futuro, se propone trabajar incrementar el tamaño de las muestras de voz, debido a que con las nuevas tecnologías que van emergiendo día con día, se puede tener un mayor poder computacional que nos permita una mejora significativa. Por el lado de los clasificadores, sería interesante utilizar los modelos de mezclas gaussianas (GMM), muy usadas en la identificación del locutor. Aplicadas sobre los dos métodos de caracterización propuestos en este trabajo. Así como, el uso de otro tipo de clasificadores que nos permitiría tener una mejoría en la verificación del idioma, como por ejemplo cadenas de Markov.

Otro trabajo sería la aplicación del método usando la transformada wavelet para la identificación de acentos regionales. Ya que esta transformada ofrece mejor resolución en frecuencias bajas, las cuales poseen la mayoría de la información en el lenguaje hablado. Un primer experimento podría realizarse con el corpus TIMIT, el cual divide al inglés americano en cinco regiones. Posteriormente, dependiendo de los resultados alcanzados se podría intentar construir un corpus para el español e incluso para las lenguas indígenas de México las cuales tienen muchas variantes como el Mixteco o el Maya.

También sería interesante comprobar el alcance de los métodos para la tarea de identificación del habla no nativa, con un corpus en donde la persona no hable su lengua materna, por ejemplo, una persona cuya lengua materna sea el español, realice una grabación hablando inglés, y a partir de esas muestras de señal de voz identificar el idioma hablado. Recordemos lo dicho por Abercombrie [28]: el “ritmo” es probablemente el rasgo de la *base articulatoria* de una lengua cuya adquisición o dominio resulta más difícil al estudiante adulto de un idioma extranjero y, aunque la inteligibilidad depende en gran parte de su correcta emisión, a éste no se le presta la atención debida en la enseñanza de idiomas extranjeros. Esto es, cuando un adulto aprende un idioma extranjero, es necesario invertir un enorme esfuerzo para adquirir el ritmo de la lengua. Aunque pronunciemos correctamente una palabra, las pautas rítmicas que emitimos al pronunciarla no corresponden al idioma. El problema para este tipo de pruebas, es que no existen corpus para ello. Por lo que un trabajo futuro sería construir ese tipo de corpus.

REFERENCIAS

[1] MPEG-7 audio and beyond, audio content indexing and retrieval. Hyoung-Gook Kim, Nicolas Moreau, Thomas Sikora.

[2] Spoken Language Recognition: From Fundamentals to Practice, By Haizhou Li, Senior Member IEEE, Bin Ma, Senior Member IEEE, and Kong Aik Lee, Member IEEE

[3] Un método para la identificación automática del lenguaje hablado basado en características suprasegmentales. Tesis Doctoral, por Ana Lilia Reyes Herrera.

[4] Reconocimiento de Palabras Aisladas habladas en español usando la unión de los Métodos MFCC y MODGDF en la Extracción de Características por Jorge C. Valverde-Rebaza, Sarah D. Amaro-Calderón. Universidad Nacional de Trujillo, Sociedad de Estudiantes de Ciencias de la Computación (SECC)

[5] Stein J.Y. (2000): *Digital Signal Processing: A computer Science Perspective*. Wiley Series in Telecommunications and Signal Processing. John G. Proakis, Series Editor. A Wiley-Interscience Publication. ISBN 0-471-29546-9.

[6] Lenguas de México http://es.wikipedia.org/wiki/Lenguas_de_México.

[7] Lenguas indígenas de México, en el sitio en internet de la Comisión Nacional para el Desarrollo de los Pueblos Indígenas, consultada el 10 de enero de 2007.

[8] Archivo de los idiomas indígenas de América latina. Colección de audios. http://www.aiila.utexas.org/site/welcome_sp.html

[9] Acervo digital de lenguas indígenas. http://lenguasindigenas.mx/nahuatl.html

[10] Audio Signal Processing and Recognition
http://neural.cs.nthu.edu.tw/jang/books/audiosignalprocessing/index.asp

[11] Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh University Press.

[12] La psicoacústica de la codificación,
http://www.lpi.tel.uva.es/~nacho/docencia/ing_ond_1/trabajos_04_05/io1/public_html/MP3.htm

[13] Bandas Criticas, http://www.eumus.edu.uy/docentes/maggiolo/acuapu/bcr.html

[14] Tesis Maestria 2011, Implementación de descriptores de música para su indexación bajo la norma MPEG-7, Martínez Cortés Alejandro

[15] Tesis Maestria 2012, Diseño de un sistema de identificación de canciones por tarareo, Salgado Estrada Martin.

[16] Support Vector Machines for Speaker and language Recognition, W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, P. A. Torres-Carrasquillo ,MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA 02420.

[17] Quilis, A. (1992). Tratado de fonología y fonética españolas. Madrid: Gredos.

[18] Etimología de consonantes http://etimologias.dechile.net/?consonante

[19] Cummins, F. (2002): "Classifying languages based on speech rhythm". In Artificial Intelligence and Cognitive Science: Proceedings of the 13th Irish International Conference (AICS 2002), volume 2464 of Lecture Notes in Computer Science. Springer Verlag.

[20] Nazzi T., Bertoncini J. and Mehler J. (1998):"Language discrimination by newborns; towards an understanding of the role of rhythm". Journal of Experimental Psychology: Human Perception and Performance, 24:756-766. APA (American Psychological Association).

[21] Ramus F., Nespor M., Mehler J., (1999):"Correlates of linguistic rhythm in the speech signal". Cognition, 73(3), pp. 265-293. Elsevier.

[22] Muthusamy Y.K., Jain N., Cole R., (1994): "Perceptual benchmarks for automatic language identification", in Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'94), Adelaide, Australia. Vol. 1.

[23] Ancho de banda crítico, Mayo 2011, http://www.labc.usb.ve/paginas/EC4514/AUDIO/PSICOACUSTICA/BANDAS_CRITICAS.html

[24] Implementación de FFT, STFT, filtro de pre énfasis y visualización del espectrograma en señales del habla, Fredy Carranza-Athó ,Universidad Nacional de Trujillo

[25] Cummins F., Gers F., Schmidhuber J., (1999): "Language Identification from Prosody without explicit Features", Proc. EUROSPEECH'99, Budapest, Hungary, 1, pp. 371-374.

[26] Rouas J-L., Farinas J., Pellegrino F., André-Obrecht R., (2003): "Modeling prosody for language identification on read and spontaneous speech" in Proc. IEEE ICASSP 2003, vol 1, pp. 40-43.

[27] Language Recognition Evaliation, National Institute of Standars and Technology. http://www.itl.nist.gov/iad/mig//tests/lang/

[28] Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh University Press

Anexo

Paso 1. Fonemas

```
clc; clear all;
%
fileList = readsfxfiles('nahuatl.txt'); %lee lista de archivos de audio
de la bd
%aqui calculamos el pitch de cada archivo5

for j=1:length(fileList);
    j
    disp('Procesando archivo....')
carpeta=fileList(j,1);
SOUNDDIR=char(strcat(carpeta, filesep, carpeta, '.wav'))     %carga el
archivo de audio
carp=char(strcat(carpeta, filesep));

 %% normaliza en audio
[musica,Fs,nb]=wavread(SOUNDDIR,448000); %1 min de audio a fs=8000
=480000 %8000*58=464000 muestras en 58seg
vec=musica';
%%%filtro de preemfasis...
a=0.95;
vecS = filter([1, -a], 1, vec);
%# obtenemos max y min
maxVec = max(vecS);
minVec = min(vecS);
%# normalizamos a -1...1
audio = ((vecS-minVec)./(maxVec-minVec) - 0.5 ) *2;

%%guardamos el archivo...
wavwrite(audio,Fs,nb,char(strcat(carpeta,filesep, carpeta,'n.wav')));    %
audio normalizado
clear musica audio %se limpian para que no hay problemas al le

disp('siguiente archivo....')
 end

%despues de crear el audio normalizado , se crea el archivo comprimido
%eliminando silencios

for j=1:length(fileList);
    disp('Procesando archivo....')
carpeta=fileList(j,1);
SOUNDDIR=char(strcat(carpeta, filesep, carpeta, 'n.wav')) ;    %carga el
archivo de audio que esta normalizado
carp=char(strcat(carpeta, filesep));

 [track,Fs,nb]=wavread(SOUNDDIR);
audio=track;%(1*Fs:15*Fs);
```

Paso 2

```
%extraccion descriptores
clc; clear all;
fileList = readsfxfiles('nahuatl.txt'); %lee lista de archivos de audio
de la bd

for j=1:length(fileList);
    disp('Procesando archivo....')
carpeta=fileList(j,1);
SOUNDDIR=char(strcat(carpeta, filesep, carpeta, 'n.wav'))     %carga el
archivo de audio
carp=char(strcat(carpeta, filesep));

que_se_va_hacer=1;

if que_se_va_hacer==1
     extractorMfcc(SOUNDDIR,carp,char(carpeta),3);
 elseif que_se_va_hacer==2
     extractorBandasCriticas(SOUNDDIR,carp,char(carpeta),3);
 elseif que_se_va_hacer==3
     pitch(SOUNDDIR,carp,char(carpeta),3);
 end
 disp('siguiente archivo....')
 end
```

Paso 3

```
clc; clear all;
fileList = readsfxfiles('nahuatl.txt'); %lee lista de archivos de audio
de la bd
%% aqui se cargan el codebook
%
for j=1:length(fileList)
    j
    disp('cargando siguiente archivo...');
filename=fileList(j,1);
%
filexlm='MFCC1.xml';
archivo=char(strcat(dir,filesep,filename,filesep,filexlm));
descriptorxmlbase=xmlread(fullfile(archivo));
xRootbase=descriptorxmlbase.getDocumentElement;
theStructbase = parseChildNodesBenja(xRootbase);
base.MFCC1{1,j}=str2num(theStructbase(1,1).Data);

%{
filexlm='EnergiaBandasCriticas.xml';
archivo=char(strcat(dir,filesep,filename,filesep,filexlm));
descriptorxmlbase=xmlread(fullfile(archivo));
xRootbase=descriptorxmlbase.getDocumentElement;
theStructbase = parseChildNodesBenja(xRootbase);
base.EnergiaBandasCriticas{1,j}=str2num(theStructbase(1,1).Data);

clear descriptorxmlbase xRootbase theStructbase
```

```
%
filexlm='PitchWavelet.xml';
archivo=char(strcat(dir,filesep,filename,filesep,filexlm));
descriptorxmlbase=xmlread(fullfile(archivo));
xRootbase=descriptorxmlbase.getDocumentElement;
theStructbase = parseChildNodesBenja(xRootbase);
base.PitchWavelet{1,j}=str2num(theStructbase(1,1).Data);

clear descriptorxmlbase xRootbase theStructbase
%}
%%
end

save('codebookMFCCyBandasCriticas', 'base')
disp('ARCHIVOS GUARDADOS CORRECTAMENTE.....');

paso 3.2
%}
%% aqui se hace el analisis
load codebookMFCCyBandasCriticas

disp('ARCHIVOS CARGADOS CORRECTAMENTE.....');
%A=7 %archivo que vamos a comparar
%B=7
MinFind=Inf;
%========================
for x=1:length(fileList)
%%%%%%%%%%
distminMFCC1 = inf;
for l = 1:length(base.MFCC1)      % si se compara contra si mismo la
distancia sale cero
        dMFCC1 = disteu(base.MFCC1{1,x}, base.MFCC1{1,l});
distMFCC1(x,l) = sum(min(dMFCC1,[],2)) / size(dMFCC1,1);
        if distMFCC1(x,l)==0
           distMFCC1(x,l)=NaN;
        end

        if distMFCC1(x,l) < distminMFCC1
            distminMFCC1 = distMFCC1(x,l
            MinFind=l;
        end
end
a=1;
RESULTADOS(x,a)=MinFind;
clear  MinFind
%%%%%%%%%%%%%%%%%%%%%%
distminEnergia = inf;
for l = 1:length(base.EnergiaBandasCriticas)
        dEnergia = disteu(base.EnergiaBandasCriticas{1,x}',
base.EnergiaBandasCriticas{1,l}');
        distEnergia(x,l) = sum(min(dEnergia,[],2)) / size(dEnergia,1);
        if distEnergia(x,l)==0;
            distEnergia(x,l)=NaN;
        end

        if distEnergia(x ,l) < distminEnergia
```

```
            distminEnergia = distEnergia(x,l);
            MinFind=l;
        end
end
a=a+1;
RESULTADOS(x,a)=MinFind;
clear  MinFind
%%
%
distminPitch = inf;
for l = 1:length(base.Pitch)      % si se compara contra si mismo la
distancia sale cero
        dPitch = disteu(base.Pitch{1,x}, base.Pitch{1,l});
        distPitch(x,l) = sum(min(dPitch,[],2)) / size(dPitch,1);

         if distPitch(x,l)==0
           distPitch(x,l)=NaN;
         end

        if distPitch(x,l) < distminPitch
            distminPitch = distPitchW(x,l);
            MinFind=l;
        end
end
a=a+1;
RESULTADOS(x,a)=MinFind;
clear  MinFind
%%
end
save('ResultadosBusqueda', 'RESULTADOS');

disp('FIN DE LA BUSQUEDA')
```

Paso 4

```
%% aqui se hace el analisis
clc; clear all;
load ResultadosBusqueda

 disp('ARCHIVOS CARGADOS CORRECTAMENTE.....');

[a b]=size(RESULTADOS);
%[37 x 8]

for x=1:a
    for y=1:b
        MinFind=RESULTADOS(x,y);
        if MinFind<=17
                ANALISIS(x,y)=1;
            elseif MinFind>=18 && MinFind<=27
                ANALISIS(x,y)=2;
            elseif MinFind>=28 && MinFind<=36
                ANALISIS(x,y)=3;
            elseif MinFind>=37 && MinFind<=47
                ANALISIS(x,y)=4;
```

```
            elseif MinFind>47
                ANALISIS(x,y)=5;
        end
    end

end

for dialecto=1:4
for x=1:a

[W Z]=find(ANALISIS(x,1:b)==dialecto);

CORRECTOS(x,dialecto)=sum(W);

end
end

%para obtener el promedio de cada caracteristica...

ANALISIS2=zeros(a,b);
for x=1:a
    for y=1:b

        dialecto=ANALISIS(x,y);
        if x<=17 && dialecto==1
                ANALISIS2(x,y)=1;
        elseif x>=18 && x<=27 && dialecto==2
                ANALISIS2(x,y)=1;
        elseif x>=28 && x<=36 && dialecto==3
                ANALISIS2(x,y)=1;
        elseif x>=37 && x<=47 && dialecto==4
                ANALISIS2(x,y)=1;
        else
                ANALISIS2(x,y)=0;
        end
    end

end
[a b]=size(ANALISIS2);
PROMEDIO=sum(ANALISIS2)/a * 100

bar(PROMEDIO);
xlabel('MFCC1    Energia   PITCH')
ylabel('PORCENTAJE DE ACIERTOS')
```