UNIVERSIDAD POLITECNICA DE VALENCIA

ESCUELA POLITECNICA SUPERIOR DE GANDIA

I.T. TELECOMUNICACIÓNES (IMAGEN Y SONIDO)

UNIVERSIDAD
POLITECNICA
DE VALENCIA

"Implementación de algoritmos para la extracción de patrones característicos en Sistemas de Reconocimiento De Voz en Matlab"

TRABAJO FINAL DE CARRERA

Autor:
David Reig Albiñana
Tutor/a:
María Asunción Pérez Pascual
GANDIA, 2014

Agradecimientos: *Agradezco a todos los profesores de la ESPG que me han ayudado a crecer y evolucionar especialmente a los que se sientan aludidos. Gracias a mi familia por creer en mi.*

Índice General:

1. Objetivos y antecedentes

El objetivo del presente trabajo es la revisión bibliográfica de algoritmos para la extracción de patrones característicos usados para la implementación de un Sistema de Reconocimiento de Voz de palabras aisladas que sea independiente del tipo de voz del hablante, y que permita efectuar un procesamiento On-Line, con el fin de que sirva como herramienta de desarrollo en estudios posteriores sobre Procesamiento de Voz.

Se ha incrementado de manera significativa el desarrollo de trabajos de investigación en este campo, tanto en las Universidades como en los Institutos especializados alrededor del mundo, los cuales aportan continuamente nuevos conocimientos sobre la materia.

Existe mucha más información sobre este tema en artículos y en medios como Internet. Se puede tener acceso a paquetes de software que constituyen herramientas de Análisis y Reconocimiento, pero sin embargo son consideradas "Cajas Negras", ya que no ofrecen mayor información sobre las técnicas y algoritmos utilizados en su construcción.

Empecé documentándome sobre la tecnología del habla, para esto consulté sitios especializados en Internet y artículos científico-técnicos recientes, lo cual me permitió obtener una visión general de las últimas técnicas utilizadas en el Procesamiento y Reconocimiento de Voz.

Se pensó implementar los algoritmos en un DSP y aprovechar así las características de este hardware, pero debido a la extensión en el estudio previo, análisis y prueba de los algoritmos Off-line para fijar variables para su funcionamiento y entender el funcionamiento de estos, se decidió hacer solo una revisión de estos algoritmos en matlab.

Primero se realizó un análisis Off-Line de la Señal de Voz y sus características. La adquisición muestras de voz en se realizó por medio de la Grabadora de Sonidos de Windows y el análisis e implementación se realizó utilizando el programa Matlab.

2. Fundamentos de la señal de voz

Una onda sonora es una onda longitudinal que transmite lo que se asocia con sonido. Si se propaga en un medio elástico y continuo genera una variación local de presión o densidad, que se transmite en forma de onda esférica periódica o cuasiperiódica. Mecánicamente las ondas sonoras son un tipo de onda elástica.

Existe un gran margen de frecuencias entre las cuales se puede generar ondas mecánicas longitudinales. Las ondas sonoras se reducen a los límites de frecuencia que pueden estimular el oído humano para ser percibidas en el cerebro como una sensación acústica. Estos límites de frecuencia se extienden de aproximadamente 20 Hz a cerca 20 KHz y se llaman límites de audición. Las ondas audibles son producidas por cuerdas en vibración, por columnas de aire en vibración.

En este apartado haremos un pequeño estudio de las ondas sonoras generadas por el sistema fonador humano, las cuales se denominan señales de voz.

En la primera parte se describe el proceso de generación de la Señal de Voz, seguidamente se describe sus propiedades en el dominio del tiempo y de la frecuencia, así como sus principales tipos. Finalmente se describe el modelo matemático del tracto vocal y los factores que afectan un Sistema de Procesamiento de Voz.

2.1 Descripción del Aparato Fonador Humano

El aparato fonador esta formado por un conjunto de órganos con la función producir la voz humana, estos son los pulmones, los cuales producen un flujo de aire; la laringe, que contiene las cuerdas vocales, la faringe, las cavidades oral y nasal y una serie de elementos articulatorios como los labios, los dientes, el alvéolo, el paladar, el velo del paladar y la lengua.

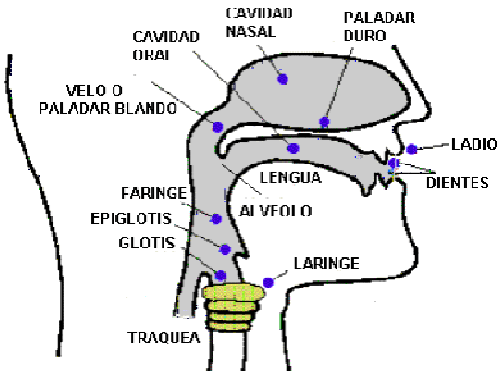


Figura 1: Sección transversal sistema fonador humano.

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

Para convertirse en sonido, el aire procedente de los pulmones debe provocar una vibración, y la laringe es el primer lugar en que se produce.

La tensión, elasticidad, altura, anchura, longitud y grosor de las cuerdas vocales pueden variar, lo que da lugar a diferentes efectos sonoros.

El efecto más importante de las cuerdas vocales es la producción de una vibración audible en los llamados sonidos sonoros, en contraste con los sonidos sordos, en cuya producción no vibran las cuerdas vocales. En español, todas las vocales y muchas consonantes (m, b, d,...) son sonoras.

2.2 Características fundamentales de la Señal de Voz

2.2.1 Forma de onda de la Señal de Voz

La representación de la Señal de Voz en función del tiempo es importante puesto que brinda información sobre características, tales como la Energía y los Cruces por Cero, las cuales facilitan su estudio y análisis.

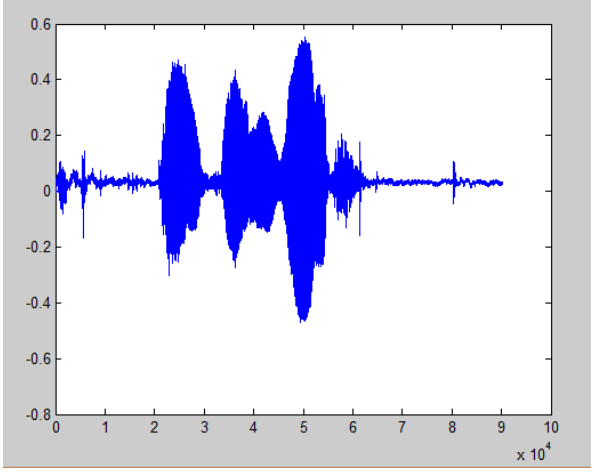


Figura 2: Forma de onda de la palabra 'Encender'

2.2.2 Energía y Cruces por Cero

En una señal continua, la Energía total E en el intervalo de tiempo t_1 a t_2 esta definida como:

$$E = \int_{t_1}^{t_2} |x(t)|^2 dt$$

Para el caso de las señales discretas la Energía se define por:

$$E = \sum_{n=m}^{N-1} x(m)^2$$

Donde: N es el número de muestras de la señal.

La Energía es útil para distinguir segmentos sordos y sonoros en la Señal de Voz, debido a que los valores de esta característica aumentan en los sonidos sonoros respecto a los sordos.

Los Cruces por Cero indican el número de veces que una señal continua toma el valor de cero. Para las señales discretas, un cruce por cero ocurre cuando dos muestras consecutivas difieren de signo, o bien una muestra toma el valor de cero. Consecuentemente, las señales con mayor frecuencia presentan un mayor valor de esta característica, el ruido también genera un gran número de cruces por cero.

La formulación matemática de la Densidad de Cruces por Cero para señales discretas esta representa en la siguiente fórmula, en la cual, sign es la función signo y N es el número de muestras de la señal.

$$z = \sum_{m=0}^{N-1} \left| sign[x(m)] - sign[x(m-1)] \right|$$

Esta fórmula analiza los cambios de signo de la señal y los va acumulando, y si el ruido de fondo es de alta frecuencia acumulará una gran cantidad de éstos. Por esto, existe otro algoritmo que no sólo relaciona los cambios de signo, sino que a la vez proporciona información sobre los cambios de Energía de la señal.

2.2.3 El Algoritmo COPER

Básicamente este algoritmo es similar al de los Cruces por Cero, pero a las funciones signo, se les ha multiplicado por la Energía de la muestra analizada, es así que la densidad acumulada no sólo dependerá del cambio de signo de las muestras sino también de su amplitud, por esto, el ruido de fondo no logrará una gran acumulación de densidad, debido a que posee una pequeña amplitud comparada con las palabras pronunciadas. Su formulación matemática se muestra a continuación:

$$C = \sum_{m=0}^{L} \left| y[m] \cdot \left| y[m] \right| - y[m-1] \cdot \left| y[m-1] \right| \right|$$

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Algoritmo COPER] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

Donde: L es el número de muestras de una trama de la señal.

De este modo solo hay un parámetro que indique si que el segmento bajo análisis es sonoro o sordo.

2.2.4 Espectro de Frecuencia

La figura muestra el espectro de una señal de voz correspondiente a la palabra "Encender".

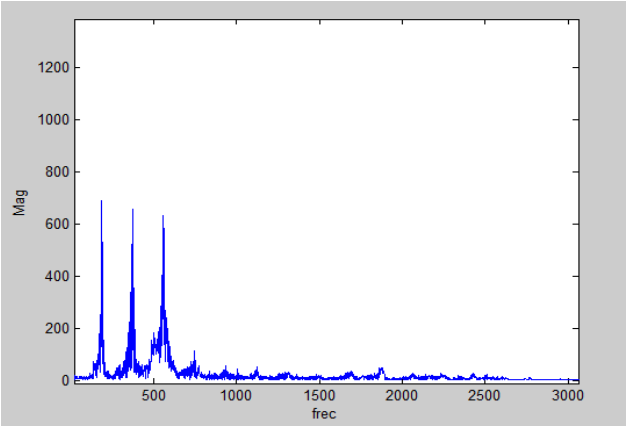


Figura 3: Densidad espectral pot. de la palabra ‘Encender'

La frecuencia fundamental (primer pico situado de izquierda a derecha) nos da información sobre la velocidad a la que vibran las cuerdas vocales al producir un sonido, el cual es generado por la rápida apertura y cierre de las cuerdas vocales.

El espectro está conformado de armónicos de la frecuencia fundamental (múltiplos enteros), el cual es el rango fundamental de frecuencia producidas por las cuerdas vocales.

Otra característica importante es la envolvente espectral. Un análisis adecuado sobre esta característica permite obtener información sobre los diferentes tipos de sonido.

2.2.5 Frecuencias Formantes

Las cavidades que conforman la cavidad supraglótica actúan como resonadores acústicos. Si se realiza un análisis espectral del sonido luego de haber atravesado estas cavidades, el efecto de la resonancia produciría un énfasis en determinadas frecuencias del espectro obtenido, a las que se les denominara Formantes (los tres picos de la figura 3).

Existen tantas Formantes como resonadores posee el tracto vocal. Sin embargo, se considera que sólo las tres primeras, asociadas a la cavidad oral, bucal y nasal respectivamente proporcionan la suficiente cantidad de información para poder diferenciar los distintos tipos de sonido. La amplificación de cada una de estas tres frecuencias depende del tamaño y forma que adopta la cavidad bucal y la cavidad oral, y si el aire pasa o no por la nariz.

2.3 *Tipos de Señales de Voz*

La Señal de Voz puede clasificarse en los siguientes tipos, Sonora, No Sonora y Plosiva.

2.3.1 Señal Sonora

La señal sonora se genera por la vibración de las cuerdas vocales manteniendo la glotis abierta, lo que permite que el aire fluya a través de ella. Estas señales se caracterizan por tener alta Energía y un contenido frecuencial en el rango de los 300 Hz a 4000 Hz presentando cierta periodicidad, es decir son de naturaleza cuasiperiódica. El tracto vocal actúa como una cavidad resonante reforzando la Energía en torno a determinadas frecuencias (formantes).

Todas las vocales se caracterizan por ser sonoras pero existen consonantes que también lo son, tales como, la b, d y la m, entre otras.

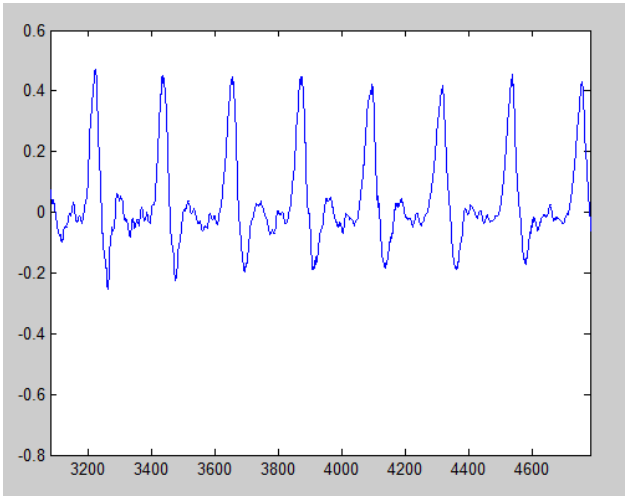


Figura 4: Forma de señal de la vocal ‘e’.

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

2.3.2 Señal No Sonora

También conocida como señal fricativa o sorda. Se caracteriza por tener un comportamiento aleatorio en forma de ruido blanco. Tienen una alta densidad de Cruces por Cero y baja Energía comparadas con las señales de tipo sonora. Durante su producción no se genera vibración de las cuerdas vocales, ya que, el aire atraviesa un estrechamiento, y genera una turbulencia. Las consonantes que producen este tipo sonidos son la 's', la 'f' y la 'z' entre otras.

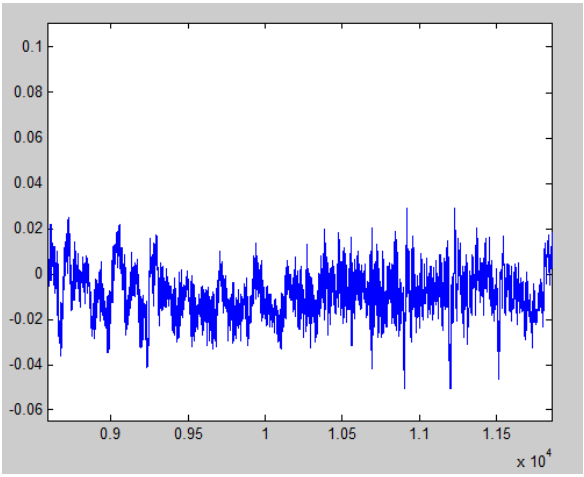


Figura 5: Forma de señal de la consonante ‘z’.

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

2.3.3 Señal Plosiva

Se genera cuando el tracto vocal se cierra en algún punto, lo que causa que el aire se acumule para después salir expulsado repentinamente (explosión). Se caracterizan por que la expulsión de aire está precedida de un silencio. Estos sonidos se generan por ejemplo, cuando se pronuncia la palabra 'campo'. La p es una consonante de carácter plosivo, y existe un silencio entre las sílabas 'cam' y 'po'. Otras consonantes que presentan esta característica son 't', y 'k', entre otras.

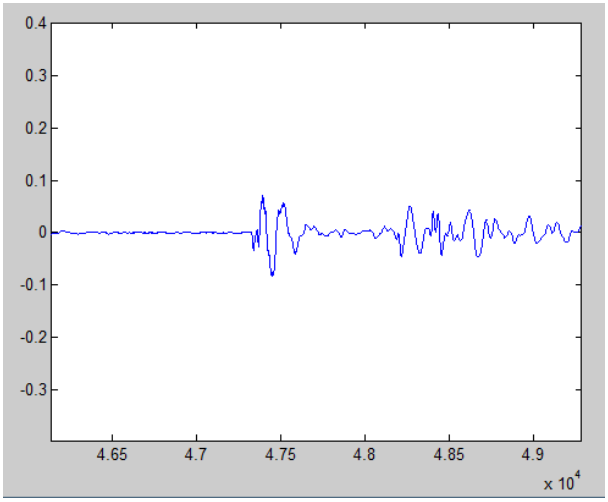


Figura 6: Forma de señal de la silaba '–po'.

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

2.4 Modelo del tracto vocal

El tracto vocal se comporta como un filtro, cuyos parámetros varían en el tiempo en función de la acción consciente que se realiza al pronunciar una palabra.

En la figura 7 se muestra el diagrama de bloques del modelo del tracto vocal. Se consideran dos posibles entradas que dependerán del tipo de señal a reproducir, sonora o no sonora. Para señales sonoras, la excitación será un tren de impulsos de frecuencia controlada, mientras que para las señales no sonoras la excitación será ruido aleatorio. La combinación de estas señales modela el funcionamiento de la glotis. El espectro de frecuencias de la Señal de Voz puede obtenerse a partir del producto del espectro de la excitación por la respuesta en frecuencia del filtro.

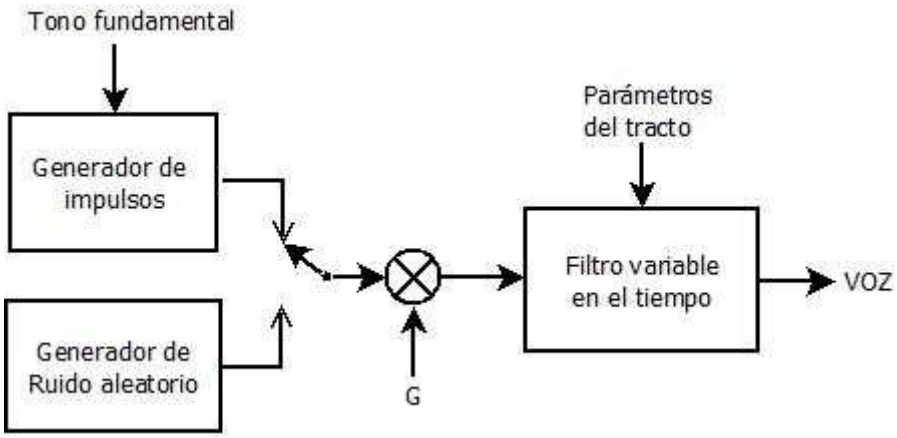


Figura 7: Modelo del tracto vocal.

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

El parámetro de ganancia (tipificado en grafico como G), determina la intensidad de la excitación. El tracto vocal manifiesta un número muy grande de resonancias, aunque sólo se consideran tres y en algunos casos cuatro, esto es debido a que las resonancias de alta frecuencia son atenuadas por la característica frecuencial del tracto que tiende a actuar como un filtro pasabajo. Este modelo es una simplificación del proceso del habla. Los sonidos fricativos,

no se filtran por el tracto con la misma extensión en que lo hacen las señales sonoras, por lo que el modelo no es muy preciso para este tipo de señales. Además, el modelo supone que las dos señales pueden separarse sin considerar ninguna interacción entre ellas, lo que no es del todo cierto, ya que la vibración de las cuerdas vocales es afectada por las ondas de presión dentro del tracto. Sin embargo, estas consideraciones pueden ser ignoradas, resultando el modelo lo suficientemente adecuado.

2.5 Factores que afectan a la Señal de Voz

Existen muchos factores que afectan la correcta percepción de las Señales de Voz, tales como el ruido, la acústica y la calidad del micrófono.

El ruido, se define como aquellos sonidos aleatorios que de forma "oculta" transforman y enmascaran el sonido. Dado que, es poco probable encontrar un entorno de audio digital en perfecto silencio, es importante conocer la cantidad de ruido, en relación con la señal que se introduce en el equipo de sonido, especialmente en la tarjeta de sonido. La fuerza de cualquier sonido comparada con la fuerza promedio del ruido, se conoce como relación señal a ruido (SNR). A medida que aumenta la relación SNR, es mejor el trabajo realizado en grabación.

Los factores que principalmente afectan a la señal de voz son los siguientes:

- Acústica de la habitación: La acústica dentro de una habitación, puede crear cambios en el espectro de la Señal de Voz, debido a las resonancias de la habitación. Puesto que, cualquier ambiente cerrado tendría resonancias inherentes, su énfasis cuando interfiere con una señal de habla puede crear rangos anormales de frecuencias. Debido a esto, se producen dos cambios básicos en la acústica de una habitación, el primero es causado por el retardo en el tiempo del retorno de la señal original de una superficie reflectante, tal como una pared o una ventana. Cuando la onda es reflejada, regresa con mucho menor amplitud, y retardada en el tiempo, ésta interactúa con la forma de onda originalmente hablada para crear un nuevo espectro compuesto del habla. El segundo, está relacionado con la reflexión de una superficie

rugosa de una pared, lo cual tiende a atenuar en altas frecuencias, pero a reforzar en el rango de bajas frecuencias.

- Ruido del ambiente: Si el usuario del sistema está operando el dispositivo en cualquier lugar que no sea una habitación tranquila, existe la posibilidad de la interferencia del ruido con las formas de onda. No obstante sin ruido externo, el sistema es susceptible de captar ruido a través del micrófono, y muchas veces el ruido proviene desde la boca durante la pronunciación del mensaje.

En el caso de los sonidos plosivos, si el micrófono es ubicado directamente enfrente de la boca del hablante, entonces es muy susceptible de ser bombardeado por pequeñas ráfagas de aire ocasionadas por los sonidos plosivos. La mejor forma de tratar el problema es de rodear el micrófono con un material esponjoso transparente acústico, que rápidamente disipe la velocidad del viento de las pronunciaciones plosivas, permitiendo a las vibraciones acústicas normales pasar a través del micrófono.

Otras fuentes de ruido externo, tal como los ventiladores en las computadoras, aire acondicionados, teléfonos, y otras personas hablando puede también causar problemas con la exactitud del sistema de reconocimiento. Otra técnica para cancelar el ruido externo es filtrar la señal de audio antes procesarla. Debido a que las frecuencias de voz que contienen información relevante están dentro de un rango relativamente estrecho desde 200 a 3000 Hz, el espectro de audio puede ser filtrado a través de un filtro pasabanda para rechazar las señales acústicas fuera de ese rango de frecuencias.

- Calidad del Micrófono: Probablemente, el factor que más influye en la adquisición electrónica de señales del habla es el tipo de micrófono que se esta usando. Existen, principalmente, cuatro tipos de micrófonos disponibles en el mercado, los cuales son el Electreto (variante del micrófono de condensador), el Dinámico, el de Cristal y el de Carbón.

Para percibir fácilmente las diferencias entre estos tipos de micrófonos, sus características principales son comparadas en la tabla.

	TIPOS DE MICROFONO			
Parámetro	Electreto	Dinámico	Cristal	Carbón
Respuesta en Frecuencia	Excelente	Excelente	Bien	Regular
Distorsión	Muy bajo	Muy bajo	Bajo	Alto
Cancelación de ruido	Excelente	Bien	Regular	Regular
Tamaño	Pequeño	Medio	Grande	Grande
Peso	Bajo	Medio	Bajo	Medio
Costo	Alto	Alto	Medio	Bajo
Nivel de Salida	Bajo (voltaje)	Medio (voltaje)	Alto (voltaje)	Alto (Resistencia)
Impedancia	Alto	Bajo	Alto	Bajo

Tabla 1: Principales diferencias de micrófonos más utilizados

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Tabla] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

Los dos parámetros más importantes en la lista, son las comparaciones de respuesta en frecuencia y la distorsión. Basados en estas comparaciones es recomendable el uso del Micrófono Dinámico y el Electreto.

3. Reconocimiento de voz

El Reconocimiento de Voz, es el proceso por el cual un conjunto de algoritmos computacionales son capaces de traducir fielmente los sonidos de una unidad lingüística (palabra, sílaba o fonema) a un código simbólico que representa al mensaje.

El sistema que se expondrá utiliza la palabra como unidad lingüística, lo que supone que el hablante pronuncia las palabras con pequeñas pausas entre ellas, las cuales son detectadas por el sistema.

3.1 Elementos del Sistema

Un sistema de reconocimiento de voz suele estar constituido por los siguientes elementos:

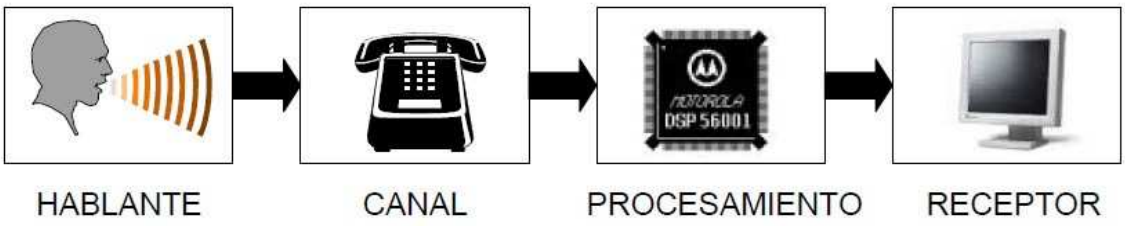


Figura 8: Elementos de un sistema de Reconocimiento de voz

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

Hablante o Locutor: Es el individuo que emite el mensaje. Este es uno de los elementos que introduce mayor variabilidad en la forma de onda de entrada. Una persona no pronuncia siempre de la misma forma, debido a distintas situaciones físicas y psicológicas.

Canal: Es el medio físico apto para la transmisión de los sonidos de voz, y que pone en contacto el sistema fonador del locutor y el sistema de procesamiento. Por ejemplo el aire o la línea telefónica.

Procesamiento: Es la etapa que se encarga de digitalizar la señal analógica del hablante a fin de extraer patrones característicos que representan a la unidad lingüística utilizada en el sistema, así como de realizar un proceso de clasificación para determinar los resultados.

Receptor: Interpreta el resultado obtenido en el Procesamiento, y dependiendo de la aplicación, puede ejecutar un comando de control, un dato de entrada a una aplicación, o simplemente mostrar en pantalla el resultado de una conversión Voz a Texto, entre otras.

3.2 Procesamiento

Las etapas en la cuales se divide el procesamiento son las siguientes:

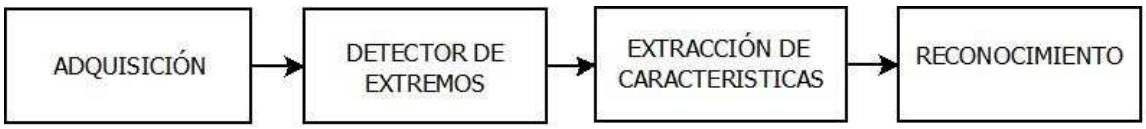


Figura 9: Etapas del procesamiento

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

Seguidamente se describen detalladamente cada una de estas etapas, excepto la etapa de reconocimiento.

3.2.1 Adquisición de la Señal de Voz

El proceso consiste en convertir la señal analógica de la voz en una cadena de datos binarios.

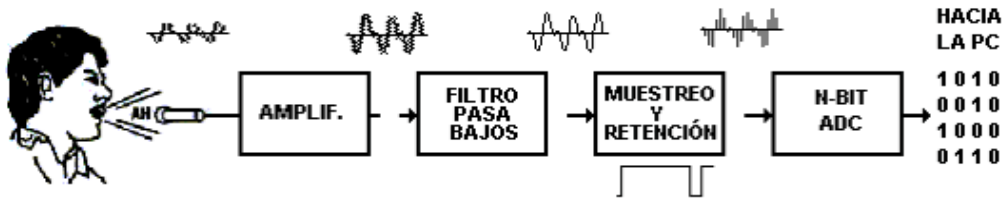


Figura 10: Proceso de adquisición de la señal de voz

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

La señal de voz es capturada a través de un micrófono que convierte las ondas acústicas del sonido en señales eléctricas, es decir, corriente o voltaje.

El proceso de muestreo, retención y cuantificación es repetido sucesivamente hasta que la forma de onda sea completamente capturada.

El número de bits utilizados en la cuantificación afecta la calidad de la señal adquirida:

$$SNR(dB) = 1.76 + 6.02N$$

donde N es el número de bits utilizados.

En aplicaciones de Procesamiento de Voz comerciales, se debería efectuar una cuantificación de por lo menos 12 bits, esto debido a que diferentes hablantes producen niveles de amplitud muy fluctuantes, por lo que la SNR sufriría importantes variaciones.

El espectro de las ondas sonoras puede tener componentes de frecuencias hasta aproximadamente 20KHz aunque en la mayoría de los Sistemas de Reconocimiento de Voz las frecuencias por encima de 3 o 4 KHz son redundantes y proveen mucho menos información que las que están debajo de ese rango.

En Sistemas de Reconocimiento a través de la línea telefónica, se debe considerar que las señales de voz transmitidas sobre la red telefónica están usualmente limitadas a un rango de frecuencias por debajo de los 3.3Khz.

Así, el teorema de Nyquist indica que la tasa de muestreo debería ser mantenida a una frecuencia de por los menos 6.5 o 7KHz.

Para aplicar técnicas de análisis y procesado, debemos limitar el segmento a procesar a un orden de ms debido al carácter pseudo-estacionario que presenta la señal sólo a corto plazo. Esto obligará al uso de TRAMAS de voz de la duración reseñada (10-20ms). La trama deberá poseer un tamaño de potencia de 2, debido a que haremos uso de la FFT en etapas posteriores.

– Para una implementación ONLINE será necesario hacer un procesado por bloques o crear un buffer que almacene las muestras de una trama.

– Tamaño: Fs x Tiempo_de_adquisicion

Ejemplo: para una frecuencia de muestreo de 46000Hz:

46000 x 0.02 = 960 muestras (20ms) que se aproximara a 1024 muestras

3.2.2 Detector Automático de Extremos

Esta etapa realiza la detección de inicio y fin de una palabra almacenada previamente en un buffer estático, entregando a la etapa posterior la palabra delimitada con su longitud exacta.

El bloque de Adquisición ira entregando tramas al de Detector de Extremos que en un principio las almacenara en un buffer de tamaño fijo para su posterior análisis y delimitación. Este buffer deberá ser de un tamaño que al menos pueda contener una palabra, por lo que se ha estimado que una palabra tiene una duración aproxi. de 2 seg:

Ejemplo: para Fs=46000Hz

Tamaño del buffer Palabra[ ]: 2seg x 46000= 92000 muestras

La Detección de Extremos se basa en el análisis de la evolución la Energía y los Cruces por Cero de las tramas. No obstante, existen alternativas como el parámetro COPER, que combinando estas dos técnicas permiten evaluar la evolución de la señal con un solo parámetro.

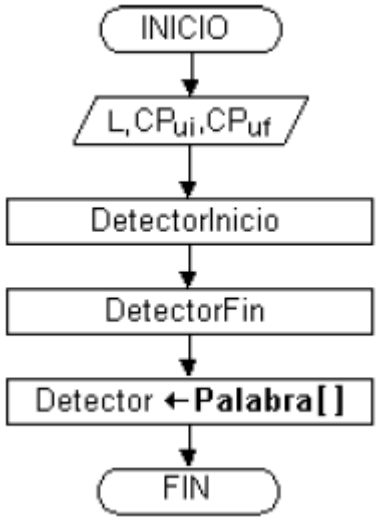


Figura 11: Diagrama de flujo de la subrutina detector automático de extremos

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

Una vez detectado el inicio de pronunciación, el sistema ira almacenando tramas asta que sea detectado el final de esta.

El proceso de detección de inicio en una implementación online estaría en bucle asta que se produjese la detección de inicio. Una vez detectado el final de pronunciación atendería una petición de interrupción del siguiente bloque, al que se le entregaría la palabra delimitada para la extracción de características. Transcurrido el procesado de los posteriores bloques se reanudaría la subrutina. Esto implica que posiblemente se pierdan muestras de voz debido al tiempo de procesado de las etapas posteriores, por lo que se exige que el procesado sea eficiente en cuanto a calculo computacional (tiempo de procesado) en dichas etapas posteriores.

3.2.2.1 Parámetro COPER

Este algoritmo es similar al de Cruces por Cero, pero a las funciones signo (1 si x>0, -1 si x<0), se les ha multiplicado por la Energía de la muestra analizada, es así que la densidad acumulada no sólo dependerá del cambio de signo de las muestras sino también de su amplitud, por esto, el ruido de fondo no logrará una gran acumulación de densidad, debido a que posee una pequeña amplitud comparada con las palabras pronunciadas. Analizando solo la evolución del parámetro COPER podremos de las tramas con el fin de saber si contiene información útil para el procesado

Se calcula el parámetro COPER de cada trama almacenada en el buffer estático Palabra[ ] sucesivamente.

$$\mathrm{COPER} = \sum_{m=0}^{L} \left| y[m] \cdot \left| y[m] \right| - y[m-1] \cdot \left| y[m-1] \right| \right|$$

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

Donde L es el tramaño de la trama.

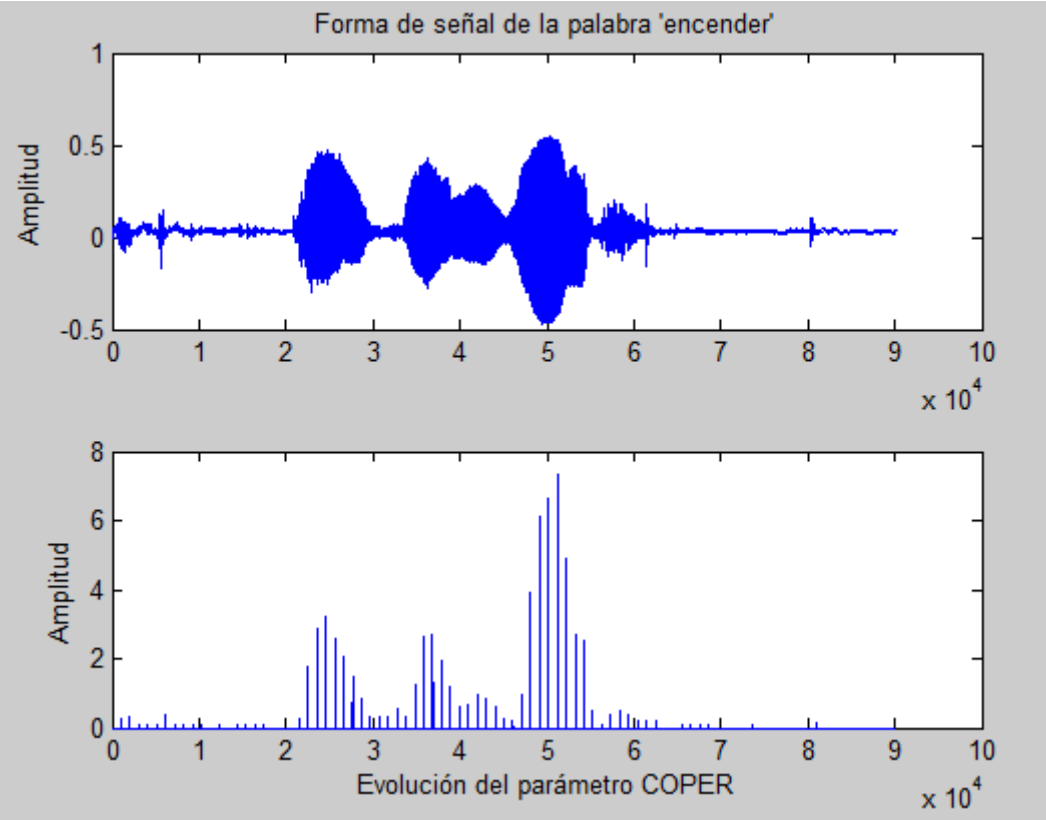
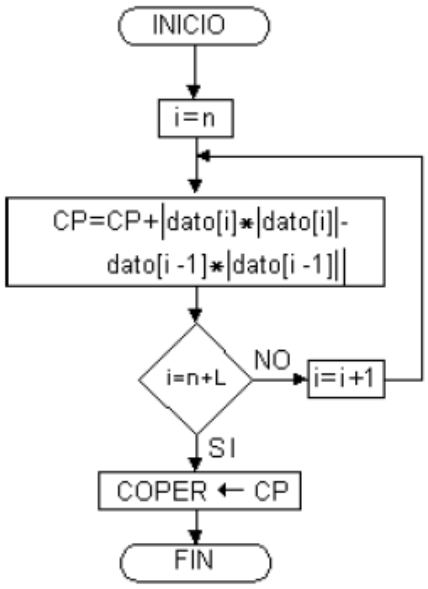


Figura 12: Evolución del parámetro COPER de la palabra ‘encender’

En la figura se muestra el diagrama de flujo que debe seguir el sistema para el calculo del parámetro COPER.



Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

Figura 13: Diagrama de flujo para el calculo del parámetro COPER

Donde:

L es el tamaño de la trama
n es el índice de la trama bajo análisis.

3.2.2.2 Detector Inicio

A partir del análisis del parámetro COPER se determina un Umbral de Inicio (Cui) para una correcta delimitación del inicio de la palabra.

Cuando un numero determinado de tramas supera el Umbral de Inicio, se empieza a almacenar dichas tramas en un buffer dinámico PalabraDelimitada[] hasta que es detectado el final de pronunciación.

3.2.2.3 Detector Fin

Una vez detectado el inicio de pronunciación se inicia el detector de fin. Para detectar el final de la palabra usaremos un método llamado NVENT. A partir del análisis de la evolución del parámetro COPER necesitaremos fijar un valor para el Umbral de Final (Cuf).El método consisten en lo siguiente:

– Se detectara el final de palabra una vez hayan transcurrido N tramas que no superen el Umbral Final (Cuf), en cuyo instante se dejaran de almacenar las tramas y quedara la palabra delimitada en un buffer estático (PalabraDelimitada[]).

– El valor de N se establece según cuantas tramas consecutivas estén contenidas dentro de un silencio intermedio de una palabra.

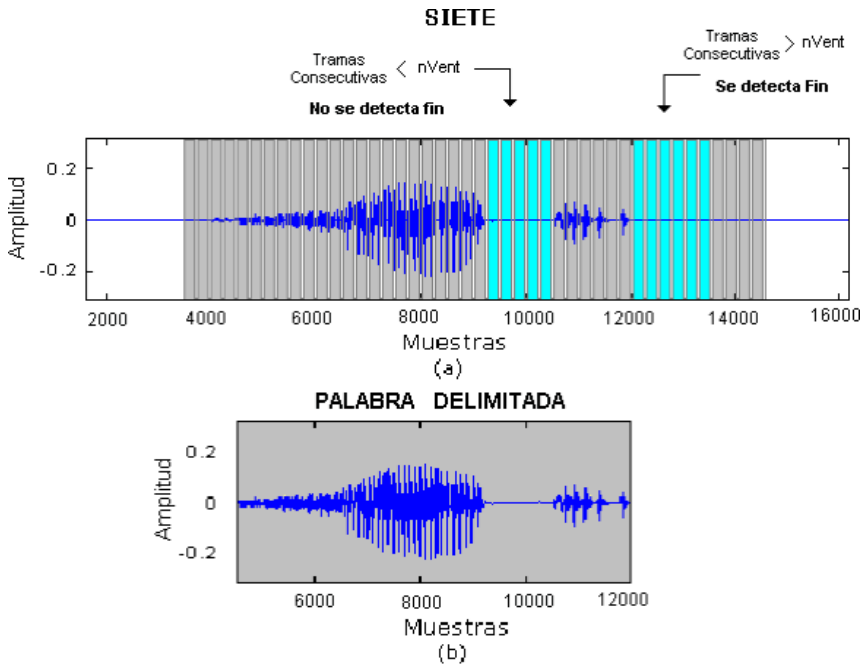


Figura 14: Grafico explicativo método Nvent.

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

3.2.3 Extracción de características

Existen varios métodos para la extracción de patrones característicos de la señal de voz con el fin de hacer más ligero el cálculo computacional. Un método muy eficiente para la extracción de características que actualmente es el más utilizado en reconocedores comerciales son los Coeficientes Cepstrales en Escala de Mel (MFCC), basadas en criterios preceptúales.

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.

En la siguente figura se muestra el diagrama de flujo para el cálculo de los coeficientes MFCC:

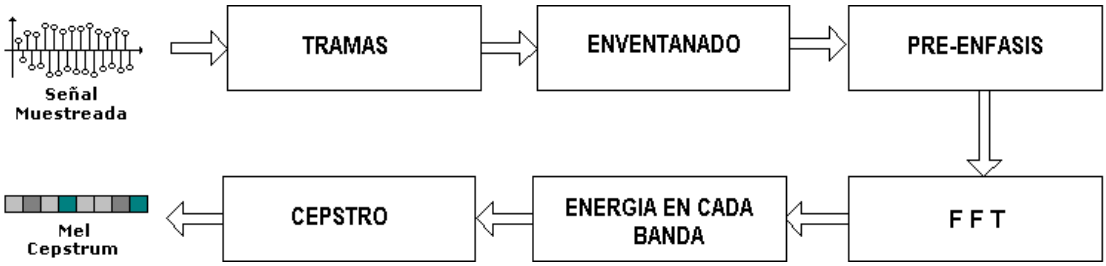


Figura 15: sub-bloques para el calculo de los MFCC.

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

Se analiza la palabra delimitada, almacenada previamente en el vector PalabraDelimitada[ ], obtenido en la etapa de detección de extremos. El análisis del vector se realiza en tramas de longitud N, en pasos de M muestras (N>M). El proceso de análisis se detiene cuando se supera el número de muestras que posee la palabra delimitada.

Del resultado del análisis se obtiene una matriz con los coeficientes MFCC para bloques posteriores de normalización y comparador de patrones o red neuronal, bloques que no revisaremos en dicho trabajo.

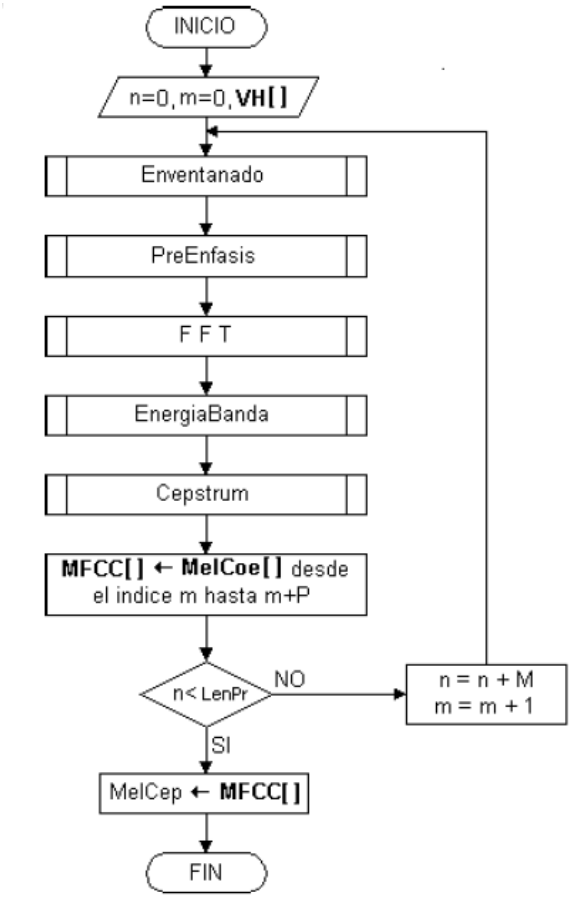


Figura 16: Diagrama de flujo para el calculo de los MFCC

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

A continuación serán vistos cada uno de los bloques con más detalles, explicando su funcionamiento:

3.2.3.1 Tramas

Se analizan tramas de 10 a 20 ms de duración, ya que es ahí, donde el análisis espectral muestra información distintiva entre los diferentes tipos de sonidos. Este bloque se reflejara en bucle externo que recorrerá la palabra

delimitada entregando al siguiente bloque tramas de longitud L con un paso de N muestras.

- Si utilizamos una frecuencia de muestreo de 46 KHz:
  - 46000 x 0.02 seg = 940 muestras que aproximamos a potencia de 2 debido a posterior uso de la FFT.
  - Esto es equivalente a M = 1024 / 46000 = 22,2 ms
    N = 512 / 46000 = 11,1 ms

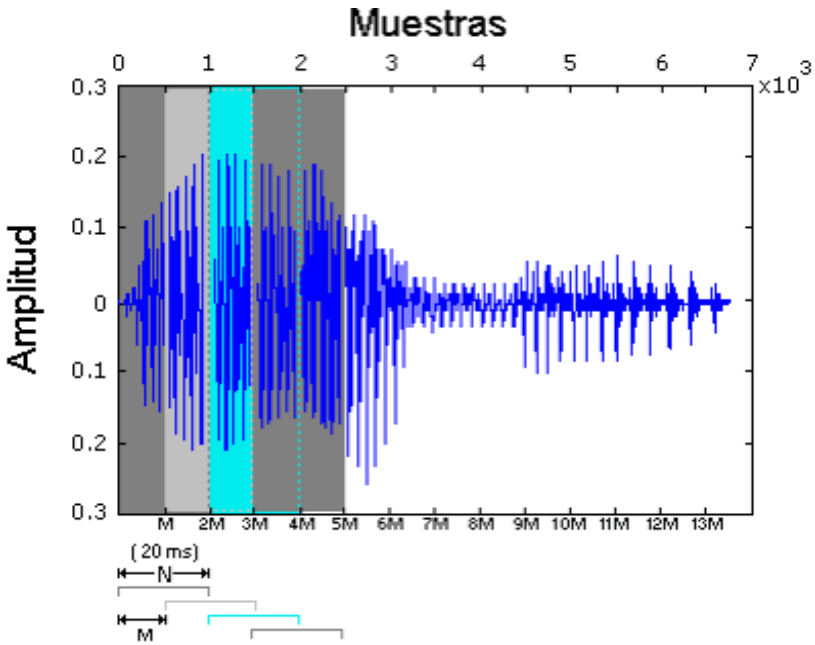


Figura 17: Gráfico explicativo de paso de muestras en el análisis.

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

La primera trama consiste de las N primeras muestras. La segunda empieza M muestras después que la primera, y la traslapa en N-M muestras. De la misma manera, la tercera trama empieza 2M muestras después de la primera trama (o M muestras luego de la segunda trama) y la traslapa en N-2M muestras. El proceso de análisis continúa hasta que se hace un barrido total de la palabra adquirida del bloque anterior (PalabraDelimitada[]).

3.2.3.2 Enventanado

Tras obtener una trama de la señal, ésta posee discontinuidades en el inicio y final:

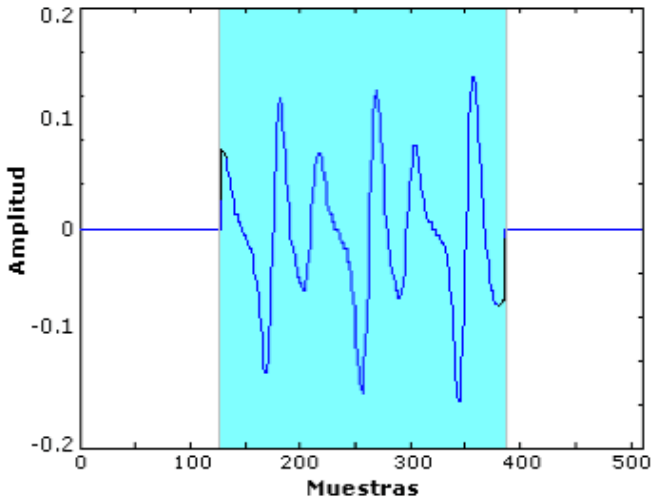


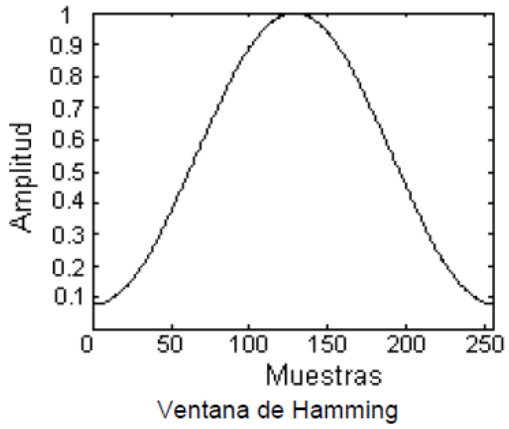
Figura 18: Trama de la señal una serñal de voz

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

Estas discontinuidades producen una distorsión de la Transformada de Fourier de la señal. A fin de minimizar las discontinuidades al inicio y al final de ésta será necesario hacer un proceso de enventanado eligiendo un tipo de ventana que produzca la menor distorsión posible.

Típicamente se utiliza la ventana de Hamming la cual tiene la siguiente formulación matemática:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$



Tras multiplicar punto a punto las muestras los extremos de la trama original quedan suavizados:

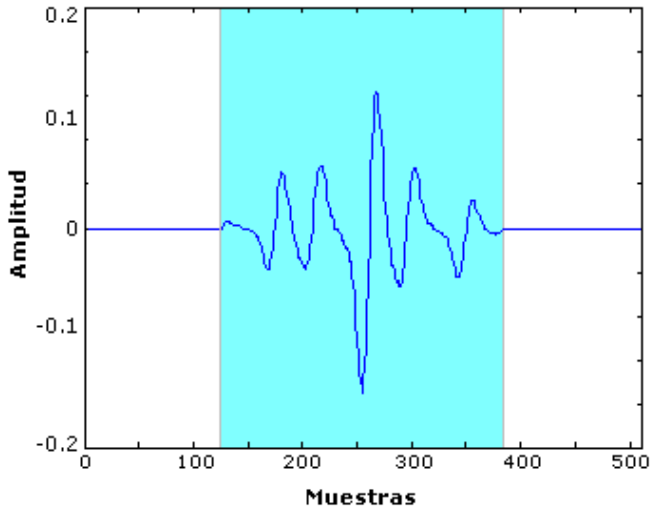


Figura 19: Trama de una señal de voz multiplicada por una ventana de hamming

3.2.3.3 Preenfasis

Debido a que la señal de voz se atenúa a 6 dB/octava conforme aumenta la frecuencia, es necesario introducir un filtrado cuya función es incrementar la relevancia de las componentes de alta frecuencia. Este proceso se conoce con el nombre de preénfasis y puede ser diseñado a través de un filtro digital paso alto. Este filtro paso alto puede implementarse con la siguiente ecuación en diferencias:

$$y[n] = x[n] - a \cdot x[n-1]$$

donde a es una constante que varía entre 0 y 1 (valor típico 0.95).

3.2.3.4 Transformada Rápida de Fourier

En esta etapa se convierte cada trama de N muestras en el dominio del tiempo al dominio de la Frecuencia.

La Transformada Rápida de Fourier (FFT), es un algoritmo rápido, que permite implementar eficientemente la DFT.

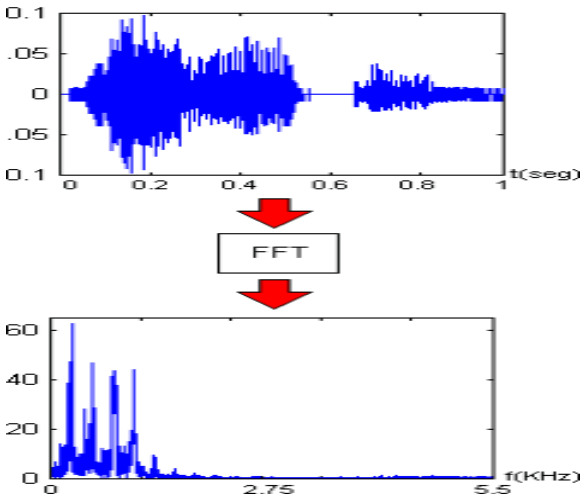


Figura 20: FFT de una señal de voz

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

Los distintos algoritmos utilizados para elaborar una FFT explotan las propiedades de simetría y periodicidad del factor de fase. Estos se basan en el uso de la estrategia “divide y vencerás”, la cual consiste en la descomposición de una DFT de N muestras en DFTs más pequeñas, donde N se puede representar por el producto de dos enteros. No entraremos en más detalle ya que es un tema sumamente difundido y no es la finalidad del presente trabajo.

3.2.3.5 Energía en cada Banda

Estudios científicos han mostrado que la percepción humana del contenido de las frecuencias de los sonidos de la señal de voz no sigue una escala lineal. Como se sabe la frecuencia es una entidad física y por tanto puede ser medida de forma objetiva por diferentes medios. Por el contrario la altura o tono de un sonido, es un fenómeno totalmente subjetivo y por tanto, no es posible medirlo de forma objetiva. Normalmente, cuando se aumenta la frecuencia de un sonido, su altura también sube, sin embargo esto no se da de forma lineal, o sea, no se corresponde la subida del valor de la frecuencia con la percepción de la subida de tono.

Por procedimientos estadísticos sobre un determinado número de personas sin conocimientos musicales se fijó el valor de la escala subjetiva mediante una ley empírica que define una nueva escala de tonos. Para medir los intervalos de esta escala se utiliza la unidad Mel, llamada también Melio. Por definición un sonido de 1000 Hz, con 40 dB por encima del umbral de percepción, tiene un tono de 1000 mels. Así, para computar las unidades Mels del sonido a una frecuencia f(Hz), se utiliza la siguiente fórmula.

– Mel(f) = 2595*log10(1+f/700)

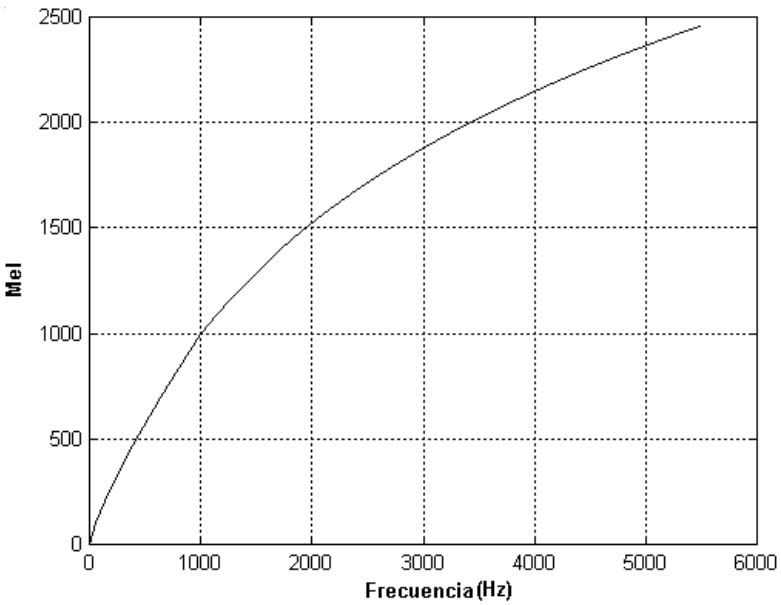


Figura 21: Gráfica Mel vs. Frec.

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

La escala de Frecuencia Mel tiende a un espaciamiento lineal de frecuencia por debajo de los 1000 Hz y a un espaciamiento logarítmico sobre los 1000Hz.

Una manera de aproximarse a este espectro subjetivo es utilizar un banco de filtros, mucho más estrechos y linealmente espaciados hasta aproximadamente 1KHz, y muy amplios y logarítmicamente espaciados a partir de 1KHz.

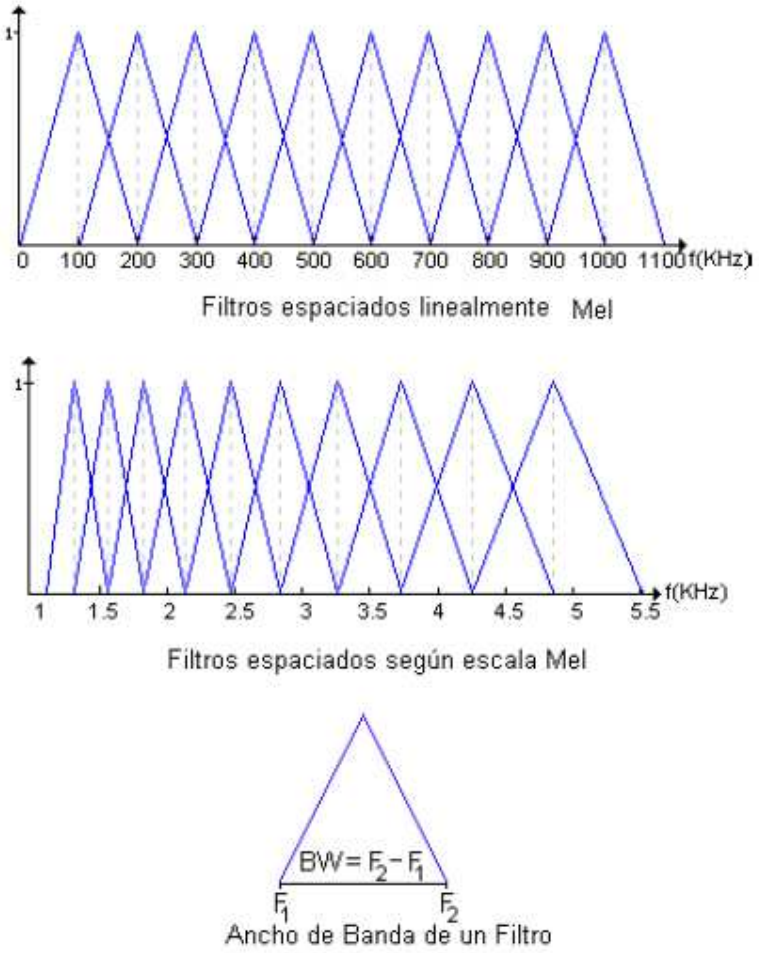


Figura 22: Banco de Filtros

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

De este modo, se da mayor importancia a la información contenida en las bajas frecuencias en concordancia con el comportamiento del oído humano.

Los filtros son aplicados directamente en el dominio de la frecuencia, y su respuesta está dada por la siguiente ecuación:

$$\omega(f) = \begin{cases} \dfrac{2f}{BW-1}, & 0 \le f \le \dfrac{BW-1}{2}\,(Hz) \\ 2 - \dfrac{2f}{BW-1}, & \dfrac{BW-1}{2} \le f \le BW-1\,(Hz) \end{cases}$$

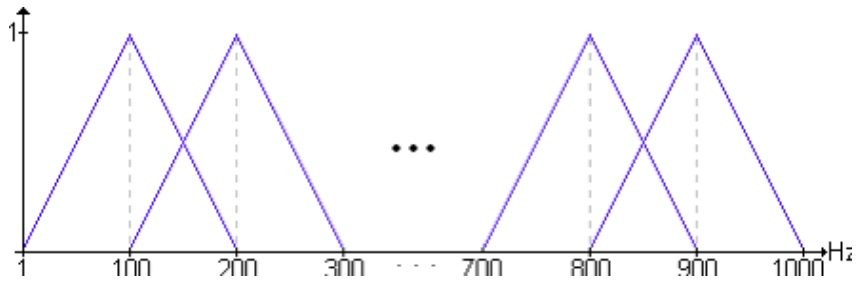
Donde BW es el ancho de banda del filtro triangular y f es la frecuencia en Hz.

Típicamente se toma un banco de 20 filtros triangulares, los diez primeros filtros se ubican hasta 1 KHz, son linealmente espaciados y dividen el espectro en 10 espacios iguales. Poseen un ancho de banda de 200Hz, y se traslapan en 100Hz, tal como se muestra en figura.

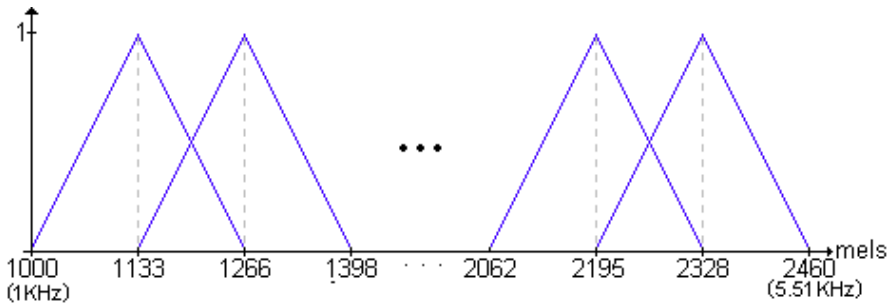
Los siguientes 10 filtros triangulares se ubican después de 1KHz, hasta 5.5KHz, 1000 y 2460 mels, respectivamente, aplicando la ecuación;

$$Mel(f) = 2595 + \log 10(1 + f/700)$$

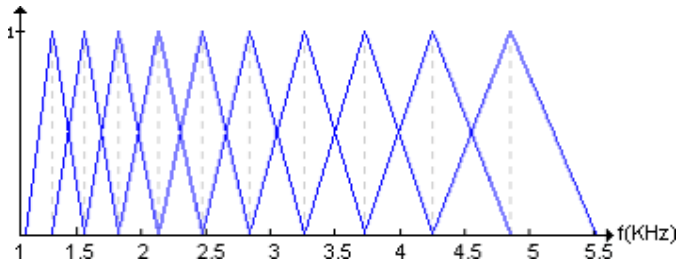
Los filtros son espaciados uniformemente en la escala de frecuencia Mel, pero en el dominio de la frecuencia, estos filtros se encuentran espaciados logarítmica- mente.



Filtros triangulares escala lineal



Filtros triangulares escala de Mel



Filtros Triangulares en el dominio de la frecuencia

Figura 22: Banco de filtros

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

3.2.3.5.1 Diseño del banco de filtros triangulares

Para diseñar el banco de filtros debe determinarse las frecuencias de Inicio (fStart), Centro (fCent) y Final (fStop) de cada filtro triangular.

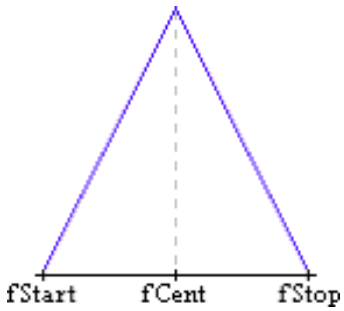


Figura 23: Filtro triangular.

Como se sabe, la trama de análisis, posee un tamaño N de 1024, luego de aplicarle la Transformada Rápida de Fourier (FFT), el espectro en magnitud de esta señal poseerá un tamaño de 1024 muestras, lo que equivale a un ancho de banda de Fs/2.

El objetivo del diseño de los filtros es establecer una correspondencia entre las frecuencias de inicio, centro y final del filtro y las muestras del espectro obtenido en la FFT.

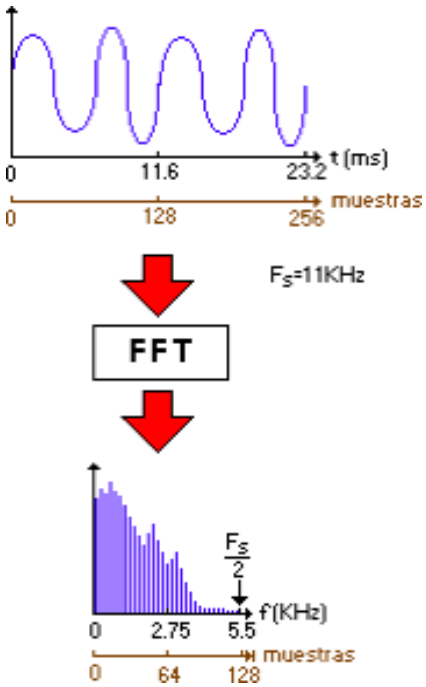


Figura 23: Paso de Frec. a numero de muestra

La siguiente fórmula establece una relación entre un intervalo de frecuencia y el número de muestras correspondiente:

$$n = \frac{N \cdot \Delta f}{Fs}$$

donde:
Δf: es el intervalo de frecuencia en Hz.
Fs: es la frecuencia de muestreo en Hz.
N: es el número total de muestras de la trama.
n: es el número de muestras.

Diseño de los filtros en escala lineal:

Inicialmente se debe establecer la frecuencia de Inicio (FrecInicio) y Final (FrecFinal) del banco de filtros:

- FrecInicio = 0Hz
- FrecFinal = 1kHz

Con estos valores se determina que el tamaño del banco de filtros, Δf es igual a 1KHz, y reemplazando en la ecuacion anterior:

$$n = \frac{N \cdot \Delta f}{Fs} = \frac{1024 \cdot 1000}{46000} = 22.26 \approx 20$$

Como se van a utilizar 10 filtros para dividir el espectro hasta 1Khz, en 10 espacios uniformes, el número n de muestras debe ser redondeado a multiplo de 10.

Al tener n=20, los índices se crean por una simple regla de tres simple:

$$Indice = \frac{n \cdot f}{T} \qquad \begin{array}{l} n = 20 \\ T = 1000 Hz. \end{array}$$

	Herzio			Muestras	
fStart	fCent	fStop	nStart	nCent	nStop
0	100	200	0	2	4
100	200	300	2	4	6
200	300	400	4	6	8
300	400	600	6	8	10
400	500	600	8	10	12
500	600	700	10	12	14
600	700	800	12	14	16
700	800	900	14	16	18
800	900	1000	16	18	20
900	1000	1100	18	20	22

Tabla 2: tabla índices muestra filtros triangulares escala lineal

Diseño de los filtros en escala Mel:

En primer lugar, se debe determinar la frecuencia de Inicio (FrecInicio) y fin (FrecFinal) del banco de filtros, y realizar su conversión a Mels:

- FrecInicio = 1 KHz > MelInicio = 1000 mels
- FrecFinal = 5.5 KHz > MelFinal = 2460.5 mels

Se divide el espacio, en escala Mel, comprendido entre MelInicio y Melfinal en 10 segmentos de igual tamaño (10 filtros triangulares), de donde se obtiene el Inicio, Centro y Final de cada filtro en Melios. Luego lo pasaremos a Herzios.

$$f = 700 \cdot (10^{Mels/2595} - 1)$$

	Melios			Herzios	
mStart	mCent	mStop	fStart	fCent	fStop
999,99	1132,76	1265,53	1000,01	1212,56	1451,68
1132,76	1265,53	1398,31	1212,56	1451,68	1720,73
1265,53	1398,31	1531,08	1451,68	1720,73	2023,39
1398,31	1531,08	1663,85	1720,73	2023,39	2363,89
1531,08	1663,85	1796,63	2023,39	2363,89	2747,00
1663,85	1796,63	1929,4	2363,89	2747,00	3177,97
1796,63	1929,40	2062,18	2747,00	3177,97	3662,87
1929,40	2062,18	2194,95	3177,97	3662,87	4208,36
2062,18	2194,95	2327,72	3662,87	4208,36	4822,05
2194,95	2327,72	2460,5	4208,36	4822,05	5512,52

Tabla 3: tabla índices muestra filtros triangulares escala Mel

Una vez obtenidas las frecuencias de Inicio, Centro y Final de los filtros, es necesario establecer a que número de índice corresponde. Ya que, en este caso, no se trata de un espaciamiento lineal, se debe obtener la distancia de una frecuencia a otra y realizar su conversión a número de muestras.

$$localización\_en\_muestra \cong floor\left(\frac{f}{Fs} \cdot N\right);$$

Aplicando las fórmulas anteriores se obtiene la siguiente tabla:

	Melios			Herzios			Muestras	
mStart	mCent	mStop	fStart	fCent	fStop	nStart	nCent	nStop
999,99	1132,76	1265,53	1000,01	1212,56	1451,68	22	27	32
1132,76	1265,53	1398,31	1212,56	1451,68	1720,73	27	32	38
1265,53	1398,31	1531,08	1451,68	1720,73	2023,39	32	38	45
1398,31	1531,08	1663,85	1720,73	2023,39	2363,89	38	45	53
1531,08	1663,85	1796,63	2023,39	2363,89	2747	45	53	61
1663,85	1796,63	1929,4	2363,89	2747	3177,97	53	61	71
1796,63	1929,4	2062,18	2747	3177,97	3662,87	61	71	82
1929,4	2062,18	2194,95	3177,97	3662,87	4208,36	71	82	94
2062,18	2194,95	2327,72	3662,87	4208,36	4822,05	82	94	107
2194,95	2327,72	2460,5	4208,36	4822,05	5512,52	94	107	123

Table 4: tabla índices muestra filtros triangulares

Una vez finalizado el diseño de los filtros, estos índices obtenidos son utilizados para obtener la función de transferencia de cada filtro y realizar el proceso de filtrado de la señal.

En la figura se muestra el diagrama de flujo, en el cual se implementa la función de transferencia del filtro triangular.

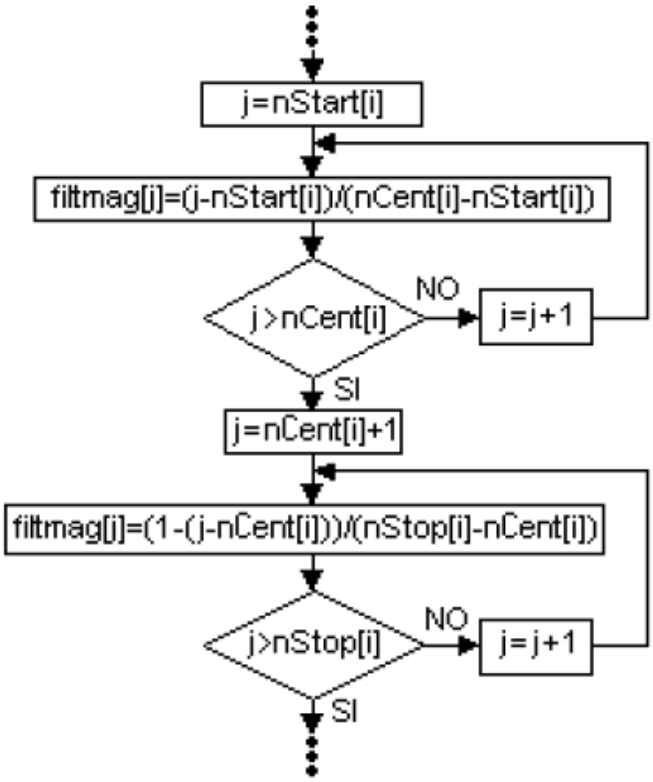


Figura 24: Diagrama de flujo para el calculo de la función de transferencia de un filtro triangular

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

Deberemos almacenarlos en forma de matriz para que no se solapen una vez se incremente el índice para el cálculo de la función de transferencia del siguiente filtro triangular.

Luego se calcula la energía a la salida de cada filtro

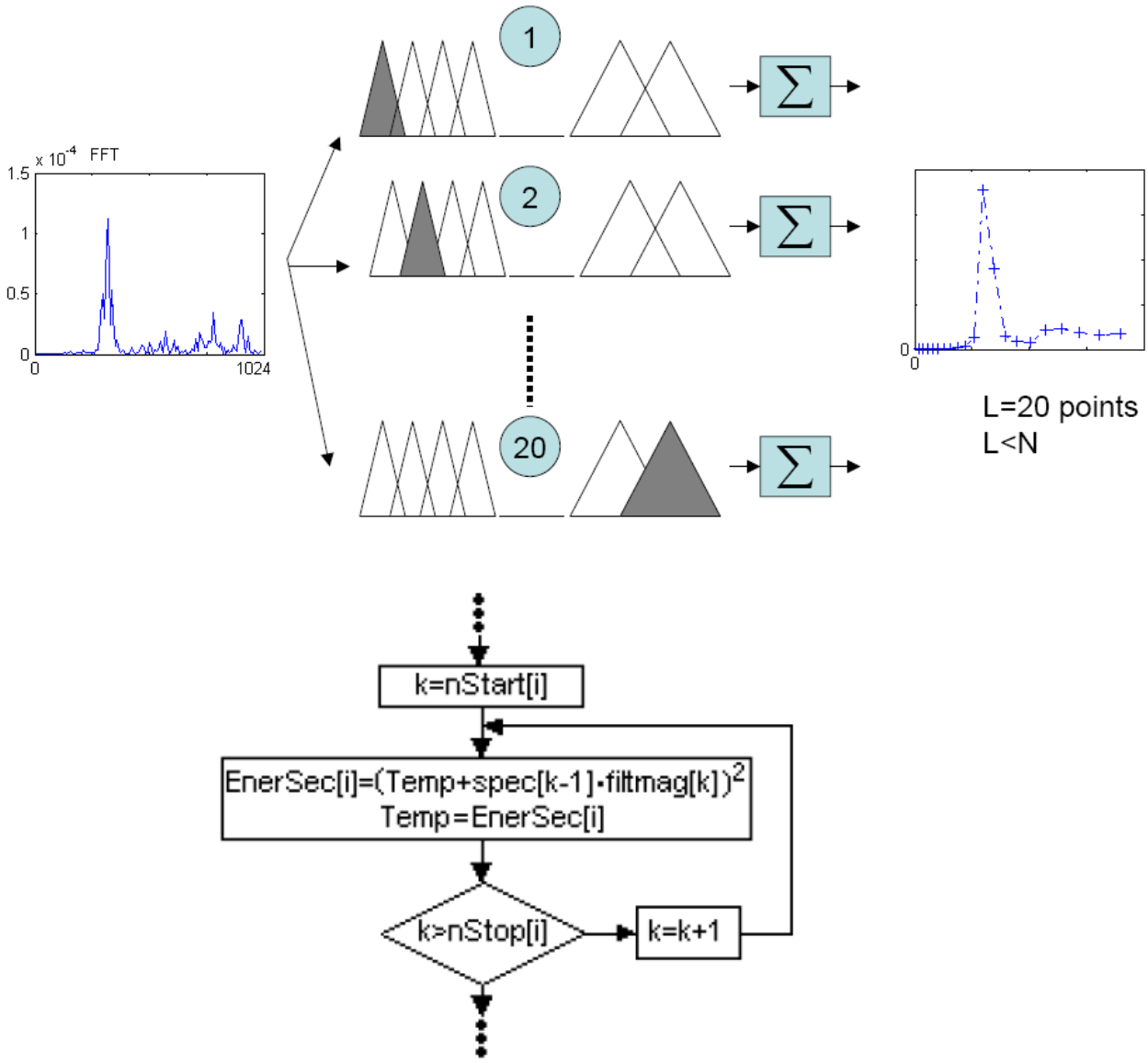


Figura 25: Diagrama de flujo para el filtrado y la obtención de la energia

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

Con esto obtenemos 20 coeficientes por trama analizada que podemos almacenarlos en forma matricial

3.2.3.6 Cepstrum

El Cepstrum se define como la transformada inversa del logaritmo del módulo de la Transformada de Fourier de la señal.

$$c(n) = F^{-1}[\log | X(w) |]$$

Donde:

$$X(w) = F[x(n)]$$

La representación cepstral del espectro del habla, provee una buena representación de las propiedades espectrales locales de la señal para cada trama.

En el modelo del tracto vocal, la voz se genera por una excitación producida por dos fuentes, la cual pasa a través de un filtro, cuya respuesta en frecuencia, modifica el espectro adicionando información lingüística al sonido. Así, para realizar el reconocimiento de las palabras pronunciadas, bastaría conocer solamente las características del tracto vocal (o el filtro que lo modela), ya que la información proveniente de las cuerdas vocales (excitación) sólo proporciona información acerca del locutor. Precisamente, este tipo de análisis denominado cepstral, se utiliza para realizar la separación de estos dos parámetros.
Las componentes bajas están relacionadas con las características del tracto vocal y las altas componentes cepstrales, con la información sobre el locutor.

En lo que respecta a la Transformada Inversa, debido a que los coeficientes del espectro y su logaritmo son número reales, se pueden convertir al dominio del tiempo utilizando la Transformada Coseno Discreta que hace las veces de Transformada Inversa de Fourier.

El cálculo de los MFCC, responde a la siguiente expresión:

$$MFCC_j(i) = \sum_{K=1}^{NF} \log\left[E(j,k)\right] \cos\left[i\left(k - \frac{1}{2}\right)\frac{\pi}{L}\right]; i = 1,2,3,\ldots,P$$

Donde :
k : es la banda de frecuencias.
j : es la trama en curso.
E(j,k) : es la energía de la banda k en la trama j.
NF : es el número de bandas o filtros.
P : es el número total de coeficientes MFCC (10, en nuestro caso).

- Típicamente, se evalúan 10 puntos de la transformada inversa, para así, sólo obtener información sobre las bajas componentes cepstrales.
- Tras el análisis de la palabra se obtendrá una matriz de patrones característicos de dimensiones 10 x numero de tramas que contenga esta.

- En la imagen se muestra el diagrama de flujo para la obtención de los MFCC, donde NF es el número de filtros utilizados y P es la cantidad de coeficientes a obtener, 10 en nuestro caso.

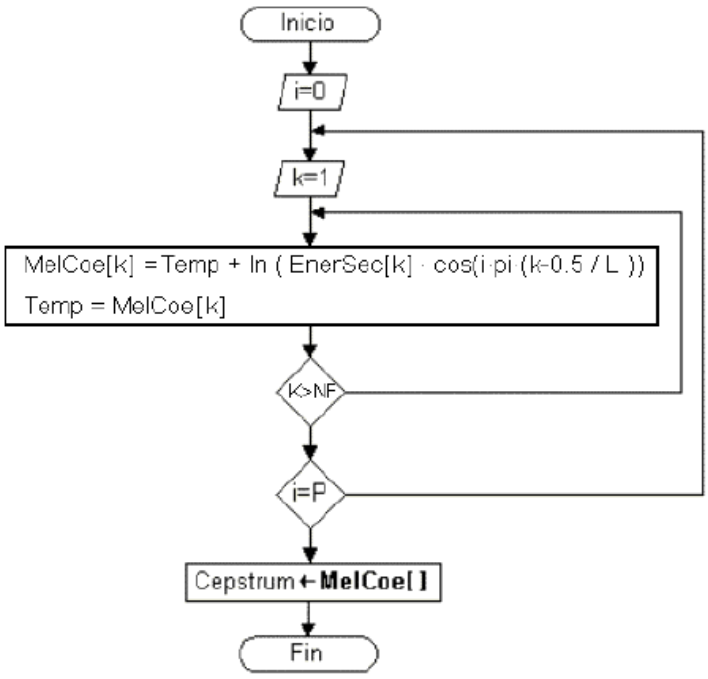


Figura 26: Diagrama de flujo para la obtención de los MFCC

Ref. Peralta, F. / Cotrina A., Tesis Reconocedor y analizador de voz facultad de ingeniería electrónica [Figura] Recuperado de UNMSM (Universidad Nacional Mayor de San Marcos)

4. Funciones implementadas en Matlab

Como materia de este trabajo se han implementado una serie de funciones en el software de tratamiento matemático Matlab con el fin de analizar y entender parte del diseño de un sistema de reconocimiento de voz.

4.1 Función AnalizaUmbral.m

La función AnalizaUmbral.m realiza un análisis del parámetro COPER de cada 1024 muestras de la variable de entrada, almacenando el resultado en la variable de salida cada 1024 posiciones. Se fijo el tamaño de la trama en 1024 por el criterio de que el tamaño de esta tenía que ser una duración de 20 ms. aproximadamente:

- Tamaño de la trama= Fs x TiempoDeAdquisición= 46000 x 0.020=920 muestras que se aproximan a 1024 (potencia de 2) debido al uso de la FFT en un bloque posterior.
- La señal de entrada debe ser unidimensional y normalizada a valor de 1.

```
function [ evolucionUmbral ] = AnalisisUmbral( Y )

 [~, columnas]=size(Y);

i=floor(columnas/1024);%numero de veces que calculara el umbral (nº
tramas, se usara para el bucle exterior)
m=2; %factor de índice de inicio, empieza por 2 porque necesitaremos
dato(x-1)
 n=1026;% factor de índice final

 for j=1:1:i,
  umbral=0;


      for k=m:1:n,

      umbral=umbral+abs(Y(k)*abs(Y(k))-Y(k-1)*abs(Y(k-1)));

      end
  evolucionUmbral(n)=umbral; %se almacena el valor cada 1024 muestras
para luego poder comparar con la señal de entrada

   m=m+1024;
   n=n+1024;

 end

end
```

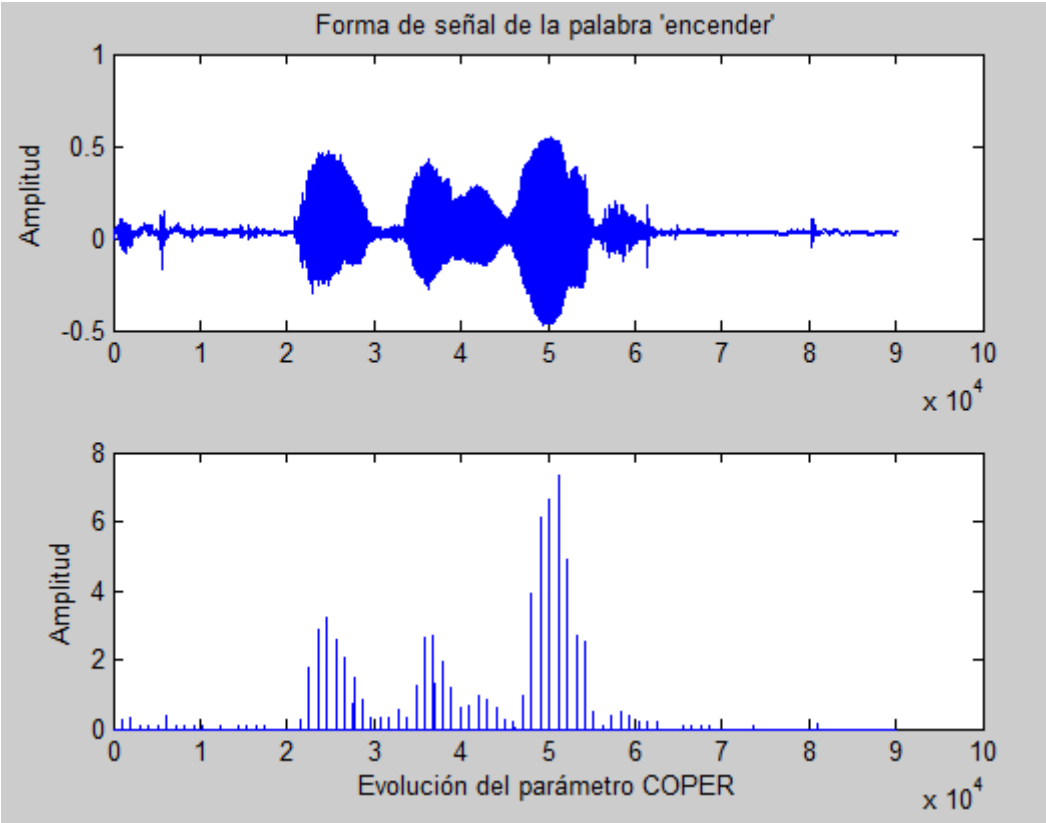


Figura 27: Señales de E/S de la función AnalizaUmbral.m

4.2 Función detectorExtremos.m

La función detectorExtremos.m realiza una segmentación de la señal de entrada basándose en la evolución del parámetro COOPER entregando la palabra delimitada y su longitud. Para ello a partir del análisis de la evolución del parámetro Cooper se ha fijado un Umbral de Inicio de 1.25 y un Umbral de Final de 1, al igual que el numero de tramas consecutivas que se deben detectar sin información para detectar el final de pronunciación en 15. Los valores los determinamos en procesos experimentales y el analisis del parametro COOPER.

```
function [ palabradelimitada , longPal ] = detectorExtremos( Y )

[~, columnas]=size(Y);
i=floor(columnas/1024);%numero de veces que calculara el umbral
                       (nºtramas)
m=2; %factor de índice de inicio, empieza por 2 porque necesitaremos
dato(x-1)
n=1026;% factor de índice final

      %esto contara las ventanas consecutivas sin info. para detectar
final pronunciación.
```

```
longPal=1;
t=0;
for j=1:1:i,            %esto segmentara la señal en tramas
  umbral=0;


    for k=m:1:n, %bucle recorre muestras de n a m y calcula el umbral
de la trama

      umbral=umbral+abs(Y(k)*abs(Y(k))-Y(k-1)*abs(Y(k-1)));

    end



  %evolucionUmbral(n)=umbral; %se almacena el valor cada 1024
posiciones

 %=DETECTOR INICIO=> si supera umbral se empieza a almacenar tramas

    if (umbral>1.25) %comprobamos si se detecta inicio, si supera
Umbral Inicio=1.25

      t=0;
    longPal=longPal+1024;
    palabradelimitada(longPal-1024:longPal)=Y(m:n); %almacenamos
tramas si t<15

    end

 %=DETECTOR FIN=> se detecta final si transcurren 15 tramas sin info
(t=15)

    if (umbral<1) %comprobamos si incrementamos t,

            if (t<10 && longPal~=1)

                t=t+1;

                longPal=longPal+1024;


                palabradelimitada(longPal-1024:longPal)=Y(m:n);
%almacenamos tramas si t<15


            end

    end

   m=m+1024;
   n=n+1024;

 end

end
```

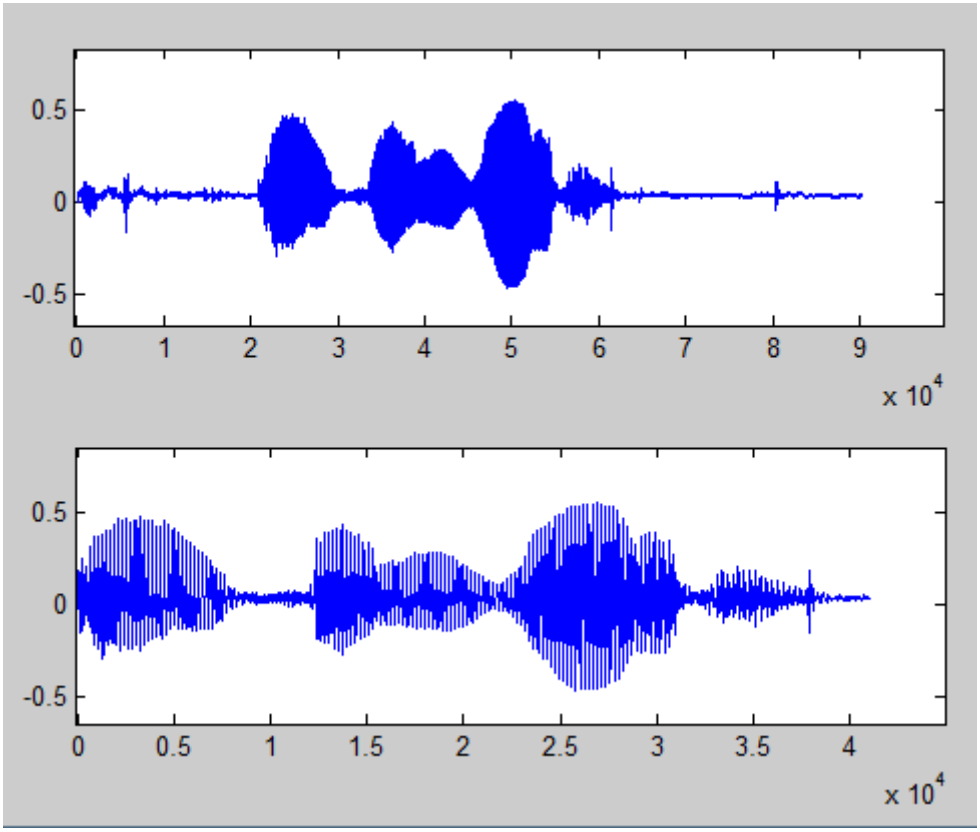


Figura 27: Señal de entrada y salida de la función DetectorExtremos.m.

4.3 Función FiltroPre.m

La función FiltroPre.m realiza un filtrado de la señal de entrada con el fin de realzar las frecuencias de interés.

```
function [ winSecPreEnf ] = filtroPreEnf( palabra )

[~, longPal]=size(palabra);

for i=2:1:longPal

    winSecPreEnf(i)=palabra(i)-0.95*palabra(i-1);

end
```

4.4 Función BancoDeFiltros.m

La función BancoDeFiltros.m construye el banco de filtros a partir de los indices de los filtros (nStart, nCent, nStop) ubicando los coeficientes de cada filtro triangular en forma de matriz para que no se solapen.

```
function [ filtroMag ] = BancoDeFiltros( nStart,nCent,nStop )


filtroMag=zeros(20,111);

for c=1:1:20

    j=nStart(c);

        while j<=nCent(c)

    filtroMag(c,j+1)=(j-nStart(c))/(nCent(c)-nStart(c));

       j=j+1;

           end

    j=nCent(c)+1;

        while j>=nStop(c)

        filtroMag(c,j+1)=(1-(j-nCent(c)))/(nStop(c)-nCent(c));

        end


end


hold on
plot(filtroMag(1,1:111),'b')
plot(filtroMag(2,1:111),'r')
plot(filtroMag(3,1:111),'g')
plot(filtroMag(4,1:111),'b')
plot(filtroMag(5,1:111),'r')
plot(filtroMag(6,1:111),'g')
plot(filtroMag(7,1:111),'b')
plot(filtroMag(8,1:111),'r')
plot(filtroMag(9,1:111),'g')
plot(filtroMag(10,1:111),'b')
plot(filtroMag(11,1:111),'r')
plot(filtroMag(12,1:111),'g')
plot(filtroMag(13,1:111),'b')
plot(filtroMag(14,1:111),'r')
plot(filtroMag(15,1:111),'g')
```

```
plot(filtroMag(16,1:111),'b')
plot(filtroMag(17,1:111),'r')
plot(filtroMag(18,1:111),'g')
plot(filtroMag(19,1:111),'b')
plot(filtroMag(20,1:111),'r')


end
```

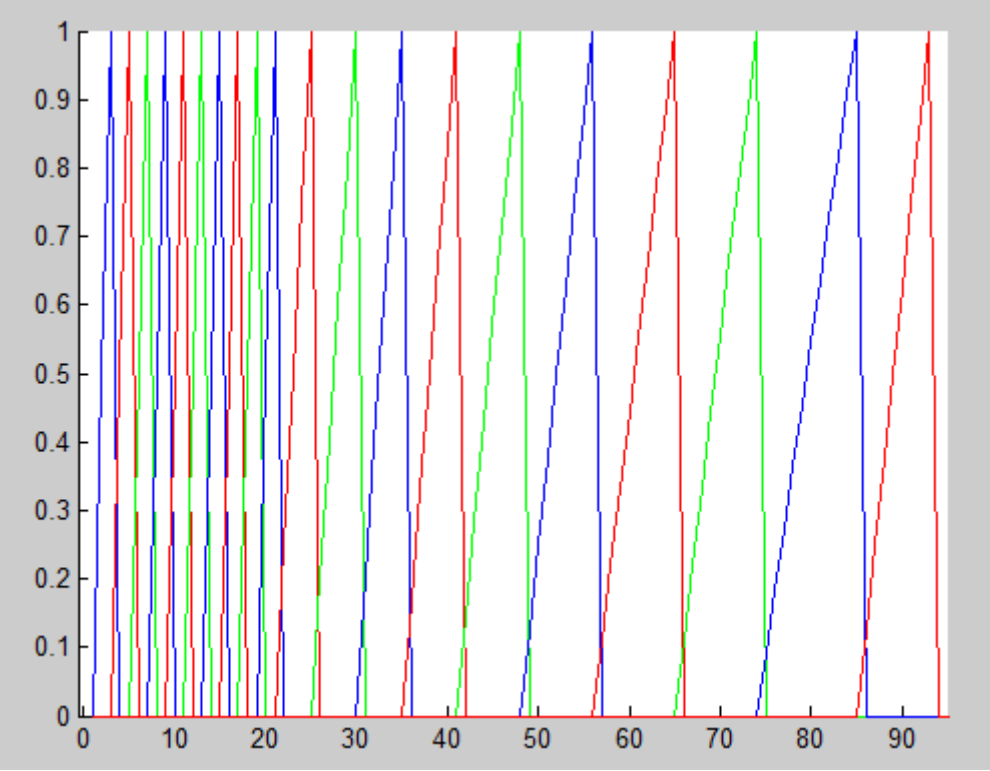


Figura 28: Banco de filtros con 20 filtros triangulares.

4.5 Función CoefMelCep.m

La función CoefMelCep.m extrae el patrón característico de las tramas de voz de la palabra delimitada almacenándolos en forma de matriz secuencialmente. La matriz obtenida pasaría a un bloque posterior para normalizar el tamaño del patrón característico con el fin de que no afectase la diferencia entre velocidades de pronunciación de los hablantes en las palabras detectadas a los bloques posteriores de decisión y comparación de patrones.

```
function [ MelCoe ] = coefMelCep( palabra , longPal )

%Se realiza analisis de tramas de longitud 1024 con un paso de 512
%muestras, dando como resultado 2*numTramas-1 bloques a analizar
```

```
nStart=[0 2 4 6 8 10 12 14 16 18 20 24 29 34 40 47 55 64 73 84;];
nCent=[2 4 6 8 10 12 14 16 18 20 25 29 35 41 48 55 64 74 85 97;];
nStop=[4 6 8 10 12 14 16 20 21 23 30 33 39 46 54 65 74 85 97 110;];

numTramas=floor((longPal-1)/1024);
m=1;
n=1024;
palabraEnv=0;
VH=hamming(1024);
palabraEnv=palabra;
Ener=zeros((2*numTramas-1),n);
EnergiaBanda=zeros((2*numTramas-1) , 20);
MelCoe=zeros((2*numTramas-1),100);
temp=0;
filtroMag=BancoDeFiltros(nStart,nCent,nStop);

  for   i=1:1:(2*numTramas-1)%índice trama bajo análisis, este  bucle
recorre las tramas de la palabra


    %Inicio analisis trama palabra(m:n)

    palabraEnv=palabra(m:n).*VH';%enventanado

    palabraPre=filter([1, 0.95],1,palabraEnv); %PreEnfasis

    Y=fft(palabraPre);

    Ener(i,1:1024)=Y.*conj(Y);


        %filtroMag(20,111) veinte filtros triangulares
        for NF=1:1:20 %este bucle recorre cada filtro del banco de
                      filtro

                    Energia=(filtroMag(NF,1:111).*(Ener(i,1:111).^2));

                    EnergiaBanda(i, NF)=sum(Energia);
        end

             for p=1:1:100 %bucle para el calculo de los coeficientes
, P es el numero de coeficientes que queremos que nos retorne

                 for NF=1:1:20

                    MelCoe(i,p)=temp+log10(EnergiaBanda(i,
NF)).*cos(p*(NF-0.5).*(3.14/20));
                    temp=MelCoe(i,p);

                 end

                 temp=0;

             end
%Fin analisis trama
m=m+512;
n=n+512;
```

```
    end

end
```

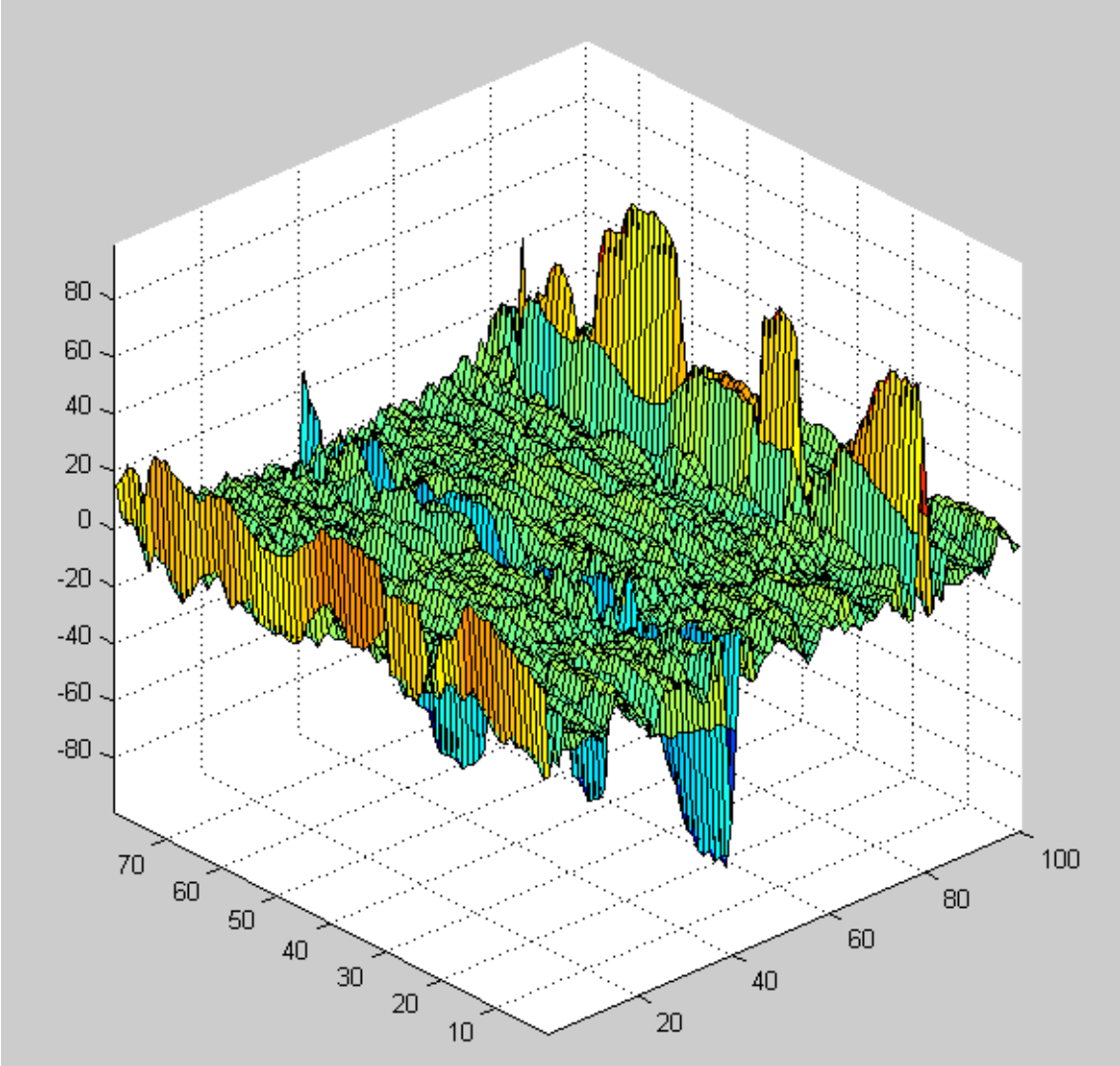


Figura 29: Representación en superficie de la matriz de los MFCC.

5. Conclusiones

Los MFCC muestran ser un algoritmo de extracción de patrones característicos muy eficiente computacionalmente, al igual que los bloques de detección automático de extremos para aislar palabras muestra una segmentación clara desechando las tramas sin información de la señal de entrada con los parámetros fijados en las pruebas experimentales en Matlab. Queda pendiente de revisión de bloques posteriores de normalización del patrón obtenido al igual que el bloque de decisión, el cual se encarga de comparar patrones obtenidos a tiempo real con un librería de patrones predefinidos, o pudiéndose implementar una red neuronal entrenada con dichos patrones para el reconocimiento de las palabras deseadas.

6. Bibliografía y referencias

- Fernando Peralta Reyes / Anibal Cotrina Atencio, Tesis RECONOCEDOR Y ANALIZADOR DE VOZ FACULTAD DE INGENIERIA ELECTRONICA – UNMSM (Universidad Nacional Mayor de San Marcos, Universidad pública, en Lima, Perú), recuperado de https://es.scribd.com/doc/216530642/TesisCap5-01JUL02

- John G. Proakis, Dimitris G. Manolakis. "Tratamiento digital de señales". 3ª Edición. Editorial Prentice Hall. 1998.

- Jesús Bernal Bermúdez, Jesús Bobadilla Sancho, Pedro Gómez Vilda. "Reconocimiento de Voz y Fonética Acústica". Ediciones Alfaomega. 2000.

- Cesar Llamas Bello, Valentín Cardeñoso. “Reconocimiento Automático del Habla. Teoría y Aplicaciones”. Universidad de Valladolid. 1995.

- Andrés Flores Espinoza, “Reconocimiento de Palabras Aisladas en Castellano”, Inictel. Dirección de Investigación y Desarrollo. 1993.

- John P. Cater. "Electronically Hearing: Computer Speech Recognition" 1st Edition. Howard W. Sams & Co., Inc. 1984.

- Freeman J.A., Skapura D.M., “Redes Neuronales, Algoritmos, Aplicaciones y Técnicas de Programación”, ADDISON-WESLEY. 1993.

- C. Crespo Casas, C. de la Torre Munilla, J.C. Torrecilla Merchán. "Detector de extremos para reconocimiento de voz". Telefónica Investigación y Desarrollo. Publicación de Telefónica I+D. S:A. Madrid España 2001.

- M, J. Poza Lara, J. F. Mateos Díaz, J. A. Siles Sánchez. " Design of an isolated-word recognition system for the Spanish Telephone Network". Telefónica Investigación y Desarrollo. Publicación de Telefónica I+D. S:A. Madrid España .2001. EDICIONES ON-LINE : http://www.tid.es/presencia/publicaciones/comsid/esp/home.html

- L. Hernández Gómez, F. J. Caminero Gil, C. de la Torre Munilla, L.

Villarrubia Grande." Estado del arte en Tecnología del Habla". Telefónica Investigación y Desarrollo. Publicación de Telefónica I+D. S:A. Madrid España .2001. EDICIONES ON-LINE : http://www.tid.es/presencia/publicaciones/comsid/esp/home.html

- M. J. Poza Lara, L. Villarrubia Grande, J A. Siles Sánchez. " Teoría y aplicaciones del reconocimiento automático del habla". Telefónica Investigación y Desarrollo. Publicación de Telefónica I+D. S:A. Madrid España .2001. EDICIONES ON-LINE : http://www.tid.es/presencia/publicaciones/comsid/esp/home.html