ALGORITMO DE DETECCIÓN DE INICIO Y FIN DE PALABRA PARA SEÑALES DE VOZ

EUCLIDES ALFONSO RUEDA DIAZ

UNIVERSIDAD INDUSTRIAL DE SANTANDER

FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS

ESCUELA INGENIERÍA DE SISTEMAS E INFORMÁTICA

BUCARAMANGA

2004

ALGORITMO DE DETECCIÓN DE INICIO Y FIN DE PALABRA PARA SEÑALES DE VOZ

EUCLIDES ALFONSO RUEDA DIAZ

Trabajo de grado para optar al título de
Ingeniero de Sistemas

Director
YEZID TORRES MORENO
Doctor en Óptica y Tratamiento de Señales

UNIVERSIDAD INDUSTRIAL DE SANTANDER

FACULTAD DE INGENIERÍAS FÍSICO-MECÁNICAS

ESCUELA INGENIERÍA DE SISTEMAS E INFORMÁTICA

BUCARAMANGA

2004

AGRADECIMIENTOS

El autor expresas agradecimientos a:

YEZID TORRES MORENO, Doctor en óptica y tratamiento de señales y director de la investigación por su orientación, don guía e invaluables concejos.

JUAN MANUEL MONTERO, Dr. Ing. de Telecomunicación y Profesor Titular de Interino en la Universidad Politécnica de Madrid. Por sus observaciones y recomendaciones

CONTENIDOS

Pág.

LISTA DE TABLAS

Pág.

LISTA DE FIGURAS

Pág.

RESUMEN

TITULO:
ALGORITMO DE DETECCIÓN DE INICIO Y FIN DE PALABRA PARA SEÑALES DE VOZ*

AUTOR:
EUCLIDES ALFONSO RUEDA DIAZ**

PALABRAS CLAVE:
VOZ, DETECCIÓN DE VOZ, DETECCIÓN DE PUNTOS DE INICIO Y FIN, PROCESAMIENTO DE LA VOZ.

RESUMEN:

En la actualidad muchas aplicaciones basadas en la voz, son desarrolladas. En estas aplicaciones es necesario conocer donde empieza y donde termina la señal de voz con exactitud "endpoint detection". En aplicaciones como la de reconocimiento de voz es necesario procesar la señal; la cual consiste de segmentos de voz, silencio y otros considerados como ruido.

Se propone un algoritmo para la solución de este problema. Se construyó un algoritmo que cumple con los siguientes requerimientos: Robustez (funcione en ambientes adversos), Baja complejidad computacional (fácil implementación), rápido tiempo de respuesta y, sobre todo exactitud a la hora de encontrar los puntos de inicio y de fin de la voz.

El algoritmo se diseño en tres etapas: La primera etapa viene asociada con el requerimiento de robustez, al ruido a través de la técnica de substracción espectral; en la segunda etapa se mejora la calidad de la señal de voz a través de filtros y la aplicación de otras técnicas; en la tercera y última etapa se encuentran los limites de la señal. Para lograrlo, lo primero que se hace es extraer los parámetros que sirven como discriminantes entre segmentos que tienen voz y los que no. Para esto se utilizan las características de energía y entropía de la señal.

Finalmente, la señal es enviada a una máquina de decisión que se encarga de clasificar los segmentos que contienen voz y los que no. Durante todo el proceso se hace una evaluación de los resultados y se compara el obtenido con los reportados para las técnicas de la energía y de la entropía.*

---

* Trabajo de grado

* *Facultad de ingenierías físico–mecánicas. Escuela de ingeniería de sistemas e informática. Universidad Industrial de Santander. Director Doctor Yezid Torres Moreno

ABSTRACT

TITLE:
ALGORITHM FOR ENDPOINT DETECTION IN SPEECH SIGNALS*

AUTHOR:
EUCLIDES ALFONSO RUEDA DIAZ**

KEY WORDS:
VOICE, SPEECH DETECTION, ENDPOINT DETECTION, SPEECH PROCESSING.

ABSTRACT:

At the present time many applications of the speech are in development. In these applications it is necessary to know where begins and where finishes the signal of speech with exactitude or "endpoint detection". In applications as those of voice recognition, it is necessary to preprocess the signal. The voice signal is composed of speech signal, silence and noise segments.

An algorithm to solve this problem is proposed. The algorithm looks for the following fulfills requirements: Robustness (it works in adverse noises), low complexity (easy computational implementation), fast time of response and mainly, accurate to find the beginning and end points.

The proposed algorithm is designed in three stages: The first stage comes associate with the robustness requirement and use the spectral subtraction technique for noise reduction; in the second stage improve the quality and the SNR ratio of the signal of voice through the filters application and others techniques; in the third one or last stage, the algorithm looks for find the limits of speech. The algoritm extract different parameters to made the speech and non speech discrimation. For this characteristics the energy and entropy of the signal are used.

Finally, the signal is sent to a decision machine to classify between speech and nonspeech. Throughout the algoritm an evaluation process is made for the results and then to compared them with the energy and the entropy algorithm.

---

* Degree project

**Faculty of Physical-Mechanical Engineering. Department of Systems Engineering. Universidad Industrial de Santander. Director: Professor Yezid Torres Moreno

INTRODUCCIÓN

Investigación, desarrollo de productos y nuevas aplicaciones que simplifiquen la interfaz Hombre/máquina se encuentran en desarrollo gracias a una demanda creciente y al rápido incremento de las capacidades de las computadoras en está época. Una de estas interfaces son las conformadas sobre el tratamiento de voz y se ha dicho que las aplicaciones en este campo pueden revolucionar la computación antes de 10 años.

En reconocimiento de voz se necesita procesar señales consistentes en segmentos de voz, silencio y en fondos que contienen ruido; la detección de la presencia de voz empotrada en eventos donde no hay voz o en ambientes de ruido se conoce como “endpoint detection” o detección de inicio y fin de trama. Estas técnicas no son algo nuevo, son técnicas que han sido estudiadas en varias épocas por sus extensas aplicaciones en la resolución de problemas. Las primeras aplicaciones de esta clase de algoritmos se hicieron para las transmisiones telefónicas.
Un algoritmo de búsqueda de inicio y fin de palabra (endpoint) busca cumplir los siguientes requerimientos: Exactitud, robustez, baja complejidad computacional, rápido tiempo de respuesta e implementación simple. Aunque en general diferentes aplicaciones para estos algoritmos definen nuevos requerimientos.

Una característica deseable es la robustez o funcionamiento del algoritmo cuando la señal está contaminada o presenta niveles de ruido, buscándose que el algoritmo funcione o responda a estas condiciones adversas que puedan afectar su desempeño. Se pretende proponer en éste trabajo proporcionarle esta

característica por medio de la utilización de técnicas de remoción de ruido conocidas como técnicas de Substracción Espectral (SS), conociendo sus fortalezas como lo es en particular la facilidad de implementación, pero también teniendo en cuenta sus falencias que serán discutidas más adelante.

En reconocimiento de voz y otros sistemas la detección del inicio y fin trama (endpoint) es algo crucial por varias razones: La primera, los algoritmos que funcionan sobre voz necesitan para realizar sus cálculos y posteriores resultados la intervención de información que no ayude a caracterizar a la señal o de lo contrario se afectaría su respuesta, por ejemplo el popular algoritmo CMS utilizado para el reconocimiento de voz y del hablante. Segundo, removiendo los segmentos que no son de voz cuando estos son un número grande puede dar como resultado reducir el tiempo de computación; también, en sistemas de comunicación en los cuales es necesario enviar los segmentos o tramas pertenecientes a la voz desde entradas de audio constante.

Los algoritmos de inicio y fin de trama (endpoint) comúnmente usados se basan en el uso de la característica principal de energía para la clasificación de los segmentos y posterior localización de los puntos de inicio y de fin, es una característica muy usada debido a la sencillez para su cálculo; sin embargo, cuando la relación SNR disminuye, el simple cálculo de la energía no es suficiente como característica de clasificación sin contar que esta técnica puede ser muy sensibles a artefactos de la voz como pueden ser una respiración o un ruido de los labios, existen otras técnicas que pueden brindar ventajas para la solución de estos problemas por lo que se hará una comparación contra el desempeño de un algoritmo tradicional de energía.

La presentación del proyecto de grado se hará de la siguiente forma: un primer capitulo para revisar algunos conceptos básicos para tener conocimiento del tema y lo aquí propuesto, en un segundo capítulo se expondrán lo diferente algoritmos, una tercera sección para la evaluación del algoritmo y su desempeño.

1. REVISIÓN DE LITERATURA

En este capítulo se harán un resumen de los más básicos y principales conceptos para lograr la compresión del trabajo desarrollado.

1.1 El SONIDO

El sonido se propaga por presión mecánica de las moléculas de aire sobre las moléculas contiguas dando lugar a un movimiento que se transmite (transmisión de energía), en una o múltiples direcciones, de unas moléculas a otras en forma de onda de presión. Cuando en el aire se produce ese tipo de oscilaciones entre 20 y 20000 veces por segundo a un umbral adecuado, nuestro cerebro puede interpretarlas como sonido por medio del oído.

Las ondas pueden verse modificadas por *reflexión*, chocan contra una superficie y rebotan cambiando su dirección inicial (eco o reverberación); *refracción*, cambiar de dirección al pasar de un medio a otro de distinta densidad (Ej: agua-aire) o *difracción*, limitarse cuando encuentran un punto de paso muy estrecho o un obstáculo en su camino.

Las características de las ondas se establecen a partir de un modelo de onda sinusoidal que sería la correspondiente a un tono puro, perfecto; además, el análisis de Fourier permite probar que cualquier otra forma real de onda puede ser considerada como una superposición ponderada de ondas sinusoidales:

- **Longitud de onda**: distancia mínima entre dos puntos que oscilan en fase (Ej: distancia entre dos crestas o entre dos valles consecutivos de una onda).
- **Frecuencia**: número de ciclos que una onda completa en un segundo y se mide en Hertz. Una onda de 1 Hz completa un solo ciclo en cada segundo. De la frecuencia depende el tono, de modo que a mayor frecuencia (más ciclos por segundo) el sonido nos parecerá más agudo y a menor frecuencia (menos ciclos por segundo) sonará más grave.
- **Amplitud**: Máximo desplazamiento respecto del punto de equilibrio que alcanza una partícula en su oscilación. Depende de la cantidad de energía que transporta la onda y está relacionada con la intensidad del sonido. Cuando gritamos estamos aplicando más energía sobre nuestras cuerdas vocales, con ello aumentamos la amplitud de la onda sonora que estamos generando. Existen umbrales para el nivel de detección del sonido y el nivel del dolor, que dependen de la frecuencia.
- **Fase**: la posición que alcanza una partícula que responde a un tono puro con respecto a su posición media. Las partículas en el mismo punto de su ciclo de movimiento se dice que están en fase.

Las medidas características del sonido son: La potencia; densidad de energía por m^3 en una unidad de tiempo que se mide en W: Wats. La sensación: medida de comparación de intensidad entre dos sonidos que se mide en dB: decibells.

1.2 LA VOZ HUMANA

La voz humana se produce voluntariamente por medio del aparato fonatorio. Éste está formado por los pulmones como fuente de energía en la forma de un flujo de aire, la *laringe*, que contiene las *cuerdas vocales*, la *faringe*, las *cavidades oral* (o bucal) y *nasal* y una serie de elementos articulatorios: los *labios*, los *dientes*, el *alvéolo*, el *paladar*, el *velo del paladar* y la *lengua.*

La frecuencia de este sonido depende de varios factores, entre otros del tamaño y la masa de las cuerdas vocales, de la tensión que se les aplique y de la velocidad del flujo del aire proveniente de los pulmones. A mayor tamaño, menor frecuencia de vibración, lo cual explica por qué en los varones, cuya glotis es en promedio mayor que la de las mujeres, la voz es en general más grave. A mayor tensión la frecuencia aumenta, siendo los sonidos más agudos. Así, para lograr emitir sonidos en el registro extremo de la voz es necesario un mayor esfuerzo vocal. También aumenta la frecuencia (a igualdad de las otras condiciones) al crecer la velocidad del flujo de aire, razón por la cual al aumentar la intensidad de emisión se tiende a elevar espontáneamente el tono de voz.

La *articulación* es una modificación principalmente a nivel temporal de los sonidos, y está directamente relacionada con la emisión de los mismos y con los fenómenos transitorios que los acompañan. Está caracterizada por el lugar del tracto vocal en que tiene lugar, por los elementos que intervienen y por el modo en que se produce, factores que dan origen a una clasificación fonética de los sonidos.

1.3 DIGITALIZACIÓN DEL SONIDO

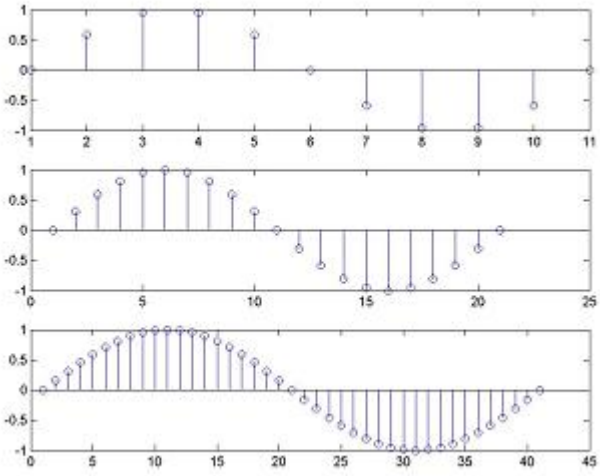
Para que la señal pueda ser tratada en el computador es necesario que esta señal sea digitalizada. Para que una señal pueda ser considerada como digital tiene que cumplir con ser cuantizada y discreta. En nuestro computador la encargada de hacer la digitalización del sonido es la tarjeta de sonido a través de un convertidor A/D.

1.3.1 Discretización

Es el proceso por el cual una señal continua es transformada en una señal discreta. Consiste en tomar una muestra cada cierto intervalo de tiempo. El número de muestras que se capture en un segundo se conoce como frecuencia de muestreo. La frecuencia de muestreo se define teniendo en cuenta el criterio de

Nyquist el cual dice: “mínimo se necesita establecer la frecuencia de muestreo del doble de la frecuencia que se quiere muestrear”.

Figura 1: La Figura de arriba presenta una señal pobremente muestreada, la Figura del medio representa la señal muestreada a la tasa de Nyquist, la de abajo sobre muestreada.



Cuando lo anterior no es tenido en cuenta pueden ocurrir dos casos: Si la frecuencia de muestreo definida es mayor a la dada según el criterio de Nyquist la señal quedará sobre muestreada conteniendo redundancia de información que en nada ayuda a representar a la señal. Pero sí la frecuencia de muestreo es más baja, la señal no será muestreada adecuadamente ocurriendo un fenómeno conocido como distorsión (aliasing) en el cual la señal es falsamente representada.

1.3.2 Cuantificación:

Las computadoras son incapaces de trabajar con muestras con valores continuos en amplitud, por lo tanto la señal es representada mediante una serie finita de niveles. La cuantificación puede se uniforme si hay igual distancia entre los niveles o de lo contrario no uniforme.

Según la localización de los niveles puede ser simétrica si hay igual numero de niveles a cada lado del nivel cero o en otro caso asimétrica.

1.4 ARCHIVOS FORMATO WAV

Formato para archivar datos de audio, diseñado por Microsoft e IBM. Este formato es estándar para Windows y puede ser utilizado en la mayoría de aplicaciones capaces de soportar sonidos y su procesamiento.

El archivo WAV es un subconjunto de Microsoft Riff, que puede incluir muchos tipos diferentes de datos. Estaba originalmente diseñado para archivos multimedia, pero su especificación permitió que fuera útil a otros formatos.

Es un formato muy flexible que puede ser comprimido y grabado en diferentes tamaños y formatos alternos. Aunque los archivos WAV pueden archivar excelente calidad de audio necesitan gran espacio para almacenarla y además lo hace ser ineficiente. Si se quiere obtener un archivo de un bajo tamaño se tendría que reducir la frecuencia de muestreo de la señal lo que llevaría a perder parte de la misma.

Los archivos WAV lo que hacen es almacenar la muestra una tras otra sin ningún tipo de compresión de datos, (a continuación de la cabecera del fichero que es la que contiene la información sobre las especificaciones del sonido ahí almacenado ejemplo: la Frecuencia de muestreo). La sencillez de este formato lo hace ideal para el tratamiento digital del sonido.

1.5 EL RUIDO

Estímulo que acompaña a la señal dificultando la adecuada transmisión, almacenamiento y compresión de la misma. Entendiéndose por señal el estímulo

que lleva una información significativa. El ruido se caracteriza por incrementar el desorden y aumentar la entropía.
Basados en las propiedades del ruido, el ruido puede ser clasificado de las siguientes maneras.

- **Ruido de fondo:** Ruido aditivo, el cual es usualmente no correlacionado con la señal y está presente en varios escenarios ambientales como lo son las oficinas, calles de ciudad, ventiladores, etc... este tipo de ruido es estacionario aunque el ruido en calles e industrias puede ser dinámico.
- **Interferencia de hablantes (voz como ruido):** Ruido aditivo compuesto por la voz de otros hablantes por ejemplo el ruido en una cafetería, un salón de clases, etc. Este ruido tiene características y un rango de frecuencia similares a la señal de voz de interés.
- **Ruido no aditivo:** Ruido debido a la no-linealidad de los micrófonos, distorsión de canales, etc...
- **Ruido correlacionado con la señal:** Ejemplos de este ruido son los ecos.

En general es más dificultoso trabajar con ruido no estacionario, ya que no hay conocimiento a priori de las características del ruido.

1.5.1 Realce de la Voz

Son técnicas para mejorar el desempeño de los sistemas de voz en ambientes de ruido a través de la eliminación de ruido. El realce de voz tiene como metas principales: mejorar la calidad e inteligibilidad de la voz corrupta con ruido; dar robustez contra el ruido a los sistemas como los codificadores y como los de reconocimiento de voz.

Los métodos de realce de voz basados en la estimación de la amplitud espectral en intervalos cortos de tiempo son colectivamente conocidos como métodos

(STSA). Estos métodos funcionan bajo el principio que la señal de voz con ruido está formada por la suma aditiva de la voz y el ruido. Ambas señales se asumen que son procesos no correlacionados y estacionarios en intervalos cortos de tiempo.

Los métodos STSA forman la base de las técnicas comunes de realce de voz encontradas. Se pueden clasificar estas técnicas en dos grupos: El primer grupo incluye métodos basados en convertir segmentos de voz al dominio de la frecuencia, donde el ruido es removido al ajustar las frecuencias ventana a ventana, esto se hace usualmente al sustraer un estimado del ruido calculado durante periodos de pausa de voz, la substracción espectral es uno de estos métodos. El segundo grupo incluye métodos donde la voz con ruido es primero usada para obtener un filtro el cual entonces es aplicado a la voz degradada.

Se analizarán algoritmos de tipo substractivo ya que serán utilizados en el desarrollo del presente trabajo, los métodos se pueden diferenciar por las reglas de supresión, estimación del ruido y otros detalles.

1.5.2 Algoritmos de Tipo Substractivo

Este conjunto de algoritmos forma una categoría que opera en el dominio de la frecuencia. La idea básica de este tipo de algoritmos es obtener la señal limpia a partir de la densidad espectral de energía de la señal con ruido y una estimación del espectro del ruido para luego obtener como resultado un mejoramiento del cociente SNR (relación señal ruido).

Sea Y(n) la señal de voz ruidosa, N(n) el ruido contaminante y S(n) la señal de voz limpia; n es el numero de muestra, n $\in$ Z+. La ecuación siguiente muestra el modelo de ruido aditivo.

$$Y(n) = S(n) + N(n) \qquad \textbf{(1)}$$

Como las señales son asumidas localmente estacionarias el proceso se llevará a cabo de forma localizada utilizando una ventana.

1.5.2.1 Substracción Espectral Lineal

El modelo de la ecuación anterior se puede escribir como:

$$S(n) = Y(n) - N(n) \tag{2}$$

Esta primera versión se le llamó substracción de la magnitud del espectro, pasando al dominio de la frecuencia o dominio de Fouier [17]:

$$|S(w)| = |Y(w)| - |N(w)| \tag{3}$$

En la ecuación anterior se puede apreciar claramente la hipótesis de que la voz y el ruido no están correlacionados. El espectro del ruido no se conoce; el espectro del ruido es un estimado calculado de los periodos donde la voz está ausente, a este estimado del ruido lo llamaremos E [N(n)].

$$|S(w)| = |Y(w)| - E\left[|N(w)|\right] \tag{4}$$

Esta ecuación no garantiza que no existan valores negativos en la voz limpia estimada debido a imprecisiones en la estimación del ruido o a valores muy bajos del nivel de voz. En este caso se puede tratar de dos formas: la primera consiste en hacer positivos estos valores en cuyo caso se habla de una rectificación total de onda y la otra manera consiste en llevar los valores negativos a cero que es el caso más utilizado en la literatura [17] con lo cual se haría una llamada rectificación media de onda.

Una vez se procede a estimar la señal de voz limpia en el dominio frecuencial. La señal de voz en el tiempo es obtenida de acuerdo a:

$$\hat{S(n)} = IDFT\left(\left|S(w)\right| * e^{(j\theta)}\right) \tag{5}$$

Donde θ es la fase de la señal, ya que es difícil una estimación de la fase de la señal de voz limpia a partir de la fase de la señal con ruido; para la reconstrucción se utiliza la fase de la señal contaminada u original; además, desde un punto de vista de la percepción se puede asumir que la fase no lleva información útil para la supresión del ruido. De la ecuación (4), se derivan otras formas de realizar la substracción espectral, una forma generalizada es:

$$\left|S(w)\right|^a = \left|Y(w)\right|^a - E\left[\left|N(w)\right|\right]^a \tag{6}$$

1.5.2.2 Substracción Espectral usando un Factor de Sobresubstracción

Una importante variación para la substracción espectral fue propuesta por Berouti [17] para la reducción de ruido musical, consiste en multiplicar el estimado del ruido por un factor, de ahí el nombre de sobresubstracción. Se puede expresar de la siguiente manera.

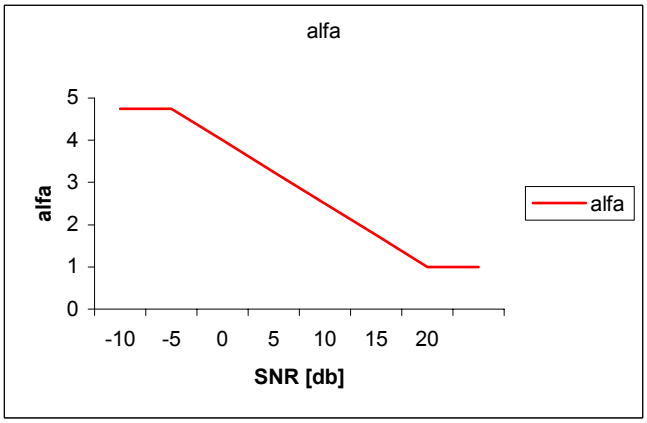
$$\left|S(k)\right|^a = \left|Y(k)\right|^a - \beta\left|E\left[N(k)\right]\right|^a \tag{7}$$

Donde β pertenece a los reales positivos puede ser cualquier constante entre 1 y n, si β=1 se habla de la substracción normal, si **β>1** se habla de sobresubstracción. Para tener mayor certeza a la hora de calcular el factor de sobre substracción β se puede hacer en términos de la relación señal ruido (SNR) que presente la señal, a este nuevo factor se le llama alfa (**α).** α Se calculara de la siguiente manera:

$$\alpha = \alpha_0 - \frac{\alpha_p}{20} SNR \qquad (8)$$

Donde, -5db ≤ SNR ≤ 20, α_0 es el valor deseado de α a 0db de SNR y α_p se ha decidido variarlo para encontrar un valor que proporcione un mayor desempeño llegado el momento de eliminar el ruido.

Figura 2: Valor de alfa a diferentes niveles de SNR.



El factor de sobresubstracción puede ser visto como un factor variable con el tiempo; el cual provee un grado de control sobre la disminución del ruido entre periodos de actualización del ruido.

1.5.2.3 Substracción con Magnitud Selectiva

Esta técnica se basa en que la magnitud resultante de la adición de dos componentes espectrales puede ser mayor o menor que la magnitud de la voz original. El esquema original sólo reduce el ruido correctamente cuando está en fase con la voz o sólo cuando la voz está ausente, cuando la voz esta presente sólo hay un 50% de probabilidad de que el ruido es constructivo resultando en una magnitud mayor, el restante 50% es destructivo resultando en una magnitud menor. Se propone que la magnitud de la substracción puede ser ejecutada sólo

cuando el ruido es aditivo, cuando el ruido es de tipo substractivo la magnitud no se altera [19].

Para determinar cuando el ruido es destructivo o selectivo para un componente frecuencial en particular las características de la voz en periodos cortos de tiempo, son tomadas en consideración. Para un solo componente espectral no es posible determinar si el ruido es constructivo o destructivo, sin embargo si se toman ventanas vecinas en consideración es posible construir una estimación significativa.

Si la información de la voz se asume como estacionaria sobre M ventanas se puede asumir que la magnitud de la voz limpia puede variar poco o seguir relativamente constante. La estimación simple de esta va hacia el uso de la media o mediana de las correspondientes magnitudes sobre las M ventanas vecinas. Esta aproximación es sólo buena si la voz limpia es verdaderamente estacionaria sobre los M ventanas escogidas [19].

$$\begin{aligned} &Sí \quad \left(\hat{S}(k) <= Y(k)\right) \text{ entonces} \\ &\text{decision} = \text{constructiva} \\ &\text{Sí } \left(\hat{S}(k) > Y(k)\right) \quad \text{entonces} \\ &\text{decision} = \text{destructiva} \end{aligned} \tag{9}$$

Este filtro es sólo aplicable cuando la voz esta presente, para periodos de voz donde su energía es insignificante, un esquema de atenuación pura es preferible; esta decisión puede ser obtenida al comparar el estimado del ruido con el estimado de la voz limpia.

$$
\begin{aligned}
&Sí \quad \left(\hat{\mathrm{S}}(\mathrm{k}) <= C * E[N(k)]\right) \\
&\quad \mathrm{E}[S(k)] = \mathrm{Max}(0, Min(\hat{S}(k), Y(k)) - E[N(k)]) \\
&\text{Else } Sí \quad \left(\hat{S}(k) <= Y(k)\right) \\
&\quad E[S(k)] = Max(0, Y(K) - E(N(k)) \\
&Else\ Sí \left(\hat{S}(k) > Y(k)\right) \\
&\quad E[S(k)] = Y(k)
\end{aligned}
\tag{10}
$$

Donde *C* es un factor de sobresubstracción el cual se determinó empíricamente en 2. La fase de la señal con ruido es entonces combinada con el resultado para obtener la voz realzada.

Una mejor aproximación es hacia el uso de un ajuste cuadrático para la estimación de los pequeños cambios en la magnitud.

$$\hat{S}_i(k) = F(Y_{i-m/2}, ..., Y_i, .., Y_{i+m/2}) \tag{11}$$

El grafico muestra la comparación de los resultados obtenido para las diferentes muestras y los resultados obtenidos a procesar la señal.

1.5.2.4 Recuperación Espectral

Se debe hacer un análisis cuantitativo del efecto del ruido aditivo sobre la amplitud del espectro de la señal de voz, para luego derivar un método de recuperación espectral del cual se estima el espectro de la voz limpia de las observaciones del ruido.

Es importante estudiar como la magnitud del espectro es afectado por el ruido. Vamos a considerar W_0 una componente frecuencial, donde la voz es desconocida y el ruido puede ser tomado como una variable aleatoria. El espectro complejo del ruido y la voz puede ser escrito de la siguiente manera.

$$
\begin{aligned}
N(w_0) &= be^{\theta 2} \\
S(w_0) &= ae^{\theta 1} \\
\left|Y(w_0)\right| &= p = \left|ae^{\theta 1} + be^{\theta 2}\right|
\end{aligned}
\tag{12}
$$

Entonces: p puede ser mayor o menor que a, dependiendo de la relación entre las fases de N(*w*) y S(*w*). La expectativa de p puede ser computada con respecto a b y θ_2. Lo primero es asumir que está uniformemente distribuida entre 0 y 2π.

$$
\begin{aligned}
E\{p\} &= \frac{1}{2\pi}\int f(b)\cdot\int_0^{2\pi}\left|ae^{\theta 1}\right| \\
E\{p\} &= \frac{1}{2\pi}\int f(b)\cdot\int_0^{2\pi}\sqrt{a^2+b^2+2abCos\,\theta}\cdot d\theta\cdot db \qquad \theta = \theta_2 - \theta_1
\end{aligned}
\tag{13}
$$

Donde *f(b)* es la función de distribución de probabilidad de b. Sí se usa la aproximación: $\mathrm{E}(\mathrm{F(b)}) = \mathrm{F}(\mathrm{E(B)}) = \mathrm{F}(\bar{\mathrm{b}}), \text{ donde } \mathrm{F(b)} \text{ es la integral con respecto de } \theta.$

La exactitud de esta aproximación depende de la distribución de b y los valores de Q(r) alrededor de b.

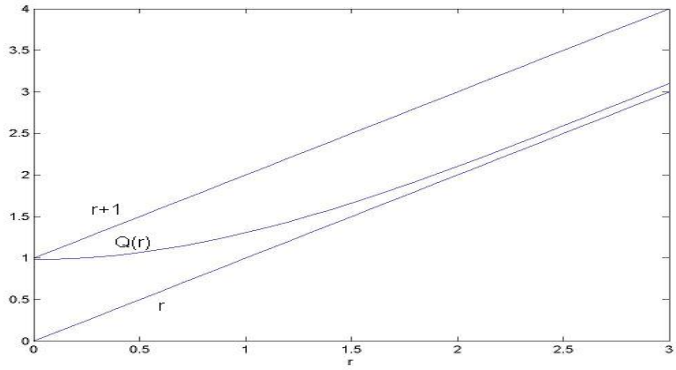
$$
\begin{aligned}
E\{p\} &= \frac{1}{2\pi}\int_0^{2\pi}\sqrt{a^2+\bar{b}^2+2a\bar{b}\,Cos(\theta)}\cdot d\theta \\
&= \frac{\bar{b}}{2\pi}\int_0^{2\pi}\sqrt{1+r^2+2rCos(\theta)}\cdot d\theta \\
&= \bar{b}\cdot Q(r)
\end{aligned}
\tag{14}
$$

Donde:

$$
\begin{aligned}
r &= a/b \\
Q(r) &= \frac{1}{2\pi}\int_0^{2\pi}\sqrt{1+r^2+2rCos(\theta)}\cdot d\theta
\end{aligned}
$$

Se derivan soluciones numéricas para *Q(r)* que se encuentran entre *r+1* y *r*, en el caso que *Q(r)= r-1* entonces *E{p}= a+b* que es la relación de la substracción espectral lineal. Si r es grande y a es mucho mayor que b entonces *Q(r) = r*.

Figura 3: Soluciones numéricas de Q(r) en el rango de 0 a 3



Si se conoce E {p} y b podemos recuperar la amplitud de la señal a ,de la siguiente manera, sí r>2 entonces Q(r) =r, o, sí 0<r<2 se aproxima Q(r) a T(r); donde T(r) es un polinomio de tercer grado como se muestra a continuación [18]:

$$Q(r) \approx {E\{p\}}/{b} \approx {p}/{\overline{b}} \approx T(r) = 0.9820 - 0.0075r + 0.3761r^2 - 0.0461r^3 \qquad (15)$$

Un límite inferior para r es dado en 0.25, ya que para r menores la señal de ruido es mucho más alta que la señal observada y la estimación de la voz no es precisa.

1.6 Técnicas para el Análisis de la Voz

1.6.1 Filtrado

El filtrado es una operación básica usada para el tratamiento de la voz. Filtrar es la operación matemática de la convolución de la transformada de Fourier del filtro con una secuencia para producir una señal. Para la implementación de un filtro digital se tiene en cuenta la siguiente expresión:

$$y(k) = \sum h(n) * x(k) \tag{16}$$

Donde h(n) son los pesos, x(k) son la muestras, y(k) son las muestras filtradas. Entonces el filtrado puede ser especificado como una secuencia en el dominio del tiempo por la respuesta percusional del filtro.

1.6.2 Transformada de Fourier

Representa una señal en una base de exponenciales complejas. La transformada de Fourier lleva una señal representada en el tiempo a su representación en frecuencias; facilitando algunos procesos y visualizaciones de la señal que son inherentemente orientados al dominio frecuencial. Se define como se muestra en la siguiente ecuación:

$$\begin{aligned} S(w) &= \sum_{n=-\infty}^{n=\infty} s(n) e^{-jwn} \\ s(n) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} S(w) e^{jwn} \end{aligned} \tag{17}$$

Donde s(n) es la representación de una secuencia en el tiempo y S(w) la respuesta frecuencial de s(n).hay dos cantidades para la representación de S(w); la primera es la magnitud que es la longitud del vector, la otra es la fase que es el ángulo que forma el vector con las abcisas, en una interpretación sencilla.

1.6.3 Segmentación y Enventanado de la Señal

La segmentación de las señales de voz consiste en tomar una señal y dividirla en n señales de menor longitud con el objetivo de realizar análisis localizados. Cuando el número de muestras (n) es pequeño la señal puede ser considerada estacionaria en ese intervalo.

Enventanar la señal consiste en multiplicar una señal por unos pesos; esos pesos son dados por una función. Al conjunto de estos pesos se le llamara ventana;

estos pesos se hallan de acuerdo a una función. La multiplicación de una secuencia en el tiempo s(n) con una ventana en el dominio del tiempo w(n) es equivalente a la convolución de S(k) y W(k) en el dominio de la frecuencia (Teorema de la convolución).

Las ventanas más comúnmente usadas son la Hamming y la Hanning, cuya función para el cálculo de pesos es la siguiente:

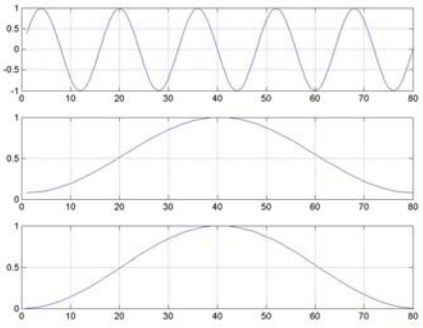
Tabla 1: Función de pesos para las ventanas de Hanning y de Hamming

$$\text{Hamming}: \quad w(n) = 0.54 - 0.46\cos(\frac{2\pi}{N})$$

$$\textit{Hanning}: \quad w(n) = 0.54 - 0.5\cos(\frac{2\pi}{N})$$

La respuesta de la ventana de Hanning, de Hamming para una señal es mostrada en la siguiente figura:

Figura 4: Señal con ventana de Hamming (medio) y de Hanning (inferior).



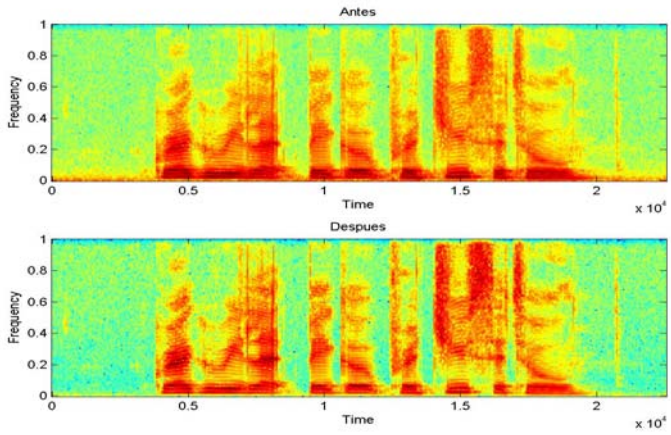
1.6.4 Preénfasis

El preénfasis es el procesamiento de la señal para hacerla menos susceptible a truncamientos y para aplanarla espectralmente; efectos causados por los pulsos glotales los cuales son afectados por los filtros del tracto vocal enfatizando algunas de las componentes, Para esto se pasa la señal de voz a través de un filtro digital pasa alto. Este filtro puede tener coeficientes fijos o ser adaptativo. El preénfasis se define de la siguiente manera, S(n) es la señal de voz de entrada:

$$\overline{S}(n) = S(n) - \overline{a} \cdot \overline{S}(n)$$
$$0 \le \overline{a} \le 1$$
$$\text{En el dominio de la transformada Z} \qquad (18)$$
$$H(z) = 1 - \overline{a} \cdot z^{-1}$$

El efecto sobre la señal después que es pasada por un filtro de preénfasis es incrementar la relevancia de las componentes de alta frecuencia. En la figura pueden verse una señal antes (superior) y después de usarse el filtro de preénfasis sobre ella.

Figura 5: Preénfasis realizado sobre una señal de voz.



1.7 Algoritmos de Detección de Inicio y Fin de Palabra para Señales de Voz

En el procesamiento de voz un problema importante es la detección de voz cuando está se encuentra empotrada en una señal sobre la cual hay varios eventos que no son voz o en ambientes de ruido. La solución para este problema se desarrolla bajo el nombre de algoritmos de detección de inicio y fin de palabra (EndPoint); es un problema que ha sido tratado por varias décadas teniendo sus primeras aplicaciones en la telefonía, aplicaciones desarrolladas por los laboratorios Bell.

Estos algoritmos han ganado importancia debido al desarrollo de aplicaciones derivadas del tratamiento digital de la voz, en estos sistemas de procesamiento de voz la inexactitud a la hora de decidir los límites de la señal es la mayor causa de error. Además, el tiempo de procesamiento y computación es mínimo cuando los puntos de inicio y fin de la señal son localizados con exactitud.

1.7.1 Requerimientos

Desde hace varios años estos algoritmos de detección han sido desarrollados para diferentes aplicaciones como la cancelación de eco, codificación de voz, verificación del hablante y otras aplicaciones. Cada una de estas aplicaciones necesita diferentes algoritmos que cumplan requerimientos específicos en términos de:

- Exactitud: Es tal vez el requerimiento más importante de los algoritmos, el detectar exactamente los puntos de inicio y fin, no es admisible para un detector descartar segmentos que contengan voz, como lo pueden ser las señales con fonemas de baja energía.
- El algoritmo tiene que funcionar en condiciones adversas como lo son las variaciones de la relación señal ruido (SNR).
- Baja complejidad computacional: es un requerimiento deseable en los algoritmos cuando forman partes de otros sistemas.
- Rápido tiempo de respuesta: Obtener una detección confiable pero rápida en el menor tiempo posible. Es un requerimiento solo posible si el algoritmo no es complejo.
- Adaptativo: El algoritmo pueda ser adaptado a otros sistemas y a cambios en el ambiente, especialmente en señales con fondos con ruido que varia.

Estos requerimientos servirán como descriptores del desempeño del algoritmo a la hora de evaluarlo.

1.7.2 Características Utilizadas para la Clasificación de la Señal

Las características utilizadas para la clasificación de voz de la señal son útiles a la hora de hacer diferencia entre los segmentos que contienen voz y aquellos que no. Dependiendo de la característica se le puede hacer más fácil a la hora de tomar una decisión. Se puede decir que la características son tomadas en intervalos cortos de tiempo donde la señal pueda ser considerada estacionaría.

1.7.2.1 Energía

Es tal vez la característica más comúnmente usada para la construcción de algoritmos de este tipo, debido a la sencillez para su cálculo, además que puede ser fácilmente adaptado a otra aplicación o ambiente. A continuación se muestra la definición de energía:

$$Energia = \sum_{k=1}^{n} S(n)_k^2 \tag{19}$$

La energía se toma como característica teniendo en cuenta que los fonemas sordos contienen más energía que los segmentos de silencio. La observación no es valida cuando los niveles de SNR disminuyen, la energía no es suficiente como medio de clasificación.

La señal de voz es una señal limitada en banda, aproximadamente entre 20 Hz y 20 Khz (región de audición para un ser humano normal). Sin embargo, la mayor parte de la energía se concentra por debajo de 2 Khz.

Se pueden utilizar variaciones en el cálculo de la energía con el fin de mejorar el desempeño; tales como el cálculo de la energía en [Db] o representaciones en escalas logarítmicas. En la ecuación siguiente se observa una variación en la forma de hallar la energía [14]:

$$E_i = 10\log_{10}\sum_{k=1}^{N} S(k)^2 \qquad (20)$$

Otra variación a la energía es la hecha por Teager [16], en la cual después de que la señal es segmentada es calculado su espectro y la energía es calculada para cada segmento como la raíz cuadrada de la suma del producto de la energía de las componentes frecuenciales por un peso.

$$TE_i = \left(\sum_{k=1}^{N} S(k)^2 * W(k)\right)^{1/2} \qquad (21)$$

1.7.2.2 Entropía

Desde la construcción de los primeros algoritmos se han buscado técnicas para reemplazar y mejorar las falencias dejadas por otra técnica de modo que se mejore el desempeño del algoritmo, una de estas es calcular la entropía de la información que representa el valor medio de las informaciones que pueden proporcionar los resultados posibles de las variables aleatorias dadas por la señal de voz. La cual se define como:

$$H(\varepsilon) = H[P(x_1), P(x_2), ...P(x_n)] = -\sum_{i=1}^{n} P(x_i)\cdot Log_2 P(x_i) \qquad (22)$$

En la entropía la base de los logaritmos es la binaria, aunque habitualmente no se indicará salvo que se indique otro proceder, según el logaritmo utilizado serán las unidades de la entropía. Puede ser pasado de una a otra base utilizando las propiedades de lo algoritmos. En esta definición las P(x) son diferentes probabilidades para los posibles eventos que al sumarse tienen que dar 1.
Según la definición en (22), la entropía cumple con las siguientes propiedades:

- La entropía siempre es mayor o igual que cero.
- Si todos los eventos son equiprobables la entropía es igual al Log(n).

- Manteniendo la equiprobabilidad la entropía aumenta con el número de datos.
- La entropía de una variable que toma n valores es máxima cuando los resultados son equiprobables.

Estas propiedades se utilizan para construir el algoritmo basado en la entropía.
La entropía puede ser utilizada tanto en el dominio del tiempo como en el dominio de la frecuencia, lo fundamental de estos algoritmos es buscar una función de probabilidad, para cada uno de los segmentos de la señal enventanada o segmentada.

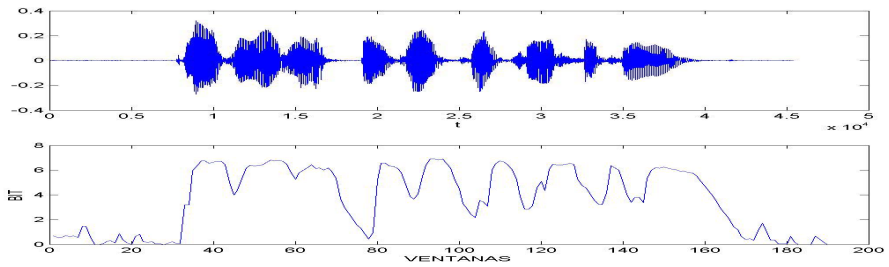
Para determinar una función de probabilidad dentro de cada "frame" se pueden utilizar diferentes técnicas, una técnica útil para calcularla es utilizar un histograma con n clases para luego ser normalizado y obtener las probabilidades, donde este n es utilizado de forma que no afecte la sensibilidad pero tampoco aumente la carga computacional; si el n es muy pequeño tenderá a un agrupamiento alrededor de algunos de los valores lo que significaría tener valores similares de entropía en todas las ventanas; si por lo contrario el n adecuado es muy grande la sola realización del histograma podría aumentar la carga computacional afectando el desempeño del sistema [12]

Otra técnica consiste en hallar la probabilidad utilizando la energía ya sea en el en la frecuencia, dividiendo el valor de la energía para una muestra i sobre el total de la energía de la ventana. Como se muestra a continuación:

$$P_i = \frac{E(m_i)}{\sum_{k=1}^{N} E(m_k)}, \quad i = 1...N \tag{23}$$

Esta forma de hallar la probabilidad es utilizada en el dominio de la frecuencia; si es combinada con otras técnicas resulta en un mejoramiento del desempeño del sistema.

Figura 6: Perfil de la entropía para una señal de voz



1.7.3 Máquina de Decisión

Es la parte del algoritmo que dependiendo de los resultados obtenidos en la extracción localizada de parámetros toma una decisión de cuales segmentos pertenecen a voz; en otras palabras se encarga de hacer una clasificación para luego producir como resultado los puntos de inicio y fin de la señal de voz.

Las máquinas de decisión básicamente están conformadas por la definición de umbrales de corte, estos umbrales pueden ser desde un simple promedio hasta cálculos más complicados, cada máquina de decisión puede contar con uno o varios umbrales dependiendo de quien diseñe el algoritmo.

Otra parte importante de la máquina de decisión es la definición de constantes como MUD (mínima distancia de pronunciación) y MUS (mínima separación entre pronunciaciones). Las cuales ayudan a reconocer cuales segmentos pertenecen a palabras. La definición del MUD y del MUS depende del idioma para el cual haya sido diseñado, aunque no varié mucho entre uno y otro.

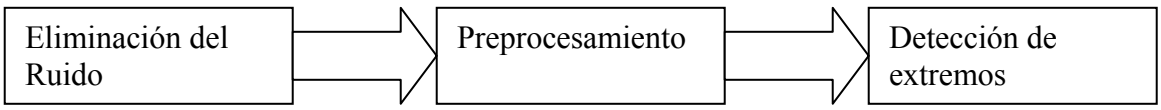
2. ALGORITMOS

En este capítulo se verán las diferentes secuencias que dan lugar cada una de las estructuras que conforman el algoritmo como lo son la eliminación de ruido, la parte de preprocesamiento y el detector de extremos. Los prototipos de los algoritmos se hicieron en MATLAB por las ventajas que nos ofrece para la construcción y evaluación del desempeño del algoritmo.

2.1 ETAPAS DEL ALGORITMO

El algoritmo definitivo se puede dividir en tres etapas principales, cada una de estas etapas cumple tareas específicas que ayudan a la siguiente a mejorar su desempeño y por lo tanto el sistema, como se puede ver a continuación:

Figura 7: Etapas del algoritmo



En la primera etapa se procesa la señal con el fin de disminuir los niveles de ruido haciendo más robusto el algoritmo contra diferentes niveles de SNR. En la segunda etapa llamada preprocesamiento lo que se intenta es mejorar la señal a través de la aplicación de filtros y otros métodos. En la última etapa se extraen las características de la señal y se procede a clasificar lo segmentos como segmentos de voz/ no voz para posteriormente definir los extremos.

2.2 ALGORITMOS FUNDAMENTALES

En esta sección se tratan los algoritmos que se pueden encontrar en una o más etapas, por ser fundamentales para el procesamiento digital de la señal de voz.

2.2.1 Algoritmo para el Cálculo Rápido de la Transformada de Fourier

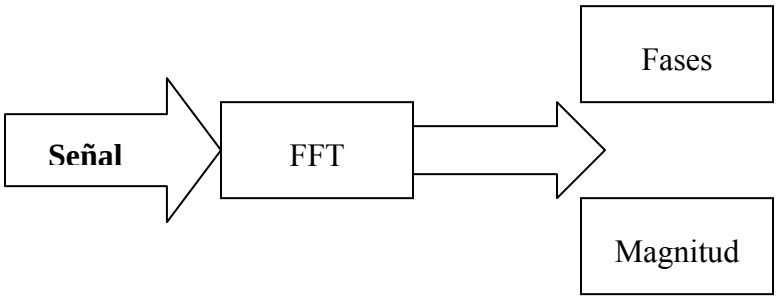
El algoritmo FFT (Fast Fourier Transform) lo único que busca es resolver de la manera más eficiente posible la siguiente expresión discreta:

$$x[n] = \frac{1}{N}\sum_{n=0}^{N-1} X[n] \cdot e^{-jk(\frac{2\pi}{N})n} \qquad (24)$$

La evaluación directa de esta sumatoria implica N^2 multiplicaciones y N(N-1) adiciones. Haciendo una serie de reordenaciones, conseguiremos con la FFT reducirlo a $N*Log_2(N)$ operaciones. El problema se reduce al cálculo de dos FFTs de tamaño N/2 y así sucesivamente hasta hacer la FFT entre dos muestras [1].

El algoritmo para la FFT se beneficia de las propiedades de simetría de la exponencial compleja discreta en el tiempo para reducir el número de multiplicaciones. Para evaluar una transformada discreta de Fourier con N muestras el algoritmo de la FFT encuentra su máxima eficiencia cuando N es una potencia de 2. Esta restricción no afecta el uso práctico de la FFT ya que la longitud de *h(n)* puede ser incrementada a la siguiente potencia de 2 aumentando el número de muestras en un adecuado número de ceros, sin embargo es necesario tener en mente que esto modifica el espectro de la señal original.

Figura 8: Proceso FFT

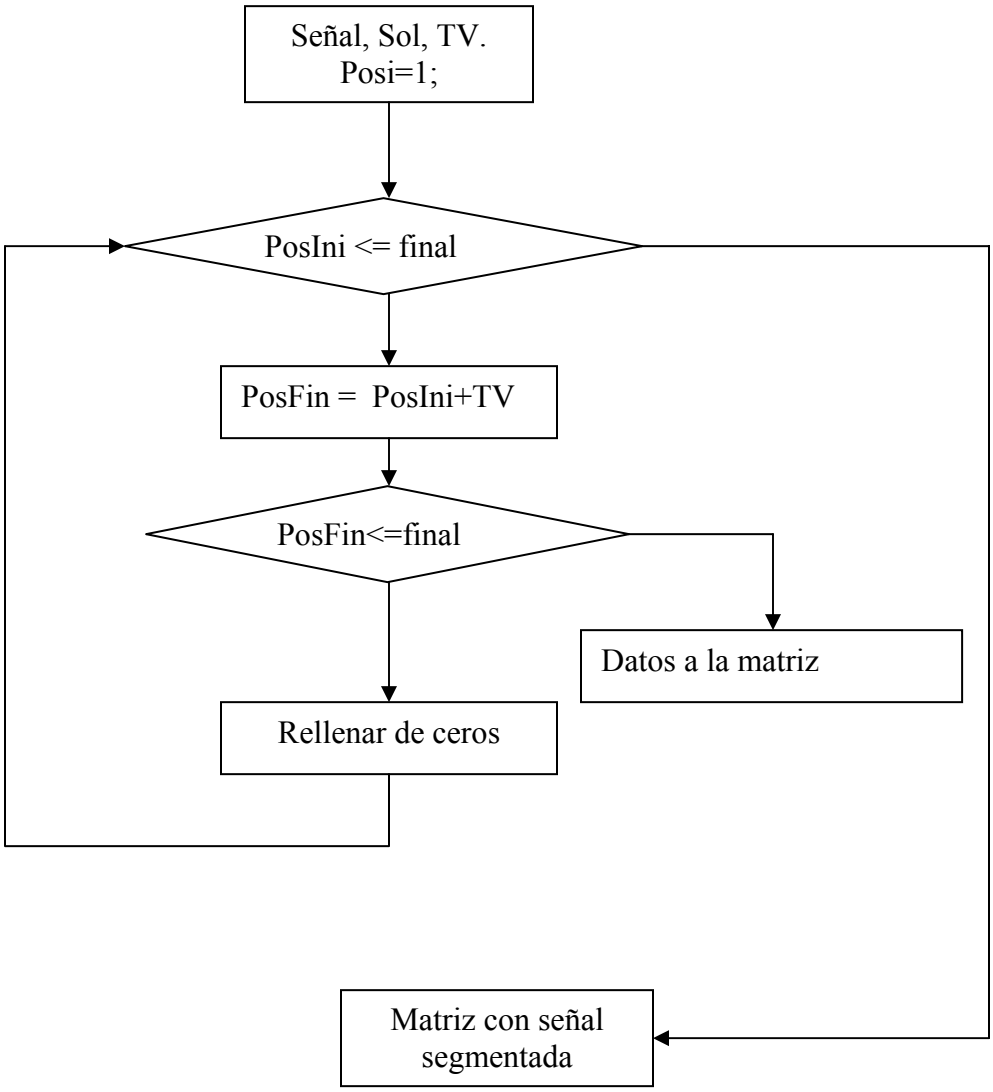


El algoritmo que se utilizó fue el que provee Matlab. Al final de pasar por el algoritmo de la FFT se obtiene una transformada compleja cuya parte real y parte imaginaria permiten obtener la magnitud y fase de la señal. La Señal puede ser reconstruida con un proceso inverso conocido como la IFFT o transformada inversa.

2.2.2 Algoritmo de Segmentación de la Señal de Voz

El algoritmo de segmentación de voz toma una señal de voz de n muestras y la divide en n' segmentos con una longitud definida por un tamaño de ventana; los cuales se almacenan en una matriz en donde cada columna es un segmento de la señal; por lo tanto el tamaño de la matriz será de [tv] X [n'] donde tv es el tamaño de la ventana que se calcula $t*F_s$ donde t es tiempo y F_s la frecuencia con la cual ha sido muestreada la señal. Si la longitud de la señal no es múltiplo de tv se hace un relleno con ceros, al final del algoritmo se obtendrá una matriz donde cada columna pertenece a un segmento de la señal original.

Figura 9: Segmentación de voz



2.3 Etapa de Eliminación de Ruido

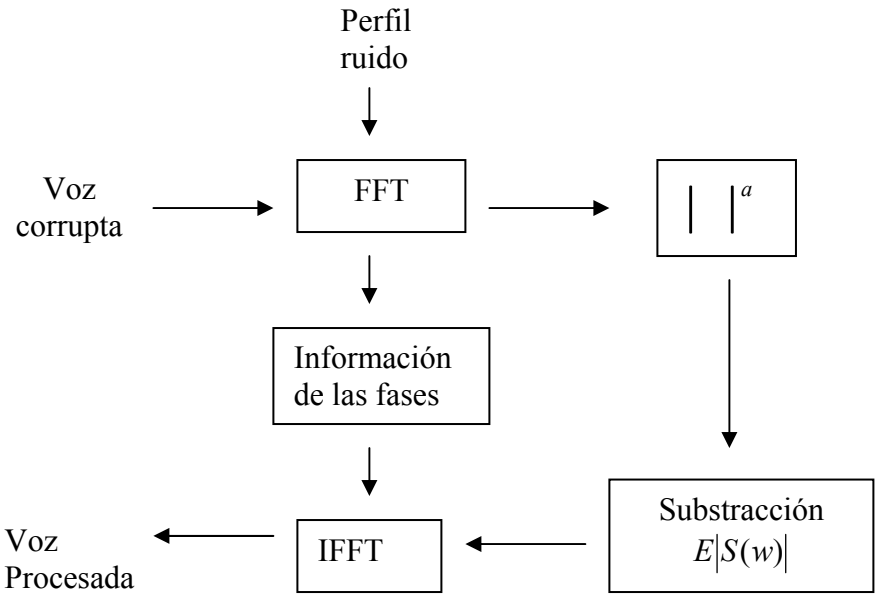
En esta fase se utilizan tres algoritmos derivados de la substracción espectral, luego en la parte de evaluación se mostrará cual ofrece un mejor desempeño para que pueda ser utilizado luego en el algoritmo final de búsqueda de puntos de inicio y fin de palabra, los algoritmos utilizados en esta etapa son: substracción

espectral, la substracción espectral con factor de sobre substracción, magnitud selectiva y recuperación espectral.

2.3.1 Substracción Espectral

A continuación se muestra el procedimiento de la substracción espectral, el primer paso es transformar tanto la señal de voz corrupta como el perfil de ruido o el estimado de ruido al dominio de la frecuencia aplicándole una FFT, entonces se obtienen magnitudes y fases de ambas señales; Con las magnitudes de las señales se realiza una substracción básica con la cual se obtiene un estimado de la amplitud de la señal de voz, luego esta amplitud es combinada con las fases para aplicarle una IFFT y obtener la señal de voz procesada.

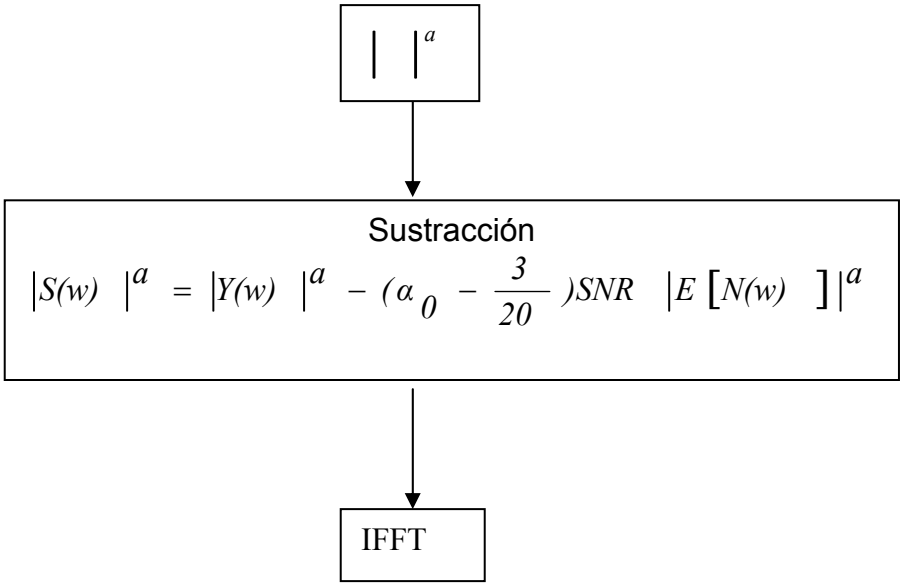
Figura 10: Diagrama Substracción espectral básica



2.3.2 Substracción Espectral con Factor de Sobresubstracción

Como se mencionó en el capitulo 1, en la sección 1.5.2.2, esta técnica es una derivación de la substracción espectral, la única diferencia con la primera es que al momento de hacer la substracción la magnitud del perfil de ruido es multiplicado por un factor, el cual varia en relación a la SNR de la señal aumentando o disminuyendo la magnitud de la substracción a la hora de hallar la magnitud de la voz con realce. El bloque de substracción se reemplazaría por el siguiente:

Figura 11: Substracción espectral con factor de sobresubstracción

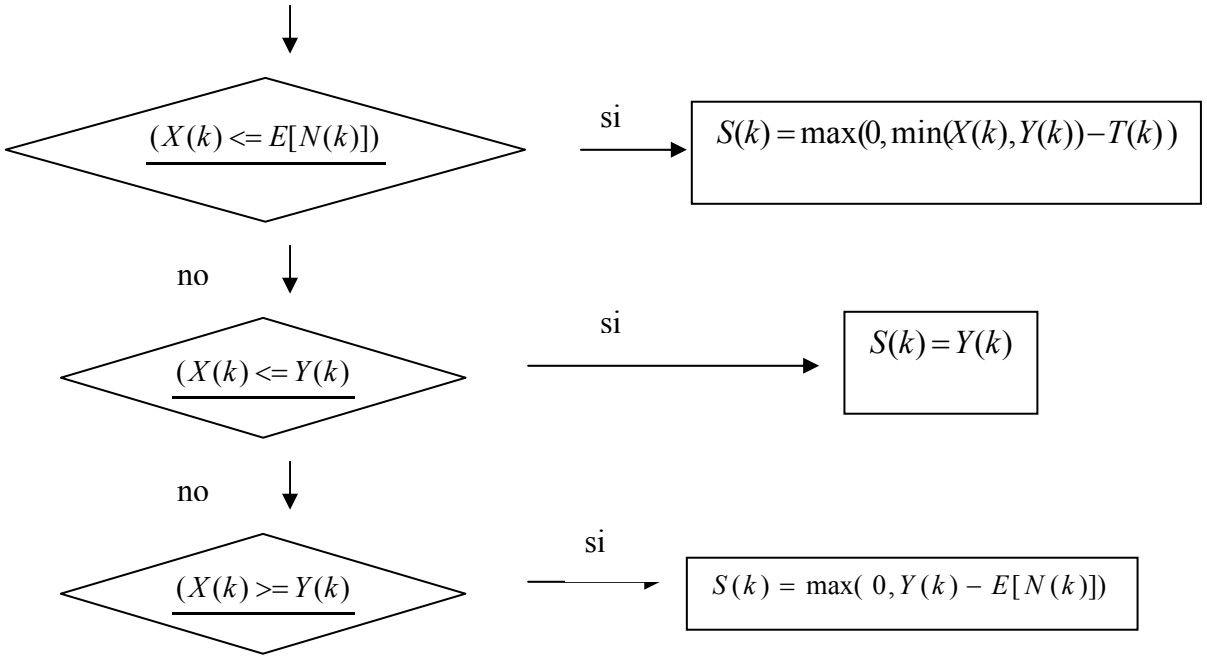


2.3.3 Substracción Espectral con Magnitud Selectiva

En está técnica la magnitud de la señal de voz realzada se obtiene comparando la magnitud de la señal de voz estimada X(n) (se halla realizando un ajuste cuadrático sobre M ventanas) con la magnitud del ruido, sí X(n) es menor que N entonces puede ser considerada como formada en su mayoría por ruido, en caso

contrario, sí la magnitud de la señal estimada Y(n) es mayor que la estimada se realiza la substracción con una reconstrucción media de onda, pero sí es menor la magnitud de la voz es igual a la magnitud de la señal corrupta. Después de tener la magnitud se sigue con el paso siguiente calculando la transformada inversa con la magnitud obtenida y con la fase de la señal corrupta.

Figura 12: Diagrama decisión magnitud selectiva



2.3.4 Recuperación Espectral

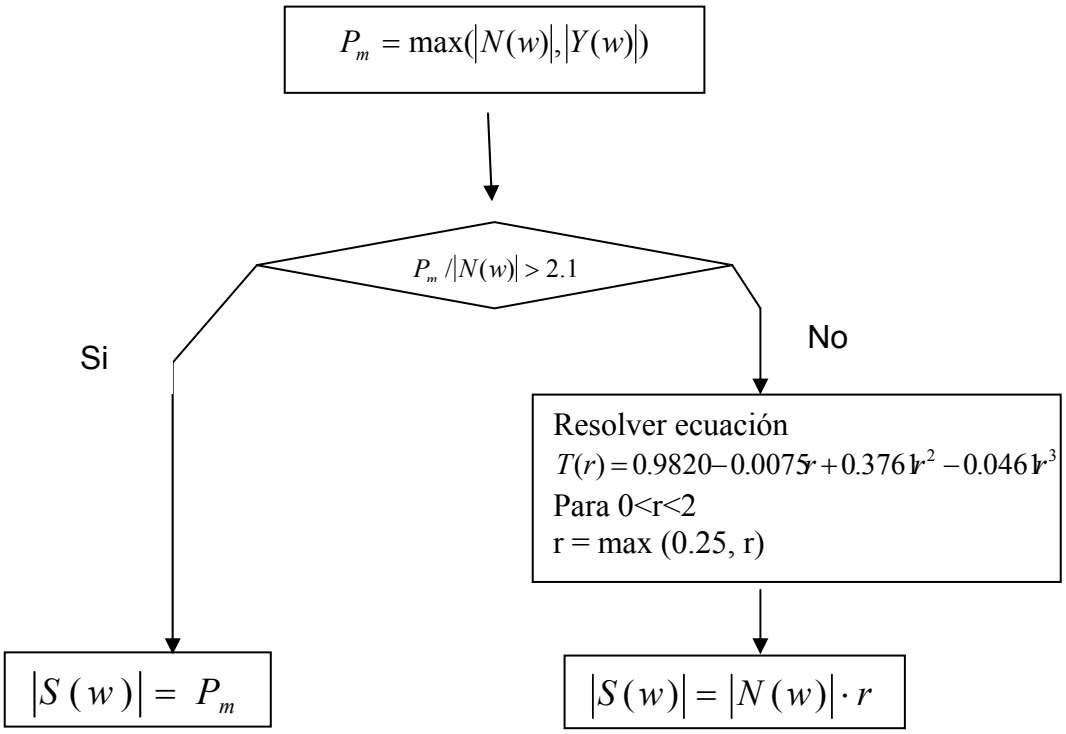
La recuperación espectral es en parte similar a la magnitud selectiva con la particularidad que la magnitud se calcula en base a unas observaciones realizadas y a unas aproximaciones hechas.

El primer paso es evitar que la amplitud de la señal corrupta sea más pequeña que la señal de ruido a este valor lo llamaremos amplitud observada, luego el parámetro r se estima dividiendo a la amplitud observada sobre la amplitud del ruido para ver si r está en el rango de 0 a 2, sí es de esta manera, se soluciona la

ecuación dada en (15) pero, sí r está fuera del rango, entonces la magnitud de la señal con realce es igual a la amplitud observada.

Para mejorar el desempeño del algoritmo a la hora de solucionar la ecuación se utilizó una tabla donde estaban las soluciones; las cuales se habían hallado previamente con métodos numéricos.

Figura 13: Diagrama recuperación espectral.



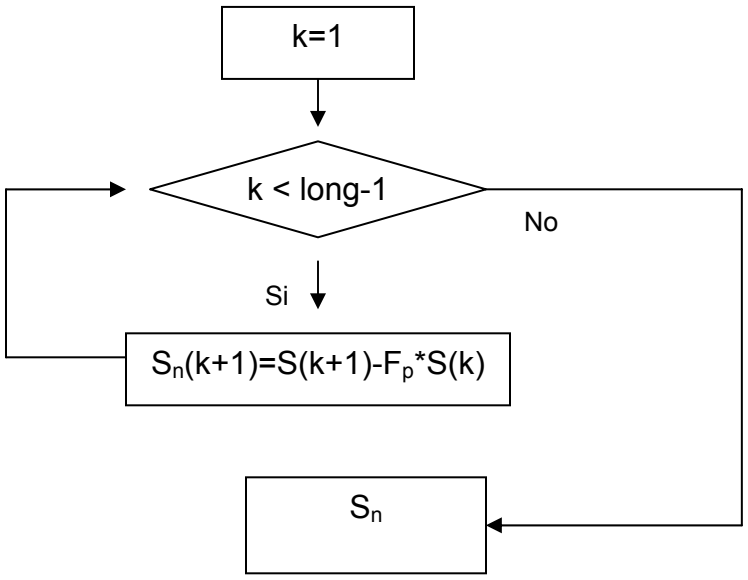
2.4 ETAPA DE PREPROCESAMIENTO

En esta etapa se agrupan los algoritmos que mejoran la calidad de voz de la señal como son los algoritmos de preénfasis y el filtro pasa banda.

2.4.1 Preénfasis

A continuación se muestra el diagrama de flujo para el algoritmo de preénfasis, donde long se refiere al número de muestras de la señal o segmento de la señal a aplicar el preénfasis.

Figura 14: Diagrama Preénfasis



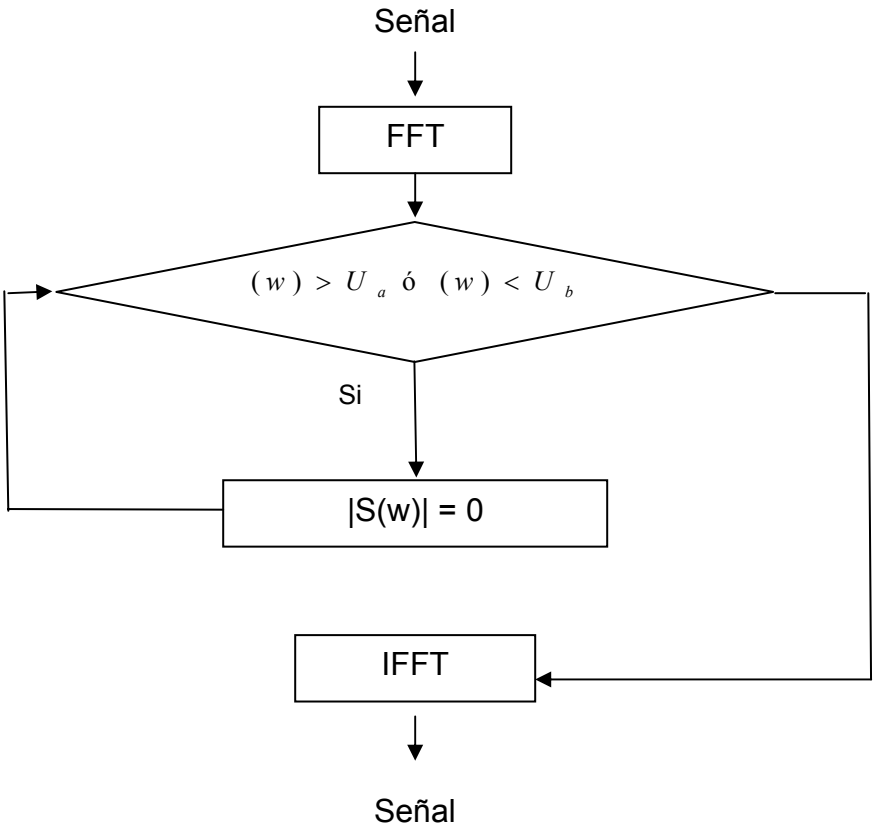
Se inicializa un contador k en 1 el cual va a ser el encargado de recorrer la señal hasta su penúltima posición, mientras recorre va obteniendo una nueva señal restándole a la componente que le sigue a k un porcentaje de la anterior, valor que se llamará factor de preénfasis.

2.4.2 Filtro Pasa Banda

La función del filtro pasa banda es remover todas aquellas componentes frecuenciales que no pueden ser consideradas como voz y que no pudieron ser removidas en el proceso de eliminación de ruido. La señal es llevada al dominio de la frecuencia al aplicar una FFT luego se definen dos umbrales denotados como umbral alto U_a y umbral bajo U_b, que son los valores para dos componentes

frecuenciales extremas, los valores por encima de este valor o por debajo de estos valores respectivamente son eliminados; luego la señal es devuelta al dominio del tiempo utilizando una transformada inversa.

Figura 15: Diagrama Filtro Pasa-Banda



2.5 DETECCIÓN DE EXTREMOS

En esta etapa se presentan los procedimientos encargados de extraer las características que ayudaran a encontrar el principio y el final de las señales de voz. En ambos casos se parte de que la señal llega segmentada.

2.5.1 Extracción de la Característica de Energía

La extracción de la característica de la energía se puede hacer tanto en el dominio del tiempo como en el dominio de la frecuencia. Después de que la señal ha pasado por las dos etapas anteriores, sí la energía va a ser calculada a partir del espectro como se hizo en nuestro caso sólo se tienen en cuenta ciertas regiones.

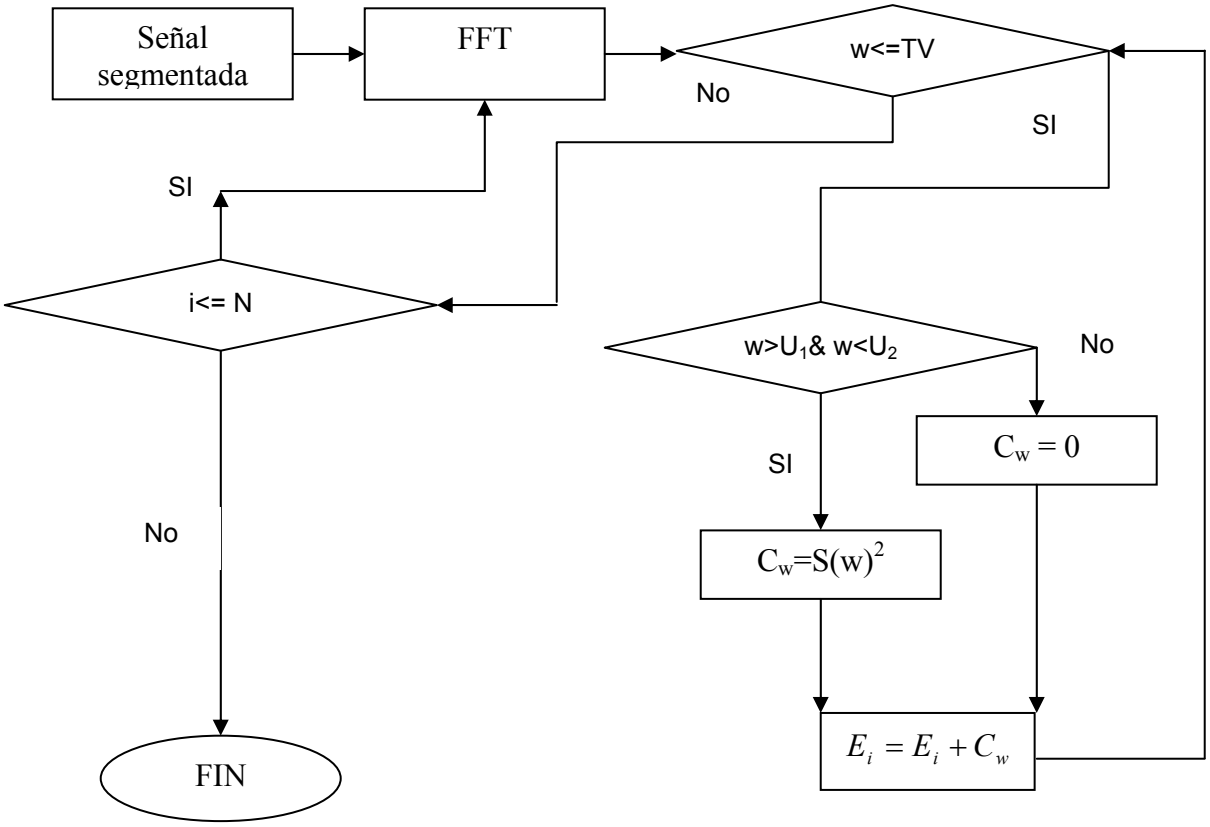


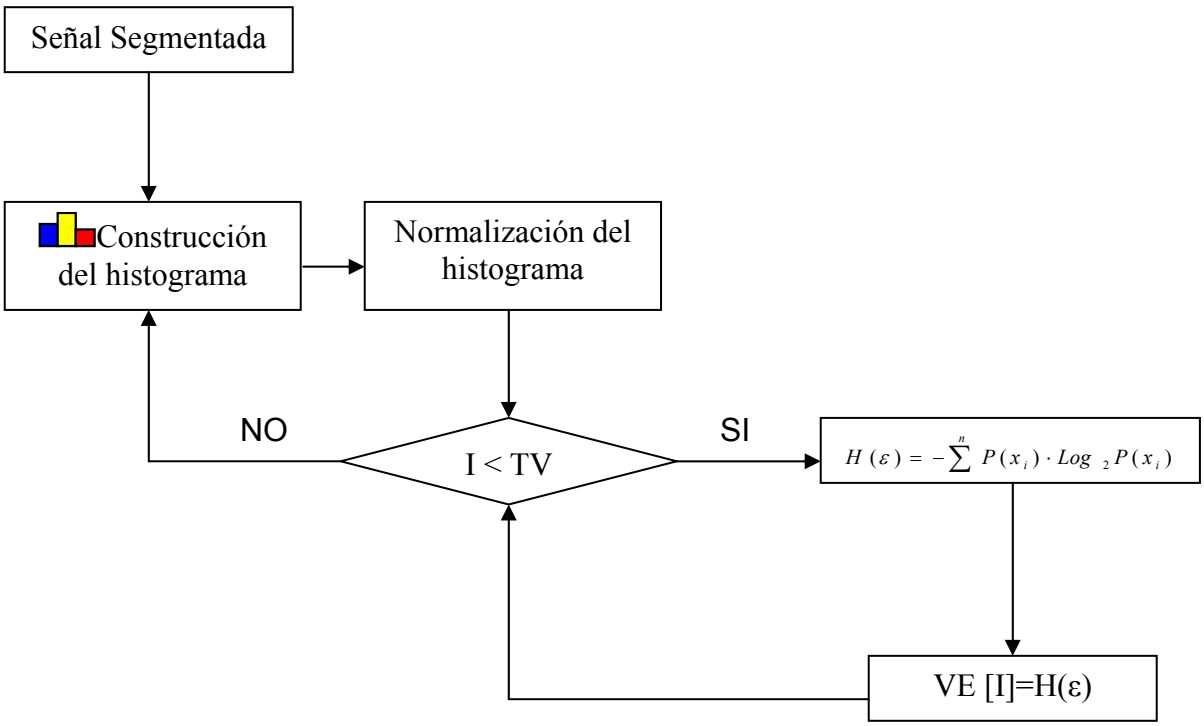
Figura 16: Algoritmo energía

Se Llamará a N el numero total de segmentos en los cuales ha sido segmentada la señal y a tv el numero de muestras por segmento. Arriba se muestra el diagrama para el cálculo de la energía.

2.5.2 Extracción de la Característica de Entropía

La aplicación del concepto de entropía a la voz se debe a que la señal de voz está más organizada durante los segmentos de voz que durante los segmentos de ruido.

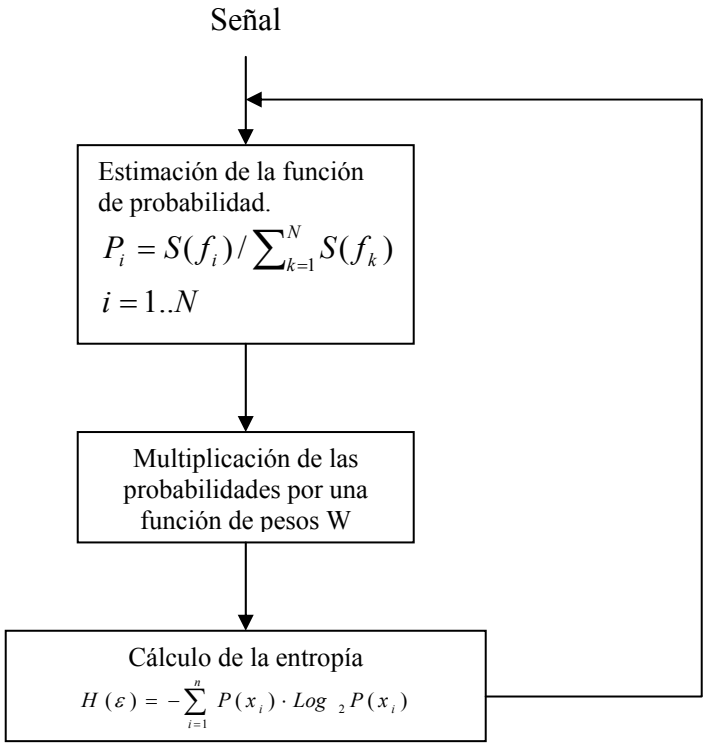
Figura 17: Algoritmo entropía



El cómputo de la entropía es llevada directamente en el tiempo, la señal es segmentada en ventanas de longitud de 25 ms y pasada por una etapa de preproceso. El paso a seguir es hallar una función de distribución de probabilidad dentro de cada ventana lo cual se hace a través de la construcción de un histograma que luego es normalizado, posteriormente la entropía es calculada para cada segmento de señal y el valor es almacenado en un vector. El proceso se repite para todos los segmentos de la señal.

Otra forma de calcular la entropía es variar la forma en que se halla la función de densidad de probabilidad, éste cálculo se hace sobre el espectro de la señal para lo cual la magnitud del componente espectral es estimado por la normalización sobre todos los componentes frecuenciales; luego, la entropía es calculada de una función de densidad de probabilidad multiplicada por una función de ponderación que tiene como objetivo dar más realce a regiones frecuenciales en las que la voz tiene mayor presencia. En la siguiente figura se presenta el diagrama para el cálculo de la entropía para un segmento de señal.

Figura 18: Cálculo de la entropía en el espectro



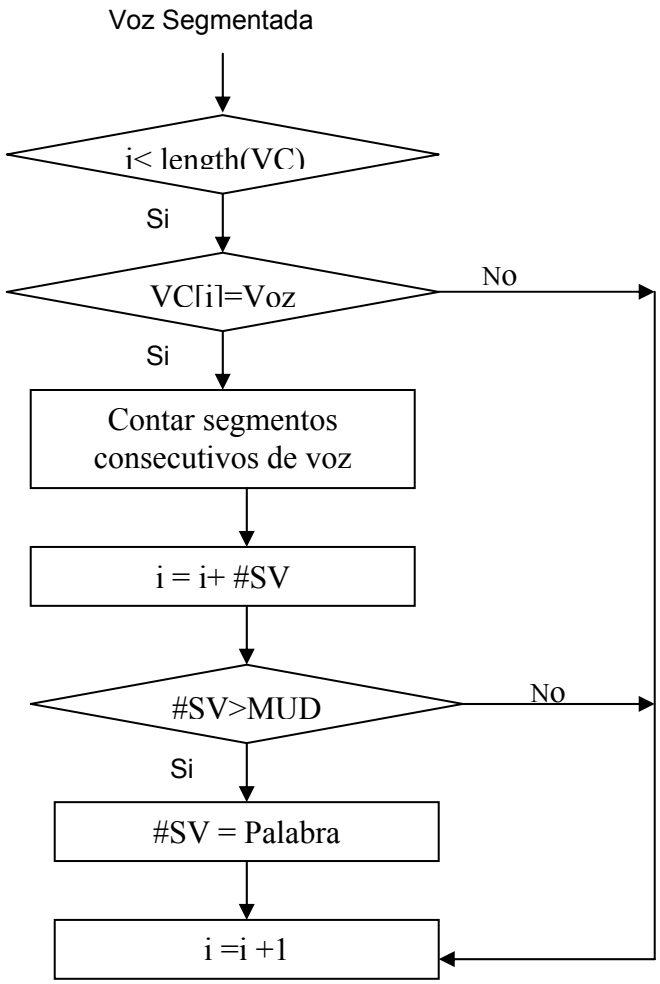
Luego que se han obtenido los valores de entropía para toda la señal el paso siguiente es hallar un umbral para el perfil de la entropía.

2.6 DETECCION DE EXTREMOS

En la etapa anterior se extraían unos parámetros que servirían como un discriminante para la variación entre segmentos de voz y aquellos que no lo son. Pero el proceso de detección de puntos de inicio y de fin no puede estar completo sin una lógica que establezca un umbral en el cual para clasificar estos segmentos y unos criterios, de mínima duración de una pronunciación (MUD) y de mínima separación entre palabras (MUS).

La función del MUD es tomar aquellos segmentos clasificados como voz y contar y clasificar como voz sólo aquellos que tengan una secuencia igual o mayor a un valor establecido.

Figura 19: Diagrama Algoritmo MUD

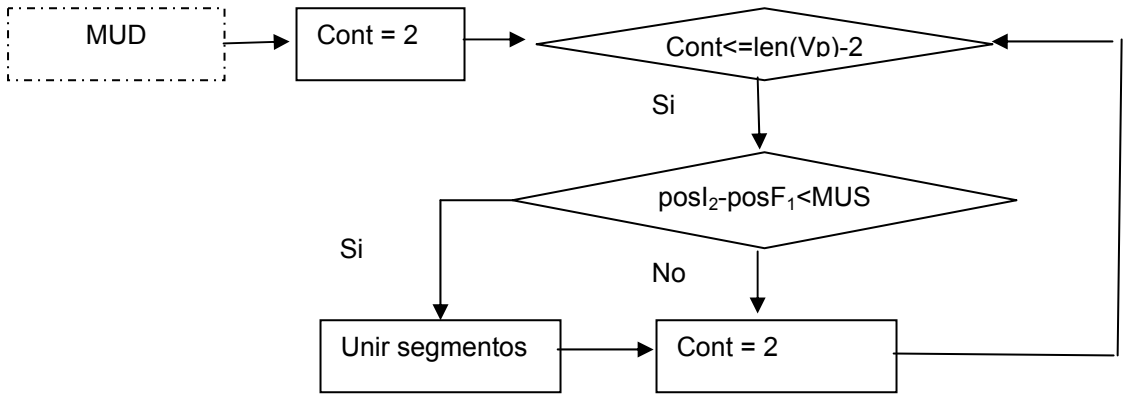


VC es un vector de la señal segmentada y clasificada como voz, #SV es el número de segmentos de voz consecutivos que salen del proceso denominado en la figura como contar # de segmentos consecutivos, si el número es mayor a una constante denominada MUD se clasifican estos segmentos realmente como voz formando parte de una segmentación de palabra, de lo contrario son descartados

El algoritmo de MUD entra en funcionamiento cuando las falsas clasificaciones de voz han sido eliminadas en el proceso anterior, este busca eliminar las distancias entre segmentos de voz con la finalidad de ir conformando grupos de segmentos o pronunciación que en su conjunto o reunión forman las palabras.

Este algoritmo se diseño de la siguiente manera: Las posiciones de punto de inicio y fin de tramas de señal se guardaron en este vector, se mide la distancia de fin de una trama con la de inicio de la siguiente sí esta separación es menor a un valor dado se elimina y la nueva trama estará conformada por el punto de inicio de la primera y el punto de fin de la trama siguiente.

Figura 20: Diagrama MUS



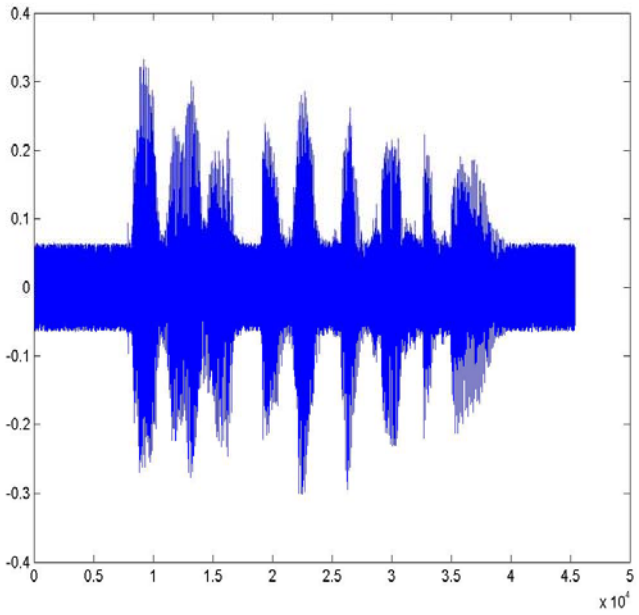
El contador se inicializa en dos por que la primera comparación se hace entre el punto final de la primera trama y el punto de inicio de la trama siguiente. Las variables que se denotan como $posI_2$ - $posF_1$ hacen referencia a las posiciones

que se van a comparar con una constante y de eso depende si las tramas son unidas.

2.7 ALGORITMO FINAL

El algoritmo comienza su funcionamiento con la carga de la señal, en este proceso las muestras de la señal de voz son extraídas de los archivos .wav, además, las características de las señales como frecuencia de muestreo, número de bits por muestras y números de canales que conforman las muestras son leídas del archivo. Al leer el archivo se tiene una señal de la siguiente forma:

Figura 21: Señal antes de ser procesada



Cuando las muestras que conforman la señal han sido cargadas, el paso siguiente es segmentarla, donde cada segmento tiene una longitud de 25ms y con un solapamiento de 10ms entre ventanas; como se dijo antes cada segmento de la señal de voz es almacenada en una matriz donde cada columna estará formada

por un segmento de voz, luego de ser segmentada la señal entra a la primera etapa o etapa de eliminación de ruido, proceso para bajar los niveles de ruido que acompañan la señal. En esta etapa un perfil de ruido de la señal contaminante es estimado para ser utilizado con la técnica de sustracción espectral; éste perfil puede ser estimado de un segmento de ruido extraído a la hora de construir la base de datos o de un segmento perteneciente a la señal donde no se encuentre voz.

Las técnicas de substracción espectral requieren que la señal sea llevada al espacio frecuencial por lo que a la matriz de segmentos o ventanas se le aplica la Transformada Rápida de Fourier al igual que al perfil de ruido ya que también es estimado en el espacio frecuencial; es entonces cuando el proceso de substracción espectral es llevado a cabo. Realizada la substracción se aplica un proceso de rectificación de media onda donde los valores negativos resultantes del proceso son llevados a cero.

Después de esta etapa tendremos una señal con un nivel de ruido más bajo que el de la señal anterior como se muestra en la siguiente figura:

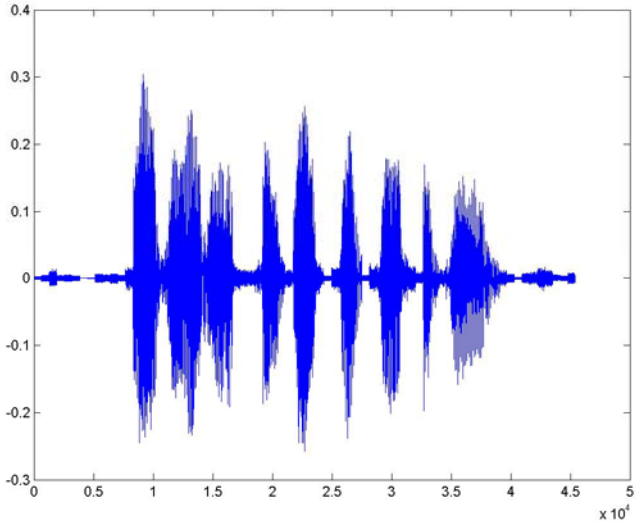


Figura 22: Señal después de la primera etapa

En la etapa de preprocesamiento, de la señal de voz es mejorada a través de la aplicación de un filtro pasabanda el cual su objetivo es eliminar aquellas componentes frecuenciales en la cual la voz no tiene presencia o no son determinantes para un proceso posterior; el rango elegido para cortar las frecuencias va desde los 0 hasta los 150 Hz y mayores a 6000Hz; estos rangos pueden ser configurados de otra manera dependiendo del uso final de la aplicación para la cual se destine el algoritmo.

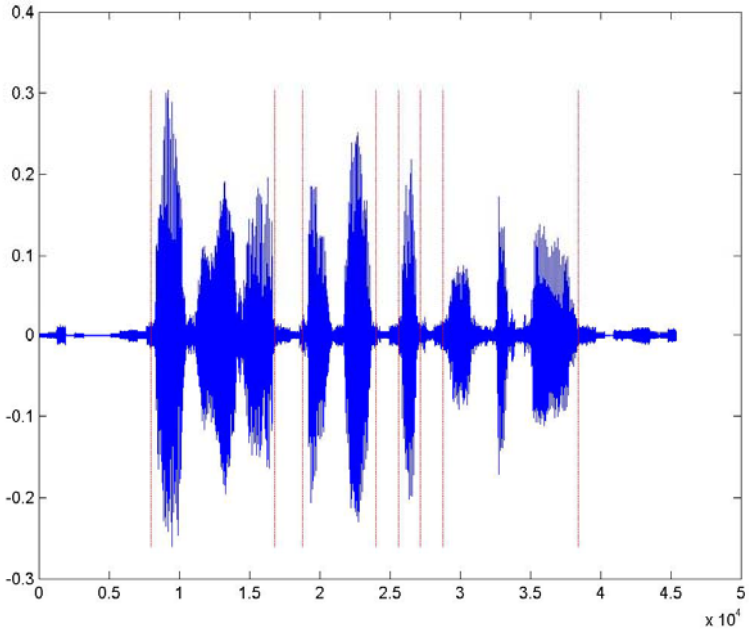
En esta etapa, aunque la señal sufre cambios; estos cambios, no son tan visibles como los vistos en la etapa anterior.

En la siguiente etapa, de la señal son extraídas las características ya sea la energía o entropía las cuales se usarán en este último proceso; sí es necesario en esta etapa la señal que es procesada es llevada al dominio del tiempo o se puede seguir representando en el dominio espectral; con las características extraídas de la señal el perfil construido caracteriza y diferencia la voz de lo que no es voz.

Luego de que el perfil es hallado la señal es clasificada como voz o no a través de un umbral que puede ser estático o dinámico.

Los segmentos de voz que han sido falsamente clasificados son eliminados, los segmentos verdaderos pertenecientes a una misma pronunciación o palabra son luego unidos para ser considerados como uno sólo.

Figura 23: Señal procesada etapa tres



3. PRUEBAS Y EVALUACIÓN

Este capítulo se divide en dos partes: La primera parte corresponderá a pruebas que se hicieron para encontrar el mejor desempeño posible en los algoritmos de eliminación de ruido. Las pruebas de optimización se realizaron a un grupo de señales con 5 diferentes cocientes de SNR. La segunda parte tiene que ver con las pruebas hechas al algoritmo final para medir su desempeño.

3.1 BASE DE DATOS

La base de datos de hablantes utilizada fue la EUSTACE speech corpus (http://www.cstr.ed.ac.uk/projects/eustace). La cual pertenece al Centre for Speech Technology Research de la universidad de Edinburgh, Inglaterra. Contiene archivos de sonido en formato WAV de la voz de 6 sujetos diferentes, 3 hombres y 3 mujeres. De cada sujeto se tienen 64 archivos donde cada archivo contiene como mínimo diez repeticiones de expresiones en idioma inglés.

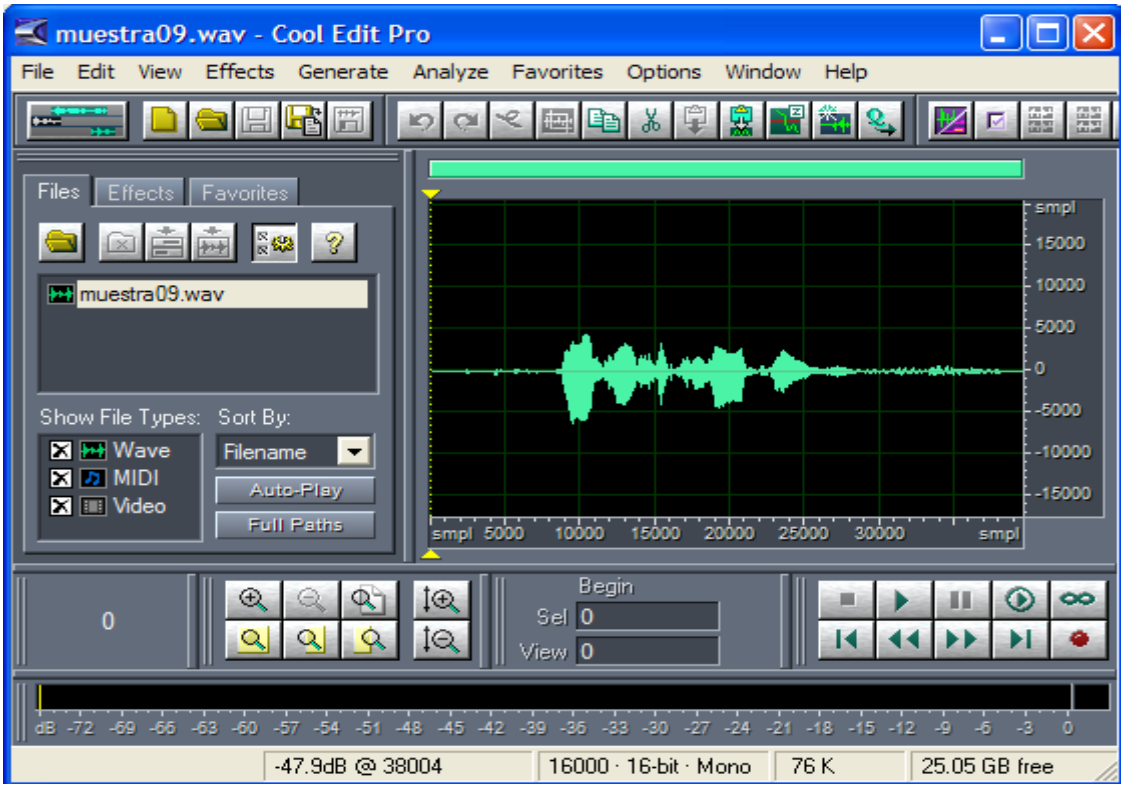
Los archivos fueron capturados con una frecuencia de muestreo de 16000 Hz en un sólo canal (mono) y a 16 bits de representación para cada muestra.

3.1.1 Marcado de las Muestras

Para poder medir la efectividad del algoritmo se debe tener un punto de comparación, este punto de comparación se hizo a través del marcado manual de la base de datos, el cual se realizó con la ayuda del software para la edición de audio Cool Edit Pro 2.0. Está aplicación permite visualizar la señal en el tiempo o

un espectrograma de ella por lo que facilita su marcado, además de poder desplazarse a través de la señal con precisión, permitir el manejo de zoom y otras características que facilitan el marcado.

Figura 24: Ventana del Cool Edit Pro 2.0



3.2 COMPUTADOR UTILIZADO

El Computador utilizado desde la etapa de diseño hasta la etapa de evaluación fue un computador Hp pavilion ze5637la con un procesador Pentium IV celeron® de 2,8 Ghz, con 512 Mb en memoria ram y sistema operativo windows Xp Home.

3.3 PRUEBAS Y OPTIMIZACIÓN DE LOS ALGORITMOS DE ELIMINACIÓN DE RUIDO

Estas pruebas nos ayudarán a optimizar los parámetros de cada uno de los algoritmos que sirven para evaluar el desempeño óptimo de los algoritmos construidos para cada una de las técnicas y elegir el que hará parte del algoritmo definitivo.

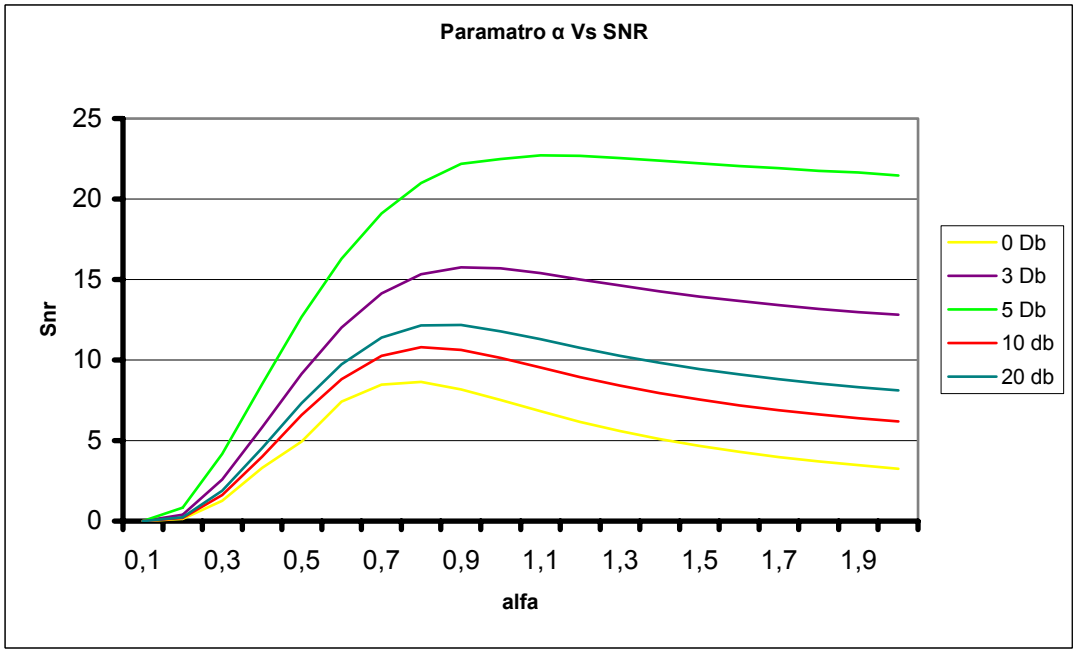
3.3.1 Alfa Óptimo

Esta prueba se realizó con la idea de optimizar el desempeño del algoritmo de substracción espectral, la prueba consistía en variar el parámetro alfa, con el objetivo de encontrar los mejores resultados para la señal en la expresión:

$$|S(w)| = |Y(w)|^a - E\left[|N(w)|\right]^a \qquad (25)$$

Donde el parámetro α varia entre 0.5 y 2. Las señales que se utilizaron estaban contaminadas con ruido de Gauss blanco. El grupo de señales con el mismo SNR tuvieron un comportamiento similar y para cada grupo de señales se grafica un comportamiento promedio el cual se presenta en la siguiente figura:

Figura 25: Comportamiento al variar alfa



Con los resultados obtenidos en la grafica, se construyo la siguiente tabla con el fin de buscar un alfa óptimo sin importar el valor del cociente SNR.

Tabla 2: Alfa óptimo

SNR [Db]	a
0	0,8
3	0,9
5	1,1
10	0,8
20	0,9
Promedio	0,9

Se puede decir que un valor de α de 0.9, sería el óptimo para la mayoría de valores. La siguiente tabla presenta los resultados de la substracción espectral realizada utilizando el valor de alfa óptimo, en la tabla se puede observar los cocientes de SNR para señales contaminadas y la variación del cociente SNR de la señal luego de ser procesada por el algoritmo:

Tabla 3: Resultados con el valor de alfa óptimo

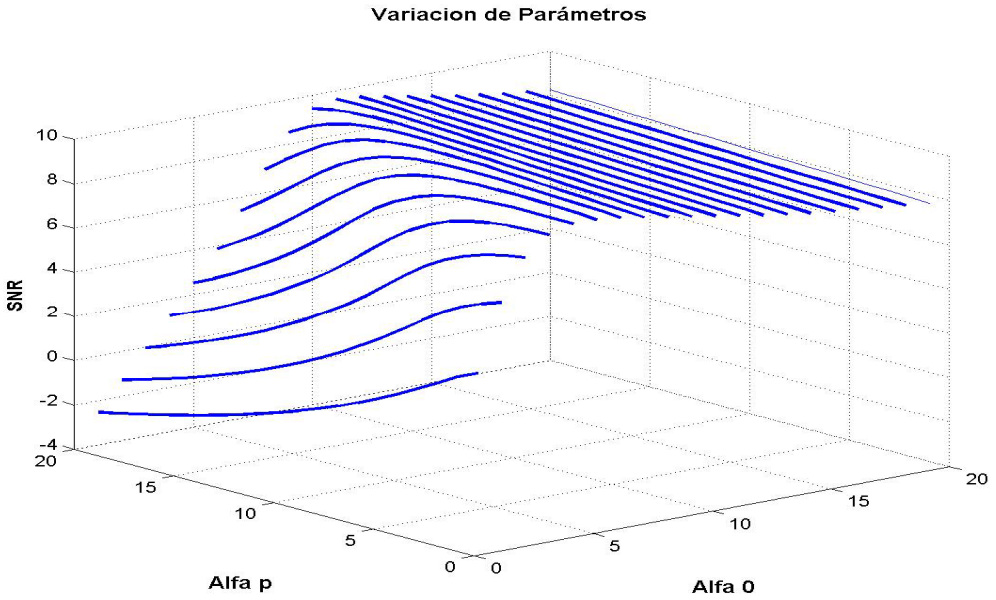
Ruido [db]	Promedio [db]
0	8,6263
3	10,7912
5	12,1735
10	15,7648
20	22,7191
Promedio	6.34232

El promedio que se muestra al final es la mejora en promedio del cociente SNR de una señal al ser procesada por el algoritmo; en general se obtienen buenos resultados con esta técnica.

3.3.2 Pruebas de Substracción Espectral Usando un Factor de Sobresubstracción

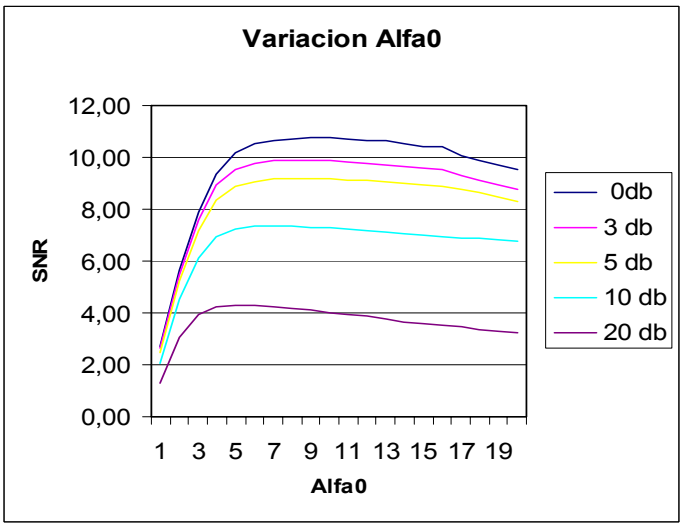
En la primera parte se busca optimizar la expresión que realiza la substracción, para ello se busca aquellos valores de α_p y α_0 para los cuales la expresión obtenga sus mejores resultados, esto se hizo variando los dos parámetros y comparando la señal procesada, contra una señal de referencia que no presentaba ruido, se calcula el cociente SNR antes y después. A continuación se muestra una superficie en la cual se pueden observar el cociente SNR al variar los parámetros, cada parámetro se varia en un rango de 0 a 20.

Figura 26: Superficie cambio SNR variando los parámetros α_p y α_0



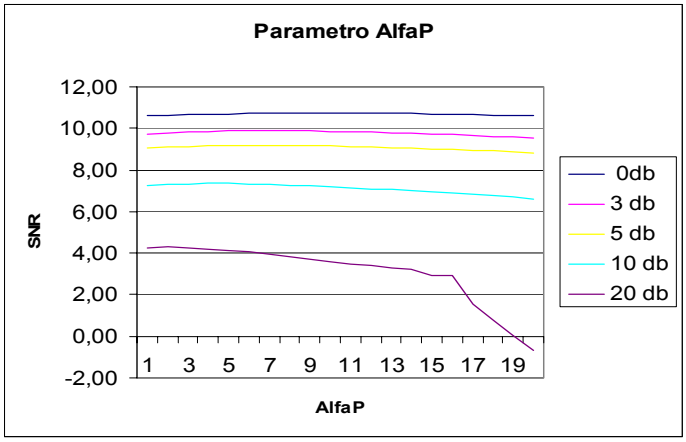
Las pruebas se hicieron sobre varias señales de voz donde cada una presentaba diferentes niveles de ruido, variando los parámetros de la manera antes indicada, se busca un α_0 optimo al variar sus valores y manteniendo constante el valor de α_p se consiguió que la señal mejore como se muestra a continuación.

Figura 27: Comportamiento al variar α_0



Según los resultados y que se muestran en la gráfica se puede escoger un α_0 con un valor de 6. Luego se varía α_p con el α_0 obtenido para encontrar un valor que optimice los resultados.

Figura 28: Variación parámetro AlfaP



Con las pruebas hechas se observó que para un valor de α_p =5 se obtienen buenos resultados como se puede ver en la superficie vista en la figura 28.

Realizadas las pruebas a diferentes muestras de voz con los valores óptimos para α_o y α_p, se muestra un promedio de las mejoras obtenidas por las señales:

Tabla 4: Resultados para el modelo de substracción con factor de sobresubstracción

Ruido [db]	Promedio [db]
0	9,74
3	11,86
5	13,21
10	16,33
20	23,13
Promedio	6.4150

3.3.3 Substracción Espectral con Magnitud Selectiva

Realizando la substracción con magnitud selectiva, usando una media como función de estimación de la señal original se obtuvieron los siguientes resultados:

Tabla 5: Resultados usando el modelo de substracción selectiva con un ajuste de media

Ruido [db]	Promedio
0	10,0445
3	11,9219
5	13,0022
10	15,8933
20	21,5218
Promedio	6.8767

Al ejecutar de nuevo el test, se utiliza un ajuste cuadrático; aproximación que es mejor para la estimación de los pequeños cambios en la magnitud:

$$\hat{S}_i(k) = F(Y_{i-m/2},...,Y_i,..,Y_{i+m/2}) \qquad (25)$$

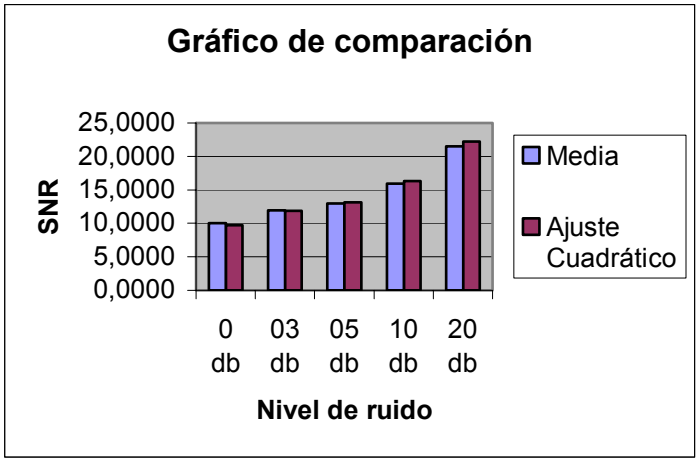
Donde, i es el índice de ventana y m es el número de ventanas escogido para el ajuste cuadrático del estimado de la voz limpia. Con este ajuste se obtuvieron los siguientes resultados:

Tabla 6: Resultados con ajuste cuadrático

Ruido [db]	Promedio
0	9,7030
3	11,8557
5	13,1387
10	16,2882
20	22,2177
Promedio	7,0407

En la siguiente gráfica se presenta una comparación entre las dos modificaciones y los resultados obtenidos para las diferentes muestras:

Figura 29: Comparación magnitud selectiva



Como se observa en la gráfica con el ajuste cuadrático se obtienen mejores resultados por lo que se aplicara un ajuste cuadrático cuando se utilice está técnica.

3.3.4 Recuperación Espectral

Esta técnica no tiene parámetros que se puedan optimizar para un mejor rendimiento. A continuación se muestra el resultado de las pruebas hechas con está técnica.

Tabla 7: Resultados recuperación espectral

Ruido [db]	Promedio [db]
0	5,5023
3	8,2035
5	9,9822
10	14,1666
20	22,0582
Promedio	4,3826

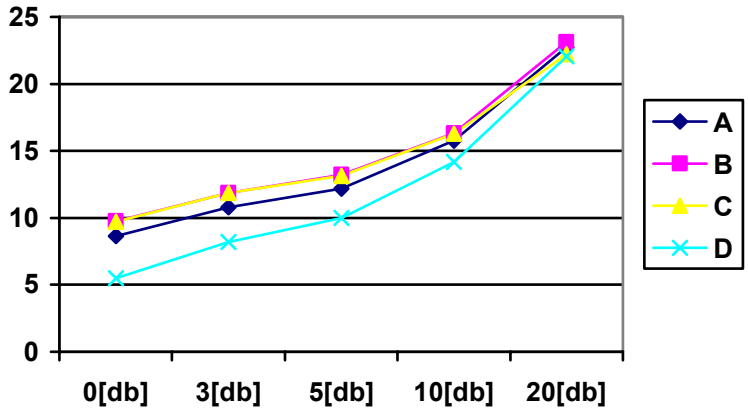
3.3.5 Comparación de las Técnicas

Con los resultados obtenidos después de encontrar las expresiones óptimas para las diferentes técnicas se evaluaron sus resultados para escoger aquella que tenga el mejor desempeño. Los resultados se presentan en una tabla donde cada técnica estará designada por una letra, la cuál será utilizada en la figura que sigue:

Tabla 8: Comparación diferentes técnicas

	SNR				
Técnica	**0[db]**	**3[db]**	**5[db]**	**10[db]**	**20[db]**
A. Substracción Espectral	8,6263	10,7912	12,1735	15,7648	22,7191
B. S.E con factor de sobresubstracción	9,74	11,86	13,21	16,33	23,13
C. S.E con magnitud selectiva	9,7030	11,8557	13,1387	16,2882	22,2177
D. Recuperación Espectral	5,5023	8,2035	9,9822	14,1666	22,0582

Figura 30: Comparaciones distintas técnicas



Como se puede observar en la tabla las técnicas B y C proporcionan resultados similares por lo tanto cualquiera de las dos seria candidata a formar parte del algoritmo final; uno de los objetivos del proyecto es tener un tiempo de respuesta rápido por lo tanto será tomado como un factor decisivo para elegir el algoritmo final; a continuación se presenta un promedio del tiempo de respuesta para varias señales y la longitud de dicha señal.

Tabla 9: Comparación tiempos de ejecución distintas técnicas**.**

	B	C	Número de muestras
Senal1	1,2410	11,9970	45397
Senal2	0,9510	12,0470	45807
Senal3	1,2220	14,4305	54687
Senal4	1,0320	12,2575	49790
Senal5	0,6910	8,8525	33077
Senal6	1,0110	12,5885	48461
Senal7	0,8110	10,8500	39076

Los resultados arrojados en la prueba muestran que la técnica B o Substracción espectral con un factor de sobresubstracción presenta un tiempo de respuesta 10 veces menor que la otra técnica; por lo que será la técnica utilizada en el algoritmo final.

3.4 PRUEBAS PARA EVALUAR EL ALGORITMO DE DETECCIÓN DE INICIO Y FIN DE PALABRA PARA SEÑALES DE VOZ

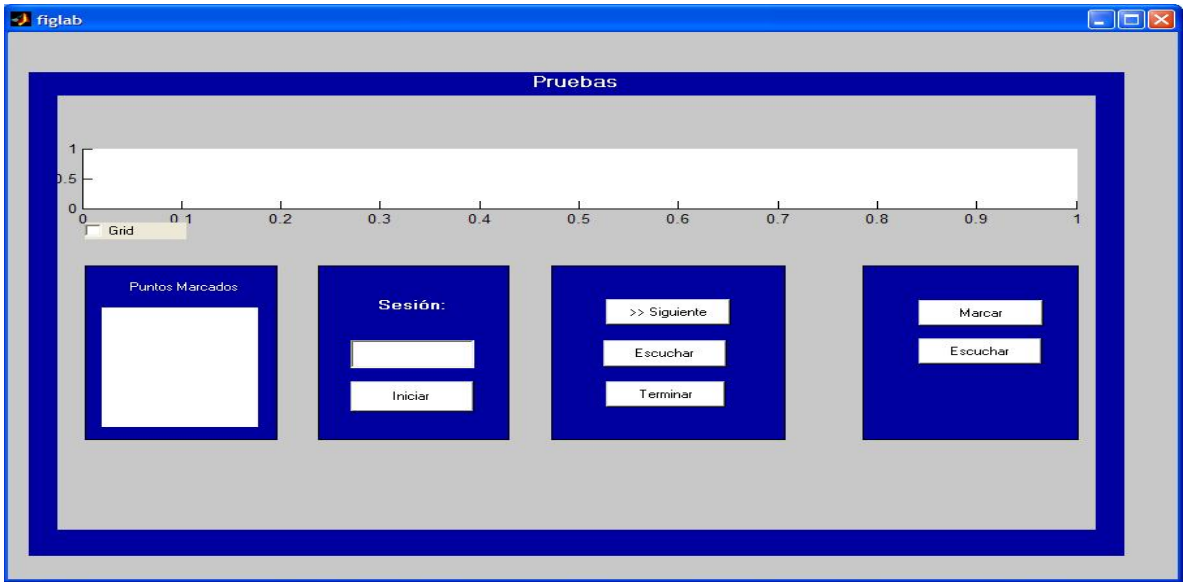
3.4.1 Prueba Para Medir la Exactitud

En esta prueba se midió que tan acertado es el algoritmo a la hora de encontrar los puntos de inicio y fin de la señal. Para hacer esta prueba, se utilizan tres datos de comparación: El primer dato de comparación son los puntos obtenidos de las señales marcadas solo al oído por las personas; el segundo, os datos de comparación son los puntos hechos por un marcado visual y auditivo a la vez el cual se hizó con la ayuda del Cool Edit pro 2.0, estos puntos se consideran como los más reales o o con mayor exactitud, y el tercer grupo de datos esta conformado por los puntos que se obtienen como resultado de la aplicación de los algoritmos.

El marcado auditivo se realizó con ayuda de un grupo de personas a través de una interfaz programada en Matlab. En esta interfaz se encuentran: Un módulo para carga de las muestras y las características de la señal son extraídas del archivo, un módulo para graficar donde la señal es representada por una barrita cuya longitud es la longitud de la señal, un módulo de reproducción que permite escuchar la señal completa o segmentos de ella y un módulo de almacenamiento donde se guardan las pruebas para cada persona en un archivo de texto.

El proceso empieza cuando las personas encargadas de marcar la base de datos crean una sesión con su nombre lo que origina un archivo donde se almacenan los datos, luego la persona establece posibles puntos de inicio y de fin de la señal; puede ir verificando estos puntos mediante la comparación de la región marcada con la original o el marcaje de nuevos puntos, una vez los puntos reciben el aval de la persona, son almacenados en el archivo de lo contrario el usuario empieza un proceso de refinamiento, la interfaz presenta los puntos que se han marcado para que el usuario pueda tener la certeza de saber si el punto buscado se encuentra antes o después. En la siguiente figura se puede observar una imagen de la interfaz utilizada:

Figura 31: Interfaz construida en MATLAB.



En la siguiente tabla se presentan los resultados de la prueba. Los resultados de la prueba auditiva presentan un promedio de los puntos marcados por las personas y su desviación estándar, se presentan los puntos marcados por el Cool Edit pro 2.0 y los obtenidos por los algoritmos elaborados.

Tabla 10: Resultados prueba

Marcado Auditivo				**Marcado auditivo y visual**		**Marcado Automático**			
P. Inicio		**P. Fin**				**Energía**		**Entropía**	
$\bar{x}$	**Σ**	$\bar{x}$	**Σ**	**P. inicio**	**P. Fin**	**P. I**	**P. Fin**	**P. I**	**P. F**
9673	471,1	32871	408,4	9990	32750	9920	33280	10080	32000
5628	301,7	31596	319,3	5605	31435	5600	31680	5760	31880
13253	383,8	37352	216,1	14750	36700	14400	36960	14720	37040
8310	87,3	32032	576,6	8200	31700	8000	32060	8320	31840
3312	175,1	20591	306,0	3390	20840	3350	20920	3120	20960
3880	171,6	29752	94,1	4280	29930	4000	30090	3840	29760
672	80,5	15941	245,2	270	14500	310	16331	480	14560
3295	177,4	26548	124,5	3310	26725	3372	26960	3360	26560
7398	878,4	28069	246,3	8840	28210	8320	28800	8800	27920
6818	251,2	29623	176,5	6780	29690	6580	29760	7040	30600
6055	144,3	26781	474,4	5720	26320	4640	26240	6400	26800
3222	32,1	29653	613,6	2915	30025	2880	30320	3040	30240
6782	358,1	29614	285,0	7125	27940	7040	27860	7200	26800
5710	147,7	31833	941,3	5812	32000	5600	32160	5920	32560
6407	295,9	30223	547,7	6885	28690	6400	28960	6880	28240
2745	111,3	28384	146,4	2570	29160	2540	29120	2240	29360
14426	285,8	35961	352,7	14927	35993	13920	36320	14800	35440
6422	177,8	31847	441,7	6180	32450	6080	33120	6000	32000
6272	91,2	25303	245,4	6450	31100	6240	25620	6240	25120
5753	106,9	31898	259,0	5700	32000	5600	32540	5760	32320

Para las gráficas utilizaremos la siguiente nomenclatura:

Tabla 11: Nomenclatura utilizada para las gráficas

Nomenclatura	
Forma 1	Marcado con el Cool Edit Pro 2.0
Forma 2	Marcado Auditivo
Forma 3	Marcado Algoritmo Energía
Forma 4	Marcado Algoritmo Entropía

Figura 32: Comparación distintos marcados para el punto de inicio.

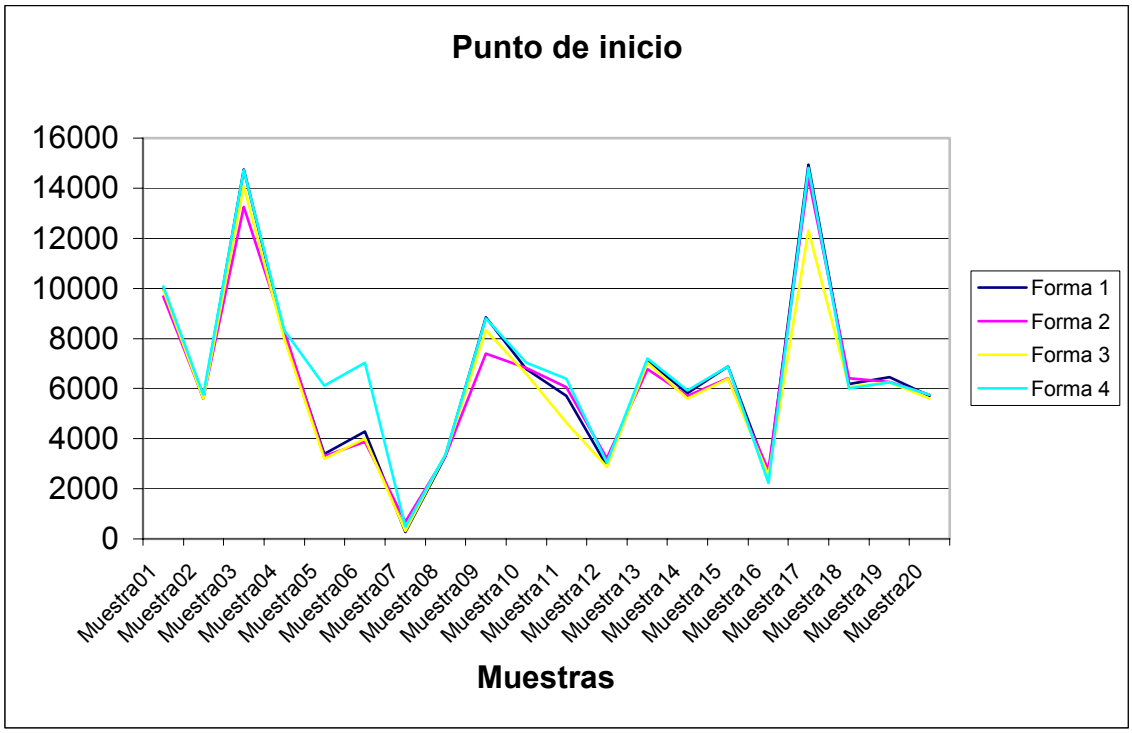
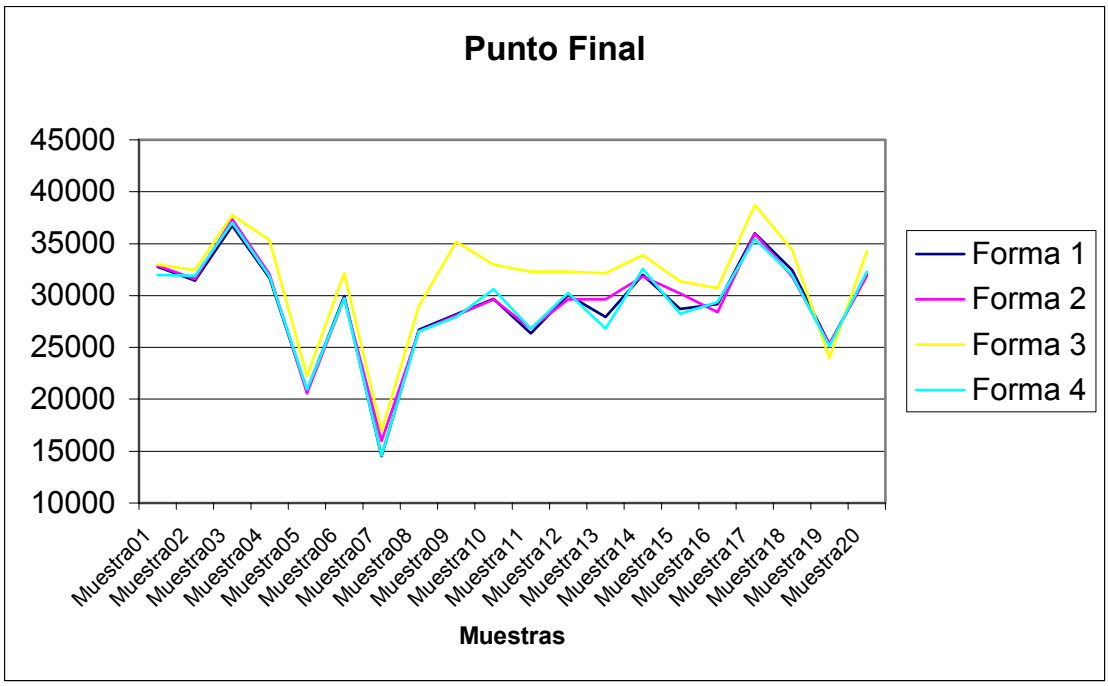


Figura 33: Gráfica comparación distintos marcados para el punto de inicio.



A continuación se presenta una tabla con los promedios de comparación de los resultados obtenidos por las formas 2,3 y 4 contra la forma 1 como propuesta para medir el grado de exactitud.

Tabla 12: Comparación promedio de las distancias con los puntos de comparación

	Punto inicio	Punto Fin
Auditivo	534	682
A. Energía	394	523,27
A. Entropía	238	482,7

Los algoritmos son más exactos que el marcaje auditivo como se puede observar en los resultados obtenidos en la tabla anterior. Por lo que es necesario comparar los dos algoritmos usando una prueba con un número mayor de señales. Se tomaron un total de 50 señales.

Tabla 13: Comparación de lo resultados de los algoritmos.

	Energía		**Entropía**	
	Muestras	Tiempo(s)	Muestras	Tiempo(s)
P. Inicio	953	0,059 s	582	0,036 s
P. Fin	865,56	0,054 s	570	0,035 s

Como se observa en la tabla se obtienen mejores resultados con el algoritmo de entropía que con el algoritmo de energía.

3.4.2 Prueba de Tiempo

Uno de los factores de mayor peso a la hora de construir un algoritmo es el tiempo de respuesta por dos consideraciones en especial sí el tiempo de respuesta es critico: La aplicación que contenga el algoritmo, el tiempo de respuesta de éste se incrementa considerablemente; y la otra consideración sí el tiempo de respuesta es alta, es más fácil hacer el marcado manual que utilizar el algoritmo.,

La prueba se hizo usando 20 muestras en la cual los usuarios marcaban las muestras, el tiempo de inicio y de fin de la prueba era cronometrado.

Tabla 14: Comparación tiempo de ejecución

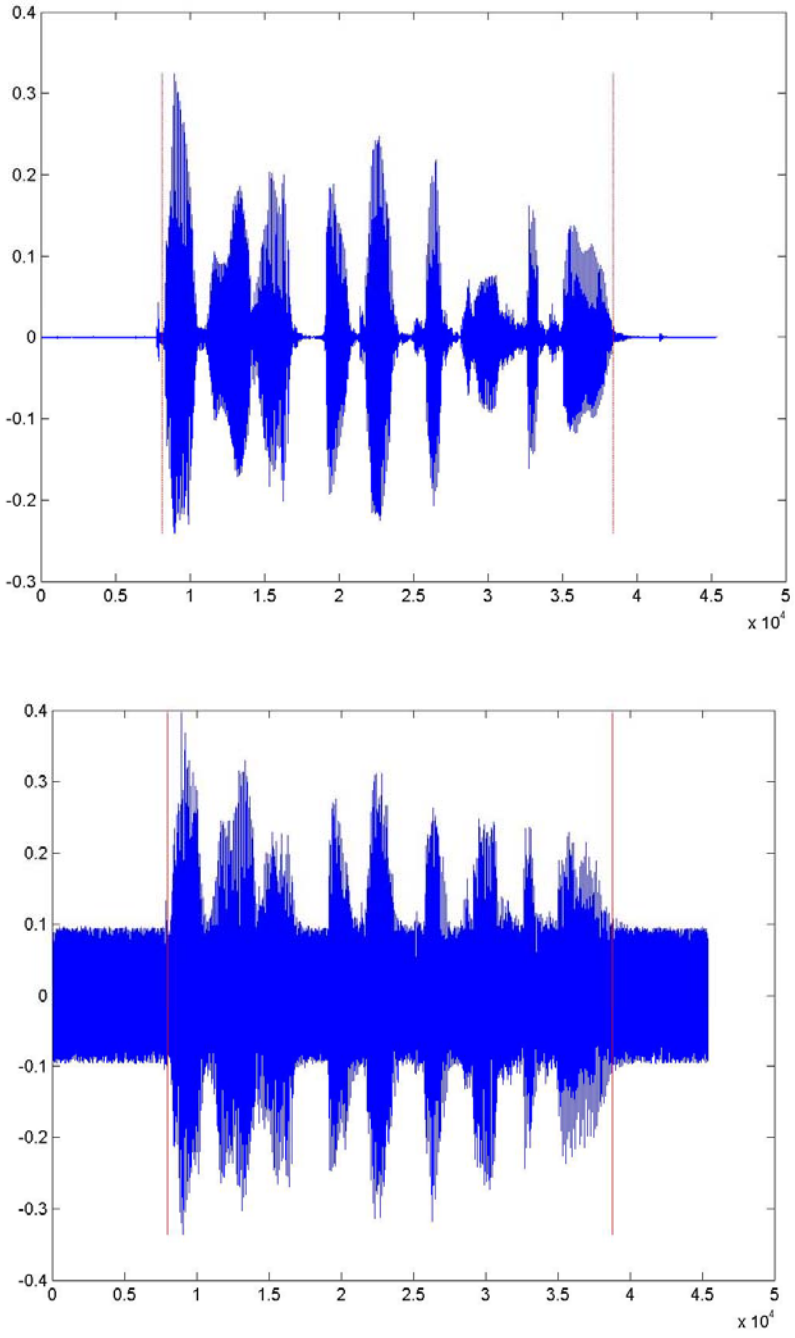
Marcado Auditivo	Marcado auditivo y visual	Marcado A. Energía	Marcado A. Entropía
1 h:50 m	1 h:30 m	50 seg.	45 seg.

En esta prueba se pudo observar la gran diferencia entre un marcado automático y uno manual, el marcado automático presentan tiempos cien veces menores a los utilizados por el marcado automático, para un ambiente no optimizado desarrollado en MATLAB.

3.4.3 Prueba de Robustez

Esta prueba consiste en contaminar una señal a diferentes niveles de ruido, a los marcar los puntos de inicio y de fin, para compararlo con unos puntos previamente marcados sobre una señal que no presentaba contaminación.

Figura 34: Prueba de robustez



En la figura se puede observar, en la parte superior la señal sin contaminación y en la parte de abajo se muestra los resultados sobre la misma señal pero con un alto contenido de ruido.

Figura 35: Comparación resultados prueba de robustez

	Energía		Entropía	
SNR	**P. Inicio**	**P. Fin**	**P. Inicio**	**P. Fin**
0 db	0.083 s	0,082 s	0,056 s	0,057
3 db	0,075 s	0,076 s	0,051 s	0,048 s
5 db	0,073 s	0,074 s	0,048 s	0,047 s
10 db	0,057 s	0,062 s	0,039 s	0,04 s
20 db	0,055s	0,059 s	0,038 s	0,039 s

En la tabla se observa en cuanto divergen los resultados dependiendo del cociente SNR, el punto se considera exacto cuando no esta alejado más de un segmento de distancia del punto marcado.

4. CONCLUSIONES Y RECOMENDACIONES

El trabajo de esté proyecto de grado va enfocado hacia la detección de los puntos de inicio y de fin de palabra o segmentos de voz, el trabajo se realizó en base a dos características de la señal la energía y la entropía. Los resultados mostrados establecieron superioridad de la entropía frente a la energía.

El algoritmo presenta robustez contra diferentes niveles de ruido al usar técnicas para realce de voz por lo que presenta buen desempeño en ambientes ruidosos permitiendo que pueda ser implementado para sistemas por ejemplos de tele conferencia.

Las técnicas de sustracción espectral funcionan con base en un perfil de ruido que puede ser extraído de la señal, estimado o tomado de otra muestra, por lo que sí se quiere mejorar los resultados obtenidos al aplicar esta técnica se pueden utilizar estimadores de ruido que hagan más exacto el perfil de ruido.

Al aplicar las técnicas de substracción aparece otra clase de ruido sobre la señal conocido como ruido musical causado por la aparición de componentes frecuenciales aislados, este ruido es más notorio entre mas bajo sea el cociente SNR de la señal. La desaparición de este tipo de ruido es algo complicado por lo que se puede implicar un aumento considerable del tiempo de respuesta del algoritmo. A futuro se debe hacer otro estudio de técnicas de

substracción espectral con los cuales se obtengan mejores resultados a la vez que reduzcan los niveles de ruido musical o lo desparezcan completamente.

Los algoritmos presentan tiempo de respuestas bajos y baja complejidad computacional por lo que pueden ser implementados en otros sistemas como circuitos electrónicos, DSP (Procesador Digital de Señales) para ser utilizados online o en sistemas que necesiten respuesta en tiempo real.

El algoritmo basado en la entropía demostró ser mejor que el algoritmo de energía en todos los aspectos, demostrando que se pueden construir algoritmos basados en características diferentes a la tradicional para mejorar la exactitud. A futuro se pueden hacer combinaciones entre estos dos métodos.

El algoritmo con pequeñas modificaciones o adaptaciones puede ser aplicado a diferentes tipos de señales acústicas.

Los algoritmos de detección de inicio y fin de palabra “endpoint” son necesarios para muchas aplicaciones de tratamiento de la voz, por lo que con este trabajo se deja una base para las personas que quieran incursionar en este importante campo.

BIBLIOGRAFÍA

[1] BERNAL BERMUDEZ, Jesús; BOBADILLA SANCHO, Jesús y GOMEZ VILDA, Pedro. Reconocimiento de Voz y fonética acústica. México DF.: Alfaomega, 2000. 332 p.

[2] FÁUNDEZ ZANUY, Marcos. Tratamiento digital de voz e imagen. México DF.: Alfaomega, 2001. 271 p.

[3] GOLBERG, Randy y RIEK, Lance. A practical handbook of speech coders. USA: CRC press LLC, 2000. 231 p.

[4] WITTEN, Ian; Principles of computer speech. USA: ACADEMIC PRESS INC, 1982. 280 p.

[5] MORA PINILLA, Benedicto. Reconocimiento de voz basado en el uso de redes neuronales. Bucaramanga, 1995. 46p. Trabajo de grado (Ingeniero de Sistemas). Universidad Industrial de Santander. Facultad de ingenierías Físico – Mecánicas.

[6] CONTRERAS, Sonia. Detección activa de señales de voz. Universidad Industrial de Santander, Simposio tratamiento de señales 2002.

[7] GANAPATHIRAJU A; WEBSTER L; TRIMBLE J; BUSH k. y KORNMAN P. Comparison of energy-based endpoint detectors for speech signal processing. Department of Electrical and Computer Engineering Mississippi State University, en http://www.isip.msstate.edu/publications/conferences/ieee_secon/1996/endpointer/presentation.pdf.

[8] GU, Lingyun; GAO, Jianbo y HARRIS, John G. Endpoint detection in noisy environment using a poincaré Recurrence metric. University of Florida, en http://www.imt.liu.se/mi/Intern/Proceedings/ICASSP03/pdfs/01-00428.pdf.

[9] KARRAY, Lamia y POLARD, Emmanuel. A wavelet denoising technique to improve endpoint detection in adverse conditions. France, en http://www.telecom.tuc.gr/paperdb/eurospeech99/PAPERS/S11P3/K003.PDF.

[10] SARIKAYA, Ruki y HANSEN, Jhon. Robust speech activity detection in the precense of noise. University of Colorado Boulder, en http://cslr.colorado.edu/rspl/PUBLICATIONS/PDFs/CP-icslp98-SAD-SL980922-Wcover.PDF.

[11] SHAFRAN, Izhak y ROSE, Richard. Robust speech detection and segmentation for real-time asr applications. AT&T Labs research, en http://ssli.ee.washington.edu/ssli/people/zak/xstatseg.pdf.

[12] SHEN, Jia-lin; HUNG Jein-weih y LEE, Lin-shan. Robust Entropy-Based Endpoint Detection for Speech Recognition in Noisy Enviroments. Institute of information Science, en http://www.ee.columbia.edu/~dpwe/papers/ShenHL98-endpoint.pdf.

[13] SHIN, Won-Ho; LEE, Byoung-Soo; LEE Yun-Keun y LEE, Jong-Seok. Speech/non-speech classification using multiple features for robust endpoint

detection. LG Corporate Institute of Technology, en http://www.icsi.berkeley.edu/~dpwe/research/etc/icassp2000/ pdf/1998_93.PDF.

[14] WAHEED, Khurram; WEAVER, Kim y SALAM, Fathi. A robust algorithm for detecting speech using an entropic contrast. Michigan State University, en http://www.egr.msu.edu/~waheedkh/khurram_cv_web.pdf.

[15] ZOU, Qiyue; ZOU, Xiaoxin; ZHANG, Ming y LIN, Zhiping . A robust speech detection algorithm in a microphone array teleconferencing system. School of Electrical and Electronics Engineering, en http://cslr.colorado.edu/array_documents/A%20robust%20speech%20detection%20algorithm%20in%20a%20microphone%20array%20teleconferencing%20system.pdf.

[16] LI, Qi; ZHENG, Jinsong; TSAI, Augustine; ZHOU, Qiru. Robust Endpoint detection and energy normaliztion for real-time espeech and speaker recognition.En: IEEE transactions on speech and audio processing. Vol 10, No 3(2002); p.146-157

[17] RENEVEY, Philippe; DRYGALJO, Andrzej. Entropy based voice activity detection in very noise conditions.

[18] GU, Lingyun; ZAHORIAN, Stephen. A new robust algorithm for isolated word endpoint detection. Department of electrical and computer engineering old dominion university.

[19] BHATNAGAR, Mukul. A modified spectral subtraction method combined with perceptual weighting for speech enhancement. Dallas, 2002. 99p. Trabajo de grado (Master of Science in Electrical Engineering). The University of Texas at Dallas.

[20] ZHU, Qifeng y ALWAN, Abeer. The effect of aditive Noise on speech amplitude spectra: a quantative analysis.IEEE Signal processing letters. Vol. 9, No. 9 (sep 2002); p. 275-277

[21] SOON, Yann; KOH; Soo Nge; YEO; Chai Kiat. Selective magnitude for speech enhancement. Proceedings of the 4th International Conference on High Performance Computing in Asia-Pacific Region, vol. 2, pp. 692-695. 2000