Laboratorio de Procesamiento Digital de Voz

Practica 3.

DETECCIÓN DE INICIO Y FIN DE PALABRA

Objetivos:

Conocer los principales problemas presentados en la detección de inicio y fin de palabras aisladas y programar un algoritmo para ello.

Introducción.

El problema de detección de inicio y fin de una palabra en presencia de ruido ambiental es complicado, para grabaciones en habitaciones a prueba de ruido se pude recurrir al uso de la energía en tiempo corto para la detección, sin embargo en ambientes con ruido es necesario tomar otras consideraciones.

El poder determinar cual es el inicio y el final de una palabra, nos da ciertas ventajas en los sistemas de reconocimiento:

- Procesar menor número de información.
- Comparar únicamente los patrones de información.
- Evitar confusiones a causa del ruido o señales de fondo.

Algunos de los problemas que se presentan en la detección:

- Espurias de ruido que se pueden confundir con la señal.
- Silencios contenidos dentro de las palabras que tienen fonemas plosivos (ej. /t/, /p/, /k/) que pueden confundirse con un falso principio o fin.
- Los fonemas fricativos (ej. /f/, /th/, /h/, etc.), ya que tienen baja energía.
- Sonidos cortos (ej. /t/, /p/, /k/).
- Detección de fonemas nasales al final de la palabra (baja energía y cruces por cero)
- Respiraciones del locutor, que pueden confundirse por su duración.
- Los micrófonos tienen resonancia después de de pronunciar una palabra (sobre todo en vocales).
- Los niveles de ruido pueden confundirse con la señal de voz.

Método para la detección de *inicio - fin* (*Rabiner–Sambur*)

Ante las dificultades para la detección del inicio y fin, se ha desarrollado un método que consiste en considerar las características de los sonidos:

Sonidos sonoros (*voiced*)	Tiene alto contenido en energía. Ocupan las frecuencias bajas del espectro de la voz humana.
Sonidos no sonoros (*unvoiced*)	Tienen bajo contenido de energía. Ocupan las frecuencias superiores del espectro de la voz humana.

De esta forma se puede implementar un detector que incluya ambas características, análisis de energía y frecuencia, cruces por ceros y magnitud promedio (o energía en tiempo corto).

Algoritmo de Detección de Inicio y Fin de palabra:

Detección de inicio

1. Por cada trama de 128 muestras, calcular las funciones: cruce por ceros { *Z[n]* } y la magnitud promedio de la señal { M[n] }, estas funciones se definen a continuación:

$$M_n = \sum_{m=0}^{N-1} |x[m]|$$

$$Z_n = \frac{\sum_{m=0}^{N-2} |sign(x[m+1]) - sign(x[m])|}{2N}$$

2. Para obtener las estadísticas del ruido ambiental, se considera que las primeras diez ventanas son ruido, con lo cual se tiene:

$$Ms_n = \{M_1, M_2, \cdots, M_{10}\}$$
$$Zs_n = \{Z_1, Z_2, \cdots, Z_{10}\}$$

3. Calcular la media y la desviación estándar para las características del ruido y obtener los siguientes umbrales:

Umbral	***Nombre del umbral***	***Valor***
UmbSupEnrg	Umbral Superior de Energía	0.5*max{M_n}
UmbInfEnrg	Umbral Inferior de Energía	μ_{Ms} + 2*σ_{Ms}
UmbCruCero	Umbral de cruces por cero	μ_{Zs} + 2*σ_{Zs}

4. Recorrer la función M_n incrementando en una unidad a ***n*** de 11 hasta que M_n > ***UmbSupEnrg***. En este punto estamos garantizando presencia de señal. A este punto lo marcaremos como ***In***.

5. Resulta lógico pensar que el inicio se la señal se encuentra en algún punto anterior a ***In***, por lo que ahora recorremos la función M_n desde ***n*** = ***In*** hasta que M_n < ***UmbInfEnrg***. Este punto lo marcaremos como ***Ie*** y lo reconocemos tentativamente como el inicio de la señal, determinado por la función de magnitud.

6. Ahora decrementamos ***n*** desde ***n*** = ***Ie*** hasta ***n*** = ***Ie – 25*** o en su defecto ***n*** = ***11***, verificando si sucede alguna de las siguientes condiciones en la función de cruces por cero, ya que lo que ahora buscamos es la posibilidad de que un sonido *no sonoro* preceda a un sonido *sonoro*.

   - Sí { Z_n < ***UmbCruCero*** } significa que no encontramos alguna porción de la señal con aumento importante de frecuencia en 25 ventanas anteriores, por lo tanto el inicio es ***Ie***.

   - Sí encontramos que { Z_n > ***UmbCruCero*** } menos de tres veces seguidas significa que solo fue una espiga de ruido, el punto de inicio sigue siendo ***Ie***.

   - Si encontramos que { Z_n > ***UmbCruCero*** } al menos tres veces seguidas hemos encontrado un sonido *no sonoro*, entonces buscamos el punto ***n*** para el cual { Z_n > ***UmbCruCero*** } la primera de las más de tres veces, es decir, el punto para el cual la

función Z_n sobrepasa el umbral, indicando el comienzo del sonido *no sonoro* y desplazamos el inicio de la palabra a ***Iz***.

Detección de fin

Para la detección de fin de la palabra, hacemos lo mismo pero en sentido inverso a partir del punto (4) de la sección anterior, como si detectáramos un inicio con la señal invertida en el tiempo.

Desarrollo

1. Elegir tres señales de voz y graficar en una misma ventana (usando la función de Matlab *subplot*) las siguientes funciones:

   - La función de energía en tiempo corto. ?
   - La función de magnitud en tiempo corto. ?
   - La función de cruces por cero ?

2. Comprobar cuales son sus características de cada uno de los sonidos principales, esto es, si son sonoros o no. Anotar sus observaciones.

3. Para las mismas señales de voz, graficar las funciones, en una misma ventana:

   - La función de magnitud en tiempo corto.
   - La función de cruces por cero.
   - La señal de voz en el tiempo.

4. Trazar una línea horizontal para los tres umbrales (*UmbSupEnrg*, *UmbInfEnrg* y *UmbCruCero*) para las gráficas correspondientes. ver umbrales

5. Trazar una línea vertical, en las tres gráficas, para los puntos utilizados para la detección de inicio y fin (*In*, *Ie, Iz*, etc).

6. Anotar sus observaciones.

Proyecto:

Programar el algoritmo de *Rabiner–Sambur*, a la entrada se introducirá el nombre del archivo de voz, como resultado entregará: los puntos del desarrollo, los puntos del *inicio–fin* y las graficas de la función de magnitud promedio, las funciones de cruces por cero, la señal original y la señal recortada.

Bibliografía

"Discrete–Time Processing of Speech Signals", John R. Deller Jr, John G. Proakis, John H. L. Hansen, Ed. Prentice Hall, pp 245-251, 1987

"Fundamentals of Speech Recognition", Lawrence Rabiner, Biing Hwang Juang, Ed. Prentice Hall, 1993

"Rabinar and Sambur algorithm", http://www.cs.wpi.edu/~dobrush/cs504/s03/projects/patel/kartik.htm, septiembre de 2003

Apéndice:

1. La función de magnitud esta dada por:

$$M_n = \sum_{m=0}^{N-1} \left| x[m] \right|$$

   Para calcular la función de energía se usa:

$$E_n = \sum_{m=0}^{N-1} \left| x[m] \right|^2$$

2. Para obtener las estadísticas del ruido ambiental, se considera que las primeras diez ventanas son ruido, con lo cual se tiene:

$$Ms_n = \{M_1, M_2, \cdots, M_{10}\}$$
$$Zs_n = \{Z_1, Z_2, \cdots, Z_{10}\}$$

   Donde Ms_n es la magnitud del ruido; las primeras diez tramas de la señal de voz.

   Zs_n son los cruces por cero del ruido: también se consideran las primeras diez tramas de voz como ruido.

3. Para el caso de UmbSupEnrg se usa M_n de las señal de voz (se considera a partir de las 10 tramas en adelante como señal de voz)

   En los siguientes dos casos, se usan los diez valores del ruido, obtenidos en el punto 2 para cada umbral. Recuerda que es la media y la desviación estándar de la magnitud del ruido, así como la media y la desviación estándar del cruce por ceros del ruido.