UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

ESTIMACIÓN DE LA CURVA DE ENTONACIÓN PARA APRENDIZAJE DE SEGUNDO IDIOMA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELECTRICISTA

ISAÍAS GABRIEL ROBLES SCHWARTZ

PROFESOR GUÍA
NÉSTOR BECERRA YOMA

MIEMBROS DE LA COMISIÓN
NELSON BALOIAN TATARYAN
JORGE WUTH SEPÚLVEDA

SANTIAGO DE CHILE
ABRIL 2009

Resumen de la Memoria para optar al
Título de Ingeniero Civil Electricista
Por: Isaías Gabriel Robles Schwartz
Prof Guía: Dr. Néstor Becerra Yoma
Santiago, abril de 2009

"Estimación de la Curva de Entonación para Aprendizaje de Segundo Idioma"

La determinación confiable de la frecuencia fundamental f_0, o pitch, no es una tarea fácil. Muchos algoritmos de detección de pitch se han desarrollado a través de los años, pero sólo son exitosos bajo ciertas circunstancias específicas. Además, la mayoría de los estimadores presentan errores como detectar el doble o la mitad de la frecuencia fundamental (conocidos como *Doubling* y *Halving* en la literatura), por lo que la investigación para encontrar un estimador confiable es aún un problema abierto.

En este contexto, la entonación se define como la variación de la frecuencia fundamental en función del tiempo. El trabajo de esta memoria se centra en un algoritmo de cálculo de la curva de entonación. Dicha curva se estima midiendo la frecuencia fundamental (o la falta de ella) en función del tiempo. Para esto se desarrolló un sistema de extracción de f_0 segmento a segmento llamado *Synthetized Speech Pitch Detector* (SSPD), basado en comparar cada segmento con la sintetización del mismo, utilizando un codificador LPC (*Linear prediction Coefficients*). La decisión sonoro-sordo se determina utilizando un modelo de Markov de dos estados (llamado modelo de Gilbert) oculto, encontrando la secuencia de estados óptima mediante el algoritmo de Viterbi. Finalmente, se realiza un post-procesamiento de la señal, de manera de corregir errores de estimación de f_0. Este post-procesamiento se realiza mediante programación dinámica, determinando el contorno óptimo del pitch, en un análisis global de toda la señal.

Se realizaron experimentos para evaluar el desempeño de las etapas antes mencionadas: SSPD, detección sonoro-sordo y post-procesamiento. Las pruebas se realizaron para señal limpia, utilizando la base de datos Keele. El resultado obtenido para SSPD fue una tasa de error de 2.63%. Para la detección sonoro-sordo, se logró un error de clasificación de 6.18%, mientras que para el post-procesamiento, se llegó a una tasa de error de 1.57%, para el mismo error de clasificación anterior. Ambos resultados son totalmente comparables con el estado del arte en extracción de pitch.

Agradecimientos.

Quisiera agradecer a mis padres, Marcelo y Judith, por todo el apoyo que me brindaron durante toda mi vida. Gracias por inculcarme valores; por ayudarme siempre; por asegurarse de que nada me faltase; y por todo el amor y cariño que he recibido de Uds. También quiero agradecer a mi hermano, Maximiliano, por toda la buena onda, las distracciones y el cariño brindado.

A mi profesor guía, Néstor Becerra, por haberme dado la oportunidad de realizar mi memoria. A todos los integrantes del laboratorio de procesamiento y transmisión de la voz, especialmente a Juan Pablo, por recibirme como lo hicieron, y por ayudarme cuando lo necesité.

A mis compañeros de la carrera. A mis amigos del Grupo Organizado BoletinSei, por las constantes juntas; por siempre estar ahí cuando se les necesitó; por tener siempre la talla precisa en el momento indicado. Gracias por haberme integrado a tan maravilloso grupo. A mis compañeros del liceo. A mis amigos de póker.

Al profesor Enrique Cordaro, por toda la confianza que depositó en mí en los 7 semestres en que fui parte de su cuerpo docente. También a los profesores David Laroze; Helmuth Thiemer; y Claudio Romero, por haberme dado la oportunidad de trabajar con ellos como auxiliar o ayudante.

A Victoria; a Catalina; y a todos los que de alguna manera, hicieron que me convirtiera en quién soy ahora. Gracias a todos ustedes.

… Dedicado a mis padres,

Marcelo y Judith

Índice General.

Índice de Figuras.

Índice de Tablas.

Capitulo 1.

Introducción.

1.1 Motivación

La prosodia es un tema muy importante para el aprendizaje de un idioma extranjero. La mayoría de los estudiantes de inglés como segundo idioma tienen buenas habilidades para escuchar; escribir; y, leer, pero usualmente presentan falencias en la pronunciación; fluidez; y, naturalidad. Esto ocurre porque no se pone énfasis en los aspectos que ayudan en esta área. Las características prosódicas más importantes son la entonación; la acentuación; y, la duración. Este trabajo se enfoca en la primera de ellas, la cual es probablemente la más importante, pues tiene muchas funciones, tales como proveer de naturalidad; la actitud que se quiere transmitir, etc.

Además, la enseñanza de segundo idioma asistida por computador ofrece enormes ventajas a los estudiantes como una herramienta complementaria a la instrucción presencial. La interactividad y el entretenimiento implícito en el software educativo hacen más efectivo el proceso de aprendizaje, ya que incrementan la motivación y permiten desarrollar actividades y ejercicios sin necesidad de la constante supervisión de un profesor.

1.2 Objetivos

En esta memoria, se pretende reforzar un sistema de evaluación de entonación mediante un algoritmo de estimación de pitch. Para esto, se desarrollará un método que conste de todas las etapas que debe tener un detector de pitch: extracción de pitch, basado en sintetización de señal mediante codificación LPC; clasificación sonoro-sordo, utilizando un modelo de Markov de dos estados (para sonoro y sordo); y post-procesamiento, consistente en un algoritmo de programación dinámica. El objetivo principal es desarrollar un método de estimación de pitch robusto, el cual, al ser evaluado debe, lograr un desempeño comparable con el estado del arte. No se pretende obtener el menor error, o el menor tiempo de procesamiento. Como se trata de la curva de entonación con fines educacionales, se permite que el sistema no sea muy preciso, en el sentido de que no se necesita el valor exacto del pitch en cada segmento, sino que se quiere la tendencia que sigue el pitch en el transcurso del tiempo.

Para lograr la meta planteada, se deben cumplir los siguientes objetivos específicos.

- Propuesta del método de estimación de pitch.
- Investigación de estimadores en el estado del arte
- Evaluar las distintas etapas con una base de datos conocida.
- Evaluar el desempeño del estimador en todas sus etapas
- Comparar el resultado con métodos en el estado del arte.

1.3 Estructura de la memoria

En el capítulo 2, el lector podrá interiorizarse en los temas concernientes a la teoría de los métodos desarrollados esta memoria. Se ofrece toda la información teórica necesaria

para comprender cada uno de los temas que se abordarán en este trabajo. Se comienza explicando el modelo de producción de la voz, y de ahí en adelante se presentan modelos del tracto vocal; se presenta el problema de estimación de la frecuencia fundamental; definiciones importantes; técnicas utilizadas en las distintas áreas de la ingeniería y ciencias; e información relevante para entender en detalle los métodos que se explicarán en el capítulo 3.

En el capítulo 3 se explica el sistema de evaluación de entonación existente, y se explica cuál es el aporte de este trabajo. Se explica en detalle todas las etapas involucradas en el proceso de estimación de la curva de entonación. En el capítulo 4 se realizarán experimentos para todas las fases mencionadas, para finalmente llegar a una configuración final, y experimentar con ella. Se comparará con métodos de extracción de pitch que están en el estado del arte y se discutirá su desempeño en términos de los indicadores que sean convenientes.

Finalmente, en el capítulo 5 se presentan las conclusiones generales que se pueden obtener de los resultados, y además se entregan las pautas a seguir para mejorar el trabajo presentado en este trabajo en un trabajo futuro

Capitulo 2.

Producción de la Voz y Estimación de la Frecuencia Fundamental.

2.1 Introducción

En este capítulo se introducirá al lector en los temas que sean necesarios para la comprensión de los temas concernientes a esta memoria de título. El tema principal consiste en extracción de pitch basado en un modelo de síntesis de voz, a través de un vocoder LPC. Por esta razón, es necesario entender el mecanismo de producción de la voz; definición de frecuencia fundamental o pitch; Algoritmos de estimación de pitch existentes; modelo del tracto vocal; modelos ocultos de Markov; clasificador Bayesiano; y vocoder LPC.

2.2 Producción de la Voz

2.2.1. Descripción de la producción de voz

Para entender el proceso de producción de voz es necesario conocer la anatomía relacionada, además de las funciones que cumplen los distintos órganos. El campo de la

fonética articulatoria estudia el cómo se producen los sonidos y cómo interactúan las estructuras del tracto vocal. Se explicará la función de la laringe y de los órganos articulatorios, que incluyen las cavidades faringeal y oral.

El aspecto fisiológico principal de la producción de la voz es el tracto vocal. Éste consiste en la faringe laringeal (bajo la epiglotis), la faringe oral (entre la epiglotis y el paladar blando), cavidad oral limitada por los labios, lengua y paladar), faringe nasal (sobre el paladar blando, al final de la cavidad nasal) y cavidad nasal.

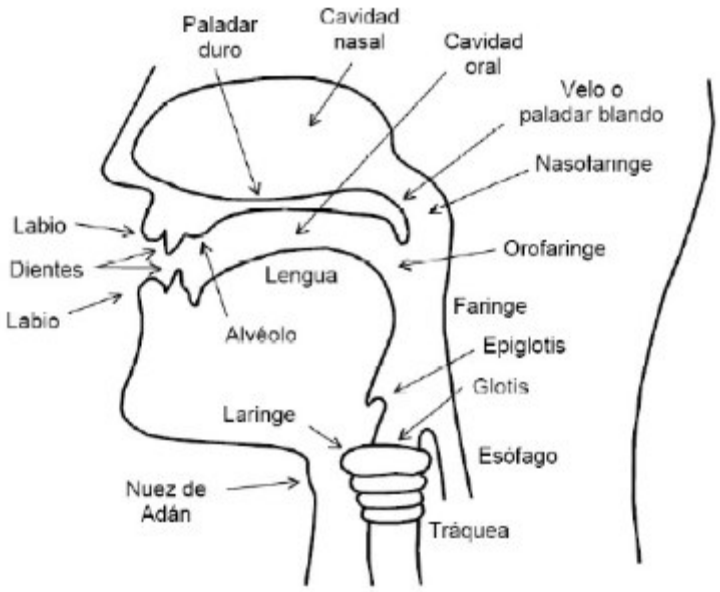

Figura 2.1. Tracto vocal.

2.2.2. Mecanismo de fonación

Las cuerdas vocales y el tracto vocal supra-segmental son las estructuras básicas para la producción de sonidos, y representan dos módulos independientes y controlables. Los pulmones proveen el flujo de aire necesario para superar la tensión de las cuerdas vocales cerradas. Las cuerdas vocales son ligamentos elásticos unidos dentro de las paredes de la laringe, las cuales se pueden abrir o cerrar a voluntad. El espacio entre las cuerdas se conoce como glotis, y su función es permitir o detener el paso de aire a través de la tráquea.

La vibración de las cuerdas vocales se conoce como fonación. Primero, las cuerdas se cierran, bloqueando el paso de aire desde los pulmones. La presión sub-glotal incrementa hasta vencer la resistencia de las cuerdas, y se vuelven a abrir. Luego se cierran rápidamente. Si el flujo de aire se mantiene relativamente constante, las cuerdas se abrirán y cerraran en forma periódica. Esta vibración viaja a través de la garganta y boca, donde es finalmente modificada para producir sonido.

Los sonidos producidos se pueden clasificar, de acuerdo a la fonación, como sonoros y sordos. Los sonidos sonoros se producen por la vibración de las cuerdas vocales. Todas las vocales, además de algunas consonantes como /m/, /n/, /l/ son sonoras. Los sonidos sordos carecen de vibración de las cuerdas vocales, y la producción se caracteriza por un flujo de aire que circule a través de la glotis abierta. Ejemplos de sonidos sordos son las consonantes /s/, /f/, /j/ (como en *jaula*). Los sonidos sordos se pueden dividir en dos tipos (Hess, 1983): no-sonoros, si existe una turbulencia en el tracto vocal produciendo una señal similar al ruido, sin vibración de las cuerdas vocales; y silencio, cuando no hay actividad del tracto vocal.

2.2.3. Tracto vocal supra-laringeal.

Muchos sonidos se producen con los órganos entre las cuerdas vocales y los labios. A este conjunto se le conoce como tracto vocal supra-laringeal, e incluye la cavidad oral; faringe; y, cavidades nasales. El tracto vocal es responsable del efecto resonante para la producción de vocales y consonantes, donde la forma adoptada por las cavidades influencia el sonido. Por ejemplo, las consonantes fricativas se producen por constricciones en el tracto vocal, como las consonantes /s/ y /z/, en las palabras en inglés *sign* y *zoo*. En estos sonidos, la parte frontal de la lengua se mueve para crear una constricción, en la cual el aire se vuelve turbulento mientras la atraviesa, creando lo que se conoce como “ruido acústico”.

2.2.4. Sonoridad y pitch.

El proceso sonoro es, casi en su totalidad, una contribución de la vibración de las cuerdas vocales, y la frecuencia de este patrón se conoce como *frecuencia fundamental*. El inverso de dicha frecuencia se define como *período fundamental*, y el abrir y cerrar de las cuerdas vocales se llama pulso glotal. La actividad de las cuerdas es una parte importante del estudio de la frecuencia fundamental. La tasa de vibración puede controlarse cambiando la tensión durante la etapa de vibración. Aumentar la tensión produce mayores frecuencias, mientras que ocurre lo contrario al dismsinuirla.

Hay estudios que definen las diferencias entre la frecuencia fundamental para hombres, mujeres y niños, dadas sus diferencias anatómicas. En lenguajes europeos,el promedio de la frecuencia fundamental es aproximadamente 120[Hz] para hombres, 220[Hz] para mujeres y 330[Hz] para niños.

El *pitch* se define como el fenómeno subjetivo de la percepción de la frecuencia fundamental, y no a un parámetro de la producción de la voz. La ecuación (2.1) muestra una relación matemática entre la frecuencia fundamental y el pitch

$$p(f) = p_{ref} + O \log_2\left(\frac{f}{f_{ref}}\right) \tag{2.1}$$

donde p_{ref} y f_{ref} corresponden a la frecuencia de un tono de referencia. La diferencia entre dos armónicos consecutivos es exactamente *O*. Asignando el valor 12 a esta variable, el pitch queda expresado en *semitonos*.

A pesar de las definiciones acústicas y perceptuales de frecuencia fundamental y pitch, ambos términos son utilizados indistintamente en la literatura.

2.2.5. Acústica de la voz.

Los parámetros acústicos, como la fonación y la periodicidad de la vibración de las cuerdas vocales, son considerados importantes en el proceso de producción de la voz. Existen cuatro procesos de la producción del habla, cuyos efectos acústicos se consideran independientes. Estos son: **fuente de sonido**, que puede ser la vibración de las cuerdas vocales, o bien una corriente turbulenta de aire; **filtro del tracto vocal**, el cual es el término acústico para la forma del tracto vocal; **pérdidas de energía**; y **radiación**, la forma en la que la onda de voz radía desde la boca.

La fuente de energía acústica está en la laringe. El ciclo abierto-cerrado de las cuerdas vocales se repite mientras exista actividad de fonación. Una representación de las características acústicas de la vibración de las cuerdas vocales es la forma de onda glotal, la cual es una representación física del volumen de aire que pasa a través de la glotis por unidad de tiempo. Consiste en tres fases: fase cerrada; fase de apertura; y fase de clausura. El período fundamental corresponde al tiempo que transcurre entre el inicio de la etapa de abertura hasta el final de la etapa cerrada.

El filtro vocal supra-laringeal actúa como un filtro acústico, que atenúa la energía de algunas componentes de frecuencia, y amplifica otras. En la producción del sonido, por ejemplo una vocal, el aire proveniente de los pulmones es interrumpido por las cuerdas vocales, lo que permite que una secuencia de pulsos de aire ingrese al tracto supra-laringeal. Dichas corrientes causan que el tracto vibre. La frecuencia y la amplitud a la que lo hacen dependen de la forma del tracto. Para explicar este fenómeno, se propone el siguiente modelo (González, 2005). El tracto vocal se puede aproximar por un conjunto de cilindros interconectados, con un largo específico y cambios insignificantes en los diámetros. El tracto vocal es representado como una función, especificada por la sección transversal y el largo de los cilindros. Mientras más cilindros se ocupen, mejor se aproximará el modelo. La forma en la que vibra el aire en el tracto, dada una forma particular, puede representarse por una función de transferencia. El modelo asume que no hay pérdidas de energía cuando el aire fluye a través del los cilindros.

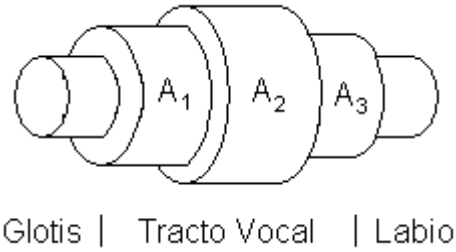

Figura 2.2. Modelo del Tracto vocal, utilizando 3 tubos.

2.3. Prosodia en la señal de voz

Las características prosódicas de la señal de voz son también denominadas características *supra-segmentales.* Esto significa que no están confinadas a una pequeña sección de la voz en particular, sino que éstas ocurren a un nivel más elevado. La prosodia intenta analizar y formalizar elementos como la entonación, acentuación, duración y ritmo. Estos últimos no son los únicos elementos prosódicos presentes en la expresión oral. En efecto, pueden ser considerados como prosodia otros elementos como aspiración, nasalidad o articulación. La producción de la prosodia se asocia con los parámetros prosódicos físicos, los cuales están constituidos principalmente por el pitch, la energía y la duración.

A continuación de describen el elemento prosódico más relevante para este trabajo, el cual es la entonación.

La entonación es la variación de pitch cuando el locutor habla. Todos los lenguajes usan semánticamente la entonación. Por ejemplo, para enfatizar, para transmitir sorpresa o ironía, o para plantear una pregunta. Lenguajes tonales, como el chino mandarín, además, usan la entonación para distinguir palabras. Entre las muchas funciones de la entonación, se pueden mencionar las siguientes (Arias et al, 2009).

- La entonación hace posible el expresar emociones y actitudes cuando se habla, y añade un tipo especial de significado al lenguaje hablado. Por ejemplo, la diferencia entre una petición y una orden recae principalmente en la entonación.

- La entonación tiene un rol significativo en asignar prominencia a sílabas que deben ser reconocidas como acentuadas. Esa sílaba, habitualmente dicha con un tono mayor, también es pronunciada con una mayor intensidad. Se utiliza para indicar contraste, o para enfatizar dicha sílaba.

- Cuando se usa información dada por entonación, el receptor reconoce con mayor facilidad la estructura sintáctica y gramatical de lo que se está diciendo. Es decir, dependiendo de la entonación, la misma frase puede tener resultados totalmente distintos.

- Considerando el acto de hablar desde una perspectiva más amplia, se puede observar que la entonación puede sugerir al receptor qué debe aceptar como *nueva* información y qué considerar como información dada, o también señalar que parte de la oración es en la que debe poner atención.

- Hay algunos idiomas, llamados lenguas tonales, en los que la función de la entonación es mucho más importante, tanto que un cambio en la melodía modifica el significado totalmente. Éste es el caso de numerosos lenguajes orientales, como por ejemplo, el chino mandarín y el vietnamita.

- Se puede incluir una sexta función, difícil de describir, pero reconocible por cualquier hablante nativo. El uso correcto de la entonación provee de *naturalidad* al discurso. El locutor nativo puede reconocer si una pronunciación la produjo uno nativo o no. Hay muchos aspectos que contribuyen a esto, de los cuales algunos son más distinguibles que otros: la elección de palabras, estructura sintáctica, características segmentales y, además, entonación y ritmo. Por muy competente que

pueda ser un hablante extranjero, si su entonación (y el ritmo) no es la que hubiera usado en la misma circunstancia uno nativo, su discurso no sonará natural.

2.4. Estimación de Frecuencia Fundamental.

La frecuencia fundamental es la correlación acústica de la tasa de vibración de las cuerdas vocales, y es directamente proporcional a ella. Es una de las características prosódicas más importantes, y se puede controlar voluntariamente por el locutor. La estimación de la frecuencia fundamental ha sido objeto de investigación por muchas décadas, lo que se observa en los numerosos métodos desarrollados.

2.4.1. Definición del Problema

La determinación confiable de la frecuencia fundamental f_0, o pitch, es una tarea difícil. Muchos algoritmos de detección de pitch (PDA, *pitch determination algorithm*) se han desarrollado a través de los años, pero sólo son exitosos bajo ciertas circunstancias, como por ejemplo, las condiciones en que se graba la voz, el ambiente en el que se está evaluando, etc. Además, algunos estimadores presentan errores, como por ejemplo, detectar el doble o la mitad de la frecuencia fundamental (conocidos como *Doubling* y *Halving* en la literatura) (Dziubinski et al, 2004), por lo que la investigación para encontrar un estimador confiable es aún un problema abierto (Hui et al, 2006).

El primer paso es segmentar la señal. Este proceso consiste en dividir la señal en sub-señales más pequeñas llamadas *segmentos*. La estimación de f_0 sobre cada intervalo de voz incluye la tarea de determinación sonoro-sordo, la cual se refiere a la clasificación de segmentos en sonoros o sordos. Estas dos tareas (detección de pitch y determinación sonoro-sordo) deben realizarla todos los PDA, pues para poder obtener f_0 para algún segmento sonoro, es necesario detectar los dichos segmentos. La vibración glotal puede presentar aperiodicidad debido a cambios en amplitud, tasa, o forma de onda del pulso

glotal, o en intervalos en los que los pulsos glotales ocurren sin regularidad en tiempo o amplitud. Estos factores hacen que la extracción de f_0 sea una tarea compleja.

2.4.2. Estimación de la frecuencia Fundamental

El objetivo es, básicamente, extraer f_0 de una señal de voz, conocido como el primer armónico de la señal. En señales periódicas, los armónicos están totalmente relacionados entre ellos, pues corresponden a múltiplos enteros del primero. El período fundamental T_0 es la brecha temporal en la señal entre dos pulsos glotales, lo que corresponde a las fases de apertura y clausura de la glotis. Conociendo el período fundamental, la frecuencia fundamental puede calcularse mediante $f_0 = 1/T_0$. En el análisis de una señal de voz, un período puede extraerse de una ventana cuya longitud debe ser aproximadamente T_0.

Un detector de pitch usualmente se compone de tres bloques principales: pre-procesamiento, extractor de pitch y post-procesamiento. La tarea básica del pre-procesamiento es la reducción de información para facilitar la extracción de pitch, tales como eliminación de ruido, sub-muestreo, etc. El extractor de pitch realiza la tarea principal, la cual es convertir una señal en una secuencia de estimaciones de f_0. El post-procesamiento realiza la corrección, detección de error y suavizado de la curva de f_0.

2.4.2.1. Métodos de extracción de Pitch.

Los algoritmos de estimación de f_0 más comunes procesan segmento a segmento la señal de voz. La propiedad principal de la estimación de f_0 es la periodicidad de la señal. Formalmente se define que una señal *s(x)* es periódica, con período *p* si

$$s(j) = s(j + n\,p) \tag{2.2}$$

donde *s(j)* corresponde a la j-ésima muestra de la señal discretizada, para $p > 0$, y para todo $n \in IN$. A continuación se describen los estimadores más utilizados

- **Tasa de cruces por cero *(ZCR, Zero-Crossing Rate*)**

ZCR es una medida de qué tan seguido una onda cruza por cero por unidad de tiempo. La idea es que ZCR entrega información sobre el contenido espectral de la onda (Kedem, 1986). Sin embargo, la presencia de armónicos y componentes de otras frecuencias hacen que en un ciclo existan varios cruces por cero, por lo que éste método ya está en desuso. Sin embargo, esta técnica es muy utilizada para detección sonoro-sordo.

- **Función de autocorrelación.**

El coeficiente de correlación (Dubnowski et al, 1976; Krubsack et al, 1999) es una medida de la similitud, o del grado de relación lineal entre dos funciones o variables de entrada . Es uno de los métodos más populares que se utilizan para estimar f_0. La función de autocorrrelación (ACF, *autocorrelation function*) se define a continuación

$$r(n) = \frac{1}{N} \sum_{j=1}^{N-n-1} s(j)\, s(j+n) \tag{2.3}$$

donde *n* es el retraso entre la señal instantánea y la misma señal retrasada. Dicha función mide la correlación entre las formas de onda de la misma señal en diferentes intervalos de tiempo. Si el segmento es periódico, la función de autocorrelación también lo será, mostrando un valor máximo para aquellos retardos iguales al período de la señal. El máximo global corresponde al retardo nulo, mientras que el segundo máximo debiese corresponde al retardo igual al período fundamental. En el caso de segmentos sordos, no se observa este efecto.

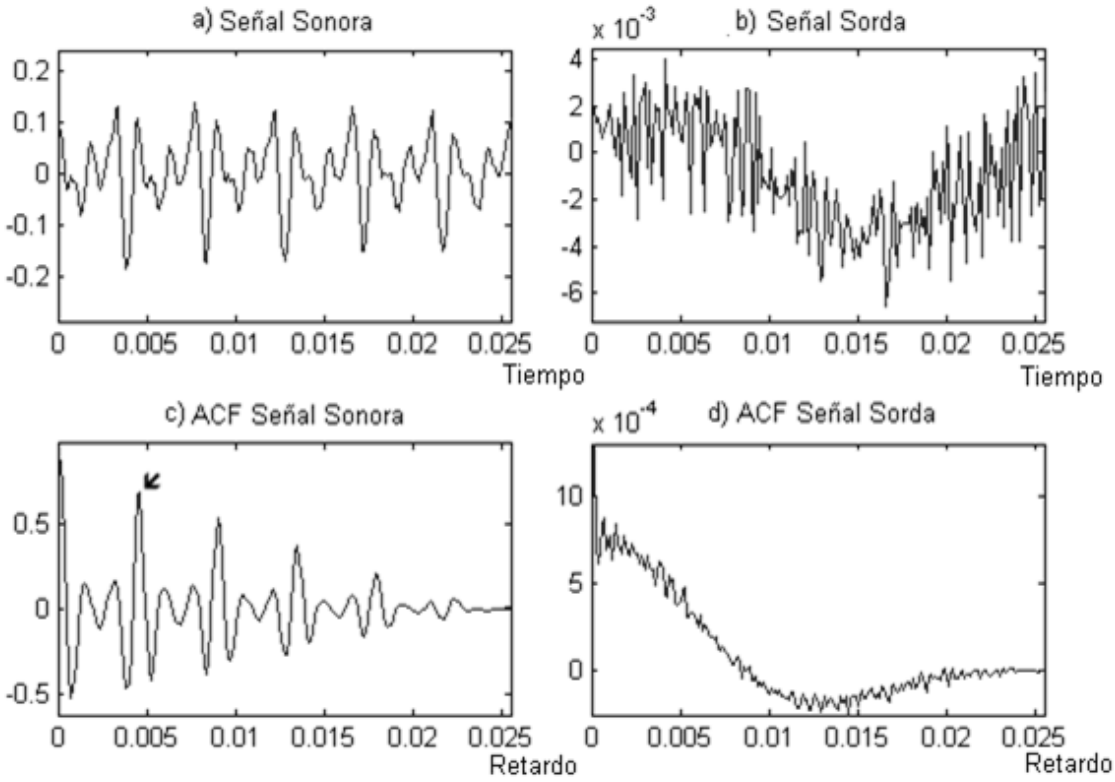

Figura 2.3. Señal sonora a) y su autocorrelación c). Se observa que el segundo peak se encuentraen t=0.0045[s]; señal Sorda b) y su autocorrelación d).

Para las señales cuasi-periódicas debe existir un peak similar para todos los múltiplos enteros del período fundamental. Para el estimador basado en ACF, la tarea principal es identificar el peak principal que corresponda al período T_0. Éste se espera que se encuentre en el rango de posibles valores de f_0 para la voz humana.

Otra importancia de la función de autocorrelación, es que los peaks de la función son distinguibles para segmentos sonoros y sordos. Es decir, dichos máximos son una buena medida de discriminación del tipo de segmento. Definiendo un umbral para el peak de la función, es posible tener una clasificación sonoro-sordo. Si el peak es superior al umbral, el segmento es sonoro. En caso contrario, el segmento es sordo, y no se considera el valor de f_0 calculado. En la Figura 2.4 se observa una señal de referencia de pitch y el peak de la autocorrelación para dichos segmentos.

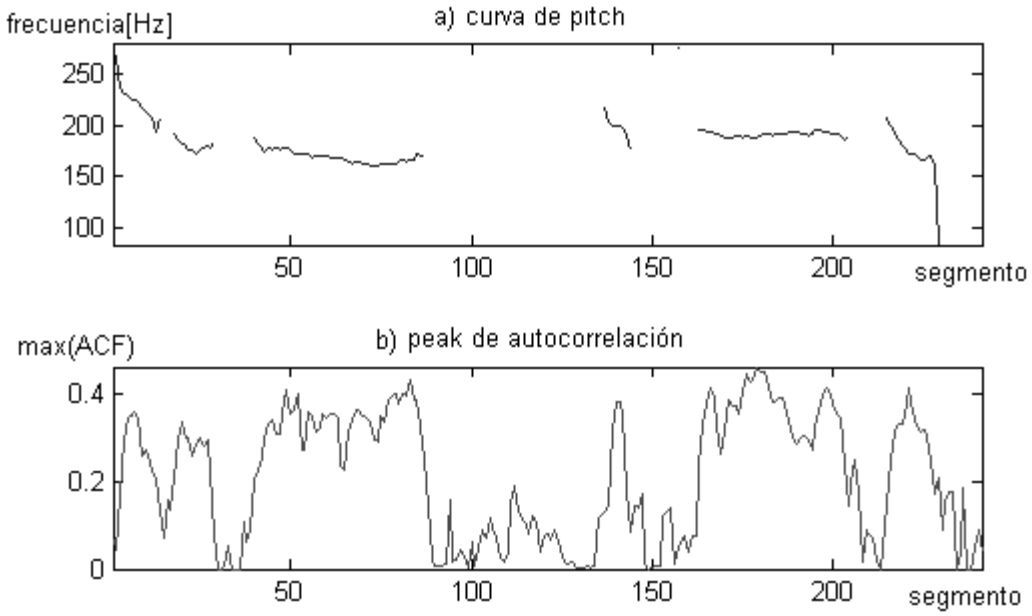

Figura 2.4. a) Pitch de referencia y b) peak de autocorrelación.

- **Función de diferencia de magnitud promedio (AMDF, *Average Magnitude Difference Function*).**

La función AMDF (Ross et al, 1974; Zeng et al, 2003), también llamada "anticorrelación", se define como

$$AMDF(n) = \frac{1}{N} \sum_{j=m}^{m+N-1} \left| s(j) - s(j+n) \right| \tag{2.4}$$

donde *N* es el tamaño del frame, *m* es la muestra inicial del frame, *n* es el retardo, y *s(j)* corresponde a las muestras de la señal de voz. Se basa en la distancia global entre dos funciones, en este caso, la señal y la misma señal desplazada en *n* muestras. La función ACF correlaciona la señal de entrada para varios retardos, mientras AMDF toma la diferencia de magnitud entre la señal retrasada y la señal original.

La función AMDF se obtiene a través de la sustracción de la señal desplazada y la original, y la suma de las diferencias de magnitudes entre ellas. Se esperaría que AMDF

tuviera un mínimo cuando el retraso n corresponde a T_0, si la señal es sonora. En el caso de una señal perfectamente periódica, el valor mínimo es cero. Si la señal es sorda, el comportamiento de la señal es indeterminado

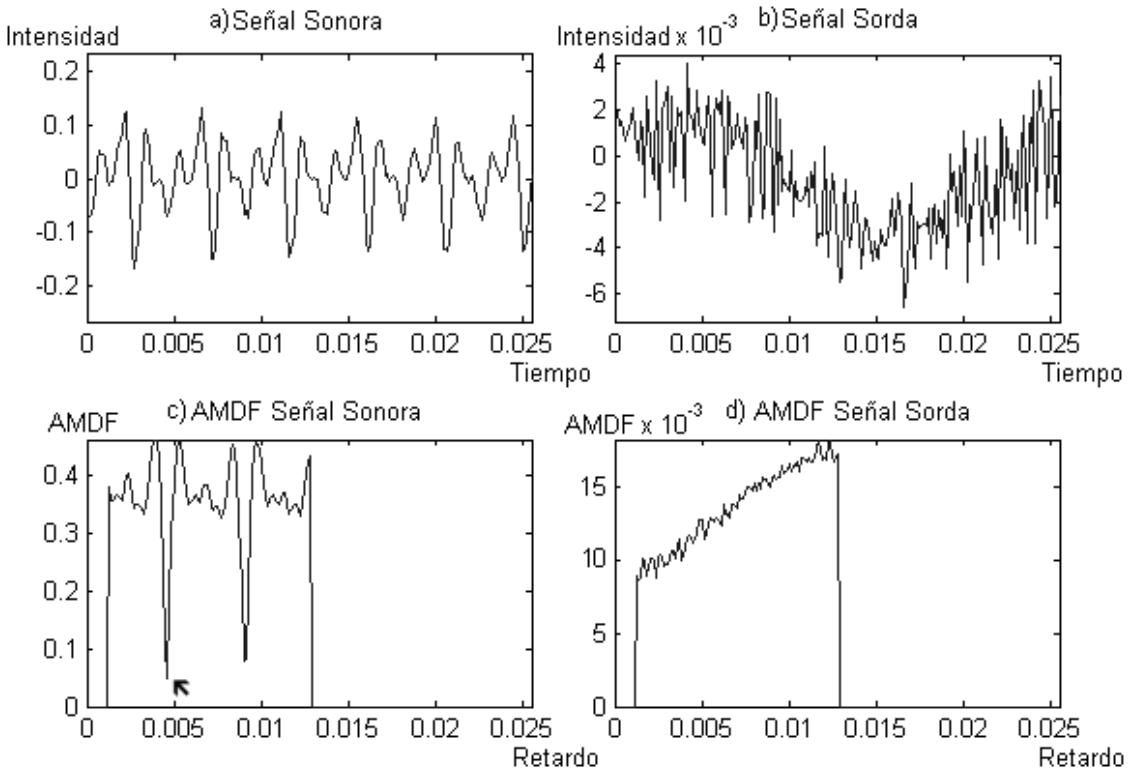

Figura 2.5. Cálculo de AMDF para señal a) sonora y b) sorda.

El algoritmo es sensible a cambios que pueden influir en la magnitud del mínimo en T_0, cambios de intensidad, ruido y señales de baja frecuencia. Además, a diferencia de ACF, AMDF no ofrece una referencia directa para clasificación sonoro-sordo. Finalmente, AMDF no requiere multiplicación matricial como ACF. Por esto, es importante recalcar que AMDF es ideal para estimadores de f_0 en tiempo real.

- **Estimador YIN**.

El estimador YIN (de Cheveigné et al, 2002) es similar a la función AMDF. Se define la función $d_t(n)$ como

$$d_t(n) = \sum_{j=1}^{N} \left(s(j) - s(j+n)\right)^2 \tag{2.5}$$

En virtud de reducir la ocurrencia de errores armónicos, YIN emplea una función auxiliar, que disminuye estos efectos indeseados.

$$d'_t(v) = \begin{cases} 1 & v = 0 \\ \dfrac{d_t(v)}{\frac{1}{v}\sum_{j=1}^{v} d_t(j)} & v \neq 0 \end{cases} \tag{2.6}$$

donde v corresponde al retardo de la señal. Finalmente, se incluye una etapa de interpolación parabólica y de eliminación de errores. Para mayor información, se insta al lector a revisar el artículo citado.

- **Análisis Cepstral.**

El *cepstrum* (Noll, 1967; Ahmadi et Al, 1999) es una transformada común utilizada para separar la señal de excitación y la función de transferencia. El cepstrum es la transformada de Fourier inversa del logaritmo del espectro de la señal.

$$c(n) = \Im^{-1}\left\{log\left(\left|\Im\left(s(n)\right)\right|\right)\right\} \tag{2.7}$$

donde $\Im(\cdot)$ corresponde a la transformada de Fourier, y $\Im^{-1}(\cdot)$ a su inversa. Como se trabaja en el dominio discreto, corresponden a la transformada discreta de Fourier (DFT, *Discrete Fourier Transform*) y a su inversa IDFT.

La variable independiente se denomina "cuefrencia" (del inglés, "*quefrency*", haciendo alusión a *frecuency*, frecuencia en inglés). La secuencia de pulsos en la señal

periódica aparece en el cepstrum como un peak marcado para un valor de cuefrencia que corresponde a T_0.

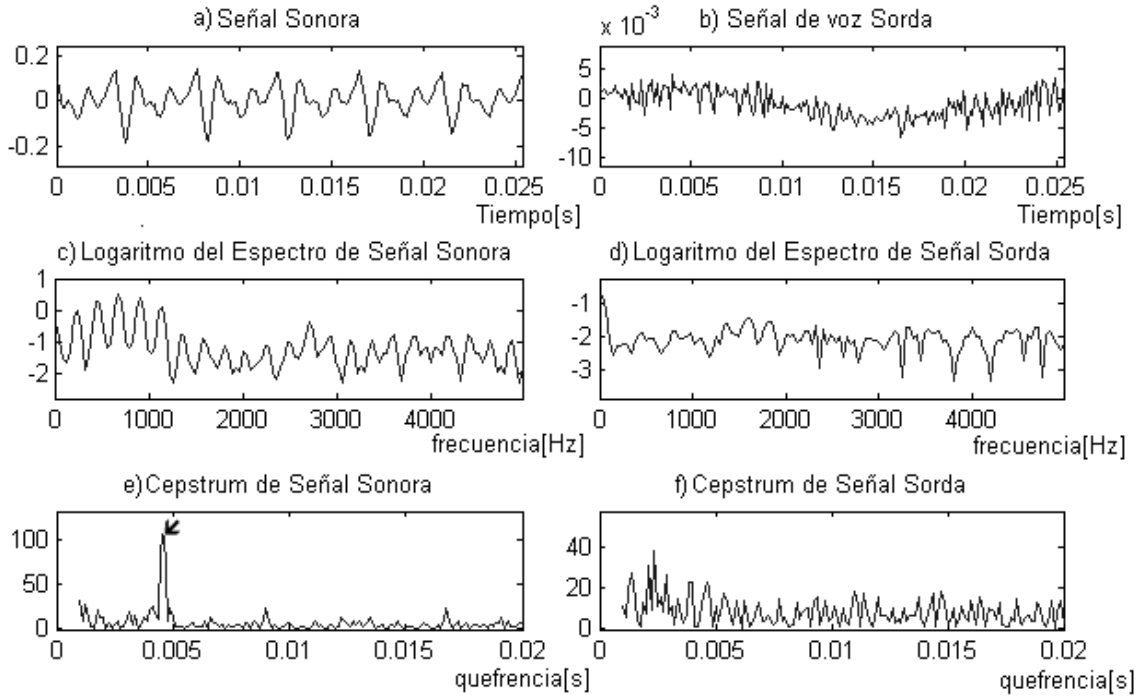

Figura 2.6. Etapas del análisis cepstral para señales a) sonora y b) sorda.

La señal sin procesar se divide en segmentos y se multiplica por una función ventana (usualmente, una ventana de Hamming). Luego, se calcula la DFT para cada segmento. Si la señal es periódica, un número regular de peaks representan el espectro armónico. El logaritmo del espectro de magnitud reduce los peaks y los traslada a una escala apropiada. La distancia entre los peaks está relacionada con la frecuencia fundamental de la señal. El paso final es aplicar un procedimiento de corrección para ajustar errores locales y para identificar transiciones de sonoro a sordo y viceversa. El error local más común es *doubling*, cuando se detecta el doble de la frecuencia real del segmento.

- **Método de multi-resolución.**

Este método (Cerdá et al, 1996) se utiliza para cualquier estimador de pitch que se calcule en el dominio de la frecuencia: Si la exactitud de un algoritmo a cierta resolución es

de alguna manera sospechoso, se pone a prueba el estimador utilizando el mismo algoritmo, pero a resoluciones distintas. Así, se puede usar, por ejemplo, una ventana de tiempo mayor (o menor) para calcular el espectro, un mayor (o menor) número de puntos para calcular la FFT, etc. Si el resultado es el mismo para la mayoría de las resoluciones, entonces se confirma el resultado. De lo contrario, se desecha la hipótesis.

2.4.2.2. Detección Sonoro-Sordo.

La determinación sonoro-sordo consiste en clasificar los segmentos de voz en sonoros y sordos. Algoritmos para esto clasifican a la fuente de la locución: si las cuerdas vocales vibran, el segmento es sonoro. De otra manera, es sordo. Algoritmos de estimación de f_0 pueden ser aplicados a segmentos sordos para obtener cierta frecuencia fundamental. En ese caso, la determinación sonoro-sordo puede implementarse después de la estimación de f_0 como parte del post-procesamiento. Los métodos más utilizados son por reconocimiento de alguna característica de la señal, ubicando un umbral de decisión que minimice los errores cometidos. Los parámetros más utilizados son.

- *Energía* (Atal et al, 1976): Los clasificadores por energía han sido muy utilizados por su simplicidad de cálculo. Los segmentos sonoros en general tienen mayor energía que los sordos. La Energía se define como

$$E = \sum_{i=1}^{N} \left| x(i) \right|^2 \tag{2.8}$$

- *Tasa de cruces por cero* (Maqsood et al, 2007): Las señales sordas, al ser poco correlacionadas, tienen un ZCR más alto que las señales sonoras. Sin embargo, es más susceptible a errores en presencia de efectos externos, como ruido, componentes continuas, etc.

- *Energía dividida por tasa de cruces por cero* (Kotnick et al, 2006). Esta característica es muy utilizada, pues mejora el método de energía sin aumentar demasiado la carga computacional. Como las señales sonoras tienen alta energía y baja tasa de cruces por cero, al revés que las señales sordas, al dividir ambas características mejora la clasificación.

- *Peak de autocorrelación* (Krubsack et al, 1999): Las señales sonoras, como son altamente correlacionadas, se espera que tengan un peak de autocorrelación en el período fundamental. Este máximo debiera ser mayor que el de una señal sorda, que casi no presenta correlación consigo misma.

- *Peak del Cepstrum* (Ahmadi et Al, 1999): El cepstrum de una señal sonora tiene un peak marcado en el período fundamental de la señal, mientras que para las señales sordas, es en general menos marcado.

2.5. Codificador de Voz

Se conoce como codificación de voz a la conversión de la señal análoga de voz a una forma digital. El objetivo principal es lograr comprimir dicha señal, es decir, emplear la menor cantidad de bits como sea posible para aquella representación digital. Dicha representación hace posible alcanzar anchos de banda muy bajos en la transmisión, o bien lograr un almacenamiento eficiente de la señal. Como la señal digital finalmente se vuelve a pasar a análoga, una consideración importante de la codificación de voz es el nivel de distorsión de la señal introducida por el proceso de digitalización.

Dentro de la gran variedad de codificadores que se han propuesto, analizado y desarrollado, los codificadores más utilizados (Deller et al, 2000), gracias a su alta calidad y bajas tasas de bits empleadas, son los que se conocen como *vocoders* (del inglés, contracción de *voice coders,* codificadores de voz), dentro de los cuales el más utilizado es el *vocoder* LPC, que se explica en la sección siguiente.

2.5.1 Vocoder LPC.

La codificación/decodificación LPC (*Linear Prediction Coefficient*, coeficientes de predicción lineal) (Atal et al, 1971) es un intento de aprovechar el hecho que en la señal de voz, luego del proceso de discretización, sus muestras presentan una fuerte correlación, asumiendo una frecuencia de muestreo apropiada. Cualquier señal de estas características es una buena candidata para predicción lineal, en la cual las muestras actuales de la voz son modeladas como una combinación lineal de muestras pasadas (Beauchaine, 2004). La minimización del error cuadrático medio entre las muestras predichas y su valor real llevan a un conjunto único de coeficientes para un orden del filtro dado.

El modelo de producción del habla para el tracto vocal humano puede ser bien modelado como la salida de un sistema lineal, variante en el tiempo, excitado por un tren de impulsos cuasi-periódicos, ruido aleatorio gaussiano, o una combinación de ambos. La mayoría de los de los modelos de producción de la voz humana consideran el tracto vocal como una serie de tubos variables en el tiempo, de varias longitudes y secciones transversales. La fuente de excitación produce la entrada en el tubo de uno de los extremos, y el sonido se propaga a través de la serie de tubos, rebotando en las paredes, reflectándose en las junturas de cada nueva sección. Esta visión simplificada del modelo de producción produce un filtro todo-polos.

Si el tracto vocal puede simplificarse de esta manera, se necesita un modelo para la excitación de manera de completar esta visión de la producción lineal de la voz. Análisis de formas de onda de la voz han mostrado que, en el caso más general, la voz se puede dividir en dos categorías principales. El primer tipo se denomina voz *sonora*[1]. Es producida por la excitación periódica del tracto vocal en la laringe, a través de la vibración de las cuerdas vocales, excitadas por el paso de aire proveniente de los pulmones. La voz sonora de caracteriza por una frecuencia fundamental seudo-estacionaria, más armónicos. La sonoridad está asociada a sonidos producidos por un tracto vocal "abierto". Las vocales proveen el mejor ejemplo de sonoridad.

[1] En inglés, *voiced speech.*

La otra categoría principal se denomina voz *sorda*[2]. Ésta se produce por una turbulencia cuando el aire es forzado a través de una constricción en el tracto vocal. Esto puede producirse al tener los labios entrecerrados, al poner la lengua contra los dientes, u otras combinaciones. La voz sorda se asemeja bastante al ruido aleatorio, mostrando poca periodicidad y muy poca correlación entre muestras.

Ahora que se conoce el modelo básico de la producción de la voz humana, se puede explicar en qué consiste el sistema básico de codificación/decodificación de voz LPC. Primero, la voz es muestreada a una frecuencia apropiada para capturar todas las componentes de frecuencia necesarias para el procesamiento. Para transmisión de voz, 10[kHz] es típicamente la frecuencia elegida. Esto pues, para la mayoría de los locutores, toda la energía significativa de la voz se encuentra bajo los 4[kHz] (aunque las señales de algunas mujeres y niños no cumplen con esta suposición). La voz es luego dividida en segmentos para el procesamiento. Los codificadores LPC simples usan segmentos de igual longitud, entre 10[ms] y 30[ms]. Menos de 10[ms] no alcanza a cubrir un período completo de algunas voces de baja frecuencia de hombres. Más de 30[ms] viola el principio básico de estacionareidad en el cual se basa el método.

Una vez segmentada, se determina si el tipo de voz es sonoro o sordo. Esta tarea es, en realidad, el tema más complicado para los vocoders LPC, como se puede apreciar en la gran cantidad de publicaciones desarrolladas en este tema en los últimos años. Si el frame es clasificado como sonoro, entonces se debe determinar la frecuencia fundamental. Esto se produce pues el oído humano, a pesar de ser tolerante a varios errores en la codificación de voz, es, por alguna razón, muy sensible a los errores en el pitch. Señales codificadas con el pitch mal calculado es bastante desagradable y típicamente suena artificial. Una vez que el tipo de voz es determinado, se estiman los parámetros LPC. Un tema crucial en el modelo de predicción es el orden de predicción del filtro (número de muestras a utilizar en el proceso de estimación). Usualmente se utilizan modelos de órdenes entre 10 y 14. Una vez que se determinan los parámetros LPC, se calcula la energía del segmento se señal. Esto es necesario porque al final del proceso de síntesis se utiliza la energía para determinar la

[2] En inglés, *unvoiced speech.*

ganancia de la señal sintetizada. Además, la energía es un buen discriminador del tipo de voz.

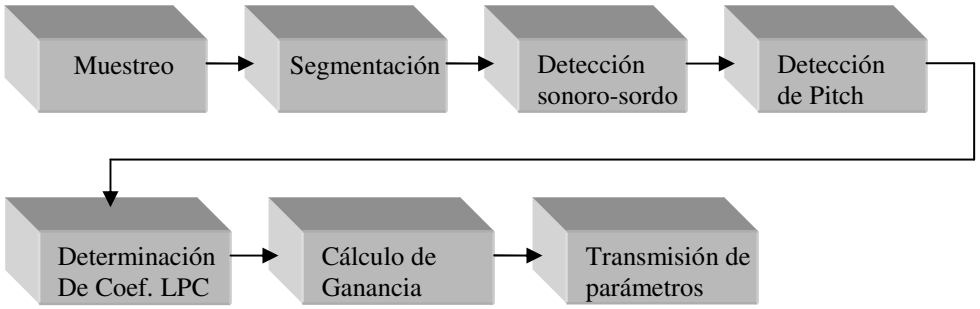

Figura 2.7. Diagrama de bloques del codificador LPC

Al final de este proceso, se tiene un modelo del segmento de voz, que determina el tipo, la energía, el pitch y los coeficientes LPC. Es importante recalcar que, para un vocoder LPC, no se transmite nada de la señal de voz original. El modelo es suficiente para el que el receptor sea capaz de sintetizar la señal, la cual puede ser una copia bastante similar a la original. Esto permite que los vocoder LPC alcancen tasas de bits muy bajas. Sistemas comerciales disponen de tasas de hasta 2.4 kbps con niveles de calidad aceptables.

En el receptor, los parámetros transmitidos son decodificados (Figura 2.8). Dependiendo del tipo de voz, la fuente de excitación puede ser una fuente de ruido aleatorio o un tren de impulsos periódicos, cuyo período es también transmitido por el codificador. La fuente de excitación es escalada apropiadamente por la ganancia transmitida, y se pasa a través del filtro formado por los coeficientes LPC para producir voz sintetizada.

2.5.2 Filtro LPC

Como se mencionó anteriormente, el filtro LPC es un filtro todo polos que modela el modelo de producción del habla del tracto vocal, tal como se muestra en la Figura 2.9.

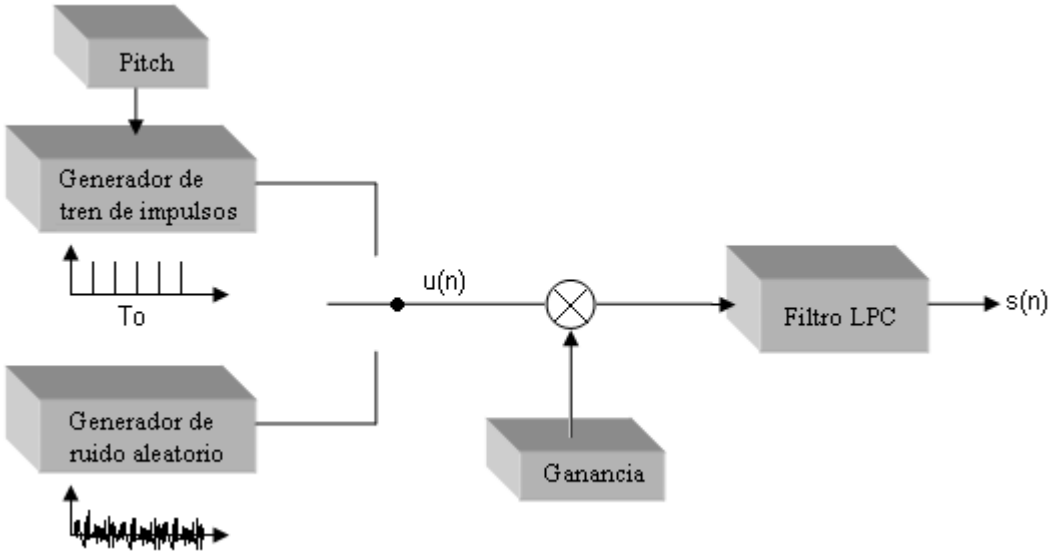

Figura 2.8. Diagrama de bloques de decodificador LPC

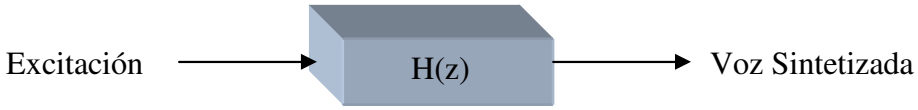

Figura 2.9. Modelo simplificado del decodificador LPC

El filtro se define mediante (2.9), donde los términos a_k corresponden a los coeficientes del filtro. En el dominio de la frecuencia, el filtro LPC modela la envolvente del espectro de la señal.

$$H(z) = \frac{1}{a_0 + a_1 z^{-1} + \cdots + a_P z^{-P}} = \frac{1}{\sum_{k=0}^{P} a_k z^{-k}} \tag{2.9}$$

Las frecuencias a las que ocurren los peaks que se observan en el espectro de una señal sonora se conocen como *frecuencias formantes*, o simplemente *formantes.* En estos se concentra la mayor parte de la energía de la señal, y sirve para diferenciar los distintos fonemas sonoros. Además, cada fonema tiene una distribución diferente para cada hablante,

por lo que hace posible que la voz de una persona sea reconocible al ser reproducida en el decodificador LPC.

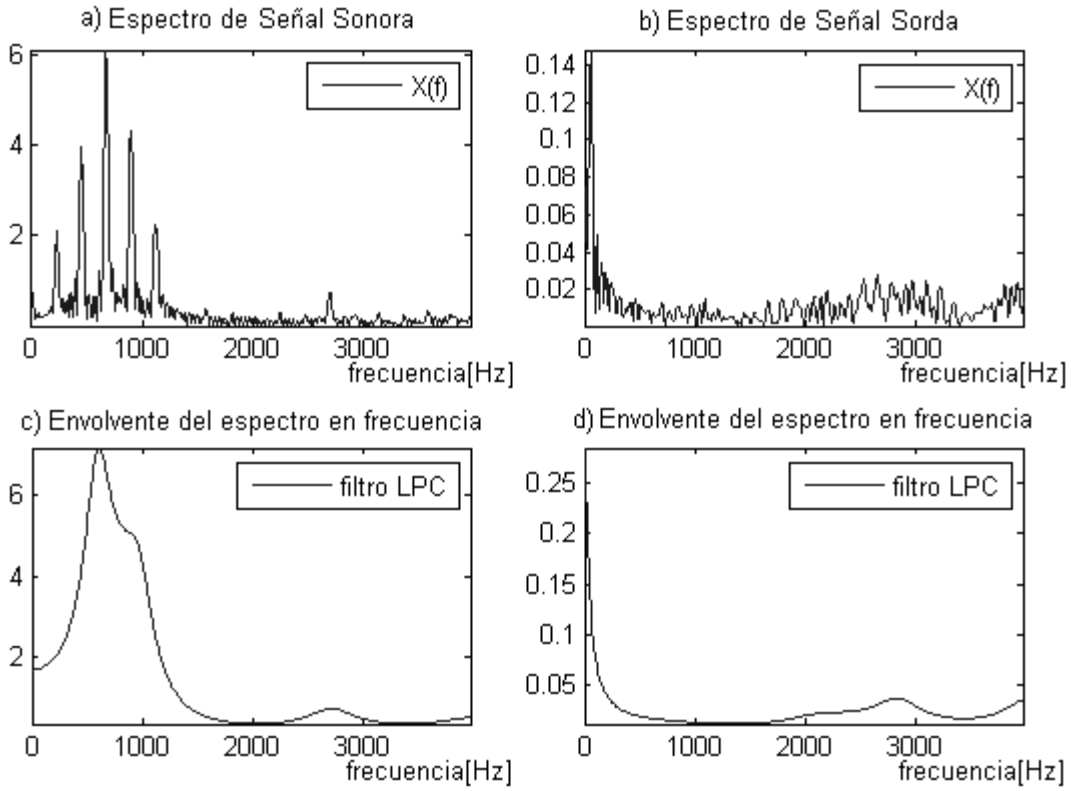

Figura 2.10. Comportamiento espectral del filtro LPC. Se muestran los espectros de segmentos a) sonoro y b) sordo, y el filtro LPC en el dominio de la frecuencia para dichos segmentos, para 8 coeficientes LPC. La señal está muestreada a 8[kHz].

2.5.3. Cálculo de coeficientes LPC.

Lo que aún no se ha explicado es como se determinan los coeficientes LPC. Sólo se mencionó que se iban a calcular mediante la minimización del error cuadrático medio. Éste se define, en este caso, como

$$\varepsilon = \frac{1}{N}\sum_{k=1}^{N}\left(s(k) - \sum_{i=1}^{P} a_i \cdot s(k-i) \right)^2 \qquad (2.10)$$

donde *P* es el orden del modelo, *N* es el largo del segmento de señal en consideración, a_i es el i-ésimo coeficiente LPC, y $s(k)$ es la muestra k-ésima del el segmento de análisis. Derivando esta ecuación respecto a los coeficientes LPC, e igualando a cero,

$$\frac{\partial \varepsilon}{\partial a_i} = 0 \quad \Rightarrow \quad \sum_{i=1}^{P} a_i \left(\sum_{k=1}^{N} s(k-i)s(k-j) \right) = \sum_{k=1}^{M} s(k)s(k-j) \tag{2.11}$$

Cambiando los índices en las sumatorias de la ventana de análisis a $+\infty$ y $-\infty$, bajo el supuesto implícito que todas las muestras fuera del segmento actual son cero.

$$\sum_{i=1}^{P} a_i \left(\sum_{k=-\infty}^{+\infty} s(k-i)s(k-j) \right) = \sum_{k=-\infty}^{+\infty} s(k)s(k-j) \tag{2.12}$$

Sustituyendo m = k-j

$$\sum_{i=1}^{P} a_i \left(\sum_{m=-\infty}^{+\infty} s(m-i+j)s(m) \right) = \sum_{m=-\infty}^{+\infty} s(m+j)s(m) \tag{2.13}$$

Definiendo el término de autocorrelación

$$R(j) = m \sum_{m=-\infty}^{+\infty} s(m)\, s(m+|\, j\,|) \tag{2.14}$$

Se obtiene finalmente que

$$\sum_{i=1}^{P} a_i R(j-i) = R(j) \tag{2.15}$$

Evaluando para $j = 1,2,..P$ se llega al siguiente sistema de ecuaciones.

$$\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & \cdots & R(p-2) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ R(P-1) & R(P-2) & R(P-3) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ \vdots \\ R(P) \end{bmatrix} \tag{2.16}$$

Escrito en su forma compacta, $R \cdot a = r$, su solución será $a = R^{-1} \cdot r$. Sin embargo, usualmente se utiliza el algoritmo de recursión de Levinson-Durbin para determinar los coeficientes LPC, pues dicho algoritmo utiliza la simetría existente en la matriz *R*, disminuyendo el número de cálculos. El algoritmo se explica a continuación

Inicialización

$$E_0 = R(0) \tag{2.17}$$

$$a_{1,1} = k_1 = \frac{R(1)}{E_0} \tag{2.18}$$

$$E_1 = E_0\left(1 - k_1{}^2\right) \tag{2.19}$$

Recursión.

Para $2 \le m \le P$ (2.20)

$$(i) \quad q_m = R(m) - \sum_{i=1}^{m-1} a_{i,(m-1)} R(m-1) \tag{2.21}$$

$$(ii) \quad k_m = \frac{q_m}{E_{(m-1)}} \tag{2.22}$$

$$(iii) \quad a_{m,m} = k_m \tag{2.23}$$

$$(iv) \quad a_{i,m} = a_{i,(m-1)} - k_m a_{(m-i),(m-1)} \quad para\ i = 1,..,m-1 \tag{2.24}$$

$$(v) \quad E_m = E_{m-1}\left[1 - k_m{}^2\right] \tag{2.25}$$

$$(vi) \quad m = m + 1 \tag{2.26}$$

donde para $a_{j,k}$, el primer subíndice corresponde a la iteración j de la recursión, mientras que el segundo indica que es el k-ésimo coeficiente.

2.5.4. Mejoramiento de la envolvente LPC.

El filtro describe bastante bien la envolvente del espectro de una señal, con la configuración de coeficientes apropiada. Sin embargo, hay ocasiones en que dicha configuración no representa de buena manera los espectros de todas ellas. Por ejemplo, si la señal es muy armónica (es decir, si su espectro tiene muchas componentes de frecuencia), se necesita un número de coeficientes LPC relativamente alto para describir la envolvente. Por otro lado, si la señal tiene pocas componentes de frecuencia (en particular, el caso en que tiene una frecuencia dominante), se necesitan pocos coeficientes LPC. En el primer caso, pocos coeficientes no serían capaces de describir bien la envolvente. En el segundo, el filtro LPC sería demasiado ajustado. Por esta razón, se desarrollo un método iterativo para resolver este problema (Villavicencio et al, 2006). El método consiste en lo siguiente: Teniendo calculados los espectros de magnitud de la señal, *X* y el filtro LPC, *H*, se realiza el siguiente método iterativo

Inicialización:

Determinar máximos de X y H al inicio

Definir H'=H y X'=X.

Iteración:

normalizar X' y H' a la unidad.

Encontrar el máximo punto a punto entre X y H', y actualizarlo en H'.

Suavizar H'.

Término:

normalizar H' al máximo inicial de H, y guardarlo en H.

Matemáticamente, se puede pobservar en las ecuaciones (2.27) a (2.36)

$$maxH = \max_{1 \le j \le M}(H(j)); \tag{2.27}$$

$$maxX = \max_{1 \le j \le M}(X(j)); \tag{2.28}$$

$$X'(j) = \frac{X(j)}{maxX} \qquad 1 \le j \le M \tag{2.29}$$

$$H'(j) = \frac{H(j)}{maxH} \qquad 1 \le j \le M \tag{2.30}$$

$$\textit{while } i < N \tag{2.31}$$

$$H'(j) = max\left(H'(j), X'(j)\right) \qquad 1 \le j \le M \tag{2.32}$$

$$smooth\left(H'\right) \qquad 1 \le j \le M \tag{2.33}$$

$$H'(j) = \frac{H'(j)}{max\left(H'(j)\right)} \qquad 1 \le j \le M \tag{2.34}$$

$$\textit{end} \tag{2.35}$$

$$H(j) = \frac{H'(j)}{max\left(H'(j)\right)} \cdot maxH \qquad 1 \le j \le M \tag{2.36}$$

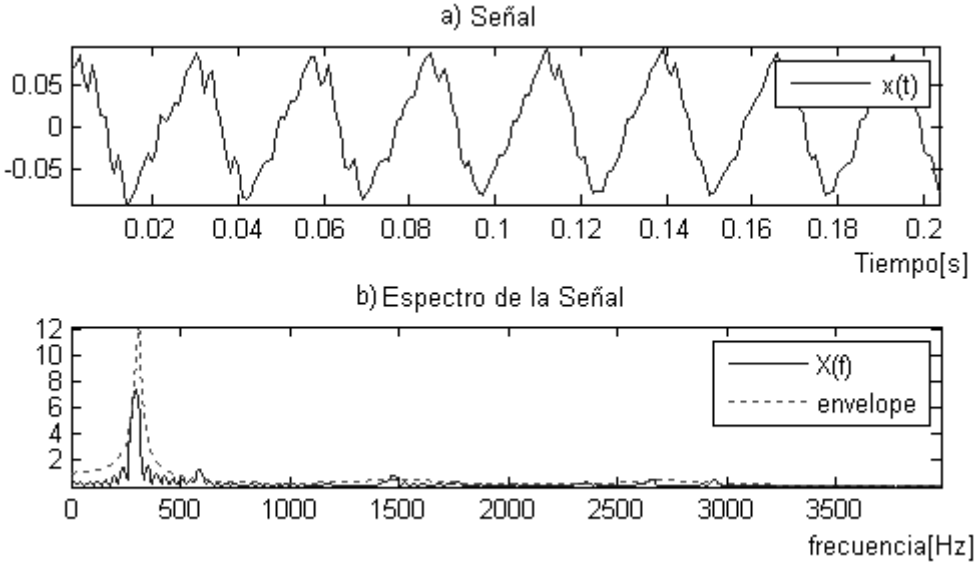

Figura 2.11. Señal a) en el dominio del tiempo, b) en el dominio de la frecuencia, con pocas componentes armónicas, codificada con muchos coeficientes LPC. Frecuencia de muestreo 8[kHz], 8 coeficientes LPC.

Donde *N* es el número de iteraciones que requiere el método para converger; y *M* es el número de puntos que definen los espectros. *smooth(·)* es una función suavizadora. A continuación, se muestra el efecto producido sobre la misma señal anterior, utilizando como función suavizadora la convolución entre la señal con una ventana gaussiana de 3 puntos.

$$w_{Gauss} = [0.0439, 1, 0.0439] \tag{2.37}$$

$$smooth(H') = conv(H', w) \tag{2.38}$$

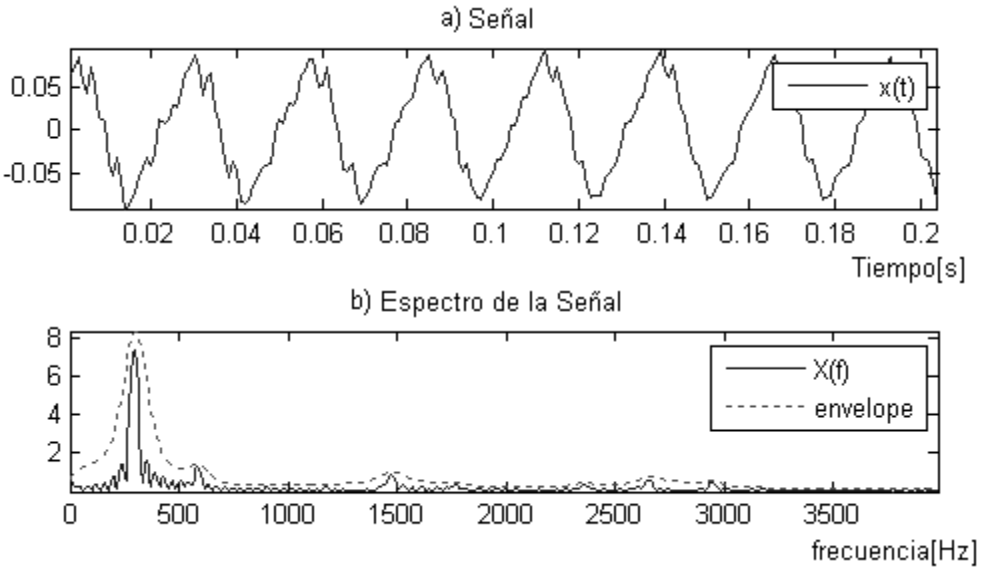

Figura 2.12: Señal a) en el dominio del tiempo, b) en el dominio de la frecuencia, con pocas componentes armónicas, codificada con muchos coeficientes LPC, después del método de mejoramiento LPC usando una ventana gaussiana de 3 puntos y 400 iteraciones. Frecuencia de muestreo 8[kHz], 8 coeficientes LPC

Notar que la ventana, a pesar de ser muy conservadora, mejora considerablemente el espectro del filtro LPC.

2.6 Teoría de Decisión de Bayes.

La teoría de decisión de Bayes es un acercamiento estadístico fundamental en el problema de reconocimiento de patrones (Duda et al, 1973). Este se basa en la suposición que el problema de decisión se puede resolver en términos probabilísticos, y que todos los valores probabilísticos relevantes son conocidos.

Se considera el caso en que se necesita un diseñar clasificador para separar dos tipos de estados diferentes. Sólo existe la posibilidad que sea uno u otro estado. Se denotan ambos estados como $\omega = \omega_1$ para el estado 1, y $\omega = \omega_2$ para el estado 2. Se considera ω como una variable aleatoria.

Lo primero es calcular las probabilidades *a priori* de que la característica pertenezca al estado 1, $P(\omega_1)$, y otra probabilidad a priori de que pertenezca al estado 2, $P(\omega_2)$. Estas probabilidades reflejan el conocimiento previo que se tiene de cuán probable es la pertenencia a alguno de los grupos sin siquiera haber realizado la observación. Obviamente, $P(\omega_1)$ y $P(\omega_2)$ son estrictamente positivos, y suman 1. Si se supone que se está forzado a tomar una decisión sobre el estado al que pertenecerá cierto patrón sin tener la posibilidad de observarlo, y que la única información que se permite utilizar es el valor de las probabilidades a priori, es razonable pensar que, con tan poca información, se utilice la siguiente regla de decisión.

$$\begin{aligned} &\textit{decidir } \omega_1 \quad \textit{si } P(\omega_1) > P(\omega_2) \\ &\textit{decidir } \omega_2 \quad \textit{en otro caso}, \end{aligned} \tag{2.39}$$

Cuán bien funciona este procedimiento depende del valor de las probabilidades a priori. Si $P(\omega_1)$ es mucho mayor que $P(\omega_2)$, la decisión de escoger ω_1 será correcta en la mayoría de los casos. Si $P(\omega_1) = P(\omega_2)$, la probabilidad de elegir correctamente es de un 50%, ya sea eligiendo ω_1 o ω_2. En este escenario, la probabilidad de error es mínima utilizando este criterio.

En la mayoría de los casos, se puede tomar decisiones con más información. Cuando se conocen observaciones de los patrones obtenidas previamente, y si sabe a cuál estado pertenecen, es natural expresarlos en términos probabilísticos. Se denomina a dichos patrones como *x*, y se consideran como variables aleatorias continuas, cuya distribución depende del estado. Sea $p(x|\omega_j),\ j = 1,2$ la función de densidad de probabilidad

condicional respecto al estado para *x*. Esto es, la densidad de probabilidad condicional para *x* dado el estado de procedencia ω_j Un ejemplo de esto, se muestra en la Figura 2.13.

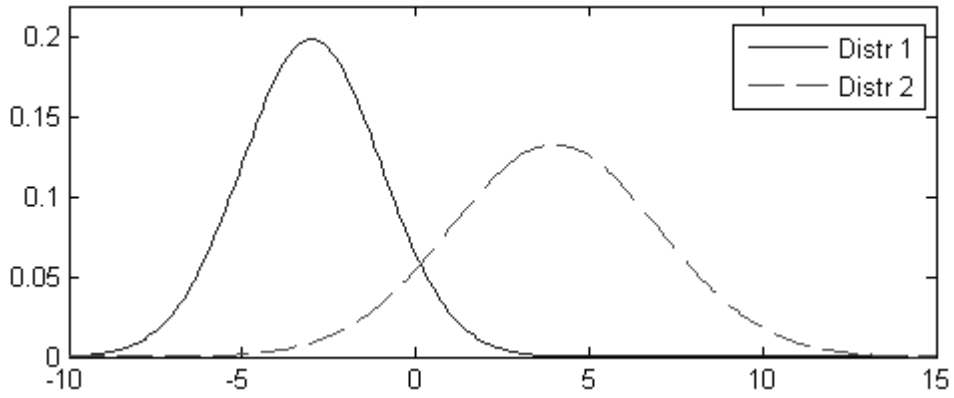

Figura 2.13. Funciones de densidad de probabilidad condicionales para cierto fenómeno. Distr 1 corresponde a $p(x|\omega_1)$, mientras que Distr 2 a $p(x|\omega_2)$.

Si ahora se supone que se conocen las probabilidades a priori $P(\omega_j)$ y las densidades condicionales $p(x|\omega_j)$, y se asume que se conoce el valor de una observación del fenómeno en cuestión, obteniendo como resultado *x*. La influencia de esta medida puede dar una noción del estado al que pertenece dicha observación a través de la regla de Bayes

$$p(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$
$$donde\ p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j) \tag{2.40}$$

La regla de Bayes muestra cómo observaciones del patrón, *x*, cambian la probabilidad a priori $P(\omega_j)$ a la probabilidad a posteriori $p(\omega_j|x)$. Si se tiene una observación *x* para la cual $p(\omega_1|x)$ es mayor que $p(\omega_2|x)$, uno se inclinaría por decidirse

por el estado de procedencia ω_1. Se puede probar que esta decisión minimiza la probabilidad de error, definida como

$$p(error|x) = \begin{cases} p(\omega_1|x) & si\ se\ decide\ \omega_2 \\ p(\omega_2|x) & si\ se\ decide\ \omega_1 \end{cases} \tag{2.41}$$

Por lo tanto, se tiene la siguiente Regla de Decisión de Bayes que minimiza la probabilidad de error,

$$\begin{array}{ll} decidir\ \omega_1 & si\ p(\omega_1|x) > p(\omega_2|x) \\ decidir\ \omega_2 & en\ otro\ caso \end{array} \tag{2.42}$$

Esta forma de la regla de decisión enfatiza el rol de las probabilidades a posteriori. Usando (2.42), se puede expresar en términos de las probabilidades a priori y de la densidad de probabilidad condicional. Notar que el término $p(x)$ es irrelevante al momento de tomar una decisión. Es básicamente un factor de escala que asegura que $p(\omega_1|x) + p(\omega_2|x) = 1$. Eliminando ese factor de escala, se obtiene la siguiente regla de decisión, totalmente equivalente a la anterior

$$\begin{array}{ll} decidir\ \omega_1 & si\ p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2) \\ decidir\ \omega_2 & en\ otro\ caso \end{array} \tag{2.43}$$

Existen algunas consideraciones especiales para casos particulares. Por ejemplo, si para algún *x*, $p(x|\omega_1) = p(x|\omega_2)$, la observación particular no ofrece ninguna información del estado del cual proviene. En este caso, la decisión recae totalmente en las probabilidades a priori. Por otro lado, si $P(\omega_1) = P(\omega_2)$, luego los estados son equiprobables a priori. En este caso, la decisión se basa completamente en $p(x|\omega_j)$, la verosimilitud de ω_j respecto a *x*. En general, ambos factores son importantes al momento

de tomar una decisión, y la decisión de Bayes los combina para lograr una mínima probabilidad de error.

2.7. Procesos Discretos de Markov.

Considerar un sistema que puede ser descrito en cualquier instante de tiempo como un estado perteneciente a un conjunto de N estados distintos $S_1, S_2,\ldots S_N$, tal como se muestra en la Figura 2.14, para 3 estados

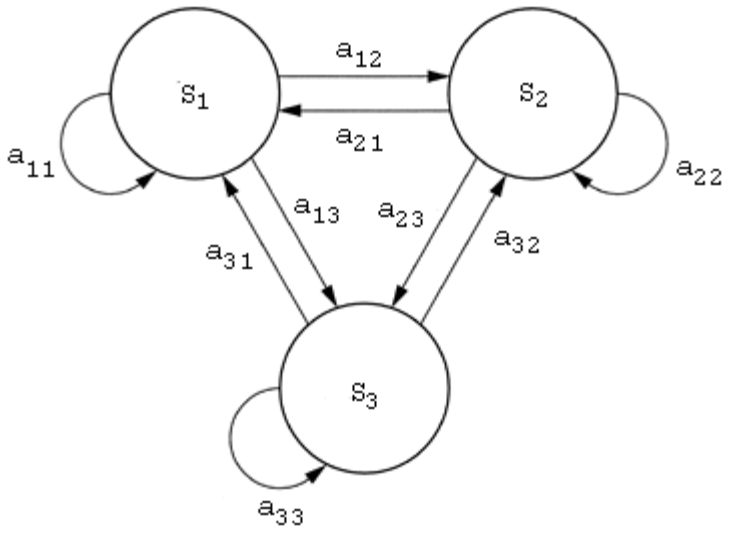

Figura 2.14. Modelo de Markov de 3 estados (N=3).

Si se mide a espacios regulares de tiempo, el sistema experimenta un cambio de estado (eventualmente, se mantiene el mismo estado). Se denotan los instantes de tiempo asociados con los cambios de estado como $t = 1,2,3,...$, y además, se denota el estado en el tiempo t como q_t. Una descripción probabilística completa del sistema presentado requeriría, en general, información del estado actual (en el instante t), así como también de todos los estados que lo preceden (Rabiner, 1989). Para el caso especial de un proceso discreto de Markov de primer orden, ésta descripción se trunca sólo para el estado actual y el anterior, es decir

$$P\left[q_t = S_j \mid q_{t-1} = S_i; q_{t-2} = S_k; \cdots\right] = P\left[q_t = S_j \mid q_{t-1} = S_i\right] \qquad (2.44)$$

Además, si se consideran sólo los procesos en los cuales el lado derecho de la ecuación es independiente del tiempo, se puede expresar el conjunto de probabilidades de transición de estado de la siguiente forma

$$a_{ij} = P\left[q_t = S_j \mid q_{t-1} = S_i\right] \qquad 1 \le i, j \le N \qquad (2.45)$$

Estas probabilidades de transición deben respetar las restricciones estocásticas estándar, es decir

$$\begin{aligned} a_{ij} &\ge 0 \\ \sum_{j=1}^{N} a_{ij} &= 1 \end{aligned} \qquad (2.46)$$

El proceso descrito puede llamarse modelo de Markov *observable*, pues la salida del proceso es el conjunto de estados para cada instante de tiempo, donde cada estado corresponde a un evento físico observable.

2.7.1 Modelos Ocultos de Markov (*Hidden Markov Models*).

Un modelo oculto de Markov (*HMM*) es una extensión de los procesos de Markov, el cual consiste en incluir el caso en el que la observación es una función probabilística del estado. Es decir, existe un proceso estocástico subyacente que no puede ser observado (o, en otras palabras, es *oculto*), pero sólo puede ser observado a través de otro conjunto de procesos estocásticos que producen una secuencia de observación.

Los elementos de un HMM son:

- *N*, el número de estados en el modelo: Aunque los estados son ocultos, generalmente los estados tienen algún significado físico, por lo que en muchas

aplicaciones prácticas los estados corresponden a resultados esperables. En el caso más general, los estados están conectados de tal manera que cualquiera de ellos puede ser alcanzado desde cualquier otro. Se denotarán los estados individuales como $S = \{S_1, S_2, \cdots, S_N\}$ y al estado en el tiempo t como q_t.

- *M*, El número de símbolos distintos de observación por cada estado: Los símbolos de observación corresponden al resultado físico del sistema que está siendo modelado. Se denotarán los símbolos como $V = \{v_1, v_2, \cdots, v_M\}$

- $A = \{a_{ij}\}$, la distribución de probabilidad de transición de estado, donde

$$a_{ij} = P[q_t = S_j \mid q_{t-1} = S_i] \qquad 1 \le i, j \le N$$

Para el caso especial donde todos los estados pueden ser alcanzados desde cualquier otro estado en una sola transición, se tiene que $a_{ij} > 0$ para todo i, j. Para otros tipos de HMM, se tendrá $a_{ij} = 0$ para algunos pares *(i , j)*.

- $B = \{b_j(k)\}$, la distribución de probabilidad del símbolo de observación en el estado j, donde

$$b_j(k) = P[v_k \ en \ t \mid q_t = S_j] \qquad \begin{matrix} 1 \le j \le N \\ 1 \le k \le M \end{matrix} \qquad (2.47)$$

- $\pi = \{\pi_j\}$, la distribución de estado inicial, donde

$$\pi_j = P[q_1 = S_j] \qquad (2.48)$$

Dados los valores de *N*, *M*, *A*, *B* y π, el HMM puede ser usado como un generador para entregar una secuencia de estados $O_1 O_2 O_3 \cdots O_T$ (donde cada observación *O* es uno de los símbolos de *V*, y *T* es el número de observaciones en la secuencia), o bien como un modelo para determinar cómo una secuencia de observación fue generada a partir de un HMM apropiado.

2.7.2. Algoritmo de Viterbi.

Para encontrar la mejor secuencia de estados $Q=\{q_1,q_2,\cdots,q_T\}$, para el conjunto de observación $O=\{O_1,O_2,\cdots,O_T\}$, se define la variable

$$\delta_{t+1}(j)=[max\{\delta_t(j)\}]b_j(O_t) \tag{2.49}$$

Es necesario llevar el rastro del argumento que maximiza la ecuación anterior para poder encontrar el camino de estados óptimos, para cada valor de *t* y *j*. Para esto se utiliza la variable $\psi_t(j)$. El procedimiento completo para encontrar la mejor secuencia de estados es la siguiente.

Inicialización:

$$\delta_1(i)=\pi_i\, b_i(O_1) \qquad 1\le i\le N \tag{2.50}$$

$$\psi_1(i)=0 \tag{2.51}$$

Recursión:

$$\delta_t(j)=\max_{1\le i\le N}\{\delta_{t-1}(i)a_{ij}\}b_j(O_t) \qquad \begin{matrix}2\le t\le T\\ 1\le j\le N\end{matrix} \tag{2.52}$$

$$\psi_t(i)=\operatorname*{argmax}_{1\le i\le N}\{\delta_{t-1}(i)a_{ij}\} \qquad \begin{matrix}2\le t\le T\\ 1\le j\le N\end{matrix} \tag{2.53}$$

Término:

$$p^*=\max_{1\le i\le N}\{\delta_T(i)\} \tag{2.54}$$

$$q_T^*=\operatorname*{argmax}_{1\le i\le N}\{\delta_T(i)\} \tag{2.55}$$

Secuencia de estados óptima:

$$q_t^*=\psi_{t+1}\left(q_{t+1}^*\right) \tag{2.56}$$

Capitulo 3.

Estimación de la Curva de Entonación

3.1 Introducción

En el presente capítulo se explica el sistema de evaluación de entonación existente, indicando el aporte de este trabajo. Se explicará en detalle los métodos utilizados para determinar la curva de entonación, incluyendo pre-procesamiento, detección de pitch, clasificación sonoro-sordo, y post-procesamiento, y la base de datos a utilizar para este propósito.

3.2 Sistema de Evaluación de Entonación.

El sistema de evaluación de entonación existente (Arias, 2008) consiste en comparar las curvas del alumno con las de una referencia previamente grabada. La interacción del alumno con el computador consta de 3 etapas: (a) el computador reproduce la señal de referencia, que debe ser emulada por el estudiante; (b) el estudiante trata de repetir el patrón de entonación presente en la señal reproducida por el computador, grabando su propia voz para ser evaluada, llamada señal de test; y (c), el computador muestra las curvas de

entonación de la referencia pre-grabada, la de test, y asigna una nota a la última. El sistema queda descrito el diagrama de bloques que se muestra en la Figura 3.1

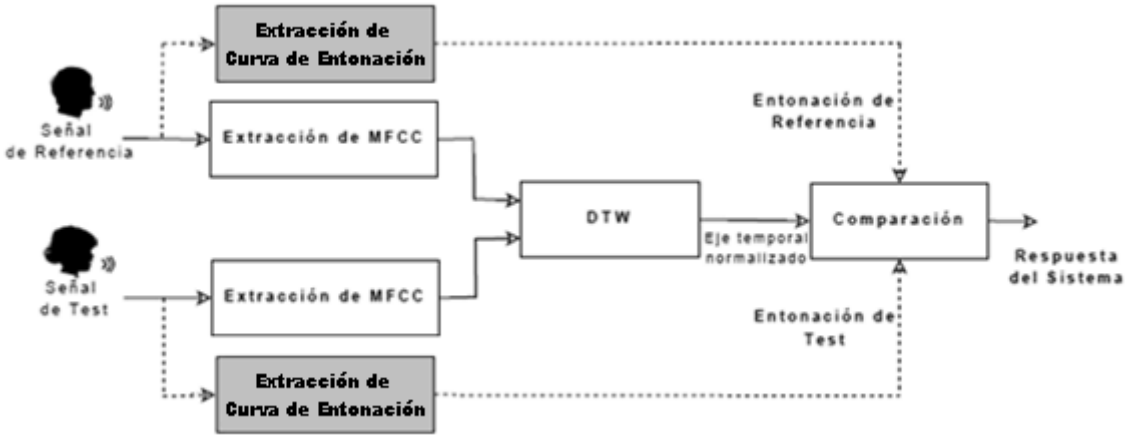

Figura 3.1: Diagrama de Bloques de sistema de evaluación de Entonación

En primer lugar, se realiza un pre-procesamiento, de manera de eliminar las componentes de frecuencia menores a 75[Hz] mediante un filtro pasa-altos. También se detecta el inicio y el fin de las elocuciones, para eliminar los intervalos de silencio al inicio y al final de la señal, los que no aportan información prosódica ni acústica. Luego se calcula la curva entonación mediante algún algoritmo de extracción de pitch. Además, como las señales no tienen el mismo largo, pues se trata de dos hablantes distintos, se deben alinear temporalmente para ser comparadas. Para este alineamiento se utiliza el algoritmo DTW (*Dinamic Time Warping*, Alineamiento Temporal Dinámico). Finalmente, con las señales en condiciones de ser contrastadas, se utiliza una medida de similitud, entregando una medida de semejanza entre ambas, a la cual se le asocia un puntaje, o bien una respuesta categórica de aceptación o rechazo.

El aporte de esta memoria de título consiste en desarrollar un algoritmo de detección de pitch, es decir, se trabajara en los bloques ennegrecidos en la Figura 3.1. El algoritmo que ocupa el sistema es un evaluador basado en ACF, cuyo clasificador sonoro-sordo consiste en un umbral para el peak de la autocorrelación. La idea es mejorar tanto el extractor de pitch, la decisión sonoro-sordo y la corrección de errores mediante post-procesamiento. A continuación se describe el método desarrollado.

3.3 Estimación de la Curva de Entonación

Un sistema de estimación de la curva de entonación debe calcular la frecuencia fundamental de una señal de voz para cada segmento, obteniendo el pitch como función del tiempo. Además, debe discriminar los segmentos sordos y sonoros. A continuación se explica en detalle el sistema propuesto en este trabajo.

3.3.1 Pre-Procesamiento.

El pre-procesamiento de la señal consiste en utilizar técnicas para tener una señal lo más limpia posible, de manera de minimizar efectos indeseados al momento de calcular el pitch. El uso de filtros pasa-bajos es muy frecuente, pues con estos se eliminan efectos no deseados en la señal de voz. Como el ancho de banda de la voz es acotado, las componentes de frecuencia fuera de este rango se producen por ruido ambiente; efectos de canal en el micrófono; etc. El uso de filtros pasa-bajos elimina estos efectos, dejando una señal mucho más limpia para ser procesada.

Además, en este caso, es necesario sub-muestrear las señales de voz. Este proceso consiste en utilizar una frecuencia de muestreo menor para discretizar la misma señal. Esto se realiza con el objeto de disminuir el tamaño en muestras de la señal de voz, pero también tiene un efecto importante en el dominio de la frecuencia. Al disminuir la frecuencia de muestreo, disminuye el ancho de banda de la señal de voz, tal como lo establece el teorema de Nyquist-Shannon. Por lo tanto, para cualquier análisis en frecuencia será relevante la aplicación de sub-muestreo a la señal de voz.

Antes de sub-muestrear, es necesario filtrar la señal de manera de eliminar las componentes de armónicas que se encuentren fuera del rango de la nueva frecuencia de muestreo. Para esto, se utilizan filtros pasa-altos, sintonizados a la mitad de la nueva frecuencia de muestreo.

Una vez submuestreada la señal, es necesario convertir la señal en segmentos. Este proceso se conoce como enventanado de la señal. Este procedimiento consiste en dividir la señal de voz en segmentos más pequeños, para evaluar la periodicidad en ese instante de tiempo. El enventanado se realiza para una ventana de 25.6[ms], y un paso de 10[ms], tal como se muestra en la figura3.2

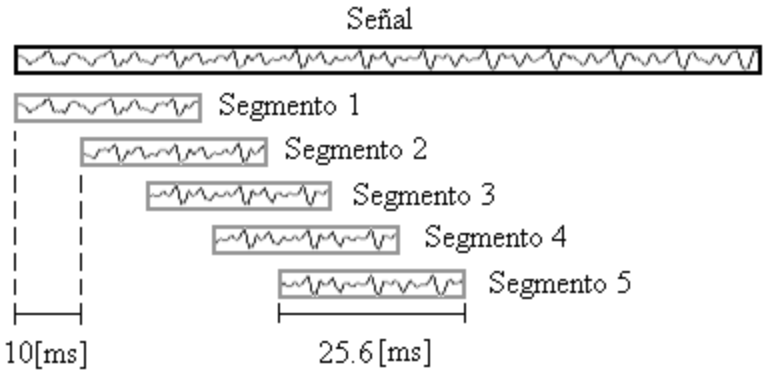

Figura 3.2. Enventanado de la señal de Voz, utilizando ventana rectangular.

La ventana utilizada es la ventana rectangular, definida como

$$Rect(n) = \begin{cases} 1 & 0 \le n \le N \\ 0 & en\ otro\ caso \end{cases} \tag{3.1}$$

Donde N corresponde al largo de la ventana en muestras. Si además, al paso de avance del proceso del enventanado se le denomina s, entonces los segmentos de voz de la señal $x(n)$ se definen como

$$segmento\, i = x(n) \cdot Rect(n - i \cdot s) \tag{3.2}$$

Las características del enventanado (largo de ventana y paso) son las que se utilizan en la base de datos Keele, la que seexplicará mas adelante.

3.3.2. SSPD

El sistema de evaluación SSPD (*Synthesized Speech Pitch Detector,* detector de pitch basado en sintetización de voz) consiste en comparar el segmento de voz con la señal sintetizada por un codificador LPC. Como se explicó en el capítulo anterior, un codificador LPC recibe como entrada un tren de pulsos cuya separación es el período fundamental de la señal a codificar. Dicho tren es convolucionado con el filtro, obteniéndose, finalmente, la señal sintetizada. Entonces, la idea es la siguiente: probar con un número alto de candidatos a frecuencia fundamental para el tren de impulsos de entrada, y elegir como frecuencia fundamental la que produzca la señal sintetizada más cercana a la señal original. Las etapas del algoritmo SSPD se enumeran a continuación.

1.- Calcular la transformada de Fourier de cada segmento de voz. Con esto se obtiene el espectro en frecuencia del segmento.

2.- Determinar los coeficientes LPC del segmento de voz.

3.- Determinar el espectro del filtro LPC a través de los coeficientes LPC. Esto se realiza calculando la transformada de Fourier de la respuesta al impulso del filtro.

4.- Crear trenes de pulsos para todas las frecuencias que se quieran evaluar.

5.- Calcular la transformada de Fourier de cada tren de pulso.

6.- Multiplicar los espectros del filtro LPC y el de los trenes de pulsos para obtener la sintetización de los segmentos de voz.

7.- Comparar el espectro del segmento de voz con la versión sintetizada, para todas las frecuencias de interés.

8.- Elegir la frecuencia que minimice (7).

Las etapas del método SSPD se muestran en el diagrama de bloques de la Figura 3.3.

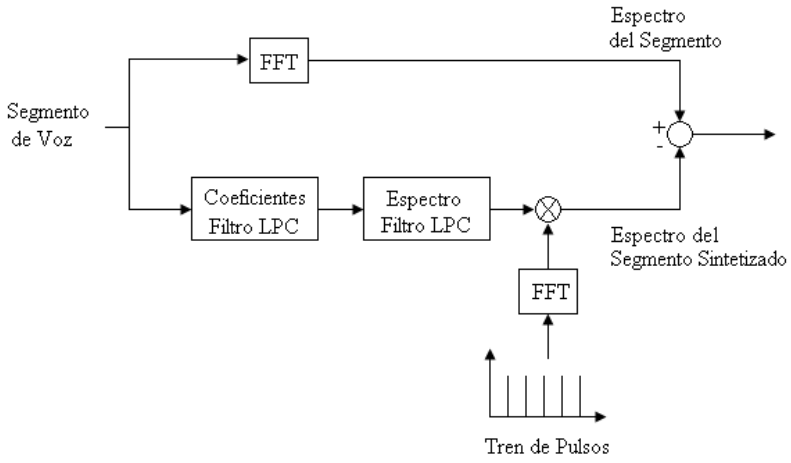

Figura 3.3. Diagrama de bloques de SSPD

En la Figura 3.4, se muestra el progreso de la sintetización de un segmento de voz sonoro. Se muestra el segmento de voz en el tiempo, en frecuencia y la señal sintetizada (a la frecuencia adecuada), además de la envolvente (filtro LPC),

La comparación entre ambas señales se mide utilizando el error cuadrático entre ellas. Para cada valor de frecuencia fundamental, se tendrá una señal sintetizada distinta. Por ende, se denominará a dicha señal sintetizada como $X_{syn}(f, f_0)$, mientras que a la señal original se designará como $X(f)$. El largo de cada señal es $N = \frac{N_{FFT}}{2} - 1$ (número de puntos no redundantes de la FFT). De esta manera, el valor de la función objetivo para alguna frecuencia fundamental f_0 será

$$FObj(f_0) = \frac{1}{N} \sum_f \left(X(f) - X_{syn}(f, f_0) \right)^2 \tag{3.3}$$

En la Figura 3.5 se muestra la función objetivo para el segmento de señal mostrado anteriormente, para cada valor de f_0

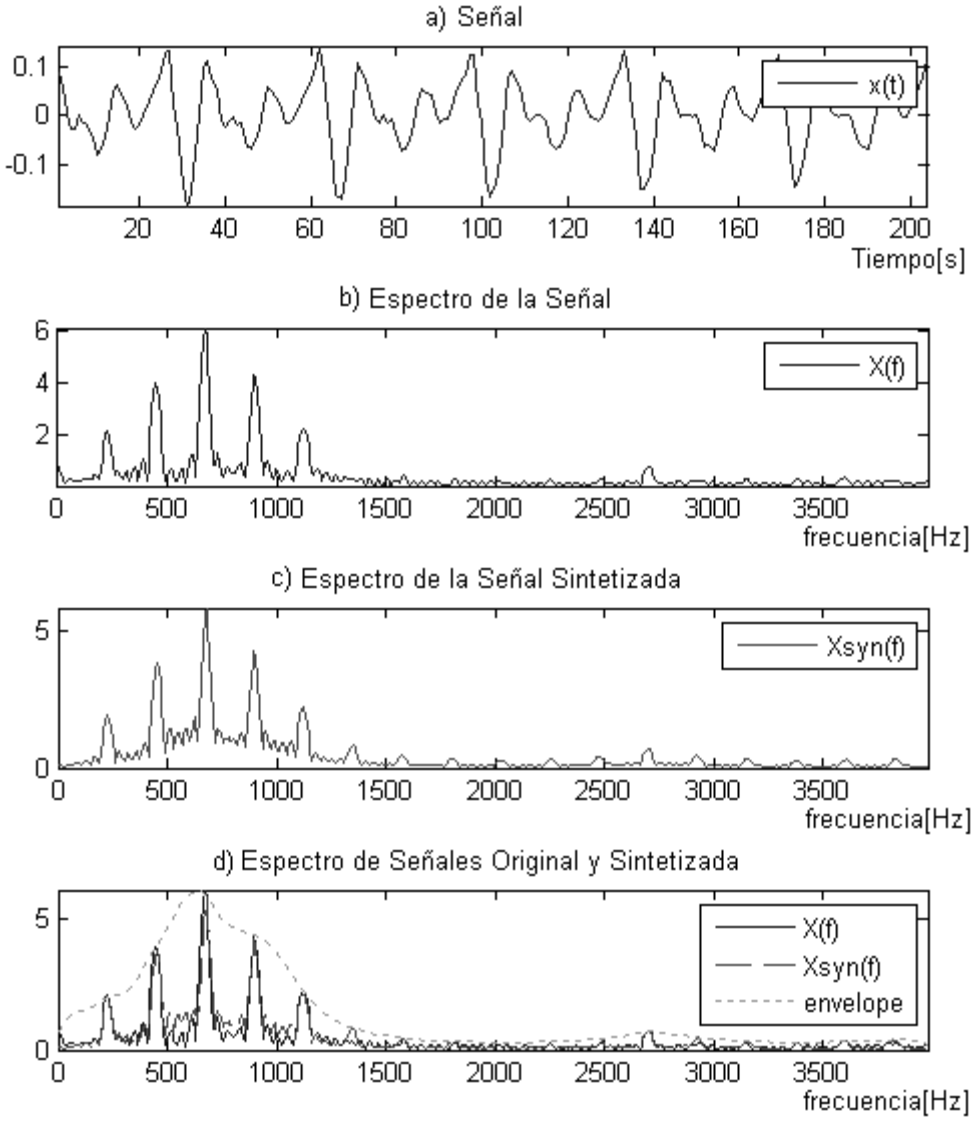

Figura 3.4. a) el segmento de señal de voz de 25,6[ms], muestreada a una frecuencia de 8[KHz]. b) la misma señal en el dominio de la frecuencia. c) señal sintetizada en el dominio de la frecuencia, a una frecuencia fundamental de 225[Hz]. d) señales original y sintetizada en el dominio de la frecuencia, calculada con 8 coeficientes LPC y con 1024 puntos para la FFT.

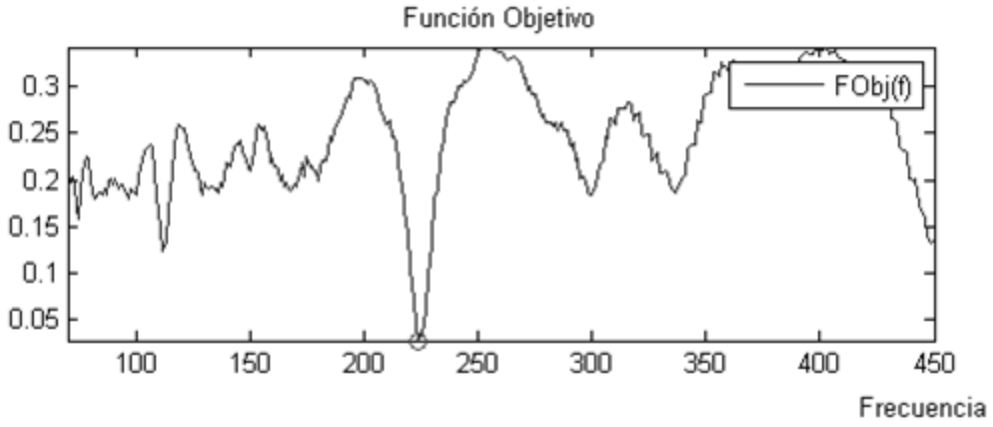

Figura 3.5: Función objetivo. El círculo representa el mínimo de la función.

De esta manera, el pitch debiese ser el argumento que minimice la función objetivo. Se observa que el mínimo global se encuentra a una frecuencia $\hat{f}_0$ de alrededor de 220[Hz].

$$\hat{f}_0 = \underset{f_0}{argmin}\{FObj(f_0)\} \tag{3.4}$$

Es importante mencionar que, en la mayoría de los casos, existen mínimos locales en el sub-armónico (la mitad de f_0) y en el armónico (el doble de f_0) de la señal. En la misma Figura 3.5 se puede observar este comportamiento para $f_0 \approx 110[Hz]$ y $f_0 \approx 440[Hz]$. En algunos casos, puede ocurrir que el mínimo global de la función objetivo se encuentre no en la frecuencia fundamental, sino que en alguno de sus armónicos. Esta conducta se conoce como *Salto de Octava* o *Error de Octava*. Si se detecta el sub-armónico, se dice que se produjo *Halving*, mientras que si lo que se detectó fue el armónico, se produjo *Doubling*. Estos errores son inherentes a cualquier detector de pitch (*Hess, 1992*). Este problema se intenta resolver con un post-procesamiento, el que será explicado más adelante.

Ahora, para determinar la frecuencia fundamental en cada pitch, utilizando el método SSPD, es necesario encontrar la mejor configuración para las siguientes variables

- *Orden del codificador LPC*: El orden del codificador, *P*, es el número de constantes que definen al filtro LPC. Este número es muy importante, pues influye directamente en el espectro de éste y, por tanto, en el espectro de la señal sintetizada. El filtro LPC define la envolvente de la señal. Si P es muy bajo, la envolvente no estará bien representada, mientras que si es muy grande, se ajustará mucho al espectro, produciendo errores de octava (principalmente *Halving*). En la Figura 3.6 se muestra cómo se modifica el filtro LPC si se varía el número de coeficientes *P*.

- *Frecuencia de muestreo*: la frecuencia de muestreo f_s también es un valor crítico. Si ésta es muy baja, puede que se pierdan componentes de frecuencia importantes de la voz. Sin embargo, si es muy alta, en particular mucho mayor que el ancho de banda en el que se encuentra la voz, el espectro estará muy concentrado en los primeros puntos, perdiéndose resolución en la transformada de Fourier para la banda de señal útil.

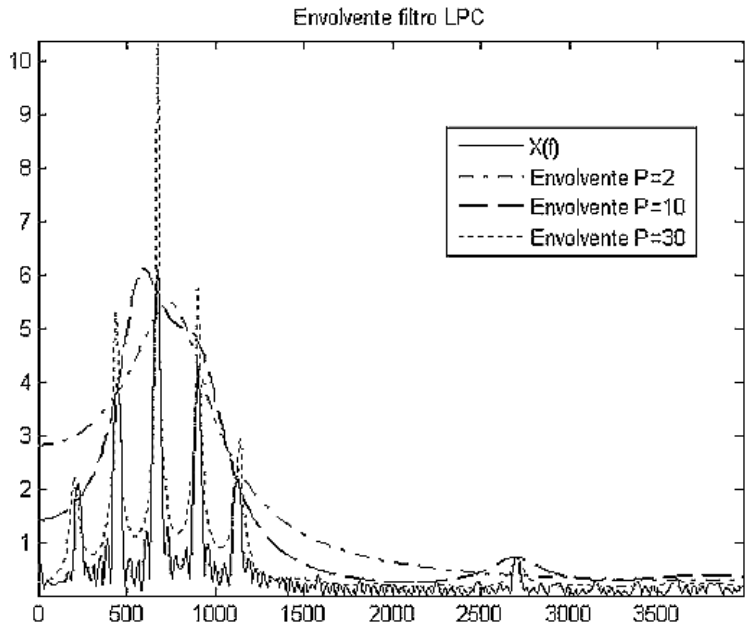

Figura 3.6: efecto de orden del filtro P en la envolvente LPC.

- *Número de puntos FFT*. El número de puntos N_{FFT} es importante, pues mientras mayor sea éste, mejor resolución tendrá el espectro de la señal. Sin embargo, también será mayor la carga computacional que requerirá el cálculo de la FFT.

Hay que mencionar que el número de coeficientes LPC está directamente relacionado con la frecuencia de muestreo. Mientras menor es f_s, se necesitan menos coeficientes para representar la envolvente de la señal, pues el espectro es más acotado.

3.3.3. Decisión Sonoro-Sordo

Como se ha explicado anteriormente, la detección de pitch se compone principalmente del algoritmo de estimación de pitch, y de la detección sonoro-sordo. Ambos elementos son fundamentales para cualquier detector.

En este trabajo se propone un detector sonoro-sordo basado en un modelo oculto de Markov de 2 estados (uno para sonoro, y otro para sordo), conocido como modelo de Gilbert (Becerra-Yoma et al, 2004). Este modelo ha sido empleado en muchas publicaciones para analizar la respuesta del modelo TCP, para evaluar algoritmos de reconocimiento de locutor, para modelar la pérdida de paquetes en redes IP, etc. El modelo se ilustra en la Figura 3.7.

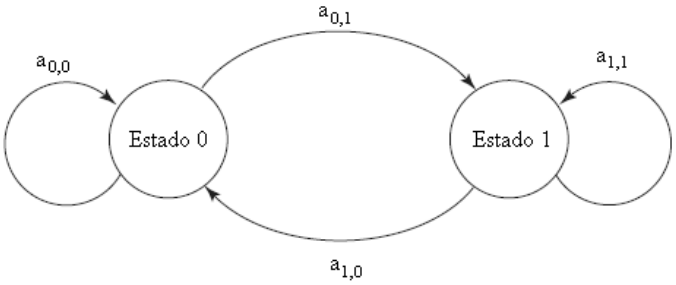

Figura 3.7: Modelo de Gilbert.

Este proceso es modelado por la **tasa de segmentos sonoros** *VR* (*Voiced Rate*), que corresponde a la razón entre número de segmentos sonoros y el número total de ellos

(sonoros y sordos), definido en (3.6); y la distribución de probabilidad de la **longitud de los segmentos sonoros** *VL* (*Voiced Length*), la cual se define como el número de segmentos sonoros consecutivos.

$$VR = \frac{\#\,segmentos\ sonoros}{\#\,segmentos\ sonoros + \#\,segmentos\ sordos} \tag{3.6}$$

El modelo básico de Gilbert usa el *Estado 0* para representar que un segmento es sonoro, y el *Estado 1* para indicar que es sordo. La probabilidad de que *VL* segmentos consecutivos están en *Estado 0* se distribuye geométricamente, y equivale a

$$Pr(largo\ tramo\ sonoro = VL\ en\ Estado\ 0) = (a_{0,0})^{VL-1} a_{0,1} \tag{3.7}$$

Donde $a_{0,0}$, $a_{0,1}$, $a_{1,0}$, $a_{1,1}$ son las probabilidades de transición del modelo mostrado en la Figura 3.7, definidas como

$$a_{0,1} = \frac{1}{E_0[VL]} \tag{3.8}$$

(3.9)

$$a_{0,0} = 1 - a_{0,1} \tag{3.10}$$

$$a_{1,0} = \frac{VR}{E_0[VL](1-VR)} \tag{3.11}$$

$$a_{1,1} = 1 - a_{1,0} \tag{3.12}$$

$E_0[VL]$ corresponde al valor esperado de *VL* en el *Estado 0*. El modelo de Gilbert ordinario, como se puede apreciar, queda totalmente determinado por dos variables, $a_{0,1}$ y $a_{1,0}$, los que se pueden estimar con $E_0[VL]$ y VR.

Es posible, además, incorporar restricciones temporales en el modelo de Gilbert. Este problema ya ha sido explorado en aplicaciones en reconocimiento de locutor. La duración de estado puede incluirse en el algoritmo de Viterbi en términos de las

probabilidades de transición $a_{i,j}^{VL} = Pr\left(s_{t+1} = j \mid s_t = s_{t-1} = \cdots = s_{t-BL+1} = i\right)$. Se puede observar que VL es el número de segmentos consecutivos en el Estado *i*, hasta un tiempo *t*; $j = i$, o $j = i+1$ (si $i = 0$), o $j = i-1$ (si $i = 0$). Luego, las probabilidades condicionales $a_{i,i}^{VL}$ y $a_{i,i\pm1}^{VL}$ se pueden estimar como

$$a_{i,i}^{VL} = \frac{Pr\left(s_{t+1} = i \mid s_t = s_{t-1} = \cdots = s_{t-BL+1} = i\right)}{Pr\left(s_t = s_{t-1} = \cdots = s_{t-BL+1} = i\right)} = \frac{D_i(VL) - d_i(VL)}{D_i(VL)} \tag{3.13}$$

$$a_{i,i\pm1}^{VL} = \frac{Pr\left(s_{t+1} = i \pm 1 \mid s_t = s_{t-1} = \cdots = s_{t-BL+1} = i\right)}{Pr\left(s_t = s_{t-1} = \cdots = s_{t-BL+1} = i\right)} = \frac{d_i(VL)}{D_i(VL)} \tag{3.14}$$

donde $d_i(VL)$ es la probabilidad de duración de estado igual a VL segmentos, y $D_i(VL)$ es la probabilidad de que el estado *i* este activo para $t \geq VL$.

$$D_i(VL) = \sum_{t=VL}^{\infty} d_i(t) \tag{3.15}$$

En caso de incluir posibles duraciones máximas $max_i(VL)$ y mínimas $min_i(VL)$, las probabilidades de transición se modifican a

$$a_{i,i}^{VL} = \begin{cases} 1 & si \quad VL < min_i(VL) \\ 0 & si \quad VL > max_i(VL) \\ \dfrac{D_i(VL) - d_i(VL)}{D_i(VL)} & Otro\ caso \end{cases} \tag{3.16}$$

$$a_{i,i\pm1}^{VL} = \begin{cases} 0 & si \quad VL < min_i(VL) \\ 1 & si \quad VL > max_i(VL) \\ \dfrac{d_i(VL)}{D_i(VL)} & Otro\ caso \end{cases} \tag{3.17}$$

donde el principal problema está en modelar la distribución de probabilidad $d_i(VL)$. Esto depende de cada estado (sordo o sonoro), pues no necesariamente tienen la misma distribución, como se verá más adelante.

Luego de determinar todas las probabilidades de transición, se encuentra la secuencia de estados más probable a través del algoritmo de Viterbi.

3.3.4. Post-procesamiento.

El post-procesamiento para la curva de pitch es el paso final de todo detector. Consiste en tomar la decisión final sobre los candidatos a la frecuencia fundamental. Por ejemplo, se puede determinar el pitch como el valor más probable, lo que no siempre lleva a resultados correctos. En el caso de esta memoria, el valor más probable es el mínimo de la función objetivo, lo que como ya se explicó, conlleva a errores de detección, principalmente *Halving* y *Doubling*. Para minimizar estos errores, se utiliza un bloque de programación dinámica para corregir posibles valores incorrectos y entregar una curva de pitch más precisa.

La programación dinámica es un método de resolución algorítmico que consiste en dividir un problema en sub-problemas, los que a su vez se dividen en sub-sub-problemas. Finalmente, resolviendo estos sub-sub-problemas, se puede llegar a la solución global del problema principal. Ésta es usada en varias áreas de la ingeniería y ciencias, pues reduce considerablemente el número de cálculos de un algoritmo. Por esta razón, se utiliza en muchas aplicaciones, como por ejemplo, encontrar el camino más corto, encontrar el camino de mínimo costo, alineamiento de señales, etc.

En este trabajo se utiliza un módulo de programación dinámica de manera de reducir el número de errores producidos por el método SSPD. Para ello, se utilizará la función objetivo obtenida luego de la etapa SSPD, pues al ver dicha función a lo largo de la longitud de la señal, se observa que el pitch debiese seguir alguno de los "valles" que se forman, como se vislumbra en la Figura 3.8.

Los segmentos en los que toda la función objetivo es de color blanco corresponden casi con seguridad a segmentos sordos. En este caso, se ve que existen muchos valles por los que se podría mover la curva de pitch. Sin embargo, existe sólo un camino de mínimo costo, que es equivalente a un camino por el valle de mayor profundidad. En el caso ilustrado, ese camino sería el que se observa en la Figura 3.9.

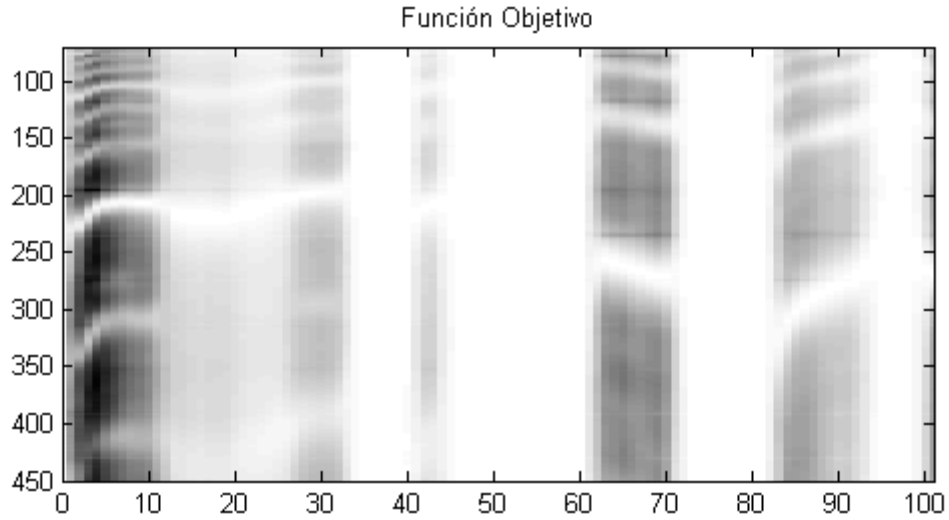

Figura 3.8: Función objetivo de parte de una señal de voz. El eje vertical corresponde a la frecuencia, mientras que el horizontal, al número de segmento. La altura se observa en el color. Mientras más oscuro, mayor altura tiene la función objetivo. Mientras más claro, menor altura.

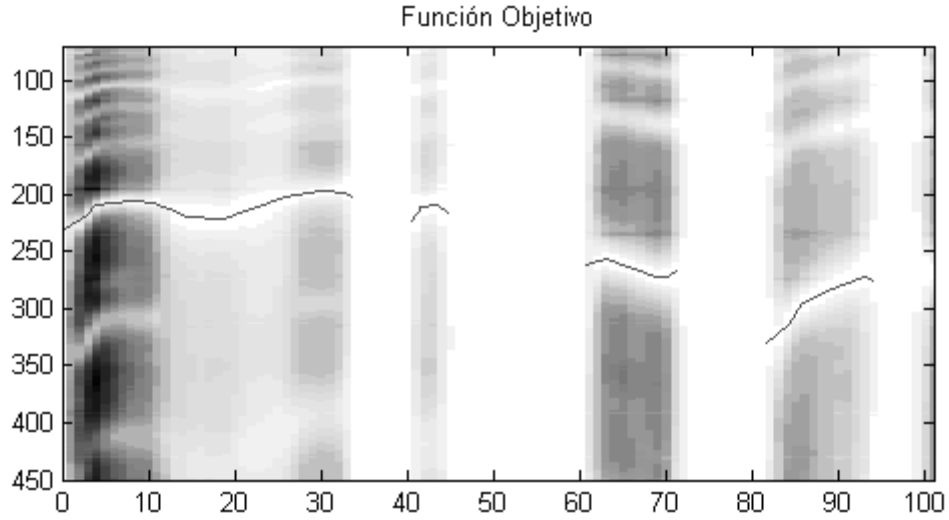

Figura 3.9: Curva de entonación en función objetivo.

El método utilizado consiste en utilizar el siguiente algoritmo: Primero, se estiman los tramos sonoros. Para cada uno de los tramos, se aplica el algoritmo. Luego para cada segmento, para cada frecuencia de la función objetivo (los que se llamarán de ahora en adelante estados), se verá cuál es el candidato más probable entre un conjunto de posibles estados. Finalmente, se acumulará el valor en dicho segmento con todos los anteriores y se sigue con el siguiente, hasta llegar al último.

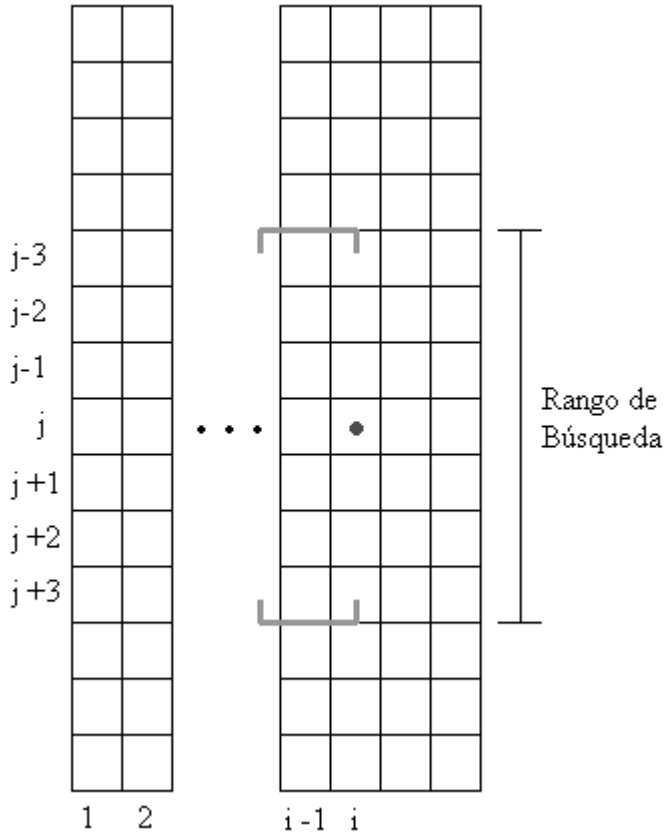

Figura 3.10. Matriz para post-procesamiento con programación dinámica.

En este momento, se quiere saber cuál es el valor de frecuencia que antecedería al segmento i, si es que en ese instante el pitch tuviera el valor j (punto en la Figura 3.10). Para esto, se restringe el conjunto de valores posibles, pues se sabe que el pitch no debería experimentar cambios demasiado abruptos. En la Figura 3.10, a modo de ejemplo, se observa que los candidatos son *{j-3,j-2,...j+2, j+3}*. De esta manera, como se desea encontrar el camino de menor costo, se debe determinar cuál de las posibles transiciones

minimiza dicho costo. Hay que considerar que los valores para cada frecuencia en el segmento *i-1* contienen los valores acumulados de todos los segmentos que lo preceden. Por lo tanto, al llegar al último segmento se tendrá el valor que minimiza el camino para cada estado. Luego, el camino de mínimo costo corresponderá al mínimo valor para ese segmento. Paralelamente, se irá guardando el camino en una variable auxiliar, de manera de poder recorrer hacia atrás el recorrido óptimo.

La formulación matemática para este algoritmo es la siguiente. Sea D la matriz de costo acumulada. $D \in \Re^m \times \Re^n$. Sea Φ la matriz de camino. $\Phi \in \Re^m \times \Re^n$, y sea q el vector de secuencia óptima q, $q \in \Re^n$.

Inicialización:

$$D(j,1) = F_{OBJ}(j,1) \quad 1 \le j \le m \tag{3.18}$$

$$\Phi(j,1) = 0 \quad 1 \le j \le m \tag{3.19}$$

Iteración:

$$D(j,i) = F_{OBJ}(j,i) + min\begin{Bmatrix} \vdots \\ D(j-2,i-1) \\ D(j-1,i-1) \\ D(j,i-1) \\ D(j+1,i-1) \\ D(j+2,i-1) \\ \vdots \end{Bmatrix} \quad \begin{matrix} 1 \le j \le m \\ 2 \le i \le n \end{matrix} \tag{3.20}$$

$$\Phi(j,i) = argmin\begin{Bmatrix} \vdots \\ D(j-2,i-1) \\ D(j-1,i-1) \\ D(j,i-1) \\ D(j+1,i-1) \\ D(j+2,i-1) \\ \vdots \end{Bmatrix} \quad \begin{matrix} 1 \le j \le m \\ 2 \le i \le n \end{matrix} \tag{3.21}$$

Finalización:

$$d^* = min\{D(j,n)\} \quad 1 \le j \le m \tag{3.22}$$

$$q^*(n) = argmin\{D(j,n)\} \quad 1 \le j \le m \tag{3.23}$$

Secuencia óptima:

$$q^*(i) = \Phi\left(i+1, q^*(i+1)\right) \quad n-1 \ge i \ge 1 \tag{3.24}$$

También es posible asignar pesos para determinar las transiciones de pitch de un segmento a otro. Esto se realiza pues en el modelo anterior todas ellas se ponderan de la misma manera, y en la práctica son más probables los cambios pequeños a los grandes. Los pesos sirven para modelar este efecto. En este caso, los cambios pequeños, como se busca minimizar el camino, deben tener una ponderación más baja que los cambios grandes. Así, la iteración se reformula como

$$D(j,i) = F_{OBJ}(j,i) + min\left\{\begin{array}{c} \vdots \\ w(-2)\cdot D(j-2,i-1) \\ w(-1)\cdot D(j-1,i-1) \\ w(0)\cdot D(j,i-1) \\ w(1)\cdot D(j+1,i-1) \\ w(2)\cdot D(j+2,i-1) \\ \vdots \end{array}\right\} \begin{array}{l} 1 \le j \le m \\ 2 \le i \le n \end{array} \tag{3.25}$$

$$\Phi(j,i) = argmin\left\{\begin{array}{c} \vdots \\ w(-2)\cdot D(j-2,i-1) \\ w(-1)\cdot D(j-1,i-1) \\ w(0)\cdot D(j,i-1) \\ w(1)\cdot D(j+1,i-1) \\ w(2)\cdot D(j+2,i-1) \\ \vdots \end{array}\right\} \begin{array}{l} 1 \le j \le m \\ 2 \le i \le n \end{array} \tag{3.26}$$

donde *w(k)* corresponde al vector de pesos antes mencionado, *k* corresponde al cambio de estado respecto al estado anterior. El resto del algoritmo es el mismo para el caso sin pesos.

Finalmente, la determinación del rango de búsqueda dependerá de las observaciones experimentales de la base de datos a utilizar, lo que se determinará en el capítulo 4.

3.4. Base de Datos Keele.

El objetivo del proyecto Keele (Meyer et al, 1995) es el desarrollo de una base de datos gratuita y abierta para comparar algoritmos de estimación de pitch. Antes de crearse esta base de datos, la mayoría de lo estimadores de pitch se evaluaban con bases de datos propias, es decir, que sólo los investigadores que creaban el método tenían acceso a ellas. Luego, sólo los detalles del algoritmo eran publicados, y no los datos que utilizaron para evaluar. Por esto, cualquier persona es capaz de replicar el algoritmo, pero no puede probar su rendimiento.

La base de datos Keele consiste de un texto fonéticamente balanceado, llamado "*The North Wind Story*", el cual es leído por 15 locutores cuyo idioma nativo es inglés: 5 mujeres adultas, 5 hombres adultos, y 5 niños (3 hombres y 2 mujeres). Los adultos fueron grabados en una sala a prueba de ruidos, mientras que los niños fueron grabados en un ambiente tranquilo, para minimizar el stress.

Grupo	Edad (años)	Duración (segundos)
Hombres	21-60	27-40
Mujeres	20-37	30-38
Niños	8-12	30-50

Tabla 3.1. Resumen de la base de datos Keele.

Las señales fueron grabadas simultáneamente con un laringógrafo y con un grabador DAT (*Digital Audio Trace Recorder*). Ambas fueron digitalizadas a 20[KHz], con 16 bits de resolución.

Para las señales de voz, la laringografía es la única referencia absoluta que existe. Ésta mide físicamente la vibración de las cuerdas vocales (si es que existe) en la producción del habla del locutor. En la Figura 2.15 se muestra una señal y la laringografía correspondiente

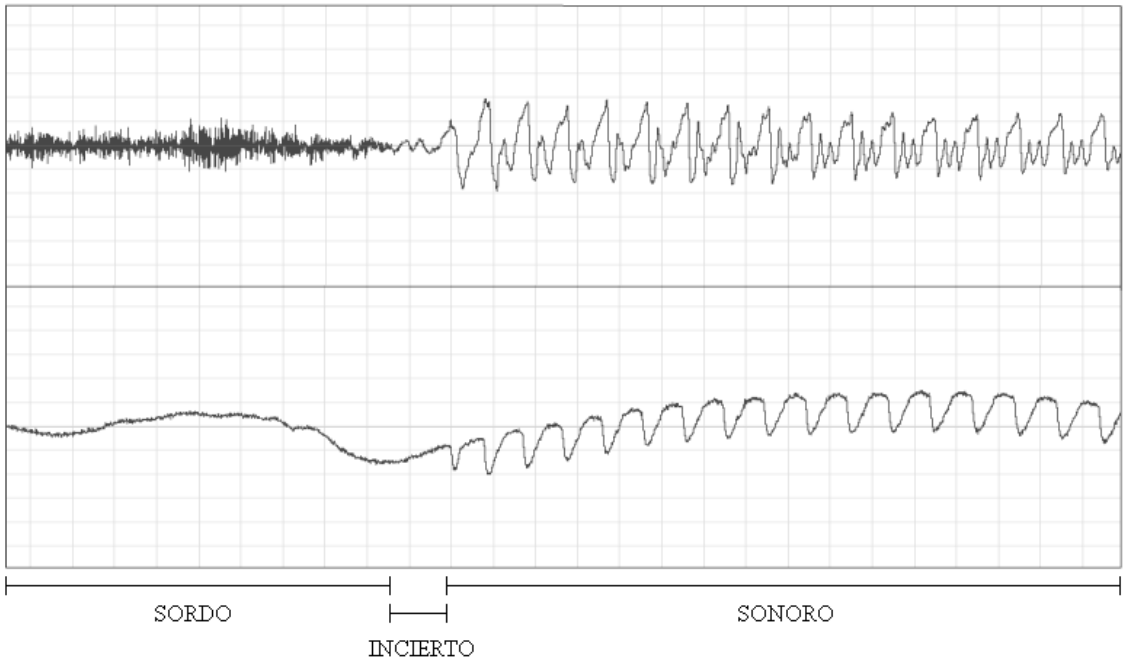

Figura 3.11. Señal de voz (arriba) y laringografía (abajo) para un mismo segmento de voz.

El tramo incierto que se muestra en la Figura 2.15 ocurre cuando existe una transición de segmentos sonoros a sordos, transición de segmentos sordos a sonoros, o también cuando existen segmentos de voz representando fonemas plosivos (como /p/, /k/, /t/, etc). Dichas incertezas representan entre un 2,5% y un 5% del número de segmentos, dependiendo del locutor.

Es importante mencionar que la base de datos también incluye una referencia pre-calculada. Ésta incluye la decisión sonoro-sordo, y una estimación de pitch, calculada desde la laringografía usando el método de autocorrelación para una ventana de 25,6[ms], desplazándose cada 10[ms].

En conclusión, la base de datos consiste en el set de señales de evaluación, además de las señales de la laringografía y la referencia pre-calculada, donde se recomienda, en una primera aproximación, no considerar los segmentos etiquetados como “incierto” (es decir, considerarlos como sordos), pues así se obtendrán mejores resultados (eliminando los segmentos problemáticos), y además, porque tampoco es claro cuál será el pitch percibido para dichos segmentos.

Capitulo 4.

Evaluación del Método de Estimación de la Curva de Entonación.

4.1 Introducción

En este capítulo se explicarán los experimentos para la evaluación de cada uno de los pasos del método de estimación de la curva de entonación. Dichos experimentos consisten en la evaluación de cada una de las etapas, SSPD, clasificación sonoro-sordo, y post-procesamiento. Para SSPD, se probabrán distintas configuraciones de frecuencias de muestreo y coeficientes LPC, de las cuales se elegirá la mejor de ellas. Para esto, se utilizarán señales de referencia para entrenar un clasificador bayesiano simple como clasificador sonoro-sordo, y se testeará con las señales restantes. Para la clasificación sonoro-sordo se ocupara la configuración que minimice SSPD. Se entrenarán señales de referencia para determinar el histograma de los largos de los tramos sonoros, con lo que se aplicará el modelo de Gilbert con y sin duración de estado para distintas características de la señal para evaluar el error de clasificación. El post-procesamiento utilizará las configuraciones de SSPD y del clasificador sonoro-sordo. Como entrenamiento se determinará el vector de pesos mediante señales de referencia, para finalmente testear la

tasa de error. Finalmente, se mostrará el resultado global de todas las etapas juntas, y se discutirán los resultados obtenidos

4.2. Indicadores para Evaluación de Estimadores de Pitch.

Para el área de estimación de la frecuencia fundamental se utilizan los siguientes errores para evaluar los métodos utilizados para ese fin

- GPE (*Gross Pitch Error*): ocurre GPE cuando el pitch detectado difiere mucho respecto a la referencia. Si el pitch estimado tiene un error de un 20% respecto a la referencia, entonces se comete un error grave. GPE se define como el porcentaje de errores graves en la señal de voz.

- VE (*Voiced Error*): Si se detecta un segmento como sordo, pero según la referencia es sonoro, se comete un error de detección sonoro. VE corresponde al porcentaje de errores de detección sonoro en la señal de voz.

- UE (*Unvoiced Error*): Si se detecta un segmento como sonoro, pero según la referencia es sordo, se comete un error de detección sordo. UE corresponde al porcentaje de errores de detección sordo en la señal de voz.

- MN (*mean*): en los segmentos en los que no se comete GPE, se mide el promedio de la diferencia entre el pitch estimado y el pitch de referencia. A este indicador se le llama MN.

- STD (*Standard Deviation*): mide la desviación estándar entre el pitch y la referencia, cuando no se comete GPE.

También es muy utilizado el error VUVE (Voiced-Unvoiced Error). Este error es la suma entre VE y UE, y se utiliza cuando se quiere determinar el desempeño de la detección

sonoro y sordo. Cuando se mide este último, son igual de importantes ambos errores (VE y UE), por lo que se busca mejorar la suma de ellos.

4.3 Experimentos.

En este capítulo se mostrarán los experimentos realizados para evaluar todas las etapas del método propuesto: SSPD; clasificador sonoro-sordo; y, post-procesamiento, además de todas las etapas en su conjunto.

4.3.1. Experimentos SSPD.

El objetivo de los experimentos para SSPD es obtener la configuración óptima de parámetros, de manera de minimizar el GPE. Para esto, se entrenan algunas señales de referencia para crear un clasificador sonoro-sordo simple basado en la teoría de Bayes. La mejor configuración se utilizará en las etapas a seguir.

4.3.1.1. Base de datos de evaluación para SSPD.

Se utilizará la base de datos Keele. Se considerarán las señales pre-grabadas de hombres y mujeres adultas, teniendo un total de 10 señales de aproximadamente 30 segundos de duración. Las señales fueron grabadas en salas de con aislación de sonido, en ausencia de ruido. El entrenamiento se realiza con 2 señales, mientras que para evaluar el algoritmo se ocupan las 8 señales restantes, tal como se explica en la Tabla 4.1.

Entrenamiento	Test
F1	F2 F3 F4 F5
M1	M2 M3 M4 M5

Tabla 4.1. Entrenamiento y test para evaluar el método SSPD

4.3.1.2. Entrenamiento de Clasificador Sonoro-Sordo Basado en SSPD.

El entrenamiento consiste en determinar las funciones de distribución para realizar una clasificación sonoro-sordo basada en la regla de Bayes. Para esto, se define la *profundidad espectral*, denotada por xFO. Ésta es una característica experimental, observable en la función objetivo, que se calcula como se muestra en (4.1).

$$xFO = F_{OBJ}\left(\frac{3}{2} f_0\right) + F_{OBJ}\left(\frac{3}{4} f_0\right) - 2 \cdot F_{OBJ}\left(f_0\right) \tag{4.1}$$

donde f_0 es la frecuencia fundamental calculada por el método SSPD. Como el rango de búsqueda de f_0 está acotado entre 70[Hz] y 450[Hz], hay que tener cuidado de no buscar los armónicos fuera del dominio de la función objetivo. Por esta razón, la profundidad espectral cambia a la siguiente expresión:

$$xFO = F_{OBJ}\left(max\left\{\frac{3}{4} f_0;70\right\}\right) + F_{OBJ}\left(max\left\{\frac{3}{2} f_0;450\right\}\right) - 2 \cdot F_{OBJ}\left(f_0\right) \tag{4.2}$$

Como ya se explicó, la función objetivo presenta mínimos locales en el armónico $\left(2 \cdot f_0\right)$ y en el sub-armónico $\left(1/2 \cdot f_0\right)$. Por esta razón, es razonable pensar que el valor medio entre el pitch y sus armónicos debiese estar cercano a un máximo local (en el caso de segmentos sonoros). Por esta razón, se calcula la distancia entre dichos valores cercanos al máximo, y el mínimo global, tal como se desprende de la Figura 4.1

Si el segmento es sordo, el rango de la función objetivo es más pequeño. Por lo tanto, el valor de xFO debiese ser también más pequeño. Además, por lo general, para los segmentos sordos el resultado de SSPD entrega valores cercanos a 70[Hz]. Por lo mismo, el sub-armónico casi nunca está en el dominio de la función objetivo, y la mayoría de las veces toma el valor 70[Hz] de acuerdo a la ecuación (4.2), como se ejemplifica en la Figura 4.2.

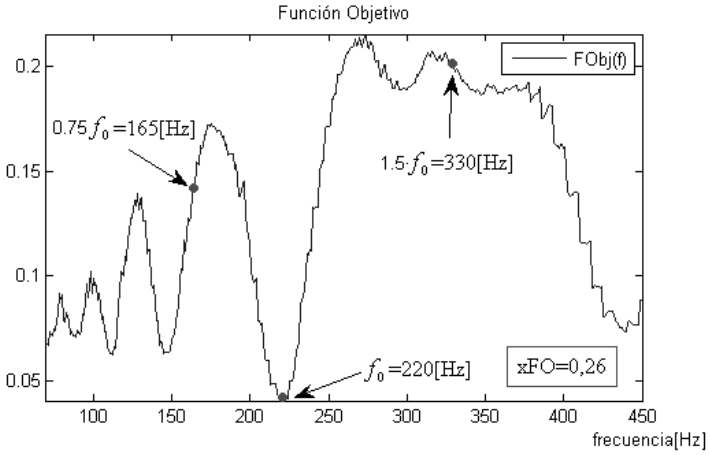

Figura 4.1: Profundidad espectral xFO para un segmento sonoro. La profundidad espectral tiene un valor de 0.26

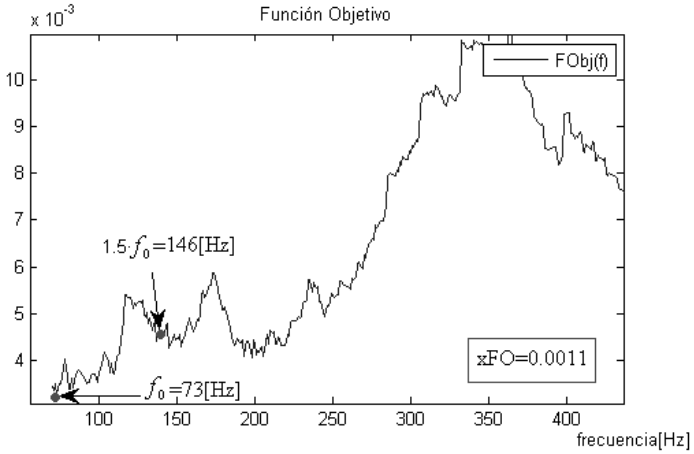

Figura 4.2. Profundidad espectral xFO para un segmento sordo. La profundidad espectral tiene un valor de 0.0011

Ahora que se conoce la característica xFO, se puede programar un clasificador simple para tomar la decisión sonoro-sordo. Esta decisión se tomará segmento a segmento, basado en las distribuciones que presenten ambos estados para dicha característica. Se encontró que la forma de distribución de la característica es log-normal.

Una vez conocida la distribución a priori, simplemente se debe tomar la siguiente determinación:

$$Si \ \frac{P_{sonoro}\left(xFO\middle|sonoro\right)}{P_{sordo}\left(xFO\middle|sordo\right)} \geq \theta \quad \Rightarrow segmento\ sonoro$$
$$en\ otro\ caso \quad \Rightarrow segmento\ sordo \tag{4.3}$$

donde θ corresponde a un umbral de decisión. Si $\theta = 1$, entonces la decisión se toma en la intersección de las curvas, siguiéndose la regla de Bayes, tal como se establece en la Figura 4.3. Si $\theta < 1$, el límite se mueve a la derecha, y si $\theta > 1$, a la izquierda. Lo último ocurre cuando se quiere privilegiar algún estado sobre el otro, dándole mayor probabilidad de ocurrencia, pero el clasificador de Bayes es el que asegura que la probabilidad de error sea mínima.

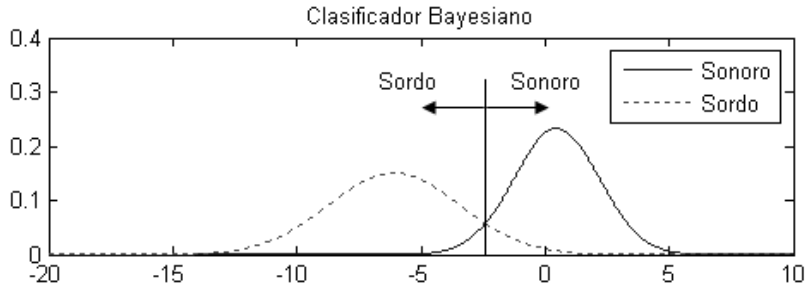

Figura 4.3. Clasificador Bayesiano ($\theta = 1$)

4.3.1.3. Resultado de Experimentos SSPD.

Los experimentos se realizan sobre la base de datos Keele. Estos consisten en detectar el pitch en un análisis segmento a segmento, y obtener una medida objetiva para medir el rendimiento del detector. Para esto, se utilizan los indicadores GPE, VE, UE, MEAN y STD, utilizados ampliamente en el área de detección de pitch.

En esta sección se calcularán los indicadores para algunas configuraciones, entre las cuales se elegirá la que entregue mejores resultados, para ser utilizada en el resto del trabajo. Los parámetros a variar serán el número de coeficientes LPC y la frecuencia de muestreo. El número de puntos para el cálculo de la FFT se dejará constante para todos los

experimentos e igual a 1024, para no aumentar la carga computacional. Además, se evaluará el método de mejoramiento de la envolvente LPC.

El criterio para elegir la mejor configuración será la siguiente: la opción que ofrezca el mínimo GPE será la que se ocupará en el resto del trabajo de esta memoria. Ese es el indicador que la mayoría de los estimadores buscan minimizar para evaluar su desempeño. Dicho criterio se evaluará sobre señales limpias (sin ruido), para luego calcular los resultados sobre la señal ruidosa. Las configuraciones utilizadas se muestran en la Tabla 4.2

Configuración	Frecuencia de Muestreo [kHz]	Número de coeficientes LPC
A	8	10
B	8	8
C	8	6
D	4	8
E	4	6
F	4	4
G	2	4

Tabla 4.2. Configuraciones utilizadas para evaluar método SSPD

Los resultados para las señales sin mejoramiento LPC se muestran en la Tabla 4.3. Además, para los experimentos realizados con el mejoramiento de la envolvente LPC, se debe fijar el número de iteraciones realizadas por el algoritmo. Experimentos realizados mostraron que con un número de 400 iteraciones las curvas muestran mejoras significativamente, y, además, con números mayores a éste la mejora es casi imperceptible. Utilizando este nuevo parámetro, los resultados se muestran en la Tabla 4.4. Además, se muestra un gráfico resumen para los resultados de GPE obtenidos en la Figura 4.4.

La configuración que entrega el mínimo GPE es la configuración B con mejoramiento LPC, es decir, utilizando una frecuencia de muestreo de 8[kHz] y un número

de coeficientes LPC igual a 8. Por lo tanto, para lo que resta de trabajo, ésta será la configuración que se utilizará para el resto de los experimentos.

Configuración	VE [%]	UE [%]	GPE [%]	MEAN [%]	STD [%]
A	6.79	1.87	**3.87**	3.13	2.73
B	7.05	1.83	**4.31**	3.18	2.86
C	7.07	1.79	**3.85**	3.15	2.77
D	5.66	2.24	**7.85**	2.17	2.15
E	6.19	1.96	**4.80**	2.29	2.31
F	6.67	1.87	**3.94**	2.34	2.40
G	5.72	2.00	**7.78**	1.99	2.05

Tabla 4.3. Resultados de SSPD para señal limpia

Configuración	VE [%]	UE [%]	GPE [%]	MEAN [%]	STD [%]
A	6.66	1.75	**3.19**	2.98	2.55
B	6.65	1.78	**2.63**	2.99	2.57
C	6.63	1.73	**2.67**	2.98	2.54
D	5.81	2.31	**7.52**	2.02	2.01
E	6.38	1.98	**5.17**	2.09	2.09
F	6.60	1.96	**4.44**	2.10	2.13
G	6.53	1.68	**5.13**	1.96	2.10

Tabla 4.4. Resultados de SSPD con señal limpia, utilizando método de mejoramiento de la envolvente LPC.

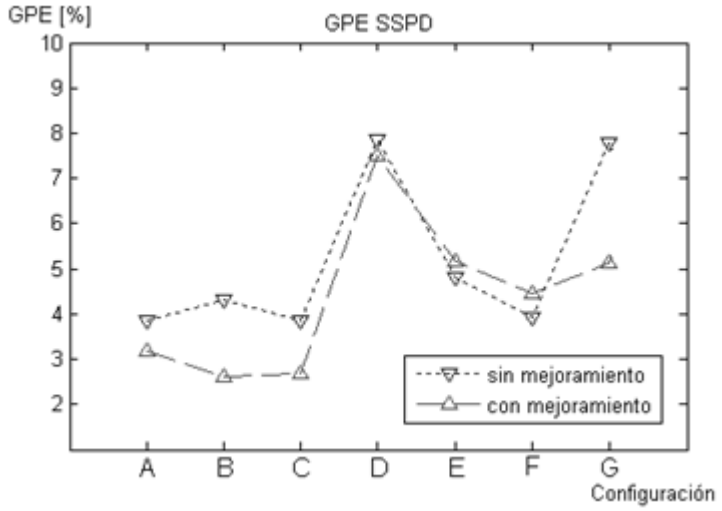

Figura 4.4. GPE para distintas configuraciones, con y sin mejoramiento espectral del filtro LPC.

4.3.2. Experimentos Clasificación Sonoro-Sordo

Se pretende evaluar una clasificación sonoro-sordo basada en el modelo de Gilbert. Para esto, se evaluará dicho algoritmo en tres características distintas, utilizadas en la literatura. Se comparará el desempeño de los modelo con y sin duración de estado, en contraste con un clasificador bayesiano simple.

4.3.2.1. Base de dato clasificación Sonoro-Sordo.

Se utilizará la base de datos Keele. El entrenamiento se realiza con 5 señales, mientras que para evaluar el algoritmo se ocupan las 5 señales restantes, tal como se explica en la Tabla 4.5.

Entrenamiento	Test
F1	F3
F2	F4
M1	F5
M2	M4
M3	M5

Tabla 4.5. Distribución de base de datos

4.3.2.2. Entrenamiento para Clasificación Sonoro-Sordo

Los experimentos para determinar el rendimiento del detector sonoro-sordo se basan en comparar un clasificador bayesiano simple, un modelo de Gilbert simple (sin duración de estado) y un modelo de Gilbert con duración de estado. La evaluación se realizará para 3 **características** de la señal

- Energía del segmento dividida por tasa de cruces por cero, xEZcr.
- Peak de Autocorrelación del segmento, xAC.
- Profundidad Espectral, xFO.

Las dos primeras características son muy utilizadas en publicaciones científicas para determinar segmentos sonoros o sordos. La última se explicó en SSPD.

Para determinar las probabilidades de observación para cada característica, se debe utilizar información conocida a priori. De esta manera, se calculan las características y, utilizando la referencia de Keele, se sabe a priori si el segmento es sonoro o sordo. Con este procedimiento, se puede determinar la distribución que sigue cada una de ellas.

En la práctica, todas esas características son bien representadas por una distribución log-normal, es decir, el logaritmo de las características sigue una distribución normal. Por esta razón, calculando la media y la varianza de dicha distribución se habrá parametrizado el comportamiento de la característica. Los gráficos de las curvas obtenidas para el logaritmo de cada característica se vislumbran en la Figura 4.5

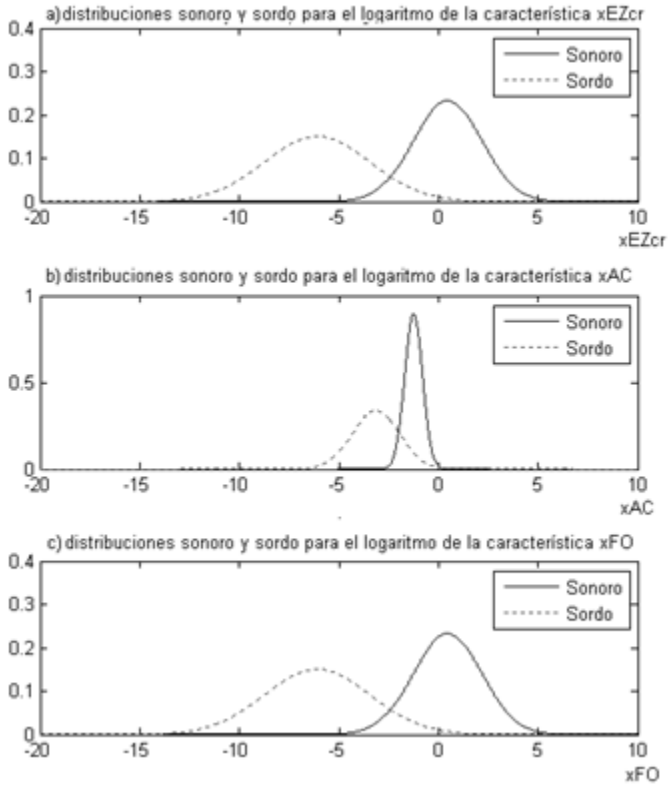

Figura 4.5. Distribuciones del logaritmo de las características a evaluar. a) Energia dividida por cruces por cero. b) peak de autocorrelación. c) profundidad espectral.

Sólo conociendo estas distribuciones de probabilidad, se puede utilizar un clasificador bayesiano para tomar la decisión sobre el estado del segmento (sonoro o sordo).

Si además se agrega la información de la longitud de los tramos sonoros y sordos, se estaría en condiciones de utilizar un modelo de Gilbert. La información del largo de los tramos se obtiene de la misma base de datos Keele. Se observa el comportamiento de la Figura 4.6.

Para el modelo simple, se asume que las distribuciones son geométricas, por lo que son modeladas solamente por $E_0[VL]$ y VR. Se conoce la cantidad de segmentos sonoros y sordos directamente de la referencia. Además, se puede calcular un estimador del valor esperado de *VL* a través de la media de la distribución de los segmentos sonoros a través de la distribución vista en la Figura 4.6 a), pues para una distribución geométrica, el estimador de máxima verosimilitud de la esperanza es, en efecto, la media muestral. Con esto, se puede calcular las probabilidades de transición $a_{0,0}$, $a_{0,1}$, $a_{1,0}$, $a_{1,1}$ mediante (3.9); (3.10); (3.11); y, (3,12).

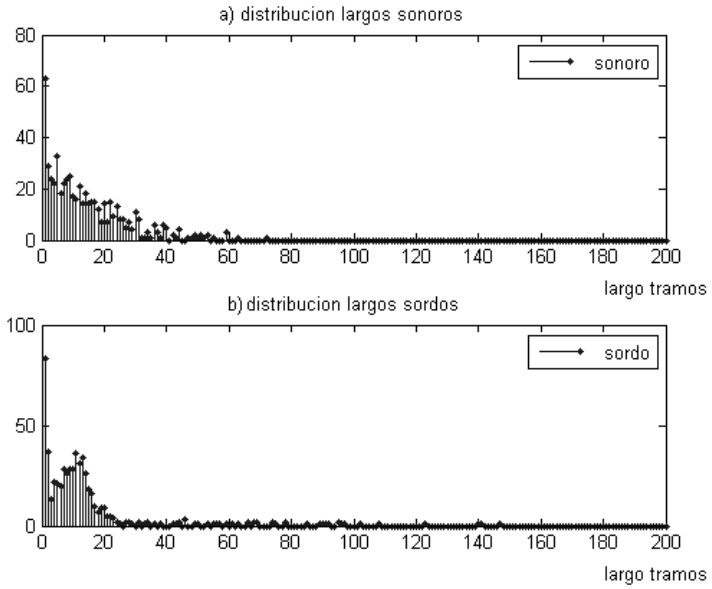

Figura 4.6. Distribución de largos a) sonoros y b) sordos.

Para el modelo de Gilbert con duración de estado se requiere conocer la distribución completa de los tramos sonoros y sordos. Para esto es necesario parametrizar dichas distribuciones. Se observa que, en ambos casos (sonoro y sordo), pareciera que existen dos distribuciones. Una presente en las duraciones más bajas, y una para las más altas. De esta manera, se plantea la siguiente hipótesis:

- Distribución de segmentos sonoros: primeros puntos siguen una distribución Geométrica, y los últimos siguen una distribución Gamma.

- Distribución de segmentos sordos: primeros puntos siguen una distribución Geométrica, y los últimos siguen una distribución Gaussiana.

- El valor asignado a $a_{i,i}^{VL=1}$ será la probabilidad del *Estado i* de tener largo 1.

donde las distribuciones mencionadas se definen como

$$Geom(x;p) = p^{x-1}(1-p) \quad x \in \text{IN} \tag{4.4}$$

$$Gamma(x;k,\theta) = x^k \frac{e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)} \tag{4.5}$$

$$Gauss(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4.6}$$

$\Gamma(\cdot)$ corresponde a la función Gamma. Los parámetros *p, k, θ, μ, σ* se obtienen utilizando los estimadores de máxima verosimilitud para cada una de ellas. Por lo tanto, cada distribución queda descrita por las siguientes fórmulas

$$Distr_0(x) = \begin{cases} Geom(x;p_0) & x \le 4 \\ Gamma(x;k,\theta) & x \ge 4 \end{cases} \tag{4.7}$$

$$Distr_1(x) = \begin{cases} Geom(x; p_1) & x \le 4 \\ Gauss(x; \mu, \sigma) & x \ge 4 \end{cases} \tag{4.8}$$

donde el subíndice 0 corresponde a la distribución sonoro, y 1 corresponde a sordo. En la Figura 4.7 se muestra el resultado de las parametrizaciones recién mencionadas

Con esto, se está en condiciones de determinar las probabilidades de transición $a_{i,i}^{VL}$ y $a_{i,i\pm1}^{VL}$. En la Figura 4.8 se observan dichas probabilidades en función del largo del tramo (sonoro o sordo, según corresponda).

4.3.2.3. Resultados de experimentos clasificación sonoro-sordo

Para medir la detección sonoro sordo, en general se utilizan los errores VE y UE. Como ambos errores son igual de importantes en este trabajo, lo que se busca es minimizar el número total de ocurrencia de ambos, por lo que se busca minimizar el VUVE.

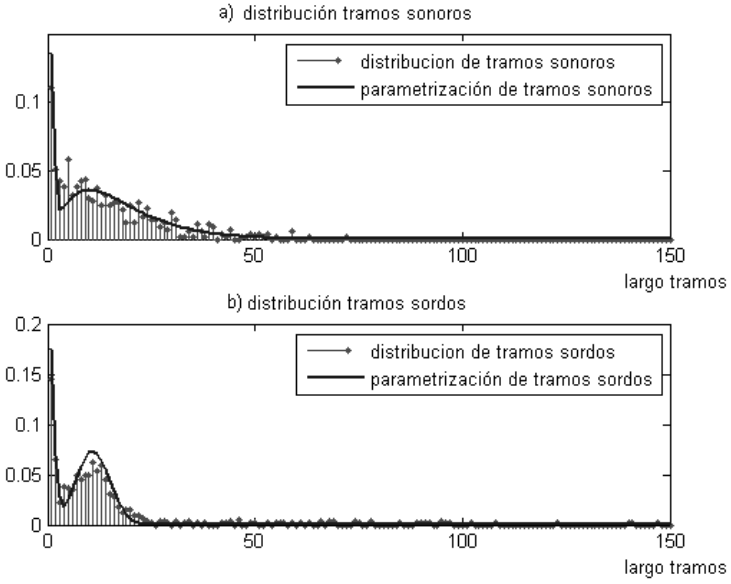

Figura 4.7. Distribuciones de tramos a) sordos, y b) sonoros. a) se parametriza con una distribución Geométrica y una Gamma, mientras que b) se parametriza con una Geométrica y una Gaussiana.

Los resultados obtenidos para cada característica, utilizando las tres alternativas planteadas, se reflejan en la Tabla 4.6.

	Característica		
	xEZcr VUVE[%]	xAC VUVE[%]	xFO VUVE[%]
Clasificador Bayesiano	9.58	10.96	6.42
Modelo de Gilbert Simple	9.09	8.15	6.19
Modelo de Gilbert con duración de estado	8.77	8.07	6.18

Tabla 4.6. Resultados clasificación sonoro-sordo para señales de test

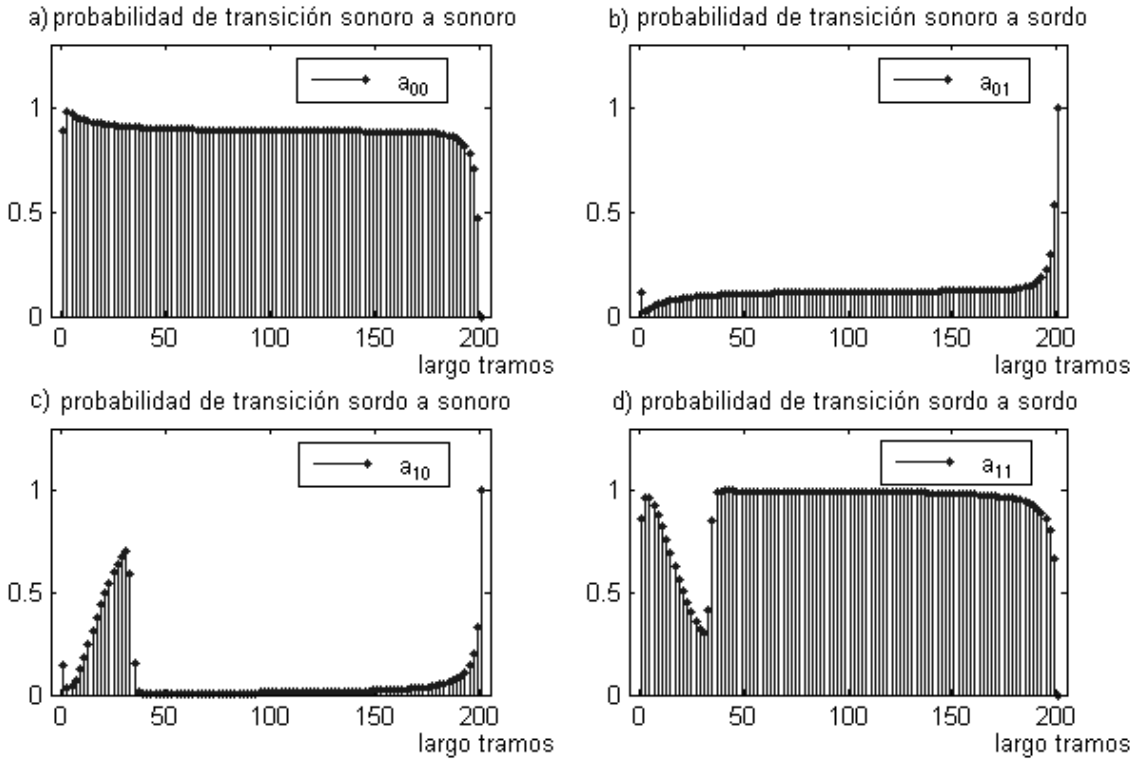

Figura 4.8. Probabilidades de transición en función del largo del tramo a) sonoro a sonoro, b) sonoro a sordo, c) sordo a sonoro, y d) sordo a sordo.

4.3.3. Experimentos post-procesamiento.

En esta sección se pretende realizar un análisis global del pitch en lugar de el análisis segmento a segmento realizado por SSPD. Se mostrarán los resultados obtenidos por el algoritmo de programación dinámica propuesto para la configuracion óptima de

SSPD, y con el clasificador sonoro-sordo basado en un modelo de Gilbert con duración de estado para la característica profundidad espectral.

4.3.3.1. Base de datos para post-procesamiento.

Se utilizará la base de datos Keele. El entrenamiento se realiza con 5 señales, mientras que para evaluar el algoritmo se ocupan las 5 señales restantes, tal como se explica en la Tabla 4.7.

Entrenamiento	Test
F1	F3
F2	F4
M1	F5
M2	M4
M3	M5

Tabla 4.7. Distribución de base de datos post-procesamiento

4.3.3.2. Entrenamiento de vector de pesos para post-procesamiento

Las funciones objetivos y la decisión sonoro-sordo fueron previamente calculadas con el método SSPD, y con el modelo de Gilbert, respectivamente. El entrenamiento para esta etapa consiste en calcular los pesos para ponderar las transiciones de un segmento a otro. Esto se realiza calculando la derivada del pitch de las señales de referencia, la cual se define como

$$D(n) = p(n+1) - p(n) \tag{4.9}$$

donde p denota al pitch de la señal en cuestión. Después de obtener la distribución de derivadas, se crea un histograma de ellas, como se muestra en la Figura 4.9.

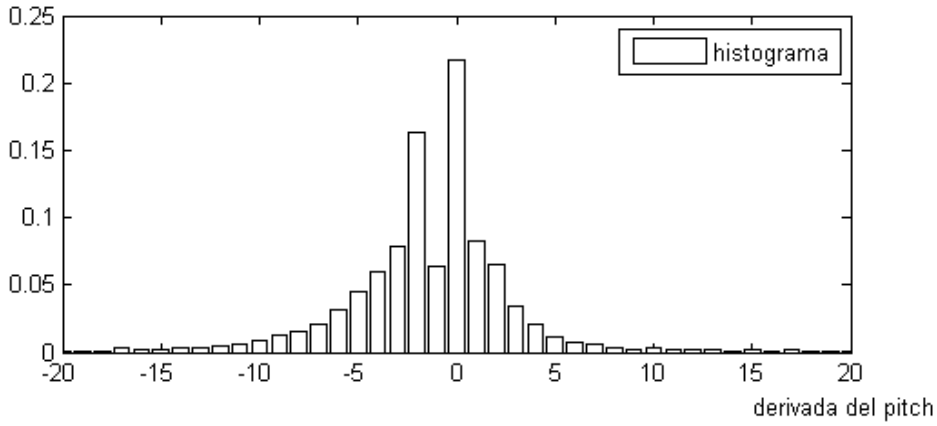

Figura 4.9. Histograma de la derivada del pitch de referencia.

Se observa que no es una distribución regular, pero sin embargo sí se aprecia que, para transiciones mayores a 3 estados, la curva disminuye mientras mayor es la transición. A pesar de la poca uniformidad del histograma, se intentó la modelación del mismo con dos distribuciones distintas: la distribución normal y la distribución t-location scale, cuyo resultado se ve en la Figura 4.10

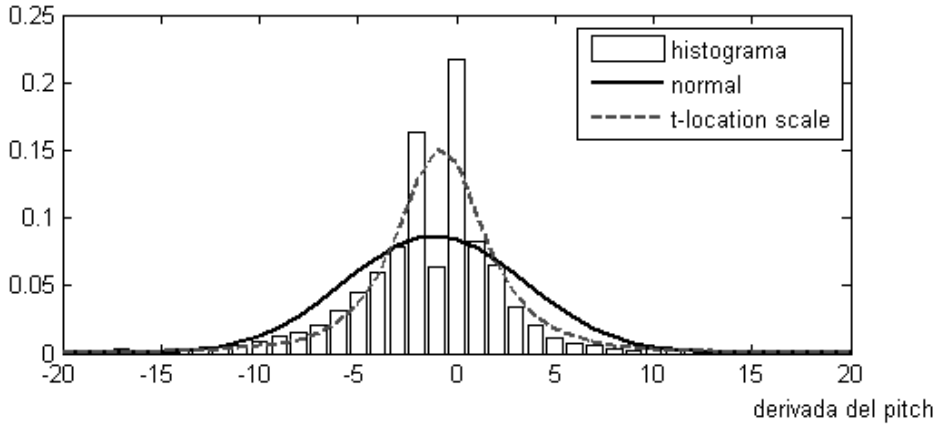

Figura 4.10. Distribuciones normal y *t-location scale* para el histograma de distribución de la derivada.

Las distribuciones tienen valores mayores para las transiciones más bajas, y valores menores para las altas. Como se quiere minimizar una función, enfatizando los cambios de estado bajos, se utiliza como función peso:

$$w_j(k) = 1 - distr_j(k) \tag{4.10}$$

El subíndice *j* indica alguna de las distribuciones utilizadas para modelar el histograma. Con esta estrategia se pretende mejorar el resultado obtenido del análisis segmento a segmento realizado por SSPD.

4.3.3.3. Resultados Experimentos de post-procesamiento

Los experimentos realizados consisten en encontrar el camino óptimo que debe recorrer la curva de pitch. Éste camino óptimo debe estar sujeto a un cambio máximo de pitch entre un segmento y el siguiente. Esta restricción define el rango de búsqueda explicado anteriormente. Observaciones realizadas en la base de datos muestran que los cambios de pitch mayores a 30[Hz] entre segmentos son muy improbables, siendo la mayoría de éstos correspondientes a saltos de octava. Por esta razón, se define el rango de búsqueda como:

$$rango\ de\ busqueda = [p - 30\ ;\ p + 30][Hz] \tag{4.11}$$

donde *p* corresponde al valor en el segmento anterior. Se subentiende que el rango debe estar en el dominio de la función objetivo. Es decir, algún valor del rango cae fuera del dominio, no se considera para el análisis.

El post-procesamiento ofrece una resolución de 1[Hz] para la curva de la frecuencia fundamental. Es decir, el pitch puede tomar valores discretos que difieren en 1[Hz]. El rango de frecuencias en análisis corresponde a las frecuencias superiores a 70[Hz] e inferiores a 450[Hz]. Por lo tanto, el rango de frecuencia comprende los valores 70[Hz], 71[Hz], … , 449[Hz], 450[Hz].

La clasificación sonoro-sordo se calcula utilizando el resultado obtenido para el modelo de Gilbert con duración de estado, para la característica xFO. Se utilizan las

mismas 5 señales de entrenamiento para determinar las distribuciones, y se testea con las restantes.

El caso base para comparar utiliza la misma detección sonoro-sordo, pero el pitch se determina encontrando el mínimo de la función objetivo segmento a segmento. Los resultados para el post-procesamiento para todas las opciones presentadas están en la Tabla 4.8.

	VE[%]	UE[%]	GPE[%]	MN[%]	STD[%]
Caso base	3.54	2.64	3.12	2.21	2.50
Sin pesos	3.54	2.64	2.01	2.29	2.72
Pesos t-location scale	3.54	2.64	1.61	2.66	3.25
Pesos normal	3.54	2.64	1.57	2.44	3.07

Tabla 4.8. Resultados de Post-procesamiento.

4.4. Discusión de resultados.

En las Tablas 4.3 y 4.4 se observa que el GPE más bajo es el obtenido por la configuración B. Ésta corresponde a una configuración típica de filtros LPC para la voz humana. La configuración recomendada por muchos autores es, para la voz humana, 8[kHz] como frecuencia de muestreo con 8 o 10 coeficientes LPC (Deller et al, 2000). Se observa que otras alternativas mejoran esta configuración sin el mejoramiento LPC realizado en la segunda sesión de experimentos (como las configuraciones A, C y F). Sin embargo, son superados por la configuración B cuando se aplica dicha técnica. Como se explicó, la técnica tiene como objetivo mejorar las características espectrales del filtro LPC, lo que se observa, ocurre en el caso de la frecuencia de muestreo igual a 8[kHz]. En general, la técnica funciona bien salvo para algunas configuraciones, donde se logran porcentajes de error más altos. La Tabla 4.9 muestra el efecto producido por la técnica de mejoramiento LPC, en término de porcentaje de mejora. A pesar de que el filtro no mejora los resultados para las configuraciones E y F, la configuración B mejora un 39.11% respecto al caso sin la técnica. Este es una mejora realmente significativa, llegando a un valor de GPE de 2.63%.

Config	% mejora
A	17.46
B	39.11
C	30.80
D	4.32
E	-7.62
F	-12.87
G	34.12

Tabla 4.9. Porcentaje de mejora de filtro LPC.

También se puede apreciar que la característica profundidad espectral es un buen discriminador del estado de voz (sonoro o sordo). Utilizando un clasificador bayesiano simple, con sólo 2 señales de entrenamiento para obtener las distribuciones, se tienen errores de detección en todos los casos de alrededor del 8.5%. Para la configuración B, que es la de interés pues con ella se hicieron los demás experimentos, se observa un VUVE de 8.43%, resultado cercano al estado del arte.

Para el detector sonoro-sordo se aplicó un modelo probabilístico basado en procesos ocultos de Markov, utilizando un clasificador de Bayes en una primera aproximación, y finalmente un modelo de Gilbert con y sin duración de estado. Los resultados de la Tabla 4.10 muestran las mejoras relativas del modelo de Gilbert respecto al clasificador bayesiano.

	Mejoras relativas [%]		
	Característica		
	xEZcr	xAC	xFO
Modelo de Gilbert Simple	5.03	25.59	3.66
Modelo de Gilbert con duración de estado	8.24	26.36	3.76

Tabla 4.10. Mejoras relativas (en porcentaje) del modelo de Gilbert respecto al clasificador Bayesiano.

Se observa que el modelo de Gilbert entrega mejores resultados que el clasificador bayesiano simple. Esto concuerda con lo que se espera de este resultado, pues el modelo de Gilbert incluye información adicional a la clasificación. El conocimiento de la longitud de los tramos sonoros y sordos ayuda, entonces, a tomar una mejor decisión y, por tanto, a realizar una mejor clasificación.

Las mejoras relativas son bastante dispares. Para la característica xAC, la mejora es bastante importante, mientras que para xFO no lo es tanto. Esto indica que, a pesar de mejorar en todos los casos, ésta depende de la característica utilizada.

También se puede determinar la mejora relativa del modelo de Gilbert con duración de estado respecto al modelo simple. La Tabla 4.11 entrega la información sobre esta comparación

	Mejoras relativas [%]		
	Característica		
	xEZcr	xAC	xFO
Modelo de Gilbert con duración de estado	3.37%	1.05%	0.13%

Tabla 4.11. Mejora de modelo de Gilbert con duración de estado, respecto al modelo simple.

Se puede apreciar que, a pesar de obtener resultados positivos, las mejoras no son muy importantes. Principalmente para xFO, la mejora es casi imperceptible. Lo más probable es que el problema radique en una mala determinación de las funciones de distribución de los tramos sonoros y sordos, pues este resultado sugiere que, para la característica xFO, asumir el modelo aquí mencionado entrega resultados casi iguales a utilizar una distribución geométrica para describir tanto el caso sonoro como sordo. Esta parte fue la que más problemas conllevó a la hora de realizar el trabajo de esta memoria en la parte de detección sonoro-sordo, pero los resultados finales llevan a pensar que no se logró obtener una mejora considerable.

El post-procesamiento, consistente en la determinación final de la curva de pitch, muestra resultados bastante alentadores. Se observan mejoras considerables, las cuales indican que para la curva de entonación, en este caso, es mejor un análisis global de la frecuencia fundamental en lugar de un análisis segmento a segmento, como lo realiza SSPD. En las Tablas 4.12 y 4.13 se muestran las diferencias porcentuales absoluta y relativa entre los casos con post-procesamiento en relación al caso base

	diferencia absoluta entre GPE [%]
Sin pesos	1.10
Pesos t-location scale	1.51
Pesos normal	1.54

Tabla 4.12. Diferencia absoluta entre caso base y casos con post-procesamiento.

	diferencia relativa entre GPE [%]
Sin pesos	35.45
Pesos t-location scale	48.47
Pesos normal	49.50

Tabla 4.13.: Diferencia relativa entre caso base y casos con post-procesamiento.

Las diferencias absolutas, hablando en términos relativos a GPE, son muy grandes, pero son las relativas las que reflejan mejor el efecto del post-procesamiento aplicado a la función objetivo. Se observan resultados alrededor del 35% más exactos para el caso sin pesos, mientras que los casos con pesos rondan el 50% de mejora. Estos valores reflejan la importancia de la etapa de post-procesamiento.

A continuación, se comparará el desempeño del método calculado en esta memoria con otros métodos (*Roa el Al, 2007*), que también utilizan la base de datos Keele para evaluar su rendimiento. En la Tabla 4.14 se muestran los resultados.

Método	VE	UE	GPE	MN
PSHF	4.51	5.06	0.61	2.46
NMF	7.7	4.6	0.9	4.3
RAPT	3.2	6.8	2.2	4.4
Método propuesto	**3.54**	**2.64**	**1.57**	**2.44**

Tabla 4.14. Comparación con métodos de estimación de pitch.

El método desarrollado en esta memoria muestra resultados totalmente comparativos con estos estimadores, los cuales están en el estado del arte. Sin entrar en detalles respecto a lo que desarrolla cada uno de ellos, se observa que NFM, si bien tiene un GPE mejor que el método propuesto, tiene un VE altísimo, por lo que muy probablemente

hay muchos segmentos conflictivos mal clasificados y, por tanto, no entran en la estadística de GPE. PSHF presenta un resultado notable en cuanto a GPE, pero no muestra un clasificador sonoro-sordo demasiado robusto. Los métodos de comparación aquí presentados presentan VUVE bastante altos, 9.57%, 12.3% y 10%, mientras que la prupuesta presentada en este trabajo tiene un VUVE de 6.18%.

Cuando se trata de estimadores de pitch, el bajar un error usualmente conlleva el aumento de otro. Se produce un *trade-off* entre los distintos tipos de errores que existen. En general, se busca minimizar el GPE en perjuicio del resto de los indicadores, tal como se refleja en la Tabla 4.14. En este trabajo se observa un compromiso entre los errores VUVE y GPE.

Capitulo 5.

Conclusiones

5.1. Conclusiones Generales

En este trabajo se propuso un método de estimación de la curva de entonación, el cual es una piedra angular en el desarrollo y funcionamiento de un evaluador de entonación. Se planteó como objetivo obtener errores GPE, VE y UE similares al estado del arte, de manera de ser comparable con ellos. Para esto, se desarrolló un sistema para la determinación de las funciones objetivo minimizar, SSPD, basado en la comparación de una señal con su símil sintetizado; un clasificador de segmentos sonoros y sordos, el cual se cimenta en un proceso oculto de Markov de dos estados; y, finalmente, una etapa de post-procesamiento, en la cual se desarrolla un algoritmo de programación dinámica para obtener un resultado para la curva de entonación en base a toda la función objetivo, en lugar de hacer un análisis segmento a segmento.

Primero se determinó la configuración de parámetros que ofrecía el menor GPE. Ésta resulto ser la denominada opción B, consistente en una frecuencia de muestreo de 8[kHz] y un número de coeficientes LPC igual a 8, obteniendo un GPE de 2.63%. Esta configuración es muy conveniente pues muchos codificadores de voz utilizan esa

frecuencia de muestreo, por lo que se evitaría la etapa de sub-muestrear la señal como parte del pre-procesamiento.

En la etapa siguiente, la clasificación sonoro-sordo permitió concluir que la mejor característica de las utilizadas para discriminar entre segmentos sonoros y sordos fue la profundidad espectral xFO, tanto para el clasificador bayesiano simple como también para los modelos de Gilbert con y sin duración de estado. Estas últimas opciones fueron las que arrojaron los resultados menos satisfactorios de este trabajo. Se esperaba una mejora significativa entre el modelo de Gilbert sin duración de estado y el modelo con. El segundo de ellos mejora, pero de manera demasiado tímida, donde incluso para xFO es casi imperceptible, de un VUVE igual a 6.19% a uno de 6.18%. De todas maneras, este valor de VUVE es bajo, comparado con los métodos utilizados en la literatura, y superior a las características energía dividida por tasa de cruces por cero, xEZcr; y peak de autocorrelación, xAC.

Finalmente, en la etapa de post-procesamiento se realizó la comparación final entre el algoritmo con y sin post-procesamiento. La mejora fue de un 49.5%, bajando un GPE de 3.11% a un 1.57%. Este valor de GPE, si bien no está entre los más bajos, es un valor bastante aceptable, sobre todo considerando el nivel de VUVE utilizado. Los métodos que superan al método desarrollado en esta memoria tienen VUVE más altos, todos sobre el 9.5%, lo que, como se explicó antes, explica el bajo GPE obtenido. En general, un método con bajo GPE tiene un alto VUVE y viceversa. En el caso del estimador desarrollado SSPD-post, se tienen un GPE razonable y un VUVE bajo (6.18%), con lo que se puede concluir que los resultados obtenidos son comparables al estado del arte.

5.2. Trabajo Futuro

El módulo más débil de todos fue la clasificación sonoro-sordo con modelo de Gilbert. Este método no mostro mejoras considerables dentro de las expectativas creadas ante tan novedoso método. Lo más seguro es que la parametrización de las distribuciones de los largos de los tramos sonoro y sordo no fue bien estimada. Por lo tanto, se propone

trabajar más a fondo en este tema, de manera de lograr un resultado más alentador que el que se obtuvo en este trabajo.

Otro tema importante que no se trató fue la utilización de otra base de datos para evaluar el algoritmo. Existe una base de datos, llamada *CSTR database* (sigla de *Centre for Speech Technology Research*, *Edinburgh University*), que se podría haber utilizado. La idea original siempre fue entrenar con una base de datos y testear con otra. Lamentable, al final de este trabajo no se logró la realización de los experimentos con ambas bases de datos.

El pitch de referencia entregado por Keele es otro aspecto a tratar en el futuro. Para ser lo más transparente posible se utilizó el archivo de comparación entregado en la base de datos. Sin embargo, la referencia muestra errores extraños en la curva de pitch, similares a errores de octava, o fallas en la clasificación sonoro-sordo, que pueden afectar la comparación. Por lo mismo, como se cuenta con las señales de laringografía, se podría estimar el pitch de referencia directamente de ellas, en lugar de utilizar dicha referencia pre-calculada.

Finalmente, al término este trabajo, no se logró la integración del método desarrollado con el sistema de evaluación. Se espera que esto se realice lo antes posible para poder evaluar el comportamiento del mencionado sistema, y verificar que realmente se realizó un aporte en el aprendizaje de segundo idioma

Referencias

Ahmadi, S., and Spanias, A.S.: "*Cepstrum-based pitch detection using a new statistical V/UV classification algorithm*", ZEEE Trans Speech and Audio Process, 7(3), 1999, 333-338.

Amado, R.G.; Filho, J.V.: "*Pitch detection algorithms based on zero-cross rate and autocorrelation function for musical notes*" Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on 7-9 July 2008 Page(s):449 - 454

Arias, J. P.: "*Evaluación de pronunciación por tono para enseñanza de segundo idioma*". Memoria de título, Departamento de Ingeniería Eléctrica, Universidad de Chile, 2008.

Arias, J.P., Yoma, N.B., Vivanco, H.:, "*Automatic intonation assessment for computer aided language learning*", 2009.

Atal, B.S. and Hanauer, S.L.: "*Speech Analysis and Synthesis by Linear Predictive Coding of the Speech Wave*", Journal of the Acoustic Society of America 50, 2,2: 637-655; in FLANAGAN, 1971

Atal, B and Rabiner, L.: "*A pattern recognition approach to voiced-unvoiced-silence classiffication with applications to speech recogntion*". IEEE transactions on Acoustics, Speech and Signal Processing, ASSP-24(3):201212, June 1976.

Bagshaw, Paul.: "*Automatic prosodic analisys for computer aided pronunciation teaching*", PhD Thesis, University of Edinburgh, 1994

Beauchaine, Bob: "*A Simple LPC Vocoder*" EE586, 2004

Cerdá, Salvador, Romero, J: "*Use of multiresolution analysis to calculate pitch in presence of noise*". J. Acoust. Soc. Am. Volume 99, Issue 4, pp. 2497-2500 (April 1996)

de Cheveigné, Alain and Kawahara, Hideki: "*YIN, a fundamental frequency estimator for speech and music"*. Acoustical Society of America Journal, Volume 111, Issue 4, pp. 1917-1930. 2002.

Deller, John, Hansen, John, Proakis, John: *Discrete-Time Processing of Speech Signals*, 2nd. edition, Wiley-IEEE Press, 2000

Demeechai, T., and Mäkeläinen, K.: "*New method for mitigation of harmonic pitch errors in speech recognition of tone languages*". Proc. IEEE Nordic Signal Processing Symp., 2000, pp. 303–306

Dubnowski, J.J., Schafer, R.W., and Rabiner, L.R.: "*Real-time digital hardware pitch detector*", IEEE Trans. Acoust., Speech and Signal Processing, Vol ASSP-24, pp. 2-8, February 1976.

Duda, R.O., Hart, P.E. and Stork, D.G.: "*Pattern classification and scene analysis*". publisher: Wiley New York. 1973

Dziubinski M. and Kostek B.: "*High Accuracy and Octave Error Immune Pitch Detection Algorithms*". Archives of Acoustics, 2004

González, Rosa.: "*Fundamental Frequency Estimation and Modeling for Speaker Recognition*". Master's thesis, Department of Computer Science, University of Joensuu. 2005

Hess, W.: "*Pitch Determination of Speech Signals: algorithms and devices"*. Springer-Verlag, Berlin, 1983.

Hui, L., Dai, B.-q., and Wei, L.: "*A pitch detection algorithm based on AMDF and ACF*" in Proc. ICASSP2006, pp.377-380, 2006.

Kedem, Benjamin. "*Spectral analysis and discrimination by zero-crossings*".: Proceedings of the IEEE, 74(11):1477–1493, November 1986.

Kotnik, B., Höge, H., and Kacic, Z.: "*Evaluation of Pitch Detection Algorithms in Adverse Conditions*". Proc. 3rd International Conference on Speech Prosody, Dresden, Germany, pp. 149-152, 2006.

Krubsack D.A., and Niederjohn, R.J.: "*An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech"*, IEEE Trans. Acoust., Speech, Signal Process. **ASSP-39** (1991) (2), pp. 319–329

Maqsood, Hania, Gudnason, J., and Naylor, P.A.: "*Enhanced Robustness to Unvoiced Speech and Noise in the DYPSA Algorithm for Identification of Glottal Closure Instants*", Proc. European Signal Processing Conference (2007)

Meyer, G.F., and Plante, F., Ainsworth, W.A.: "*A pitch extraction reference database*". In proceedings EUROSPEECH, vol. 1, pp. 837.840. 1995.

Noll, A. M.: "*Cepstrum pitch determination*" J. Acoust. Soc. Amer., vol. 41, pp. 293–309, Feb. 1967.

Rabiner, L.R.: "A tutorial on hidden Markov models and selected applications in speech recognition". *Proceedings of the IEEE*, 77(2): 257-286. 1989

Roa, S., Bennewitz, M., and Behnke, S.: "*Fundamental Frequency Estimation Based on Pitch-Scaled Harmonic Filtering*". In Proc. of the ICASSP'07*, pp.* IV-397-IV-400

Roebel A., Villavicencio, F. and Rodet, X.: "*Improving LPC Spectral Envelope extraction of Voiced Speech by True-Envelope Estimation*". in Proc. of the ICASSP'06, France, 2006.

Ross M.J., Shaffer, H.L., Cohen, A., Freudberg, R., and Manley,H.J.: "*Average Magnitude Difference Function Pitch Extractor*" IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 22, no. 5, pp. 353362, 1974.

Yoma, N.B, Busso, C and Soto, I.: "*Packet-loss modelling in IP networks with state-duration constraints*", IEE Proc.-Commun., Vol. 152, No. 1, February 2005

Yoma, N.B., McInnes, F., Jack, M., Stump, S., and Ling, L.: "*On including temporal constraints in Viterbi alignment for speech recognition in noise*", IEEE Trans. Speech Audio Process., 2001, 9, (2), pp. 179–182

Zeng, Yu-Min, Wu, Zhen-Yang, Liu, Hai-Bin, Zhou, Lin.: "*Modified AMDF pitch detection algorithm*" Machine Learning and Cybernetics, 2003 International Conference on Volume 1, 2-5 Nov. 2003 Page(s):470 - 473