UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE INGENIERÍA

ANÁLISIS DE CODIFICACIONES PERCEPTUALES DE LA VOZ Y SU COMPARACIÓN EN EL RECONOCIMIENTO DE COMANDOS

T E S I S

QUE PARA OBTENER EL TÍTULO DE
INGENIERO ELÉCTRICO ELECTRÓNICO
P R E S E N T A :
ALAN EDGAR BALLADO DE LA RIVERA

DIRECTOR:
DR. JOSÉ ABEL HERRERA CAMACHO

Ciudad Universitaria
México, D.F. 2012

JURADO ASIGNADO:

Presidente: Dr. Francisco Javier García Ugalde
Vocal: Dr. José Abel Herrera Camacho
Secretaria: Ing. Gloria Mata Hernández
1er Suplente: Dr. Pablo Roberto Pérez Alcázar
2º Suplente: Dr. Bohumil Psenicka

Facultad de Ingeniería, Ciudad Universitaria.

TUTOR DE TESIS
Dr. José Abel Herrera Camacho

_____

FIRMA

ÍNDICE

CAPÍTULO 8
PRUEBAS CON EL SIMULADOR DE RECONOCIMIENTO DE VOZ

APÉNDICES

BIBLIOGRAFÍA

ÍNDICE DE FIGURAS

CAPÍTULO 2 SONIDOS Y VOZ

CAPÍTULO 3 PROCESAMIENTO / ANÁLISIS

CAPÍTULO 4 LPC

CAPÍTULO 5 MFCC

CAPÍTULO 6 PLP

CAPÍTULO 7 VQ

CAPÍTULO 8 PRUEBAS / RECONOCIMIENTO

ÍNDICE DE TABLAS

CAPÍTULO 2 SONIDOS Y VOZ

CAPÍTULO 3 PROCESAMIENTO / ANÁLISIS

CAPÍTULO 4 LPC

CAPÍTULO 5 MFCC

CAPÍTULO 6 PLP

CAPÍTULO 8 PRUEBAS / RECONOCIMIENTO

CAPÍTULO

1

PREFACIO

1.1 INTRODUCCIÓN

El lenguaje es la capacidad básicamente humana de asociar significado a determinados sonidos, con los cuales el hombre elabora, expresa y comunica sus pensamientos. Como se desprende de esta definición, el lenguaje es el elemento clave para la vida social e intelectual del hombre, el desarrollo de la ciencia y la cultura, puesto que cualquier tipo de conocimiento se encuentra coexistiendo en él.

Ahora bien, hay que hacer la indispensable distinción entre lengua y lenguaje, pues cabe precisar que lengua es el sistema de signos concretos con el que una comunidad pone en práctica su capacidad de lenguaje. Se emplea la expresión "lengua natural" para definir a todo código de signos fonéticos utilizado por un grupo social para su comunicación cotidiana, para distinguirla de cualquier otro tipo de sistema de signos no verbales, como por ejemplo, aquellos utilizados en las matemáticas. Por otro lado, los lingüistas oponen la lengua al habla, ya que la lengua es un sistema gramatical virtualmente existente en las mentes de todos los individuos de una misma comunidad lingüística, mientras que el habla es algo particular, la peculiaridad individualizada de la lengua.

Efectivamente, la extraordinaria habilidad o capacidad del hombre para comunicarse mediante el lenguaje lo distingue de todas las demás criaturas terrenales y con frecuencia se toma como una prueba de su espiritualidad.

Siendo esta la forma más natural de comunicación para los seres humanos, el habla y la voz han sido objetos de mucha atención e interés, aún desde la antigüedad. La estructura del habla, su producción y los mecanismos para su percepción han mantenido ocupados por largo tiempo a los lingüistas, a los fisiólogos y a los psicólogos.

Por otro lado, los inventores, los científicos y los ingenieros han emprendido la tarea de producir dispositivos que puedan ser capaces de reproducir, reconocer y comprender la voz humana. Ahora bien, esta meta ha resultado ser mucho más difícil de conseguir de lo que se esperaba. Sin embargo, no hace muchos años atrás, a finales del siglo XX, se logró comenzar a materializar parcialmente esta meta, pues todavía queda un largo camino por recorrer, porque a estos sistemas aún les falta mucho para poder emular estas capacidades naturales del ser humano.

Estos sistemas de reconocimiento de voz, son capaces de reconocer palabras con una exactitud razonable, pero con la restricción de un vocabulario limitado. Por lo general, su desempeño se degrada significativamente cuando se les demanda manejar un vocabulario extenso, o si se pretende manejar una amplia variedad de locutores. Pese a estas dificultades, se ha estado trabajando para mejorar el desempeño de estos sistemas y paulatinamente se ha ido consiguiendo esto, de tal manera que actualmente existen varias aplicaciones comerciales del procesamiento de voz que ya se están utilizando.

La conversión analógico-digital de la señal de voz se suele denominar como codificación de voz, o simplemente codificación. Sin embargo, la codificación también tiene que ver con el desarrollo de las técnicas de análisis que aprovechan las características de la señal de voz, de tal manera que se pueda representar eficientemente mediante una cantidad reducida de parámetros. Esto es importante cuando se tiene en cuenta el espacio de memoria que requiere el almacenamiento de los archivos de voz en los medios electrónicos digitales, o también para mejorar la eficiencia en el uso del ancho de banda disponible, para la transmisión de una señal por algún medio de telecomunicaciones, como sucede por ejemplo, en los casos de los buzones de voz y en los canales disponibles para los teléfonos inalámbricos. Además, estas técnicas de codificación se pueden emplear para extraer las características o la información acústica de la señal de voz para su representación, por lo que constituyen los componentes fundamentales de los sistemas de reconocimiento de voz.

1.2 OBJETIVOS

Principalmente, el presente es un trabajo de investigación sobre tres técnicas de codificación orientadas al reconocimiento de voz: la codificación de predicción lineal (LPC), la codificación o análisis mel cepstral (MFCC) y la codificación de predicción lineal perceptual (PLP). Mediante este trabajo, se espera contribuir con un documento que pueda servir de apoyo para los alumnos interesados en el área del procesamiento digital de voz.

Como una de las características de la ingeniería es la aplicación de los conocimientos, también se proponen algunas pruebas o experimentos mediante la simulación de un sistema de reconocimiento de voz, basado en la técnica de cuantización vectorial multiseccionada (MSVQ), con la finalidad de examinar las propiedades, características o cualidades de estas técnicas de codificación, y al comparar los resultados obtenidos, determinar cuales técnicas proporcionaron la mayor cantidad de palabras reconocidas.

Se asume que para el reconocimiento de voz, las codificaciones perceptuales MFCC y PLP, representan mejor la información acústica de la señal de voz en comparación con otras técnicas de análisis que no toman en cuenta algunos aspectos de la percepción de los sonidos, como es el caso de la codificación LPC. Entonces, se esperaría conseguir la mayor cantidad de palabras reconocidas con PLP, pues esta técnica propone ciertos procesos adicionales, luego con MFCC y por último con LPC.

1.3 DEFINICIÓN DEL PROBLEMA

La compresión de la señal de voz es uno de los objetivos principales de la codificación, esto es, el emplear tan pocos parámetros como sea posible en la representación de una señal. Durante los últimos cuarenta o cincuenta años se han propuesto, analizado y desarrollado muchas técnicas de codificación de voz. De manera general, estas se pueden clasificar en dos categorías: como codificadores de la forma de onda (wave coders) y como codificadores de voz (vocoders: voice coders).

En los codificadores de la forma de onda se busca codificar directamente la forma de onda de la señal de voz aprovechando sus características. En contraste, los codificadores de voz tienen que ver con su representación mediante un conjunto reducido de parámetros, que además proporcionan una forma de análisis preliminar. La estimación de éstos se realiza a partir de segmentos cortos o tramas de la señal de voz.

Sin embargo, uno de los grandes retos a los que hay que enfrentarse en los sistemas de reconocimiento es la variabilidad inherente de la voz. La técnica de codificación más conocida y mejor desarrollada es la de predicción lineal (LPC). Aunque esta técnica puede representar muy bien la señal de voz y se puede utilizar en los sistemas de reconocimiento tiene sus limitaciones.

En general, se asume que entre mejor se comprenda el procesamiento de las señales de voz en el sistema auditivo humano, también se estará más cerca de diseñar un sistema que en verdad pueda entender tanto el significado como el contenido de la señal de voz. Por esta razón, muchos investigadores han sugerido otras técnicas motivadas en la percepción de los sonidos, como por ejemplo, la codificación o análisis mel cepstral (MFCC) y la codificación de predicción lineal perceptual (PLP). Estas codificaciones perceptuales tratan de imitar algunas propiedades conocidas del sistema auditivo humano, con el objetivo de representar mejor las características de la señal de voz, con lo que se busca obtener una mejora evidente en el desempeño de los sistemas de reconocimiento.

1.4 ALCANCE

A partir de 1960, las computadoras digitales y el desarrollo del procesamiento digital de voz hicieron posible el avance en las técnicas de análisis y codificación, las cuales constituyen una parte fundamental en cualquier aplicación de voz. De tal manera que aunque se han propuesto muchos métodos de análisis, en esta tesis solamente se abarcan las siguientes tres técnicas: la codificación de predicción lineal (LPC), la codificación o análisis mel cepstral (MFCC) y la codificación de predicción lineal perceptual (PLP). No se pretende la demostración ni la optimización de estas técnicas, más bien, se da más énfasis a la obtención de resultados prácticos, esto es, a la obtención de los vectores de coeficientes, así como la obtención de los procedimientos que permitan realizar los programas para su cálculo, mediante cualquier software o lenguaje de programación.

Puesto que primordialmente, ésta es una tesis sobre codificación, no se trata de implementar en forma un sistema de reconocimiento de voz, como los que ya existen actualmente, de otra forma, éste tendría que ser un sistema de reconocimiento de voz continua, para el cual intervienen otras variables exógenas. Sin embargo, se emplea la simulación de un sistema de reconocimiento, solamente como un medio o herramienta para examinar las características de estas tres técnicas de codificación, en pocas palabras, lo que se busca es una forma directa para implementar las codificaciones, con la que se esperan obtener diferencias evidentes en los resultados.

Por lo tanto, por simplicidad, en la simulación del sistema de reconocimiento de voz, se empleará la técnica de cuantización vectorial multiseccionada (MSVQ), para el reconocimiento de comandos o palabras aisladas, recortadas manualmente (detección manual del inicio y fin de las palabras), con un vocabulario reducido de doce palabras o clases, de doce locutores, seis hombres y seis mujeres de diferentes edades.

Debido a la gran cantidad de variables involucradas en la simulación del sistema de reconocimiento de voz, se propone mantener la mayoría de los parámetros fijos y modificar solamente uno de ellos, esto es, la cantidad o el número de coeficientes del vector de características, para así poder comparar la variación en el reconocimiento, es decir, la cantidad de aciertos o palabras reconocidas en función del número de coeficientes para cada una de las tres codificaciones.

Entonces, se propone utilizar básicamente las mismas condiciones de prueba para cada uno de los experimentos, siendo la modificación de la técnica de codificación la única diferencia entre las pruebas, obteniendo así, tres versiones diferentes para la prueba con el simulador de reconocimiento de voz, que corresponden a cada una de las tres codificaciones.

Al comparar los resultados obtenidos en estas tres pruebas, se espera contar con los elementos necesarios y suficientes para evaluar si se obtuvieron los resultados esperados y por consiguiente, sacar conclusiones.

El contenido de los capítulos que componen está tesis se describe a continuación:
Capítulo 1. Prefacio. Consiste en la presentación y descripción general de la tesis. Contiene algunos antecedentes del origen y el propósito del procesamiento de voz.

Capítulo 2. Los sonidos y la señal de voz. Aborda algunos conceptos básicos relacionados con la voz, tales como su producción, clasificación, representación, así como algunas de las características que dificultan su reconocimiento.

Capítulo 3. Procesamiento o análisis de la señal de voz. Trata acerca de los procesos comunes que se emplean en las diferentes técnicas de análisis o codificación, así como su aplicación en la elaboración de espectrogramas.

Capítulo 4. LPC. Codificación de predicción lineal. La descripción y análisis de esta técnica de codificación y un procedimiento para la obtención de los coeficientes LPC.

Capítulo 5. MFCC. Codificación o análisis mel cepstral. El análisis de esta técnica de codificación y un procedimiento para la obtención de los coeficientes MFCC.

Capítulo 6. PLP. Codificación de predicción lineal perceptual. El análisis de esta técnica de codificación y un procedimiento para la obtención de los coeficientes PLP.

Capítulo 7. VQ. Cuantización vectorial. La descripción de las técnicas de cuantización vectorial VQ y de cuantización vectorial multiseccionada MSVQ.

Capítulo 8. Pruebas con el simulador de reconocimiento de voz. Consiste en una introducción al reconocimiento de voz e incluye una descripción y análisis de las pruebas propuestas para examinar las técnicas de codificación, la presentación de los resultados obtenidos así como de las conclusiones.

1.5 APORTACIONES

Este documento no se limita solamente a mostrar el cálculo de los vectores de coeficientes LPC, MFCC y PLP, sino que además presenta un análisis detallado que permite una mejor comprensión de como se realizan cada una de estas tres codificaciones, el cual difícilmente podría encontrarse en tal extensión en la literatura del procesamiento digital de voz (con la única excepción de LPC) y mucho menos en español.

Con la finalidad de facilitar la asimilación y la comprensión de los conceptos expuestos, se ha hecho un esfuerzo para presentar los temas en un orden lógico y progresivo, de una manera sencilla, con una notación coherente y consistente, además de incluir algunas figuras ilustrativas.

Estos argumentos, confirman la originalidad de este trabajo de investigación, ya que este no se trata simplemente de una copia ni tampoco de una mera traducción de otros libros o artículos sobre codificación de voz.

En cuanto a las pruebas o experimentos propuestos, estos también muestran una manera singular para examinar, comparar y evaluar las propiedades, características o cualidades de las técnicas de codificación. Evidentemente, estas pruebas o experimentos se pueden plantear y analizar de muchas otras formas distintas, esto sin tomar en cuenta las posibles mejoras o modificaciones que también se les pueden incluir.

De tal manera que esta tesis, no solamente puede servir de apoyo en la formación de más alumnos, sino que también puede servir como una base para examinar minuciosamente las diferentes codificaciones de voz, así como para ampliar y/o profundizar en las ideas y conceptos del procesamiento digital de voz. De modo que esta tesis también puede aportar algunas ideas que resulten útiles para los investigadores.

Aunque el reconocimiento de comandos o palabras aisladas mediante la cuantización vectorial multiseccionada no sea algo novedoso, este todavía se puede utilizar como un medio o herramienta directa para implementar las codificaciones de voz. Además, si aun con este modelo se pueden obtener buenos resultados, entonces con un modelo o sistema más complejo y robusto, se deben seguir obteniendo buenos resultados o incluso mejores.

CAPÍTULO

2

LOS SONIDOS Y LA SEÑAL DE VOZ

2.1 INTRODUCCIÓN

La lengua es uno de los órganos más activos al hablar, aunque no es el único implicado. En este proceso también intervienen los labios, los músculos faciales y los de la garganta. Efectivamente, en la producción de la voz y en su percepción intervienen muchos órganos y partes del cuerpo, como por ejemplo: el cerebro, los oídos, los pulmones, la boca, la nariz, etc. Ahora bien, hay que tomar en cuenta que estas partes del cuerpo no se emplean solamente para hablar, sino que además desempeñan otras funciones adicionales (comer, respirar, etc.).

Así que el procesamiento de voz depende enteramente de la investigación básica de las ciencias del habla y de la audición, algunas de las cuales, como la lingüística, la gramática y la fisiología ya son centenarias. Sin embargo, pocos ingenieros disponen del tiempo necesario para volverse expertos en estas disciplinas fundamentales, por lo tanto el procesamiento de voz continúa siendo un área multidisciplinaria. En cualquier caso, se necesita un conocimiento de los conceptos básicos de estas áreas para poder analizar y modelar la señal de voz.

2.2 PRODUCCIÓN DE LA VOZ

La producción de la voz se realiza cuando se exhala un flujo de aire desde los pulmones a través de la tráquea hacia la laringe, que se encuentra en la parte media de la garganta. Dentro de la laringe, hay dos pliegues musculares en lados opuestos, éstos son las cuerdas vocales. Al hablar, los músculos tensan las cuerdas vocales y estas vibran con el aire que empujan los pulmones a través de ellas, creándose así el sonido. Por lo tanto, este flujo de aire constituye la fuente de energía para la generación de la voz.

Cuanto más se tensan las cuerdas vocales, más rápida es su vibración y el sonido que producen es más agudo, pero cuanto más relajadas estén, el tono resultante es más grave. Al salir de la laringe, la onda sonora entra en la faringe, que es la parte superior de la garganta. De ahí pasa a la boca y a la cavidad nasal, en donde se le da resonancia, es decir, se le añaden los matices que la modifican. El paladar, la lengua, los dientes y los labios se combinan para conformar las ondas sonoras para producir cualquiera de los sonidos empleados al hablar. El sonido particular de la voz de cada persona está determinado por el tamaño y la forma de las cuerdas vocales, de la garganta, de la nariz, de la boca, etc. De tal manera que los sonidos se ven afectados por la forma del tracto vocal.

2.3 CLASIFICACIÓN Y REPRESENTACIÓN

El elemento más pequeño o mínimo, mediante el cual se denota un significado diferente para algún sonido, se denomina fonema. De manera inicial, los sonidos presentes en la voz se dividen en dos categorías básicas: vocales y consonantes, o lo que es lo mismo, en sonidos sonoros y sonidos sordos.

Si el flujo de aire proveniente de los pulmones causa la vibración de las cuerdas vocales, entonces se genera un sonido sonoro (voiced speech), este es el caso de las vocales. La tensión ejercida sobre las cuerdas vocales determina el tono fundamental (pitch). Pero si este flujo se vuelve turbulento, debido a una restricción significativa en algún punto del tracto vocal, entonces se genera un sonido sordo (unvoiced speech), este es el caso de las consonantes. Para estas últimas, el punto de constricción y la posición de los órganos de articulación de la voz determinan el sonido que es producido.

Los configuraciones del tracto vocal para los sonidos sordos generalmente tienen un mayor grado de constricción que para los sonidos sonoros. En algunos casos se pueden requerir de movimientos dinámicos precisos de los órganos de articulación de la voz para su producción, pero para otras no, puesto que su sonido es sostenido como el de las vocales.

Las consonantes se pueden clasificar de acuerdo a su modo de producción, así como por la presencia o la ausencia de fonación: la vibración u oscilación de las cuerdas vocales. Las fricativas ocurren al excitar el tracto vocal mediante un flujo de aire constante que se vuelve turbulento al frente de la boca, el punto de constricción.

Las plosivas suceden cuando las cuerdas vocales se cierran e interrumpen brevemente el flujo de aire, de modo que la presión del aire va aumentando hasta que se libera súbitamente. Las africativas se producen cuando dos consonantes, una plosiva y una fricativa, se acortan y se combinan. Las laterales suceden cuando el tracto vocal se cierra en el punto de articulación. Para las vibrantes, ocurre una abertura y cierre intermitente del punto de constricción, acompañado de una exhalación gradual que produce la turbulencia. Por último, las nasales se producen con la excitación de las cuerdas vocales con el tracto vocal totalmente cerrado en el punto de articulación, pero el velo se cierra para que el sonido sea irradiado desde la nariz. De modo que la cavidad oral actúa como un resonador alterno que atrapa la energía acústica a ciertas frecuencias, por lo que estas frecuencias se presentan como antiresonancias o ceros [12].

El español que se habla en México consta de 30 letras: 5 vocales y 25 consonantes. Los fonemas básicos asociados a estas letras se muestran en la tabla 2.1, así como su clasificación asignada de acuerdo a la manera en que se producen [8].

TABLA 2.1 Fonemas básicos del alfabeto español de México

Fonema	Clasificación	Letras / Casos
/i/	Vocal frontal	i , y (vocal)
/e/	Vocal frontal	e
/a/	Vocal central	a
/o/	Vocal trasera	o
/u/	Vocal trasera	u
/f/	Fricativa sonora	f
/j/	Fricativa sonora	y
/θ/	Fricativa sonora	z
/s/	Fricativa sorda	s ; ce, ci ; x ; z
/χ/	Fricativa sorda	j , ge, gi, x
/b/	Plosiva sonora	b , v
/d/	Plosiva sonora	d
/g/	Plosiva sonora	g, ga, gue, gui, go, gu ; w
/p/	Plosiva sorda	p
/t/	Plosiva sorda	t
/k/	Plosiva sorda	c, ca, co, cu ; k ; que, qui
/c/	Africativa sorda	ch
/l/	Semivocal lateral sonora	l
/ʎ/	Semivocal lateral sonora	ll
/r/	Semivocal vibrante sonora	r , rr
/m/	Nasal sonora	m
/n/	Nasal sonora	n
/ɳ/	Nasal sonora	ñ

La letra ‘h’ no tiene un fonema asociado porque se dice que es “muda”. Pero la letra ‘x’ puede leerse como una ‘s’ o como una ‘j’, pero también puede pronunciarse como la combinación de dos fonemas: /k/ y /s/.

Durante el habla normal o cotidiana, también tienen lugar algunas pausas o silencios, pero a diferencia de lo que podría esperarse, estas no ocurren entre las palabras, pues no existen límites claros o evidentes al final de una palabra y al inicio de la siguiente. Estos silencios ocurren con mayor frecuencia justo antes de las plosivas.

En general, los sonidos sonoros tienen una mayor amplitud que los sonidos sordos y además presentan un patrón regular o periódico, con algunos valores pico que corresponden a los instantes de máxima excitación. En cuanto a los sonidos sordos, éstos no presentan algún patrón periódico, más bien se componen por valores de poca amplitud, sin frecuencias dominantes, por lo que se parecen mucho al ruido blanco.

Sin embargo, estas características no siempre son tan fáciles de distinguir al observar la forma de onda de una señal de voz, por lo tanto, a veces resulta algo complicado distinguir las diferencias entre estos tres, especialmente si se trata de sonidos débiles.

Se dice que la señal de voz varía lentamente en el tiempo, en el sentido que, al examinarla sobre un breve periodo de tiempo (de 5 a 100 ms) sus características permanecen prácticamente estacionarias. No obstante, para periodos de tiempo más largos (de más de 200 ms) las características de la señal cambian para expresar los diferentes sonidos que se producen al hablar.

En la figura 2.1 se pueden apreciar algunas de las diferencias que existen entre los sonidos sonoros y los sonidos sordos. Aquí se muestra la forma de onda de la señal de voz para la palabra "*fénix*", dicha por un varón. Se puede observar un primer segmento para los sonidos sonoros, un silencio y otro segmento para los sonidos sordos, en donde la amplitud de la señal es menor y pareciera ser ruido. También se destacan otros dos segmentos de 60 ms, en donde se pueden observar mejor tanto el patrón periódico para un sonido sonoro como la semejanza a ruido para un sonido sordo.

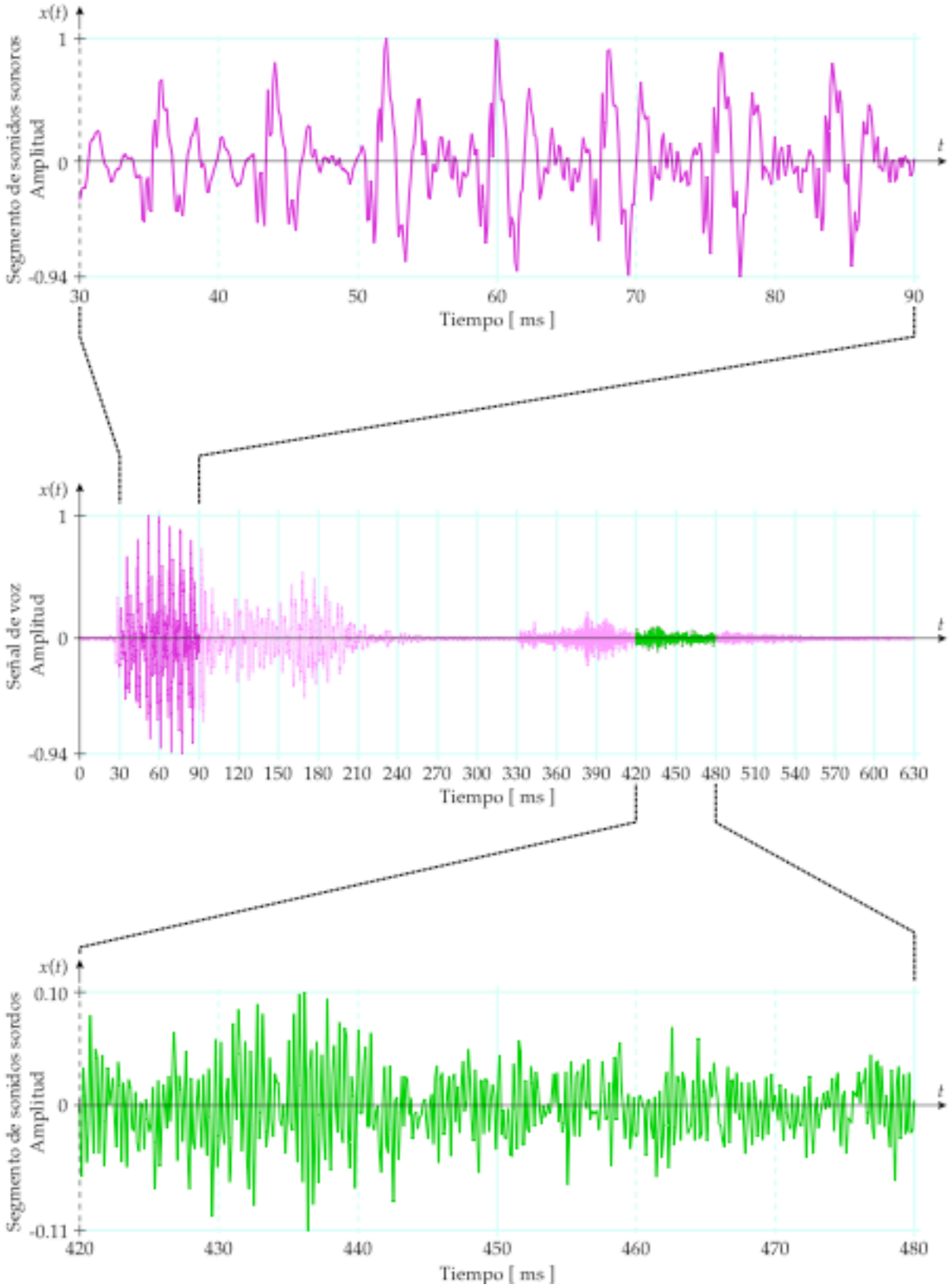


FIGURA 2.1 Sonidos sonoros y sonidos sordos en una señal de voz

2.4 EL MODELO FUENTE-FILTRO

Una forma bastante simplificada pero efectiva para comprender el mecanismo para la producción de la voz es mediante el modelo fuente-filtro, propuesto originalmente por Muller en 1848 [15]. En dicho modelo, se supone que la fuente o señal de excitación es linealmente separable de las características del tracto vocal, que se representan por medio de un filtro. Por lo tanto, la señal de voz es la salida de este filtro en respuesta a la fuente de excitación, la cual puede representarse como un generador de un tren de impulsos cuasi periódicos para los sonidos sonoros, o como un generador de ruido aleatorio (ruido blanco) para los sonidos sordos.

Puesto que el filtro debe tener las mismas características que el tracto vocal, no es posible mantener constantes los parámetros del filtro, los cuales tienen que ser modificados para que correspondan a las variaciones que suceden al hablar. Por lo tanto, el filtro tiene parámetros variantes en el tiempo. No obstante, en la práctica, esta tasa de cambio varía lentamente en el tiempo, por lo que los parámetros tienen que ser actualizados en intervalos de 5 a 30 ms. La figura 2.2 muestra un diagrama de bloques del modelo fuente-filtro. El modelo es relevante porque para cada tipo de excitación, un fonema se identifica principalmente por la forma del tracto vocal, por lo que su configuración puede ser estimada al identificar el filtrado realizado por este.

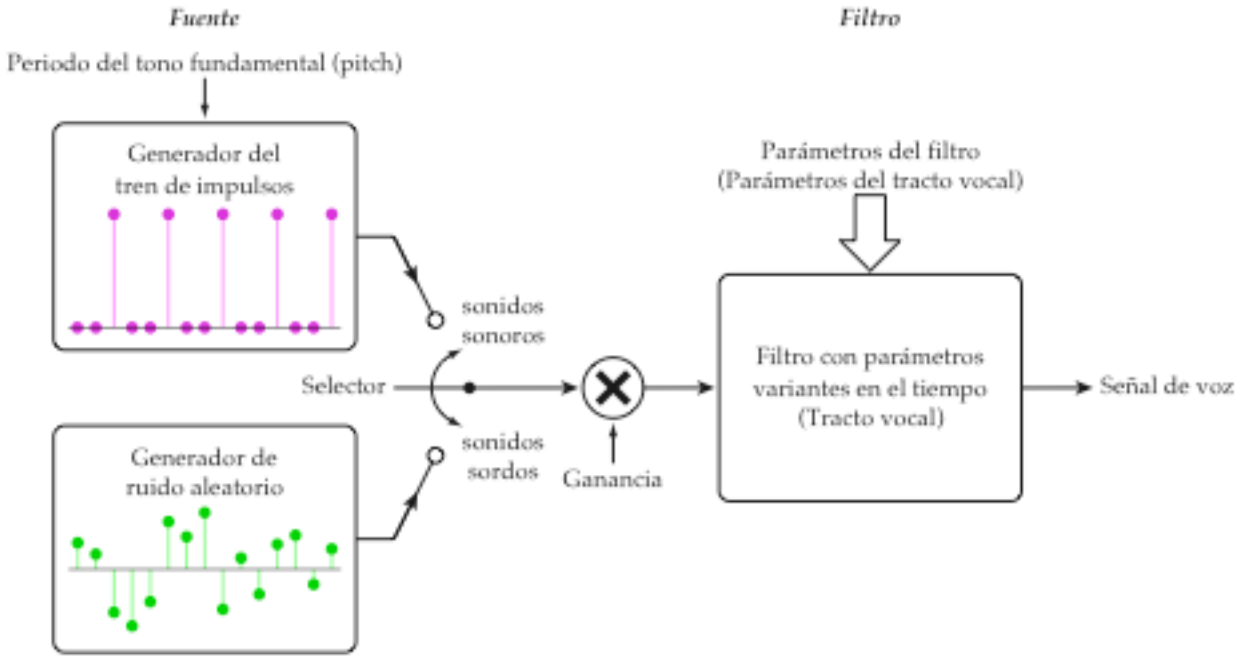


FIGURA 2.2 Modelo fuente-filtro para la producción de la voz

2.5 CARACTERÍSTICAS DE LA VOZ

De una manera simplista, la voz se puede considerar como una secuencia de sonidos tomados a partir de un conjunto de elementos básicos llamados fonemas. Los diferentes sonidos se producen al modificar la forma del tracto vocal y de los demás órganos de articulación de la voz. Sin embargo, esta no es una secuencia de sonidos discretos discernibles, sino más bien, consiste en una serie de sonidos que fluyen suavemente uno sobre el otro.

Además, al hablar normalmente se anticipa lo que se va a decir, así que al articular un fonema ya se está preparando para articular el siguiente. De tal manera que el sonido producido para un mismo fonema puede variar dependiendo de la posición en la que se encuentre dentro de una palabra. Estos movimientos por anticipado alteran el sonido del fonema que se está diciendo. Esta interacción entre sonidos adyacentes se conoce como co-articulación y es uno de los más grandes obstáculos a los que hay que enfrentarse en el reconocimiento de voz.

Otro problema significativo es el hecho de que en la señal de voz no hay limites claramente definidos entre las palabras ni mucho menos entre los sonidos. Aún la tarea relativamente simple de determinar el inicio y el final de una palabra, presenta sus dificultades y es propensa a errores, especialmente al operar bajo condiciones ruidosas.

Por otra parte, el sonido particular de la voz de una persona es muy diferente a la de cualquier otra, ya sea debido a sus antecedentes lingüísticos o simplemente debido a las diferencias físicas. Lo que es más, aun un mismo individuo produce versiones acústicamente diferentes de una ocasión a otra, aunque es posible obtener resultados muy parecidos, de ninguna manera serán idénticos [15].

También ocurren una amplía gama de variaciones acústicas cuando el mismo fonema es dicho por diferentes personas, debido a las diferencias de sus aparatos para la producción de la voz. Por ejemplo, las diferencias en la longitud de las cuerdas vocales y en la musculatura ocasionan la voz grave en los hombres y una voz más aguda en las mujeres. La frecuencia promedio del tono fundamental (pitch) es aproximadamente de 125 Hz para los hombres y de 200 Hz para las mujeres, aunque puede alcanzar 300 Hz para el caso de algunos niños. Las cuerdas vocales de los hombres tienen una longitud de 17 a 24 mm, mientras que las de las mujeres miden de 13 a 17 mm [15].

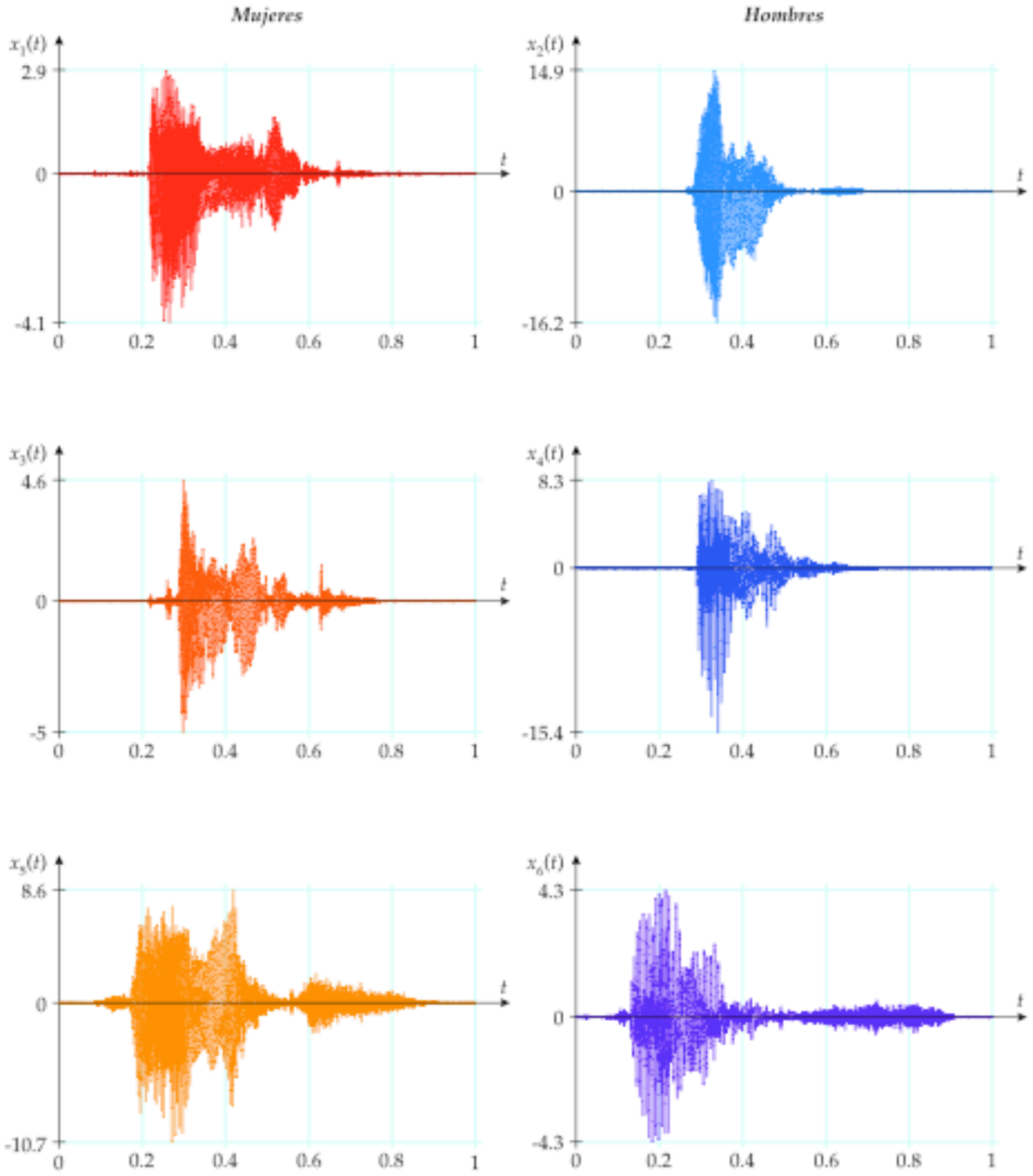
Por lo general, la señal de voz tiende a ser continua, sin límites obvios entre una palabra y la siguiente. Además, la velocidad con la que se habla, así como la duración de las palabras y de los sonidos también son muy variables. En la figura 2.3 se pueden observar algunas de las diferencias o variaciones que se presentan en las señales de voz. Se muestran seis ejemplos para la misma palabra *"fénix"*, pero dichas por seis locutores distintos, tres mujeres y tres hombres. Claramente las formas de onda, la amplitud de las señales y hasta su duración no son iguales, pues estas características manifiestan las diferencias entre las voces de los locutores, así como su manera de hablar, etc.

Otro problema adicional es que la señal de voz pudiera estar contaminada por el ruido del entorno, por las voces de otras personas o incluso por algunos sonidos que pudiera hacer la misma persona que está hablando, incluyendo disfluencias, muletillas, toser, etc.

Y por si esto fuera poco, también es posible que la voz sea ambigua, pues con frecuencia las palabras pueden tener varios significados, o las diferencias acústicas entre algunas palabras resulten ser muy pequeñas. Incluso puede suceder que las vocales y las consonantes no se pronuncien correctamente, puesto que hay personas que suelen omitir algunos sonidos o hasta sílabas completas.

Además, para controlar la producción de la voz, ésta se monitorea constantemente. Lo cual se realiza de muchas maneras diferentes, como por ejemplo, mediante el tacto, el oído, la sensación de movimiento de los órganos de articulación y por la realimentación central interna en el cerebro. Esta realimentación auditiva tiene el propósito de garantizar que se están produciendo los sonidos correctos y de que estos son de la intensidad correcta considerando el entorno. En esencia cada locución es única y es esta singularidad la que ocasiona que sea muy complicado el decodificarla.

Finalmente, en el proceso de la percepción de la voz también interviene la memoria. Pues la habilidad para entender las palabras del lenguaje oral se basa en un conocimiento previo de estas, es decir, en los recuerdos o ejemplos de haber oído esas mismas palabras. Por ejemplo, se puede comparar la habilidad para comprender la lengua materna con la habilidad para entender otro lenguaje distinto, con el cual no se está bien familiarizado, en este caso, solo se pueden entender algunas palabras ocasionalmente, pero se vuelve muy difícil si se trata de un lenguaje totalmente desconocido, en el que ninguna palabra resulta familiar.



Eje horizontal : Escala de tiempo [s]
Eje vertical : Escala de amplitud de la señal de voz (x 1000)

FIGURA 2.3 Ejemplos de una misma palabra dichos por diferentes locutores

CAPÍTULO

3

PROCESAMIENTO O ANÁLISIS DE LA SEÑAL DE VOZ

3.1 INTRODUCCIÓN

La finalidad del procesamiento o análisis de la señal de voz es el obtener una representación más útil o conveniente de la información contenida en esta. Como ya se ha mostrado en el capítulo anterior, la señal de voz es una composición compleja de sonidos, pero mediante un análisis, se puede mostrar que esta señal esta formada por la combinación de componentes a diferentes frecuencias, producidas por la excitación periódica o aperiódica de las resonancias en el tracto vocal, principalmente entre 80 y 8000 Hz [15].

Las técnicas de análisis se pueden clasificar en dos categorías: en el dominio del tiempo o en el dominio de la frecuencia. Los procedimientos en el dominio del tiempo, se denominan así, porque éstos tienen que ver directamente con la señal de voz. El atractivo de estas representaciones es la facilidad con la que se pueden implementar, no obstante, estas suministran un medio útil para estimar otras características importantes de la señal. En contraste, los métodos en el dominio de la frecuencia involucran, ya sea de manera explícita o implícita, alguna forma de representación del espectro en frecuencia. La mayoría de estas se basan en el modelo fuente-filtro para la producción de la voz. Así que esencialmente, el análisis es el proceso en el cual, a partir de una señal de voz, se estiman los parámetros variantes en el tiempo del filtro. De tal manera que su objetivo principal es el estimar la respuesta en frecuencia del filtro que representa al tracto vocal.

3.2 ANCHO DE BANDA Y MUESTREO

El ancho de banda para las señales de voz es de aproximadamente unos 16,000 Hz, pero la mayor parte de la energía se concentra en las frecuencias menores a 7000 Hz [1]. No obstante, un ancho de banda de 3500 Hz es suficiente para poder entender el mensaje contenido en la señal de voz.

El valor para la frecuencia de muestreo se determina mediante el teorema del muestreo de Nyquist, que establece que si la componente en frecuencia más alta presente en la señal de voz es f_{max}, entonces para poder reconstruir apropiadamente la señal a partir de las muestras obtenidas, la frecuencia de muestreo f_s debe satisfacer la siguiente condición:

$$f_s \geq 2f_{max} \qquad f_N = f_{max} \qquad f_s \geq 2f_N \tag{3.1}$$

En donde a f_N se le conoce como la frecuencia de Nyquist. Por lo tanto, el periodo de muestreo T_s es igual al inverso de la frecuencia de muestreo f_s:

$$T_s = \frac{1}{f_s} \tag{3.2}$$

Si se emplea una frecuencia de muestreo menor (menos muestras por segundo) que el doble de la frecuencia de Nyquist, entonces sucede un fenómeno conocido como aliasing, para el cual una señal senoidal a cierta frecuencia pudiera parecer otra señal de menor frecuencia, como si esta fuera un alias de la original. Si el aliasing ocurriera en una señal compleja, como es la señal de voz, se insertarían algunas componentes en frecuencia no deseadas que la distorsionarían.

Puesto que el ancho de banda de una señal de voz se reduce durante la conversión analógico-digital, para su procesamiento se suelen utilizar básicamente dos frecuencias de muestreo, la primera de $f_s = 16{,}000\,\text{Hz}$ para obtener un ancho de banda o una frecuencia de Nyquist de $f_N = 8000\,\text{Hz}$, y la segunda frecuencia de muestreo es de $f_s = 8000\,\text{Hz}$ que corresponde a un ancho de banda o una frecuencia de Nyquist de $f_N = 4000\,\text{Hz}$.

Si se descartan todas las frecuencias mayores de los 16,000 Hz, entonces se dice que la señal de voz es ortofónica (orthophonic). Pero si esta solamente contiene las frecuencias dentro de la banda entre los 300 y 3400 Hz, entonces se dice que la señal de voz es telefónica, la cual es de menor calidad [1].

En cualquier caso, el muestreo periódico de una señal analógica de voz $x(t)$, proporciona una representación de esa señal mediante una secuencia de muestras $x(n)$, que se pueden obtener mediante la siguiente relación:

$$x(n) = x(t)\big|_{t=nT_s} \qquad n = 0,1,2,3,\ldots,N_x - 1 \qquad \textbf{(3.3)}$$

En donde N_x es la cantidad o el número total de muestras presentes en la señal de voz. La figura 3.1 ilustra el muestreo de una señal analógica de voz a una frecuencia de muestreo de $f_s = 8000$ Hz, así que cada una de las $N_x = 21$ muestras se toman cada $T_s = 0.125$ ms.

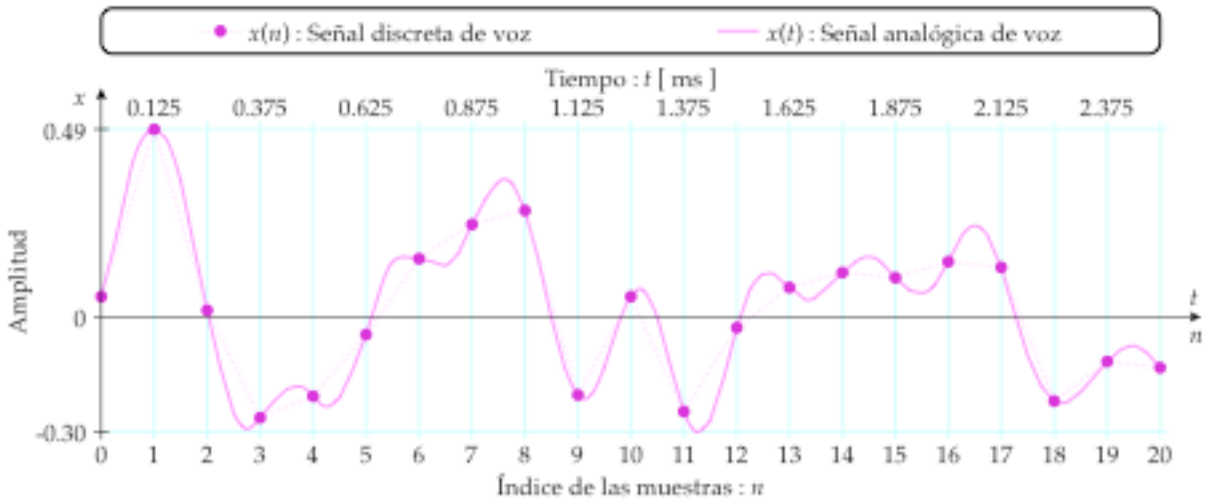


FIGURA 3.1 Muestreo de una señal de voz

La secuencia o la señal discreta de voz $x(n)$ esta indizada por la variable entera n, que de hecho proporciona una normalización de la señal, es decir, que ésta no contiene ninguna información explícita acerca de la frecuencia de muestreo. Aunque las muestras definen una señal discreta, se acostumbra unir éstas para que formen una curva continua para representar la señal de voz.

3.3 PRE-ÉNFASIS

Las componentes de alta frecuencia presentes en la señal de voz, normalmente tienen una amplitud menor con respecto a las componentes de baja frecuencia, debido a la atenuación que ocurre durante el mecanismo de producción de la voz, por lo que se necesita realizar un pre-énfasis de las componentes de alta frecuencia, a fin de obtener una amplitud similar para todas las componentes, de tal manera que el espectro de la señal sea lo más plano posible. Esto se puede conseguir mediante el filtrado de la señal de voz digitalizada $x(n)$ mediante un filtro FIR paso altas de primer orden, definido por la siguiente ecuación en diferencias:

$$x_p(n) = x(n) - ax(n-1) \qquad \begin{matrix} 0.9 \le a \le 1 \\ n = 0,1,2,3,\ldots,N_x - 1 \end{matrix} \qquad \textbf{(3.4)}$$

Donde $x_p(n)$ es la respuesta del filtro o la señal de voz con pre-énfasis y a es el parámetro de pre-énfasis. Si se obtienen las transformadas z, la función de transferencia del sistema es:

$$H(z) = 1 - az^{-1} \qquad 0.9 \le a \le 1 \qquad \textbf{(3.5)}$$

El valor más común para el parámetro de pre-énfasis es $a = 0.95$, el cual proporciona un incremento de unos 20 dB de amplificación para las componentes de alta frecuencia del espectro [1]. No obstante, el valor exacto de este parámetro no es tan crítico [14]. La figura 3.2 muestra la magnitud de la respuesta en frecuencia de un filtro de pre-énfasis para una frecuencia de muestreo de $f_s = 8000$ Hz y un parámetro de pre-énfasis $a = 0.95$.

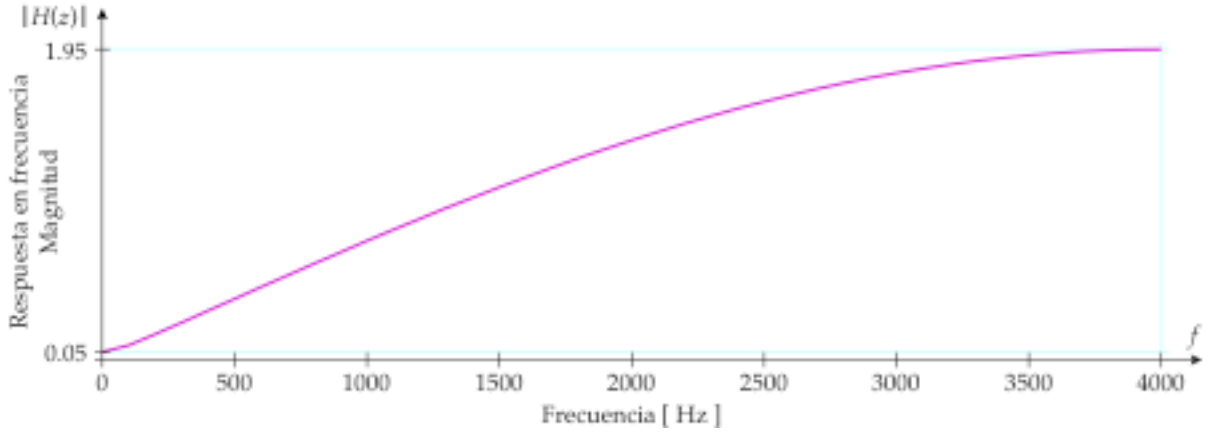


FIGURA 3.2 Magnitud de la respuesta en frecuencia del filtro de pre-énfasis

3.4 DIVISIÓN EN TRAMAS

Los métodos tradicionales de análisis en frecuencia son bastante confiables para las señales estacionarias, pero este no es el caso de la señal de voz. Entonces, para poder emplear estas técnicas, resulta necesario dividir la señal en segmentos cortos o tramas para poder realizar un análisis en tiempo-corto. Estas tramas pueden ser separadas y procesadas de manera individual, como si fueran segmentos cortos de un sonido sostenido con propiedades fijas. Este proceso se repite tantas veces como sea necesario hasta que se haya abarcado completamente la señal.

Estos segmentos o tramas de análisis, son de longitud finita, consisten de N muestras y normalmente se traslapan entre si, teniendo S muestras de desplazamiento o separación entre el inicio de una trama y la siguiente:

$$x_\ell(n) = x(n + \ell \cdot S) \qquad \begin{matrix} \ell = 0,1,2,3,\ldots,L-1 \\ n = 0,1,2,3,\ldots,N-1 \end{matrix} \qquad \textbf{(3.6)}$$

En donde $x_\ell(n)$ es la ℓ-ésima trama y L es la cantidad o el número total de tramas presentes en la señal de voz $x(n)$. La figura 3.3 ilustra la división en tramas con traslape de una señal de voz, para una separación entre tramas de $S = N/3$ muestras.

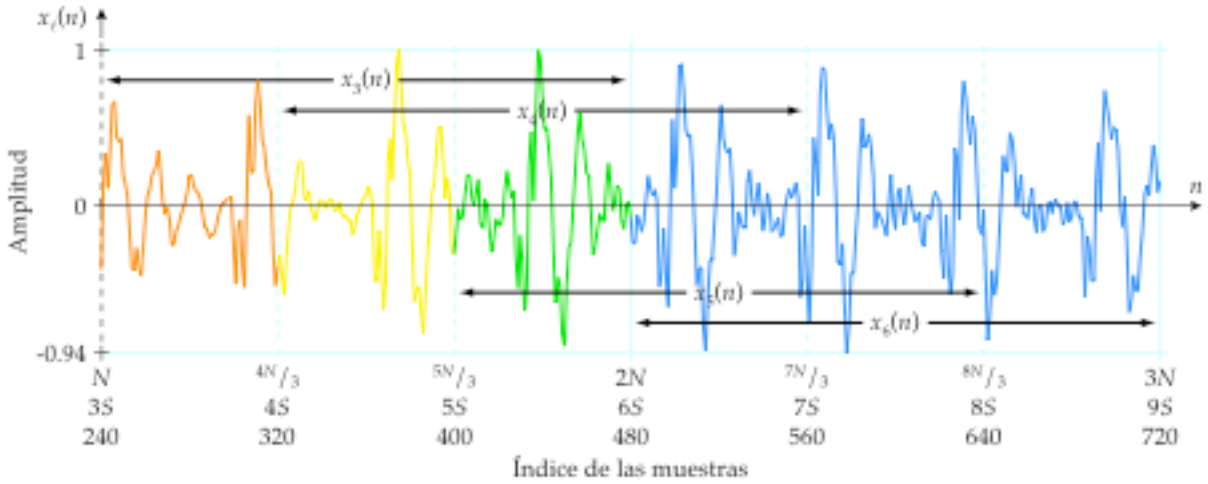


FIGURA 3.3 División en tramas con traslape de una señal de voz

A partir de la figura anterior, se puede observar que la cantidad o el número total de tramas en la señal de voz L, se puede calcular mediante la siguiente ecuación:

$$L = \frac{N_x - N}{S} + 1 \qquad \textbf{(3.7)}$$

Si $N > S$, entonces sí ocurre el traslape y $N - S$ muestras al final de las tramas son duplicadas al inicio de la siguiente trama. Entre mayor sea el traslape, la correlación entre las tramas adyacentes también será mayor y los cambios serán más suaves. Por otra parte, si $N < S$, entonces no hay traslape, pero además algunas porciones de la señal de voz se perderán, ya que no aparecerán en ninguna de las tramas. Por lo tanto, esta última situación resulta inaceptable en cualquier tipo de procesamiento de voz [14].

Entonces, la señal de voz se divide en L tramas para su análisis. Cada una de éstas se compone por N muestras con S muestras de separación entre tramas adyacentes. Sin embargo, para no perder las propiedades dinámicas o variantes en el tiempo de la señal de voz, es necesario volver a calcular otros coeficientes para la siguiente trama.

3.5 PONDERADO POR LA VENTANA DE HAMMING

La más simple de las ventanas tiene una forma rectangular y se define como:

$$w(n) = \begin{cases} 0 & \text{para} \quad n < 0 \\ 1 & \text{para} \quad 0 \le n \le N-1 \\ 0 & \text{para} \quad n > N-1 \end{cases} \qquad \textbf{(3.8)}$$

Esta ventana rectangular se usa de manera implícita cuando se extrae una trama de N muestras a partir de una señal de voz:

$$x_\ell(m) = x(n+m) \cdot w(m) \qquad m = 0,1,2,3,\ldots,N-1 \qquad \textbf{(3.9)}$$

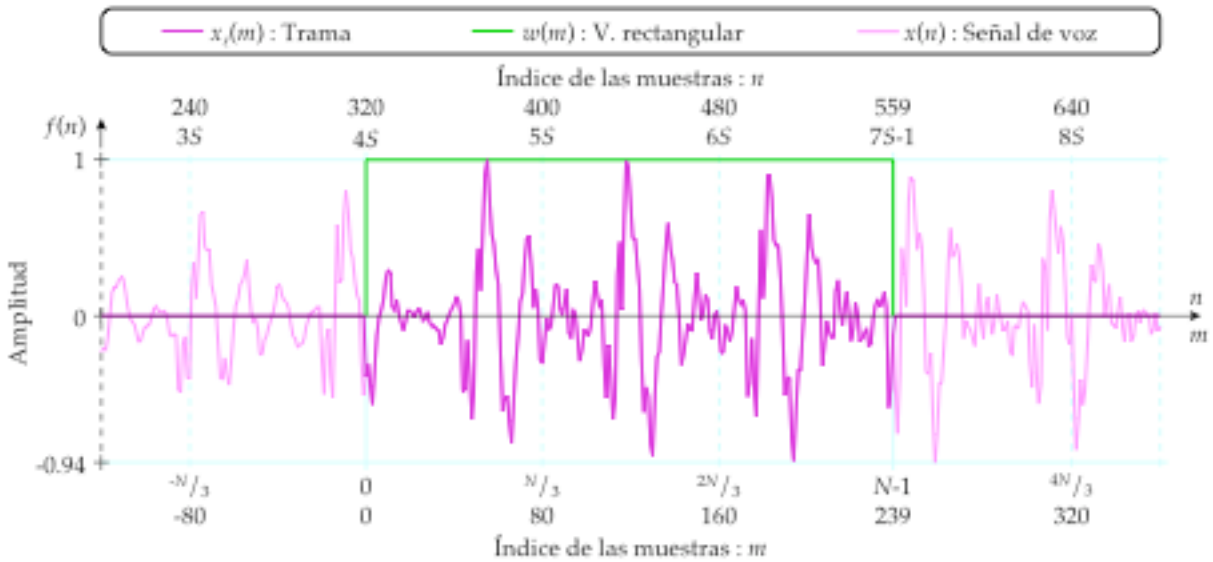


FIGURA 3.4 Extracción de una trama mediante una ventana rectangular

Sin embargo, la presencia de ésta ventana ocasiona la distorsión del espectro de la señal

de voz, pero si se pondera la trama por una ventana con la forma adecuada, entonces se puede reducir la distorsión en la frecuencia, aunque se deforme la señal en el tiempo. En el reconocimiento de voz, la ventana que más se utiliza es la de Hamming:

$$w(n) = \begin{cases} 0 & \text{para} \quad n < 0 \\ 0.54 - 0.46\cos\left(\dfrac{2\pi n}{N-1}\right) & \text{para} \quad 0 \le n \le N-1 \\ 0 & \text{para} \quad n > N-1 \end{cases} \qquad \textbf{(3.10)}$$

FIGURA 3.5 Ventana de Hamming

La figura 3.5 es la representación gráfica de la ventana de Hamming, la cual tiene valores que decaen suavemente hacia cero, aunque en los extremos, tiene una transición abrupta para el primer y el último valor de la ventana, los que son iguales a 0.08.

El segmento o la trama ponderada $x_w(n)$ se obtiene mediante el producto de cada una de las muestras de la trama de análisis $x_\ell(n)$ por el peso correspondiente de la ventana de Hamming $w(n)$:

$$x_w(n) = x_\ell(n) \cdot w(n)$$

$$x_w(n) = x_\ell(n) \cdot \left[0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)\right] \qquad n = 0,1,2,3,\ldots,N-1 \qquad \textbf{(3.11)}$$

El ponderado de la trama de análisis contribuye a minimizar las discontinuidades al inicio y al final de las tramas, al atenuar los valores en los extremos. De otro modo la disparidad entre la primera y la última muestra de la trama ocasionaría algunos efectos indeseables en los valores del espectro en frecuencia [15]. No obstante, si las muestras que se encuentran cerca de los extremos de la trama describen algún evento significativo de corta duración, entonces este evento prácticamente no afecta a la trama, dado que recibe un ponderado bajo. Así que normalmente, las tramas se traslapan para asegurar que todos estos eventos se tomen en cuenta al ser cubiertos en las tramas adyacentes.

3.6 LA LONGITUD DE LA TRAMA

La selección de la longitud de la trama N es una consideración muy importante, pues por lo general, el trabajo de computo requerido durante el análisis es proporcional a esta, por lo que resulta ventajoso mantenerla tan pequeña como sea posible. Sin embargo, se deben abarcar varios periodos del tono fundamental (pitch) para asegurar que los resultados sean confiables. Además, debido al ponderado, la longitud de la trama debe ser lo suficientemente larga como para que los efectos de atenuación de la ventana de Hamming no afecten seriamente los resultados.

Adicionalmente, la longitud o la duración de la trama se determina mediante un compromiso entre las resoluciones en el tiempo y en la frecuencia, ya que la longitud de la trama es proporcional a la resolución en frecuencia, pero inversamente proporcional a la resolución en el tiempo. De manera similar, el traslape es proporcional al número de tramas en la señal de voz, pero también es proporcional a la correlación de las tramas subsecuentes. Por lo tanto, se debe buscar un balance o compromiso entre estas características contrapuestas.

Una manera para determinar la longitud de la trama N, es haciendo el producto de la duración de la trama T por la frecuencia de muestreo f_s de la señal de voz:

$$N = \left(T \cdot f_s\right)_{\mathbf{int}} \qquad \textbf{(3.12)}$$

No obstante, como el número de muestras por trama N debe ser un valor entero, se debe redondear o truncar el resultado. El número de muestras de separación entre tramas adyacentes S, generalmente se selecciona de tal manera que se obtenga una tasa de 100 tramas por segundo ($Fps = 100$), es decir una trama cada 10 ms.

$$Fps = \frac{f_s}{S} \qquad \Rightarrow \qquad S = \frac{f_s}{100} \qquad \textbf{(3.13)}$$

Los valores típicos para la duración de las tramas son de 10 a 30 ms. Las ventanas cortas han sido propuestas para estimar los parámetros del tracto vocal que varían rápidamente, mientras que las ventanas largas se usan para estimar la frecuencia del tono fundamental (pitch). Las tramas de 20 a 30 ms de duración generalmente tienen una buena relación entre ambas resoluciones [1].

3.7 TRANSFORMADA DISCRETA DE FOURIER

El espectro en frecuencia de una señal discreta se puede estimar mediante la transformada discreta de Fourier (DFT):

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j\frac{2\pi nk}{N}} \qquad \textbf{(3.14)}$$

Esta transformada proporciona un conjunto de N valores en el dominio de la frecuencia, en donde $X(k)$ es la transformada discreta de Fourier en el dominio de la frecuencia, de la señal discreta $x(n)$ en el dominio del tiempo. Sin embargo, en la práctica, la transformada discreta de Fourier se calcula de una manera mucho más eficiente mediante la transformada rápida de Fourier (FFT).

3.8 RELLENADO CON CEROS

Para facilitar el uso de la transformada rápida de Fourier, la longitud de la trama N debería de ser igual al número de puntos para la transformada rápida de Fourier N_{fft}, un número que debe ser igual a la potencia de dos más próxima:

$$N_{fft} = 2^{\left[\log_2(N)\right]_{\text{int}}} \qquad \textbf{(3.15)}$$

Pero si no se cumple con esta condición, entonces se puede rellenar con ceros, es decir, añadir las muestras nulas que hagan falta.

$$x_{zp}(n) = \begin{cases} x_w(n) & \text{para} \quad n = 0,1,2,3,\ldots,N-1 \\ 0 & \text{para} \quad n = N, N+1,\ldots,N_{fft}-1 \end{cases} \qquad \textbf{(3.16)}$$

A partir de la definición de la transformada discreta de Fourier:

$$X(k) = \sum_{n=0}^{N_{fft}-1} x(n)\, e^{-j\frac{2\pi nk}{N_{fft}}} \qquad \textbf{(3.17)}$$

Se puede apreciar que el agregar valores nulos a $x(n)$ no afecta significativamente la sumatoria para cualquier $X(k)$ [15]. Sin embargo, la longitud aumentada de la trama N_{fft} solamente proporciona una mejor descripción de la transformada discreta de Fourier, pues no incrementa la resolución en frecuencia, pues este es el único propósito de la longitud de la trama N [9].

3.9 OBTENCIÓN DEL ESPECTRO EN FRECUENCIA

El espectro en frecuencia de una trama de análisis se obtiene mediante el algoritmo de la transformada rápida de Fourier (FFT):

$$X(k) = \mathbf{FFT}\left\{x_{zp}(n), N_{fft}\right\} \qquad \begin{matrix} k = 0,1,2,3,\ldots,N_{fft}-1 \\ n = 0,1,2,3,\ldots,N_{fft}-1 \end{matrix} \qquad \textbf{(3.18)}$$

Dado que la señal de voz $x(n)$ es una secuencia de números reales, $X(k)$ y $X(N-k)$ son conjugados complejos. Puesto que la magnitud de los conjugados complejos es igual, el espectro en frecuencia resultante tiene una simetría par con respecto a la mitad de la frecuencia de muestreo, es decir, que la magnitud de la mitad superior del espectro es un reflejo de la mitad inferior.

3.10 OBTENCIÓN DEL ESPECTRO DE POTENCIA

Ahora bien, no importa tanto la respuesta en frecuencia, sino la envolvente de la respuesta en frecuencia. Las características del tracto vocal se pueden estimar mediante el espectro de potencia $P(k)$ que se calcula como la magnitud del espectro al cuadrado, o lo que es lo mismo, a la suma de los cuadrados de la parte real e imaginaria del espectro:

$$\begin{aligned} P(k) &= \left| X(k) \right|^2 \\ P(k) &= \mathbf{real}\left[X(k)\right]^2 + \mathbf{imag}\left[X(k)\right]^2 \end{aligned} \qquad k = 0,1,2,3,\ldots,N_{fft}-1 \qquad \textbf{(3.19)}$$

El espectro de potencia se compone de valores reales, enfatiza los picos en el espectro y conserva la simetría par del espectro en frecuencia. Por lo tanto, normalmente solamente se ocupa la mitad inferior del espectro, es decir los primeros N_p valores:

$$Np = 0.5N_{fft} \qquad \textbf{(3.20)}$$

$$P(k) = \mathbf{real}\left[X(k)\right]^2 + \mathbf{imag}\left[X(k)\right]^2 \qquad k = 0,1,2,3,\ldots,N_p \qquad \textbf{(3.21)}$$

Hay que destacar que el espectro de potencia solamente proporciona información acerca de la magnitud, pues la información relacionada con la fase se descarta. Lo que no constituye una limitante, pues la fase no es tan importante como lo es la amplitud en la mayoría de las aplicaciones de voz [3]. Esto es consistente con el hecho de que la fase no aporta ninguna información que resulte útil. De hecho, Se ha comprobado experimentalmente que la percepción de una señal reconstruida con una fase distinta es prácticamente indistinguible de la señal original, esto si se mantiene la continuidad secuencial de la fase entre las tramas [1].

3.11 ESPECTROGRAMAS

Una de las primeras extensiones del análisis de voz al dominio de la frecuencia fue el desarrollo de los espectrogramas. Puesto que la forma del tracto vocal se modifica en función del tiempo para producir los sonidos deseados para comunicarse, también lo deben hacer las propiedades espectrales de la señal. Un espectrograma es una representación gráfica de la variación de la magnitud del espectro con respecto al tiempo. De tal manera que una señal de voz en dos dimensiones (tiempo , amplitud) se transforma en un patrón tridimensional (tiempo , frecuencia , magnitud).

La idea fundamental en la que se basan los espectrogramas, es el calcular la transformada discreta de Fourier para cada una de las tramas de análisis, mostrando la magnitud para cada punto de tiempo y frecuencia. El tiempo se representa en el eje de la variable independiente, la frecuencia se representa en el eje de la variable dependiente y la magnitud se representa por medio de una escala de colores (o grises). Existen dos tipos básicos de espectrogramas: los de banda angosta (narrow band) y los de banda ancha (wide band).

Los espectrogramas de banda angosta, son los que se obtienen cuando la resolución en frecuencia es alta. Estos emplean tramas relativamente largas ($T > 20$ ms) con las que se obtiene una buena resolución en la frecuencia a expensas de la resolución en el tiempo. Se utilizan para estimar los valores de la frecuencia del tono fundamental (pitch) y de sus armónicas.

Los espectrogramas de banda ancha, son los que se obtienen cuando la resolución en frecuencia es baja. Estos emplean tramas relativamente cortas ($T < 10$ ms) con las que se obtiene una buena resolución en el tiempo a expensas de la resolución en la frecuencia. Se utilizan para estimar los parámetros del tracto vocal que varían rápidamente en el tiempo.

La figura 3.6 ilustra la relación entre la resolución en el tiempo y la resolución en la frecuencia para los espectrogramas. Para mayor claridad, se omite el traslape entre las tramas. La resolución en el tiempo Δt es baja cuando hay pocas tramas, pero es alta si se tienen muchas tramas. Por otra parte, la resolución en la frecuencia Δf es alta cuando la longitud de la trama N es grande, pero es baja si la longitud de la trama N es pequeña.

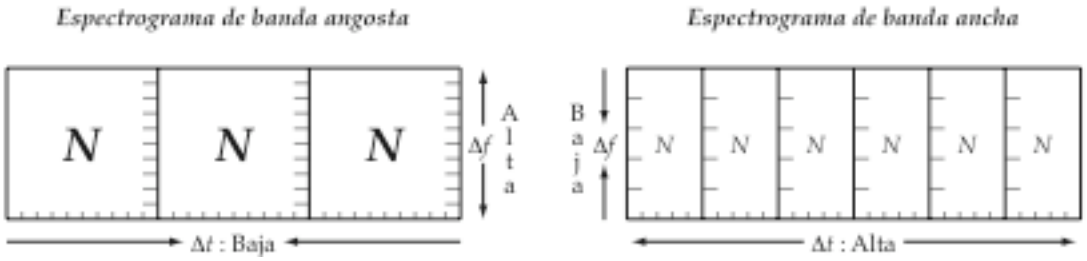


FIGURA 3.6 Relación entre las resoluciones para los espectrogramas

Si se fija la frecuencia de muestreo, la resolución en frecuencia Δf es inversamente proporcional a la longitud de la trama N:

$$\Delta f = \frac{f_s}{N} \tag{3.22}$$

Aumentar la resolución en frecuencia (obtener valores más pequeños para Δf) es equivalente a usar secuencias más largas, aunque esto está en conflicto con el requerimiento de analizar segmentos estacionarios de la señal. Las ventanas de más de 70 ms tienen una resolución en frecuencia más alta, con lo que es posible identificar las frecuencias individuales de las armónicas. Sin embargo en tal caso las transiciones súbitas en el espectro, como por ejemplo, de las plosivas, no son detectadas [1].

3.12 PROCEDIMIENTO DE CÁLCULO PARA LOS ESPECTROGRAMAS

A continuación se describe un procedimiento para obtener los espectrogramas de una señal de voz en dos y tres dimensiones. La figura 3.7 muestra el diagrama de bloques para este procedimiento. Los bloques muestran las operaciones necesarias y a la izquierda de éstos se muestran los parámetros de entrada requeridos en cada paso. A partir del cuarto bloque, el ponderado por la ventana de Hamming, las operaciones se realizan sobre las tramas de la señal de voz, también se ilustra el número de parámetros resultantes para cada trama mediante la cantidad de flechas.

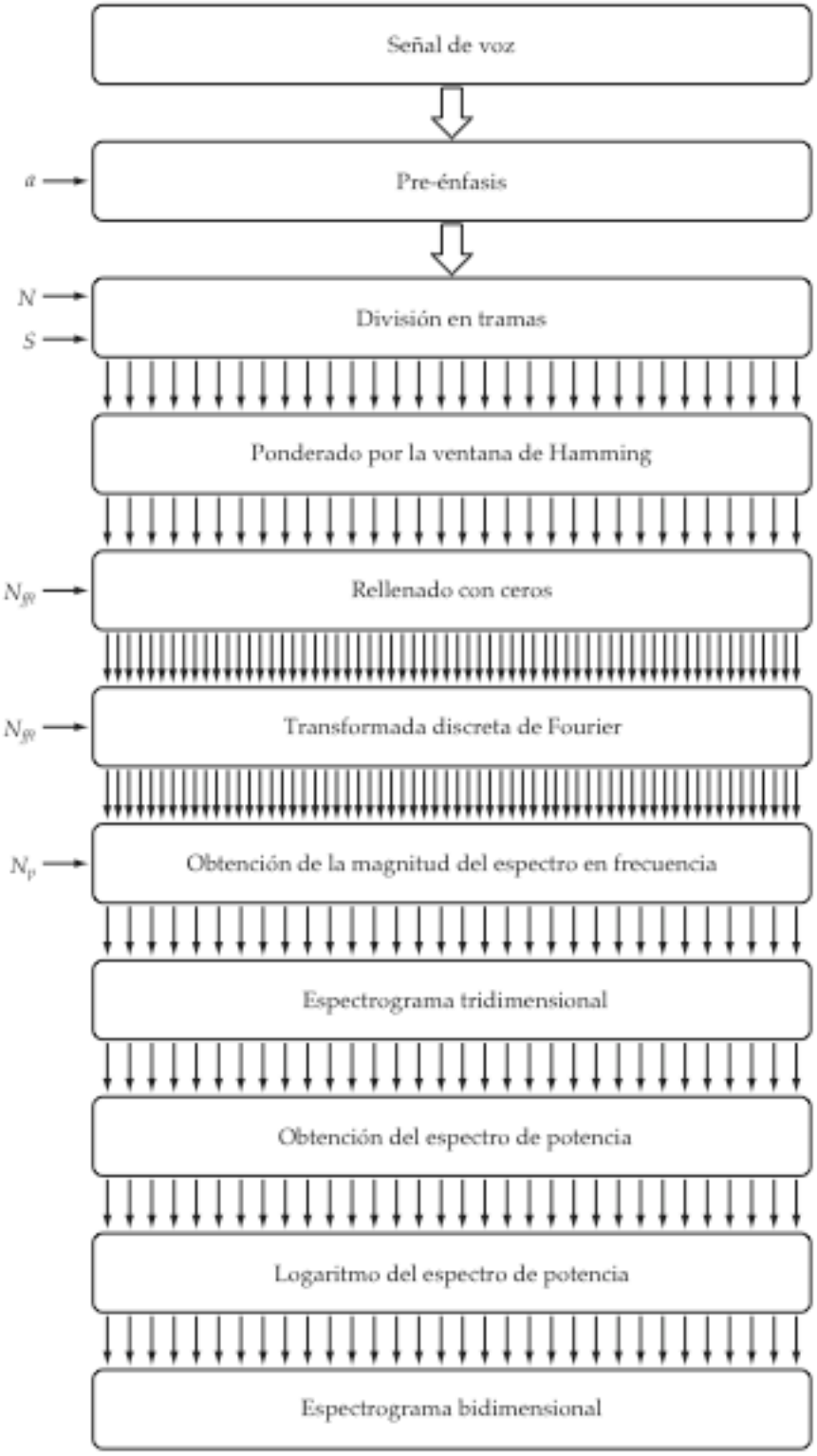


FIGURA 3.7 Procedimiento de cálculo para los espectrogramas

3.12.1 PRE-ÉNFASIS

Para hacer que los espectrogramas sean más fáciles de interpretar, se puede realizar un pre-énfasis de la señal de voz para amplificar las altas frecuencias, las cuales normalmente son atenuadas al hablar [9].

$$x_p(n) = x(n) - ax(n-1) \qquad \begin{matrix} 0.9 \le a \le 1 \\ n = 0,1,2,3,\ldots,N_x - 1 \end{matrix} \qquad \textbf{(3.23)}$$

3.12.2 DIVISIÓN EN TRAMAS

La señal de voz se divide en L tramas de N muestras para su análisis, con S muestras de separación entre tramas adyacentes:

$$x_\ell(n) = x(n + \ell \cdot S) \qquad \begin{matrix} N \ge S \\ \ell = 0,1,2,3,\ldots,L-1 \\ n = 0,1,2,3,\ldots,N-1 \end{matrix} \qquad \textbf{(3.24)}$$

3.12.3 PONDERADO POR LA VENTANA DE HAMMING

El ponderado de la trama de análisis por la ventana de Hamming, contribuye a minimizar las discontinuidades al inicio y al final de las tramas:

$$x_w(n) = x_\ell(n) \cdot w(n)$$

$$x_w(n) = x_\ell(n) \cdot \left[0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)\right] \qquad n = 0,1,2,3,\ldots,N-1 \qquad \textbf{(3.25)}$$

3.12.4 RELLENADO CON CEROS

Se añaden las muestras nulas que hagan falta para que la longitud de la trama de análisis sea igual a N_{fft}, la potencia de dos más próxima:

$$N_{fft} = 2^{\left[\log_2(N)\right]_{\text{int}}} \qquad \textbf{(3.26)}$$

$$x_{zp}(n) = \begin{cases} x_w(n) & \text{para} \quad n = 0,1,2,3,\ldots,N-1 \\ 0 & \text{para} \quad n = N, N+1,\ldots,N_{fft}-1 \end{cases} \qquad \textbf{(3.27)}$$

3.12.5 TRANSFORMADA DISCRETA DE FOURIER

La transformada discreta de Fourier de la trama de análisis se calcula mediante la transformada rápida de Fourier de N_{fft} puntos:

$$X(k) = \mathbf{FFT}\left\{x_{zp}(n), N_{fft}\right\} \qquad \begin{matrix} k = 0,1,2,3,\ldots,N_{fft}-1 \\ n = 0,1,2,3,\ldots,N_{fft}-1 \end{matrix} \qquad \textbf{(3.28)}$$

3.12.6 OBTENCIÓN DEL ESPECTRO DE POTENCIA

El espectro de potencia es igual a la suma de los cuadrados de la parte real e imaginaria del espectro en frecuencia de la trama de análisis:

$$N_p = 0.5N_{fft} \qquad \textbf{(3.29)}$$

$$P(k) = \mathbf{real}\left[X(k)\right]^2 + \mathbf{imag}\left[X(k)\right]^2 \qquad k = 0,1,2,3,\ldots,N_p \qquad \textbf{(3.30)}$$

3.12.7 LOGARITMO DEL ESPECTRO DE POTENCIA

Como último paso, se obtiene el logaritmo de los valores del espectro de potencia:

$$Y(k) = \mathbf{log}\left[P(k)\right] \qquad k = 0,1,2,3,\ldots,N_p \qquad \textbf{(3.31)}$$

3.12.8 EJEMPLOS DE LOS ESPECTROGRAMAS

Los espectrogramas pueden ser de gran utilidad para estimar los valores de la frecuencia del tono fundamental (pitch) y de sus armónicas, así como para identificar las regiones correspondientes para los sonidos. Generalmente, los sonidos sonoros tienen una mayor concentración de energía a baja frecuencia, pero los sonidos sordos la tienen para alta frecuencia.

De manera alternativa, se puede obtener la representación gráfica tridimensional de un espectrograma. En este caso se muestra la magnitud del espectro en frecuencia en lugar del logaritmo del espectro de potencia. Aunque se pueden apreciar mejor ciertas características, en general este tipo de representación es menos práctica que un espectrograma convencional [12]. En la figura 3.8 se muestran la forma de onda y sus respectivos espectrogramas para la palabra en inglés "*six*", dicha por una mujer.

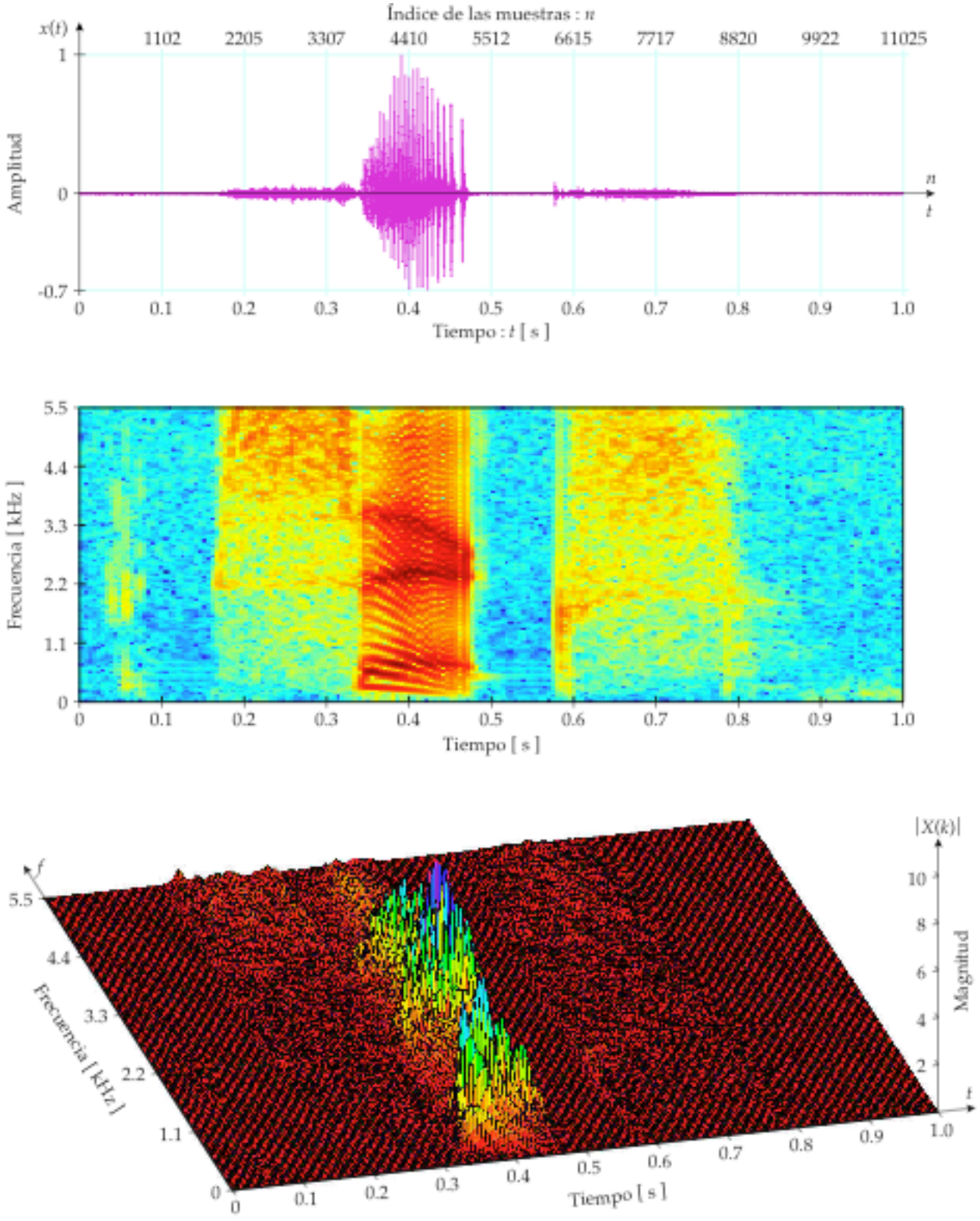


FIGURA 3.8 Forma de onda de una señal de voz y sus espectrogramas en 2 y 3 dimensiones

Para resaltar los sonidos sordos, se realizó un pre-énfasis de la señal de voz con $a = 0.95$. En los dos espectrogramas se pueden observar más claramente las correspondencia de los segmentos, dos para los sonidos sordos, uno para un sonido sonoro y tres para los silencios, así como sus diferencias entre éstos. Evidentemente, el sonido sonoro porta una mayor cantidad de energía que los dos sonidos sordos. También se puede apreciar una diferencia entre los dos fonemas /s/, al inicio y al final de la palabra, así como una liberación súbita de una plosiva /k/ después del silencio que se encuentra dentro de la palabra. La frecuencia de muestreo es $f_s = 11{,}025$ Hz, la longitud de la trama es de $N = 220$ muestras (20 ms) con $S = 55$ muestras (5 ms) de separación entre tramas adyacentes, el número de puntos para la FFT es de $N_{fft} = 256$ puntos para el espectrograma en 2D, pero para el espectrograma en 3D es de $N_{fft} = 512$ puntos.

3.13 FILTRADO E INTEGRACIÓN O RE-MUESTREO POR BANDAS CRÍTICAS

El análisis espectral revela aquellas características de la señal de voz que son consecuencia directa de la configuración del tracto vocal. Asimismo, estas características de la voz por lo general se obtienen como las respuestas o salidas de un banco de filtros.

Ahora bien, se ha encontrado que la percepción de una frecuencia en particular por parte del sistema auditivo humano está influenciada por la energía dentro de una banda crítica de frecuencias a su alrededor [3]. En la psicoacústica, a esta propiedad se le conoce como la resolución espectral por bandas críticas [5].

Un método muy utilizado para implementar estos bancos, consiste en efectuar el filtrado directamente en el dominio de la transformada discreta de Fourier. Los filtros simplemente son versiones desplazadas en la escala de frecuencia de alguna ventana. Estas ventanas no están equiespaciadas en la escala de frecuencia en Hertz, de hecho se utiliza otra escala no lineal (logarítmica) para conseguir una mejor relación entre las resoluciones en el tiempo y en la frecuencia, pues los filtros paso banda angostos para baja frecuencia hacen posible la detección de las frecuencias del tono fundamental (pitch), pero el incremento en el ancho de banda de los filtros paso banda para alta frecuencia permiten una mejor resolución en el tiempo, con la que se pueden detectar las plosivas. Además, la parte del espectro que se encuentra por debajo de 1000 Hz se procesa por una mayor cantidad de filtros, pues ésta contiene más información [3].

La respuesta en frecuencia del banco de filtros simula el proceso de percepción que tiene lugar dentro del oído interno del humano, por eso a este proceso de filtrado también se le denomina como ponderado perceptual [1]. La figura 3.9 muestra la estructura general de estos bancos de filtros perceptuales, en donde *BC* es la cantidad o el número de filtros o bandas críticas en el banco.

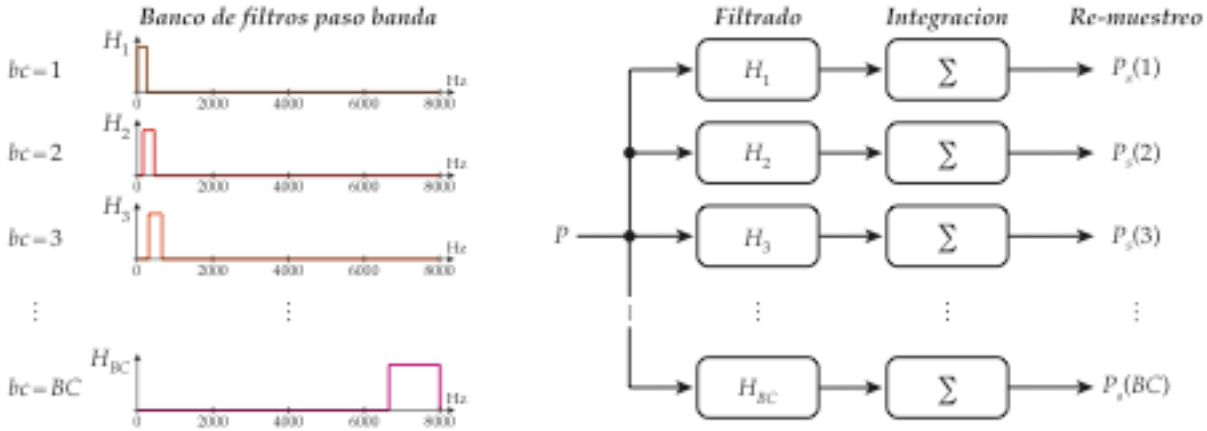


FIGURA 3.9 Estructura general de los bancos de filtros perceptuales

Entonces, al integrar las salidas individuales de los filtros, se obtiene una muestra por cada una de las bandas críticas, por lo que el espectro es re-muestreado a determinados rangos de frecuencia, proporcionando así, una compresión o reducción en el número de parámetros. El efecto neto de esta etapa es reducir la sensibilidad del espectro a la frecuencia, particularmente para las altas frecuencias, las cuales de alguna manera son enfatizadas debido al ancho de banda más amplio de los filtros del banco [5].

3.14 CAMBIO DE LA ESCALA DE FRECUENCIA

Existen muchas formas diferentes para las ventanas que se pueden emplear en los bancos de filtros, pero todas estas se basan en alguna escala de frecuencia que se considera aproximadamente lineal para frecuencias menores de 1,000 Hz, pero para frecuencias mayores que ésta, presenta un comportamiento no lineal, logarítmico.

Este cambio en la escala de frecuencia en inglés se denomina frequency warping, es decir, que se realiza un arqueado o una transformación de la escala de frecuencia en Hertz, a otra escala de frecuencia que exhibe un comportamiento (no lineal) logarítmico. Las escalas mel y Bark son dos ejemplos de estas escalas. Puesto que están basadas en datos experimentales de la percepción humana, existen varias aproximaciones que se pueden usar para estas [5].

3.14.1 LA ESCALA MEL

La escala mel fue propuesta por Stevens, Volksman y Newmann en 1940. Un mel es una unidad de medición de la frecuencia o del tono percibido. Esta no corresponde de manera lineal a la frecuencia real del tono, pues el sistema auditivo humano no parece percibir la frecuencia de una manera lineal. Se designó como punto inicial de referencia, que una frecuencia de 1000 Hz también es de 1000 mels. Los observadores juzgan tonos espaciados exponencialmente como tonos equiespaciados. De esta manera fue posible determinar un mapeo entre la escala de la frecuencia real (en Hertz) y la escala de la frecuencia percibida (en mels) [3].

La siguiente ecuación proporciona una equivalencia entre la escala logarítmica de la frecuencia mel f_{mel} con respecto a la escala de frecuencia en Hertz f_{Hz} [3]:

$$f_{mel}(f_{Hz}) = \frac{1000}{\mathbf{Ln}(2)}\mathbf{Ln}\left(1+\frac{f_{Hz}}{1000}\right) \qquad \textbf{(3.32)}$$

No obstante, se acostumbra utilizar otra función de aproximación para realizar la transformación no lineal de la frecuencia en Hertz a la escala de frecuencia mel [9]:

$$f_{mel}(f_{Hz}) = 1125\mathbf{Ln}\left(1+\frac{f_{Hz}}{700}\right) \qquad \textbf{(3.33)}$$

La figura 3.10 muestra la representación gráfica de la ecuación anterior, en donde se puede apreciar el arqueado de la frecuencia al cambiar de la escala de frecuencia (lineal) en Hertz f_{Hz} a la escala logarítmica de la frecuencia mel f_{mel}:

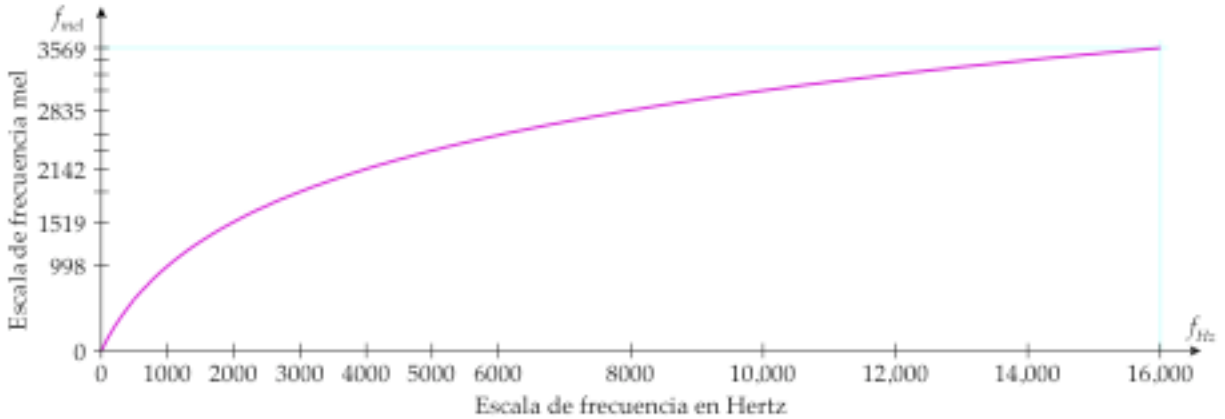


FIGURA 3.10 Gráfica del arqueado de la frecuencia con la escala mel

3.14.2 LA ESCALA BARK

La escala Bark fue propuesta por Zwicker en 1961. La cóclea actúa como si estuviera construida por filtros traslapados, siendo sus anchos de banda idénticos a los de las bandas críticas. El término banda crítica fue introducido por Fletcher en la década de 1940 al realizar experimentos sobre enmascaramiento, que consiste en reproducir simultáneamente dos sonidos a diferentes frecuencias y que uno de ellos impida la percepción del otro. También se puede utilizar ruido de banda angosta para enmascarar un tono, donde el ancho de banda que contribuye al enmascaramiento se denomina banda crítica. Sólo el ruido dentro de una banda crítica contribuye al enmascaramiento, sin embargo, el umbral decae conforme la frecuencia central del ruido o del tono se aleja de las frecuencias centrales de las bandas críticas. La escala va de 1 a 24 Barks, que corresponden a las primeras 24 bandas críticas del sistema auditivo. Como puede apreciarse en la tabla 3.1 el ancho de cada banda varía desde los 100 Hz a baja frecuencia hasta los 3500 Hz en alta frecuencia.

TABLA 3.1 Bandas críticas de la escala Bark

Número de banda Bark	Límite inferior [Hz]	Frecuencia central [Hz]	Límite superior [Hz]	Ancho de banda [Hz]
1	0	50	100	100
2	100	150	200	100
3	200	250	300	100
4	300	350	400	100
5	400	450	510	110
6	510	570	630	120
7	630	700	770	140
8	770	840	920	150
9	920	1000	1080	160
10	1080	1170	1270	190
11	1270	1370	1480	210
12	1480	1600	1720	240
13	1720	1850	2000	280
14	2000	2150	2320	320
15	2320	2500	2700	380
16	2700	2900	3150	450
17	3150	3400	3700	550
18	3700	4000	4400	700
19	4400	4800	5300	900
20	5300	5800	6400	1100
21	6400	7000	7700	1300
22	7700	8500	9500	1800
23	9500	10,500	12,000	2500
24	12,000	13,500	15,500	3500

La siguiente ecuación proporciona una equivalencia entre la escala logarítmica de la frecuencia Bark f_{Bark} con la escala de frecuencia en Hertz f_{Hz} [9]:

$$f_{Bark}(f_{Hz}) = 13\,\mathbf{arctan}(0.00076 f_{Hz}) + 3.5\,\mathbf{arctan}\left[\left(\frac{f_{Hz}}{7500}\right)^2\right] \qquad \textbf{(3.34)}$$

No obstante, la transformación no lineal de la frecuencia en Hertz a la escala de frecuencia Bark, se puede realizar mediante otra función de aproximación, como la función del arqueado de la frecuencia de Schroeder [7]:

$$f_{Bark}(f_{Hz}) = 6\,\mathbf{Ln}\left[\left(\frac{f_{Hz}}{600}\right) + \sqrt{\left(\frac{f_{Hz}}{600}\right)^2 + 1}\right] \qquad \textbf{(3.35)}$$

La figura 3.11 muestra la representación gráfica de la ecuación anterior, en donde se puede apreciar el arqueado de la frecuencia al cambiar de la escala de frecuencia (lineal) en Hertz f_{Hz} a la escala logarítmica de la frecuencia Bark f_{Bark}:

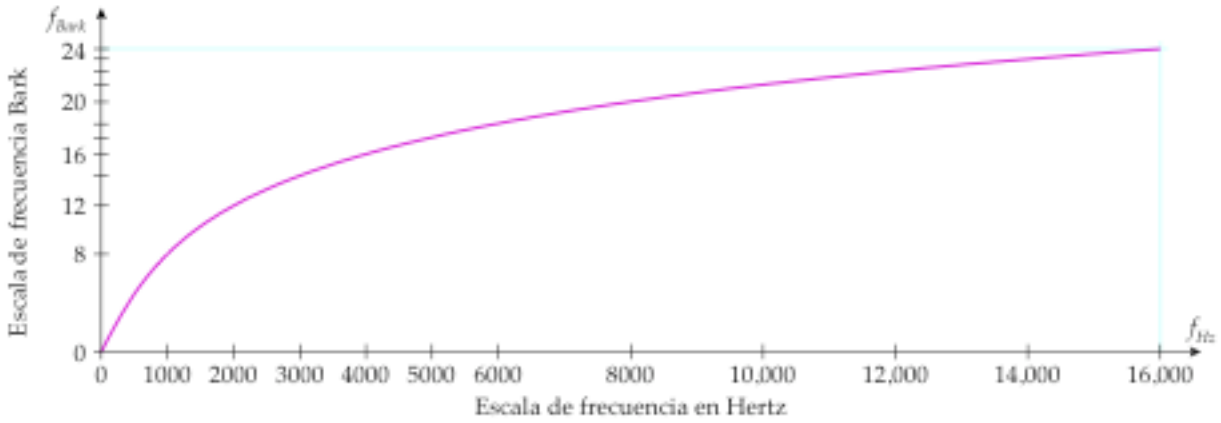


FIGURA 3.11 Gráfica del arqueado de la frecuencia con la escala Bark

CAPÍTULO

4

LPC

CODIFICACIÓN DE PREDICCIÓN LINEAL

4.1 INTRODUCCIÓN

Las técnicas y los métodos de predicción lineal ya tienen un buen tiempo que han estado disponibles en la ingeniería. Estos conceptos ya se habían estado utilizado en el área de control y en la teoría de la información, en donde se les denomina estimación de sistemas e identificación de sistemas [13]. Pero a partir de 1968, varios investigadores empezaron a publicar algunos artículos en los que mostraban que la técnica de predicción lineal también podía aplicarse a la señal de voz [3].

El término "predicción lineal" en el procesamiento de voz, se refiriere a las diferentes formas de abordar el problema del modelado de la señal de voz. Esta técnica permite estimar los parámetros de la señal de voz con una gran exactitud, puede representarla de una forma compacta mediante un número reducido de parámetros, que además se obtienen de una manera rápida, simple y eficiente. También es muy importante porque ha sido la base para muchos otros desarrollos teóricos y prácticos de tal manera que sería difícil concebir la tecnología actual para el procesamiento de voz prescindiendo de esta técnica [3].

El análisis de predicción lineal provee un método robusto, confiable y exacto para estimar los parámetros que caracterizan el modelo LPC para la producción de la voz, un sistema lineal con parámetros variantes en el tiempo. Este conjunto de técnicas de análisis suele denominarse como codificación de predicción lineal o **L**inear **P**redictive **C**oding (LPC), pues en los sistemas de comunicaciones, estos parámetros (u otros relacionados) se usan para codificar la señal de voz [3].

4.2 DESCRIPCIÓN

Un análisis de la señal de voz, permite apreciar que las muestras adyacentes están altamente correlacionadas, especialmente para el caso de los sonidos sonoros. Ahora bien, se puede aprovechar esta correlación, pues hace posible la predictibilidad entre las muestras adyacentes.

La idea fundamental detrás del método LPC es que el valor para una muestra de la señal de voz puede ser predicho o estimado a partir de unas cuantas de sus predecesoras. Esta predicción se puede realizar mediante la combinación lineal de las p muestras más recientes, por está razón, al proceso se le denomina predicción lineal. Matemáticamente, la siguiente ecuación en diferencias describe al predictor lineal:

$$\tilde{x}(n) = a_1 x(n-1) + a_2 x(n-2) + a_3 x(n-3) + \ldots + a_p x(n-p)$$

$$\tilde{x}(n) = \sum_{c=1}^{p} a_c x(n-c) \qquad \textbf{(4.1)}$$

En donde $x(n)$ es la señal de voz y $\tilde{x}(n)$ es la señal predicha o estimada, los coeficientes del predictor $a_c = [a_1, a_2, a_3, \ldots, a_p]$ son los pesos o los coeficientes de ponderado empleados en la combinación lineal.

Con estos coeficientes se pueden estimar con una exactitud aceptable los valores para las muestras de la señal de voz. Sin embargo, como no es posible predecir exactamente el valor correspondiente para cada una de las muestras de la señal, se obtiene como consecuencia, un error de predicción $e(n)$ para cada instante de muestreo:

$$e(n) = x(n) - \tilde{x}(n)$$

$$e(n) = x(n) - \sum_{c=1}^{p} a_c x(n-c) \qquad \textbf{(4.2)}$$

Si se obtienen las transformadas z, la función de transferencia del sistema es:

$$A(z) = \frac{E(z)}{X(z)}$$

$$A(z) = 1 - \sum_{c=1}^{p} a_c z^{-c} \qquad \textbf{(4.3)}$$

La función $A(z)$ se conoce como el filtro del error de predicción.

La estructura del filtro del error de predicción $A(z)$ corresponde a un filtro digital FIR (con respuesta finita al impulso), como se muestra en la figura 4.1.

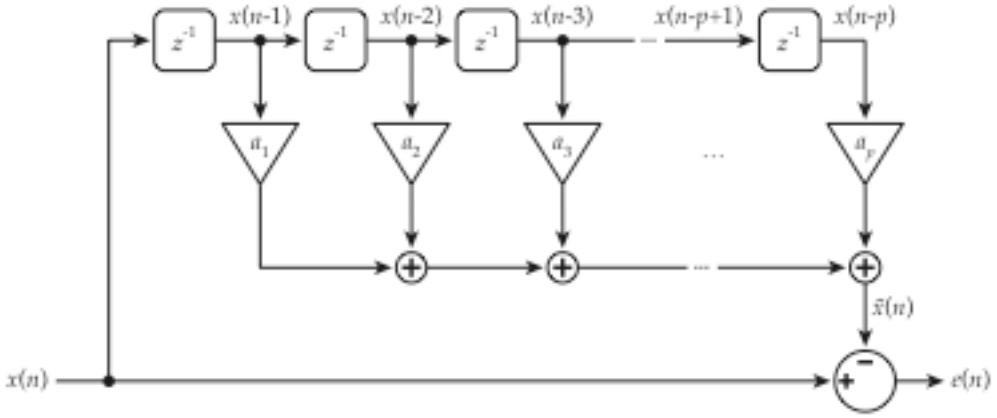


FIGURA 4.1 Predictor LPC (FIR) para el análisis de voz

El sistema consiste de p elementos de retardo, que almacenan los p valores más recientes de la señal de voz $x(n)$, los cuales se emplean en una combinación lineal de acuerdo a los coeficientes del predictor $a_c = [a_1, a_2, a_3, \ldots, a_p]$ para determinar el valor de la señal predicha o estimada $\tilde{x}(n)$. Al obtener la diferencia entre la señal de voz y la señal predicha se obtiene la señal de error $e(n)$, que también se conoce como la señal residual.

Ahora bien, si se conoce la señal de error, entonces se puede reconstruir una replica exacta de la señal de voz a partir de la señal predicha o estimada:

$$x(n) = e(n) + \tilde{x}(n)$$

$$x(n) = e(n) + \sum_{c=1}^{p} a_c x(n-c) \qquad \textbf{(4.4)}$$

Si se obtienen las transformadas z, la función de transferencia del sistema ahora es:

$$H(z) = \frac{X(z)}{E(z)}$$

$$H(z) = \frac{1}{1 - \sum_{c=1}^{p} a_c z^{-c}} \qquad \textbf{(4.5)}$$

La estructura del filtro sintetizador $H(z)$ corresponde a un filtro digital puramente recursivo, el cual es un caso particular de un filtro digital IIR (con respuesta infinita al impulso), como se muestra en la figura 4.2.

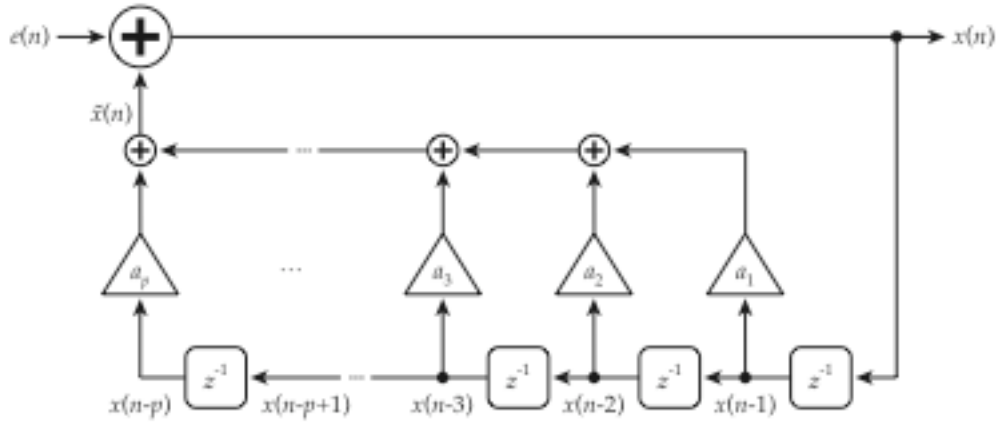


FIGURA 4.2 Predictor LPC (IIR) para la síntesis de voz

De manera similar, el sistema consiste de p elementos de retardo, que almacenan los p valores más recientes de la señal de voz reconstruida $x(n)$, los cuales se emplean en una combinación lineal de acuerdo a los coeficientes del predictor $a_c = [a_1, a_2, a_3, \ldots, a_p]$ para determinar el valor de la señal predicha o estimada $\tilde{x}(n)$. Pero ahora, al realizar la suma de la señal predicha o estimada y la señal de error se obtiene la señal de voz reconstruida.

La función de transferencia $H(z)$ de este sistema corresponde a un filtro digital que solamente tiene polos, potencias de z en su denominador. El término identificación de sistemas, resulta especialmente descriptivo del método LPC, pues una vez obtenidos los coeficientes del predictor, el sistema ha sido identificado a tal grado que puede ser modelado como un sistema lineal todos polos [13].

Desde el punto de vista de sistemas, la señal de voz reconstruida $x(n)$ puede verse como la respuesta o salida del filtro de síntesis cuando su excitación o entrada es la señal de error. De tal manera que el filtro $H(z)$ modela la respuesta del tracto vocal y la señal de error $e(n)$ denota la excitación del tracto vocal. De manera similar, la señal de error $e(n)$ puede verse como la respuesta o salida del filtro de análisis $A(z)$ cuando su excitación o entrada es la señal de voz original $x(n)$.

Debido a la naturaleza variante en el tiempo de la voz, estos coeficientes tienen que calcularse para segmentos cortos, de 10 a 30 ms, durante los cuales sus características se mantienen relativamente estacionarias, por lo que la secuencia de las muestras de la señal $x(n)$ pueden ser representadas por los p coeficientes del predictor $a_c = [a_1, a_2, a_3, \ldots, a_p]$ y por la señal de error $e(n)$ correspondiente.

Para obtener la pareja de filtros compatibles análisis-síntesis, los coeficientes del filtro de síntesis deben de ser actualizados periódicamente a los valores más recientes de los coeficientes del filtro de análisis. La semejanza entre las ecuaciones (4.5) y (4.3), se debe a que las formas de los filtros de análisis y de síntesis se escogieron de tal manera que fueran compatibles, pues son filtros inversos.

$$H(z) = \frac{1}{A(z)} \qquad \textbf{(4.6)}$$

Puesto que el filtro $A(z)$ es el inverso del filtro todos polos $H(z)$, al proceso en el cual se obtiene la señal de error se le denomina como filtrado inverso.

El problema fundamental en el análisis de predicción lineal es, determinar el conjunto de coeficientes del predictor directamente de la señal de voz, de tal manera que las propiedades espectrales del filtro todos polos sean un buen estimado de las propiedades espectrales de la señal de voz.

La señal de error para los sonidos sonoros por lo general contiene valores muy pequeños la mayor parte del tiempo, a excepción de una discontinuidad que coincide con el inicio de cada periodo del tono fundamental (pitch). Para los sonidos sordos, la señal de error se asemeja a ruido aleatorio. Ahora bien, la magnitud de la señal de error resultante del análisis de una trama de voz, puede ser reducido a valores muy pequeños mediante la selección apropiada de los coeficientes del predictor.

Por lo tanto, el enfoque básico es buscar los coeficientes del predictor tales que minimicen el error cuadrático medio para un número específico de muestras, es decir, para una trama o segmento corto de la señal de voz.

La minimización del error cuadrático medio resulta útil porque proporciona una forma eficaz de resolver el problema. Pero más importante aún, los resultados obtenidos conforman una representación muy útil y exacta de la señal de voz. La siguiente ecuación define al error cuadrático medio E:

$$E = \sum_m [e(m)]^2$$
$$E = \sum_m \left[x(m) - \sum_{c=1}^{p} a_c x(m-c) \right]^2 \qquad \textbf{(4.7)}$$

En donde $x(m) = x(n+m)$ representa un segmento o trama de la señal de voz y $e(m) = e(n+m)$ su error de predicción correspondiente. Cabe señalar que para obtener un promedio se debería dividir el error entre la longitud del segmento. Sin embargo, esta constante resulta irrelevante con respecto a la solución del problema y por lo tanto se puede omitir [13]. Si lo que se busca es minimizar el error cuadrático medio por la selección apropiada de los coeficientes del predictor, entonces se deben obtener las derivadas parciales del error con respecto a cada uno de los coeficientes e igualarlas a cero:

$$\frac{\partial E}{\partial a_i} = 0 \qquad i = 1,2,3,\ldots,p \qquad \textbf{(4.8)}$$

Obteniendo así la siguiente ecuación:

$$\sum_{c=1}^{p} a_c \sum_m x(m-i)\cdot x(m-c) = \sum_m x(m-i)\cdot x(m) \qquad \begin{matrix} i = 1,2,3,\ldots,p \\ c = 1,2,3,\ldots,p \end{matrix} \qquad \textbf{(4.9)}$$

Si se define la siguiente función:

$$\phi(i,c) = \sum_m x(m-i)\cdot x(m-c) \qquad \textbf{(4.10)}$$

Entonces, la ecuación (4.9) se puede reescribir de una forma más compacta:

$$\sum_{c=1}^{p} a_c \phi(i,c) = \phi(i,0) \qquad \begin{matrix} i = 1,2,3,\ldots,p \\ c = 0,1,2,3,\ldots,p \end{matrix} \qquad \textbf{(4.11)}$$

La ecuación anterior representa al sistema de ecuaciones de predicción lineal, que consiste de p ecuaciones lineales con p incógnitas. La solución de este sistema de ecuaciones proporciona los valores óptimos de los coeficientes del predictor $a_c = [a_1, a_2, a_3, \ldots, a_p]$ que minimizan el error cuadrático medio para la trama de análisis.

Sin embargo, el resolver un sistema de ecuaciones de 10 a 20 incógnitas no es un problema trivial, ni siquiera para el caso de las ecuaciones lineales. La inversión de una matriz de ese orden es un proceso costoso en términos de computo, no obstante, la solución para este sistema de ecuaciones se puede obtener de otra manera mucho más eficiente.

4.2.1 SOLUCIÓN DEL SISTEMA DE ECUACIONES LPC

Para poder implementar de manera efectiva un sistema de análisis de predicción lineal, es necesario resolver el sistema de ecuaciones lineales de una manera rápida y eficiente. Aunque se puede emplear cualquier técnica para obtener la solución, es posible hacerlo de una manera mucho más eficiente al aprovechar las características particulares de este sistema de ecuaciones.

Para obtener los coeficientes óptimos del predictor, primero se tienen que calcular los valores de la función $\phi(i,c)$ para $i = 1,2,3,\ldots,p$ y $c = 0,1,2,3,\ldots,p$. Una vez hecho esto, solo hace falta resolver el sistema de ecuaciones lineales descrito por la ecuación (4.11):

$$\sum_{c=1}^{p} a_c \phi(i,c) = \phi(i,0) \qquad \begin{matrix} i = 1,2,3,\ldots,p \\ c = 0,1,2,3,\ldots,p \end{matrix} \tag{4.11}$$

De tal manera que, en esencia, el análisis de predicción lineal es un método muy directo.

Ahora bien, para obtener un procedimiento en tiempo corto, los límites de las sumatorias para las ecuaciones:

$$E = \sum_{m} [e(m)]^2 \tag{4.7}$$

$$\phi(i,c) = \sum_{m} x(m-i) \cdot x(m-c) \tag{4.10}$$

Deben de estar dentro de un intervalo finito. Las cuestiones concernientes a la definición del segmento o trama de la señal y la de los límites de las sumatorias, dan como resultado dos métodos de análisis de predicción lineal: el método de la autocorrelación y el método de la covarianza. Puesto que los métodos son diferentes, las propiedades de los coeficientes obtenidos también resultan distintas. Sin embargo, el método de la covarianza por lo general no se utiliza en los sistemas de reconocimiento de voz [9].

4.2.2 EL MÉTODO DE LA AUTOCORRELACIÓN

Una forma para determinar los límites de la sumatorias es el suponer que todos los valores de las muestras que se encuentran fuera del intervalo $0 \le m \le N-1$ son iguales a cero. Lo cual se puede representar convenientemente mediante el ponderado de la señal de voz por una ventana $w(m)$ de longitud finita:

$$x(m) = x(n+m) \cdot w(m) \qquad \textbf{(4.12)}$$

Donde $w(m)$ puede ser una ventana de Hamming, como se muestra en la figura 4.3.

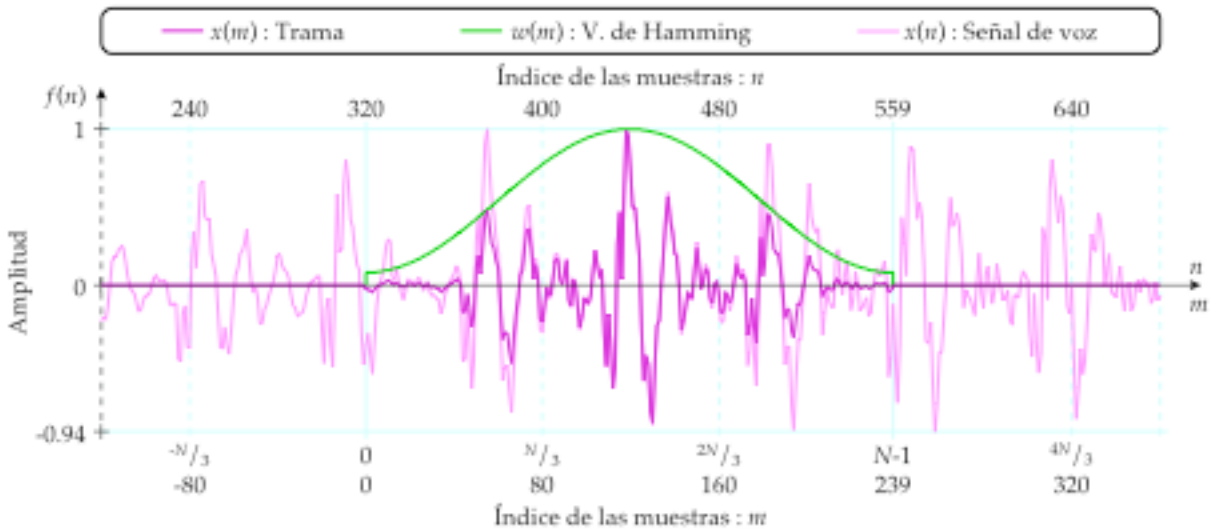


FIGURA 4.3 Trama de análisis para el método de la autocorrelación

El efecto de esta suposición en los límites de las sumatorias se puede apreciar al considerar sus efectos en la expansión de la ecuación del error de predicción:

$$e(m) = x(m) - a_1 x(m-1) - a_2 x(m-2) - a_3 x(m-3) - \ldots - a_p x(m-p) \qquad \textbf{(4.13)}$$

Si solamente las muestras comprendidas dentro del intervalo $0 \le m \le N-1$ son diferentes de cero, entonces el error de predicción para un predictor de orden p, también será diferente de cero pero solamente dentro del intervalo $0 \le m \le N-1+p$. Sin embargo, existe la posibilidad de que el error de predicción sea grande en los extremos del intervalo. Esto es así porque para el inicio del intervalo, esto es $0 \le m \le p-1$, el predictor está tratando de estimar la señal con ceros, a partir de muestras nulas que se encuentran fuera del intervalo. De manera similar, para el final del intervalo, esto es $N \le m \le N-1+p$, el predictor está tratando de predecir los ceros, las muestras nulas fuera del intervalo, a partir de muestras que son diferentes de cero. Estos efectos son especialmente prominentes para los sonidos sonoros, para cuando el inicio del periodo del tono fundamental (pitch) ocurre muy cerca del inicio o del final del intervalo o la trama.

Ahora bien, el ponderado por la ventana de Hamming contribuye a reducir el error de predicción al disminuir la amplitud en los extremos de la señal. Sin embargo, este problema no ocurre para los sonidos sordos.

Entonces, para el método de la autocorrelación, los límites de la sumatoria del error cuadrático medio son:

$$E = \sum_{m=0}^{N-1+p} \left[e(m)\right]^2 \qquad \textbf{(4.14)}$$

Incluso, se puede decir que la sumatoria se efectúa para todos los valores diferentes de cero de $-\infty$ a $+\infty$ [13].

Los límites de la sumatoria para la función $\phi(i,c)$ son los mismos que los de la ecuación anterior:

$$\phi(i,c) = \sum_{m=0}^{N-1+p} x(m-i)\cdot x(m-c) \qquad \textbf{(4.15)}$$

Sin embargo, si se hace un cambio de variable:

$$n = m - i \qquad \textbf{(4.16)}$$

$$\phi(i,c) = \sum_{n=-i}^{N-1+p-i} x(n)\cdot x(n+i-c) \qquad \textbf{(4.17)}$$

Puesto que fuera del intervalo todas las muestras son iguales a cero, los límites de la sumatoria también se pueden cambiar, obteniendo así la siguiente ecuación:

$$\phi(i,c) = \sum_{n=0}^{N-1-(i-c)} x(n)\cdot x(n+(i-c)) \qquad \textbf{(4.18)}$$

La ecuación anterior es idéntica a la función de autocorrelación en tiempo corto, pero evaluada para $k = (i-c)$:

$$R(k) = \sum_{n=0}^{N-1-k} x(n)\cdot x(n+k) \qquad \textbf{(4.19)}$$

Entonces, al usar la función de autocorrelación y tomando en cuenta que es una función par, se tiene que:

$$\phi(i,c) = R(|i-c|) \qquad \begin{matrix} i = 1,2,3,\ldots,p \\ c = 0,1,2,3,\ldots,p \end{matrix} \qquad \textbf{(4.20)}$$

Por lo tanto, el sistema de ecuaciones de predicción lineal se puede expresar de la siguiente manera:

$$\sum_{c=1}^{p} a_c R(|i-c|) = R(i) \qquad \begin{matrix} i = 1,2,3,\ldots,p \\ c = 1,2,3,\ldots,p \end{matrix} \qquad \textbf{(4.21)}$$

Este sistema de ecuaciones lineales también se puede expresar en forma matricial de la siguiente manera:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & \cdots & R(p-2) \\ R(2) & R(1) & R(0) & \cdots & R(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{bmatrix} \qquad \textbf{(4.22)}$$

La matriz de los valores de autocorrelación corresponde a una matriz Toeplitz, es decir, que la matriz es simétrica y que todos los elementos en su diagonal principal son iguales. Esta propiedad especial se puede aprovechar para obtener la solución del sistema mediante un algoritmo eficiente.

Finalmente, el método de la autocorrelación proporciona una expresión muy simple para el cálculo del error cuadrático medio mínimo, en términos de los valores de autocorrelación y de los coeficientes del predictor:

$$E = R(0) - \sum_{c=1}^{p} a_c R(c) \qquad \textbf{(4.23)}$$

4.2.3 EL ALGORITMO DE LEVINSON-DURBIN

En 1947 Levinson publicó un algoritmo para resolver la ecuación matricial $\mathbf{Ax} = \mathbf{b}$, para cuando la matriz **A** es Toeplitz, definida positiva y donde **b** es un vector columna arbitrario. Las ecuaciones de la autocorrelación satisfacen estas condiciones, pero además el vector **b** tiene una relación especial con los elementos de la matriz **A**. En 1960 Durbin publicó un algoritmo recursivo mejorado para este caso en particular. Por esta razón, al método frecuentemente se le denomina como el algoritmo de Levinson-Durbin [3]. Una virtud del algoritmo es su eficiencia computacional. Su empleo resulta en un gran ahorro en el número de operaciones y de localidades de memoria empleadas, al compararlo contra otros métodos estándar para la inversión de matrices.

A continuación se muestra el procedimiento y las ecuaciones empleadas en la solución del sistema de ecuaciones:

- Inicialización del error cuadrático medio:

$$m = 0 \qquad E^{(0)} = R(0) \tag{4.24}$$

- Método recursivo para $m = 1,2,3,\ldots,p$:

  1er Paso. Computo del *m*-ésimo coeficiente de reflexión k_m:

$$k_m = \frac{R(m) - \sum_{c=1}^{m-1} \alpha_c^{(m-1)} R(m-c)}{E^{(m-1)}} \qquad m = 1,2,3,\ldots,p \tag{4.25}$$

  Condición inicial: la sumatoria se omite para el primer coeficiente k_1.

  2o Paso. Cálculo de los *c* coeficientes del predictor de *m*-ésimo orden $\alpha_c^{(m)}$:

$$\alpha_m^{(m)} = k_m \qquad m = 1,2,3,\ldots,p \tag{4.26}$$

$$\alpha_c^{(m)} = \alpha_c^{(m-1)} - k_m \alpha_{m-c}^{(m-1)} \qquad \begin{matrix} m = 2,3,4,\ldots,p \\ c = 1,\ldots,m-1 \end{matrix} \tag{4.27}$$

  Condición inicial: la ecuación (4.27) se omite para primer coeficiente $\alpha_1^{(1)}$.

  3er Paso. Computo del error cuadrático medio mínimo asociado al predictor de *m*-ésimo orden:

$$E^{(m)} = \left[1 - (k_m)^2\right] E^{(m-1)} \qquad m = 1,2,3,\ldots,p \tag{4.28}$$

  4o Paso. Incremento del orden del predictor:

$$m = m + 1 \tag{4.29}$$

  5o Paso. Se vuelve a iterar hasta que el orden sea igual al orden deseado:

$$m = p \tag{4.30}$$

- La solución final, los coeficientes del predictor a_c que minimizan el error cuadrático medio, están dados por el conjunto de coeficientes del predictor de orden *p*:

$$a_c = \alpha_c^{(p)} \qquad c = 1,2,3,\ldots,p \tag{4.31}$$

Este es un método de solución recursivo en el orden del modelo, lo que significa que la solución deseada para el modelo de orden *p*, se obtiene a partir de los demás modelos de menor orden, de tal manera que $\alpha_c^{(m)}$ y $E^{(m)}$ son los coeficientes del predictor y el error cuadrático medio para un predictor de orden *m*.

4.3 PROCEDIMIENTO DE ANÁLISIS LPC

A continuación se describe un procedimiento para obtener los coeficientes LPC mediante el método de la autocorrelación. La figura 4.4 muestra el diagrama de bloques para este procedimiento. Los bloques muestran las operaciones necesarias y a la izquierda de éstos se muestran los parámetros de entrada requeridos en cada paso. A partir del cuarto bloque, el ponderado por la ventana de Hamming, las operaciones se realizan sobre las tramas de la señal de voz. La consiguiente reducción en el número de parámetros resultantes para cada trama se ilustra mediante la cantidad de flechas.

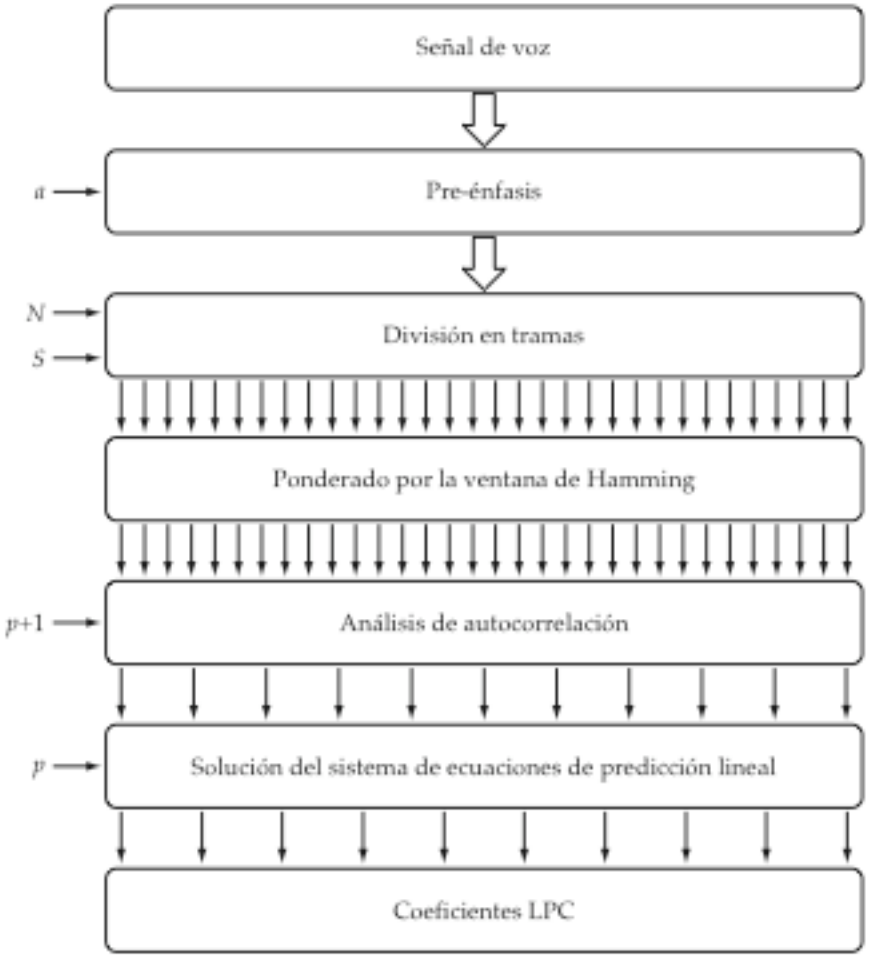


FIGURA 4.4 Procedimiento para la obtención de los coeficientes LPC

4.3.1 EL ORDEN DEL PREDICTOR

El orden del predictor p, determina la cantidad o el número de coeficientes del predictor LPC. Su selección depende primordialmente de la frecuencia de muestreo f_s, de forma independiente del método que se utilice. Existe una regla (thumb rule) que el espectro de voz a analizar puede representarse conteniendo una densidad promedio de un polo por cada 1000 Hertz, debido a la contribución del tracto vocal, pero además se necesitan de otros 3 a 4 polos extras para representar adecuadamente el espectro de la fuente de excitación y la carga de radiación. Así que para una frecuencia de muestreo de $f_s = 10{,}000\,\text{Hz}$, se necesitan por lo menos 10 polos para representar la respuesta del tracto vocal, pero un predictor de 13º o 14º orden es suficiente. Para verificar esta conclusión Atal y Hanauer simularon el comportamiento del error de predicción RMS normalizado, para diferentes valores del orden del predictor, para tramas de sonidos sonoros y sordos, a una frecuencia de muestreo $f_s = 10{,}000\,\text{Hz}$ [13].

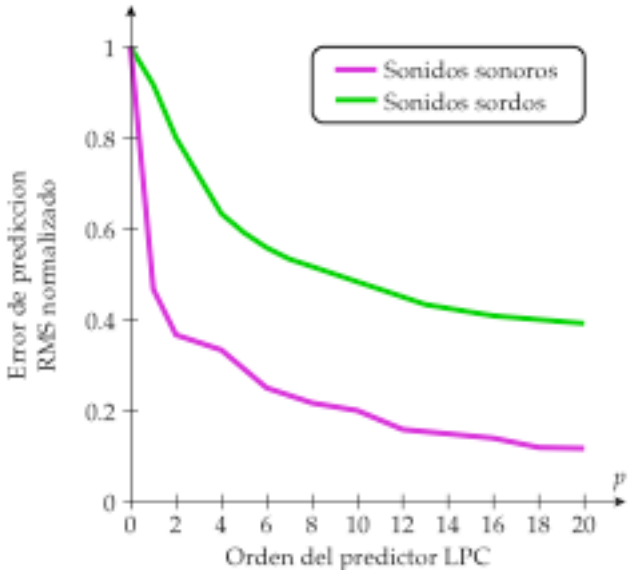


FIGURA 4.5 Variación en el error de predicción en función del orden del predictor LPC

Pese a que el error de predicción disminuye paulatinamente al incrementar el orden del predictor, para un orden de 13 a 14 coeficientes, el error prácticamente ya se ha estabilizado, pues solo presenta pequeñas disminuciones al seguir incrementando el orden [13]. Es interesante notar que a partir de esta figura, se puede concluir que el error de predicción, RMS normalizado, para los sonidos sordos es significativamente más grande que para los sonidos sonoros. Lo cual era de esperarse ya que el modelo LPC representa mucho mejor a los sonidos sonoros que a los sonidos sordos [13].

4.3.2 PRE-ÉNFASIS

La señal de voz digitalizada $x(n)$, se pasa a través del filtro FIR de pre-énfasis de primer orden, definido por la siguiente ecuación en diferencias:

$$x_p(n) = x(n) - ax(n-1) \qquad \begin{matrix} 0.9 \le a \le 1 \\ n = 0,1,2,3,\ldots,N_x - 1 \end{matrix} \qquad \textbf{(4.32)}$$

Un beneficio adicional del pre-énfasis es el disminuir los efectos de los errores numéricos, como los que pueden ocurrir durante la obtención de los coeficientes de autocorrelación, los cuales pueden afectar la estabilidad del filtro de síntesis LPC [13].

4.3.3 DIVISIÓN EN TRAMAS

La señal de voz se divide en L tramas de N muestras para su análisis, con S muestras de separación entre tramas adyacentes:

$$x_\ell(n) = x(n + \ell \cdot S) \qquad \begin{matrix} N \ge S \\ \ell = 0,1,2,3,\ldots,L-1 \\ n = 0,1,2,3,\ldots,N-1 \end{matrix} \qquad \textbf{(4.33)}$$

Para una frecuencia de muestreo de 10 kHz, se han estado usando tramas de 100 a 400 muestras (de 10 a 40 ms de duración), siendo la tendencia hacia los valores más altos en la mayoría de los sistemas [13].

La tabla 4.1 muestra los valores típicos para los sistemas de análisis LPC, para tres frecuencias de muestreo que se emplean comúnmente en el reconocimiento de voz [14].

TABLA 4.1 Valores típicos para los sistemas LPC

fs	*p*	*N*	*N*	*S*	*S*	*Fps*
[Hz]	**[orden]**	**[muestras]**	**[ms]**	**[muestras]**	**[ms]**	**[tramas / s]**
6667	8	300	45	100	15	66.67
8000	10	240	30	80	10	100
10,000	10	300	30	100	10	100

4.3.4 PONDERADO POR LA VENTANA DE HAMMING

El ponderado de la trama de análisis por la ventana de Hamming contribuye a minimizar las discontinuidades al inicio y al final de las tramas:

$$x_w(n) = x_\ell(n) \cdot w(n)$$
$$x_w(n) = x_\ell(n) \cdot \left[0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)\right] \qquad n = 0,1,2,3,\ldots,N-1 \qquad \textbf{(4.34)}$$

4.3.5 ANÁLISIS DE AUTOCORRELACIÓN

Se obtienen los primeros $p+1$ valores de los coeficientes de autocorrelación para la trama ponderada:

$$R(k) = \sum_{n=0}^{N-1-k} x_w(n) \cdot x_w(n+k) \qquad k = 0,1,2,3,\ldots,p \qquad \textbf{(4.35)}$$

Un beneficio adicional del análisis de autocorrelación es que el valor de autocorrelación cero $R(0)$, es igual a la energía de la trama, la cual es un parámetro importante en los sistemas de detección de voz [14].

4.3.6 SOLUCIÓN DEL SISTEMA DE ECUACIONES

El paso final es la solución del sistema de ecuaciones lineales:

$$\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & \cdots & R(p-2) \\ R(2) & R(1) & R(0) & \cdots & R(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{bmatrix} \qquad \textbf{(4.22)}$$

En el que los $p+1$ valores de los coeficientes de autocorrelación son convertidos en los parámetros LPC. El método formal para esta conversión es mediante el algoritmo de Levinson-Durbin. Por lo tanto, la solución final es el vector o el conjunto de coeficientes LPC:

$$\mathbf{V}_{lpc} = a_c$$
$$\mathbf{V}_{lpc} = \left[a_1, a_2, a_3, \ldots, a_p\right] \qquad c = 1,2,3,\ldots,p \qquad \textbf{(4.36)}$$

4.4 CARACTERÍSTICAS DEL ANÁLISIS LPC

La codificación de predicción lineal LPC es, esencialmente, una técnica eficiente de codificación en el dominio del tiempo, la cual permite que una trama de voz, de 100 a 300 muestras, sea representada por una cantidad mucho menor de p coeficientes, de 10 a 15. Lo que es más, los coeficientes del predictor pueden ser utilizados para estimar la respuesta del tracto vocal [14].

Existe una dependencia directa entre los coeficientes LPC y el orden de su predictor, esto es, para un mismo parámetro a_x se obtienen valores diferentes para ordenes distintos, es decir, que:

$$a_x^{(n)} \neq a_x^{(m)} \qquad \text{si} \qquad n \neq m \tag{4.37}$$

Por esta razón se denota el orden como un superíndice entre paréntesis y se lee de la siguiente manera: el coeficiente x de orden m [16].

La mayor ventaja del modelo es que los coeficientes LPC proporcionan una buena representación de la voz, y se pueden estimar de una forma directa, sencilla y computacionalmente eficiente. Además, la teoría detrás del método ha sido objeto de una investigación intensiva y como resultado, está muy desarrollada. Tomando como base esta teoría, han surgido una gran variedad de aplicaciones del análisis de predicción lineal para el procesamiento de voz.

El proceso de minimización del error cuadrático medio entre las muestras de la señal de voz y las muestras estimadas linealmente, tiende a producir una señal de error que tiene un espectro blanco, es decir, amplio y plano. Esto está en función de que tan bueno sea el predictor modelando la señal.

El modelo simplificado todos polos es una representación natural de los sonidos sonoros no nasales, pero para los sonidos nasales y fricativos, hacen falta tanto los ceros como los polos de la función de transferencia del tracto vocal. No obstante, si el orden del predictor es suficientemente alto, el modelo produce una buena representación para la mayoría de los sonidos de la voz. Para el caso particular de los sonidos sordos, éstos se pueden representar adecuadamente mediante un modelo de bajo orden.

Para un predictor ideal, la señal de error para los sonidos sonoros consistiría de un tren de impulsos a la frecuencia del tono fundamental (pitch), pero para el caso de los sonidos sordos, la señal de error sería semejante a un ruido blanco. Así que el sistema LPC se basa en el modelo fuente-filtro, como se muestra en la figura 4.6.

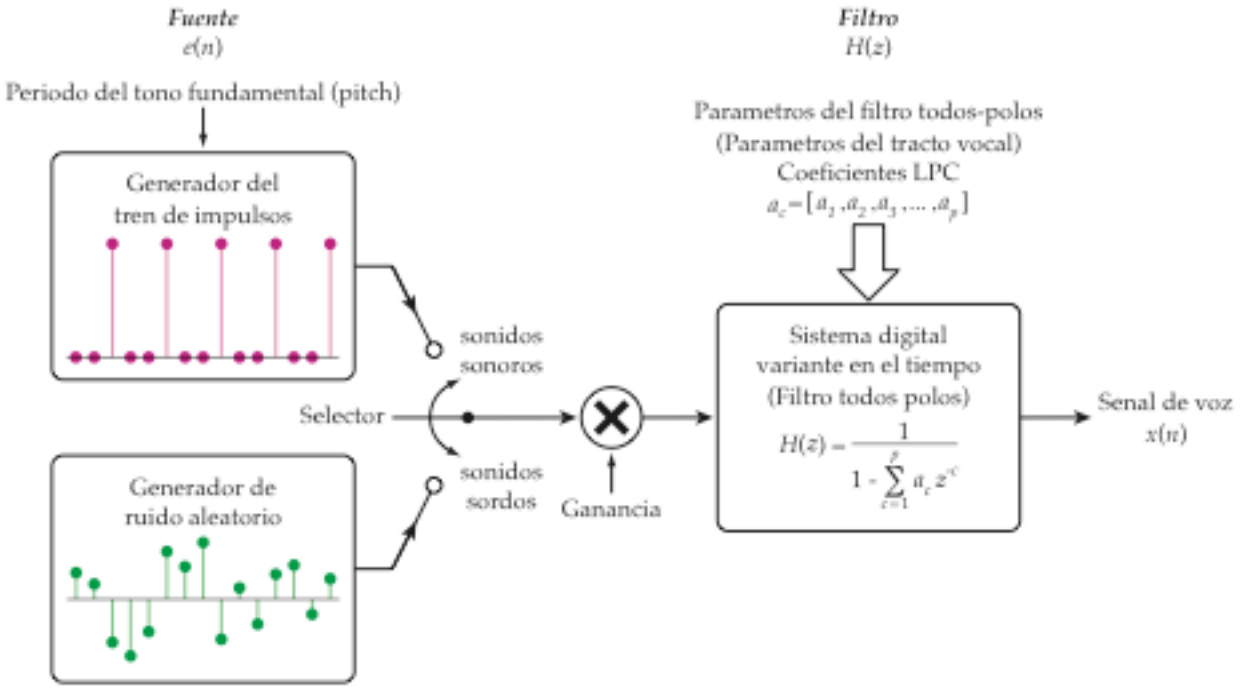


FIGURA 4.6 Modelo fuente-filtro LPC para la síntesis de voz

Los coeficientes LPC proporcionan un buen modelo para la señal de voz. Sin embargo, es menos efectivo para los sonidos sordos y las regiones de transición de la voz, pero de todos modos proporciona un modelo apropiado para propósitos del reconocimiento de voz [14].

CAPÍTULO

5

MFCC

CODIFICACIÓN O ANÁLISIS MEL CEPSTRAL

5.1 INTRODUCCIÓN

La motivación para investigar los métodos de análisis espectral, que se basan en la psicoacústica, es el obtener una mejor comprensión de la manera en la que se procesa la señal de voz en el sistema auditivo del humano, para poder diseñar e implementar métodos más robustos y eficientes para el análisis de la voz y su representación.

En 1980 Davis y Mermelstein propusieron un nuevo tipo de representación del cepstrum, en el cual combinaron sus ventajas con los conocimientos de la percepción no lineal de la frecuencia en el sistema auditivo humano, tomando como base los estudios realizados por Zwicker en 1961 [3]. En este método de análisis, se utiliza la parte real del cepstrum, pero con una transformación no lineal de la frecuencia, pasando de la escala en Hertz a la escala de frecuencia mel [9]. Por eso, a estos coeficientes se les ha denominado como los coeficientes mel cepstral o los **M**el **F**requency **C**epstral **C**oefficients (MFCC).

Así es que a partir de 1980, los MFCC comenzaron a substituir a los parámetros de predicción lineal, pues con estos otros parámetros se consiguió mejorar el desempeño de los sistemas de reconocimiento de voz [3]. Por consiguiente, la mayor parte de los sistemas han estado convergiendo hacia el uso de estos vectores, que se obtienen a partir de un procesamiento de la señal de voz empleando un banco de filtros que ha sido diseñado tomando en cuenta algunas propiedades de la percepción del sonido en el sistema auditivo humano, con lo que se busca imitar su comportamiento [5].

5.2 DEFINICIÓN DE LAS BANDAS CRÍTICAS TRIANGULARES CON LA ESCALA MEL

El diseño de los filtros para las bandas críticas primero se realiza en la escala mel. Los filtros utilizados en este método presentan una forma triangular, que es simétrica con respecto a su frecuencia central y tienen el mismo ancho de banda en la escala mel. Aunque a estas ventanas triangulares se les suele llamar filtros, estas simplemente se usan para promediar el espectro de potencia con respecto a la frecuencia central de cada una de las bandas críticas [9]. La función que las define está dada por la siguiente ecuación y su representación gráfica se muestra en la figura 5.1.

$$H_{bc}(f)=\begin{cases} 0 & \text{para} \quad f \le f_l \\ \dfrac{1}{(f_c-f_l)}(f-f_l) & \text{para} \quad f_l < f < f_c \\ 1 & \text{para} \quad f = f_c \\ \dfrac{-1}{(f_h-f_c)}(f-f_h) & \text{para} \quad f_c < f < f_h \\ 0 & \text{para} \quad f \ge f_h \end{cases} \qquad \textbf{(5.1)}$$

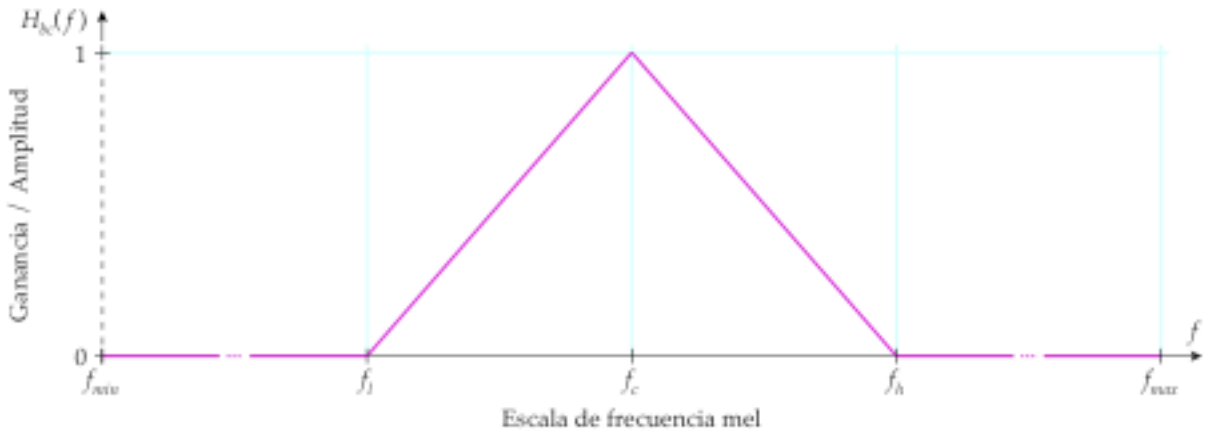


FIGURA 5.1 Curva para las bandas críticas triangulares

En la ecuación (5.1) se puede apreciar que los primeros términos (los cocientes) son las pendientes de la rectas que forman la ventana triangular, y los otros términos del producto contienen a la variable independiente con su respectiva abscisa al origen.

Ahora bien, la transformación no lineal de la frecuencia en Hertz a la escala de frecuencia mel, se acostumbra realizar mediante la siguiente función de aproximación:

$$f_{mel}(f) = 1125\,\mathbf{Ln}\left(1 + \frac{f}{700}\right) \qquad \textbf{(5.2)}$$

El rango de frecuencias de interés abarca desde los 0 Hz hasta la frecuencia de Nyquist (la mitad de la frecuencia de muestreo), sin embargo, se puede determinar un límite inferior de frecuencia f_{min}, y un límite superior de frecuencia f_{max}. Para la señal de voz, se recomienda que $f_{min} > 100\,\text{Hz}$ y que el límite superior de frecuencia sea menor a la frecuencia de Nyquist, aunque para una frecuencia mayor de 6800 Hz ya no hay mucha información en la señal de voz. Además, si se fija el límite inferior de frecuencia arriba de 50 o 60 Hz, se puede evitar el ruido procedente de la red de suministro o distribución de ca (ac) [10]. Ahora bien, puesto que estos límites están definidos en Hertz, hace falta transformarlos a la escala mel.

$$(f_{min})_{mel} = 1125\,\mathbf{Ln}\left(1 + \frac{f_{min}}{700}\right) \qquad \textbf{(5.3)}$$

$$(f_{max})_{mel} = 1125\,\mathbf{Ln}\left(1 + \frac{f_{max}}{700}\right) \qquad \textbf{(5.4)}$$

Las frecuencias centrales deben estar equiespaciadas en la escala mel, entonces se puede calcular la separación en frecuencia *SF* entre éstas mediante la siguiente fórmula:

$$SF = \frac{(f_{max})_{mel} - (f_{min})_{mel}}{BC + 1} \qquad \textbf{(5.5)}$$

En donde *BC* es la cantidad o el número de filtros o bandas críticas en el banco de filtros. Las frecuencias centrales están separadas por múltiplos enteros de este parámetro más el valor del límite inferior de frecuencia:

$$f_c(bc)_{mel} = bc \cdot SF + (f_{min})_{mel} \qquad bc = 1,2,3,\ldots,BC \qquad \textbf{(5.6)}$$

En donde *bc* denota el número correspondiente de la banda crítica. Ahora sólo hace falta transformar las frecuencias centrales de la escala mel a la escala en Hertz mediante la siguiente fórmula, que es la función inversa de la ecuación (5.2):

$$f_{Hz}(f_{mel}) = 700\left[\boldsymbol{e}^{\left(\frac{f_{mel}}{1125}\right)} - 1\right] \qquad \textbf{(5.7)}$$

Con estos valores de las frecuencias centrales en Hertz, ya se pueden determinar las frecuencias inferior y superior para cada una de las bandas críticas. La frecuencia inferior para una banda crítica *bc* es igual a la frecuencia central de la banda crítica anterior:

$$f_l(bc) = f_c(bc-1) \qquad \textbf{(5.8)}$$

La frecuencia superior para una banda crítica *bc* es igual a la frecuencia central de la siguiente banda crítica:

$$f_h(bc) = f_c(bc+1) \qquad \textbf{(5.9)}$$

La frecuencia inferior de la primera banda crítica es igual al límite inferior de frecuencia:

$$f_l(1) = f_{min} \qquad \textbf{(5.10)}$$

De manera similar, la frecuencia superior de la última banda crítica es igual al límite superior de frecuencia.

$$f_h(BC) = f_{max} \qquad \textbf{(5.11)}$$

La tabla 5.1 muestra los valores para las frecuencias inferior, central y superior para un banco de $BC = 17$ filtros, tanto en la escala mel como en la escala en Hertz. El límite inferior de frecuencia es $f_{min} = 0$ Hz y el límite superior de frecuencia es $f_{max} = 4000$ Hz.

TABLA 5.1 Valores de las frecuencias para las bandas críticas triangulares

bc [banda c.]	*fl* [mel]	*fc* [mel]	*fh* [mel]	*fl* [Hz]	*fc* [Hz]	*fh* [Hz]
1	0.000	119.015	238.030	0.00	78.11	164.94
2	119.015	238.030	357.045	78.11	164.94	261.46
3	238.030	357.045	476.059	164.94	261.46	368.75
4	357.045	476.059	595.074	261.46	368.75	488.01
5	476.059	595.074	714.089	368.75	488.01	620.58
6	595.074	714.089	833.104	488.01	620.58	767.94
7	714.089	833.104	952.119	620.58	767.94	931.75
8	833.104	952.119	1071.134	767.94	931.75	1113.84
9	952.119	1071.134	1190.148	931.75	1113.84	1316.24
10	1071.134	1190.148	1309.163	1113.84	1316.24	1541.23
11	1190.148	1309.163	1428.178	1316.24	1541.23	1791.33
12	1309.163	1428.178	1547.193	1541.23	1791.33	2069.34
13	1428.178	1547.193	1666.208	1791.33	2069.34	2378.37
14	1547.193	1666.208	1785.223	2069.34	2378.37	2721.88
15	1666.208	1785.223	1904.237	2378.37	2721.88	3103.72
16	1785.223	1904.237	2023.252	2721.88	3103.72	3528.18
17	1904.237	2023.252	2142.267	3103.72	3528.18	4000.00

En la figura 5.2 se muestra la representación gráfica, en la escala mel, del banco de filtros triangulares mostrado en la tabla 5.1. Las bandas críticas están equiespaciadas y todas las bandas críticas tienen el mismo ancho de banda.

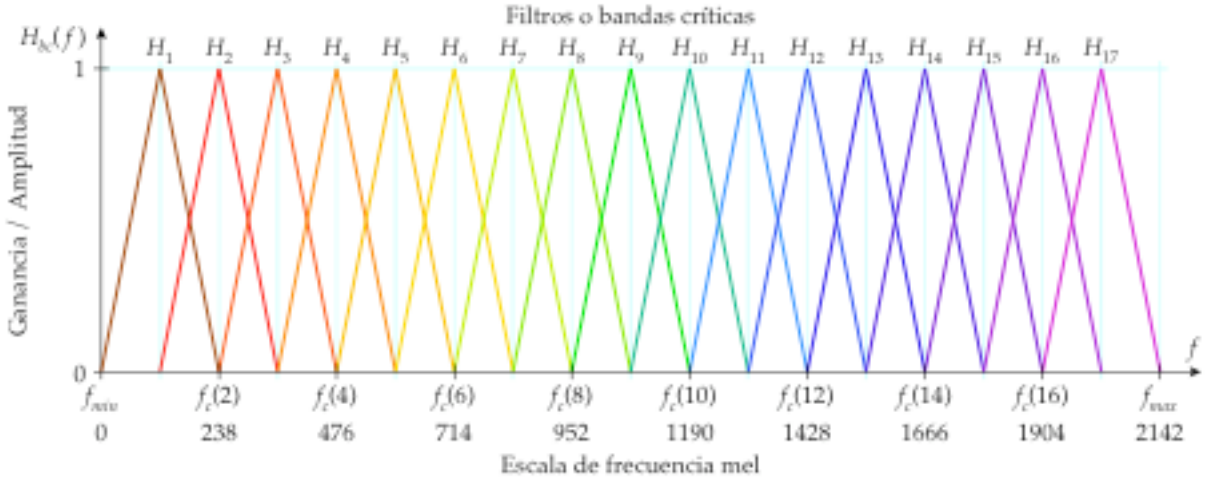


FIGURA 5.2 Banco de filtros o ventanas triangulares equiespaciados en la escala mel

La figura 5.3 muestra el mismo banco de filtros pero ahora en la escala en Hertz. Puede apreciarse como la distribución de las bandas críticas ahora sigue una distribución logarítmica y también como se incrementa el ancho de banda al aumentar la frecuencia.

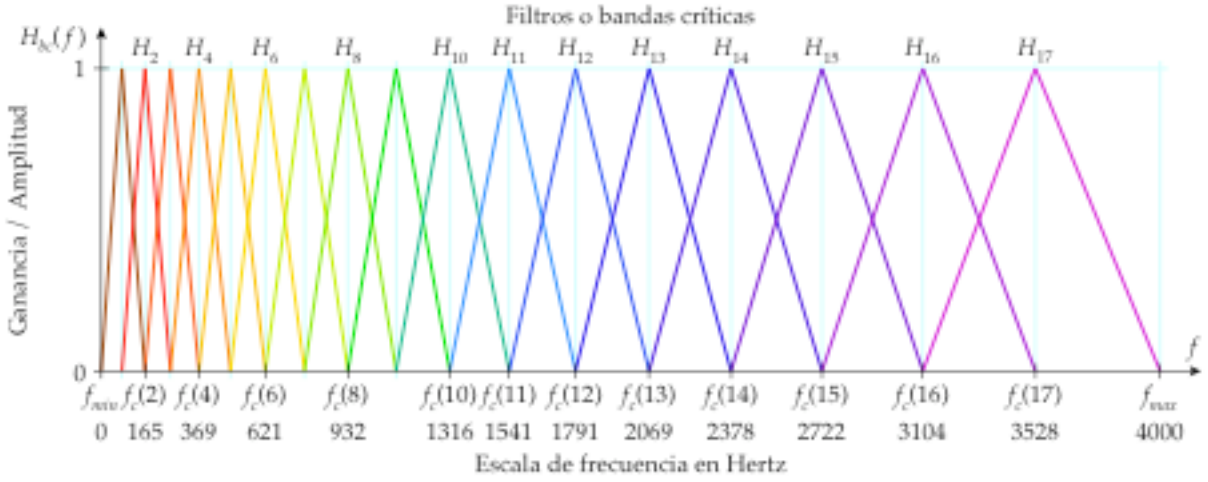


FIGURA 5.3 Banco de filtros o ventanas triangulares en Hertz con distribución logarítmica

Finalmente, sólo hace falta hacer un muestreo de las bandas críticas de acuerdo al número de puntos del espectro de potencia N_p. Ahora bien, como las bandas críticas contienen unos cuantos valores diferentes de cero, no hace falta calcular todos los valores correspondientes a cada uno de los puntos del espectro de potencia, basta con las muestras comprendidas entre la frecuencia inferior y la superior para cada banda crítica.

A cada muestra del espectro de potencia $P(k)$ le corresponde un valor de frecuencia en Hertz, el cual es un múltiplo de la resolución en frecuencia del espectro:

$$f = k \cdot \Delta f \qquad k = 0,1,2,3,\ldots,N_p \tag{5.12}$$

La resolución en frecuencia del espectro Δf, se puede calcular mediante el cociente de la frecuencia de muestreo f_s, entre el número de puntos para la transformada rápida de Fourier N_{fft}:

$$\Delta f = \frac{f_s}{N_{fft}} \tag{5.13}$$

Si se invierte este cociente, el nuevo factor es la resolución por muestra del espectro Δk:

$$\Delta k = \frac{N_{fft}}{f_s} \tag{5.14}$$

El cual se puede utilizar para determinar que índice o muestra corresponde a determinada frecuencia, al multiplicar esa frecuencia por este factor de equivalencia:

$$k = \left(f \cdot \Delta k\right)_{\mathbf{int}} \tag{5.15}$$

Puesto que los índices de las muestras deben ser números enteros, se tiene que redondear el resultado a un número entero. El índice para la frecuencia inferior k_l se puede redondear hacia arriba, pero el índice para la frecuencia superior k_h se puede redondear hacia abajo. También, si es necesario, hay que tomar en cuenta que la primera muestra del espectro de potencia representa a la frecuencia cero.

Con los índices correspondientes a las frecuencias inferior k_l y superior k_h, ya se pueden calcular los N_{bc} valores diferentes de cero para cada una de las bandas críticas.

$$N_{bc} = k_h(bc) - k_l(bc) + 1 \tag{5.16}$$

Para esto, primero se obtiene el valor correspondiente en frecuencia para la muestra k, comprendida dentro del intervalo $k_l(bc) \le k \le k_h(bc)$ de una banda crítica bc:

$$f = k \cdot \Delta f \tag{5.17}$$

Entonces, el valor correspondiente a la muestra k para el filtro o la banda crítica H_{bc} se determina mediante la ecuación (5.1), según sea el caso.

En la tabla 5.2 se muestran los índices correspondientes a las frecuencias inferior y superior para las $BC = 17$ bandas críticas de la tabla 5.1. El número de puntos para la transformada rápida de Fourier es de $N_{fft} = 256$ puntos y la frecuencia de muestreo es de $f_s = 8000$ Hz. Por lo tanto, la resolución en frecuencia es de $\Delta f = 31.25$ Hz/muestra pero la resolución por muestra es de $\Delta k = 0.032$ muestras/Hz.

TABLA 5.2 Correspondencia entre índices y frecuencias para las bandas críticas mel

bc	*fl*	*kl*	*(kl) int*	*fh*	*kh*	*(kh) int*	*Nbc*
[banda c.]	**[Hz]**	**[índice]**	**[índice]**	**[Hz]**	**[índice]**	**[índice]**	**[muestras]**
1	0.00	0.00	0	164.94	5.28	5	6
2	78.11	2.50	3	261.46	8.37	8	6
3	164.94	5.28	6	368.75	11.80	11	6
4	261.46	8.37	9	488.01	15.62	15	7
5	368.75	11.80	12	620.58	19.86	19	8
6	488.01	15.62	16	767.94	24.57	24	9
7	620.58	19.86	20	931.75	29.82	29	10
8	767.94	24.57	25	1113.84	35.64	35	11
9	931.75	29.82	30	1316.24	42.12	42	13
10	1113.84	35.64	36	1541.23	49.32	49	14
11	1316.24	42.12	43	1791.33	57.32	57	15
12	1541.23	49.32	50	2069.34	66.22	66	17
13	1791.33	57.32	58	2378.37	76.11	76	19
14	2069.34	66.22	67	2721.88	87.10	87	21
15	2378.37	76.11	77	3103.72	99.32	99	23
16	2721.88	87.10	88	3528.18	112.90	112	25
17	3103.72	99.32	100	4000.00	128.00	128	29

La tabla 5.3 muestra la correspondencia entre los índices y las frecuencias para los puntos que definen la octava banda crítica mel H_8. Se muestran los índices k para las muestras del espectro de potencia y sus valores correspondientes de frecuencia en Hertz. La resolución en frecuencia es $\Delta f = 31.25$ Hz/muestra. Para fines ilustrativos, se incluyen los valores exactos de los tres límites para los cinco casos de la ecuación (5.1).

TABLA 5.3 Correspondencia entre los valores para la octava banda crítica mel

k	*f*	**CASO**	*H8(k)*
[índice]	**[Hz]**		
24	750.00	$1 : f \le fl$	0
kl(8)	***767.94***	$\mathbf{1 : f \le fl}$	***0***
25	781.25	$2 : fl < f < fc$	0.0812
26	812.50	$2 : fl < f < fc$	0.2720
27	843.75	$2 : fl < f < fc$	0.4628
28	875.00	$2 : fl < f < fc$	0.6536
29	906.25	$2 : fl < f < fc$	0.8443
kc(8)	***931.75***	$\mathbf{3 : f = fc}$	***1***
30	937.50	$4 : fc < f < fh$	0.9684
31	968.75	$4 : fc < f < fh$	0.7968
32	1,000.00	$4 : fc < f < fh$	0.6252
33	1,031.25	$4 : fc < f < fh$	0.4536
34	1,062.50	$4 : fc < f < fh$	0.2819
35	1,093.75	$4 : fc < f < fh$	0.1103
kh(8)	***1,113.84***	$\mathbf{5 : f \ge fh}$	***0***
37	1,156.25	$5 : f \ge fh$	0

Hay otra función que se puede emplear para definir estas ventanas triangulares:

$$H_{bc}(f)=\begin{cases}\left(\dfrac{2}{f_h-f_l}\right)\dfrac{1}{(f_c-f_l)}(f-f_l) & \text{para} \quad f_l \le f < f_c \\ 1 & \text{para} \quad f = f_c \\ \left(\dfrac{2}{f_h-f_l}\right)\dfrac{-1}{(f_h-f_c)}(f-f_h) & \text{para} \quad f_c < f \le f_h\end{cases} \qquad \textbf{(5.18)}$$

A el banco de filtros obtenido mediante la ecuación anterior se le denomina normalizado, debido al factor de escala adicional que modifica la amplitud de la ventana triangular o la ganancia del filtro, como se muestra en la figura 5.4 .

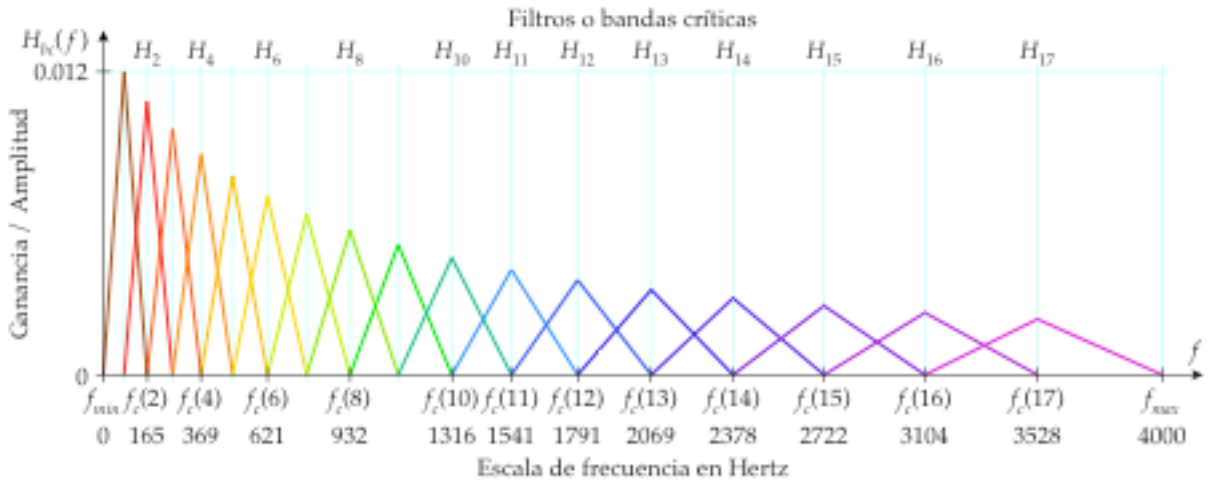


FIGURA 5.4 Banco de filtros o ventanas triangulares normalizados

La diferencia entre los MFCC obtenidos mediante la ecuación (5.1) y los obtenidos con la ecuación (5.18) es un vector constante para todas las tramas, de tal manera que no importa cuál de las dos ecuaciones se elija, siempre y cuando se utilicen los mismos filtros tanto en el entrenamiento como en el sistema de reconocimiento de voz [9].

5.3 PROCEDIMIENTO DE ANÁLISIS MFCC

A continuación se describe un procedimiento para obtener los coeficientes MFCC. La figura 5.5 muestra el diagrama de bloques para este procedimiento. Los bloques muestran las operaciones necesarias y a la izquierda de éstos se muestran los parámetros de entrada requeridos en cada paso. A partir del cuarto bloque, el ponderado por la ventana de Hamming, las operaciones se realizan sobre las tramas de la señal de voz. La consiguiente reducción en el número de parámetros para cada trama se ilustra mediante la cantidad de flechas.

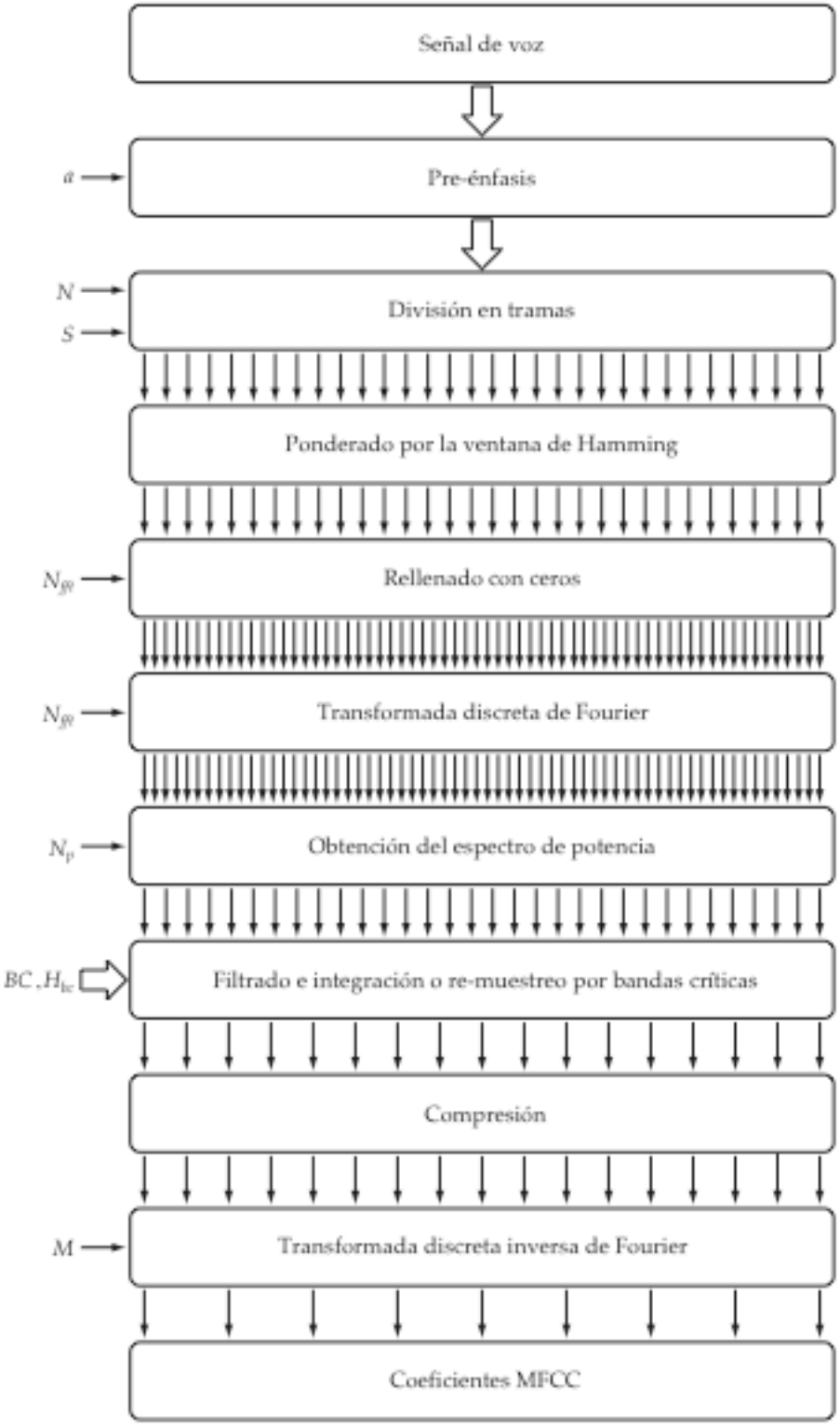


FIGURA 5.5 Procedimiento para la obtención de los coeficientes MFCC

5.3.1 EL NÚMERO DE COEFICIENTES Y DE FILTROS

En cuanto al número de filtros o bandas críticas *BC*, para una frecuencia de muestreo de $f_s = 8000$ Hz (un ancho de banda de 4000 Hz) se emplean hasta 24 filtros [2]. Pero normalmente se usan sólo $BC = 20$ bandas críticas [11], aunque en algunos casos se han llegado a usar hasta $BC = 40$ [9].

Los MFCC se calculan para una cantidad de coeficientes M que es menor que el número de filtros o bandas críticas *BC*. Por ejemplo, en los sistemas de reconocimiento de voz, la cantidad de coeficientes que se utilizan siempre es menor de $M = 15$ [1]. Por lo general solamente se usan los primeros $M = 13$ coeficientes obtenidos a partir de 20 o 24 filtros o bandas críticas [2].

5.3.2 PRE-ÉNFASIS

La señal de voz digitalizada $x(n)$, se pasa a través del filtro FIR de pre-énfasis de primer orden, definido por la siguiente ecuación en diferencias:

$$x_p(n) = x(n) - ax(n-1) \qquad \begin{matrix} 0.9 \le a \le 1 \\ n = 0,1,2,3,\ldots,N_x - 1 \end{matrix} \qquad \textbf{(5.19)}$$

5.3.3 DIVISIÓN EN TRAMAS

La señal de voz se divide en L tramas de N muestras para su análisis, con S muestras de separación entre tramas adyacentes:

$$x_\ell(n) = x(n + \ell \cdot S) \qquad \begin{matrix} N \ge S \\ \ell = 0,1,2,3,\ldots,L-1 \\ n = 0,1,2,3,\ldots,N-1 \end{matrix} \qquad \textbf{(5.20)}$$

5.3.4 PONDERADO POR LA VENTANA DE HAMMING

El ponderado de la trama de análisis por la ventana de Hamming, contribuye a minimizar las discontinuidades al inicio y al final de las tramas:

$$x_w(n) = x_\ell(n) \cdot w(n)$$

$$x_w(n) = x_\ell(n) \cdot \left[0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)\right] \qquad n = 0,1,2,3,\ldots,N-1 \qquad \textbf{(5.21)}$$

5.3.5 RELLENADO CON CEROS

Se añaden las muestras nulas que hagan falta para que la longitud de la trama de análisis sea igual a N_{fft}, la potencia de dos más próxima:

$$N_{fft} = 2^{\left[\log_2(N)\right]_{\text{int}}} \tag{5.22}$$

$$x_{zp}(n) = \begin{cases} x_w(n) & \text{para} \quad n = 0,1,2,3,\ldots,N-1 \\ 0 & \text{para} \quad n = N,N+1,\ldots,N_{fft}-1 \end{cases} \tag{5.23}$$

5.3.6 TRANSFORMADA DISCRETA DE FOURIER

La transformada discreta de Fourier de la trama de análisis se calcula mediante la transformada rápida de Fourier de N_{fft} puntos:

$$X(k) = \mathbf{FFT}\left\{x_{zp}(n), N_{fft}\right\} \qquad \begin{matrix} k = 0,1,2,3,\ldots,N_{fft}-1 \\ n = 0,1,2,3,\ldots,N_{fft}-1 \end{matrix} \tag{5.24}$$

5.3.7 OBTENCIÓN DEL ESPECTRO DE POTENCIA

El espectro de potencia es igual a la suma de los cuadrados de la parte real e imaginaria del espectro en frecuencia de la trama de análisis:

$$N_p = 0.5 N_{fft} \tag{5.25}$$

$$P(k) = \mathbf{real}\left[X(k)\right]^2 + \mathbf{imag}\left[X(k)\right]^2 \qquad k = 0,1,2,3,\ldots,N_p \tag{5.26}$$

5.3.8 FILTRADO E INTEGRACIÓN O RE-MUESTREO POR BANDAS CRÍTICAS

Para extraer las características de la envolvente del espectro, se emplea un banco de filtros o ventanas triangulares traslapadas, con sus frecuencias centrales equiespaciadas en la escala mel. La integración por bandas criticas se realiza mediante la sumatoria de los productos de las muestras del espectro de potencia $P(k)$ por cada uno de los filtros H_{bc} para las bandas críticas:

$$P_s(bc) = \sum_{k=k_l(bc)}^{k_h(bc)} P(k) \cdot H_{bc}(k) \qquad bc = 1,2,3,\ldots,BC \tag{5.27}$$

5.3.9 COMPRESIÓN

En lugar de simplemente usar la magnitud del espectro re-muestreado mediante la escala mel para calcular los MFCC, algunos investigadores han sugerido que es más provechoso el emplear una función logaritmo para calcular la energía total de las bandas críticas alrededor de las frecuencias centrales:

$$P_{sc}(bc) = \mathbf{log}\left[P_s(bc)\right] \qquad bc = 1,2,3,\ldots,BC \qquad \textbf{(5.28)}$$

El procesamiento de la magnitud y la obtención de su logaritmo también se realizan en el oído humano. Lo que es más, el empleo de la magnitud descarta la información innecesaria que proporciona la fase mientras que el logaritmo efectúa una compresión, dando como resultado que la extracción de características sea menos sensible a las variaciones de la amplitud producidas por las resonancias espectrales.

5.3.10 TRANSFORMADA INVERSA DE FOURIER

El último paso en el cálculo de los MFCC consiste en realizar la transformada discreta inversa de Fourier del espectro muestreado y comprimido P_{sc}. En este paso es donde se obtienen los MFCC. Esta etapa del procedimiento tiene grandes ventajas. Primero, puesto que los valores del espectro son números reales y presentan una simetría par con respecto a la frecuencia de Nyquist, entonces se puede reducir a una transformada coseno discreta (DCT) en lugar de la transformada discreta inversa de Fourier:

$$c_m = \sum_{bc=1}^{BC} P_{sc}(bc)\,\mathbf{cos}\left[\frac{\pi}{BC}\left(bc - \frac{1}{2}\right)m\right] \qquad m = 0,1,2,3,\ldots,M \qquad \textbf{(5.29)}$$

La transformada coseno discreta tiene la propiedad de producir características que prácticamente no están correlacionadas entre sí. De tal manera que se pueden usar matrices diagonales de covarianza en lugar de las matrices ordinarias de covarianza, con lo que se reduce en gran manera el trabajo de computo y el número de parámetros a estimar [1]. Además la transformada coseno discreta proporciona un suavizado del espectro si solamente se conservan los primeros coeficientes. Por lo tanto, la solución final es el vector o el conjunto de coeficientes MFCC:

$$\begin{aligned} \mathbf{V}_{mfcc} &= c_m \\ \mathbf{V}_{mfcc} &= \left[c_1, c_2, c_3, \ldots, c_M\right] \end{aligned} \qquad m = 1,2,3,\ldots,M \qquad \textbf{(5.30)}$$

El coeficiente inicial c_0 representa el logaritmo de la energía promedio en la trama y con frecuencia se descarta [11], pues ésta se puede calcular directamente a partir de la señal de voz [1].

5.4 CARACTERÍSTICAS DEL ANÁLISIS MFCC

La extracción de características confiables es uno de los principales problemas en el área del reconocimiento de voz. En general, las características en el dominio de la frecuencia, como los coeficientes MFCC, se desempeñan mejor que las características en el dominio del tiempo. Las mejoras obtenidas con los sistemas de reconocimiento usando estos coeficientes indican que las representaciones motivadas por la percepción del sonido contribuyen a mejorar el reconocimiento de voz [9]. En general, con estos coeficientes se han obtenido mejores resultados con respecto a otras técnicas alternativas de procesamiento, pero además, mediante una complejidad computacional menor[1].

Se ha estudiado experimentalmente el efecto de la cantidad de coeficientes, ahora bien, no se consiguió disminuir el error con más de $M = 13$ coeficientes, lo que significa que estos primeros coeficientes son los que portan la información más relevante para el reconocimiento [9].

Los coeficientes MFCC se han consolidado como los vectores de características básicos para muchos problemas de reconocimiento de patrones, tanto de los acústicos como de los de voz. Por eso, es de interés seguir buscando nuevas formas eficientes para calcularlos [2].

CAPÍTULO

6

PLP

CODIFICACIÓN DE PREDICCIÓN LINEAL PERCEPTUAL

6.1 INTRODUCCIÓN

Los bancos de filtros, el análisis de predicción lineal y el análisis cepstral son unas herramientas muy útiles para el análisis de la voz y su representación. Por esta razón, tanto los investigadores como los desarrolladores de sistemas las han estado utilizado ampliamente. No obstante, existen otras representaciones que han sido desarrolladas específicamente para alguna de las áreas del procesamiento digital de voz, particularmente para el área del reconocimiento de voz [5].

Ahora bien, Hynek Hermansky estudió algunas técnicas de transformación del espectro de la voz, con el objetivo de imitar la dinámica no lineal del oído interno, que determina la percepción humana de la voz, estimando así un espectro auditivo o perceptual, que posteriormente es aproximado por los coeficientes de predicción lineal [7], los cuales son una representación compacta de la forma del espectro auditivo en términos de sus polos [6]. Por lo tanto, a esta técnica de codificación se le ha denominado predicción lineal perceptual o **P**erceptual **L**inear **P**rediction (PLP) [6].

La implementación de las aproximaciones de tres conceptos bien conocidos de la psicoacústica permiten un medio diferente y generalmente más comprensible para representar la compleja señal de voz. Los tres conceptos que se emplean en la estimación del espectro auditivo PLP son las siguientes [7]:

1. La resolución espectral por bandas críticas.
2. Las curvas isofónicas o de igual nivel de sonoridad.
3. La ley de potencia de la audición.

6.2 DEFINICIÓN DE LAS BANDAS CRÍTICAS TRAPEZOIDALES CON LA ESCALA BARK

El diseño de los filtros para las bandas críticas primero se realiza en la escala Bark. Los filtros utilizados en este método presentan una forma trapezoidal, que no es simétrica con respecto a su frecuencia central, pero tienen el mismo ancho de banda de 1 Bark. Estos filtros son planos en su parte superior pero con caídas exponenciales distintas, siendo una más pronunciada que la otra y están truncadas a –40 dB. Esta ventana trapezoidal es una aproximación de lo que se conoce acerca de la forma de los filtros auditivos, se basa en la curva de enmascaramiento asimétrica propuesta por Schroeder y aprovecha la propuesta de Zwicker de que la forma de los filtros auditivos es aproximadamente constante en la escala Bark [7]. La función que las define está dada por la ecuación (6.2), la cual está en términos de la diferencia de frecuencias Bark f_d, y su representación gráfica se muestra en la figura 6.1.

$$f_d = (f)_{Bark} - [f_c(bc)]_{Bark} \tag{6.1}$$

$$H_{bc}(f_d) = \begin{cases} 0 & \text{para} \quad f_d < -2.5 \\ 10^{(f_d+0.5)} & \text{para} \quad -2.5 \le f_d < -0.5 \\ 1 & \text{para} \quad -0.5 \le f_d \le 0.5 \\ 10^{-2.5(f_d-0.5)} & \text{para} \quad 0.5 < f_d \le 1.3 \\ 0 & \text{para} \quad f_d > 1.3 \end{cases} \tag{6.2}$$

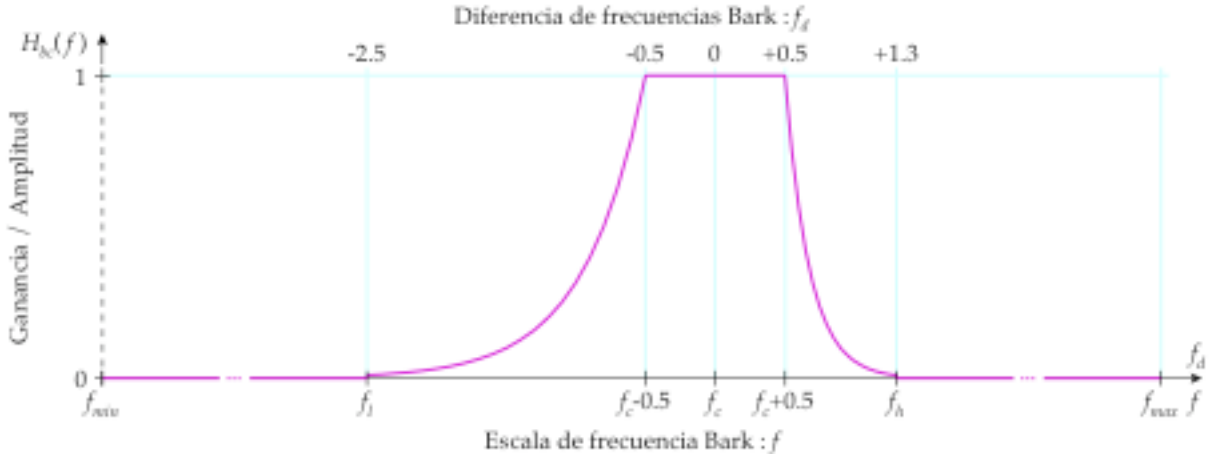


FIGURA 6.1 Curva para las bandas críticas trapezoidales

Ahora bien, la transformación no lineal de la frecuencia en Hertz a la escala de frecuencia Bark, se puede realizar mediante la función del arqueado de la frecuencia de Schroeder [7]:

$$f_{Bark}(f) = 6\,\mathbf{Ln}\left[\left(\frac{f}{600}\right) + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right] \qquad \textbf{(6.3)}$$

El rango de frecuencias de interés abarca desde los 0 Hz hasta la frecuencia de Nyquist (la mitad de la frecuencia de muestreo), pero como este rango esta definido en Hertz, hace falta transformarlo a la escala Bark:

$$(f_{max})_{Bark} = 6\,\mathbf{Ln}\left[\left(\frac{0.5 f_s}{600}\right) + \sqrt{\left(\frac{0.5 f_s}{600}\right)^2 + 1}\right] \qquad \textbf{(6.4)}$$

Las frecuencias centrales deben estar equiespaciadas en la escala Bark, entonces se puede calcular la separación en frecuencia *SF* entre éstas mediante la siguiente fórmula:

$$SF = \frac{(f_{max})_{Bark}}{BC + 1} \qquad \textbf{(6.5)}$$

En donde *BC* es la cantidad o el número de filtros o bandas críticas en el banco de filtros. Las frecuencias centrales están separadas por múltiplos enteros de este parámetro:

$$f_c(bc)_{Bark} = bc \cdot SF \qquad bc = 1,2,3,\ldots,BC \qquad \textbf{(6.6)}$$

En donde *bc* denota el número correspondiente de la banda crítica. Con las frecuencias centrales definidas en la escala Bark ya se pueden calcular las frecuencias inferior y superior para cada una de las bandas críticas. La frecuencia inferior para una banda crítica *bc* es igual a su frecuencia central menos 2.5 Barks:

$$f_l(bc)_{Bark} = f_c(bc)_{Bark} - 2.5 \qquad \textbf{(6.7)}$$

La frecuencia superior para una banda crítica *bc* es igual a su frecuencia central más 1.3 Barks:

$$f_h(bc)_{Bark} = f_c(bc)_{Bark} + 1.3 \qquad \textbf{(6.8)}$$

Ahora sólo hace falta transformar estas frecuencias de la escala Bark a la escala en Hertz mediante la siguiente fórmula, que es la función inversa de la ecuación (6.3):

$$f_{Hz}(f_{Bark}) = 300\left[e^{\left(\frac{f_{Bark}}{6}\right)} - e^{-\left(\frac{f_{Bark}}{6}\right)}\right] \qquad \textbf{(6.9)}$$

La tabla 6.1 muestra los valores para las frecuencias inferior, central y superior para un banco de $BC = 15$ filtros, tanto en la escala Bark como en la escala en Hertz. Los 15 filtros corresponden a un ancho de banda o una frecuencia de Nyquist de $f_N = 4000$ Hz.

TABLA 6.1 Valores de las frecuencias para las bandas críticas trapezoidales

bc	*fl*	*fc*	*fh*	*fl*	*fc*	*fh*
[banda c.]	**[Bark]**	**[Bark]**	**[Bark]**	**[Hz]**	**[Hz]**	**[Hz]**
1	-1.527	0.973	2.273	-154.31	97.77	232.82
2	-0.553	1.947	3.247	-55.39	198.12	340.77
3	0.420	2.920	4.220	42.07	303.70	457.70
4	1.394	3.894	5.194	140.63	417.29	586.71
5	2.367	4.867	6.167	242.91	541.89	731.20
6	3.341	5.841	7.141	351.59	680.78	894.98
7	4.314	6.814	8.114	469.55	837.63	1,082.36
8	5.288	7.788	9.088	599.90	1,016.58	1,298.30
9	6.261	8.761	10.061	746.07	1,222.34	1,548.48
10	7.234	9.734	11.034	911.92	1,460.35	1,839.52
11	8.208	10.708	12.008	1,101.83	1,736.88	2,179.08
12	9.181	11.681	12.981	1,320.81	2,059.23	2,576.12
13	10.155	12.655	13.955	1,574.62	2,435.90	3,041.12
14	11.128	13.628	14.928	1,869.98	2,876.83	3,586.34
15	12.102	14.602	15.902	2,214.67	3,393.66	4,226.18

En la tabla anterior, se puede apreciar que las frecuencias inferiores para las primeras bandas críticas pueden tomar valores menores que cero, si este es el caso, se tiene que truncar su frecuencia inferior para esta que sea igual a cero.

$$f_l(bc) \geq 0 \qquad \textbf{(6.10)}$$

De manera similar, las frecuencias superiores para las últimas bandas críticas pueden tomar valores mayores que la frecuencia de Nyquist, si este es el caso, se tiene que truncar su frecuencia superior para que esta sea igual al límite superior de frecuencia.

$$f_h(bc) \leq f_{max} \qquad \textbf{(6.11)}$$

La figura 6.2 muestra la representación gráfica, en la escala Bark, del banco de filtros trapezoidales mostrado en la tabla 6.1. Las bandas críticas están equiespaciadas y todas las bandas críticas tienen el mismo ancho de banda.

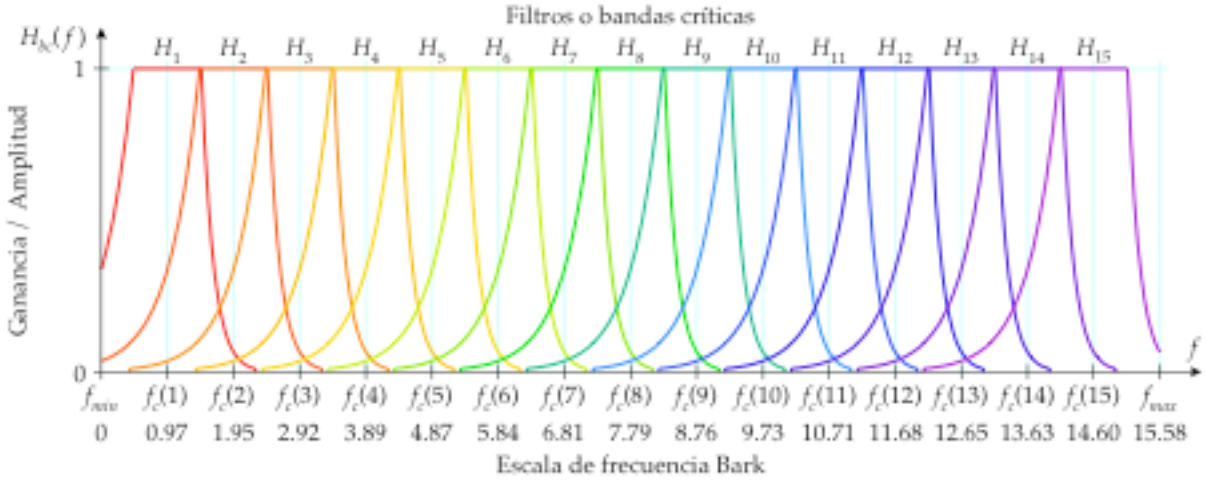


FIGURA 6.2 Banco de filtros o ventanas trapezoidales equiespaciadas en la escala Bark

La figura 6.3 muestra el mismo banco de filtros pero ahora en la escala en Hertz. Puede apreciarse como la distribución de las bandas críticas ahora sigue una distribución logarítmica y también como se incrementa el ancho de banda al aumentar la frecuencia.

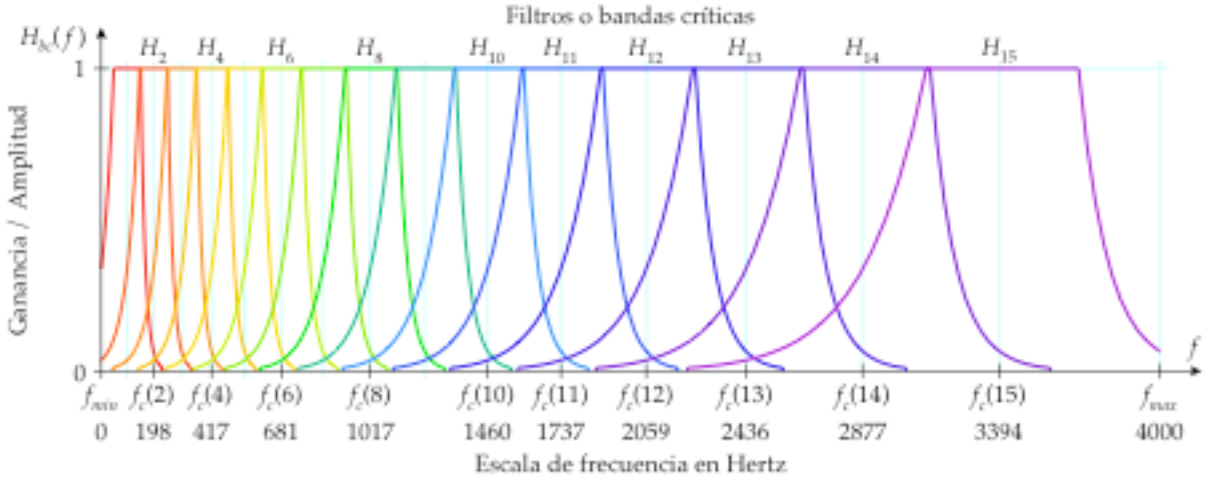


FIGURA 6.3 Banco de filtros o ventanas trapezoidales en Hertz con distribución logarítmica

Finalmente, sólo hace falta hacer un muestreo de las bandas críticas de acuerdo al número de puntos del espectro de potencia N_p. Ahora bien, como las bandas críticas contienen unos cuantos valores diferentes de cero, no hace falta calcular todos los valores correspondientes a cada uno de los puntos del espectro de potencia, basta con las muestras comprendidas entre la frecuencia inferior y la superior para cada banda crítica.

A cada muestra del espectro de potencia $P(k)$ le corresponde un valor de frecuencia en Hertz, el cual es un múltiplo de la resolución en frecuencia del espectro:

$$f = k \cdot \Delta f \qquad k = 0,1,2,3,\ldots,N_p \tag{6.12}$$

La resolución en frecuencia del espectro Δf, se puede calcular mediante el cociente de la frecuencia de muestreo f_s, entre el número de puntos para la transformada rápida de Fourier N_{fft}:

$$\Delta f = \frac{f_s}{N_{fft}} \tag{6.13}$$

Si se invierte este cociente, el nuevo factor es la resolución por muestra del espectro Δk:

$$\Delta k = \frac{N_{fft}}{f_s} \tag{6.14}$$

El cual se puede utilizar para determinar que índice o muestra corresponde a determinada frecuencia, al multiplicar esa frecuencia por este factor de equivalencia:

$$k = \left(f \cdot \Delta k\right)_{\mathbf{int}} \tag{6.15}$$

Puesto que los índices de las muestras deben ser números enteros, se tiene que redondear el resultado a un número entero. El índice para la frecuencia inferior k_l se puede redondear hacia arriba, pero el índice para la frecuencia superior k_h se puede redondear hacia abajo. También, si es necesario, hay que tomar en cuenta que la primera muestra del espectro de potencia representa a la frecuencia cero.

Con los índices correspondientes a las frecuencias inferior k_l y superior k_h, ya se pueden calcular los N_{bc} valores diferentes de cero para cada una de las bandas críticas.

$$N_{bc} = k_h(bc) - k_l(bc) + 1 \tag{6.16}$$

En la tabla 6.2 se muestran los índices correspondientes a las frecuencias inferior y superior para las $BC = 15$ bandas críticas de la tabla 6.1. El número de puntos para la transformada rápida de Fourier es de $N_{fft} = 256$ puntos y la frecuencia de muestreo es de $f_s = 8000$ Hz. Por lo tanto, la resolución en frecuencia es de $\Delta f = 31.25$ Hz/muestra pero la resolución por muestra es de $\Delta k = 0.032$ muestras/Hz.

TABLA 6.2 Correspondencia entre índices y frecuencias para las bandas críticas Bark

bc	*fl*	*kl*	*(kl) int*	*fh*	*kh*	*(kh) int*	*Nbc*
[banda c.]	**[Hz]**	**[índice]**	**[índice]**	**[Hz]**	**[índice]**	**[índice]**	**[muestras]**
1	0.00	0.00	0	232.82	7.45	7	8
2	0.00	0.00	0	340.77	10.90	10	11
3	42.07	1.35	2	457.70	14.65	14	13
4	140.63	4.50	5	586.71	18.77	18	14
5	242.91	7.77	8	731.20	23.40	23	16
6	351.59	11.25	12	894.98	28.64	28	17
7	469.55	15.03	16	1,082.36	34.64	34	19
8	599.90	19.20	20	1,298.30	41.55	41	22
9	746.07	23.87	24	1,548.48	49.55	49	26
10	911.92	29.18	30	1,839.52	58.86	58	29
11	1,101.83	35.26	36	2,179.08	69.73	69	34
12	1,320.81	42.27	43	2,576.12	82.44	82	40
13	1,574.62	50.39	51	3,041.12	97.32	97	47
14	1,869.98	59.84	60	3,586.34	114.76	114	55
15	2,214.67	70.87	71	4,000.00	128.00	128	58

Para calcular los valores diferentes de cero para cada una de las bandas críticas, primero se tiene que obtener el valor correspondiente en frecuencia para la muestra *k*, comprendida dentro del intervalo $k_l(bc) \le k \le k_h(bc)$ de una banda crítica *bc*:

$$f = k \cdot \Delta f \qquad \textbf{(6.17)}$$

Luego se obtiene su equivalencia en la escala Bark:

$$(f)_{Bark} = 6\,\mathbf{Ln}\left[\left(\frac{f}{600}\right) + \sqrt{\left(\frac{f}{600}\right)^2 + 1}\right] \qquad \textbf{(6.18)}$$

Ahora se calcula la diferencia de frecuencias Bark f_d, que es igual a la diferencia de la frecuencia obtenida menos la frecuencia central de la banda crítica *bc*, ambas expresadas en la escala Bark:

$$f_d = (f)_{Bark} - (f_c(b))_{Bark} \qquad \textbf{(6.19)}$$

Entonces, el valor correspondiente a la muestra *k* para la banda crítica H_{bc} se determina mediante la ecuación (6.2), según sea el caso.

$$H_{bc}(f_d) = \begin{cases} 0 & \text{para} \quad f_d < -2.5 \\ 10^{(f_d + 0.5)} & \text{para} \quad -2.5 \le f_d < -0.5 \\ 1 & \text{para} \quad -0.5 \le f_d \le 0.5 \\ 10^{-2.5(f_d - 0.5)} & \text{para} \quad 0.5 < f_d \le 1.3 \\ 0 & \text{para} \quad f_d > 1.3 \end{cases} \qquad \textbf{(6.2)}$$

La tabla 6.3 muestra la correspondencia entre los índices y las frecuencias para los puntos que definen la séptima banda crítica Bark H_7. Se muestran los índices k para las muestras del espectro de potencia y sus valores correspondientes de frecuencia, primero en Hertz y luego en la escala Bark. También se muestra el resultado de la diferencia de frecuencias. La resolución en frecuencia es $\Delta f = 31.25$ Hz/muestra. Para fines ilustrativos, se incluyen los valores de los límites para los cinco casos de la ecuación (6.2).

TABLA 6.3 Correspondencia entre los valores para la séptima banda crítica trapezoidal

k [índice]	*f* [Hz]	*f* [Bark]	*fd* [Bark]	CASO	*H7(k)*
15	468.75	4.308	-2.506	*1 : fd < -2.5*	0.00
kl(7)	***469.55***	***4.314***	***-2.5***	***2 : -2.5 ≤ fd < -0.5***	***0.01***
16	500.00	4.551	-2.263	*2 : -2.5 ≤ fd < -0.5*	0.02
17	531.25	4.788	-2.026	*2 : -2.5 ≤ fd < -0.5*	0.03
18	562.50	5.019	-1.795	*2 : -2.5 ≤ fd < -0.5*	0.05
19	593.75	5.244	-1.570	*2 : -2.5 ≤ fd < -0.5*	0.09
20	625.00	5.463	-1.351	*2 : -2.5 ≤ fd < -0.5*	0.14
21	656.25	5.677	-1.137	*2 : -2.5 ≤ fd < -0.5*	0.23
22	687.50	5.885	-0.929	*2 : -2.5 ≤ fd < -0.5*	0.37
23	718.75	6.088	-0.726	*2 : -2.5 ≤ fd < -0.5*	0.59
24	750.00	6.286	-0.529	*2 : -2.5 ≤ fd < -0.5*	0.94
fc(7)-0.5	***754.58***	***6.314***	***-0.5***	***3 : -0.5 ≤ fd ≤ +0.5***	**1.00**
25	781.25	6.478	-0.336	*3 : -0.5 ≤ fd ≤ +0.5*	1.00
26	812.50	6.666	-0.148	*3 : -0.5 ≤ fd ≤ +0.5*	1.00
kc(7)	837.63	6.81	0	*3 : -0.5 ≤ fd ≤ +0.5*	1.00
27	843.75	6.850	0.036	*3 : -0.5 ≤ fd ≤ +0.5*	1.00
28	875.00	7.029	0.214	*3 : -0.5 ≤ fd ≤ +0.5*	1.00
29	906.25	7.203	0.389	*3 : -0.5 ≤ fd ≤ +0.5*	1.00
fc(7)+0.5	***926.50***	***7.314***	***0.5***	***3 : -0.5 ≤ fd ≤ +0.5***	**1.00**
30	937.50	7.374	0.560	*4 : +0.5 < fd ≤ +1.3*	0.71
31	968.75	7.540	0.726	*4 : +0.5 < fd ≤ +1.3*	0.27
32	1,000.00	7.703	0.889	*4 : +0.5 < fd ≤ +1.3*	0.11
33	1,031.25	7.862	1.048	*4 : +0.5 < fd ≤ +1.3*	0.04
34	1,062.50	8.017	1.203	*4 : +0.5 < fd ≤ +1.3*	0.02
kh(7)	***1,082.36***	***8.114***	***1.3***	***4 : +0.5 < fd ≤ +1.3***	***0.01***
35	1,093.75	8.169	1.355	*5 : fd > +1.3*	0.00

6.3 PROCEDIMIENTO DE ANÁLISIS PLP

A continuación se describe un procedimiento para obtener los coeficientes PLP. La figura 6.4 muestra el diagrama de bloques para este procedimiento. Los bloques muestran las operaciones necesarias y a la izquierda de éstos se muestran los parámetros de entrada requeridos en cada paso. A partir del tercer paso, el ponderado por la ventana de Hamming, las operaciones se realizan sobre las tramas de la señal de voz. La consiguiente reducción en el número de parámetros para cada trama se ilustra mediante la cantidad de flechas.

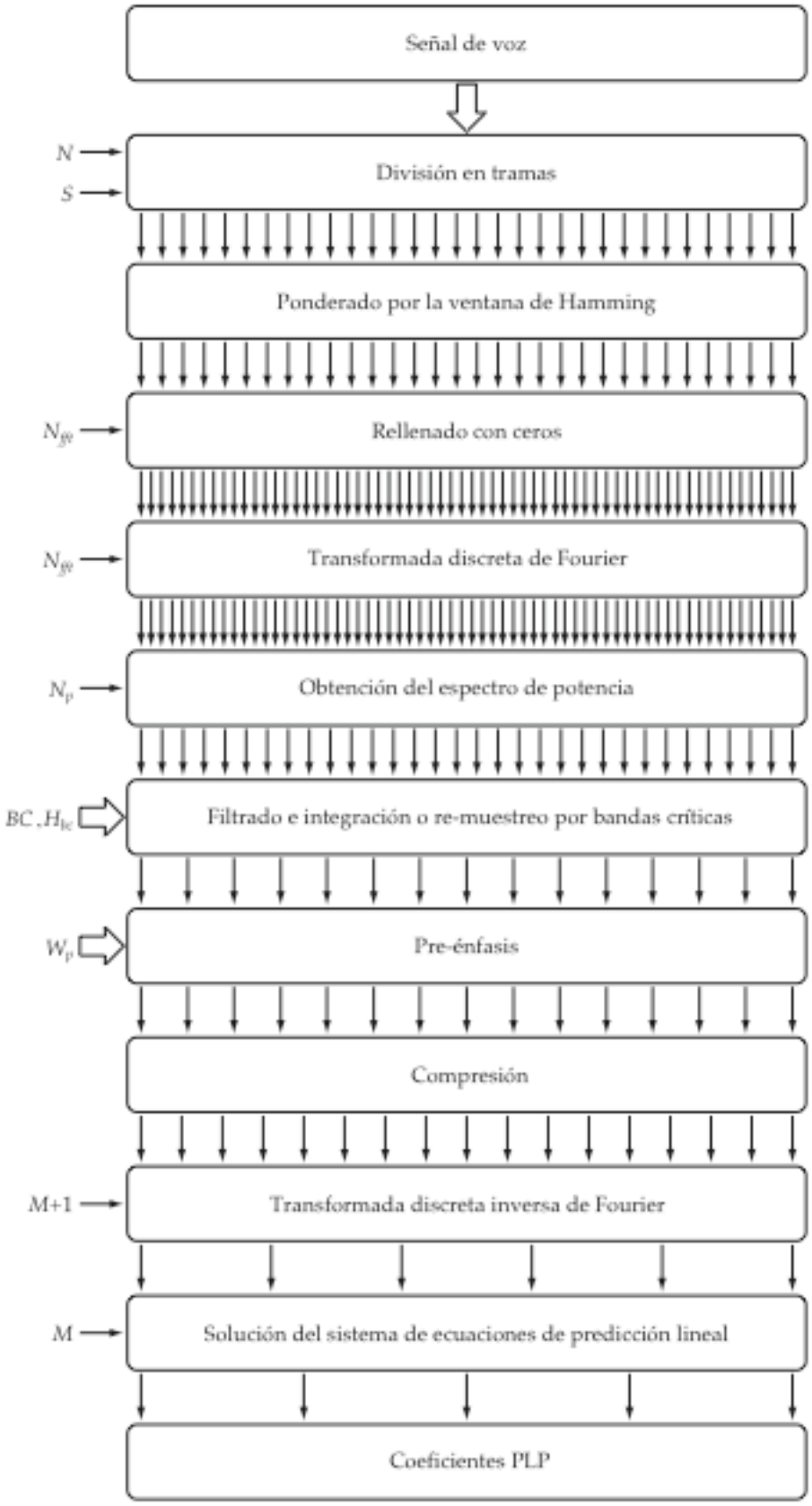


FIGURA 6.4 Procedimiento para la obtención de los coeficientes PLP

6.3.1 EL NÚMERO DE COEFICIENTES Y DE FILTROS

Los coeficientes PLP se calculan para una cantidad de coeficientes M que es menor que el número de filtros o bandas críticas BC. En cuanto al número de filtros o bandas críticas BC, este es igual al límite superior de frecuencia en la escala Bark:

$$BC = \left(\left(f_{max} \right)_{Bark} \right)_{\mathbf{int}} \tag{6.20}$$

Pero como el número de filtros debe ser un número entero, hay que truncar o redondear hacia abajo el resultado. Ahora bien, el número de filtros o bandas críticas se incrementa en dos para incluir las dos bandas críticas centradas en $f = 0$ y en la frecuencia de Nyquist f_{max}.

$$BC = BC + 2 \tag{6.21}$$

Dado que que la aproximación es menos exacta para estas bandas, los valores de éstas se igualan a los valores de sus vecinos más cercanos. De tal manera que se comienza y se termina con dos muestras con valores idénticos [7].

6.3.2 DIVISIÓN EN TRAMAS

La señal de voz se divide en L tramas de N muestras para su análisis, con S muestras de separación entre tramas adyacentes:

$$x_\ell(n) = x(n + \ell \cdot S) \qquad \begin{array}{l} N \geq S \\ \ell = 0,1,2,3,\ldots,L-1 \\ n = 0,1,2,3,\ldots,N-1 \end{array} \tag{6.22}$$

6.3.3 PONDERADO POR LA VENTANA DE HAMMING

El ponderado de la trama de análisis por la ventana de Hamming, contribuye a minimizar las discontinuidades al inicio y al final de las tramas:

$$\begin{aligned} x_w(n) &= x_\ell(n) \cdot w(n) \\ x_w(n) &= x_\ell(n) \cdot \left[0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1} \right) \right] \qquad n = 0,1,2,3,\ldots,N-1 \end{aligned} \tag{6.23}$$

6.3.4 RELLENADO CON CEROS

Se añaden las muestras nulas que hagan falta para que la longitud de la trama de análisis sea igual a N_{fft}, la potencia de dos más próxima:

$$N_{fft} = 2^{\left[\log_2(N)\right]_{\text{int}}} \tag{6.24}$$

$$x_{zp}(n) = \begin{cases} x_w(n) & \text{para} \quad n = 0,1,2,3,\ldots,N-1 \\ 0 & \text{para} \quad n = N,N+1,\ldots,N_{fft}-1 \end{cases} \tag{6.25}$$

6.3.5 TRANSFORMADA DISCRETA DE FOURIER

La transformada discreta de Fourier de la trama de análisis se calcula mediante la transformada rápida de Fourier de N_{fft} puntos:

$$X(k) = \mathbf{FFT}\left\{x_{zp}(n), N_{fft}\right\} \qquad \begin{matrix} k = 0,1,2,3,\ldots,N_{fft}-1 \\ n = 0,1,2,3,\ldots,N_{fft}-1 \end{matrix} \tag{6.26}$$

6.3.6 OBTENCIÓN DEL ESPECTRO DE POTENCIA

El espectro de potencia es igual a la suma de los cuadrados de la parte real e imaginaria del espectro en frecuencia de la trama de análisis:

$$N_p = 0.5 N_{fft} \tag{6.27}$$

$$P(k) = \mathbf{real}\left[X(k)\right]^2 + \mathbf{imag}\left[X(k)\right]^2 \qquad k = 0,1,2,3,\ldots,N_p \tag{6.28}$$

6.3.7 FILTRADO E INTEGRACIÓN O RE-MUESTREO POR BANDAS CRÍTICAS

En esta etapa es donde se emplea el primero de los tres conceptos de la psicoacústica, la resolución espectral por bandas críticas. Para extraer las características de la envolvente del espectro, se emplea un banco de filtros o ventanas trapezoidales traslapadas, con sus frecuencias centrales equiespaciadas en la escala Bark. La integración por bandas criticas se realiza mediante la sumatoria de los productos de las muestras del espectro de potencia $P(k)$ por cada uno de los filtros H_{bc} para las bandas críticas:

$$P_s(bc) = \sum_{k=k_l(bc)}^{k_h(bc)} P(k) \cdot H_{bc}(k) \qquad bc = 1,2,3,\ldots,BC \tag{6.29}$$

El resultado que proporciona esta etapa es un espectro integrado por bandas críticas, o re-muestreado $P_s(bc)$ que proporciona un espectro relativamente suavizado, con uno o dos grandes picos [6]. Se obtiene una muestra por cada una de las bandas críticas, de tal manera que el espectro pasa por una primera compresión o reducción en el número de parámetros.

6.3.8 PRE-ÉNFASIS

En esta etapa es donde se emplea el segundo de los tres conceptos de la psicoacústica, las curvas isofónicas o de igual nivel de sonoridad. El oído humano no percibe las frecuencias de la misma manera, tiene una sensibilidad selectiva o variable a diferentes frecuencias, por esta razón, en esta técnica se busca imitar esta propiedad mediante la siguiente ecuación [5]:

$$W_p(f) = \frac{\left[(f)^4\right]\left[(f)^2 + 1200^2\right]}{\left[(f)^2 + 400^2\right]^2\left[(f)^2 + 3100^2\right]} \qquad \textbf{(6.30)}$$

El pre-énfasis se hace para cada una de las bandas críticas. Con la ecuación anterior, se calcula el peso correspondiente para la frecuencia central de la banda crítica y con ese peso, se ponderan todas las muestras de esa misma banda crítica [7]. En la tabla 6.4 se muestran los valores de los pesos del pre-énfasis correspondientes a las frecuencias centrales de las $BC = 15$ bandas críticas de la tabla (6.1).

TABLA 6.4 Pesos del pre-énfasis para las bandas críticas Bark

bc [banda c.]	*fc(bc)* [Hz]	*Wp(bc)*
1	97.77	0.0005
2	198.12	0.0059
3	303.70	0.0211
4	417.29	0.0448
5	541.89	0.0733
6	680.78	0.1044
7	837.63	0.1377
8	1,016.58	0.1743
9	1,222.34	0.2156
10	1,460.35	0.2633
11	1,736.88	0.3183
12	2,059.23	0.3808
13	2,435.90	0.4498
14	2,876.83	0.5228
15	3,393.66	0.5966

Puesto que el peso del pre-énfasis para una banda crítica es un escalar, hay tres alternativas equivalentes para realizar el pre-énfasis:

1. Ponderar los valores del banco de filtros.

$$H_{bc}(k) = W_p(bc) \cdot H_{bc}(k) \tag{6.31}$$

2. Ponderar los productos para el espectro muestreado.

$$P_s(bc) = \sum_{k=k_l(bc)}^{k_h(bc)} W_p(bc) \cdot P(k) \cdot H_{bc}(k) \tag{6.32}$$

3. Ponderar los resultados del espectro muestreado.

$$P_s(bc) = W_p(bc) \cdot \sum_{k=k_l(bc)}^{k_h(bc)} P(k) \cdot H_{bc}(k)$$
$$P_s(bc) = W_p(bc) \cdot P_s(bc) \tag{6.33}$$

En cualquiera de los tres casos anteriores, el pre-énfasis modifica la amplitud de los valores del espectro muestreado.

Para el primer caso, la figura 6.5 muestra el banco de $BC = 15$ bandas críticas ponderadas por la ecuación de pre-énfasis. Puede observarse en la figura que la amplitud de las bandas críticas se incrementa a medida que aumenta la frecuencia.

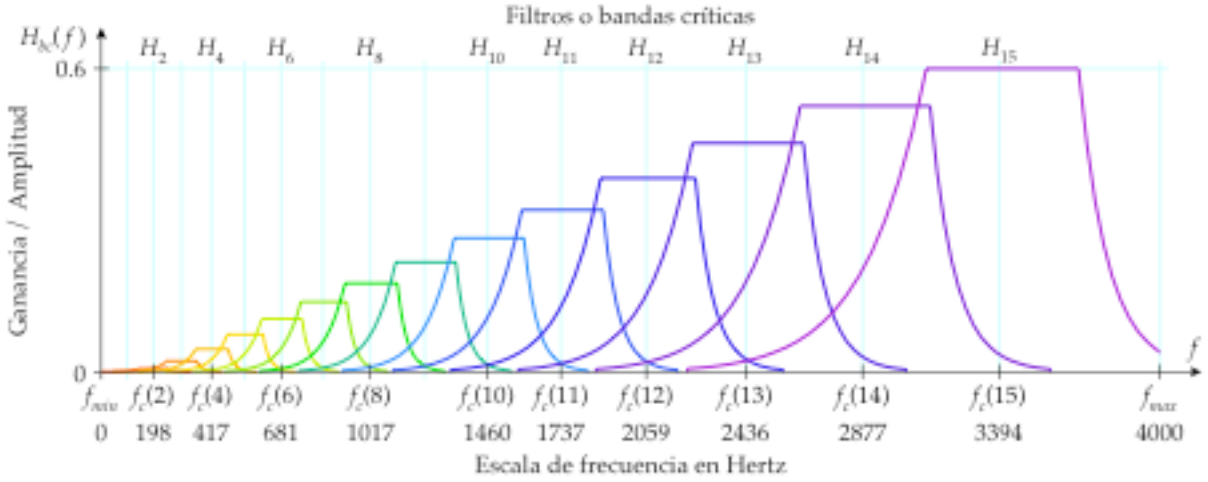


FIGURA 6.5 Banco de filtros o ventanas ponderados por los pesos del pre-énfasis

6.3.9 COMPRESIÓN

En esta etapa es donde se emplea el tercer concepto de la psicoacústica, la ley de potencia de la audición. En el análisis cepstral se obtiene el logaritmo del espectro después de su integración, pero en PLP se usa otra función no lineal en su lugar, lo que constituye una aproximación de la relación no lineal de la ley de potencia entre la intensidad del sonido $I(\omega)$ y el volumen percibido o sonoridad $L(\omega)$.

$$L(\omega) = \sqrt[3]{I(\omega)} \qquad \textbf{(6.34)}$$

Aparte de asemejar esta propiedad del oído humano, en este paso se comprime la amplitud del espectro por medio de la raíz cúbica:

$$\begin{aligned} P_{sc}(bc) &= \sqrt[3]{P_s(bc)} \\ P_{sc}(bc) &= P_s(bc)^{0.33} \end{aligned} \qquad \textbf{(6.35)}$$

La compresión mediante la raíz cúbica contribuye a disminuir las variaciones de amplitud producidas por la resonancias espectrales, de tal manera que el modelado por predicción lineal se pueda hacer mediante un modelo de un orden bajo [7].

6.3.10 TRANSFORMADA INVERSA DE FOURIER

A diferencia del método de predicción lineal, los coeficientes de autocorrelación no se obtienen en el dominio del tiempo mediante la función de autocorrelación en tiempo corto. Más bien se obtienen mediante la transformada inversa de Fourier [9]. Puesto que no se calculó el logaritmo del espectro, sus elementos son muy parecidos a los coeficientes de autocorrelación, aunque todavía pertenecen al espectro comprimido [5]. La transformada inversa de Fourier es una mejor opción que la FFT inversa, ya que solamente se necesitan unos cuantos coeficientes, basta con los primeros $M+1$ valores, que son los necesarios para resolver el sistema de ecuaciones de predicción lineal [7]. Puesto que el espectro tiene simetría par y sus valores son reales, solo hace falta calcular las componentes coseno de la transformada discreta inversa de Fourier [5]:

$$R(m) = \frac{1}{2(BC-1)} \sum_{bc=0}^{2(BC-1)} P_{sc}(bc)\cos\left[\frac{\pi}{BC-1}(bc)m\right] \qquad m = 0,1,2,3,\ldots,M \quad \textbf{(6.36)}$$

6.3.11 SOLUCIÓN DEL SISTEMA DE ECUACIONES

Aunque durante el re-muestreado del espectro en bandas críticas se eliminan algunos detalles, se ha notado que otro nivel de integración resulta conveniente para reducir los efectos de otras fuentes no lingüísticas de variación en la señal de voz. En este análisis también se emplea la técnica LPC para suavizar el espectro. El sistema de ecuaciones de predicción lineal construido con los coeficientes de autocorrelación del paso anterior, se resuelve mediante el algoritmo de Levinson-Durbin, para obtener la solución de la misma manera que con el LPC convencional [5]. Por lo tanto, la solución final es el vector o el conjunto de coeficientes PLP:

$$\mathbf{V}_{plp} = a_m$$
$$\mathbf{V}_{plp} = [a_1, a_2, a_3, \ldots, a_M] \qquad m = 1,2,3,\ldots,M \qquad \textbf{(6.37)}$$

6.4 CARACTERÍSTICAS DEL ANÁLISIS PLP

Al igual que en LPC, el análisis PLP proporciona una envolvente suavizada del espectro de potencia, que se ajusta mejor a los picos que a los valles. Muchos investigadores han encontrado que este enfoque contribuye a la robustez contra el ruido y a una mayor independencia del locutor que aquella obtenida por el truncamiento cepstral [5].

Una de las principales desventajas del modelo todos polos de predicción lineal es que aproxima de igual manera todas las frecuencias comprendidas en el rango de análisis. Esta propiedad es inconsistente con la audición humana, porque para las frecuencias mayores de 800 Hz, la resolución espectral del oído disminuye con la frecuencia. Y lo que es más, el oído es más sensible a los niveles de amplitud presentes en las conversaciones normales, las cuales se encuentran a la mitad del rango del espectro audible. En comparación con el análisis de predicción lineal convencional, el análisis PLP es más consistente con el oído humano [7].

Tanto el análisis mel cepstral como el PLP proporcionan una representación equivalente a un espectro en tiempo corto suavizado, que ha sido comprimido y equalizado, de una manera similar a como sucede en el sistema auditivo humano. De hecho, las dos técnicas son bastante parecidas. Cada una tiene la resolución reducida para altas frecuencias, que es característica de los métodos basados en los bancos de filtros auditivos. No obstante, la principal diferencia entre estos métodos radica en la naturaleza del suavizado del espectro, cepstral para el primero y mediante predicción lineal para el segundo [5].

De tal manera que el PLP puede ser visto ya sea como un análisis mel cepstral con un suavizado del espectro mediante LPC, o como un análisis LPC para una versión de voz que ha sido modificada de acuerdo a algunas propiedades auditivas, en particular la resolución espectral por bandas críticas y al modelo de la ley de potencia de la audición. En la práctica, se ha encontrado que como resultado de esta modificación, basta con un modelo PLP de un orden mucho menor en comparación con un LPC estándar [5].

En términos de computo, las operaciones más costosas son: el cálculo de la FFT, seguida por la integración espectral por bandas críticas y la compresión mediante la raíz cúbica. El costo del modelado por predicción lineal resulta insignificante debido al número reducido de muestras del espectro auditivo a ser aproximadas [7].

En los sistemas actuales de reconocimiento de voz, se ha tenido la meta de encontrar alguna representación que sea relativamente estable o consistente para diferentes ejemplos de una misma palabra o locución, a pesar de las diferencias entre locutores o de las características del entorno. Esto resulta algo complicado, ya que la voz humana contiene una gran cantidad de información, pero en este caso, la más relevante es la información acústica. De tal manera que, datos como el género del hablante, su identidad, etc. resultan irrelevantes. Los modelos PLP y LPC de orden elevado son buenas técnicas en los experimentos dependientes del locutor, pero en los experimentos con locutores mixtos, el modelo PLP de quinto orden se desempeña mucho mejor que el LPC convencional. De hecho, los modelos LPC de orden elevado nunca alcanzan la exactitud del modelo PLP de quinto orden. Además, con PLP se logra una mejor independencia del locutor [7].

Parece ser que las claves lingüísticamente relevantes para la independencia del locutor yacen en la forma general del espectro auditivo, la cual puede ser caracterizada mediante uno o dos de los picos del espectro auditivo del modelo PLP de quinto orden, que es el óptimo para eliminar la información dependiente del locutor de la voz. Los otros detalles más sutiles del espectro, pueden ser modelados por los polos adicionales de los modelos PLP de mayor orden, los cuales contienen la información dependiente del locutor [7].

Una consecuencia del modelo todos polos de bajo orden es que, los picos menores del espectro son removidos. Además se consiguen otros dos efectos importantes, la reducción en el número de parámetros del espectro auditivo y el aumento de la resolución en frecuencia del resultado. La resolución en frecuencia es incrementada efectivamente porque las posiciones de los picos en el modelo todos polos no están restringidos a las frecuencias centrales de las bandas criticas [6].

De tal manera que el modelo PLP óptimo es el de quinto orden. Un modelo LPC de quinto orden no arroja información fonética consistente. En cambio, tanto un LPC de orden 13 como un PLP de quinto orden preservan la mayoría de la información, como se muestra en la figura 6.6 [7].

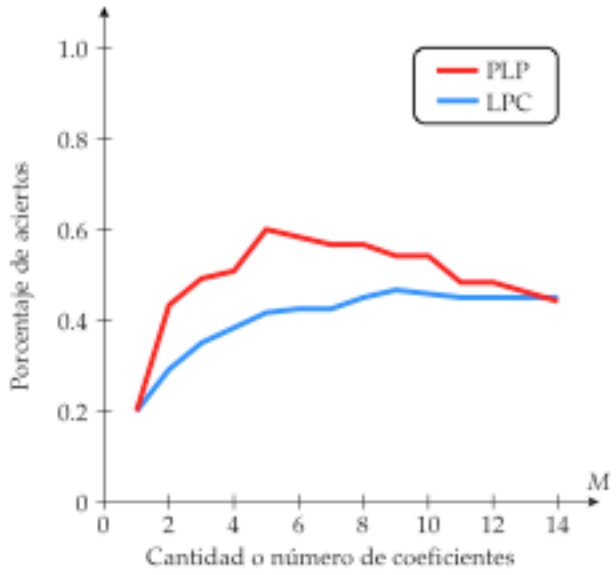


FIGURA 6.6 Comparación de la variación en el reconocimiento entre PLP y LPC

CAPÍTULO

7

VQ

CUANTIZACIÓN VECTORIAL

7.1 INTRODUCCIÓN

La cuantización vectorial o **V**ectorial **Q**uantization (VQ), es la generalización de la cuantización escalar, ya que en lugar de cuantizar solamente un valor escalar, ahora se hace con un vector completo, es decir, sobre un conjunto ordenado de valores reales. Además, el cambio de una a varias dimensiones, es un paso trascendental que origina y hace posibles un gran número de nuevas ideas, conceptos, técnicas y aplicaciones [4].

Mientras que la cuantización escalar se utiliza principalmente en la conversión analógico-digital, la cuantización vectorial se enfrenta a las sofisticadas técnicas del procesamiento digital de señales, pues por ejemplo, esta se emplea en la compresión de imágenes y de voz, pero además puede aplicarse en el reconocimiento de patrones, por lo que también se puede usar para el reconocimiento de voz. Por lo tanto, la cuantización vectorial es mucho más que la generalización formal de la cuantización escalar [4].

7.2 DESCRIPCIÓN

La mayoría de las técnicas de codificación de voz emplean una secuencia de tramas en el tiempo, a su vez, cada una de estas tramas se representa mediante un vector o conjunto de M valores numéricos o coeficientes. De tal manera que cada uno de estos vectores de coeficientes se puede representar como un vector $\mathbf{V}$ de M-dimensiones [12]:

$$\mathbf{V} = [v_1, v_2, v_3, \ldots, v_M] \qquad \textbf{(7.1)}$$

Así es que los resultados obtenidos mediante la codificación de la señal de voz consisten en una secuencia de vectores con las características espectrales variantes en el tiempo de la señal de voz [14].

La idea fundamental en la que se basa la cuantización vectorial, es que un espacio vectorial de M dimensiones, se puede dividir en K subconjuntos, particiones, grupos o contenedores, los cuales se pueden representar mediante un sólo vector, que a su vez, proporciona el centroide del grupo. El conjunto $\mathbf{Z}$ de los K centroides se conoce como "libro código" (codebook):

$$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3, \ldots, \mathbf{Z}_K] \qquad (7.2)$$

A cada uno de estos vectores se le asigna una etiqueta, índice o dirección que es única [12]. La figura 7.1 muestra un caso para cuando $M = 2$ dimensiones, con $K = 7$ particiones hexagonales para los centroides, de tal manera que los vectores y los centroides se representan por medio de puntos o pares de coordenadas.

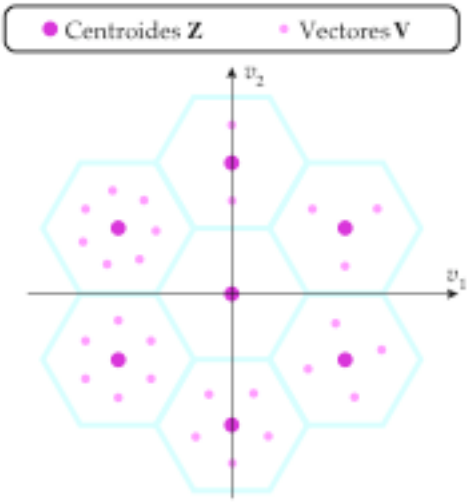


FIGURA 7.1 Ilustración de la cuantización vectorial en dos dimensiones

Debido a que se tiene solamente un número finito de vectores en el libro código o codebook, el proceso de elegir la mejor representación es equivalente a cuantizar el vector, obteniendo como consecuencia cierto nivel de error de cuantización, que aunque no sea posible anular, se puede reducir al incrementar el número de centroides. Además existe un compromiso inherente entre el error de cuantización, el proceso para elegir los centroides y el espacio de memoria necesario para su almacenamiento [14].

Una de las ventajas que ofrece la cuantización vectorial es la flexibilidad en el tamaño del libro código que se puede usar. Pues a diferencia de la cuantización escalar, es posible tener cualquier número entero de particiones arbitrarias. Ahora bien, el objetivo principal en el diseño de un cuantizador vectorial, es encontrar los centroides óptimos del libro código, que maximicen su desempeño al tomar en cuenta los vectores a procesar [4].

Los elementos necesarios para su implementación son los siguientes: una distancia o medida de distorsión, un procedimiento de clasificación, un conjunto de vectores de entrenamiento y un método de agrupamiento.

7.3 DISTANCIAS O MEDIDAS DE DISTORSIÓN

Un componente clave en la mayoría de los algoritmos de comparación de patrones, es una medida de distancia o de la semejanza entre un par de vectores de características. Mediante estas distancias, es posible obtener los centroides al agrupar los vectores de entrenamiento, o también asociarlos con algún centroide en particular, esto es, el clasificar o asignar un vector dentro de un subconjunto, grupo o partición.

Para que una función de distancia pueda tratarse con rigor matemático, debe satisfacer las siguientes condiciones:

$$d(\mathbf{X},\mathbf{Y}) > 0 \quad \text{si} \quad \mathbf{X} \neq \mathbf{Y} \tag{7.3}$$

$$d(\mathbf{X},\mathbf{Y}) = 0 \quad \text{si} \quad \mathbf{X} = \mathbf{Y} \tag{7.4}$$

$$d(\mathbf{X},\mathbf{Y}) = d(\mathbf{Y},\mathbf{X}) \tag{7.5}$$

$$d(\mathbf{X},\mathbf{Y}) \leq d(\mathbf{X},\mathbf{Z}) + d(\mathbf{Y},\mathbf{Z}) \tag{7.6}$$

Ahora bien, para que una medida de distancia sea útil en el procesamiento digital de voz, hay otra consideración que también es muy importante, pues una medida de distancia debe tener una alta correlación entre su valor numérico y su relevancia fonética, es decir, que los cambios espectrales que ocasionan que los sonidos comparados parezcan diferentes, deberían estar asociados con distancias grandes, de manera similar, los cambios espectrales que no alteran un sonido, deberían estar asociados con distancias pequeñas [14].

Sin embargo, este requisito de la relevancia fonética muy a menudo no se puede satisfacer con las medidas de distancia que se pueden tratar con un rigor matemático, por lo que resulta inevitable cierto compromiso entre ambas. Por esta razón, en lugar de medidas de distancia se suele hablar de medidas de distorsión, porque generalmente, estas no cumplen con todas las propiedades que debe satisfacer una función de distancia, pues existe cierta relajación en las condiciones matemáticas para las distintas propiedades. Por lo tanto, de acuerdo a lo que se acostumbra en el procesamiento digital de voz, el término distancia se emplea como un sinónimo de la semejanza entre dos vectores o objetos [14].

7.3.1 LA DISTANCIA CUADRÁTICA MEDIA

La medida de distancia más conveniente y más ampliamente utilizada para determinar la distancia entre dos vectores $\mathbf{X}, \mathbf{Y}$ es la distancia cuadrática media [4]:

$$d(\mathbf{X},\mathbf{Y}) = (\mathbf{X}-\mathbf{Y})^{\mathrm{T}}(\mathbf{X}-\mathbf{Y}) \qquad \mathbf{X} = [x_1, x_2, x_3, \ldots, x_M]$$
$$d(\mathbf{X},\mathbf{Y}) = \sum_{m=1}^{M} (x_m - y_m)^2 \qquad \mathbf{Y} = [y_1, y_2, y_3, \ldots, y_M] \qquad \textbf{(7.7)}$$

A esta distancia también se le conoce como la distancia Euclidiana cuadrática o también como el error cuadrático medio.

7.3.2 LA DISTANCIA DE ITAKURA

Probablemente, la mejor medida de distorsión para determinar la semejanza entre dos vectores LPC es la distancia propuesta por Itakura:

$$d(\mathbf{X},\mathbf{Y}) = \log\left(\frac{\mathbf{X}\mathbf{R}_{\mathbf{Y}}\mathbf{X}^{\mathrm{T}}}{\mathbf{Y}\mathbf{R}_{\mathbf{Y}}\mathbf{Y}^{\mathrm{T}}}\right) \qquad \mathbf{X} = [-1, a_{x1}, a_{x2}, a_{x3}, \ldots, a_{xp}]$$
$$\mathbf{Y} = [-1, a_{y1}, a_{y2}, a_{y3}, \ldots, a_{yp}] \qquad \textbf{(7.8)}$$

En donde $\mathbf{Y}$ es un vector de referencia y $\mathbf{X}$ es el vector de prueba a comparar. Ambos son vectores aumentados de coeficientes LPC, para los cuales $a_0 = -1$. $\mathbf{R}_{\mathbf{Y}}$ es la matriz de los coeficientes de autocorrelación para la trama correspondiente al vector $\mathbf{Y}$:

$$\mathbf{R}_{\mathbf{Y}} = \begin{bmatrix} R(0) & R(1) & R(2) & R(3) & \cdots & R(p) \\ R(1) & R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(2) & R(1) & R(0) & R(1) & \cdots & R(p-2) \\ R(3) & R(2) & R(1) & R(0) & \cdots & R(p-3) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p) & R(p-1) & R(p-2) & R(p-3) & \cdots & R(0) \end{bmatrix} \qquad \textbf{(7.9)}$$

A partir de la multiplicación de las tres matrices, se pueden obtener las siguientes ecuaciones:

$$\mathbf{X}\mathbf{R}_{\mathbf{Y}}\mathbf{X}^{\mathrm{T}} = \sum_{i=0}^{p} a_{xi} \sum_{c=0}^{p} a_{xc} \cdot R(|i-c|) \qquad \textbf{(7.10)}$$

$$\mathbf{Y}\mathbf{R}_{\mathbf{Y}}\mathbf{Y}^{\mathrm{T}} = \sum_{i=0}^{p} a_{yi} \sum_{c=0}^{p} a_{yc} \cdot R(|i-c|) \qquad \textbf{(7.11)}$$

7.3.3 EL ERROR CUADRÁTICO PONDERADO

Otra medida de distorsión, que es de particular interés, es la del error cuadrático ponderado:

$$d(\mathbf{X},\mathbf{Y}) = (\mathbf{X}-\mathbf{Y})^{\mathrm{T}}\,\mathbf{W}(\mathbf{X}-\mathbf{Y}) \qquad \begin{matrix}\mathbf{X} = [x_1, x_2, x_3, \ldots, x_M] \\ \mathbf{Y} = [y_1, y_2, y_3, \ldots, y_M]\end{matrix} \qquad \textbf{(7.12)}$$

En donde $\mathbf{W}$ es una matriz de ponderado simétrica y definida positiva, en cuanto a los vectores $\mathbf{X}$ y $\mathbf{Y}$, éstos se tratan como si fueran vectores columna. Cuando la matriz de ponderado es una matriz diagonal, la ecuación anterior se puede reescribir como:

$$d(\mathbf{X},\mathbf{Y}) = \sum_{m=1}^{M} w_{mm}(x_m - y_m)^2 \qquad \begin{matrix}\mathbf{X} = [x_1, x_2, x_3, \ldots, x_M] \\ \mathbf{Y} = [y_1, y_2, y_3, \ldots, y_M]\end{matrix} \qquad \textbf{(7.13)}$$

Aunque consiste en una modificación simple de la distancia cuadrática media, esta resulta muy útil, pues permite dar distintos pesos por separado a las diferentes componentes de los vectores.

7.3.4 LA DISTANCIA DE MAHALANOBIS

Para algunas aplicaciones, la matriz de ponderado se elige de acuerdo a las características estadísticas del vector de entrada. Por ejemplo, en la distancia de Mahalanobis, la matriz de ponderado es la inversa de la matriz de la covarianza del vector de entrada.

$$d(\mathbf{X},\mathbf{Y}) = (\mathbf{X}-\mathbf{Y})^{\mathrm{T}}\,\mathbf{W}^{-1}(\mathbf{X}-\mathbf{Y}) \qquad \begin{matrix}\mathbf{X} = [x_1, x_2, x_3, \ldots, x_M] \\ \mathbf{Y} = [y_1, y_2, y_3, \ldots, y_M]\end{matrix} \qquad \textbf{(7.14)}$$

Para este caso en particular, si las componentes del vector de entrada no están correlacionadas, la matriz de ponderado de Mahalanobis se reduce a una matriz diagonal, cuyos elementos son iguales al inverso de la varianza:

$$w_{mm} = \frac{1}{\sigma_m^2} \qquad \textbf{(7.15)}$$

$$d(\mathbf{X},\mathbf{Y}) = \sum_{m=1}^{M} \sigma_m^{-2}(x_m - y_m)^2 \qquad \begin{matrix}\mathbf{X} = [x_1, x_2, x_3, \ldots, x_M] \\ \mathbf{Y} = [y_1, y_2, y_3, \ldots, y_M]\end{matrix} \qquad \textbf{(7.16)}$$

7.4 PROCEDIMIENTO DE CLASIFICACIÓN

El procedimiento de clasificación para cualquier vector arbitrario, es básicamente una búsqueda en todo el libro código para encontrar el centroide más cercano o que más se parezca al vector de entrada. Entonces, mediante alguna distancia o medida de distorsión, se puede realizar la comparación de todas las distancias $\mathbf{D} = [d_1, d_2, d_3, \ldots, d_K]$ entre un vector de entrada $\mathbf{V}$ y cada uno de los K centroides del libro código $\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3, \ldots, \mathbf{Z}_K]$, para asignar el vector al grupo correspondiente al centroide que proporcione la mínima distancia o la medida de distorsión más pequeña:

$$i = \arg\min\{d_1, d_2, d_3, \ldots, d_K\} \qquad d_k = d(\mathbf{V}, \mathbf{Z}_k) \qquad k = 1,2,3,\ldots,K \quad \textbf{(7.17)}$$

En donde $\arg\min$ denota el valor del índice i del mínimo de los valores que se encuentran entre de las llaves. Por lo tanto, la salida del cuantizador vectorial es solamente el valor del índice del centroide más cercano de los K vectores almacenados.

Parte de la complejidad del cuantizador vectorial surge de la necesidad de realizar una búsqueda, porque para poder hacer una sola identificación, se necesitan hacer tantas comparaciones como palabras existan dentro del libro código. Si su tamaño es muy grande (como por ejemplo, de más de 1024 centroides), el procedimiento puede resultar computacionalmente costoso, dependiendo de como se calculen las distancias [14].

7.5 ENTRENAMIENTO

Un aspecto muy importante en cualquier esquema de cuantización vectorial, es la provisión de un conjunto de MV vectores de entrenamiento.

$$\mathbf{V}^{(m)} = [\mathbf{V}^{(1)}, \mathbf{V}^{(2)}, \mathbf{V}^{(3)}, \ldots, \mathbf{V}^{(MV)}] \qquad m = 1,2,3,\ldots,MV \qquad \textbf{(7.18)}$$

El entrenamiento es el proceso en el cual se obtienen y se guardan los centroides o los vectores más representativos u óptimos del conjunto, para usarse como patrones de referencia para las distintas palabras o clases. Una de las motivaciones para considerar el empleo de un libro código era la suposición de que, en el límite, de manera ideal se obtendría alrededor de un centroide por fonema [14].

Si K denota el número de centroides, entonces se necesita que $MV > K$ para poder encontrar el mejor conjunto de centroides. En la práctica, se ha encontrado que para obtener un libro código que trabaje razonablemente bien, el número de vectores de entrenamiento debe ser al menos 10 veces mayor que el número de centroides [14].

$$MV \geq 10K \qquad \textbf{(7.19)}$$

7.6 EL MÉTODO DE AGRUPAMIENTO K-MEDIAS

El método de agrupamiento permite obtener el conjunto de los K centroides del libro código, mediante la separación y clasificación de los MV vectores de entrenamiento. Los elementos que pertenecen a un subconjunto o grupo, tienen una mayor semejanza entre sí en comparación con los otros elementos de los demás grupos. Uno de estos métodos de agrupamiento es el algoritmo K-medias, y su procedimiento se muestra a continuación:

- Inicialización.

Se seleccionan de manera arbitraria K vectores directamente del conjunto de vectores de entrenamiento $\mathbf{V}^{(m)}$ como los centroides iniciales $\mathbf{Z}_k^{(0)}$:

$$\mathbf{Z}_k^{(0)} = \left[\mathbf{Z}_1^{(0)}, \mathbf{Z}_2^{(0)}, \mathbf{Z}_3^{(0)}, \ldots, \mathbf{Z}_K^{(0)}\right] \qquad \mathbf{Z}_k^{(0)} \in \mathbf{V}_m \tag{7.20}$$

- Método recursivo.

1er Paso. Búsqueda y agrupamiento de los vectores. En la iteración ℓ, se tienen que calcular todas las distancias $\mathbf{D} = [d_1, d_2, d_3, \ldots, d_K]$ entre cada vector de entrenamiento $\mathbf{V}^{(m)}$ y cada uno de los K centroides $\mathbf{Z}_k^{(\ell)}$. Después se busca el centroide que proporcione la mínima distancia o la medida de distorsión más pequeña, para entonces asignar el vector de entrenamiento $\mathbf{V}^{(m)}$ al grupo correspondiente G_i:

$$i = \mathbf{arg\,min}\{d_1, d_2, d_3, \ldots, d_K\} \qquad d_k = d\left(\mathbf{V}^{(m)}, \mathbf{Z}_k^{(\ell)}\right) \qquad k = 1,2,3,\ldots,K \tag{7.21}$$

Esto se tiene que hacer para cada uno de los MV vectores de entrenamiento.

2o Paso. Actualización de los centroides. Se calcula un nuevo conjunto de K centroides $\mathbf{Z}_k^{(\ell+1)}$ utilizando los vectores que se asignaron dentro de cada uno de los grupos:

$$\mathbf{Z}_k^{(\ell+1)} = \frac{1}{N_i} \sum_{\mathbf{v} \in G_i} \mathbf{V} \qquad i = 1,2,3,\ldots,K \tag{7.22}$$

En donde N_i es el número o la cantidad de vectores asignados al grupo G_i.

- Criterio de convergencia.

Se continua repitiendo los pasos de búsqueda, agrupamiento de los vectores y la actualización de los centroides hasta que la diferencia entre los centroides sea mínima, es decir, que no sobrepase cierto umbral predeterminado de tolerancia TOL:

$$\mathbf{Z}_k^{(\ell+1)} - \mathbf{Z}_k^{(\ell)} \leq TOL \qquad k = 1,2,3,\ldots,K \tag{7.23}$$

Entonces se puede asegurar la convergencia del algoritmo para terminar.

7.7 ENTRENAMIENTO CASUAL

Cuando la cantidad de palabras o clases en el vocabulario no es muy grande y el sistema ha sido diseñado para un usuario en particular, un procedimiento simple de entrenamiento que se puede realizar es emplear cada ejemplo de una palabra dicha durante la sesión de entrenamiento para usarla como patrón de referencia. Normalmente, cada palabra o clase se representa por varios ejemplos de la misma locución, obteniendo así múltiples patrones de referencia. Esto es lo que se denomina como entrenamiento casual [14].

Para un vocabulario simple, un locutor puede ser capaz de producir un conjunto consistente de ejemplos que pueden ser empleados como patrones de referencia. Sin embargo, puesto que el método no intenta estimar o predecir la variabilidad de los patrones, fácilmente puede fallar en aquellos sistemas en los cuales las palabras del vocabulario se prestan a confusión. Lo que es más, debido a la intención de mantener la simplicidad inherente, los errores incurridos durante el entrenamiento, tales como una articulación incorrecta, una mala pronunciación, el ruido del entorno u otros problemas se toleran y aún así se aceptan como patrones de referencia validos pues no hay posibilidad de corregirlos. Obviamente todo esto ocasiona un desempeño pobre bajo ciertas circunstancias [14].

7.8 CUANTIZACIÓN VECTORIAL MULTISECCIONADA

El enfoque convencional de la cuantización vectorial no esta diseñado para conservar las características secuenciales de los sonidos presentes en las palabras. No obstante, este problema se puede solucionar al tratar cada palabra o clase como si fuera la concatenación de varias fuentes secundarias de información, cada una de las cuales se representa mediante un conjunto de centroides o libros código distintos. Esta forma de realizar la cuantización vectorial por segmentos se conoce como la cuantización vectorial multiseccionada o **M**ulti**S**ectional **V**ector **Q**uantization (MSVQ) [14].

La manera más sencilla para descomponer una palabra en una concatenación de varias fuentes secundarias de información, es el dividirla en SC segmentos o secciones iguales. Se puede optar por determinar la cantidad de secciones para que sea igual a la cantidad de fonemas presentes en la palabra.

Con un determinado conjunto de vectores de entrenamiento para una misma palabra o clase, se forman los conjuntos de datos de entrenamiento correspondientes y se utilizan para calcular los *SC* libros código. Estos tienen de manera implícita una secuencia temporal porque corresponden a diferentes porciones de las palabras. De modo que cada conjunto de *SC* libros código sucesivos representa una palabra o clase, y la medida de distorsión obtenida al codificar una palabra desconocida con estos cuantizadores vectoriales sucesivos es el criterio de discriminación para la toma de decisión en el reconocimiento de voz [14].

En la figura 7.2 se muestra como se realiza la cuantización vectorial multiseccionada y su entrenamiento. También muestra la división en $SC = 5$ secciones de varias palabras con diferente longitud o duración y además muestra cómo se obtienen los centroides para cada una de las secciones.

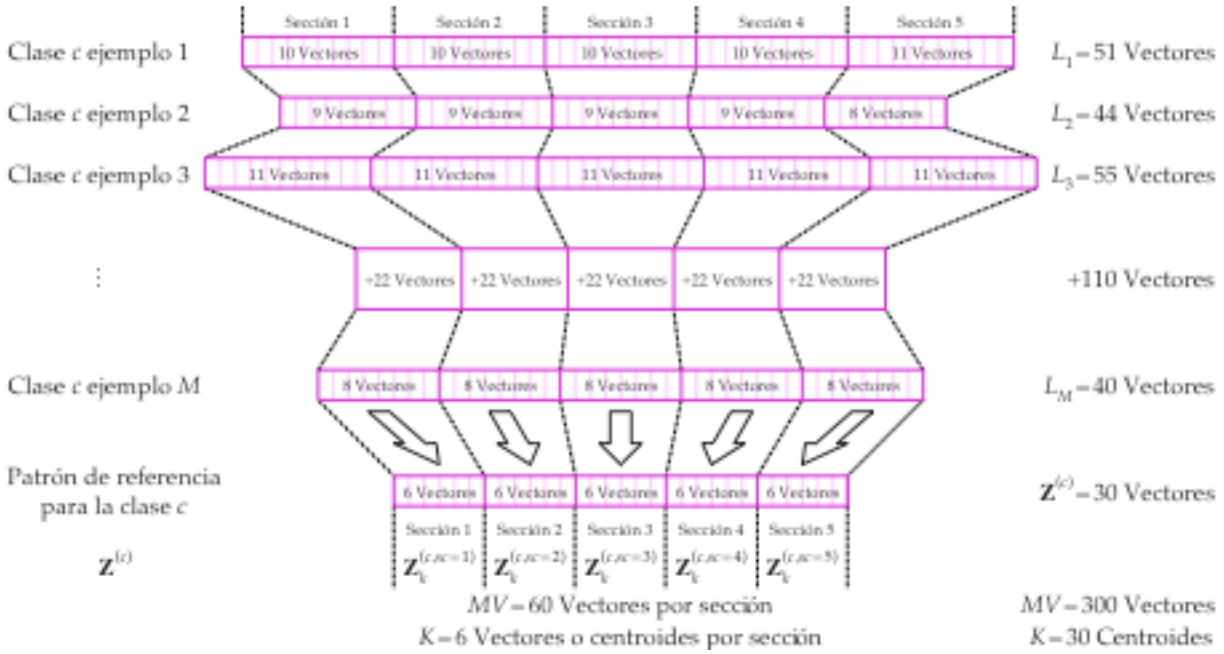


FIGURA 7.2 Ilustración de la cuantización vectorial multiseccionada y su entrenamiento

El mantener esta relación secuencial con la concatenación de los libros código a menudo resulta ser muy beneficioso en el reconocimiento de voz. Este método proporciona una simplicidad computacional para las aplicaciones de reconocimiento de voz [14].

CAPÍTULO

8

PRUEBAS CON EL SIMULADOR DE RECONOCIMIENTO DE VOZ

8.1 INTRODUCCIÓN

Durante un largo tiempo, las máquinas inteligentes capaces de entender la voz existían solamente en las obras de ciencia ficción, pues muy frecuentemente, la imaginación del hombre sobrepasa su capacidad tecnológica. Sin embargo, no hace mucho tiempo que algunos laboratorios y centros de investigación emprendieron la ardua tarea de hacer realidad el reconocimiento de voz.

La meta suprema de esta investigación, es el producir un sistema de reconocimiento de voz continua, en tiempo real, sin restricciones de vocabulario e independiente del locutor, es decir, una máquina o dispositivo que pueda ser capaz de reconocer de una manera exacta y confiable, cualquier palabra dicha durante el habla normal o cotidiana, sin importar quien sea la persona que este hablando, tal como lo haría cualquier ser humano.

Desafortunadamente, como ya se ha indicado, el problema que presenta el reconocimiento de voz ha resultado ser mucho más complicado de lo que pudiera parecer, pues a pesar de todos los esfuerzos realizados, todavía dista mucho de estar completamente resuelto. No obstante, estos trabajos de investigación han dado como resultado el desarrollo de varios sistemas de reconocimiento de voz, que se desempeñan de una manera aceptable para algunas aplicaciones, pero bajo ciertas limitaciones o restricciones.

8.2 SISTEMAS DE RECONOCIMIENTO DE VOZ

Para poder implementar efectivamente un sistema de reconocimiento de voz, primero se tiene que restringir la cantidad de variables o parámetros con los que se ha de trabajar, de tal manera que se ha de permitir solamente un cierto número de palabras para el vocabulario y también se ha de limitar la cantidad de locutores. De tal manera que de acuerdo a las restricciones, un sistema de reconocimiento puede clasificarse ya sea como dependiente o independiente del locutor, para el reconocimiento de palabras aisladas o de voz continua.

Un sistema dependiente del locutor, se hace para reconocer la voz de una sola persona, en cambio, un sistema independiente del locutor, se hace para reconocer las voces de varias personas. Además, un sistema dependiente del locutor se entrena antes de realizar el reconocimiento de voz, mediante la obtención y el almacenamiento de los patrones de referencia para cada una de las palabras del vocabulario, empleando la voz del usuario. Pero en un sistema independiente del locutor, puesto que ya se almacenó un patrón promedio para cada palabra del vocabulario, los usuarios del sistema pueden usarlo sin la necesidad de efectuar un entrenamiento previo.

En un sistema de reconocimiento de palabras aisladas (o de comandos), cada palabra del vocabulario se tiene que decir por separado, haciendo pausas entre las palabras, mientras que en los sistemas de reconocimiento de voz continua, las pausas no son necesarias y pueden aceptar oraciones completas.

8.3 RECONOCIMIENTO DE PATRONES

El reconocimiento de voz mediante la cuantización vectorial se puede realizar mediante los vectores de características, pues un vector de entrada puede ser comparado y aproximado por cualquiera de los patrones de referencia previamente almacenados. Este tipo de reconocimiento permite encontrar el patrón de referencia que más se parezca al vector de entrada [4]. En los esquemas de reconocimiento de patrones existen tres etapas básicas:

1. La obtención de parámetros.
2. La comparación contra los patrones de referencia.
3. La toma de decisión.

En la figura 8.1 se muestra un diagrama de bloques con un procedimiento para efectuar el reconocimiento de voz mediante la técnica de cuantización vectorial multiseccionada. Los bloques muestran las operaciones necesarias y a la izquierda de éstos se muestran los parámetros de entrada requeridos en cada paso.

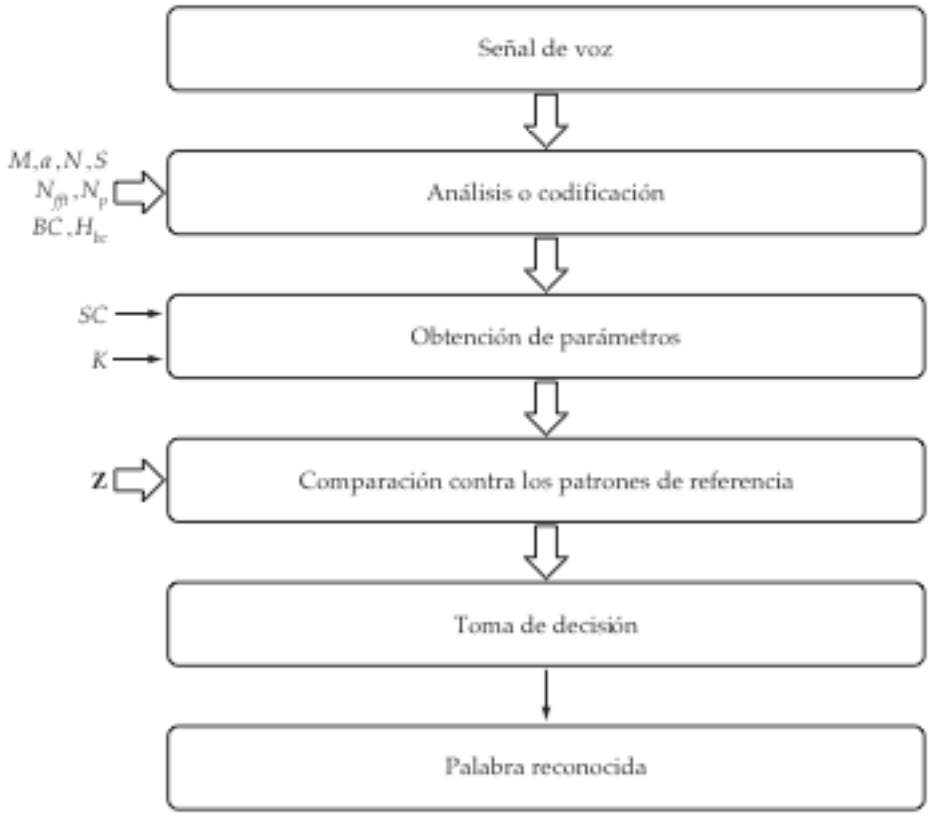


FIGURA 8.1 Procedimiento para el reconocimiento de voz mediante la técnica MSVQ

Como primer paso, se debe efectuar un análisis o codificación de la señal de voz $x(n)$ para poder obtener un conjunto de vectores de características $\mathbf{V}$, esto es, una secuencia de L vectores con M coeficientes:

$$\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \mathbf{V}_3, \ldots, \mathbf{V}_L] \qquad \ell = 1,2,3,\ldots,L \tag{8.1}$$

$$\mathbf{V}_\ell = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_M \end{bmatrix} \qquad \begin{matrix} \ell = 1,2,3,\ldots,L \\ m = 1,2,3,\ldots,M \end{matrix} \tag{8.2}$$

Al dividir esta secuencia de L vectores en SC secciones, se pueden calcular los K centroides por sección, para conseguir un vector de entrada $\mathbf{V}^{(x)}$, esto es, un conjunto de SC vectores con K componentes, de M coeficientes cada uno:

$$\mathbf{V}^{(x)} = \begin{bmatrix} \mathbf{V}^{(x,1)} \\ \mathbf{V}^{(x,2)} \\ \mathbf{V}^{(x,3)} \\ \vdots \\ \mathbf{V}^{(x,SC)} \end{bmatrix} \qquad sc = 1,2,3,\ldots,SC \tag{8.3}$$

$$\mathbf{V}^{(x,sc)} = \left[\mathbf{V}_1^{(x,sc)}, \mathbf{V}_2^{(x,sc)}, \mathbf{V}_3^{(x,sc)}, \ldots, \mathbf{V}_K^{(x,SC)}\right] \qquad \begin{matrix} sc = 1,2,3,\ldots,SC \\ k = 1,2,3,\ldots,K \end{matrix} \tag{8.4}$$

$$\mathbf{V}_k^{(x,sc)} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_M \end{bmatrix} \qquad \begin{matrix} sc = 1,2,3,\ldots,SC \\ k = 1,2,3,\ldots,K \\ m = 1,2,3,\ldots,M \end{matrix} \tag{8.5}$$

La función de la etapa de obtención de parámetros, es la de representar una secuencia de vectores de características $\mathbf{V}$ en términos de un vector de entrada $\mathbf{V}^{(x)}$, es decir, mediante un conjunto más compacto de parámetros.

En la etapa de comparación contra los patrones de referencia, se calcula un conjunto de distancias o medidas de distorsión $\mathbf{D}$ entre el vector de entrada $\mathbf{V}^{(x)}$ y cada uno de los patrones de referencia $\mathbf{Z}^{(c)}$, correspondientes a cada una de las C palabras o clases contenidas dentro del vocabulario:

$$\mathbf{D} = \left[d^{(1)}, d^{(2)}, d^{(3)}, \ldots, d^{(C)}\right] \qquad c = 1,2,3,\ldots,C \tag{8.6}$$

En donde $d^{(c)}$ son las distancias por palabra o clase. Estas se calculan como la suma de las SC distancias por sección $d^{(c,sc)}$ para una palabra o clase:

$$\begin{aligned} d^{(c)} &= d^{(c,1)} + d^{(c,2)} + d^{(c,3)} + \ldots + d^{(c,SC)} \\ d^{(c)} &= \sum_{sc=1}^{SC} d^{(c,sc)} \end{aligned} \qquad \begin{matrix} c = 1,2,3,\ldots,C \\ sc = 1,2,3,\ldots,SC \end{matrix} \tag{8.7}$$

Las distancias por sección para una palabra o clase $d^{(c,sc)}$, se pueden determinar al seleccionar el valor mínimo de las distancias entre los K centroides del vector de entrada y de los K centroides del patrón de referencia, para una misma sección sc:

$$d^{(c,sc)} = \mathbf{min}\{d_1, d_2, d_3, \ldots, d_K\} \qquad d_k = d\left(\mathbf{V}_j^{(x,sc)}, \mathbf{Z}_k^{(c,sc)}\right) \qquad \begin{matrix} j = 1,2,3,\ldots,K \\ k = 1,2,3,\ldots,K \end{matrix} \tag{8.8}$$

La etapa de toma de decisión, es básicamente, la regla para decidir cual de los patrones de referencia se parece más al vector de entrada desconocido. Una de las reglas que se emplean con mayor frecuencia es la del vecino más cercano (NN : Nearest Neighbor), esto es, elegir el índice i del patrón de referencia con la mínima distancia como la palabra o clase reconocida:

$$i = \arg\min\{\mathbf{D}\}$$
$$i = \arg\min\left\{d^{(1)}, d^{(2)}, d^{(3)}, \ldots, d^{(C)}\right\} \qquad i = 1,2,3,\ldots,C \qquad \textbf{(8.9)}$$

Lo que se busca es una representación que proporcione la mejor discriminación entre diferentes clases de patrones, que además sean insensibles a sus variaciones, de tal manera que se puedan minimizar los errores en el reconocimiento.

8.4 DISEÑO DE LAS PRUEBAS

Para poder darle una aplicación a las codificaciones de voz, se propone realizar tres pruebas, una por cada una de estas técnicas de codificación, empleando un programa en computadora que simule un sistema de reconocimiento de voz, para así poder evaluar y comparar cuales técnicas proporcionaron la mayor cantidad de palabras reconocidas.

El primer paso consiste en definir las características del sistema de reconocimiento de voz. Como se pretende emplear las voces de diferentes personas para realizar un reconocimiento de unas cuantas palabras, se trata de un sistema de reconocimiento de palabras aisladas (o de comandos) independiente del locutor.

El vocabulario es bastante reducido, se proponen $C = 12$ palabras o clases distintas compuestas por cinco fonemas. Estas se escogieron de tal modo que en cada palabra no se repitan los mismos fonemas, pero que en conjunto, formen una muestra, pues incluyen por lo menos un ejemplo de cada fonema. Además se trato de que estas siguieran un patrón semejante, con dos vocales y tres consonantes alternadas (es decir, sin diptongos), con una sola excepción. En la tabla 8.1 se muestra la descripción por fonemas de las palabras del vocabulario. Por otra parte, la cantidad de locutores también es reducida, se propone utilizar como locutores a $P = 12$ personas diferentes, seis mujeres y seis hombres de distintas edades, cada uno de los cuales se puede agrupar dentro de uno de los seis rangos de edad que se muestran en la tabla 8.2, donde se muestra la descripción de los locutores.

TABLA 8.1 Descripción por fonemas de las palabras del vocabulario

	Chispa	Digno	Flecha	Fugaz	Líder	Lluvia	Mejor	Monte	Punto	Soñar	Subir	Yunque	
Fonema	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	**Cant.**
/ i /	✓	✓			✓	✓					✓		5
/ e /			✓		✓		✓	✓				✓	5
/ a /	✓		✓	✓		✓				✓			5
/ o /		✓					✓	✓	✓	✓			5
/ u /				✓		✓			✓		✓	✓	5
/ f /			✓	✓									2
/ j /												✓	1
/θ/				✓									1
/ s /	✓									✓	✓		3
/χ/							✓						1
/ b /						✓					✓		2
/ d /		✓			✓								2
/ g /		✓		✓									2
/ p /	✓								✓				2
/ t /								✓	✓				2
/ k /												✓	1
/ c /	✓		✓										2
/ l /			✓		✓								2
/λ/						✓							1
/ r /					✓		✓			✓	✓		4
/ m /							✓	✓					2
/ n /		✓						✓	✓			✓	4
/η/										✓			1

TABLA 8.2 Descripción de los locutores

Locutor	Iniciales	Género	Edad [años]	Grupo
1	KVS	Femenino	5 ≤ edad < 12	I
2	KPM	Femenino	12 ≤ edad < 20	II
3	AMA	Femenino	20 ≤ edad < 30	III
4	AMG	Femenino	30 ≤ edad < 45	IV
5	GRQ	Femenino	45 ≤ edad < 60	V
6	EBP	Femenino	edad ≥ 60	VI
7	EVB	Masculino	5 ≤ edad < 12	I
8	AYG	Masculino	12 ≤ edad < 20	II
9	MMC	Masculino	20 ≤ edad < 30	III
10	ABR	Masculino	30 ≤ edad < 45	IV
11	FBP	Masculino	45 ≤ edad < 60	V
12	EBC	Masculino	edad ≥ 60	VI

Para cada uno de los locutores, se realizo una grabación por cada palabra o clase, por lo que se cuenta con doce ejemplos distintos para cada una de las $C = 12$ clases, de tal manera que en total, se realizaron 144 grabaciones a una frecuencia de muestreo de $f_s = 8000$ Hz.

El segundo paso consiste en el análisis o la codificación de las señales de voz para obtener los vectores de características LPC, MFCC y PLP. A fin de mantener las mismas condiciones de prueba, se ha de procurar que las tres pruebas sean lo más parecidas en la medida de lo posible.

Para LPC y MFCC el parámetro de pre-énfasis es $a = 0.95$, PLP usa otra función de pre-énfasis, pero se supone que tiene un efecto equivalente al del filtro FIR paso altas de primer orden [5].

Para obtener una división en tramas consistente, en todas las pruebas se usa la misma longitud para las tramas $N = 240$ muestras $(30\,\text{ms})$, con $S = 80$ muestras $(10\,\text{ms})$ de separación entre tramas adyacentes, que proporcionan un traslape entre tramas adyacentes del 33% con una tasa de $Fps = 100$ tramas por segundo.

Para MFCC y PLP se emplean $N_{fft} = 256$ puntos para la transformada rápida de Fourier a fin de obtener la misma resolución en frecuencia. La cantidad de filtros en PLP está en función de la frecuencia de muestreo, entonces para una frecuencia de muestreo de $f_s = 8000\,\text{Hz}$, se tiene que el banco consiste de $BC = 15 + 2$ filtros trapezoidales. Pero para MFCC se suelen usar de 20 a 40 filtros. A pesar de que las formas de los filtros no son iguales, para que éstos resulten equivalentes, también se emplea la misma cantidad de $BC = 17$ filtros triangulares, con los que además, se abarca el mismo rango de frecuencias, desde $f_{min} = 0\,\text{Hz}$ hasta $f_{max} = 4000\,\text{Hz}$. Estos parámetros no aplican para LPC porque no se efectúa ningún análisis en frecuencia.

Para MFCC y PLP el número de coeficientes M tiene que ser menor o igual a la cantidad de filtros. En tal caso el número máximo de coeficientes es de $M = 17$. Aunque se pueden obtener más coeficientes LPC, un predictor de mayor orden no proporciona ninguna mejora en el reconocimiento. Por lo tanto, se emplean como máximo, hasta $M = 17$ coeficientes para todas las pruebas. En la tabla 8.3 se resumen los valores propuestos para los parámetros de las codificaciones.

TABLA 8.3 Resumen de los valores propuestos para los parámetros de las codificaciones

Parámetro	LPC	MFCC	PLP
a	0.95	0.95	N / A
N	240	240	240
S	80	80	80
Nfft	N / A	256	256
fs	8000	8000	8000
fmin	N / A	0	0
fmax	N / A	4000	4000
BC	N / A	17	15 + 2
Hbc	N / A	Triangulares / mel	Trapezoidales / Bark
M	1 : 1 : 17	1 : 1 : 17	1 : 1 : 17

El tercer y último paso consiste en definir las características del cuantizador vectorial. Puesto que las palabras o clases se componen por cinco fonemas, los vectores de características se tienen que dividir en $SC = 5$ secciones iguales. La locución más larga se compone por $L_{max} = 111$ tramas, el valor promedio es de $\overline{L} = 60$ tramas por palabra, pero la más corta se compone de $L_{min} = 40$ tramas. Entonces para la primera, cada sección consiste de $MV_{max} \approx 22$ vectores, para el promedio, cada sección consiste de $MV_{avg} = 12$ vectores, pero para la tercera, cada sección consiste de $MV_{min} = 8$ vectores. Por lo tanto, solamente se utiliza $K = 1$ centroide por sección.

Aunque cada codificación tiene asociada una medida de distancia, se propone utilizar la misma medida de distancia para las tres pruebas, la distancia cuadrática media, de tal modo que lo único que se cambie sea la técnica de codificación.

De tal manera que el resultado obtenido en la etapa de obtención de parámetros es un vector o patrón de entrada $\mathbf{V}^{(x)}$ compuesto por $SC = 5$ vectores de hasta $M = 17$ coeficientes:

$$\mathbf{V}^{(x)} = \left[\mathbf{V}^{(x,1)}, \mathbf{V}^{(x,2)}, \mathbf{V}^{(x,3)}, \mathbf{V}^{(x,4)}, \mathbf{V}^{(x,5)}\right]$$

$$\mathbf{V}^{(x,sc)} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ \vdots \\ v_M \end{bmatrix} \qquad \begin{matrix} sc = 1,2,3,4,5 \\ m = 1,2,3,\ldots,17 \end{matrix} \qquad \textbf{(8.10)}$$

Para el entrenamiento, se propone tomar a uno de los locutores, como si se tratara de un tipo de entrenamiento casual, para obtener los patrones de referencia $\mathbf{Z}^{(c)}$ para cada palabra o clase, los cuales también están compuestos por $SC = 5$ vectores de hasta $M = 17$ coeficientes:

$$\mathbf{Z}^{(c)} = \left[\mathbf{Z}^{(c,1)}, \mathbf{Z}^{(c,2)}, \mathbf{Z}^{(c,3)}, \mathbf{Z}^{(c,4)}, \mathbf{Z}^{(c,5)}\right]$$

$$\mathbf{Z}^{(c,sc)} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_M \end{bmatrix} \qquad \begin{aligned} c &= 1,2,3,\ldots,12 \\ sc &= 1,2,3,4,5 \\ m &= 1,2,3,\ldots,17 \end{aligned} \qquad \textbf{(8.11)}$$

La razón para efectuar un mínimo de entrenamiento, es para poder apreciar el desempeño individual de cada uno de los locutores como patrón de referencia. Con estos patrones de referencia ya es posible realizar el reconocimiento de voz.

8.5 DESCRIPCIÓN DE LAS PRUEBAS

Una sola prueba con el programa de simulación de reconocimiento de voz abarca X experimentos, en los que se varia la cantidad o el número de coeficientes M del vector de características, desde 1 hasta el valor de X. A su vez, un sólo experimento se compone de P pasos diferentes, esto es, un paso por cada uno de los locutores. A su vez, un sólo paso consiste en C casos distintos, es decir, un caso por cada palabra o clase en el vocabulario. Finalmente un sólo caso está formado por $(P-1)$ ejercicios de reconocimiento.

De está manera, se emplea cada una de las palabras de cada uno de los locutores como patrón de referencia para realizar el reconocimiento de las otras palabras dichas por los demás locutores. Puesto que es prácticamente imposible tener dos locuciones idénticas, no tiene sentido comparar el patrón de referencia contra si mismo, por esta razón solamente se efectúan $(P-1)$ ejercicios de reconocimiento.

A continuación se describe paso a paso, el procedimiento mediante el cual se obtuvieron los resultados para un experimento con el simulador de reconocimiento de voz, empleando $M = 8$ coeficientes para la técnica de codificación MFCC.

8.5.1 UN CASO

En la figura 8.2 se presentan los resultados correspondientes al caso número 7 del paso número 12 del experimento en cuestión, lo cual significa que se está utilizando al locutor número 12 (EBC) como patrón de referencia para las clases (el paso número 12), para realizar el reconocimiento de la palabra o clase número 7 *"mejor"* (el caso número 7), empleando los otros $(P-1)=11$ ejemplos de la misma palabra pero de los demás locutores.

```
CODIFICACION MFCC CON  8 COEFICIENTES
PASO 12 CASO  7   REFERENCIA : LOCUTOR 12 EBC , PALABRA 7 : mejor
ENTRADA RESULTADO PRUEBA  chispa digno                  mejor          soñar
 1 KVS    mejor   Ok      [  447  636  394  525  483  434 254  476  507  304  816  514 ]
 2 KPM    mejor   Ok      [  346  455  339  821  529  388 273  510  435  338  947  501 ]
 3 AMA    mejor   Ok      [  500  776  455  451  340  481 152  467  727  219  568  469 ]
 4 AMG    mejor   Ok      [  483  513  469  753  374  411 291  329  484  339  842  335 ]
 5 GRQ    mejor   Ok      [  433  523  334  483  375  433 169  454  450  350  843  468 ]
 6 EBP    soñar   ERROR!  [  639  807  523  365  372  477 299  466  759  283  496  512 ]
 7 EVB    chispa  ERROR!  [  330  483  404  981  616  567 359  540  509  519 1106  486 ]
 8 AYG    mejor   Ok      [  350  328  355  913  509  389 294  419  336  476 1005  382 ]
 9 MMC    digno   ERROR!  [  479  403  632 1239  841  542 453  571  415  694 1329  562 ]
10 ABR    mejor   Ok      [  425  406  358  702  449  389 236  402  362  411  970  416 ]
11 FBP    mejor   Ok      [  470  517  500  739  480  480 248  470  501  516  978  449 ]
12 EBC    .....
```

FIGURA 8.2 Resultados para un caso con el simulador de reconocimiento de voz

Los resultados se agrupan en columnas. La primera, corresponde al locutor de origen del vector de entrada. La segunda, es el resultado o la palabra reconocida. La tercera, es el resultado de una prueba de comprobación, las leyendas `Ok` y `ERROR!` especifican si la palabra se identifico correctamente o erróneamente. Por último, se muestran otras doce columnas que corresponden a las distancias para cada una de las $C=12$ palabras o clases en el vocabulario. El criterio de decisión es el del vecino más cercano, así que la palabra reconocida es la que tiene la mínima distancia. Para mayor claridad, se incluyen cuatro etiquetas para señalar las columnas correspondientes para las palabras reconocidas, las distancias mínimas y la palabra correcta se destacan en color negro, pero las otras distancias y las palabras incorrectas se muestran con un color gris.

8.5.2 UN PASO

En la tabla 8.4, se presentan los resultados correspondientes al paso número 12 del experimento en cuestión, en donde se está utilizando al locutor número 12 (EBC) como patrón de referencia para cada uno de los $C = 12$ casos.

TABLA 8.4 Resultados para un paso con el simulador de reconocimiento de voz

Palabra		Chispa	Digno	Flecha	Fugaz	Líder	Lluvia	Mejor	Monte	Punto	Soñar	Subir	Yunque	
Entrada		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*AL*
KVS	*1*			✓				✓		✓	✓		✓	5
KPM	*2*	✓		✓				✓	✓	✓	✓		✓	7
AMA	*3*	✓		✓		✓		✓	✓	✓	✓	✓		8
AMG	*4*	✓						✓	✓	✓			✓	5
GRQ	*5*			✓		✓	✓	✓	✓	✓	✓	✓		8
EBP	*6*		✓	✓	✓	✓	✓		✓	✓	✓	✓		9
EVB	*7*			✓		✓			✓	✓			✓	5
AYG	*8*	✓	✓	✓		✓		✓	✓	✓			✓	8
MMC	*9*	✓	✓	✓		✓			✓	✓	✓		✓	8
ABR	*10*	✓	✓	✓	✓		✓	✓	✓	✓	✓			9
FBP	*11*	✓	✓	✓				✓	✓	✓		✓		7
EBC	*12*													
AC		7	5	10	2	6	3	8	10	11	7	4	6	*79*

Una palomita (✓) indica un acierto en el reconocimiento, para mayor claridad, se han omitido los errores (×), dejando las celdas correspondientes en blanco. Las columnas representan los resultados obtenidos en cada uno de los casos, de tal manera que los resultados para el caso anterior (el caso 7) corresponden a la séptima columna de la tabla. Los renglones se pueden interpretar como los resultados del reconocimiento de las palabras o clases de un locutor en particular contra el de referencia. Por lo tanto, la columna del extremo derecho es la cantidad de aciertos por locutor (*AL*), pero el renglón inferior es la cantidad de aciertos por clase (*AC*). La celda de la esquina inferior derecha es el total de aciertos o palabras reconocidas para el paso. El total de palabras a reconocer para un paso es de $C \cdot (P-1) = 132$ palabras. Evidentemente, algunas palabras y ciertas voces son más difíciles de reconocer que otras.

8.5.3 UN EXPERIMENTO

En la tabla 8.5 se presentan los resultados obtenidos con vectores MFCC de ocho coeficientes, correspondientes al experimento número 8, en el que se están utilizando todos los locutores como patrones de referencia para cada uno de los $P = 12$ pasos.

TABLA 8.5 Resultados para un experimento con el simulador de reconocimiento de voz

Palabra		Chispa	Digno	Flecha	Fugaz	Líder	Lluvia	Mejor	Monte	Punto	Soñar	Subir	Yunque	
Entrada		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*SAR*
KVS	*1*	11	10	3	10	2	3	6	3	8	8	5	2	71
KPM	*2*	7	11	1	11	3	0	7	11	8	10	8	4	81
AMA	*3*	10	8	5	10	6	9	6	1	9	2	6	1	73
AMG	*4*	10	5	8	11	6	5	8	7	11	3	7	3	84
GRQ	*5*	4	10	2	7	7	8	9	6	11	3	4	9	80
EBP	*6*	11	11	4	8	9	4	2	1	4	5	7	10	76
EVB	*7*	10	9	7	11	9	4	8	5	11	6	9	6	95
AYG	*8*	7	8	10	11	9	6	7	9	11	7	8	6	99
MMC	*9*	6	8	10	11	8	1	5	8	11	6	3	6	83
ABR	*10*	10	9	4	7	5	3	8	3	9	9	10	8	85
FBP	*11*	8	10	7	11	9	3	8	4	10	5	4	2	81
EBC	*12*	7	5	10	2	6	3	8	10	11	7	4	6	79
SAC		101	104	71	110	79	49	82	68	114	71	75	63	**987**

Los renglones representan los resultados obtenidos en cada uno de los pasos, de tal modo que la cantidad de aciertos por clase (*AC*) del paso anterior (el paso 12) corresponde al doceavo renglón. Las columnas representan la cantidad de aciertos obtenidos al comparar una palabra en particular contra un locutor de referencia. Por lo tanto, el renglón inferior es la suma de los aciertos por clase (*SAC*), pero la columna del extremo derecho es la suma de los aciertos por locutor de referencia (*SAR*). La celda de la esquina inferior derecha es el total de aciertos o palabras reconocidas para el experimento. El total de palabras a reconocer para un experimento es de $P \cdot C \cdot (P-1) = 1584$ palabras.

8.5.4 UNA PRUEBA

En las siguientes dos tablas se indican los resultados obtenidos para una prueba con vectores MFCC con $X = 8$ experimentos con el simulador de reconocimiento de voz. En la primera, la tabla 8.6, los datos se encuentran ordenados por locutor de referencia, pero en la segunda, la tabla 8.7, los datos se encuentran ordenados por palabra o clase.

TABLA 8.6 Resultados ordenados por locutor para un a prueba con el simulador

	KVS	KPM	AMA	AMG	GRQ	EBP	EVB	AYG	MMC	ABR	FBP	EBC	
M	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***A***
1	35	45	51	58	41	56	53	57	46	45	56	51	594
2	71	75	72	74	67	68	84	92	78	77	72	66	896
3	67	75	80	82	75	69	86	91	79	77	68	76	925
4	71	80	76	81	79	76	93	97	81	79	79	82	974
5	73	81	76	82	79	76	93	99	79	79	80	81	978
6	73	82	75	82	78	76	91	100	83	81	82	83	986
7	70	81	74	83	77	77	93	99	84	80	82	82	982
8	71	81	73	84	80	76	95	99	83	85	81	79	987
TAR	531	600	577	626	576	574	688	734	613	603	600	600	***7322***

TABLA 8.7 Resultados ordenados por palabra o clase para un a prueba con el simulador

	Chispa	Digno	Flecha	Fugaz	Líder	Lluvia	Mejor	Monte	Punto	Soñar	Subir	Yunque	
M	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***A***
1	78	58	40	86	49	27	32	30	92	38	37	27	594
2	86	103	62	104	83	40	77	62	96	54	74	55	896
3	89	100	65	109	73	46	80	66	115	60	67	55	925
4	96	107	70	109	79	47	85	68	115	70	71	57	974
5	98	106	70	109	76	48	88	66	114	69	71	63	978
6	99	104	71	110	79	50	86	67	116	68	73	63	986
7	100	105	70	109	77	49	85	67	114	69	72	65	982
8	101	104	71	110	79	49	82	68	114	71	75	63	987
TAR	747	787	519	846	595	356	615	494	876	499	540	448	***7322***

Los renglones representan los resultados obtenidos en cada uno de los experimentos, de tal manera que las sumas de los aciertos del experimento anterior (el experimento número 8) corresponden al octavo renglón, para la primera tabla, se trata de la suma de los aciertos por locutor de referencia (*SAR*), pero para la segunda, ahora se trata de la suma de los aciertos por clase (*SAC*). Las columnas representan la cantidad de aciertos obtenidos al emplear un locutor o una clase en particular como patrón de referencia. Por lo tanto, las columnas del extremo derecho proporcionan la cantidad de aciertos o palabras reconocidas (*A*), pero los renglones inferiores suministran el total de aciertos por locutor de referencia (*TAR*) para la primera tabla, así como el total de aciertos por clase (*TAC*) para la segunda tabla. El total de palabras a reconocer para esta prueba es de $X \cdot P \cdot C \cdot (P-1) = 12{,}672$ palabras.

8.5.5 LA MATRIZ DE CONFUSIÓN

Durante cada ejercicio de reconocimiento se puede llevar un registro de las palabras que causaron confusión, es decir, de las palabras que se reconocieron erróneamente. La tabla 8.8 muestra la matriz de confusión resultante para el experimento número 8.

TABLA 8.8 Matriz de confusión para un experimento con el simulador

		Chispa	Digno	Flecha	Fugaz	Líder	Lluvia	Mejor	Monte	Punto	Soñar	Subir	Yunque	
		1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***EC***
Chispa	***1***		12	23		5	1	3			8	2		54
Digno	***2***	7		4		8	9	13	1	8	2	1	2	55
Flecha	***3***	15	4			11	12	1	1		6	4	3	57
Fugaz	***4***			5		5	5	19	2		26	10		72
Líder	***5***		3	3			1							7
Lluvia	***6***		1	10	1	10		5	10	5	9	6	19	76
Mejor	***7***	2	4	5	2	2	1		7	4			1	28
Monte	***8***		1			2	9	2				1	28	43
Punto	***9***		3	1		2	11	5	9				1	32
Soñar	***10***	7		7	19	2	7	2	2			22	10	78
Subir	***11***						3		1		6		5	15
Yunque	***12***			3		6	24		31	1	4	11		80
ER		31	28	61	22	53	83	50	64	18	61	57	69	***597***

Las celdas representan la cantidad de errores cometidos al confundir una palabra por otra durante el reconocimiento. La cabecera de las columnas corresponde al patrón de referencia y la cabecera de los renglones corresponde a la palabra reconocida erróneamente. Para mayor claridad, se han dejando en blanco las celdas con valores nulos. Por lo tanto, el renglón inferior corresponde a la cantidad de errores por patrón de referencia (*ER*), pero la columna del extremo derecho corresponde a la cantidad de errores por palabra o clase (*EC*). La celda de la esquina inferior derecha es el total de errores para el experimento, el cual es el complemento del total de aciertos o palabras reconocidas durante el experimento. Es fácil observar que algunas palabras causan mayor confusión que otras.

8.6 RESULTADOS DE LAS PRUEBAS

En las tablas que se muestran a continuación en las siguientes secciones, se presentan los resultados obtenidos para las diferentes pruebas con el simulador de reconocimiento de voz. Para cada una de estas, básicamente, sólo se modificó la técnica de codificación empleada. De tal manera que se realizaron tres pruebas, una por cada técnica de codificación. La primera fue con LPC, la segunda con MFCC y la tercera con PLP.

Cada una de las pruebas consiste en $X = 17$ experimentos con $P = 12$ locutores y para un vocabulario con $C = 12$ palabras o clases. En la tabla 8.9 se muestra la relación entre las distintas etapas de una prueba así como la cantidad correspondiente de palabras a reconocer.

TABLA 8.9 Relación entre las etapas de una prueba y la cantidad de palabras a reconocer

Descripción	Equivalente	Cantidad de palabras a reconocer
Una prueba	X experimentos	$X \cdot P \cdot C \cdot (P-1) = 26{,}928$
Resultados por clase para una prueba	X experimentos $\div$ C clases	$X \cdot P \cdot (P-1) = 2244$
Resultados por locutor para una prueba	X experimentos $\div$ P locutores	$X \cdot C \cdot (P-1) = 2244$
Un experimento	P pasos	$P \cdot C \cdot (P-1) = 1584$
Un paso	C casos	$C \cdot (P-1) = 132$
Un caso	$(P-1)$ ejercicios de reconocimiento	$(P-1) = 11$
Un ejercicio de reconocimiento	N / A	1

Los renglones y las columnas correspondientes a los valores máximos se destacan en las tablas mediante letras cursivas y negritas.

8.6.1 PRUEBA DEL SIMULADOR CON LPC

En la tabla 8.10 se muestran los resultados obtenidos para la prueba con el simulador de reconocimiento de voz empleando la técnica de codificación LPC. Estos datos se encuentran ordenados de acuerdo a los locutores.

TABLA 8.10 Resultados ordenados por locutor para la prueba del simulador con LPC

	KVS	KPM	AMA	AMG	GRQ	EBP	EVB	*AYG*	MMC	ABR	FBP	EBC	
M	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***A***
1	51	48	55	58	61	56	59	***63***	49	46	68	59	673
2	51	45	55	56	59	56	54	***62***	51	58	73	64	684
3	47	44	55	63	61	60	58	***71***	56	55	71	63	704
4	41	53	60	66	56	54	67	***63***	55	55	63	54	687
5	44	56	61	68	60	60	58	***64***	58	59	67	51	706
6	48	62	62	73	59	64	64	***67***	60	67	72	55	753
7	42	67	63	72	59	68	72	***73***	67	69	78	61	791
8	47	68	63	77	65	69	75	***78***	78	67	72	70	829
9	48	74	58	77	67	68	71	***74***	80	70	71	67	825
10	***47***	***77***	***60***	***80***	***69***	***68***	***67***	***78***	***79***	***71***	***74***	***69***	***839***
11	51	70	61	72	68	67	67	***79***	83	73	69	70	830
12	50	65	60	73	66	69	64	***79***	83	75	70	67	821
13	53	67	59	71	64	70	66	***77***	84	74	70	66	821
14	52	69	60	69	66	70	65	***79***	81	73	68	73	825
15	49	70	58	70	67	70	66	***80***	83	77	68	75	833
16	47	72	61	69	66	70	64	***78***	82	73	65	76	823
17	42	70	60	72	62	71	64	***75***	81	74	65	74	810
TAR	810	1077	1011	1186	1075	1110	1101	***1240***	1210	1136	1184	1114	**13,254**

La mayor cantidad de aciertos o palabras reconocidas fue de $A = 839$ palabras de un total de 1584 palabras a reconocer por experimento, este resultado se obtuvo utilizando vectores con $M = 10$ coeficientes LPC. Por otra parte, el valor más alto para el total de aciertos por locutor de referencia fue de $TAR = 1240$ palabras, para el locutor número 8 (AYG), pero el valor más bajo fue de $TAR = 810$ palabras, para la locutora número 1 (KVS), en ambos casos estos locutores se usaron como patrones de referencia para tratar de reconocer un total de 2244 palabras por locutor de referencia.

En la Tabla 8.11 se muestran los mismos datos de la tabla anterior, pero ahora se encuentran ordenados de otra manera diferente, de acuerdo a las doce palabras o clases del vocabulario.

TABLA 8.11 Resultados ordenados por palabra o clase para la prueba del simulador con LPC

	Chispa	Digno	Flecha	*Fugaz*	Líder	Lluvia	Mejor	Monte	Punto	Soñar	Subir	Yunque	
M	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***A***
1	85	58	56	***63***	53	34	56	49	113	41	40	25	673
2	85	69	58	***84***	55	34	74	34	80	48	39	24	684
3	86	78	58	***82***	45	33	60	52	88	53	37	32	704
4	77	73	48	***96***	43	37	62	47	75	56	30	43	687
5	73	72	52	***101***	47	32	64	47	83	56	26	53	706
6	72	81	56	***100***	55	32	70	52	87	54	30	64	753
7	71	87	57	***100***	62	37	73	57	91	57	31	68	791
8	74	92	62	***100***	54	49	77	57	103	65	28	68	829
9	76	89	60	***102***	50	48	81	57	104	62	27	69	825
10	***78***	***85***	***55***	***103***	***61***	***53***	***86***	***57***	***98***	***64***	***31***	***68***	***839***
11	76	83	61	***102***	54	53	87	54	99	67	28	66	830
12	75	84	56	***105***	47	51	85	52	102	66	29	69	821
13	76	84	55	***104***	49	53	85	51	104	64	27	69	821
14	74	87	55	***102***	61	55	86	49	101	62	26	67	825
15	77	86	55	***102***	59	54	85	50	103	64	29	69	833
16	76	87	53	***99***	57	52	86	52	100	65	26	70	823
17	75	81	53	***99***	57	50	86	51	102	63	26	67	810
TAC	1306	1376	950	***1644***	909	757	1303	868	1633	1007	510	991	**13,254**

La palabra más fácil de reconocer con $TAC = 1644$ aciertos fue la número 4 *"fugaz"*, pero la más difícil de reconocer con $TAC = 510$ aciertos fue la número 11 *"subir"*, en ambos casos el total de palabras a reconocer por cada clase era de 2244 palabras por clase.

En las siguientes dos tablas, se muestran los resultados obtenidos para el mejor experimento con el simulador de reconocimiento de voz, con el que se consiguió el mayor número de aciertos, utilizando el número óptimo de $M = 10$ coeficientes LPC. La tabla 8.12 indica los resultados o aciertos y la tabla 8.13 muestra la matriz de confusión.

TABLA 8.12 Resultados del experimento para el número óptimo de coeficientes LPC

Palabra Entrada		Chispa 1	Digno 2	Flecha 3	*Fugaz* 4	Líder 5	Lluvia 6	Mejor 7	Monte 8	Punto 9	Soñar 10	Subir 11	Yunque 12	*SAR*
KVS	1	6	7	0	**7**	0	0	6	1	9	7	1	3	47
KPM	2	0	10	2	**9**	8	4	10	6	7	8	4	9	77
AMA	3	8	9	2	**10**	2	9	3	5	2	4	2	4	60
AMG	4	**9**	**6**	**8**	**11**	**2**	**5**	**8**	**8**	**8**	**7**	**3**	**5**	**80**
GRQ	5	4	6	3	**6**	6	8	9	5	10	1	3	8	69
EBP	6	11	7	3	**10**	5	6	1	2	9	1	3	10	68
EVB	7	8	3	6	**11**	7	3	8	1	10	5	3	2	67
AYG	8	4	7	10	**11**	5	4	8	9	11	5	1	3	78
MMC	9	7	10	6	**10**	7	5	6	7	7	6	2	6	79
ABR	10	9	5	2	**9**	8	1	8	3	7	8	4	7	71
FBP	11	5	9	7	**9**	7	4	10	3	9	4	3	4	74
EBC	12	7	6	6	**0**	4	4	9	7	9	8	2	7	69
SAC		78	85	55	**103**	61	53	86	57	98	64	31	68	**839**

La mayor cantidad de aciertos fue de $SAR = 80$ palabras, para la locutora número 4 (AMG), pero la menor fue de $SAR = 47$ palabras, para la locutora número 1 (KVS). Por otra parte, la palabra más fácil de reconocer con $SAC = 103$ aciertos fue la número 4 *"fugaz"*, pero la más difícil de reconocer con $SAC = 31$ aciertos fue la número 11 *"subir"*.

TABLA 8.13 Matriz de confusión para el número óptimo de coeficientes LPC

		Chispa 1	Digno 2	Flecha 3	Fugaz 4	Líder 5	Lluvia 6	Mejor 7	Monte 8	Punto 9	Soñar 10	Subir 11	Yunque 12	*EC*
Chispa	1		15	20	2			5	1		11			54
Digno	2	9		3		4	2	23		5			1	47
Flecha	3	19	2		5	11	16	2	1	3	8		10	77
Fugaz	4						7	4			18			29
Líder	5	12	15	19	4		5	1	3	5	1		6	71
Lluvia	6	1		7	8	5		6	8	13	7	2	22	79
Mejor	7	3	11		15	1	3		6	4	3			46
Monte	8				5		18	9		9	3		31	75
Punto	9	1	4	1		1	8	6	11				2	34
Soñar	10	10		7	31		7			1		6	6	68
Subir	11	4		7	21		16		6	1	35		11	101
Yunque	12		1	3		2	20	2	19	2	13	2		64
ER		59	48	67	91	24	102	58	55	43	99	10	89	**745**

En la tabla 8.14 se muestra la variación en la cantidad de aciertos o palabras reconocidas en función del número de coeficientes al emplear la técnica de codificación LPC, y en la figura 8.3 se muestra su representación gráfica. El total de palabras a reconocer en cada uno de los experimentos era de 1584 palabras.

TABLA 8.14 Variación en el reconocimiento en función del número de coeficientes LPC

LPC	Aciertos o palabras reconocidas	
M	*A*	%
1	673	42.49
2	684	43.18
3	704	44.44
4	687	43.37
5	706	44.57
6	753	47.54
7	791	49.94
8	829	52.34
9	825	52.08
10	***839***	***52.97***
11	830	52.40
12	821	51.83
13	821	51.83
14	825	52.08
15	833	52.59
16	823	51.96
17	810	51.14
TOTAL	**13,254**	**49.22**

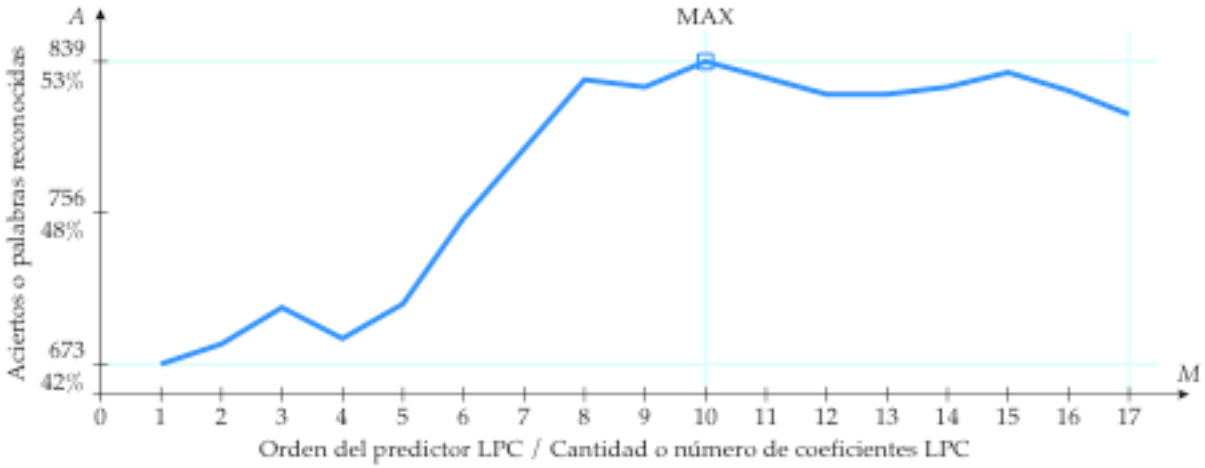


FIGURA 8.3 Variación en el reconocimiento en función del número de coeficientes LPC

Entonces, la cantidad total de aciertos para la prueba con LPC fue de 13,254 palabras reconocidas de un total de 26,928 palabras. En promedio, se reconocieron correctamente unas 780 palabras por experimento, obteniendo así un porcentaje de reconocimiento del 49.22%.

8.6.2 PRUEBA DEL SIMULADOR CON MFCC

En la tabla 8.15 se presentan los resultados obtenidos para la prueba con el simulador de reconocimiento de voz empleando la técnica de codificación MFCC. Estos datos se encuentran ordenados de acuerdo a los doce locutores.

TABLA 8.15 Resultados ordenados por locutor para la prueba del simulador con MFCC

	KVS	KPM	AMA	AMG	GRQ	EBP	EVB	*AYG*	MMC	ABR	FBP	EBC	
M	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***A***
1	35	45	51	58	41	56	53	***57***	46	45	56	51	594
2	71	75	72	74	67	68	84	***92***	78	77	72	66	896
3	67	75	80	82	75	69	86	***91***	79	77	68	76	925
4	71	80	76	81	79	76	93	***97***	81	79	79	82	974
5	73	81	76	82	79	76	93	***99***	79	79	80	81	978
6	73	82	75	82	78	76	91	***100***	83	81	82	83	986
7	70	81	74	83	77	77	93	***99***	84	80	82	82	982
8	***71***	***81***	***73***	***84***	***80***	***76***	***95***	***99***	***83***	***85***	***81***	***79***	***987***
9	72	82	73	84	81	76	92	***99***	84	80	82	81	986
10	69	81	73	84	82	76	91	***100***	84	81	82	81	984
11	69	78	72	85	82	76	90	***99***	82	81	79	81	974
12	69	77	72	85	82	76	89	***101***	83	81	79	84	978
13	70	77	72	85	82	77	90	***101***	82	81	79	84	980
14	70	78	72	86	82	77	89	***101***	82	81	79	84	981
15	70	79	72	86	82	77	89	***101***	82	81	81	84	984
16	70	78	72	86	82	77	89	***101***	82	81	81	84	983
17	70	78	72	86	82	77	89	***101***	82	81	81	84	983
TAL	1160	1308	1227	1393	1313	1263	1496	***1638***	1356	1331	1323	1347	**16,155**

La mayor cantidad de aciertos o palabras reconocidas fue de $A = 987$ de un total de 1584 palabras a reconocer por experimento, este resultado se obtuvo utilizando vectores con $M = 8$ coeficientes MFCC. Por otra parte, el valor más alto para la cantidad total de aciertos por locutor fue de $TAL = 1638$ palabras, para el locutor número 8 (AYG), pero el valor más bajo fue de $TAL = 1160$ palabras, para la locutora número 1 (KVS), en ambos casos estos locutores se usaron como patrones de referencia para tratar de reconocer un total de 2244 palabras por locutor de referencia.

En la tabla 8.16 se muestran los mismos datos de la tabla anterior, pero ahora se encuentran ordenados de otra manera diferente, de acuerdo a las doce palabras o clases del vocabulario.

TABLA 8.16 Resultados ordenados por palabra para la prueba del simulador con MFCC

	Chispa	Digno	Flecha	Fugaz	Líder	Lluvia	Mejor	Monte	*Punto*	Soñar	Subir	Yunque	
M	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***A***
1	78	58	40	86	49	27	32	30	***92***	38	37	27	594
2	86	103	62	104	83	40	77	62	***96***	54	74	55	896
3	89	100	65	109	73	46	80	66	***115***	60	67	55	925
4	96	107	70	109	79	47	85	68	***115***	70	71	57	974
5	98	106	70	109	76	48	88	66	***114***	69	71	63	978
6	99	104	71	110	79	50	86	67	***116***	68	73	63	986
7	100	105	70	109	77	49	85	67	***114***	69	72	65	982
8	***101***	***104***	***71***	***110***	***79***	***49***	***82***	***68***	***114***	***71***	***75***	***63***	***987***
9	100	104	71	111	80	50	81	67	***114***	70	72	66	986
10	97	106	72	111	80	50	81	67	***113***	71	71	65	984
11	96	105	72	111	79	49	82	65	***113***	69	70	63	974
12	95	105	72	111	77	49	83	65	***113***	69	70	69	978
13	97	105	72	111	77	49	84	65	***113***	69	70	68	980
14	97	105	71	111	78	50	84	65	***113***	69	70	68	981
15	97	105	71	111	78	50	84	67	***113***	69	71	68	984
16	97	105	71	111	77	50	84	67	***113***	69	71	68	983
17	97	105	71	111	77	50	84	67	***113***	69	71	68	983
TAC	1620	1732	1162	1845	1298	803	1362	1089	***1894***	1123	1176	1051	**16,155**

La palabra más fácil de reconocer con $TAC = 1894$ aciertos fue la número 9 *"punto"*, pero la más difícil de reconocer con $TAC = 803$ aciertos fue la número 6 *"lluvia"*, en ambos casos el total de palabras a reconocer era de 2244 palabras por clase.

En las siguientes dos tablas, se muestran los resultados obtenidos para el mejor experimento con el simulador de reconocimiento de voz, con el que se consiguió el mayor número de aciertos, utilizando el número óptimo de $M = 8$ coeficientes MFCC. La tabla 8.17 muestra los resultados o aciertos y la tabla 8.18 la matriz de confusión.

TABLA 8.17 Resultados del experimento para el número óptimo de coeficientes MFCC

Palabra		Chispa	Digno	Flecha	Fugaz	Líder	Lluvia	Mejor	Monte	*Punto*	Soñar	Subir	Yunque	
Entrada		***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***SAR***
KVS	***1***	11	10	3	10	2	3	6	3	***8***	8	5	2	**71**
KPM	***2***	7	11	1	11	3	0	7	11	***8***	10	8	4	**81**
AMA	***3***	10	8	5	10	6	9	6	1	***9***	2	6	1	**73**
AMG	***4***	10	5	8	11	6	5	8	7	***11***	3	7	3	**84**
GRQ	***5***	4	10	2	7	7	8	9	6	***11***	3	4	9	**80**
EBP	***6***	11	11	4	8	9	4	2	1	***4***	5	7	10	**76**
EVB	***7***	10	9	7	11	9	4	8	5	***11***	6	9	6	**95**
AYG	***8***	***7***	***8***	***10***	***11***	***9***	***6***	***7***	***9***	***11***	***7***	***8***	***6***	***99***
MMC	***9***	6	8	10	11	8	1	5	8	***11***	6	3	6	**83**
ABR	***10***	10	9	4	7	5	3	8	3	***9***	9	10	8	**85**
FBP	***11***	8	10	7	11	9	3	8	4	***10***	5	4	2	**81**
EBC	***12***	7	5	10	2	6	3	8	10	***11***	7	4	6	**79**
SAC		**101**	**104**	**71**	**110**	**79**	**49**	**82**	**68**	***114***	**71**	**75**	**63**	**987**

La mayor cantidad de aciertos fue de $SAR = 99$ palabras, para el locutor número 8 (AYG), pero la menor fue de $SAR = 71$ palabras, para la locutora número 1 (KVS). Por otra parte, la palabra más fácil de reconocer con $SAC = 114$ aciertos fue la número 9 *"punto"*, pero la más difícil de reconocer con $SAC = 49$ aciertos fue la número 6 *"lluvia"*.

TABLA 8.18 Matriz de confusión para el número óptimo de coeficientes MFCC

Palabra		Chispa	Digno	Flecha	Fugaz	Líder	Lluvia	Mejor	Monte	Punto	Soñar	Subir	Yunque	
		1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***EC***
Chispa	***1***		12	23		5	1	3			8	2		54
Digno	***2***	7		4		8	9	13	1	8	2	1	2	55
Flecha	***3***	15	4			11	12	1	1		6	4	3	57
Fugaz	***4***			5		5	5	19	2		26	10		72
Líder	***5***		3	3			1							7
Lluvia	***6***		1	10	1	10		5	10	5	9	6	19	76
Mejor	***7***	2	4	5	2	2	1		7	4			1	28
Monte	***8***		1			2	9	2				1	28	43
Punto	***9***		3	1		2	11	5	9				1	32
Soñar	***10***	7		7	19	2	7	2	2			22	10	78
Subir	***11***						3		1		6		5	15
Yunque	***12***			3		6	24		31	1	4	11		80
ER		31	28	61	22	53	83	50	64	18	61	57	69	**597**

En la tabla 8.19 se muestra la variación en la cantidad de aciertos o palabras reconocidas en función del número de coeficientes al emplear la técnica de codificación MFCC, y en la figura 8.4 se muestra su representación gráfica. El total de palabras a reconocer en cada uno de los experimentos era de 1584.

TABLA 8.19 Variación en el reconocimiento en función del número de coeficientes MFCC

MFCC	Aciertos o palabras reconocidas	
M	**Total**	*%*
1	594	37.50
2	896	56.57
3	925	58.40
4	974	61.49
5	978	61.74
6	986	62.25
7	982	61.99
8	***987***	***62.31***
9	986	62.25
10	984	62.12
11	974	61.49
12	978	61.74
13	980	61.87
14	981	61.93
15	984	62.12
16	983	62.06
17	983	62.06
TOTAL	**16,155**	59.99

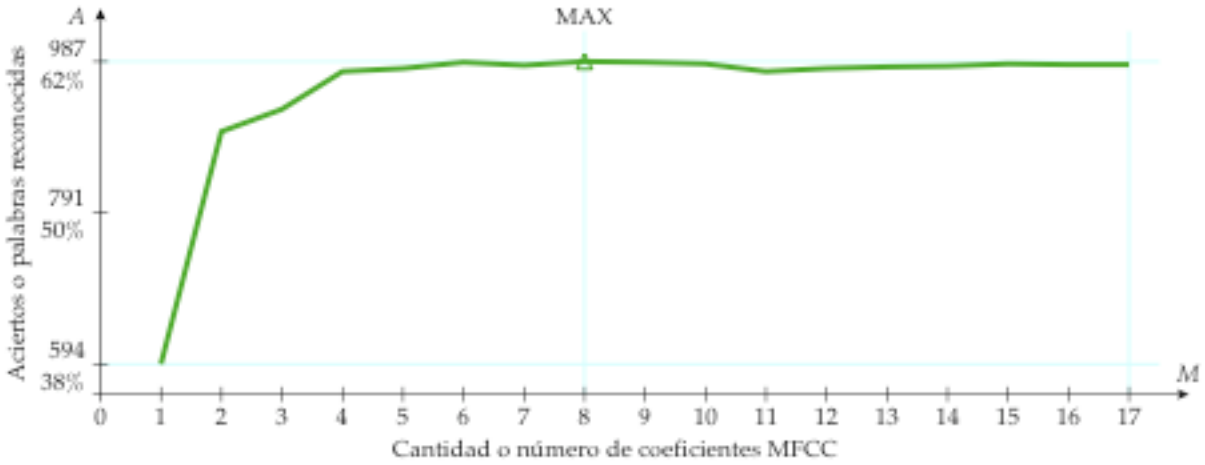


FIGURA 8.4 Variación en el reconocimiento en función del número de coeficientes MFCC

Entonces, la cantidad total de aciertos para la prueba con MFCC fue de 16,155 palabras reconocidas de un total de 26,928 palabras. En promedio, se reconocieron correctamente unas 950 palabras por experimento, obteniendo así un porcentaje de reconocimiento del 59.99%.

8.6.3 PRUEBA DEL SIMULADOR CON PLP

En la tabla 8.20 se muestran los resultados obtenidos para la prueba con el simulador de reconocimiento de voz empleando la técnica de codificación PLP. Estos datos se encuentran ordenados de acuerdo a los locutores.

TABLA 8.20 Resultados ordenados por locutor para la prueba del simulador con PLP

	KVS	KPM	AMA	AMG	GRQ	EBP	EVB	*AYG*	MMC	ABR	FBP	EBC	
M	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***A***
1	42	43	52	59	47	53	51	***61***	48	45	59	46	606
2	63	73	73	73	61	70	71	***89***	76	75	77	70	871
3	65	73	66	69	57	64	69	***86***	76	64	72	60	821
4	67	70	77	75	69	76	79	***93***	80	81	76	71	914
5	***66***	***72***	***72***	***79***	***71***	***77***	***78***	***96***	***78***	***81***	***78***	***73***	***921***
6	61	70	74	79	70	75	80	***92***	81	79	75	72	908
7	64	70	77	81	71	75	80	***97***	78	78	73	74	918
8	59	70	76	79	70	75	82	***95***	74	82	75	78	915
9	59	68	76	81	66	74	82	***94***	72	78	76	75	901
10	57	65	75	82	66	73	81	***90***	73	77	75	74	888
11	55	65	74	82	63	71	80	***87***	71	74	73	73	868
12	56	65	75	82	63	69	80	***87***	69	74	73	73	866
13	53	62	74	76	63	67	77	***87***	69	72	73	75	848
14	52	61	73	78	61	66	76	***86***	69	70	73	76	841
15	49	59	72	80	61	59	77	***86***	69	70	72	76	830
16	46	58	73	76	59	58	78	***85***	69	70	72	75	819
17	47	58	73	78	59	57	79	***85***	69	70	72	75	822
TAR	961	1102	1232	1309	1077	1159	1300	***1496***	1221	1240	1244	1216	**14,557**

La mayor cantidad de aciertos o palabras reconocidas fue de $A = 921$ palabras de un total de 1584 palabras a reconocer por experimento, este resultado se obtuvo utilizando vectores con $M = 5$ coeficientes PLP. Por otra parte, el valor más alto para el total de aciertos por locutor de referencia fue de $TAL = 1496$ palabras, para el locutor número 8 (AYG), pero el valor más bajo fue de $TAL = 961$ palabras, para la locutora número 1 (KVS), en ambos casos estos locutores se usaron como patrones de referencia para tratar de reconocer un total de 2244 palabras por locutor de referencia.

En la Tabla 8.21 se muestran los mismos datos de la tabla anterior, pero ahora se encuentran ordenados de otra manera diferente, de acuerdo a las doce palabras o clases del vocabulario.

TABLA 8.21 Resultados ordenados por palabra o clase para la prueba del simulador con PLP

	Chispa	Digno	Flecha	*Fugaz*	Líder	Lluvia	Mejor	Monte	Punto	Soñar	Subir	Yunque	
M	***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***A***
1	76	55	52	***75***	46	32	33	33	102	48	31	23	606
2	82	105	61	***106***	66	40	79	72	116	52	49	43	871
3	71	88	55	***105***	60	39	74	65	110	58	49	47	821
4	85	104	63	***109***	64	46	85	69	110	69	57	53	914
5	***87***	***107***	***64***	***110***	***69***	***45***	***88***	***65***	***111***	***72***	***54***	***49***	***921***
6	86	102	60	***110***	66	45	89	65	113	69	53	50	908
7	85	103	64	***110***	72	47	87	63	110	74	55	48	918
8	86	106	63	***110***	76	48	85	63	104	74	55	45	915
9	87	102	62	***110***	74	47	85	62	101	69	54	48	901
10	87	101	58	***110***	73	48	83	62	98	67	52	49	888
11	84	100	59	***110***	72	48	81	56	95	65	51	47	868
12	84	98	59	***110***	70	51	80	57	93	66	52	46	866
13	80	94	59	***108***	69	51	77	55	93	64	52	46	848
14	79	93	59	***108***	68	50	78	54	93	63	52	44	841
15	79	91	60	***107***	66	50	76	53	94	63	48	43	830
16	77	88	57	***106***	65	49	75	52	93	65	46	46	819
17	77	87	59	***106***	66	46	76	53	93	67	44	48	822
TAC	1392	1624	1014	***1810***	1142	782	1331	999	1729	1105	854	775	**14,557**

La palabra más fácil de reconocer con $TAC = 1810$ aciertos fue la número 4 *"fugaz"*, pero la más difícil de reconocer con $TAC = 775$ aciertos fue la número 12 "yunque", en ambos casos el total de palabras a reconocer era de 2244 palabras por clase.

En las siguientes dos tablas, se muestran los resultados obtenidos para el mejor experimento con el simulador de reconocimiento de voz, con el que se consiguió el mayor número de aciertos, utilizando el número óptimo de $M = 5$ coeficientes PLP. La tabla 8.22 muestra los resultados o aciertos y la tabla 8.23 la matriz de confusión.

TABLA 8.22 Resultados del experimento para el número óptimo de coeficientes PLP

Palabra		Chispa	Digno	Flecha	Fugaz	Líder	Lluvia	Mejor	Monte	*Punto*	Soñar	Subir	Yunque	
Entrada		***1***	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***SAR***
KVS	***1***	6	10	1	9	0	2	7	6	***8***	10	6	1	66
KPM	***2***	7	10	2	10	1	1	9	8	***8***	8	6	2	72
AMA	***3***	10	9	3	10	6	9	6	1	***7***	4	4	3	72
AMG	***4***	8	6	10	11	5	4	9	4	***11***	3	5	3	79
GRQ	***5***	5	9	1	8	4	6	10	6	***11***	1	2	8	71
EBP	***6***	11	11	4	8	8	4	4	1	***6***	5	5	10	77
EVB	***7***	8	9	3	11	9	3	8	5	***11***	4	5	2	78
AYG	***8***	***6***	***9***	***10***	***11***	***8***	***5***	***7***	***9***	***11***	***8***	***6***	***6***	***96***
MMC	***9***	6	9	9	11	8	1	3	9	***10***	5	3	4	78
ABR	***10***	9	9	6	7	6	4	8	3	***8***	10	6	5	81
FBP	***11***	4	10	9	11	9	3	8	2	***9***	7	4	2	78
EBC	***12***	7	6	6	3	5	3	9	11	***11***	7	2	3	73
SAC		87	107	64	110	69	45	88	65	***111***	72	54	49	**921**

La mayor cantidad de aciertos fue de $SAR = 96$ palabras, para el locutor número 8 (AYG), pero la menor cantidad de aciertos fue de $SAR = 66$ palabras, para la locutora número 1 (KVS). Por otra parte, la palabra más fácil de reconocer con $SAC = 111$ aciertos fue la número 9 *"punto"*, pero la más difícil de reconocer con $SAC = 45$ aciertos fue la número 6 *"lluvia"*.

TABLA 8.23 Matriz de confusión para el número óptimo de coeficientes PLP

Palabra		Chispa	Digno	Flecha	Fugaz	Líder	Lluvia	Mejor	Monte	Punto	Soñar	Subir	*Yunque*	
		1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***11***	***12***	***EC***
Chispa	***1***		4	23				4			13		***1***	45
Digno	***2***	4		3		2	3	7		5			***1***	25
Flecha	***3***	30	3		5	5	8	4			8		***5***	68
Fugaz	***4***						2	3			17		***0***	22
Líder	***5***	13	10	11	1		11	2	5	2			***8***	63
Lluvia	***6***	***0***	***4***	***10***	***5***	***2***		***3***	***18***	***11***	***7***	***3***	***24***	***87***
Mejor	***7***	3	14	3	16		2		3	3			***0***	44
Monte	***8***		1				13	7		11	1	1	***33***	67
Punto	***9***		8			1	6	4	1				***1***	21
Soñar	***10***	17	1	8	20		2	1	1			3	***7***	60
Subir	***11***	2		6	11		7		5		30		***17***	78
Yunque	***12***		3	2		7	13	4	41	1	9	3		83
ER		69	48	66	58	17	67	39	74	33	85	10	***97***	**663**

En la tabla 8.24 se muestra la variación en la cantidad de aciertos o palabras reconocidas en función del número de coeficientes al emplear la técnica de codificación PLP, y en la figura 8.5 se muestra su representación gráfica. El total de palabras a reconocer en cada uno de los experimentos era de 1584.

TABLA 8.24 Variación en el reconocimiento en función del número de coeficientes PLP

PLP	Aciertos o palabras reconocidas	
M	Total	%
1	606	38.26
2	871	54.99
3	821	51.83
4	914	57.70
5	***921***	***58.14***
6	908	57.32
7	918	57.95
8	915	57.77
9	901	56.88
10	888	56.06
11	868	54.80
12	866	54.67
13	848	53.54
14	841	53.09
15	830	52.40
16	819	51.70
17	822	51.89
TOTAL	**14,557**	**54.06**

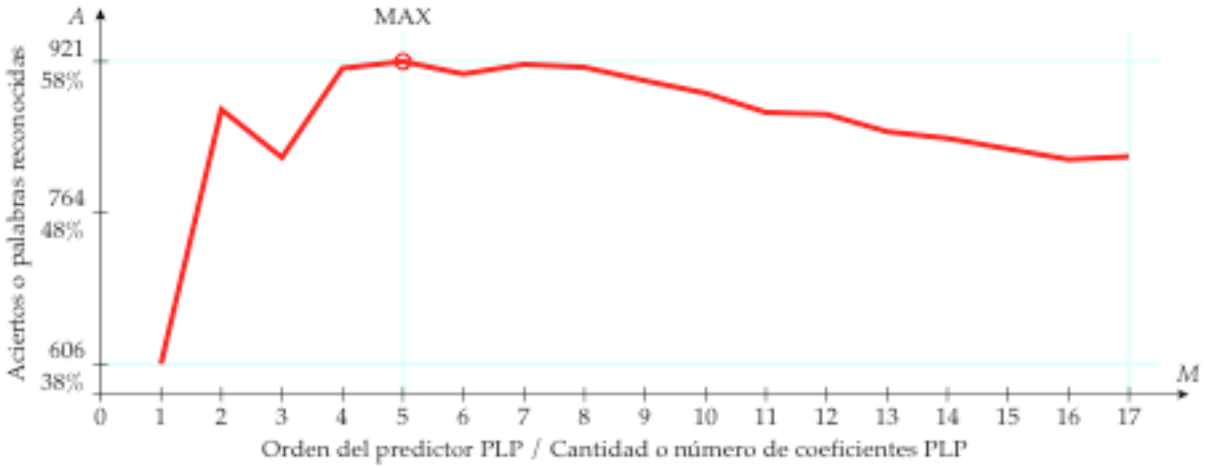


FIGURA 8.5 Variación en el reconocimiento en función del número de coeficientes PLP

Entonces, la cantidad total de aciertos para la prueba con PLP fue de 14,557 palabras reconocidas de un total de 26,928 palabras. En promedio, se reconocieron correctamente unas 856 palabras por experimento, obteniendo así un porcentaje de reconocimiento del 54.06%.

8.7 COMPARACIÓN DE LOS RESULTADOS

Para poder determinar si las codificaciones perceptuales proporcionaron una mayor cantidad de palabras reconocidas, se tienen que comparar los resultados obtenidos en las tres diferentes pruebas. Además, a fin de contar con más elementos para sacar conclusiones sobre los resultados, se toman en cuenta tres aspectos al realizar las comparaciones, a saber:

1. La variación en el reconocimiento en función del número de coeficientes.
2. El análisis estadístico del total de aciertos por locutor para las pruebas.
3. El análisis estadístico del total de aciertos por clase para las pruebas.

Los resultados se pueden ordenar ya sea por lo locutor o por clase. En ambos casos se trata del mismo número de aciertos. Mediante un análisis estadístico, se puede calcular la media y la desviación estándar de los datos de las tres pruebas en conjunto, de modo que se pueda comprobar si se obtuvo una mejora en el reconocimiento de palabras al cambiar de técnica de codificación. De tal manera que si el valor absoluto de la diferencia entre los resultados de dos codificaciones es menor o igual que la mitad de una desviación estándar, se considera que prácticamente permaneció igual, pero si la diferencia es mayor que la mitad de una desviación estándar, entonces si es positiva se considera que hubo una mejora en el reconocimiento, pero si es negativa entonces se considera que hubo un deterioro en el reconocimiento. En la tabla 8.25 se muestra un resumen de estos umbrales propuestos para tomar una decisión sobre la variación en el reconocimiento de palabras.

TABLA 8.25 Umbrales de decisión para la comparación de los resultados de las pruebas

Rango	Reconocimiento	Símbolo
$\alpha > +2.5$	Mejoro mucho	> > >
$+1.5 < \alpha \leq +2.5$	Mejoro	> >
$+0.5 < \alpha \leq +1.5$	Mejoro un poco	>
$-0.5 \leq \alpha \leq +0.5$	Sin cambios	=
$-1.5 \leq \alpha < -0.5$	Empeoro un poco	<
$-2.5 \leq \alpha < -1.5$	Empeoro	< <
$\alpha < -2.5$	Empeoro mucho	< < <

8.7.1 LA VARIACIÓN EN EL RECONOCIMIENTO

La figura 8.6 muestra la representación gráfica de la variación en el reconocimiento (la cantidad de aciertos o palabras reconocidas) en función del número de coeficientes, así como el número óptimo de coeficientes para cada una de las tres técnicas de codificación.

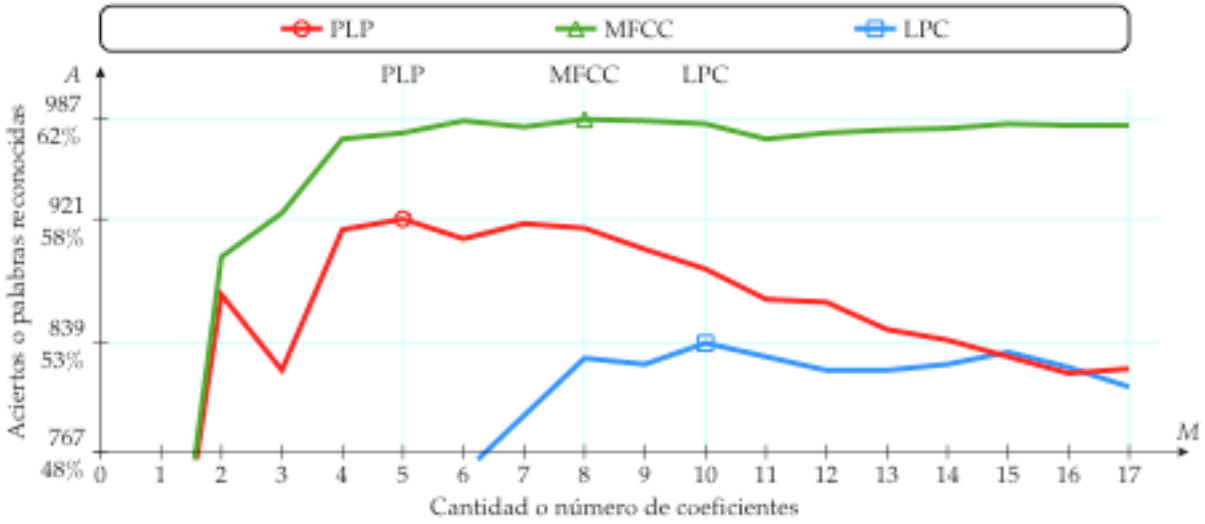


FIGURA 8.6 Detalle de la comparación de la variación en el reconocimiento

La tabla 8.26 resume los resultados de las tres pruebas. Claramente, MFCC proporciona la mayor cantidad de palabras reconocidas, en segundo lugar PLP y en tercer lugar LPC.

TABLA 8.26 Resumen de los resultados de las tres pruebas con el simulador

Posición	Codificación	*M*	*BC*	*A*	%
1	MFCC	1 : 1 :17	17	16,155 de 26928	59.99
2	PLP	1 : 1 :17	15+2	14,557 de 26928	54.06
3	LPC	1 : 1 :17	N/A	13,254 de 26928	49.22

La tabla 8.27 resume los resultados de los experimentos con el número óptimo de coeficientes para cada una de las tres codificaciones. Aunque PLP proporciona el número óptimo de coeficientes más bajo, no proporciona la mayor cantidad de palabras reconocidas.

TABLA 8.27 Resultados de los experimentos con el número óptimo de coeficientes

Posición	Codificación	*M óptimo*	*BC*	*A*	%
1	MFCC	8	17	987 de 1584	62.31
2	PLP	5	15+2	921 de 1584	58.14
3	LPC	10	N/A	839 de 1584	52.97

8.7.2 ACIERTOS POR LOCUTOR

En la tabla 8.28 se presentan los resultados del total de aciertos por locutor de referencia (*TAR*) para las tres pruebas y en la figura 8.7 se muestra su representación gráfica. Los datos se encuentran ordenados en orden ascendente de acuerdo a su valor promedio de aciertos.

TABLA 8.28 Resumen del total de aciertos por locutor

Locutor		LPC	MFCC	PLP	Promedio
KVS	***1***	810	1160	961	977.00
GRQ	***5***	1075	1313	1077	1155.00
AMA	***3***	1011	1227	1232	1156.67
KPM	***2***	1077	1308	1102	1162.33
EBP	***6***	1110	1263	1159	1177.33
EBC	***12***	1114	1347	1216	1225.67
ABR	***10***	1136	1331	1240	1235.67
FBP	***11***	1184	1323	1244	1250.33
MMC	***9***	1210	1356	1221	1262.33
AMG	***4***	1186	1393	1309	1296.00
EVB	***7***	1101	1496	1300	1299.00
AYG	***8***	1240	1638	1496	1458.00
		σ	156.55	μ	1217.18

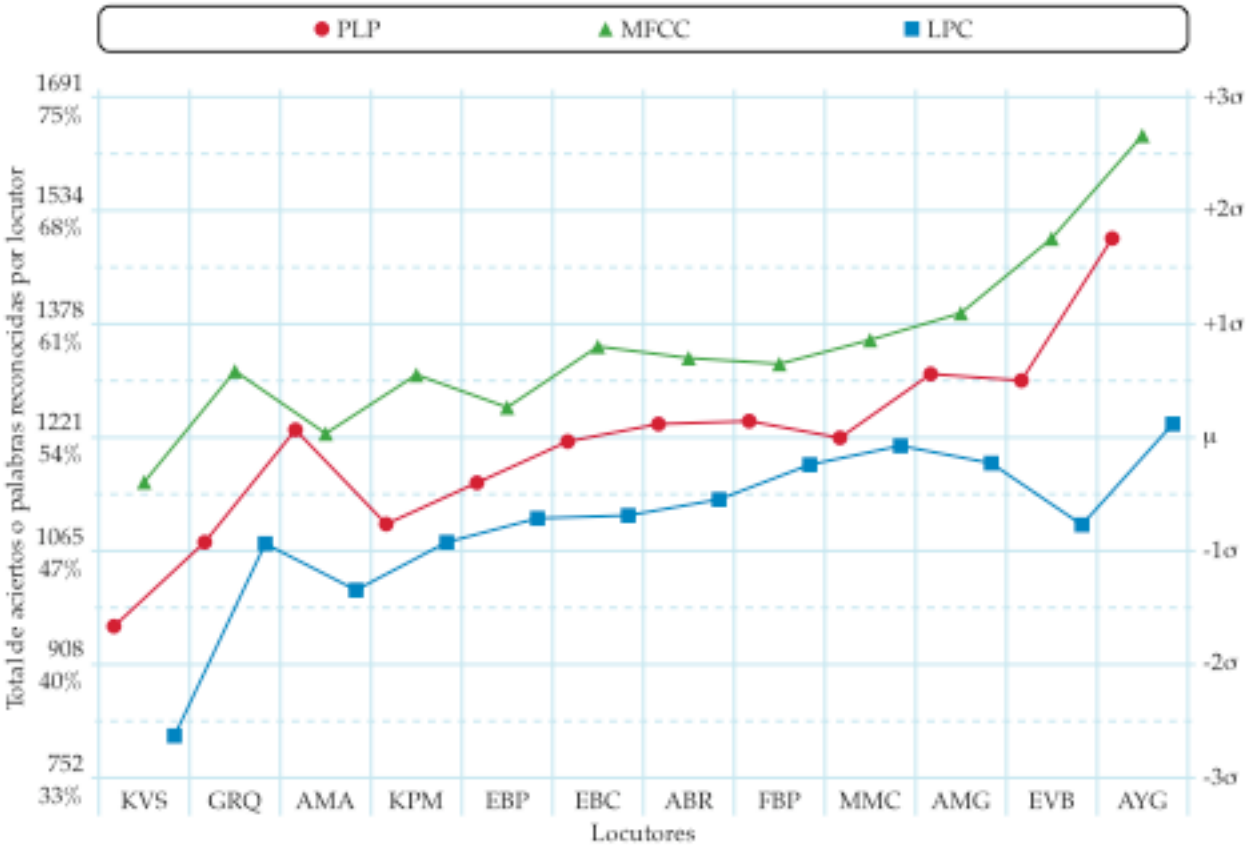


FIGURA 8.7 Distribución del total de aciertos por locutor de referencia

En las siguientes tres tablas se muestran las comparaciones del total de aciertos por locutor de referencia (*TAR*) para verificar si se obtuvo una mejoría en el reconocimiento de palabras al cambiar de técnica de codificación, en la tabla 8.29 de LPC a MFCC, en la tabla 8.30 de MFCC a PLP y en la tabla 8.31 de LPC a PLP.

TABLA 8.29 Comparación del total de aciertos por locutor de MFCC contra los de LPC

Locutor		MFCC	LPC	MFCC - LPC	α σ	Peor	Sin cambios	Mejor
KVS	***1***	1160	810	350	2.24			> >
GRQ	***5***	1313	1075	238	1.52			> >
AMA	***3***	1227	1011	216	1.38			>
KPM	***2***	1308	1077	231	1.48			>
EBP	***6***	1263	1110	153	0.98			>
EBC	***12***	1347	1114	233	1.49			>
ABR	***10***	1331	1136	195	1.25			>
FBP	***11***	1323	1184	139	0.89			>
MMC	***9***	1356	1210	146	0.93			>
AMG	***4***	1393	1186	207	1.32			>
EVB	***7***	1496	1101	395	2.52			> > >
AYG	***8***	1638	1240	398	2.54			> > >

TABLA 8.30 Comparación del total de aciertos por locutor de PLP contra los de MFCC

Locutor		PLP	MFCC	PLP - MFCC	α σ	Peor	Sin cambios	Mejor
KVS	***1***	961	1160	-199	-1.27	<		
GRQ	***5***	1077	1313	-236	-1.51	< <		
AMA	***3***	1232	1227	5	0.03		=	
KPM	***2***	1102	1308	-206	-1.32	<		
EBP	***6***	1159	1263	-104	-0.66	<		
EBC	***12***	1216	1347	-131	-0.84	<		
ABR	***10***	1240	1331	-91	-0.58	<		
FBP	***11***	1244	1323	-79	-0.50	<		
MMC	***9***	1221	1356	-135	-0.86	<		
AMG	***4***	1309	1393	-84	-0.54	<		
EVB	***7***	1300	1496	-196	-1.25	<		
AYG	***8***	1496	1638	-142	-0.91	<		

TABLA 8.31 Comparación del total de aciertos por locutor de PLP contra los de LPC

Locutor		PLP	LPC	PLP - LPC	α σ	Peor	Sin cambios	Mejor
KVS	***1***	961	810	151	0.96			>
GRQ	***5***	1077	1075	2	0.01		=	
AMA	***3***	1232	1011	221	1.41			>
KPM	***2***	1102	1077	25	0.16		=	
EBP	***6***	1159	1110	49	0.31		=	
EBC	***12***	1216	1114	102	0.65			>
ABR	***10***	1240	1136	104	0.66			>
FBP	***11***	1244	1184	60	0.38		=	
MMC	***9***	1221	1210	11	0.07		=	
AMG	***4***	1309	1186	123	0.79			>
EVB	***7***	1300	1101	199	1.27			>
AYG	***8***	1496	1240	256	1.64			> >

8.7.3 ACIERTOS POR PALABRA O CLASE

En la tabla 8.32 se presentan los resultados del total de aciertos por palabra o clase (*TAC*) para las tres pruebas y en la figura 8.8 se muestra su representación gráfica. Los datos se encuentran ordenados en orden ascendente de acuerdo a su valor promedio de aciertos.

TABLA 8.32 Resumen del total de aciertos por palabra o clase

Palabra		LPC	MFCC	PLP	Promedio
Lluvia	***6***	757	803	782	781
Subir	***11***	510	1176	854	847
Yunque	***12***	991	1051	775	939
Monte	***8***	868	1089	999	985
Flecha	***3***	950	1162	1014	1042
Soñar	***10***	1007	1123	1105	1078
Líder	***5***	909	1298	1142	1116
Mejor	***7***	1303	1362	1331	1332
Chispa	***1***	1306	1620	1392	1439
Digno	***2***	1376	1732	1624	1577
Punto	***9***	1633	1894	1729	1752
Fugaz	***4***	1644	1845	1810	1766
		σ	357.49	μ	1217.18

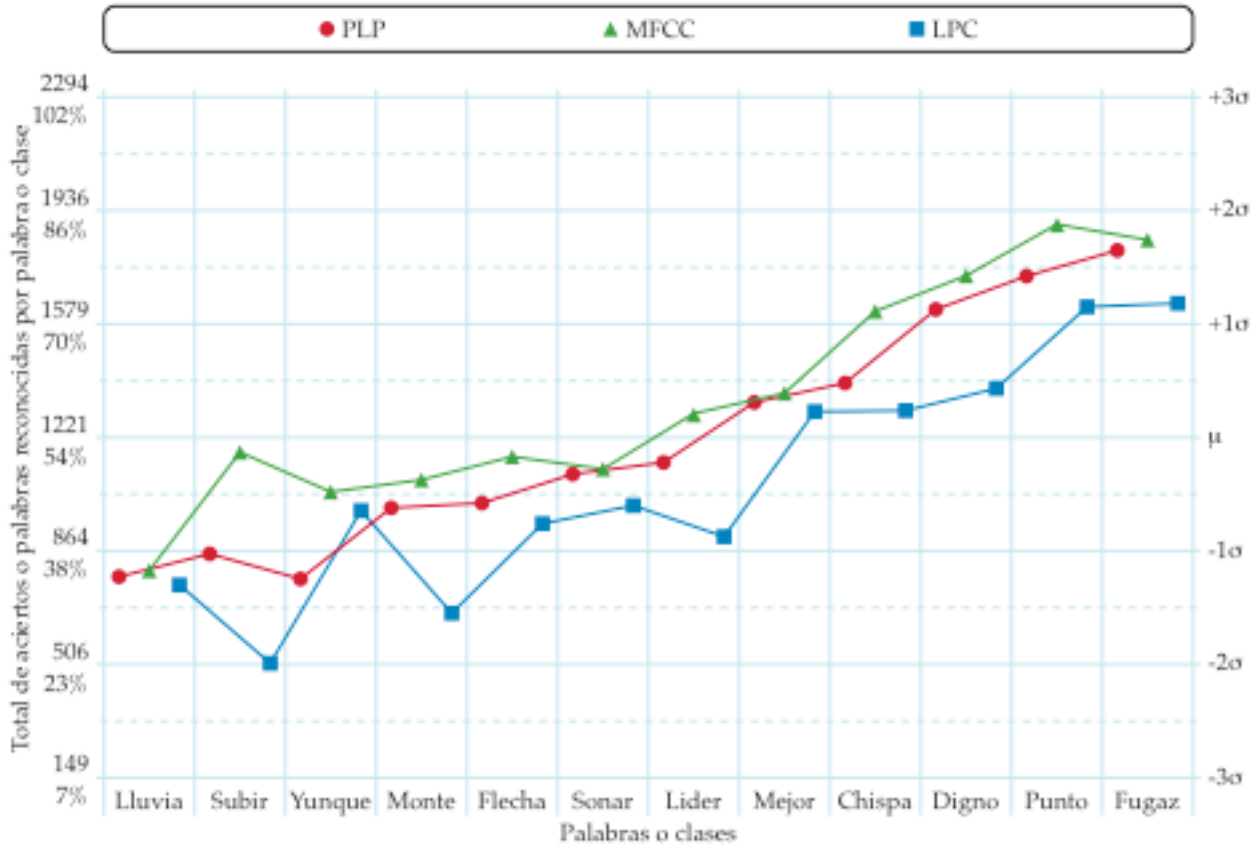


FIGURA 8.8 Distribución del total de aciertos por palabra o clase

En las siguientes tres tablas se muestran las comparaciones del total de aciertos por por palabra o clase (*TAC*) para verificar si se obtuvo una mejoría en el reconocimiento de palabras al cambiar de técnica de codificación, en la tabla 8.33 de LPC a MFCC, en la tabla 8.34 de MFCC a PLP y en la tabla 8.35 de LPC a PLP.

TABLA 8.33 Comparación del total de aciertos por palabra de MFCC contra los de LPC

Palabra		MFCC	LPC	MFCC - LPC	α σ	Peor	Sin cambios	Mejor
Lluvia	*6*	803	757	46	0.29		=	
Subir	*11*	1176	510	666	4.25			> > >
Yunque	*12*	1051	991	60	0.38		=	
Monte	*8*	1089	868	221	1.41			>
Flecha	*3*	1162	950	212	1.35			>
Soñar	*10*	1123	1007	116	0.74			>
Líder	*5*	1298	909	389	2.48			> >
Mejor	*7*	1362	1303	59	0.38		=	
Chispa	*1*	1620	1306	314	2.01			> >
Digno	*2*	1732	1376	356	2.27			> >
Punto	*9*	1894	1633	261	1.67			> >
Fugaz	*4*	1845	1644	201	1.28			>

TABLA 8.34 Comparación del total de aciertos por palabra de PLP contra los de MFCC

Palabra		PLP	MFCC	PLP - MFCC	α σ	Peor	Sin cambios	Mejor
Lluvia	*6*	782	803	-21	-0.13		=	
Subir	*11*	854	1176	-322	-2.06	< <		
Yunque	*12*	775	1051	-276	-1.76	< <		
Monte	*8*	999	1089	-90	-0.57	<		
Flecha	*3*	1014	1162	-148	-0.95	<		
Soñar	*10*	1105	1123	-18	-0.11		=	
Líder	*5*	1142	1298	-156	-1.00	<		
Mejor	*7*	1331	1362	-31	-0.20		=	
Chispa	*1*	1392	1620	-228	-1.46	<		
Digno	*2*	1624	1732	-108	-0.69	<		
Punto	*9*	1729	1894	-165	-1.05	<		
Fugaz	*4*	1810	1845	-35	-0.22		=	

TABLA 8.35 Comparación del total de aciertos por palabra de PLP contra los de LPC

Palabra		PLP	LPC	PLP - LPC	α σ	Peor	Sin cambios	Mejor
Lluvia	*6*	782	757	25	0.16		=	
Subir	*11*	854	510	344	2.20			> >
Yunque	*12*	775	991	-216	-1.38	<		
Monte	*8*	999	868	131	0.84			>
Flecha	*3*	1014	950	64	0.41		=	
Soñar	*10*	1105	1007	98	0.63			>
Líder	*5*	1142	909	233	1.49			>
Mejor	*7*	1331	1303	28	0.18		=	
Chispa	*1*	1392	1306	86	0.55			>
Digno	*2*	1624	1376	248	1.58			> >
Punto	*9*	1729	1633	96	0.61			>
Fugaz	*4*	1810	1644	166	1.06			>

8.8 CONCLUSIONES

Puesto que los resultados se obtuvieron a partir de una muestra muy pequeña, no se puede asegurar que las características observadas en estas pruebas siempre se cumplan en cualquier otro caso. Pero aún así, se pueden sacar varias conclusiones útiles e interesantes. Por ejemplo, ciertos locutores funcionaron mejor que otros como patrones de referencia, por lo que podría ser ventajoso evaluarlos primero antes de emplearlos en el entrenamiento, con el propósito de evitar la incorporación de patrones de referencia deficientes. Sin embargo, la manera óptima para garantizar el buen funcionamiento de un sistema de reconocimiento es mediante la realización de una sesión de entrenamiento con los principales usuarios del mismo. Por otra parte, puesto que en las comparaciones por clases los cambios no fueron considerables, parece ser que la dificultad para reconocer una palabra no depende tanto de la técnica de codificación, sino más bien de las palabras que componen el vocabulario.

Como era de esperarse, la restricción del mínimo entrenamiento ocasionó que los resultados de las pruebas tendieran a ser muy similares, por lo que las diferencias no fueron tan significativas. No obstante, sí se pudo confirmar el número óptimo esperado de coeficientes para las tres técnicas de codificación y de acuerdo a los resultados obtenidos para estos experimentos, las dos codificaciones perceptuales, MFCC y PLP, sí fueron mejores que LPC, pues estas técnicas proporcionan una mayor independencia del locutor. Por lo tanto, el desempeño del reconocimiento con PLP no fue mejor que con MFCC, pues el desempeño con MFCC fue mejor que con PLP y LPC. Ahora bien, una forma en la que se podrían mejorar las pruebas sería mediante la variación de la cantidad de vectores de entrenamiento, a partir de un vector en adelante, de este modo también se podría evaluar su contribución para mejorar el desempeño del reconocimiento, así como para comprobar que tanto pueden mejorar las codificaciones e incluso determinar un número óptimo de vectores de entrenamiento.

En cuanto a la compresión de la señal de voz, PLP proporciona el número óptimo de coeficientes más bajo para las tres codificaciones. Sin embargo, probablemente no se le haya prestado mucha atención debido a que tiene la gran desventaja de que a partir de cierta cantidad de coeficientes, incluso la codificación LPC, que es mucho más simple, puede resultar igual o hasta mejor que PLP. A diferencia de esta, MFCC se ha convertido en la técnica básica empleada en el campo del reconocimiento de voz, seguramente debido a su excelente desempeño y a que su fundamento teórico está basado en el análisis cepstral.

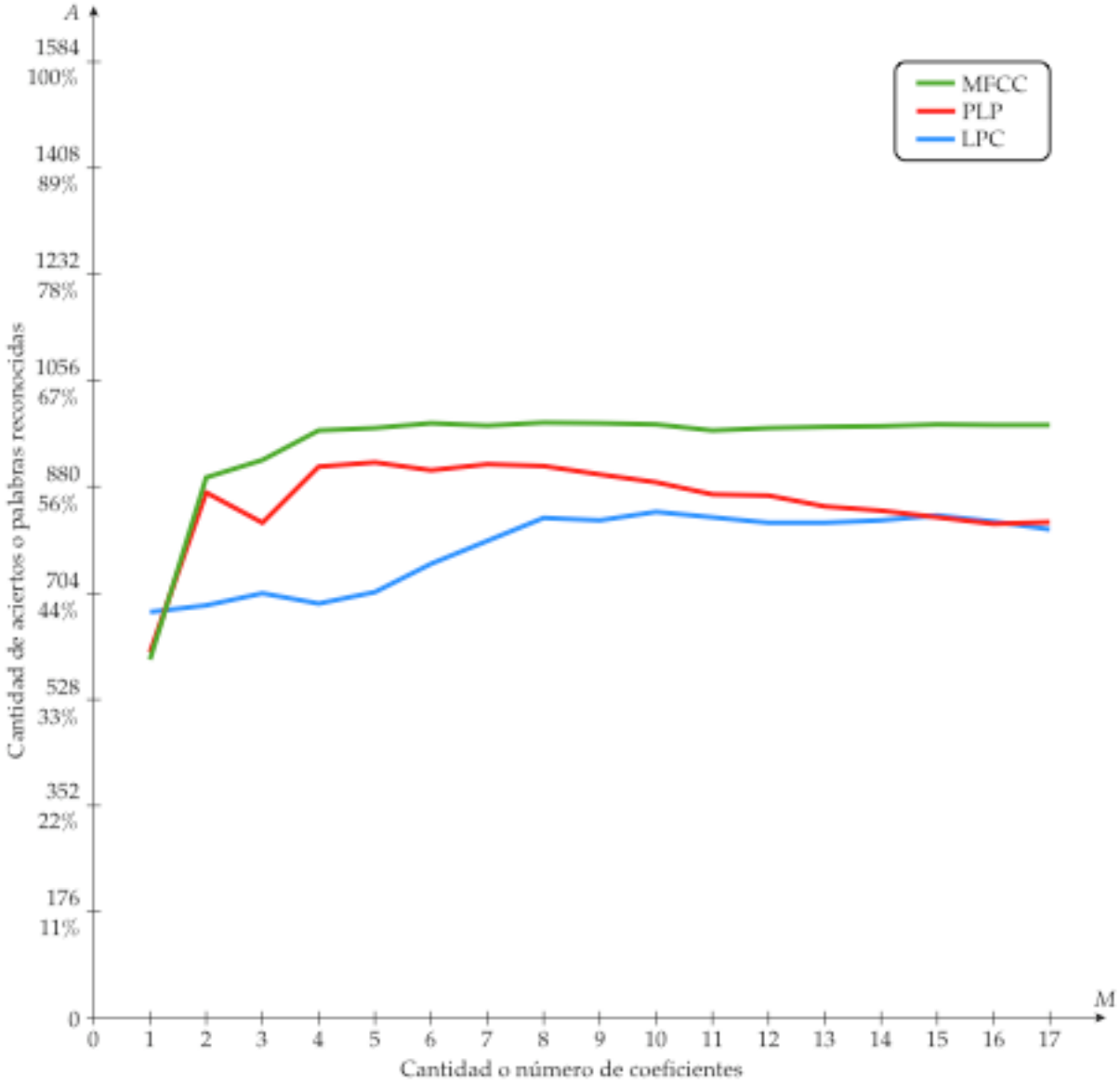


FIGURA 8.9 Comparación de la variación en el reconocimiento entre LPC, MFCC y PLP

APÉNDICE I

PROGRAMA DE MATLAB PARA OBTENER UN ESPECTROGRAMA EN 2D

```
function SG2D=Espectrograma2D(x,fs,N,S,Nfft)

%Espectrograma2D(x,fs,N,S,Nfft)
%   Programado por Alan Edgar Ballado de la Rivera UNAM FI 2012
%   Obtiene el espectrograma bidimensional de la señal de voz x, en donde:
%   fs   : es la frecuencia de muestreo.
%   N    : es la longitud o el numero de muestras de la trama.
%   S    : es el numero de muestras de separacion entre tramas adyacentes.
%   Nfft : es el numero de puntos para la transformada rapida de Fourier.

%INICIALIZACION
Nx=length(x);
L=round((Nx-N)/S+1);
Np=0.5*Nfft+1;
SG2D=zeros(Np,L);
w=hamming(N);

%ESPECTROGRAMA
for l=1:1:L,
   xf=zeros(Nfft,1);
   ini=1+(l-1)*S;
   fin=ini+N-1;
   if fin<=Nx,
      xf(1:N)=x(ini:fin);
   else
      xf(1:Nx-ini+1)=x(ini:Nx);
   end
   xf(1:N)=xf(1:N).*w;
   X=fft(xf,Nfft);
   P=real(X(1:Np)).^2+imag(X(1:Np)).^2;
   SG2D(1:Np,l)=log(P);
end

%GRAFICA DE LA FORMA DE ONDA DE LA SEÑAL DE VOZ
figure(gcf);
clf;
subplot(2,1,1);
Rt=(Nx./fs).*(0:1/(Nx-1):1);
line(Rt,x);
axis tight;
xlabel('tiempo [ s ]');
ylabel('amplitud');

%ESPECTROGRAMA EN 2D
colormap('jet');
subplot(2,1,2);
Rt=(Nx./fs).*(0:1/(L-1):1);
Rf=(0.5*fs).*(0:1/(Np-1):1);
h=pcolor(Rt,Rf,SG2D);
set(h,'EdgeColor','none');
xlabel('tiempo [ s ]');
ylabel('frecuencia [ Hz ]');
```

APÉNDICE II

PROGRAMA DE MATLAB PARA OBTENER UN ESPECTROGRAMA EN 3D

```
function SG3D=Espectrograma3D(x,fs,N,S,Nfft)

%Espectrograma3D(x,fs,N,S,Nfft)
%   Programado por Alan Edgar Ballado de la Rivera UNAM FI 2012
%   Obtiene el espectrograma tridimensional de la señal de voz x, en donde:
%   fs   : es la frecuencia de muestreo.
%   N    : es la longitud o el numero de muestras de la trama.
%   S    : es el numero de muestras de separacion entre tramas adyacentes.
%   Nfft : es el numero de puntos para la transformada rapida de Fourier.

%INICIALIZACION
Nx=length(x);
L=round((Nx-N)/S+1);
Np=0.5*Nfft+1;
SG3D=zeros(Np,L);
w=hamming(N);

%ESPECTROGRAMA
for l=1:1:L,
   xf=zeros(Nfft,1);
   ini=1+(l-1)*S;
   fin=ini+N-1;
   if fin<=Nx,
      xf(1:N)=x(ini:fin);
   else
      xf(1:Nx-ini+1)=x(ini:Nx);
   end
   xf(1:N)=xf(1:N).*w;
   X=fft(xf,Nfft);
   P=real(X(1:Np)).^2+imag(X(1:Np)).^2;
   SG3D(1:Np,l)=sqrt(P);
end

%ESPECTROGRAMA EN 3D
figure(gcf);
clf;
colormap('hsv');
Rt=(Nx./fs).*(0:1/(L-1):1);
Rf=(0.5*fs).*(0:1/(Np-1):1);
surf(Rt,Rf,SG3D);
axis tight;
xlabel('tiempo [ s ]');
ylabel('frecuencia [ Hz ]');
zlabel('magnitud');
set(gca,'XGrid','off');
set(gca,'YGrid','off');
set(gca,'ZGrid','off');
set(gca,'Color','none');
```

APÉNDICE III
ECUACIÓN DE YULE-WALKER

El análisis de predicción lineal de las muestras de una señal de voz, es un proceso de identificación de un sistema todos polos, en donde se asume que la forma de onda de la señal de voz es un proceso auto-regresivo (AR) [16].

La ecuación básica del método de autocorrelación:

$$\sum_{c=1}^{p} a_c R(|i-c|) = R(i) \qquad i = 1,2,3,\ldots,p$$

$$\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & \cdots & R(p-2) \\ R(2) & R(1) & R(0) & \cdots & R(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{bmatrix}$$

Es semejante a la ecuación de Yule-Walker, empleada en la teoría moderna de identificación de sistemas [16], a excepción del signo negativo del vector columna de los coeficientes de autocorrelación:

$$\sum_{c=1}^{p} a_c R(|i-c|) = -R(i) \qquad i = 1,2,3,\ldots,p$$

$$\begin{bmatrix} R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & \cdots & R(p-2) \\ R(2) & R(1) & R(0) & \cdots & R(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \cdots & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{bmatrix}$$

Por lo tanto, los coeficientes obtenidos mediante la ecuación Yule-Walker difieren solamente por un signo negativo en comparación de los obtenidos por el método de autocorrelación.

BIBLIOGRAFÍA

[1] Claudio Becchetti; Lucio Prina Ricotti.
Speech recognition theory and C++ implementation.
John Wiley & Sons 2000.

[2] Jacob Benesty; M. Mohan Sondhi; Yiteng Huang.
Springer handbook of speech processing.
Springer 2008.

[3] John R. Deller; John G. Proakis; John H. L. Hansen.
Discrete-time processing of speech signals.
MacMillan 1993.

[4] Allen Gersho; Robert M. Gray.
Vector quantization and signal compression.
Kluwer academic publishers 1997.

[5] Ben Gold; Nelson Morgan.
Speech and audio signal processing:
Processing and perception of speech and music.
John Wiley & Sons 2000.

[6] Hynek Hermansky; Brian A. Hanson; Hisashi Wakita.
Perceptually based linear predictive analysis of speech.
IEEE 1985.

[7] Hynek Hermansky.
Perceptual linear predictive (PLP) analysis of speech.
Journal of the Acoustical Society of America. April 1990, 87(4), 1738-1752.

[8] Abel Herrera Camacho.
Notas de procesamiento digital de voz.
Universidad Nacional Autónoma de México. Facultad de Ingeniería 2007.

[9] Xuedong Huang; Alex Acero; Hsiao-Wuen Hon.
Spoken language processing:
A guide to theory, algorithm, and system development.
Prentice Hall 2001.

[10] Sunil Kumar Kopparapu; M. Laxminarayana.
Choice of mel filter bank in computing MFCC of a resampled speech.
IEEE 2010.777.

[11] Douglas O'Shaughnessy.
Speech communications: human and machine.
IEEE Press 2000.

[12] Frank J. Owens.
Signal processing of speech.
McGraw Hill 1993.

[13] Lawrence R. Rabiner; Ronald W. Schafer.
Digital processing of speech signals.
Prentice Hall 1978.

[14] Lawrence R. Rabiner; Biing-Hwang Juang.
Fundamentals of speech recognition.
Prentice Hall 1993.

[15] Chris Rowden.
Speech processing.
McGraw Hill 1992.

[16] Shuzo Saito; Kazuo Nakata.
Fundamentals of speech signal processing.
Academic press 1985.