UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA ELÉCTRICA Y ELECTRÓNICA

COMPRESIÓN DE VOZ – CÓDEC G.729

INFORME DE SUFICIENCIA

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE TELECOMUNICACIONES

PRESENTADO POR:

CRISTIAN EVERARDO ESPINOZA GIRALDO

PROMOCIÓN
2006 – II

LIMA – PERÚ
2012

COMPRESIÓN DE VOZ – CÓDEC G.729

... a Jehová mi Dios
por brindarme las fuerzas necesarias
para alcanzar mis distintos objetivos en esta vida

SUMARIO

El presente informe pretende dar a entender el funcionamiento del codificador de voz basado en la recomendación G.729 de la ITU-T, "Codificación de la voz a 8kbit/s mediante Predicción Lineal por Excitación Algebraica de Código de Estructura Conjugada – CS-ACELP".

Éste codificador está diseñado para trabajar con señales digitales y se basa en una variante del modelo de codificación por Predicción Lineal por Excitación de Código (CELP), que a su vez nace de la necesidad de optimizar el modelo de Codificación por Predicción Lineal (LPC) basado en un modelo muy simplificado de producción sintetizada de la voz.

Opera con tramas vocales de 10 ms correspondientes a 80 muestras a una velocidad de muestreo de 8000 muestras por segundo. En cada trama de 10 ms se analiza la señal vocal para extraer sólo parámetros como coeficientes del filtro de predicción lineal, ganancias e índices de las tablas de códigos adaptativos y fijos. Estos parámetros e índices se empaquetan y se transmiten. En el decodificador, se reciben estos parámetros e índices para recuperar los parámetros de excitación y del filtro de síntesis.

En este informe revisaremos las definiciones previas de los principales conceptos para proceder con la descripción de las características más. importantes de la estructura codificador G.729 y así alcanzar el objetivo de entender su funcionamiento.

CONTENIDO

INTRODUCCION **1**

CAPÍTULO I:
DEFINICIONES PREVIAS **2**

1. El Sonido 2

1.1. El Infrasonido 2

1.2. El Ultrasonido 3

2. La voz humana 3

3. Las Señales de Voz 6

3.1. Función de Auto-correlación de la Voz 7

3.2. La Frecuencia Fundamental o Pitch 8

3.3. Frecuencias formantes 8

3.4. Envolvente Espectral 9

4. Sistema Auditivo Humano 10

4.1. Estructura del sistema auditivo humano 11

4.2. El umbral absoluto 12

4.3. Enmascaramiento 13

4.4. Fase de percepción 15

5. Conversión Analógico-Digital 15

5.1. Muestreo 16

5.2. Cuantificación 16

5.3. Codificación 16

5.4. Ventajas de la señal digital 18

5.5. Inconvenientes de la señal digital 18

CAPÍTULO II:
CODIFICACION DE VOZ **20**

1. Estructura General de un codificador de Voz 20

2. Tipos de Codificadores 21

3. Codificadores de Forma de Onda 22

4. Codificadores Paramétricos (Vocoders) 23
4.1. Vocoder LPC 24
a) Filtro LPC 25
b) Excitación del Modelo LPC 28
c) Cuantización de Parámetros LPC 28
4.2. Codificadores Sinusoidales 29
5. Codificadores Híbridos 29
5.1. El Análisis por Síntesis 30
5.2. Codificadores Multipulso 31
a) Filtro de Pitch o LTP 33
b) Filtro de Peso 34
5.3. Codificadores CELP 35
a) La Técnica ACELP 36
5.4. Codificadores Híbridos en Aplicaciones VoIP 36

CAPÍTULO III
CODEC G.729 39

1. Introducción 39
2. Retardo 40
3. Estructura del Libro de Códigos Algebraico 40
4. Libro de Códigos Adaptivo 42
4.1. Filtro de ponderación perceptual 42
4.2. Estimación del Período de Pitch de Lazo Abierto 43
4.3. Rango de búsqueda para el Período de Pitch 44
4.4. Señales involucradas en la Búsqueda del Libro de Códigos Adaptivo 44
a) Matriz de Libro de Códigos Adaptivo o Secuencia de Excitación: 45
b) Elección de la secuencia para la Búsqueda del Libro de Códigos Adaptivo 45
4.5. Búsqueda del Retardo Entero 45
4.6. Búsqueda del Retardo Fraccionario 46
4.7. Generación del Codevector Adaptivo 46
4.8. Resumen 47
5. Codificación y Decodificación 48
5.1. Señales involucradas en la Codificación y Actualización del Filtro de Estado 48
5.2. Revisión de las operaciones de codificación 50
5.3. Resumen de Operaciones de Decodificación 51

CAPÍTULI IV
SIMULACIONES 52
1. Captura y Reproducción de Tramas RTP codificadas con G.711 53
1.1. Esquema utilizado: 53
1.2. Captura de trama con Wireshark 54
1.3. Filtrado de las llamadas VoIP. 54
1.4. Gráfica de análisis de la llamada 55
1.5. Reproducción de la llamada (Las tramas RTP) 55
2. Captura y Reproducción de Tramas RTP codificadas con G.729 56
2.1. Esquema: 57
2.2. Captura de la trama de una llamada SIP (G729) 57
2.3. Filtrado de las llamadas VoIP. 58
2.4. Gráfica para analizar la llamada 58
2.5. Reproducción de la llamada (Las tramas RTP) con wireshark 59
2.6. Exportar las tramas codificadas en G.729 59
2.7. Decodificando con el codificador G.729 de VoiceAge. 60
2.8. Reproduciendo las señales de voz con el reproductor Audacity. 60

ANEXO A
G729 ANEXO A (G.729A) 62
ANEXO B
G729 ANEXO B (G.729B) 65
ANEXO C
ESTANDAR G.729.1 68
ANEXO D
SOFTWARE WIRESHARK 71
ANEXO E
CODIFICADOR G.729 DE VOICEAGE 74
ANEXO F
REPRODUCTOR AUDACITY 76
CONCLUSIONES 78
BIBLIOGRAFIA 80

INTRODUCCIÓN

El objetivo de este informe es comprender el funcionamiento del códec G.729.

Para alcanzar el objetivo planteado, realizaremos un breve estudio de acuerdo a la siguiente estructura.

En el Capítulo I, estudiaremos los conceptos básicos necesarios para comprender los componentes involucrados en la síntesis de la voz, iniciando por el sonido, el cual da origen a la voz, de la cual revisaremos también sus señales y componentes principales como la frecuencia fundamental, frecuencias formantes y la envolvente de estas señales, al no ser el sistema fonador humano el único involucrado, analizaremos también el sistema auditivo, del cual se aprovechan ciertas características como el umbral absoluto, su propiedad se enmascaramiento y la fase de percepción. En este capítulo también revisaremos la digitalización de las señales analógicas que es el primer paso antes de poder ingresar a la codificación y compresión de la voz.

En el capítulo II, habiendo revisado los conceptos básicos podremos analizar ahora sí a los Codificadores de Voz, su estructura, y tipos.

Entre los tipos nos detendremos un poco en los conceptos que nos permitan llegar a entender el funcionamiento del códec G.729.

En el capítulo III, entramos a detallar en sí el comportamiento y las características de los elementos del codificador de voz.

CAPÍTULO I.
DEFINICIONES PREVIAS

1. El Sonido

El sonido es el fenómeno originado por variaciones de presión o vibraciones de moléculas de aire, las cuales llegan hasta nuestro sentido del oído en forma de ondas acústicas. Estas ondas pueden ser audibles o no, las ondas cuya frecuencia se limite a un rango superior a los 20 Hertz o ciclos por segundo para los sonidos más graves e inferior de los 20 mil Hertz (20 kHz) o ciclos por segundo para los más agudos serán audibles. Cualquier cuerpo que vibre dentro de esa gama de frecuencias, generará sonidos que podrán ser captados por nuestro sentido del oído como una onda sonora audible, las ondas con frecuencias por debajo de los 20Hz son consideradas como infrasonidos y por encima de los 20 kHz se convierten en ultrasonidos, las cuales son imperceptibles por el oído humano.

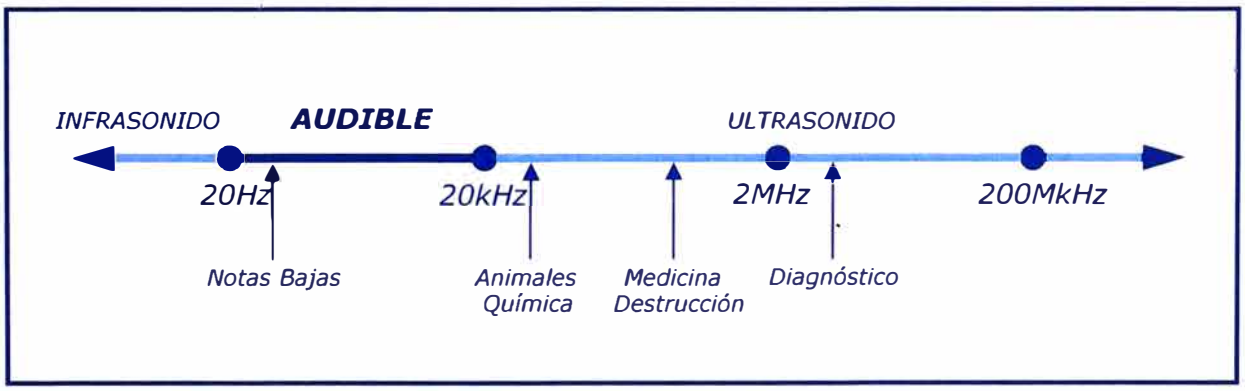

Fig. 1.1: Rango de Frecuencias Acústicas

1.1. El Infrasonido

El infrasonido es la onda acústica cuya frecuencia es tan baja que no puede ser percibida por el oído humano. Generalmente aplica para las frecuencias inferiores a 16 o 17 Hertz hasta los 0,001 Hertz. Algunos animales pueden percibir el infrasonido y usarlo para su beneficio. Posiblemente la mayoría de los animales pueden captar el infrasonido generado por un terremoto, lo cual hace que sean los primeros en inquietarse ante estos desastres.

El infrasonido tiene la característica de poder cubrir grandes distancias y traspasar objetos con poca disipación, se utiliza en sismografía para monitorear terremotos.

1.2. El Ultrasonido

El ultrasonido es la onda acústica imperceptible por el oído humano debido a que pertenece a un rango de frecuencias superior a los 20kHz. Animales como murciélagos, delfines y perros, logran oír estas frecuencias, e incluso utilizarlas como radar para orientarse y cazar.

El ultrasonido se utiliza en muchos ámbitos de la ciencia y la tecnología, por ejemplo:

En la medicina se emplea para el diagnóstico por ultrasonido ó econografía que permite apreciar diferencias en la densidad de un órgano a diferencia de los rayos X que sólo aportan datos sobre el contorno y forma del mismo, la fisioterapia, la limpieza dental, la liposucción, etc.

En la industria se utiliza para medir distancias, ensayos no destructivos, caracterización interna de materiales, limpieza de superficies, etc.

En el ámbito militar el ultrasonido puede utilizarse como arma.

2. La voz humana

Realizaremos una breve descripción de la anatomía del aparato fonatorio humano

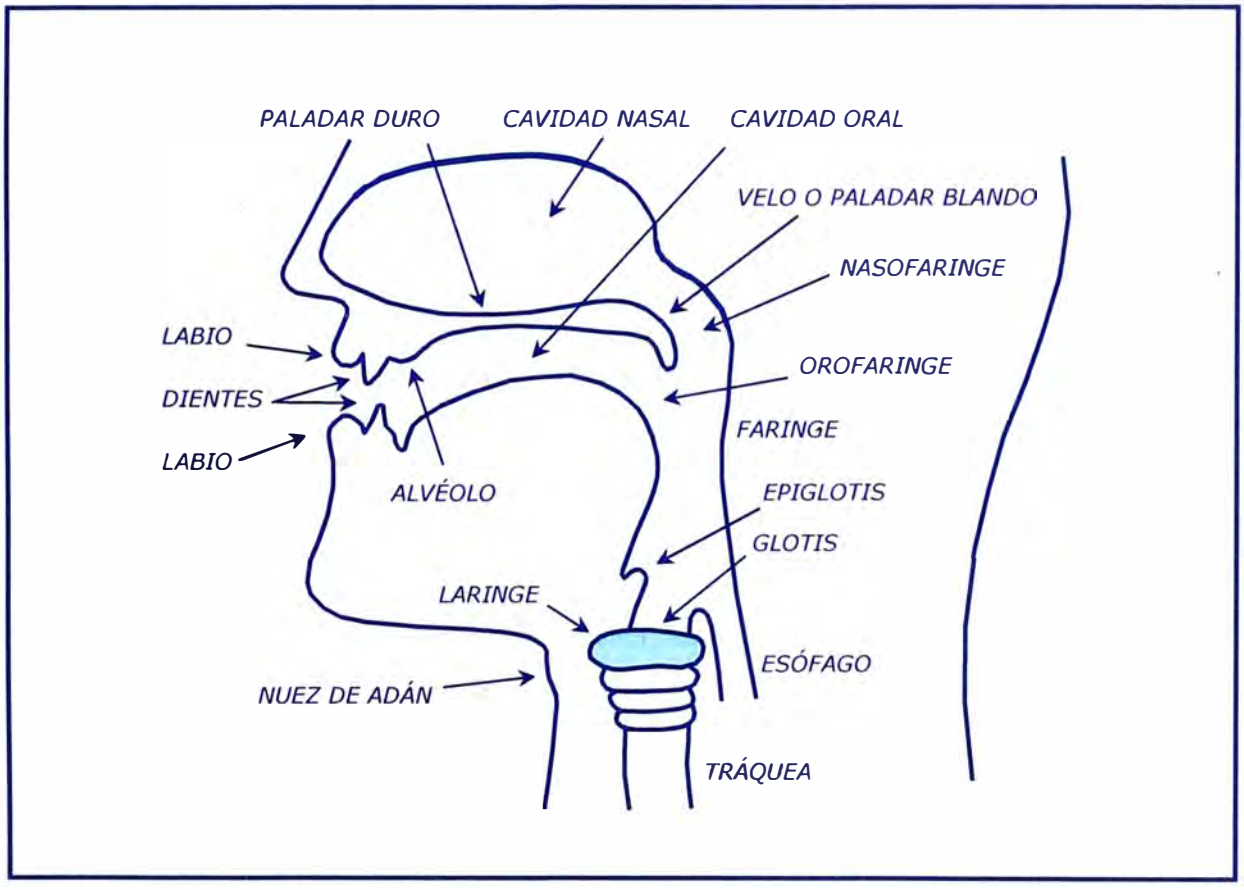

Fig. 1.2: Corte esquemático del aparato fonatorio humano.

La voz humana se produce voluntariamente por medio del aparato fonatorio, éste está formado por los pulmones como fuente de energía en la formación de un flujo de aire, la laringe, que contiene las cuerdas vocales, la faringe, las cavidades oral (o bucal) y nasal y

una serie de elementos articulatorios: los labios, los dientes, el alvéolo, el paladar, el velo del paladar y la lengua (Fig. 1.2).

Las cuerdas vocales son dos membranas dentro de la laringe orientadas de adelante hacia atrás como lo muestra la Fig. 1.3, por delante se unen en el cartílago tiroides (que puede palparse sobre el cuello, en los varones suele apreciarse como una protuberancia conocida como nuez de Adán), por detrás, cada una está sujeta a uno de los dos cartílagos aritenoides, los cuales pueden separarse voluntariamente por medio de músculos, la abertura entre ambas cuerdas se denomina glotis.

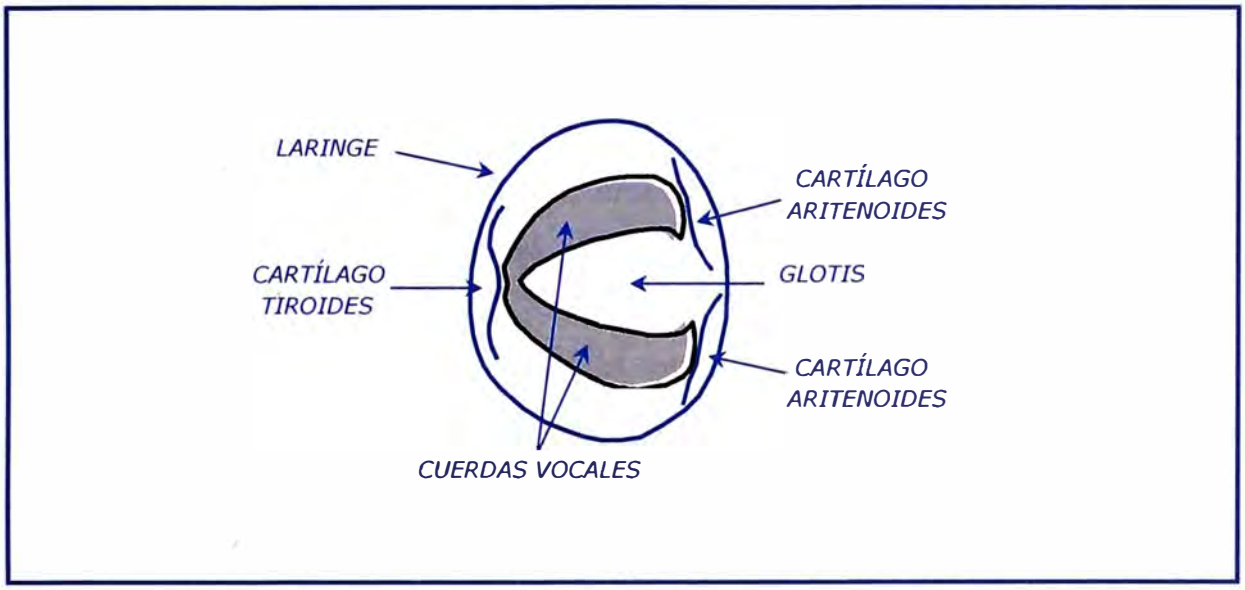

Fig. 1.3: Corte esquemático de la laringe según un plano horizontal

Cuando las cuerdas vocales se encuentran separadas, la glotis adopta una forma triangular, permitiendo el paso libre del aire, con lo que prácticamente no se produce sonido, tal es el caso de la respiración. Cuando la glotis empieza a cerrarse, el aire que la atraviesa experimenta una turbulencia, al cerrarse más, las cuerdas vocales comienzan a vibrar, produciéndose un sonido tonal, es decir periódico. La frecuencia de este sonido depende de varios factores, el tamaño y la masa de las cuerdas vocales, la tensión que se les aplique y de la velocidad del flujo del aire proveniente de los pulmones entre otros. A mayor tamaño, menor frecuencia de vibración, lo cual explica por qué en los varones, cuya glotis es en promedio mayor que la de las mujeres, la voz es en general más grave. A mayor tensión, la frecuencia aumenta, siendo los sonidos más agudos, es por tal motivo que para lograr emitir sonidos en el registro extremo de la voz es necesario un mayor esfuerzo vocal. La frecuencia también aumenta al aumentar la velocidad del flujo de aire, razón por la cual al aumentar la intensidad de emisión se tiende a elevar espontáneamente el tono de voz. Es posible obturar la glotis completamente, en ese caso no se produce sonido.

Sobre la glotis se encuentra la epiglotis, un cartílago en la faringe que permite tapar la glotis durante la deglución para evitar que el alimento ingerido se introduzca en el tracto respiratorio. Durante la respiración y la fonación (emisión de sonido) la epiglotis está separada de la glotis permitiendo la circulación del flujo de aire. Durante la deglución, en cambio, la laringe ejecuta un movimiento ascendente de modo que la glotis apoya sobre la epiglotis.

La porción que incluye las cavidades faríngea, oral y nasal junto con los elementos articulatorios se denomina genéricamente cavidad supraglótica, en tanto que los espacios por debajo de la laringe, es decir la tráquea, los bronquios y los pulmones, se denominan cavidades infraglóticas. Varios de los elementos de la cavidad supraglótica se controlan a voluntad, permitiendo modificar dentro de márgenes muy amplios los sonidos producidos por las cuerdas vocales o agregar partes distintivas a los mismos, e inclusive producir sonidos propios. Todo esto se efectúa por dos mecanismos principales: el filtrado y la articulación.

El filtrado actúa modificando el espectro del sonido. Tiene lugar en las cuatro cavidades supraglóticas principales: la faringe, la cavidad nasal, la cavidad oral y la cavidad labial. Las mismas constituyen resonadores acústicos que enfatizan determinadas bandas frecuenciales del espectro generado por las cuerdas vocales, conduciendo al concepto de formantes, es decir una serie de picos de resonancia ubicados en frecuencias o bandas de frecuencia que, según veremos, son bastante específicas para cada tipo de sonido.

Tabla 1.1: Clasificación de las consonantes de la lengua castellana según el lugar y el modo de articulación y la sonoridad.

Lugar de articulación	Modo de Articulación								
	Oral								Nasal
	Oclusiva		Fricativa		Africada	Lateral	Vibrante	Aproximante	
	Sorda	Sonora	Sorda	Sonora	Sorda	Sonora	Sonora	Sonora	Sonora
Bilabial	p	b,v		b,v				w	m
Dentilabial			f						
Dentilingual			z	d					
Alveolar	t	d	s	y	ch	l	r, rr		n
Palatal				(y)	(ch)	ll		i	ñ
Velar	k	g							
Glotal									

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

Tabla 1.2: Clasificación de las vocales según la posición de la lengua.

Posición Vertical	Tipo de Vocal	Posición Horizontal		
		Anterior	Central	Posterior
Alta	Cerrada	i		u
Media	Media	e		o
Baja	Abierta		a	

3. Las Señales de Voz

La señal de voz se constituye por un conjunto de sonidos generados por el aparato fonador. Esta señal acústica puede ser transformada en una señal eléctrica a través de un micrófono. La misma que puede representarse en el dominio del tiempo en un par de ejes cartesianos.

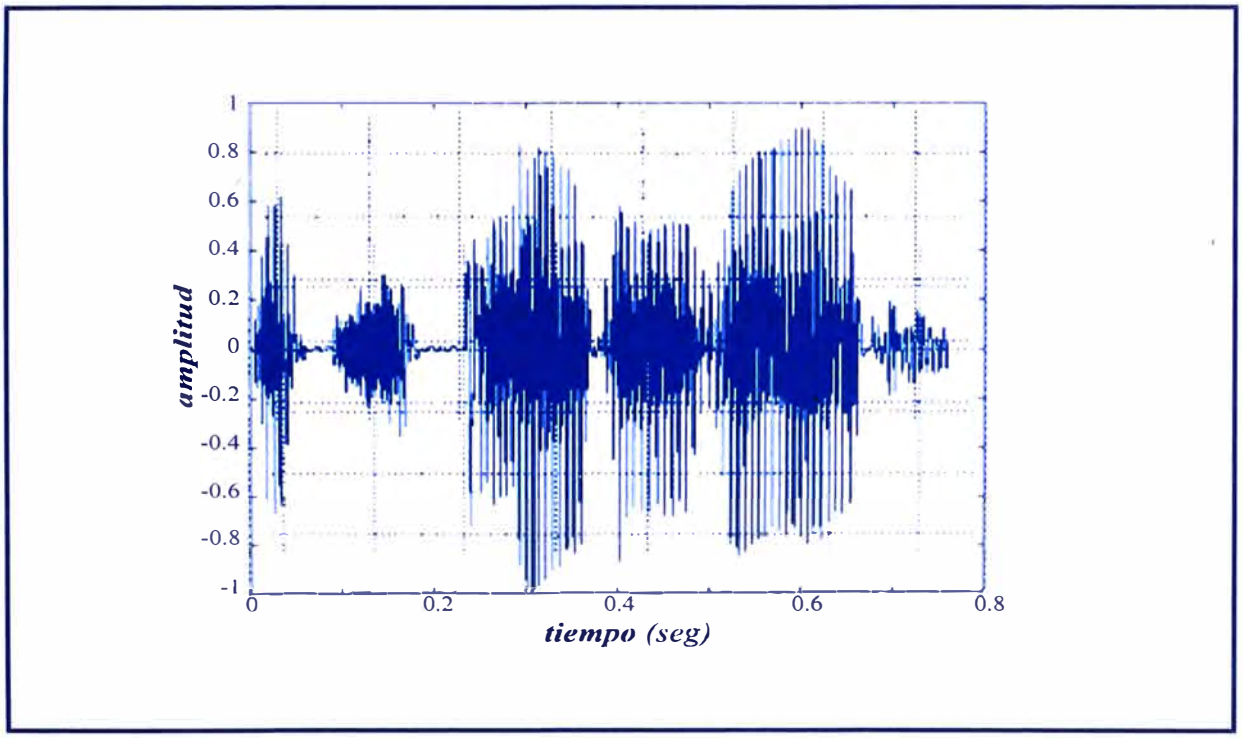

Fig. 1.4: Forma de onda de la palabra "explorador"

Como todos los sonidos, está formada esencialmente por curvas elementales (senos y cosenos) pero las posibles combinaciones de éstas pueden ser muy complejas. En la Fig. 1.4 se muestra la representación de la señal de voz en función del tiempo, la cual brinda información sobre características importantes como la energía y los cruces por cero, las cuales facilitan su estudio y análisis.

Podemos realizar el estudio de la señal de voz en el dominio de la frecuencia, con la finalidad de conocer sus características espectrales.

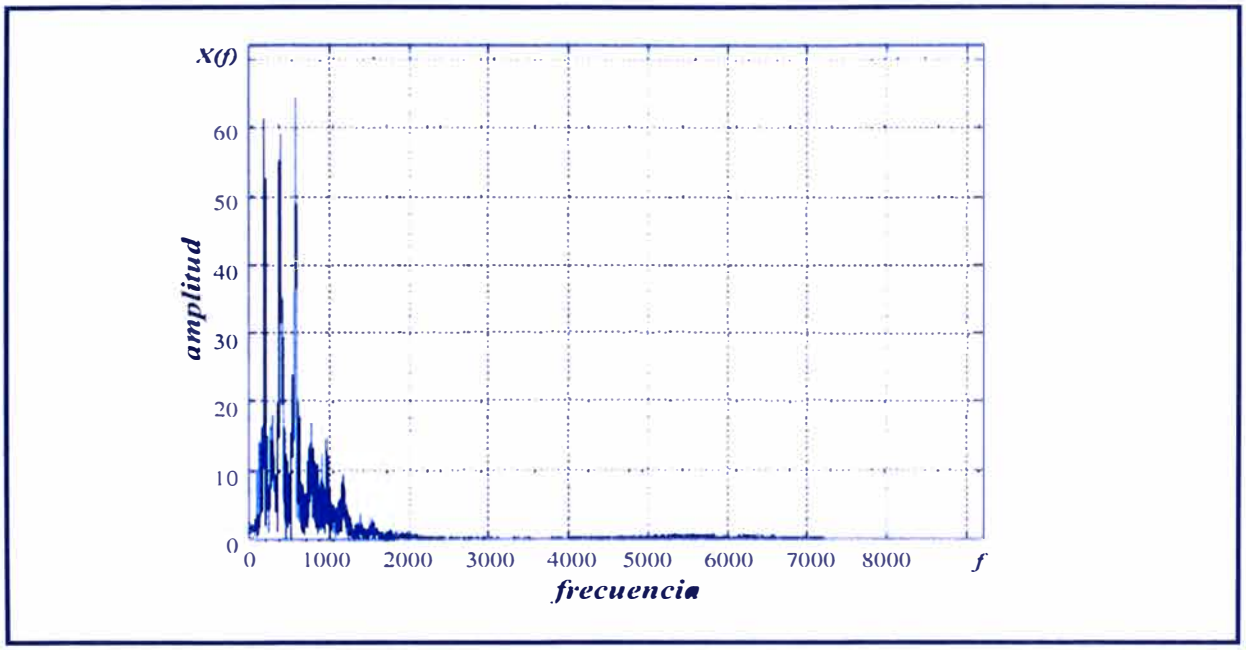

Fig. 1.5: Espectro de frecuencia de la palabra 'dos'.

3.1. Función de Auto-correlación de la Voz

La función de auto-correlación es el resultado de comparar la señal de voz consigo misma desplazada m muestras en el tiempo, dándonos una indicación de la estructura de dependencia entre esta y las muestras sucesivas. La señal de voz presenta tanto correlaciones importantes a corto como a largo plazo (muestras próximas y lejanas respectivamente). Siendo las primeras debidas al tracto vocal y las segundas a la vibración de las cuerdas vocales.

Podemos estimar las auto-correlaciones de un tono y de un ruido.

$$y = cos(2 * pi * (1:1000)/25);$$

$$y = randn(1,1000);$$

En el primer caso, observaremos la periodicidad de la auto-correlación obtenida, mientras que en el segundo vemos una función completamente aleatoria.

Una de las aplicaciones de dicha comparación es la determinación del valor de la frecuencia fundamental o pitch en el caso de que la señal de voz sea sonora teniendo en cuenta que la auto-correlación adquiere un valor máximo cuando el desplazamiento es igual cero.

Podemos aplicar esta medida a la voz y analizar el resultado. Especialmente deberemos fijarnos en las características de segmentos sonoros y sordos, identificando tanto la correlación a corto plazo, como la presencia o no de correlación a largo plazo.

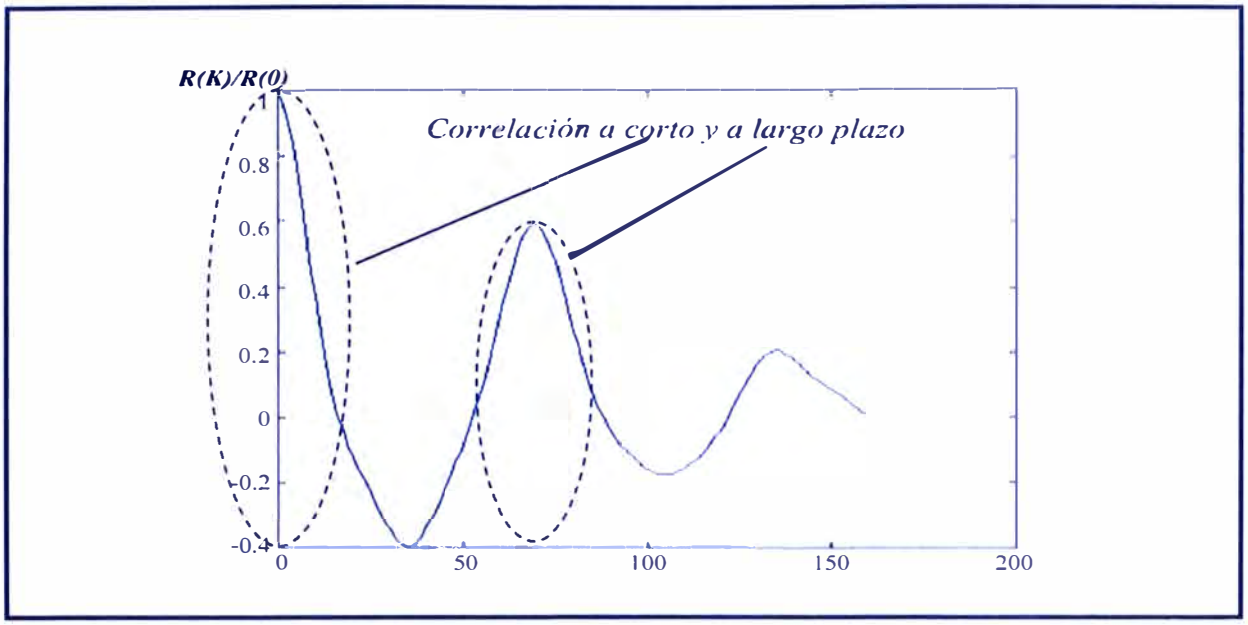

Fig. 1.6: Diagrama donde se muestra la correlación a corto y largo plazo de una señal de voz

Tras analizar la información que nos muestra la función de auto-correlación aplicada a la señal de voz, podremos identificar tramos sordos y sonoros, y en los tramos sonoros podremos medir el periodo fundamental para estimar la frecuencia fundamental. Esta identificación y medida de la frecuencia fundamental está vinculada tanto a la presencia de un valor importante de auto-correlación, que señala la decisión sonoro/sordo, así como a su posición, que indica el valor del periodo fundamental.

3.2. La Frecuencia Fundamental o Pitch

La frecuencia fundamental o también denominada pitch, brinda información sobre la velocidad a la que vibran las cuerdas vocales al producir un sonido, produciendo un espectro de frecuencia similar al mostrado en la Fig. 1.6. Este espectro podría ser obtenido si se colocara un micrófono de amplio rango directamente en la garganta, encima de las cuerdas vocales, pero debajo de las estructuras resonantes del tracto vocal.

El espectro está conformado de armónicos del periodo pitch, el cual es el rango fundamental de frecuencia producidas por las cuerdas vocales. Si bien el espectro lleva un gran componente cerca de la frecuencia pitch, también tiene gran cantidad de armónicos y así tiene componentes de frecuencia.

3.3. Frecuencias formantes

Los componentes de la cavidad supraglótica actúan como resonadores acústicos. Si se realiza un análisis espectral del sonido luego de haber atravesado estas cavidades, el efecto de la resonancia produciría un énfasis en determinadas frecuencias del espectro obtenido, a las que se les denominara 'formantes'. Existen tantas formantes como resonadores posee el

tracto vocal. Sin embargo se considera que sólo las tres primeras, asociadas a la cavidad oral, bucal y nasal respectivamente y proporcionan la suficiente cantidad de información para poder diferenciar los distintos tipos de sonido. En la Fig. 1.7 se muestra el espectro de la palabra 'uno', y se denominan F1, F2 y F3 a sus tres principales frecuencias formantes. La amplificación de cada una de estas tres frecuencias depende del tamaño y forma que adopta la cavidad bucal y la cavidad oral, y si el aire pasa o no por la nariz.

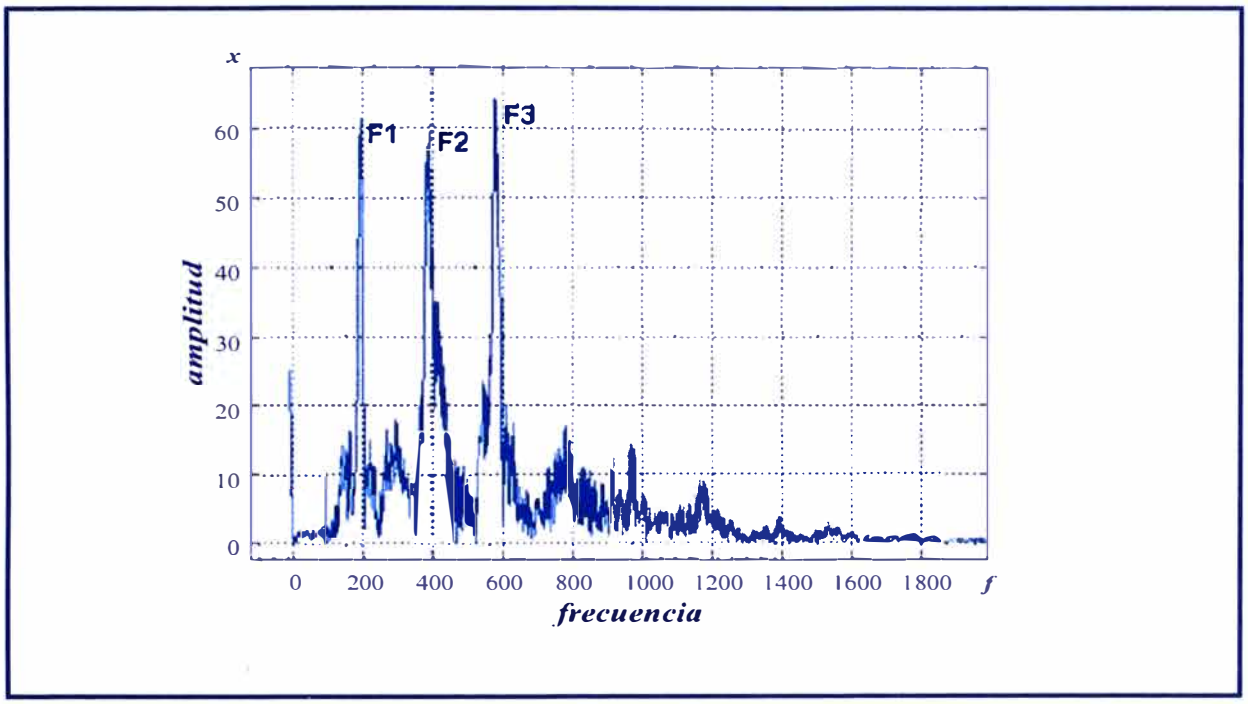

Fig. 1.7: Frecuencias Formantes de la palabra 'dos'.

3.4. Envolvente Espectral

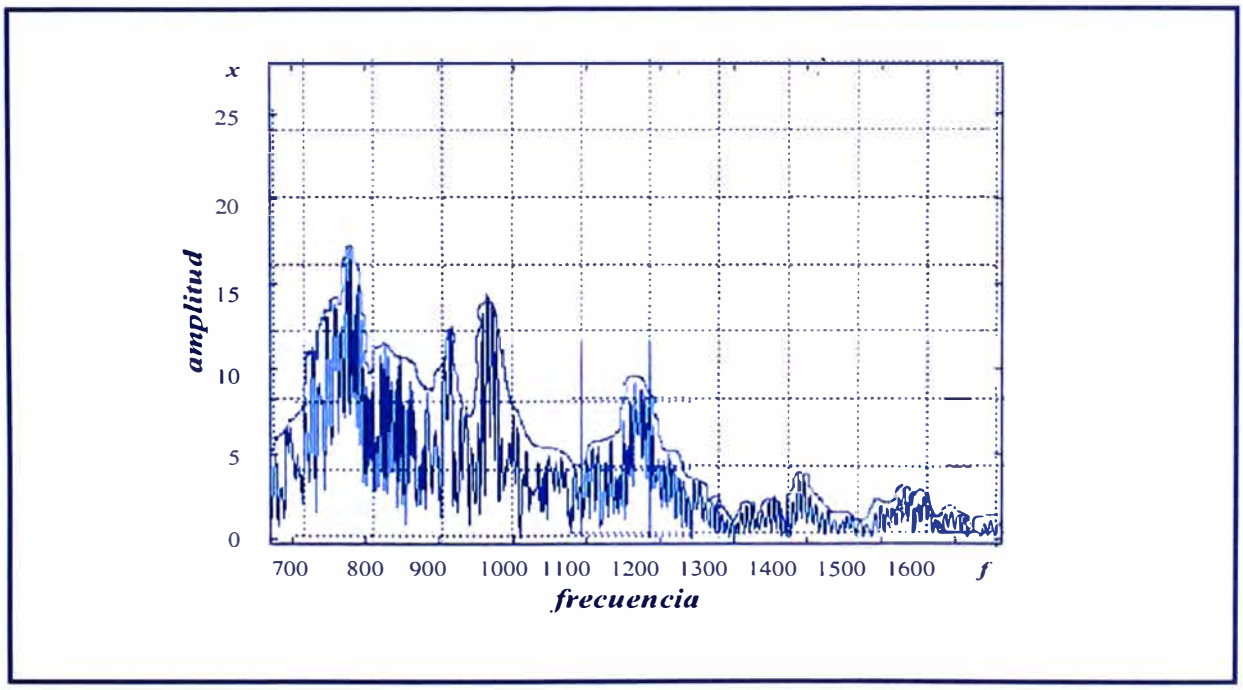

Fig. 1.8: Envolvente espectral.

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

Las señales de voz, no son más que una secuencia de símbolos acústicos elementales conocidos como fonemas, además de su función en la comunicación con información lingüística, llevan consigo gran cantidad de información de otros tipos. Como hemos visto no es igual para todas las personas debido a las características del individuo, tanto las constitucionales, anatómicas como anímicas. Algunas de estas características son: edad, intensidad, emoción, humor, el estado de salud, etc.

Se caracterizan por ser no estacionarias y con variaciones lentas en el dominio del tiempo y se procesan normalmente en segmentos cortos de tiempo, entre 5 y 30 ms. En estos segmentos la señal puede aproximarse a una señal cuasi-periódica. Cada segmento o trama puede ser clasificado como sonoro, no sonoro o silencio:

Los sonidos sonoros tienen una naturaleza cuasi-periódica en el dominio del tiempo y una estructura armónica fina en el dominio de la frecuencia, provocada por la vibración de las cuerdas vocales, su espectro decae en las frecuencias altas, posee una energía es alta debido a que el aire encuentra poca obstrucción al pasar por el tracto vocal. Estos sonidos consisten en una frecuencia fundamental (frecuencia de pitch) y una serie de componentes armónicos de la misma, producidos por las cuerdas vocales. El tracto vocal modifica la señal de excitación provocando frecuencias formantes, las cuales son las frecuencias de resonancia del espectro, es decir, son los picos de la envolvente del espectro de la señal de voz que representan las frecuencias de resonancia del tracto vocal. Cada formante tiene una amplitud y un ancho de banda. La frecuencia fundamental y las frecuencias de los formantes son los conceptos más importantes para la síntesis y procesamiento de la voz.

4. Sistema Auditivo Humano

La forma en la que funciona el sistema auditivo humano juega un papel importante en el diseño de sistemas de codificación de voz.

Al comprender cómo se perciben los sonidos, se pueden asignar los recursos del sistema de codificación de la manera más eficiente.

Muchos de estándares de codificación de voz están diseñados para tomar ventaja de las propiedades del sistema auditivo humano.

Esta sección proporciona una visión general, resumiendo varios temas, incluyendo la estructura del sistema auditivo humano, el umbral absoluto, el enmascaramiento, y la percepción de fase.

4.1. Estructura del sistema auditivo humano

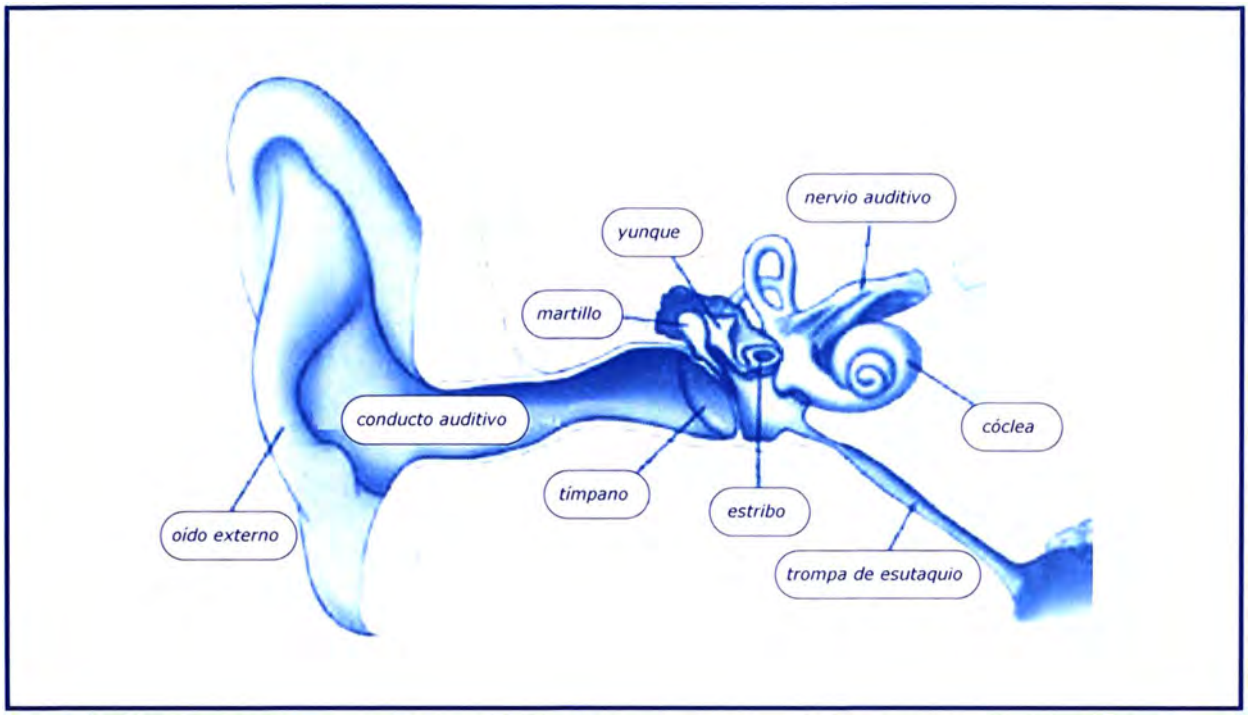

Fig. 1.9: Sistema Auditivo Humano.

En la Fig. 1.9 se presenta un diagrama simplificado del sistema auditivo humano. El pabellón de la oreja es la superficie que rodea el canal en el cual se canaliza el sonido. Las ondas sonoras son guiadas por el canal hacia el tímpano, una membrana que actúa como un transductor acústico-a-mecánico. Las ondas sonoras se traducen en vibraciones mecánicas que se transmiten a la cóclea a través de una serie de huesos conocidos como la "cadena de huesecillos". La presencia de la cadena de huesecillos mejora la propagación del sonido mediante la reducción de la cantidad de reflexión y se lleva a cabo por el principio de adaptación de impedancia.

La cóclea es un órgano rígido en forma de caracol, lleno de líquido. Oscilaciones mecánicas causadas por una membrana interna inciden en la cadena de huesecillos, esta membrana interna es conocida como la membrana basilar, que vibra a diferentes frecuencias. La membrana basilar se caracteriza por un conjunto de respuestas de frecuencia en diferentes puntos a lo largo de la membrana, para describir su comportamiento se utiliza una técnica de modelado simple de un banco de filtros. Movimiento a lo largo de la membrana basilar es detectado por las células ciliadas internas y hace que las actividades neuronales que se transmiten al cerebro a través del nervio auditivo.

Los diferentes puntos a lo largo de la membrana basilar reaccionan de forma diferente en función de las frecuencias de las ondas sonoras que se reciben. De este modo, las células

ciliadas ubicadas en diferentes posiciones a lo largo de la membrana son excitadas por los sonidos de diferentes frecuencias. Las neuronas hacen contacto con las células ciliadas y transmiten la excitación a los centros auditivos superiores.

Debido a esta disposición, el sistema auditivo humano tiene un comportamiento muy parecido a un analizador de frecuencias, y caracterización del sistema es más simple si se hace en el dominio de la frecuencia.

4.2. El umbral absoluto

El umbral absoluto de un sonido es el nivel mínimo detectable de sonido en ausencia de cualquier otro sonido externo. Es decir, caracteriza la cantidad de energía necesaria en un tono puro de modo que pueda ser detectado en un entorno sin ruido. La Fig. 1.10 muestra una curva típica de umbral absoluto, donde el eje horizontal es la frecuencia medida en hercios (Hz), mientras que el eje vertical es el umbral absoluto medido en decibelios (dB), relacionado a una intensidad de referencia de 10^{12} vatios por metro cuadrado (una cantidad estándar para la medición de la intensidad del sonido.

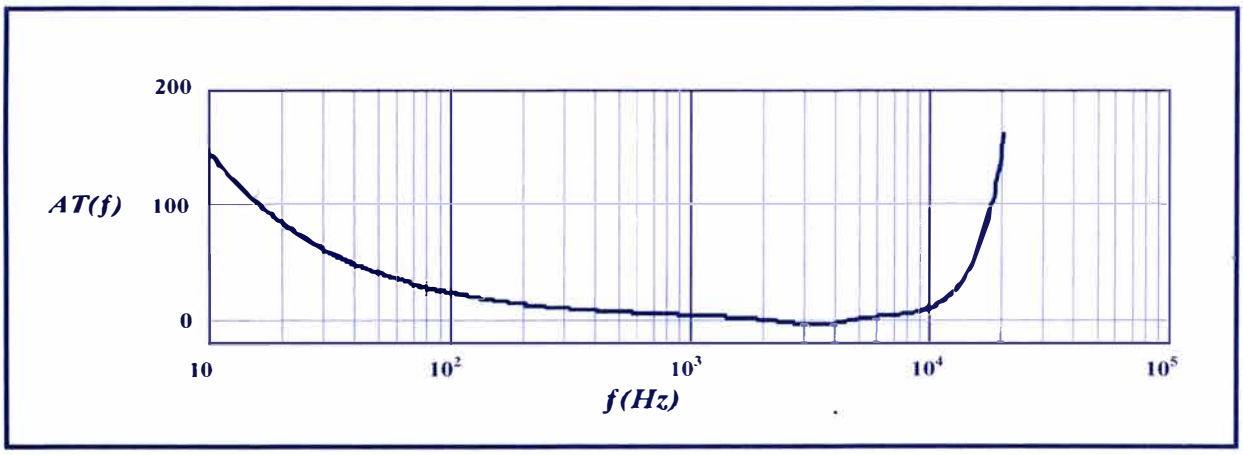

Fig. 1.10: Una curva típica de umbral absoluto.

Notamos que la curva de umbral absoluto, como se muestra en la Fig. 1.10, refleja sólo el comportamiento promedio; la forma real varía de una persona a otra y se mide mediante la presentación de un tono de una cierta frecuencia en un sujeto, con la intensidad siendo sintonizada hasta que el sujeto ya no perciba su presencia. La curva de umbral absoluto resulta de la repetición de las mediciones usando un gran número de valores de frecuencia.

Como podemos ver, los seres humanos tienden a ser más sensibles en frecuencias en el rango de 1 a 4 kHz, mientras que los umbrales incrementan rápidamente a frecuencias muy altas y muy bajas. Se considera comúnmente que por debajo de 20 Hz y por encima de 20 kHz, el sistema auditivo es esencialmente disfuncional. Estas características son debido a la estructura del sistema auditivo humano: la selectividad acústica del pabellón de la oreja

y el canal, las propiedades mecánicas de la membrana timpánica y los huesecillos, la elasticidad de la membrana basilar, y así sucesivamente.

Se puede tomar ventaja de la curva de umbral absoluto en el diseño de codificador de voz. Algunos enfoques son los siguientes:

- Cualquier señal con una intensidad por debajo del umbral absoluto no necesita ser considerada, ya que no tiene ningún impacto sobre la calidad final del codificador.
- Más recursos deben asignarse para la representación de las señales en el rango de frecuencia más sensible, más o menos 1 a 4 kHz, ya que las distorsiones en este rango son más evidentes.

4.3. Enmascaramiento

El enmascaramiento se refiere al fenómeno donde se representa un sonido audible debido a la presencia de otros sonidos. La presencia de un solo tono, por ejemplo, puede enmascarar las señales de vecinos con la capacidad de enmascaramiento inversamente proporcional a la diferencia absoluta en frecuencia.

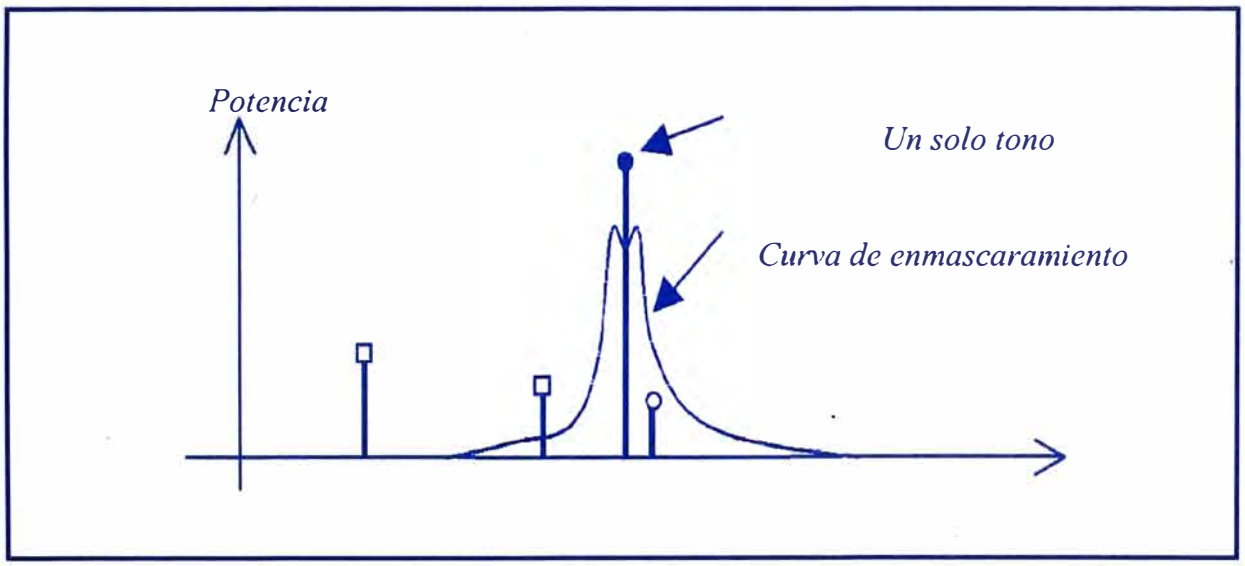

Fig. 1.11: Ejemplo de curva de enmascaramiento asociado a un solo tono.

Se muestra sobre la base de la curva de enmascaramiento ejemplos de tonos audibles (□) y no audibles (O), que dependerán de si la potencia está por encima o por debajo de la curva de enmascaramiento, respectivamente.

Este es un ejemplo donde un solo tono está presente; el tono genera una curva de enmascaramiento que hace que cualquier señal con una potencia por debajo de ella se convierta en imperceptible. En general, enmascarando la capacidad aumenta con la intensidad de la señal de referencia, o el tono único en este caso.

Las características de la curva de enmascaramiento dependen de cada individuo y se

pueden medir en la práctica, poniendo un objeto en un entorno de laboratorio y pedir su percepción de un cierto sonido sintonizado en cierta amplitud y frecuencia en presencia de un tono de referencia.

El enmascaramiento es explorado en el desarrollo de codificadores de voz. Analizando el contenido espectral de una señal, es posible localizar las regiones de frecuencia que son más susceptibles a la distorsión, tal como se muestra en la Fig. 1.12. En donde se muestra un espectro típico, que consta de una serie de regiones de alta y baja potencia, conocidos como picos y valles respectivamente. Existe una curva de enmascaramiento asociada que sigue los altibajos del espectro original. Las señales por debajo de la curva de enmascaramiento son inaudibles, por lo que, en general, los picos pueden tolerar más distorsión o ruido que los valles.

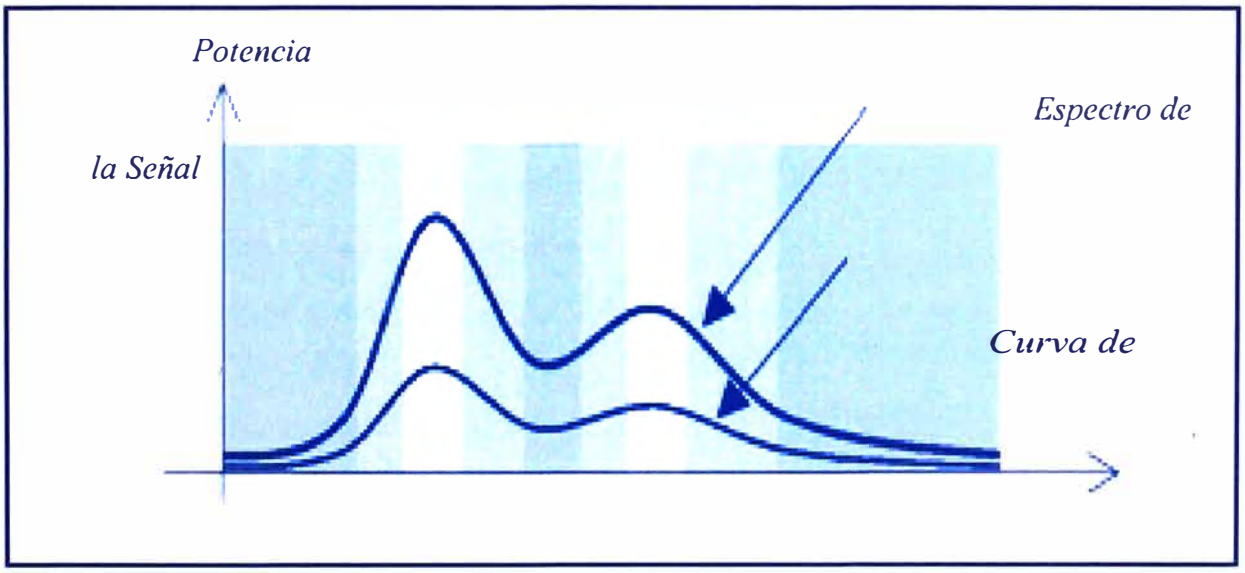

Fig. 1.12: Curva de enmascaramiento asociada de un espectro de señal.

En la figura mostrada, las áreas oscuras corresponden a regiones con relativamente baja tolerancia a la distorsión, mientras que las zonas claras corresponden a regiones con tolerancia relativamente alta a la distorsión.

Un sistema de codificación bien diseñado debe asegurarse de que los valles se conserven libres de distorsiones, mientras que los picos pueden tolerar una mayor cantidad de ruido. Al seguir este principio, la eficacia del algoritmo de codificación se mejora, dando lugar a una mejor calidad de voz sintetizada.

Como observaremos más adelante, los codificadores CELP se basan en un filtro de ponderación perceptual para ponderar el espectro de error durante la codificación; cuya respuesta de frecuencia es variable en el tiempo y depende del espectro original del la señal de entrada, este mecanismo es muy eficiente y se aplica ampliamente en la práctica.

4.4. Fase de percepción

Las tecnologías modernas de codificación de voz dependen en gran medida la aplicación de las características perceptuales del sistema auditivo humano en diversos aspectos del diseño de un cuantificador y la arquitectura en general. En la mayoría de los casos, sin embargo, el enfoque de la percepción se limita en gran medida a la información de la magnitud de la señal, en contraparte la mayoría ha abusado negligentemente del supuesto de que los seres humanos tienen fases sordas.

Un ejemplo de la fase sorda es que un solo tono y su versión desplazada en el tiempo producen la misma sensación; por otro lado, la percepción del ruido es principalmente determinado por la magnitud del espectro. Este último ejemplo es la base de los primeros codificadores de voz, tales como el LPC.

A pesar de que la fase tiene un papel menor en la percepción, sigue siendo deseable un cierto nivel de preservación de fase en el proceso de codificación, ya que la naturalidad es normalmente incrementada. El algoritmo CELP, por ejemplo, tiene un mecanismo para retener la información de fase de la señal.

5. Conversión Analógico-Digital

La conversión analógica-digital consiste en la transcripción de señales analógicas en señales digitales, con el propósito de facilitar su procesamiento (codificación, compresión, etc.) y hacer de la señal resultante (la digital) más inmune al ruido y otras interferencias a las que son más sensibles las señales analógicas.

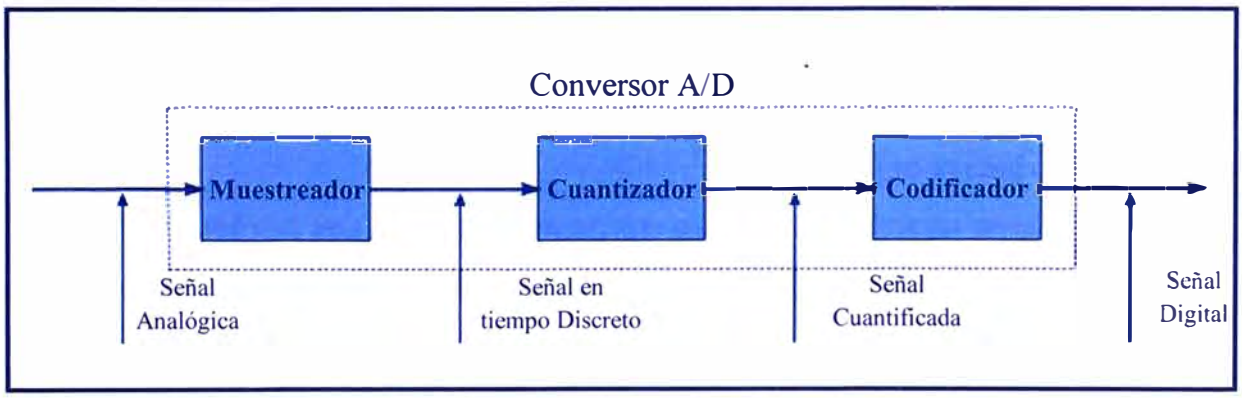

Fig. 1.13: Procesos de la Conversión A/D.

El objetivo fundamental de la codificación de voz es la conversión de la señal de voz a una secuencia binaria o representación digital. Dado el carácter analógico (señal continua en tiempo y amplitud) de la señal de voz, la codificación de voz conlleva un proceso básico de muestreo y cuantificación para conseguir una representación digital (conversión analógico/digital A/D).

5.1. Muestreo

Mediante el muestreo discretizamos la señal en el tiempo, para que no exista pérdida de información en este proceso de digitalización, debemos muestrear la señal a una velocidad (f_m) que como mínimo sea el doble de la frecuencia más alta presente en la señal que estamos discretizando.

$$f_m \geq 2f_{MAX} \tag{1.1}$$

5.2. Cuantificación

Mediante la cuantificación discretizamos la señal en amplitud. En el proceso de discretización en amplitud debemos utilizar un número (N) de bits por muestra que resulte adecuado para la calidad deseada.

$$niveles\ de\ cuantización = 2^N \tag{1.2}$$

Así, por ejemplo, una señal de voz con una frecuencia máxima de f_{MAX} kHz, lo que supone una frecuencia de muestreo mínima de $2f_{MAX}$ kHz y donde se utiliza una representación con N bits por muestra (2^N niveles de cuantización) con una distribución logarítmica de los niveles de cuantificación a lo largo del margen dinámico de la señal, lo que supone una velocidad de transmisión, almacenamiento o grabación de:

$$velocidad\ de\ transmisión = (N \times f_m)\frac{kb}{s} \tag{1.3}$$

En la práctica se aplican muestreos de $N = 8$ bits por muestra para señales de audio de calidad telefónica y de $N = 16$ bits por muestra para señales de audio de alta calidad HiFi.

5.3. Codificación

El proceso de codificación de voz permite transmitir y almacenar la señal de voz eficientemente y sin pérdida de calidad en forma digital.

Desde el punto de vista de la transmisión de la señal de voz, la codificación de voz permite optimizar la utilización del canal de comunicación, transmitiendo el máximo de información, por ejemplo, transmitir varias comunicaciones por un solo canal, con la mínima pérdida de calidad optimizando la relación entre velocidad de transmisión (bits/segundo) e inteligibilidad del mensaje.

Desde el punto de vista de almacenar señal de voz en formato digital, la codificación de voz permite minimizar el número de bits necesarios para el almacenamiento, manteniendo un nivel de calidad adecuado.

Como valor añadido al proceso, la codificación digital de voz permite incorporar algoritmos de cifrado para establecer comunicaciones privadas seguras o realizar

grabaciones indescifrables para terceras personas.

El proceso de codificación propiamente dicho, toma como señal de entrada la obtenida de la conversión A/D a una velocidad de $N \times f_m$ ($bits/s$), utilizando ciertas propiedades de la señal de voz se obtiene una nueva codificación con una velocidad de R ($bits/s$) inferior a la inicial.

Los codificadores de voz trabajan explotando las propiedades tanto temporales como frecuenciales de la señal de voz y del sistema auditivo humano puesto que en último término va a ser el sistema auditivo humano quien va a juzgar la calidad de la señal. Así, la correlación de la señal es utilizada para reducir el margen dinámico de la señal y de este modo poderla cuantificar con un número menor de bits, como por ejemplo en la codificador ADPCM (Adaptive Differential Pulse Code Modulation / Modulación por Impulsos Codificados Diferencial Adaptativa) que permite reducir a 32 kb/s la velocidad de transmisión sin degradar la calidad de la señal.

Una parte importante de los esquemas de codificación de voz hacen uso extensivo del modelo digital de producción de voz, en el cual se linealiza el proceso de producción y se supone que la señal de voz es generada mediante la excitación de forma adecuada de un filtro que modela la envolvente del espectro de una señal de voz en segmentos cortos de entre 20 y 30 ms. Este modelo es la base de algunos sistemas de síntesis de señal de voz y del codificador conocido como vocoder LPC. Conjugando este modelo con el conocido efecto de enmascaramiento producido en el sistema auditivo humano de una señal de alta energía sobre otra de más baja, se puede modificar el espectro del ruido producido en la cuantificación y hacerlo perceptualmente inaudible con el consiguiente incremento en las prestaciones de los codificadores con velocidades de transmisión bajas.

El modelo digital de producción de la señal de voz ha dado lugar, según el tipo de excitación utilizado, a los codificadores CELP (Code Excited Linear Prediction), MPLPC (Multipulse Linear Predictive Coder) y RELP (Residual Excited Linear Prediction).

Desde el punto de vista de señales de audio HiFi (ancho de banda de 20 kHz), los requisitos de calidad para la señal codificada son mayores que para la codificación de voz a 4 kHz. Por ello, los esquemas de codificación son mucho más complejos y obedecen todos ellos a sistemas basados en el análisis frecuencial de la señal de audio, puesto que en este caso no es válido el modelo de producción de voz al ser una señal mucho más compleja. Para reducir la velocidad de transmisión se utiliza la propiedad de enmascaramiento de sonidos del sistema auditivo humano. La señal de audio analizada en 32 bandas

frecuenciales, se cuantifica con una asignación de bits que depende del grado de enmascaramiento de cada banda. Bajo esta idea básica se consiguen reducciones de 1.41 Mb/s necesarios para calidad Compact Disc estéreo a 128 kb/s sin pérdida significativa en la calidad. Un ejemplo ampliamente utilizado es el esquema de codificación tipo MPEG y su variante Capa 3 o MP3.

5.4. Ventajas de la señal digital

A continuación detallamos algunas de las ventajas de digitalizar la señal.

- Cuenta con sistemas de detección y corrección de errores, que se utilizan cuando la señal llega al receptor; entonces comprueban (uso de redundancia) la señal, primero para detectar algún error, y, algunos sistemas, pueden luego corregir alguno o todos los errores detectados previamente.
- La señal digital permite la multigeneración infinita sin pérdidas de calidad.
- Cuando una señal digital es atenuada o experimenta perturbaciones leves, puede ser reconstruida y amplificada mediante sistemas de regeneración de señales.
- Facilidad para el procesamiento de la señal. Cualquier operación es fácilmente realizable a través de cualquier software de edición o procesamiento de señal.
- Es posible aplicar técnicas de compresión de datos sin pérdidas o técnicas de compresión con pérdidas basados en la codificación perceptual mucho más eficientes que con señales analógicas.

5.5. Inconvenientes de la señal digital

Se necesita una conversión analógica-digital previa y una decodificación posterior, en el momento de la recepción. Si no se emplean un número suficientes de niveles de cuantificación en el proceso de digitalización, la relación señal a ruido resultante se reducirá con relación a la de la señal analógica original que se cuantificó. Esto es una consecuencia de que la señal conocida como error de cuantificación que introduce siempre el proceso de cuantificación sea más potente que la del ruido de la señal analógica original, en cuyo caso, además, se requiere la adición de un ruido conocido como "dither" más potente aún con objeto de asegurar que dicho error sea siempre un ruido blanco y no una distorsión. En los casos donde se emplean suficientes niveles de cuantificación, la relación señal a ruido de la señal original se conservará esencialmente porque el error de cuantificación quedará por debajo del nivel del ruido de la señal que se cuantificó. Esto, naturalmente, es lo normal.

Se hace necesario emplear siempre un filtro activo analógico pasa bajo sobre la señal a

muestrear con objeto de evitar el fenómeno conocido como "aliasing", que podría hacer que componentes de frecuencia fuera de la banda de interés quedaran registrados como componentes falsos de frecuencia dentro de la banda de interés. Asimismo, durante la reconstrucción de la señal en la posterior conversión D/A, se hace también necesario aplicar un filtro activo analógico del mismo tipo (pasa bajo) conocido como filtro de reconstrucción. Para que dicho filtro sea de fase lineal en la banda de interés, siempre se debe dejar un margen práctico desde la frecuencia de Nyquist (la mitad de la tasa de muestreo) y el límite de la banda de interés (por ejemplo, este margen en los CD es del 10%, ya que el límite de Nyquist es en este caso 44,1 kHz / 2 = 22,05 kHz y su banda de interés se limita a los 20 kHz).

CAPÍTULO II:
CODIFICACIÓN DE VOZ

El objetivo de la codificación de voz es obtener una representación de la señal con el menor número de bits posibles, pero que a la vez permita la reconstrucción de ésta con un cierto nivel de calidad de modo que se pueda realizar su transmisión o almacenamiento de una forma eficiente.

El ***codificador*** se encarga de realizar las operaciones correspondientes de ***análisis*** sobre la voz original, mientras que el ***decodificador*** realiza la ***síntesis*** cuando extrae la señal de voz a partir de la secuencia de bits recibidos. No obstante, ambos procesos no son totalmente inversos ya que parte de la información se pierde, ya sea por razones intrínsecas del propio proceso de codificación, o bien por las degradaciones introducidas por el canal. Esta pérdida de información introducida por el códec de voz se puede reducir mediante la selección de un esquema de codificación apropiado.

Normalmente, la voz se codifica mediante parámetros que representan eficientemente la señal, de modo que, mediante un reducido conjunto de bits, se puede reconstruir la voz con una pérdida mínima de calidad. Para realizar eficientemente esta función, los codificadores aprovechan las características intrínsecas de la voz y del sistema de percepción auditivo. Una vez llevada a cabo esta transformación, la señal de voz se convierte en una fuente de información, por lo que antes de ser enviada a través del canal, se le introduce una cierta redundancia mediante la codificación de canal, haciéndola así más inmune a las distorsiones introducidas por el medio de transmisión.

1. Estructura General de un codificador de Voz

A continuación se muestran los diagramas de bloques genéricos de un codificador y un decodificador de voz, para el codificador, la voz de entrada se procesa y se analiza con el fin de extraer un número de parámetros, los cuales se codifican y se cuantizan en índices binarios para ser enviados como un flujo comprimido de bits, los bits se colocan de acuerdo a cierto orden predeterminado y se transmite al decodificador.

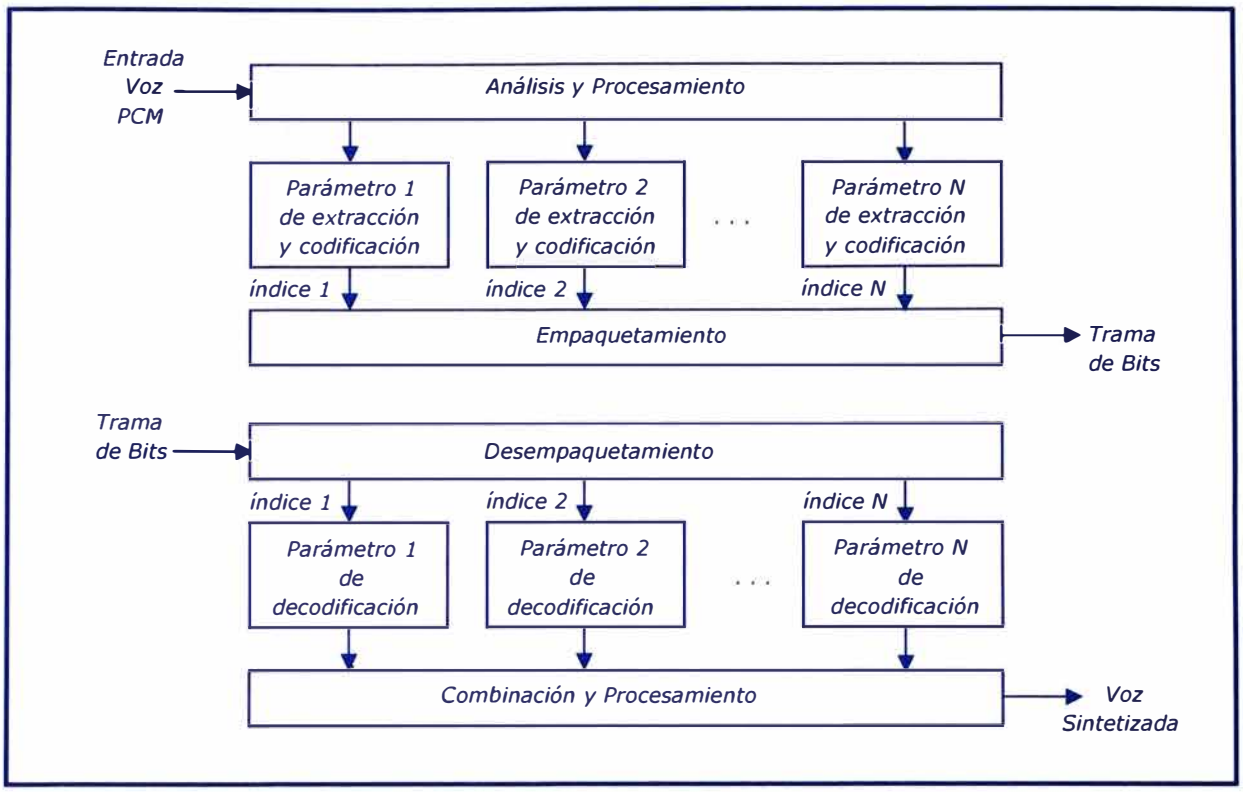

Fig. 2.1: Estructura general de un codificador de voz
Parte Superior (codificador), Parte Inferior (decodificador)

El decodificador desempaqueta el flujo de bits, donde los índices binarios recuperados se dirigen al parámetro decodificador correspondiente a fin de obtener los parámetros cuantificados. Estos parámetros decodificados se combinan y se procesan para generar la voz sintetizada.

Es la responsabilidad del diseñador del algoritmo el decidir la funcionalidad y características de diversos bloques de procesamiento, análisis y cuantización. Sus decisiones determinarán el rendimiento y la característica del codificador de voz.

2. Tipos de Codificadores

Históricamente, el ancho de banda utilizado para los codificadores de voz ha sido el del canal telefónico (300-3400 Hz), lo que da lugar a una primera clasificación en función del ancho de banda utilizado para la transmisión de voz. Los codificadores que hacen uso del ancho de banda del canal telefónico son denominados de banda estrecha, mientras que aquellos que hacen uso de una banda mayor reciben el nombre de codificadores de banda ancha. El hecho de utilizar una banda mayor dota a la señal reconstruida de una mayor naturalidad, aunque el uso de codificadores de voz de banda estrecha es suficiente para establecer la inteligibilidad del mensaje, así como las principales características del locutor. Los principales codificadores utilizados en las arquitecturas actuales de VoIP siguen haciendo uso de codificadores de banda estrecha, principalmente por la

compatibilidad con las redes de telefonía fija, por lo que orientaremos el desarrollo del presente informe a estos últimos, hasta llegar al G.729.

En las últimas décadas se han desarrollado numerosos esquemas de codificación de la señal de voz, divididos típicamente en los siguientes tres grupos:

- Codificadores de Forma de Onda.
- Codificadores Paramétricos o Vocoders.
- Codificadores Híbridos.

3. Codificadores de Forma de Onda

Presentan un modelo basado en el procesamiento muestra a muestra de la voz. El objetivo final de estos codificadores es conseguir reconstruir una señal que se aproxime lo máximo posible a la original, proporcionando por tanto una reproducción muy aproximada a la original, operan habitualmente en rangos de razón de transmisión superiores a 2 bits por muestra. Obteniendo para el caso de banda estrecha (utilizando una frecuencia de muestreo de 8 kHz), tasas superiores a los 16 kbps.

Los Codificadores de Forma de Onda se basan en el empleo de la codificación PCM (Pulse Code Modulation / Modulación por Impulsos Codificados) de las muestras de la señal de voz en el dominio temporal.

Un ejemplo es el estándar G.711 desarrollado por la CCITT (Comité Consultatif International Téléphonique et Télégraphique / Comité Consultivo Internacional Telegráfico y Telefónico), actual ITU (International Telecommunication Union / Unión Internacional de Telecomunicaciones). En este caso, se trata de un codificador de banda estrecha que adquiere la señal de voz mediante un proceso de muestreo a 8kHz. Este codificador aplica un cuantizador logarítmico (ley μ en EE.UU. y ley A en Europa) que asigna códigos binarios de 8 bits a cada muestra, consiguiendo un tasa final de 64 kbps.

Siendo la necesidad la de reducir la tasa de bits de estos codificadores, aparecen los codificadores DPCM (Differential Pulse Code Modulation / Modulación por Impulsos Codificados Diferencial) con distintos órdenes de predicción, cuyos coeficientes se determinan a partir de estimaciones de la función de auto-correlación de la voz. Para aplicar estas técnicas eficientemente, y debido que la señal de voz no es estacionaria, los coeficientes deben de ir modificándose a medida que cambien las funciones de auto-correlación a corto plazo. Estos esquemas de codificación reciben el nombre de técnicas APC (Adaptive Prediction Coding / Codificación con Predicción Adaptiva) y comprenden un gran número de variantes.

Si la información de predicción se transmite explícitamente al receptor DPCM da lugar a las denominadas técnicas APF (Adaptive Prediction Forward / Predicción Adaptiva hacia Adelante), mientras que si es derivada de la historia de la señal recientemente cuantizada, se denominan técnicas APB (Predicción Adaptiva hacia Atrás).

Además, si a la adaptación de los predictores se une la adaptación de los cuantizadores a la señal de entrada, se generan unas técnicas denominadas técnicas ADPCM (Adaptive Differential Pulse Code Modulation / Modulación por Impulsos Codificados Diferencial Adaptiva), ampliamente utilizadas y que constituyen la base de los estándares como el G.726.

Una técnica muy relacionada con la anterior es la codificación en sub-bandas, donde se aproxima la señal en el dominio de la frecuencia en lugar de en el dominio temporal. En esta técnica la banda de la señal de voz se divide normalmente en cuatro o más sub-bandas mediante un banco de filtros Pasa-Banda. La salida correspondiente a cada filtro es entonces diezmada, para tener una frecuencia de muestreo acorde con su ancho de banda, para finalmente ser codificada mediante un codificador PCM o DPCM. Asignando los bits de forma apropiada en las diferentes bandas se puede controlar el espectro del error de reconstrucción. Pudiendo de esta manera utilizar más bits en las bandas de frecuencia más baja para representar con mayor precisión la estructura de los formantes y el tono fundamental o pitch, y, en general, hacer adaptable dicha asignación de bits.

Un ejemplo es el estándar G.722, también desarrollado por la ITU.

4. Codificadores Paramétricos (Vocoders)

Presentan un modelo que pretende alcanzar una señal perceptualmente equivalente a la original, para esto se determinan ciertos parámetros que caracterizan a los distintos segmentos de voz y que constituyen la información enviada por el codificador. A partir de esos parámetros el decodificador genera una señal, con forma de onda distinta a la original, pero que el oído humano la percibe como similar.

La principal ventaja obtenida es la reducción drástica de la tasa a transmitir., y como contraparte, la voz reproducida pierde cierta naturalidad, operan a velocidades de transmisión menores que los codificadores de forma de onda, en el rango de los 0.5 bits por muestra e inferiores, pero la calidad de la señal de voz reproducida, aunque resulta inteligible, normalmente sufre de una pérdida de naturalidad y de las características propias del locutor.

Estos codificadores también son conocidos como vocoders, en comparación con los

codificadores de forma de onda, los vocoders no persiguen la reproducción de la señal de voz original, sino que su objetivo consiste en obtener una señal sintetizada que perceptualmente sea similar a la original.

Como parte de nuestro objetivo, el entender el comportamiento del códec G.729, ingresaremos un segmento importante de su historia y origen, revisaremos el vocoder LPC.

4.1. Vocoder LPC

Es el vocoder más ampliamente estudiado, basado en el modelo LPC (Linear Prediction Coding / Codificación por Predicción Lineal), el cual es un modelo muy simplificado para la producción de voz, como el que se muestra en la Fig. 2.2.

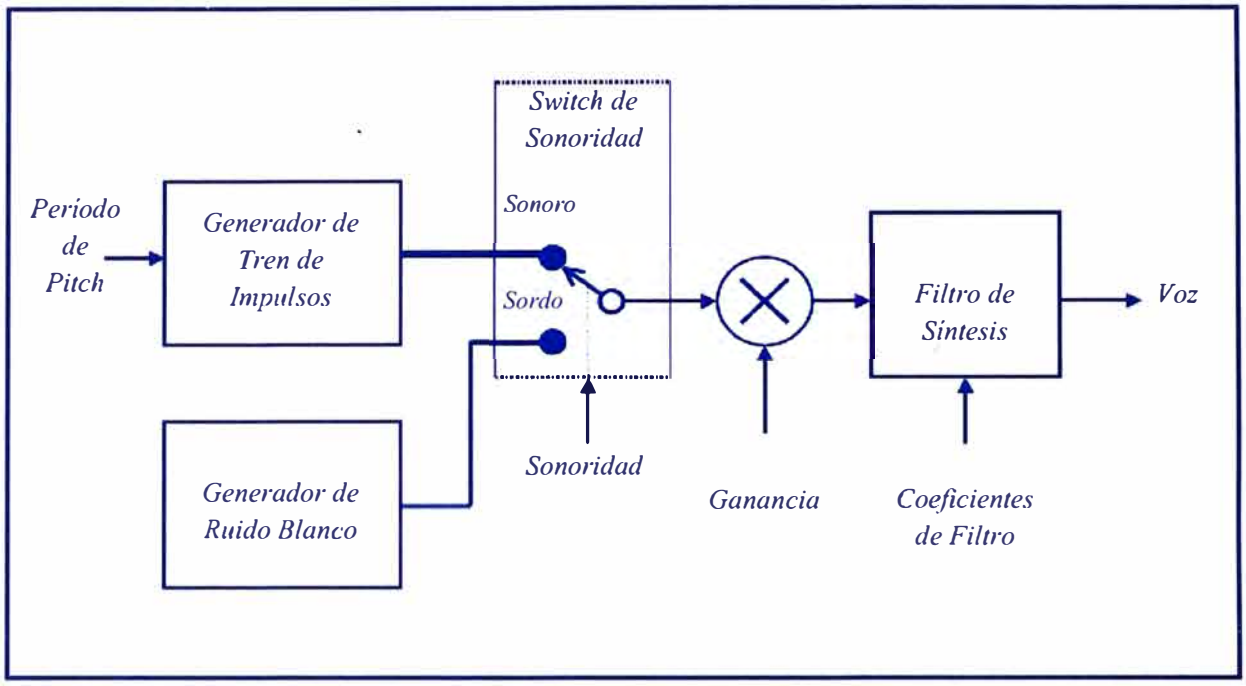

Fig. 2.2: Modelo LPC de Producción de Voz

Este modelo está inspirado en las observaciones de las propiedades básicas de las señales de la voz humana y representa un intento de imitación del mecanismo de producción de la misma. Las contribuciones espectrales combinadas del flujo glotal, el tracto vocal, y la vibración de los labios están representadas por el filtro de síntesis. La señal de entrada del filtro o señal de excitación se modela bien como un tren de impulsos (sonidos sonoros) o un ruido aleatorio (sonidos sordos). Por lo tanto, dependiendo del estado de sonoridad de la señal (sonoro o sordo), el interruptor se establece en la ubicación que selecciona la entrada apropiada. El nivel de energía de la salida es controlado por el parámetro de ganancia.

Para explicar el funcionamiento de este modelo, se consideran muestras separadas de voz en tramas que no se superponen. Para una longitud de trama suficientemente corta, las propiedades de la señal esencialmente permanecen constantes.

En cada trama se calculan los siguientes parámetros del modelo a partir de las muestras de voz:

- Sonorización: si la trama es sonora o sorda.
- Ganancia: principalmente relacionada con el nivel de energía del marco.
- Coeficientes del filtro: especifican la respuesta del filtro de síntesis.
- Periodo de Pitch: en el caso de los marcos sonoros, es la longitud de tiempo entre los impulsos de excitación consecutivos.

El proceso de estimación de parámetros se repite para cada cuadro, obteniendo los resultados que representan información de los mismos, de este modo, en lugar de transmitir las muestras PCM, se transmiten los parámetros del modelo, logrando una impresionante relación de compresión mediante una cuidadosa asignación de bits para cada parámetro a fin de minimizar la distorsión.

El precio a pagar es la pérdida irreversible de la calidad. Sin embargo, en muchas aplicaciones prácticas la alta calidad de reproducción no se requiere.

La estimación de los parámetros es la responsabilidad del codificador. El decodificador toma los parámetros estimados y utiliza el modelo de producción del habla para sintetizar la voz. Las formas de onda de salida utilizando el mismo conjunto de parámetros y filtros son diferentes a las de las condiciones iniciales ya que el generador de ruido blanco es aleatorio.

El enfoque de este modelo se basa en que la PSD (Power Spectral Density / Densidad espectral de Potencia) de la voz original sea capturada por el filtro de síntesis, por lo tanto de voz sintética está cerca de la original.

El método desecha toda la información de fase de la forma de onda original, conservando sólo la magnitud del espectro de frecuencias, por lo que para un oyente humano, la forma de onda sintética suena como la original.

a) Filtro LPC

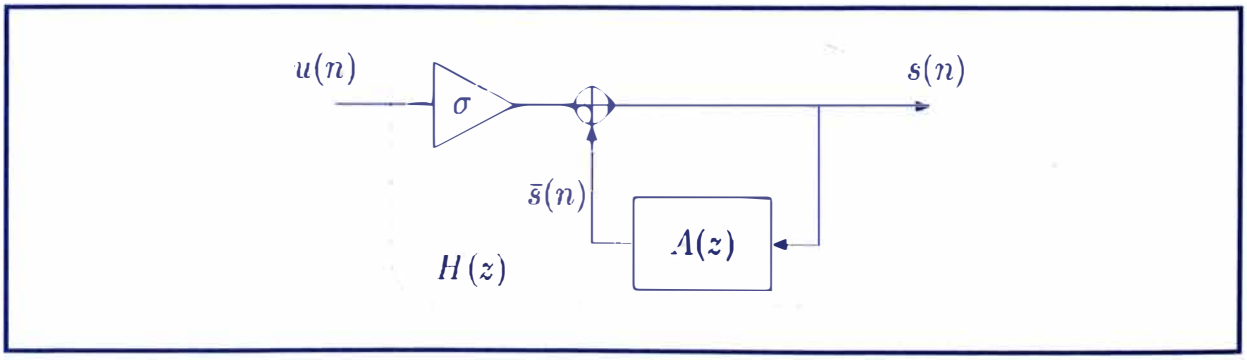

Fig. 2.3: Representación del tracto vocal a través de un filtro LPC

Se muestra la representación del tracto vocal a través de un filtro LPC, donde la señal de voz $s(n)$ se forma como la respuesta de un sistema lineal a una cierta entrada $u(n)$,

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + \sigma \sum_{l=0}^{q} b_l u(n-l), \qquad b_0 = 1 \tag{2.1}$$

donde $\boldsymbol{a_k}, 1 \leq k \leq p$; $\boldsymbol{b_l}, 1 \leq l \leq q$, y la ganancia $\boldsymbol{\sigma}$ son los parámetros del sistema.

La ecuación (2.1) indica que la salida $s(n)$ es una función lineal de las salidas anteriores y de las entradas anteriores y la actual. Es decir, la señal $s(n)$ se puede predecir como una combinación lineal de anteriores salidas y entradas, de aquí el nombre de predicción lineal.

También podemos expresar la ecuación (2.1) en el dominio de la frecuencia haciendo la transformada z a ambos lados de la igualdad. Si $H(z)$ es la función de transferencia del sistema, entonces obtenemos la siguiente expresión,

$$H(z) = \frac{S(z)}{U(z)} = \sigma \frac{1 - \sum_{l=1}^{q} b_l z^{-l}}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2.2}$$

donde $S(z)$ y $U(z)$ son la transformada z de $s(n)$ y $u(n)$ respectivamente.

H(z) representa el modelo polos-ceros más general. Las raíces de los polinomios del numerador y del denominador son los ceros y los polos del modelo, respectivamente.

- El modelo todo-ceros o de media móvil (MA, moving average)
- $a_k = 0, \quad 1 \leq k \leq p$
- El modelo todo-polos o auto-regresivo (AR, auto-regressive)
- $b_l = 0, \quad 1 \leq l \leq q$

El modelo más utilizado para representar la señal de voz de forma paramétrica es el modelo todo-polos, donde la señal $s(n)$ viene dada como una combinación lineal de los valores anteriores y de una entrada $u(n)$,

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + \sigma u(n) \tag{2.3}$$

donde σ es un factor de ganancia, a_k los denominados coeficientes de predicción y p es el orden de predicción. Obteniendo una función de transferencia $H(z)$ como sigue:

$$H(z) = \frac{\sigma}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{\sigma}{1 - A(z)} \tag{2.4}$$

Los coeficientes del filtro de predicción lineal $A(z)$ se obtienen llevando a cabo un proceso de minimización del error de predicción sobre un conjunto de N muestras.

Se considera que el filtro lineal permanece constante durante cierto intervalo de tiempo,

que suele durar de 10 a 20 ms, modelando el tracto vocal como una sucesión de filtros lineales estacionarios ya que la velocidad de articulación del tracto vocal no es muy superior a 10 veces por segundo. Expresando la predicción realizada como,

$$\bar{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{2.5}$$

Por tanto, el error entre el valor actual de $s(n)$ y el predicho $\bar{s}(n)$ es,

$$e(n) = s(n) - \bar{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{2.6}$$

a partir del cual podemos obtener el error cuadrático E definido como,

$$E = \sum_{n=0}^{N-1} e^2(n) \tag{2.7}$$

Los coeficientes LPC se calculan minimizando este error cuadrático, es decir,

$$\frac{\partial E}{\partial a_k} = 0 \qquad k = 1, \dots, p \tag{2.8}$$

expresión a partir de la cual obtenemos el sistema de ecuaciones normales,

$$\sum_{k=1}^{p} a_k R_s(|i-k|) = R_s(i) \qquad 1 \leq i \leq p \tag{2.9}$$

donde R_s representa la función de auto-correlación de la señal $s(n)$.

Como se observa en la expresión (2.9), para el cálculo de los distintos coeficientes a_k se precisa de una descripción matemática de cada segmento de $s(n)$ que no se encuentra disponible.

Por este motivo, es necesario estimar los valores de la matriz de auto-correlación $R(|i-k|)$ a partir de las observaciones consecutivas de $s(n)$ con $0 \leq n \leq N-1$.

El método para la resolución del modelo todo-polos descrito anteriormente se denomina método de auto-correlación. Este método considera que se dispone de un conjunto infinito de muestras para determinar la función de auto-correlación R_s y, por consiguiente, dicha función solo depende de la diferencia $|i-k|$, obteniéndose así una matriz Toeplitz.

Basándose en dicha característica se han desarrollado métodos muy eficaces de resolución de dicho sistema de ecuaciones, como por ejemplo, el algoritmo recursivo de Levinson y Durbin. Sin embargo, en la práctica no se dispone de infinitas muestras de un segmento, por lo que es necesario multiplicar la señal $s(n)$ por una función de ventana w_n para obtener otra señal $s_w(n)$ que es cero fuera del intervalo $0 \leq n \leq N-1$, siendo N el

tamaño del segmento de análisis. Este enventanado hace del método de autocorrelación un método aproximado.

Una alternativa más precisa la proporciona el método de covarianza. En este método se calcula el error E en un intervalo finito de tamaño igual al del segmento de análisis, $0 \leq n \leq N-1$, obteniendo la siguiente expresión para los elementos de la matriz de covarianza,

$$R_s(i,k) = \sum_{n=0}^{N-1} s(n-k)\, s(n-i) \tag{2.10}$$

Hay que señalar que según se desprende de esta última expresión, en este método se ha de conocer la señal $s(n)$ en el intervalo $-p \leq n \leq N-1$.

En este caso la matriz del sistema de ecuaciones ya no es una matriz Toeplitz aunque sí simétrica. Para resolver dicho sistema de ecuaciones se utiliza la denominada descomposición de Cholesky.

El método de auto-correlación tiene la ventaja, frente al de covarianza, de asegurar la estabilidad del filtro de síntesis usado durante la reconstrucción de la voz. Además, el procedimiento de covarianza requiere un mayor esfuerzo computacional. Por estas razones, el método de covarianza es menos utilizado que el método de auto-correlación.

b) Excitación del Modelo LPC

Como ya hemos visto, la excitación en el modelo LPC se modela por medio de la señal $u(n)$. Como se mencionó líneas atrás, en caso de que el segmento de voz sea sonoro, la señal de excitación adopta la forma de un tren de pulsos. Por contraparte, cuando se trata de un segmento sordo, la excitación se hace corresponder con ruido blanco. La excitación en el decodificador del modelo LPC es generada sólo a partir de una especificación relativamente simple del carácter general del segmento de voz actual, sin enviar información que especifique realmente la forma de la excitación.

La única información que se envía consiste en los parámetros del filtro (coeficientes LPC y ganancia), información de si el segmento es sordo o sonoro y una estimación del valor del periodo de pitch en el caso de que el segmento sea sonoro.

c) Cuantización de Parámetros LPC

Existen numerosas técnicas para cuantizar los parámetros LPC, estos cuantizadores pueden ser uniformes o no uniformes, siendo los no uniformes los más utilizados por presentar una menor distorsión, no solo se busca que al cuantizar los parámetros LPC se introduzca la menor distorsión posible, también es necesario que el filtro todo-polos se mantenga estable

tras la cuantización. Normalmente, no se utiliza la cuantización directa de los coeficientes LPC ya que pequeños errores pueden producir errores espectrales relativamente grandes, además de inestabilizar los filtros todo-polos $H(z)$. Por este motivo, se utilizan otras representaciones de los coeficientes LPC que aseguran la estabilidad.

Dentro de estas representaciones destacan los RC (Reflections Coefficients), los ASRC (ArcSine Reflection Coefficients), los coeficientes LAR (Log Area Ratio), y los LSF (Line Spectral Frecuency), siendo estos dos últimos los más extendidos por presentar las mejores características de cuantización. Aunque estos codificadores reducen sustancialmente la tasa de bits, en la actualidad son poco utilizados debido a su baja calidad, la señal sintetizada suena artificial o poco natural y la identidad del locutor resulta difícil de reconocer, la calidad tiende a degradarse todavía más si la señal de voz original contiene ruido de fondo. Un ejemplo es el estándar del gobierno estadounidense FS1015, que muestrea la señal de voz a 8kHz sobre tramas de 22.5ms, obteniendo una tasa final de 2.4kbps.

4.2. Codificadores Sinusoidales

Los codificadores sinusoidales constituyen un tipo de vocoder, en el que se toma la voz como el resultado de pasar una excitación generada por la glotis a través de un filtro lineal variante en el tiempo (caracterizado por la secuencia de coeficientes LPC) que modela las características resonantes del tracto vocal.

A diferencia del carácter rígido del vocoder LPC, los codificadores sinusoidales generan la excitación en el decodificador generando una suma de sinusoides, cuyas frecuencias y fases son modificadas en segmentos sucesivos para representar el carácter cambiante del espectro de la señal de voz original.

El enfoque de esta representación viene dado por la asunción de que la excitación es periódica, pudiendo de esta manera representarse por una descomposición en series de Fourier en la que cada componente armónica corresponda a una única onda sinusoidal. Donde la calidad de la voz sintetizada dependerá de la codificación de las fases y amplitudes de las ondas sinusoidales utilizadas.

Como referencia tenemos al IMBE (Improved Multi-Band Excitation), una de las aplicaciones más exitosas de esta técnica de modelado para codificación a baja tasa, utilizado para el sistema INMARSAT-M (1990) y para el INMARSAT-Mini-M (1994).

5. Codificadores Híbridos

Los codificadores híbridos surgen como una mezcla de los codificadores de forma de onda y los codificadores paramétricos. Éstos pueden considerarse codificadores paramétricos

atendiendo a que emplean un modelo paramétrico, aunque también intentan preservar la forma de onda de la señal de voz. Actualmente la mayoría de los codificadores utilizados en los sistemas de transmisión digitales de voz utilizan distintas variantes enmarcadas en esta tipología.

En el caso de que consideremos un modelo LPC, podemos plantear que el objetivo que persigue un codificador paramétrico sería conseguir que la señal de excitación, $\hat{e}(n) = \sigma\hat{u}(n)$, haga que la síntesis se aproxime lo máximo posible a la señal de voz original. El principal problema que surge es cómo determinar los parámetros óptimos de la nueva excitación $\hat{e}(n)$ una vez determinado el tipo de modelo que se utilizará. Normalmente, esta cuestión se resuelve mediante un procedimiento denominado análisis por síntesis, del que se muestra un diagrama básico en la Fig. 2.5.

5.1. El Análisis por Síntesis

El Análisis por Síntesis consiste en obtener aquellos parámetros que logran ajustar mejor la señal sintetizada $\hat{s}(n)$ a la señal original $s(n)$. Normalmente, esta tarea se realiza aplicando un criterio LSE (Least Square Error/Menor Error Cuadrático), es decir, minimizando el error cuadrático definido como,

$$E = \sum_{n=0}^{N-1} (s(n) - \hat{s}(n))^2 \tag{2.11}$$

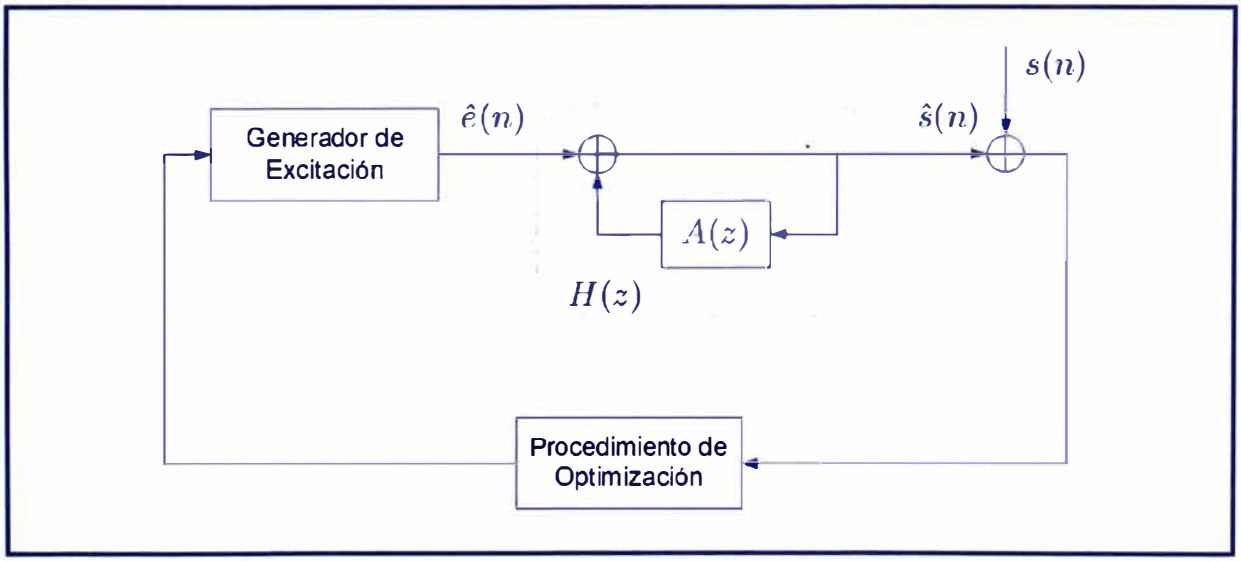

Fig. 2.4: Esquema de selección de parámetros de un codificador híbrido mediante el proceso de analisis por síntesis.

Cada muestra de la señal de excitación $\hat{e}(n)$ afecta a muchas muestras de la señal de voz reconstruida $\hat{s}(n)$, debido a la estructura recursiva del filtro de síntesis expresada en la ecuación (2.3). Por consiguiente, la elección de una señal de excitación se realiza midiendo

su efecto sobre la señal de voz reconstruida durante más de una muestra. Es decir, la decisión sobre la mejor representación cuantizada no se realiza de forma instantánea sino que se tarda un intervalo que incluye varias muestras. Este tipo de aproximación se denomina codificación de decisión retardada. Como dicha decisión depende del error entre la señal original y la sintetizada, estas técnicas requieren síntesis durante el análisis, y el procedimiento se conoce como Codificación Adaptable por Predicción y Análisis por Síntesis.

La característica principal de un esquema de codificación basado en análisis por síntesis es que el codificador incluye al propio decodificador. De esta forma, el codificador puede conocer en cada momento la señal decodificada $\hat{s}(n)$. Así, mediante un proceso iterativo representado por el bucle de la Fig. 2.4, se lleva a cabo el proceso de optimización. Finalmente, el codificador envía la información referente al tracto vocal (coeficientes del filtro LPC) junto a los parámetros que definen la excitación.

Este tipo de codificadores obtienen una buena calidad de la señal de voz en el rango de razones de transmisión que va de 0.5 a 2 bits/muestra, lográndose así tasas de bits entre 4 y 16kbps. Debido a las buenas relaciones entre calidad perceptual y las moderadas tasas de bit conseguidas, la mayoría de los codificadores actualmente utilizados en telefonía móvil y VoIP hacen uso de esta tipología.

5.2. Codificadores Multipulso

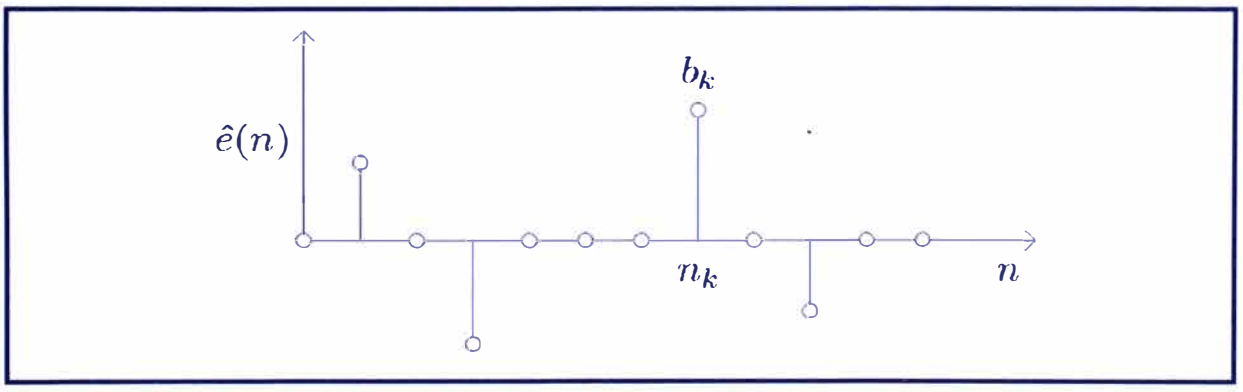

Fig. 2.5: Ejemplo de excitación multipulso.

Este esquema construye la señal de excitación $\hat{e}(n)$ mediante una serie de L pulsos dados por unas ciertas amplitudes b_l y posiciones n_l $(l = 0, \ldots, L-1)$, de modo que podemos expresar la excitación como,

$$\hat{e}(n) = \sum_{l=0}^{L-1} b_l \delta(n - n_l) \qquad (2.12)$$

donde $\delta(n)$ define la función impulso unidad.

La Fig. 2.5 muestra un ejemplo de este tipo de excitación en el que podemos ver que la señal de excitación está constituida por ceros, excepto en algunas posiciones donde se ubican los pulsos.

Una vez que los parámetros LPC del modelo son seleccionados es necesario determinar los parámetros de la excitación, es decir, las posiciones y amplitudes de los pulsos. Atendiendo al criterio LSE, podemos decir que los parámetros de la excitación han de ser aquellos que minimicen la expresión,

$$E = \sum_{n=0}^{N-1} (s(n) - h(n) * \hat{e}(n))^2 \tag{2.13}$$

respecto a las amplitudes b_k y las posiciones n_k $(k = 0, \dots, L-1)$.

Así mismo, h(n) corresponde con la respuesta impulsiva del filtro LPC utilizado, que normalmente suele corresponderse con un filtro todo-polos como el expresado en la ecuación (2.4)

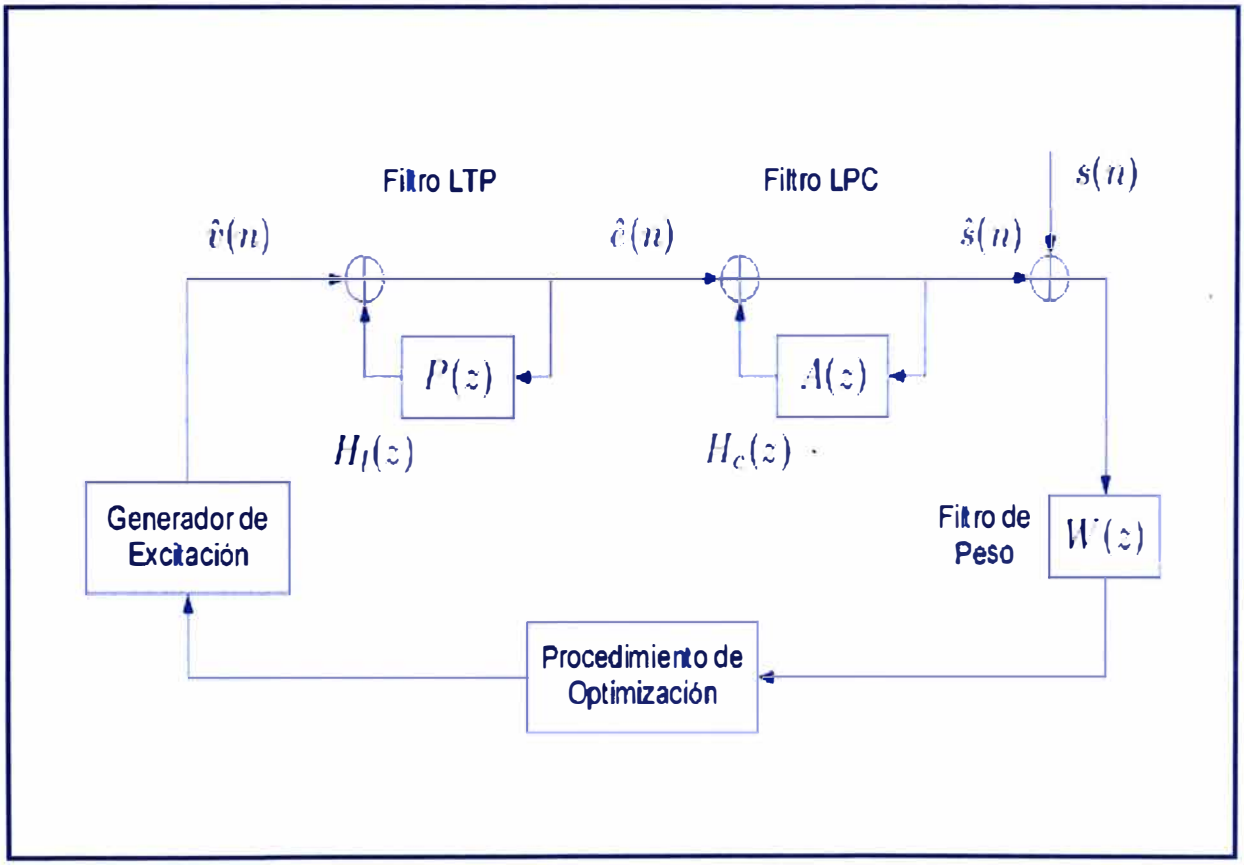

Fig. 2.6: Esquema de codificación mejorado basado en el procedimiento de análisis por síntesis.

El diagrama presentado en la Fig. 2.6 muestra un esquema de codificación de análisis por síntesis mejorado, donde se han introducido, respecto al diagrama presentado en la Fig. 2.4, dos nuevos bloques: el ***filtro de pitch*** y el ***filtro de peso***.

a) Filtro de Pitch o LTP

El filtro de pitch introduce un nuevo termino de filtrado $H_l(z)$ sobre la respuesta del filtro LPC $H_c(z)$, resultando en una respuesta total $H(z) = H_l(z)H_c(z)$.

Como vimos, el filtro LPC consiste en un predictor a corto plazo (STP, Short-Term Predictor) que modela la envolvente del espectro, del mismo modo, el filtro de pitch está constituido por un predictor, pero que en este caso, se encarga de modelar las correlaciones a largo plazo debidas al pitch, con esto, el predictor a largo plazo (LTP, Long-Term Predictor) intenta eliminar las redundancias a largo plazo del residuo o error de predicción $e(n)$, de forma que el nuevo residuo $v(n)$ pueda ser correctamente representado por una aproximación $\hat{v}(n)$ con menos pulsos.

La forma general de un filtro LTP es:

$$P(z) = 1 - \sum_{k=-(q-1)/2}^{(q-1)/2} b_k z^{-(M+k)} \tag{2.14}$$

donde M es el retardo en muestras y b_k son los coeficientes de predicción de retardo largo, mientras que q determina el orden del filtro

El valor de M, cuyo rango suele ser de 2 a 20ms, coincide con el periodo de pitch en los segmentos sonoros, mientras que es un valor aleatorio en los segmentos sordos.

El retardo M y los coeficientes b_k se determinan a partir de la señal de voz o a partir de la señal diferencia después de eliminar las correlaciones de retardo corto.

El procedimiento para determinar el valor de los coeficientes b_k para un M dado, es similar que para el caso de los coeficientes de predicción de retardo corto a_k, pudiendo determinarse mediante el método de auto-correlación o covarianza. Sin embargo, es necesario determinar por adelantado el valor del retardo M.

Típicamente se utilizan de uno a tres coeficientes de predicción y sus valores se adaptan con el tiempo a un ritmo que va de 50 a 200 veces por segundo. Si se aumenta el orden del predictor se obtiene una mejor predicción, pero a cambio se necesitan más bits para codificar los coeficientes adicionales.

Para segmentos periódicos, la razón principal de disponer de múltiples coeficientes es suministrar interpolación entre las muestras en caso de que el periodo no corresponda con un número entero de muestras. En lugar de usar un predictor de orden alto se puede también usar un predictor de primer orden con un retardo no entero, lo cual permite una cuantización más eficiente de los parámetros del predictor.

b) Filtro de Peso

Como revisamos en el capítulo anterior, el oído presenta una propiedad denominada enmascaramiento, que consiste en que el sistema auditivo tiene poca capacidad de detectar ruido en las bandas de frecuencia en las cuales la señal de voz tiene una alta energía, tal y como ocurre en las bandas donde aparecen los formantes de la voz.

Cuando se reduce el número de bits para la codificación de la excitación se aumenta la energía del ruido, lo que hace que el ruido sea más audible. Sin embargo, teniendo en cuenta la propiedad de enmascaramiento del oído es posible reducir este efecto. Para ello se realiza el filtrado de la señal excitación mediante un filtro $W(z)$ antes de ser codificada, de forma que se reduzca el ruido de cuantización en los valles y se aumente en los picos del espectro.

Así, al aplicar el proceso de minimización del error cuadrático medio a través del filtro $W(z)$ se consigue aumentar la calidad de la voz sintetizada, a pesar de que la SNR sea menor.

Dicho procedimiento no afecta ni a la razón de transmisión ni a la complejidad del procedimiento de síntesis. Aumenta sólo la complejidad del codificador.

El uso de un procedimiento de enmascaramiento del ruido fue inicialmente investigado para los esquemas de codificación adaptable predictiva y se obtuvo que un filtro de peso apropiado que responde a la siguiente forma,

$$W(z) = \frac{F(z/\gamma_1)}{F(z/\gamma_2)} = \frac{1 - \sum_{k=1}^{p} f_k \gamma_1^k z^{-k}}{1 - \sum_{k=1}^{p} f_k \gamma_2^k z^{-k}} \qquad 0 \leq \gamma_2 \leq \gamma_1 \leq 1 \qquad (2.15)$$

donde $F(z)$ es un predictor de retardo corto.

Para la mayor parte de las aplicaciones se puede seleccionar $F(z) = A(z)$.

Los parámetros γ_1 y γ_2 tienen un valor entre 0 y 1 que controla la energía del error en las regiones formantes.

Nótese que al decrementar γ aumenta el ancho de banda de los ceros de $F(z/\gamma)$. Los valores de γ_1 y γ_2 se determinan heurísticamente mediante los correspondientes pruebas auditivas.

Cuando se utilizan tasas de transmisión pequeñas el efecto de enmascarar el error es menos perceptible. Una razón que explica este efecto es que el nivel de ruido es tan alto que, a pesar del enmascaramiento, el ruido sigue siendo audible. Además, la suposición de que el ruido de cuantización tiene un espectro plano ya no es válida, lo que hace que el resultado del proceso de enmascaramiento sea menos predecible.

5.3. Codificadores CELP

De todas las técnicas hibridas, la más extendida es la denominada codificación CELP (Code Excited Linear Prediction / Predicción Lineal por Código Exitado), la cual se deriva de la codificación multipulso.

Veíamos que el esquema de análisis por síntesis mejorado (Fig. 2.6) tiene un doble objetivo, en primer lugar intenta obtener una señal de excitación $\hat{v}(n)$ con el menor número de redundancias posibles, por lo que se introducen dos etapas de predicción que consisten en un predictor a corto plazo (filtro LPC) y otro a largo plazo (filtro LTP), de este modo, en caso de que la excitación $\hat{v}(n)$ fuera igual a la señal residual $v(n)$, se reproduciría la señal de voz original $s(n)$, el segundo objetivo es la búsqueda de una excitación que represente el residuo de la forma más precisa posible, en este sentido, la innovación que introducen los codificadores CELP frente a los multipulso consiste en incorporar un cierto diccionario de M códigos fijos de excitación $c_k(n)$ $(k = 0, \dots, M-1)$. La formulación básica de la codificación CELP supone que las M entradas del diccionario están formadas por secuencias aleatorias Gausianas de varianza unitaria.

De este modo, el esquema final de un codificador CELP se puede representar mediante el diagrama de la Fig. 2.7.

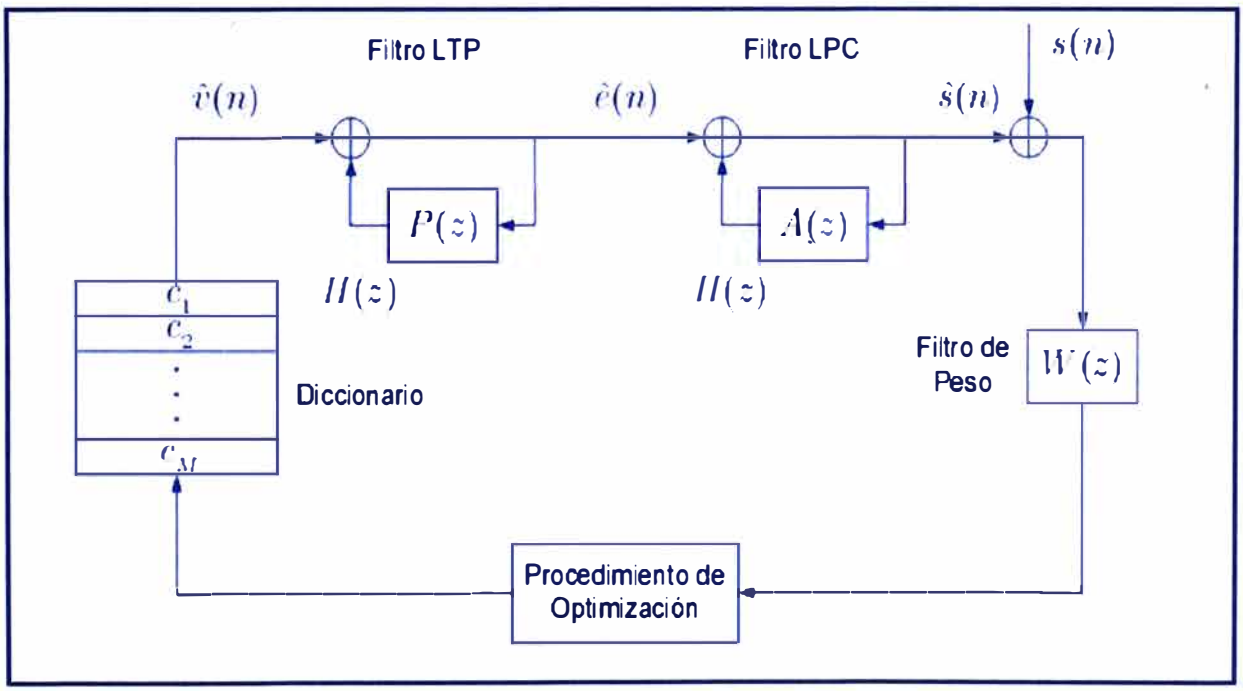

Fig. 2.7: Esquema general de un codificador CELP.

De este esquema, la excitación se construye como $\hat{v}(n) = g_c c(n)$, donde la ganancia g_c se obtiene durante el proceso de optimización junto a $c(n)$ (código óptimo de los M posibles del diccionario).

En este caso, tanto el codificador como el decodificador almacenan un diccionario de M secuencias posibles para la excitación, donde cada secuencia tiene una longitud L. La excitación en cada segmento está completamente descrita por el índice del vector apropiado dentro del diccionario. Este índice se determina mediante una búsqueda exhaustiva sobre todos los vectores del diccionario y corresponde al índice de aquel vector del diccionario que produce el error más pequeño entre la señal original y la reconstruida.

La transmisión del índice requiere $(log_2 M)/L$ bits/muestra, con valores típicos en el rango de 0.5 a 2 bits/muestra. Considerando una frecuencia de muestreo de 8 kHz, las tasas totales de transmisión obtenidas se sitúan entre 4 y 16 kbit/s.

La técnica CELP ha derivado en un amplio número de variantes, como por ejemplo los esquemas VSELP (Vector Sum Excited Linear Prediction), adoptado por los estándares IS-54/136 y GSM-HR (GSM Half-Rate); QCELP (Qualcomm CELP), adoptado por el estándar IS-95 de telefonía móvil, y el ACELP (Algebraic CELP), siendo esta última variante la que ha tenido más éxito durante los últimos años.

a) La Técnica ACELP

La técnica ACELP (Algebraic Code Excited Linear Prediction / Predicción Lineal por Excitación Algebraica de Código) se basa en la selección de códigos ligeros, formados principalmente por ceros, lo que permite una rápida búsqueda en el diccionario. Los valores distintos de cero se hallan permutando una serie de pulsos en una serie de posiciones prefijadas. De este modo, los codificadores ACELP recurren a una solución similar a la codificación multipulso, en la que se introducen ciertas restricciones en la amplitud y ubicación de los pulsos, de ahí que la búsqueda por medio de diccionario de los codificadores CELP se mantenga.

Éste fue el esquema de codificación más utilizado durante la década de los 90, destacando los codificadores EFR (Enhanced Full Rate / Tasa Completa Mejorada) con una tasa de 12.2 kbps y AMR (Adpative Multi-Rate / Multi Tasa Adaptiva) el cual dispone de 8 tasas comprendidas entre 4.75 y 12.2 kbps.

5.4. Codificadores Híbridos en Aplicaciones VoIP

Mientras que los codificadores anteriormente mencionados fueron principalmente desarrollados para su uso en redes de telefonía celular, también existen otros estándares de codificación, tales como G.723.1 y G.729 comúnmente utilizados para aplicaciones VoIP.

En particular, el codificador G.723.1 es un códec con dos posibles tasas, 5.3 kbps y 6.3 kbps, en el que la representación de la excitación depende de la tasa escogida. Así, el modo

de 5.3 kbps lleva a cabo una codificación CELP, mientras que la tasa de 6.3 kbps establece una representación de excitación multipulso. La longitud de la trama es de 30 ms, utilizando un pre-análisis de 7.5 ms.

El estándar G.729 emplea un esquema de codificación CS-ACELP (Conjugate Structure ACELP / Predicción Lineal con Excitación por Código Algebraico de Estructura Conjugada) con una tasa de 8 kbps y una longitud de trama de 10 ms. En este caso el pre-análisis es de 5 ms, consiguiendo una importante reducción del retardo algorítmico sobre el estándar G.723.1. Los estándares de G.723.1 y G.729 incluyen sus correspondientes algoritmos para la mitigación de pérdidas de tramas.

Aunque, como acabamos de ver, los codificadores CELP son utilizados en redes de paquetes, éstos fueron originalmente desarrollados para ser utilizados en comunicaciones móviles, en las que es posible adaptar la codificación en función de las degradaciones presentes en el medio de transmisión.

Así, por ejemplo, el estándar de codificación AMR permite la selección de distintas tasas de bits para el codificador de voz de modo que, si las características del canal empeoran, puede reducir la tasa del codificador para aumentar la tasa de protección frente a errores. Sin embargo, este esquema no puede ser abordado en una red como Internet, en la que la información se encuentra fragmentada en paquetes individuales que pueden seguir caminos distintos. En este caso, la unidad de menor información codificada es la correspondiente a un paquete, el cual puede ser recibido o no en función del estado del canal. Cuando se produce la pérdida de un paquete implícitamente se pierde una porción de información que es igual para todos los codificadores. Pero aquí no acaban los efectos negativos, ya que tal y como se ha estudiado en los codificadores anteriormente vistos, algunos de ellos presentan fuertes dependencias temporales, como observamos en el filtro de pitch o LTP, que provocan la expansión del error a las tramas consecutivas a la pérdida.

Con el objetivo de eliminar las dependencias inter-trama aparece el esquema de codificación iLBC (internet Low Bit-Rate Codec), reduciendo así los efectos negativos de la pérdida de una trama. A pesar de su reciente creación en el año 2004, este códec ha tenido un éxito bastante relevante y se ha impuesto en algunas aplicaciones VoIP tan conocidas como Skype o Google Talk.

iLBC se encuadra dentro de los codificadores híbridos basados en el modelo LPC, aunque en este caso no responde a una estructura CELP de análisis por síntesis. Puesto que pretende evitar la dependencia inter-trama, rechaza el uso del filtro de pitch para replicar

esta periodicidad a largo plazo. Por el contrario, detecta la zona de mayor energía en la excitación de la trama actual, la cual es llamada estado inicial (donde normalmente se encuentra el pulso de pitch), para llevar a cabo sobre ella una codificación de forma de onda DPCM. En este punto, parecería razonable pensar que el codificador responde a un esquema de forma de onda. Sin embargo, el resto de la excitación no es codificada con DPCM, sino que se realiza a partir de las muestras del estado inicial, el cual es utilizado mediante un proceso de tres pasadas para aproximar el resto de la excitación. Este diccionario adaptativo se utiliza para codificar el resto de muestras de la excitación, empleando para ello predictores a largo plazo hacia delante y hacia atrás (pero siempre dentro de la trama actual). De este modo, la decodificación de las muestras correspondientes a una trama no depende de las muestras previas. No obstante, la robustez frente a pérdidas de paquetes presenta el inconveniente de una mayor tasa de bits. iLBC presenta dos modos de funcionamiento denominados en función de su longitud de bloque o trama, siendo éstas de 20 o 30 ms. En el primero de los casos la tasa de bits obtenida es 15.2 kbps, mientras que en el segundo caso es 13.3 kbps.

El diseño de iLBC está directamente orientado a las características de las redes IP, teniendo como objetivo incrementar la robustez ante la pérdida de paquetes. Ésta característica lo distingue del resto de codificadores actualmente en uso y lo convierte en un caso de especial interés para el desarrollo de la presente tesis. Los detalles de implementación de este codificador se pueden encontrar en el documento RFC 3951.

CAPÍTULO III
CÓDEC G.729

1. Introducción

La Codificación de la voz a 8 kbit/s Mediante Predicción Lineal con Excitación por Código Algebraico de Estructura Conjugada (CS-ACELP), es un estándar desarrollado por la ITU aprobado en 1996, con la denominación "Recomendación G.729", y es de especial interés pues se encuentra junto con otros estándares de la ITU-T (G.711, G.723, G.728...), amparado por la recomendación H.323, la cual define los protocolos necesarios para establecer sesiones de comunicación audio-visual sobre redes de conmutación de paquetes. Concretamente, G.729 presenta una tasa de codificación de 8 kbps, aunque posteriormente se han añadido anexos que ofrecen bajo la misma arquitectura tasas de 6.4 y 11.8 kbps, y según la recomendación H.323, se adecúa perfectamente a aquellos tipos de aplicaciones de audio con un bajo requerimiento de ancho de banda.

Los codificadores desarrollados por la ITU son adaptables a la carga útil de la trama RTP mediante el RFC 3551, el cual define el número mínimo de tramas por paquete para cada codificador. Para G.729, el empaquetado mínimo es de 2 tramas por paquete.

Tabla 3.1: Resumen de las características del códec G.729

	Tasa (kbps)	Trama (ms)	Pre-Análisis (ms)	Empaquetado (tramas/paq)	Retardo Alg. (ms)
G.729	8	10	5	2	15

El codificador G.729 fue desarrollado originalmente para sistemas de comunicaciones personales, sistemas digitales por satélite y otras aplicaciones tales como la paquetización de la voz en equipos de circuito multiplexado.

A continuación revisaremos las características principales del codificador G.729 como la estructura de la tabla de codificación algebraica, seguido por la metodología de construcción y búsqueda del libro de códigos adaptivo. Descripción de las operaciones de codificación y decodificación, y las técnicas de búsqueda de libro de códigos algebraico.

En principio, el codificador CS-ACELP es una variante de la familia de los codificadores CELP y siendo el grupo de investigadores que realizó el desarrollo del codificador G.729 es el mismo que realizó desarrollos de otros estándares ACELP, por lo tanto comparten muchos atributos comunes. El concepto de ACELP tiene una gran influencia sobre la dirección de la evolución de codificación de voz, por lo que es necesario comprender su estructura.

2. Retardo

El códec G.729 codifica la voz y otras señales de audio con tramas de 10 ms. Se produce además un pre-análisis de 5 ms, por lo que el retardo algorítmico total es de 15 ms. Los demás retardos producidos por la aplicación práctica de este códec tienen por causa:

- El tiempo de procesamiento necesario para las operaciones de codificación y decodificación;
- El tiempo de transmisión en el enlace de comunicación;
- El retardo de multiplexación por la combinación de datos de señales de voz y otros.

3. Estructura del Libro de Códigos Algebraico

Como observamos en el capítulo anterior, un codificador CELP contiene un libro de códigos de excitación fija que contiene vectores de códigos que sirven como entrada al filtro de síntesis. El libro de códigos es consultado durante la codificación para localizar el mejor vector de código de una trama particular de voz.

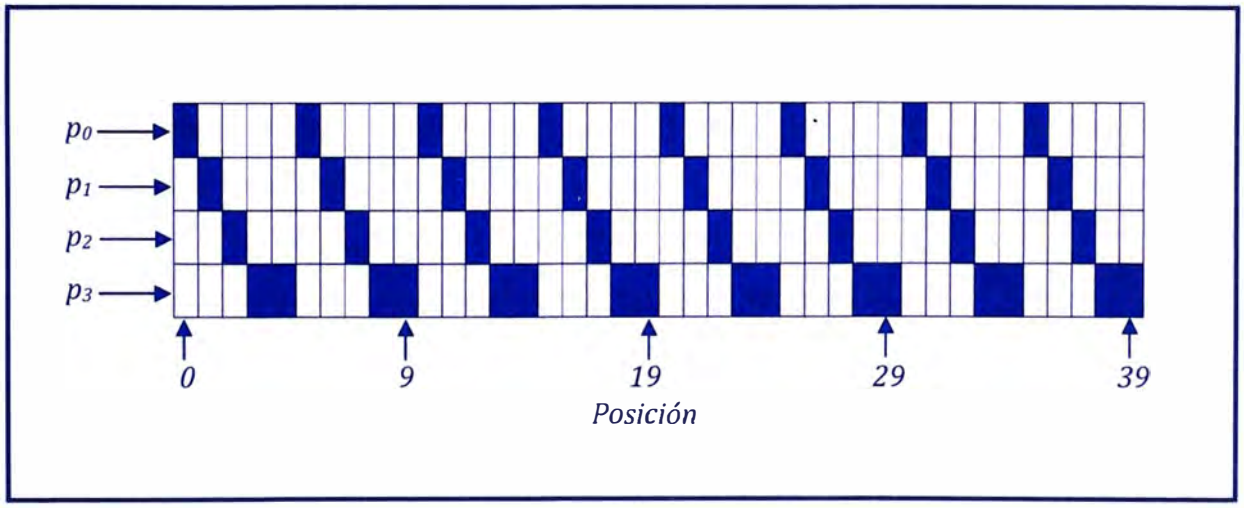

Fig. 3.1: Libro de Códigos Algebraico

El codificador G.729 tiene una estructura algebraica para el libro de códigos fijo, denominado como el Libro de Códigos Algebraico, en este esquema cada vector de códigos consta de cuatro pulsos distintos de cero. Cada impulso puede tener la amplitud de 1 ó -1 y puede asumir las posiciones indicadas en la Fig. 3.1.

Cada vector de códigos de excitación tiene 40 muestras, que es la longitud de una subtrama. El vector de códigos de excitación $v[n]$ se construye mediante la suma de los cuatro pulsos de acuerdo a:

$$v[n] = \sum_{i=0}^{3} p_i[n] = \sum_{i=0}^{3} s_i \delta[n - m_i]\,; \qquad n = 0, \ldots, 39 \tag{3.1}$$

donde $s_i = \pm 1$ es el signo de la señal del pulso y m_i es la posición. Así, cada impulso requiere un bit por señal. Para $p_0, p_1\ y\ p_2$ son necesarios 3 bits por posición, mientras que para p_3, son necesarios 4 bits. Por lo tanto, son necesarios un total de 17 bits para el índice de todo el libro de códigos.

Tabla 3.2: Valores a asignar en el libro de Códigos Algebraico

Impulso	Signo	Posiciones
i_0	$s_0 = \pm 1$	s_i: 0, 5, 10, 15, 20, 25, 30, 35
i_1	$s_1 = \pm 1$	s_i: 1, 6, 11, 16, 21, 26, 31, 36
i_2	$s_2 = \pm 1$	s_i: 2, 7, 12, 17, 22, 27, 32, 37
i_3	$s_3 = \pm 1$	s_i: 3, 8, 13, 18, 23, 28, 33, 38 4, 9, 14, 19, 24, 29, 34, 39

El índice del libro de códigos se compone de dos partes: signo (4 bits) y posición (13 bits). Utilizando una función de truncamiento, $f(x) = x$ si $x > 0$ y $f(x) = 0$ en otro caso. El índice de señal viene dado por

$$s_{indice} = f(s_0) + 2f(s_1) + 4f(s_2) + 8f(s_3) \tag{3.2}$$

representado con 4 bits, mientras que para la posición, tenemos

$$p_{indice} = \frac{m_0}{5} + 8\left[\frac{m_1}{5}\right] + 64\left[\frac{m_2}{5}\right] + \left(2\left[\frac{m_3}{5}\right] + ms3\right) \tag{3.3}$$

representado con 13 bits, donde

$$ms3 = \begin{cases} 0; & si \quad m_3 = 3, 8, \ldots, 38 \\ 1; & si \quad m_3 = 4, 9, \ldots, 39 \end{cases} \tag{3.4}$$

Notemos que (3.2) y (3.3) son los índices reales de transmisión como parte de la trama de bits G.729. La ventaja de este método es que no se requiere almacenamiento físico para el libro de códigos fijo, sin embargo, la cantidad de bits asignados es relativamente alta en comparación con los estándares CELP, tales como el FS1016.

4. Libro de Códigos Adaptivo

El codificador G.729 tiene algunas de las únicas metodologías aplicadas al libro de códigos adaptivo, es uno de los más complejos, con una estructura afinada para mejorar la calidad de la voz sintetizada, debido a los avances en la tecnología DSP, la complejidad adicional puede ser tolerada, de hecho, G.729 en comparación con las generaciones anteriores de codificadores CELP muestra un aumento en la complejidad, la cual es asequible con las mejoras constantes en la parte de hardware.

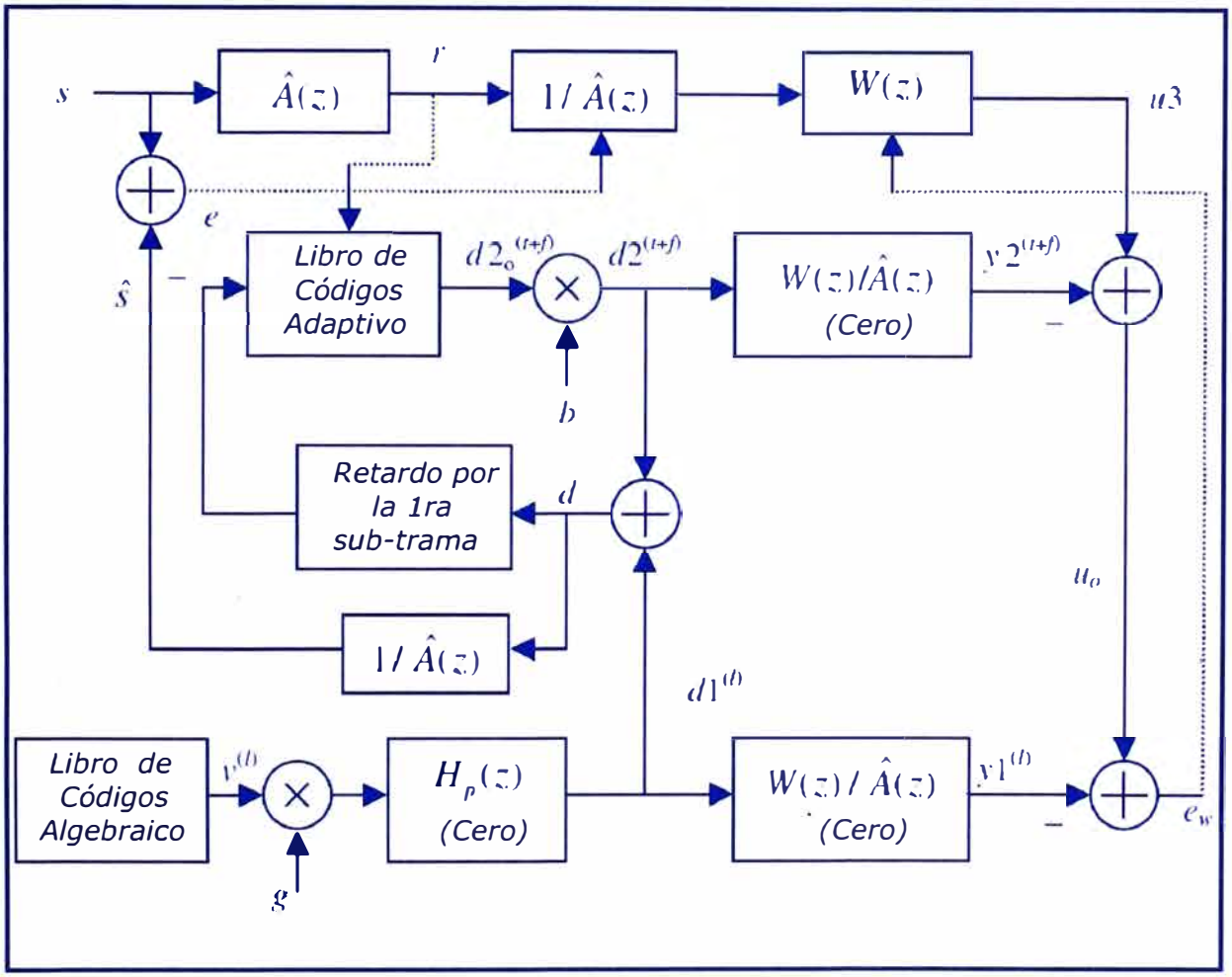

Fig. 3.2: Señales implicadas en el codificador G.729

4.1. Filtro de ponderación perceptual

G.729 utiliza un sofisticado filtro de ponderación perceptual, la función del sistema es:

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} = \frac{1 + \sum_{i=1}^{10} a_i \gamma_1^i z^{-i}}{1 + \sum_{i=1}^{10} a_i \gamma_2^i z^{-i}} \qquad (3.5)$$

Los a_i son coeficientes no cuantificados y se actualizan una vez cada sub-trama. Para cada sub-trama, los LPCs se obtienen a través de la interpolación en el dominio LSF.

El uso de LPCs no cuantificados en la ecuación (3.5) permite una mejor captura del espectro de la señal original, dando lugar a una calidad subjetiva elevada. Esto contrasta

con las generaciones pasadas de los codificadores CELP, tales como el FS1016, donde se utilizan coeficientes cuantificados debido a la simplicidad. El filtro de ponderación perceptual y el filtro de síntesis de formantes se fusionan para formar el filtro de síntesis de formantes modificada. Esta técnica no es aplicable al codificador G.729, donde la reducción modesta en el coste computacional se abandona en favor de la integridad de la señal más alta.

Los parámetros γ_1 y γ_2 determinan la respuesta de frecuencia del filtro y se hacen adaptivos dependiendo de las características espectrales. La adaptación se basa en una medida de espectro aplanado obtenido a través de las dos primeras RCs, así como la cercanía de los LSFs. Hay dos posibles valores de γ_1 0.94 ó 0.98 y γ_2 pertenece al rango [0.4,0.7]. Las reglas para la determinación de γ_1 y γ_2 se basan en pruebas subjetivas y la mejora en la calidad final de la voz ha sido reportada.

4.2. Estimación del Período de Pitch de Lazo Abierto

La estimación del Período de Pitch de Lazo Abierto es el primer paso hacia la búsqueda del Libro de Códigos Adaptivo. El procedimiento se basa en la ponderación de la señal de voz y se realiza una vez por cuadro, es decir, cada 10ms. La Fig. 3.3 resume los principales pasos. La voz perceptualmente ponderada, denotada por $u[n]$, se utiliza en la expresión de autocorrelación siguiente:

$$R[l] = \sum_{n=0}^{79} u[n]u[n-l] \tag{3.6}$$

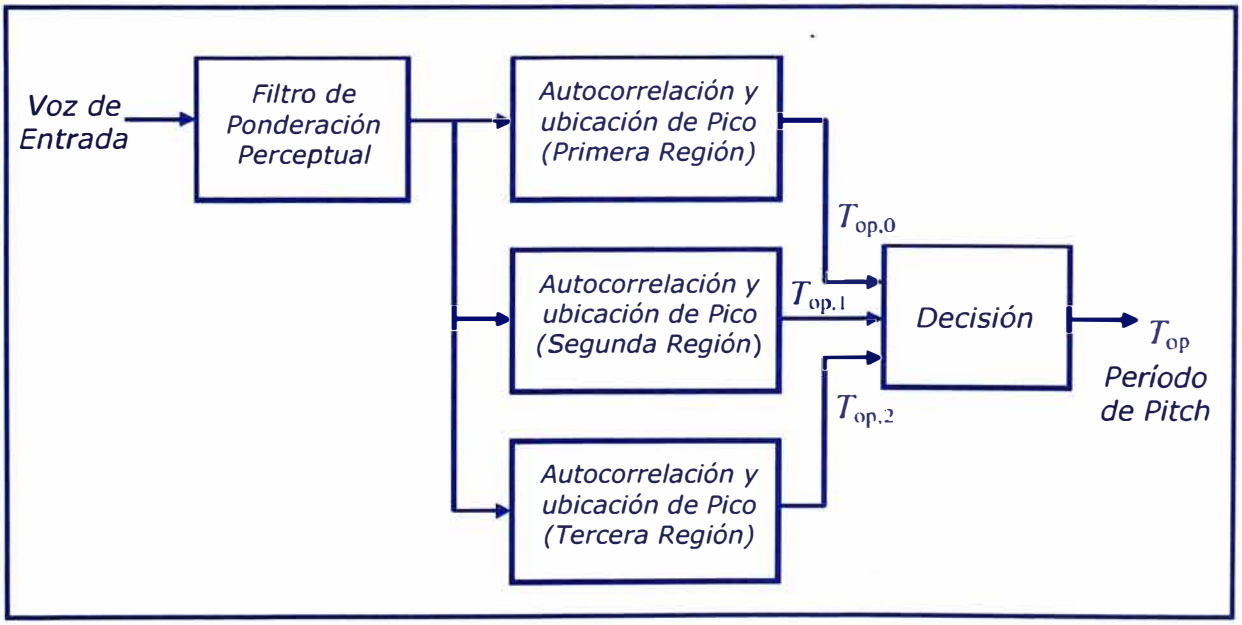

Fig. 3.3: Procedimiento de estimación de Período de Pitch de Lazo Abierto para el codificador G.729

Notemos que 80 muestras están implicadas correspondiente a dos sub-tramas. Tres picos se encuentra separado en tres rangos de retraso; estos resultados se denotan por

$$R_0 = R[T_{op,0}] \mid R[l] \leq R[T_{op,0}], \quad 20 \leq l, T_{op,0} \leq 39, \tag{3.7}$$

$$R_1 = R[T_{op,1}] \mid R[l] \leq R[T_{op,1}], \quad 40 \leq l, T_{op,1} \leq 79, \tag{3.8}$$

$$R_2 = R[T_{op,2}] \mid R[l] \leq R[T_{op,2}], \quad 80 \leq l, T_{op,0} \leq 143, \tag{3.9}$$

Los valores máximos acumulados de auto-correlación se normalizan de acuerdo a:

$$r_i = \frac{R_i}{\sqrt{\sum_n u^2[n - T_{op,i}]}}; \quad i = 0, 1, 2 \tag{3.10}$$

Se selecciona uno de estos tres resultados, siendo el seleccionado el que ofrezca menor retraso. Así, el procedimiento divide el rango de retraso en tres regiones, con el resultado final (T_{op}) orientado hacia la zona inferior, esto se realiza para evitar la elección de los múltiplos de Pitch.

4.3. Rango de búsqueda para el Período de Pitch

El período de pitch o índice de libro de códigos adaptivo se codifica con 8 bits para la primera sub-trama, para la segunda sub-trama se codifica con 5 bits, un valor relativo con respecto a la primera sub-trama.

Para la primera sub-trama, se utiliza un período de pitch fraccionado con una resolución de $1/3$ en el intervalo de $[19\ 1/3; 84\ 2/3]$ y sólo enteros en el intervalo $[85; 143]$.

Para la segunda sub-trama, un retardo con una resolución de $\frac{1}{3}$ se utiliza siempre en el rango de $T_2 \in \left[round(T_1) - 5\frac{2}{3}; round(T_1) + 4\frac{2}{3}\right]$, donde la función $round$ redondea al entero más cercano. Donde T_1 y T_2 son los períodos de pitch de la primera y segunda sub-trama, respectivamente.

Para cada sub-trama, el período de pitch se determina mediante la búsqueda a través del libro de códigos adaptivo de tal manera que el error cuadrático ponderado se reduzca al mínimo.

El rango de búsqueda se determina a partir del T_{op} estimado de lazo abierto: se encuentra aproximadamente en $T_{op} \pm 3$ y $T_{op} \pm 5$ para la primera y segunda sub-trama, respectivamente.

4.4. Señales involucradas en la Búsqueda del Libro de Códigos Adaptivo

Después de conocer el rango de Búsqueda del Libro de Códigos Adaptivos, se recogen las siguientes señales antes de buscar en el libro de códigos.

a) Matriz de Libro de Códigos Adaptivo o Secuencia de Excitación:

Denotada como $d2[n]$, para $n < 0$, la matriz contiene muestras de excitación pasada (d en la Fig. 3.2), para $0 \leq n < N$, cuando $N = 0$ es la longitud de la sub-trama, la matriz contiene error de predicción de la sub-trama actual, que se obtiene filtrando la voz de entrada utilizando el Filtro de Predicción de Error. Los coeficientes del filtro son los LPC cuantificados e interpolados, con la función del sistema denotado por $\hat{A}(z)$. Con esta disposición, el codevector de adaptación preliminar para el periodo de pitch t entero está dado por:

$$d2'^{(t)}[n] = d2[n-t]; \quad 0 \leq n < N \tag{3.11}$$

En la formulación tradicional del Libro de Códigos Adaptivo, el codevector repite las t muestras de $d2[n]$ en el rango $n \in [n-t, -1]$ cuando $t < n$. Un inconveniente de repetir las t muestras es que la convolución recursiva no se aplica cuando $t < N$.

Para mejorar la eficiencia computacional, $d2[n]$, $0 \leq n < N$ se llena de error de predicción de modo que la convolución recursiva puede ser utilizado para todo t.

Debemos tener en cuenta que los codevectores (3.11) son de carácter preliminar, no son la versión final utilizada para la síntesis de habla.

b) Elección de la secuencia para la Búsqueda del Libro de Códigos Adaptivo

La secuencia $u3[n]$ se encuentra utilizando la voz perceptualmente ponderada y la respuesta de Entrada Cero de la conexión en cascada entre el Filtro de Ponderación Perceptual y el Filtro de Síntesis de Formantes.

La función del sistema de Filtro de Síntesis Formante se denota por $1/\hat{A}(z)$, donde los coeficientes son los LPCs cuantizados e interpolados.

La respuesta al impulso de la conexión en cascada entre el Filtro de Ponderación Perceptual y el Filtro de Síntesis de Formantes, se denota por $h[n]$.

4.5. Búsqueda del Retardo Entero

Dado el codevector preliminar $d2'^{(t)}[n]$, es procesado por los filtros en cascada para generar la secuencia de salida $y2'^{(t)}[n]$ como sigue:

$$y2'^{(t)}[n] = d2'^{(t)}[n] * h[n] = \sum_{k=0}^{n} h[k] d2'^{(t)}[n-k] \tag{3.12}$$

para $t = T_{mín} - 4$ hasta $T_{máx} + 4$ y $n = 0$ hasta $N - 1$, con los valores de $T_{mín}$ y $T_{máx}$ encontrados en el paso anterior. Notamos que el rango de t se ha extendido para permitir la interpolación. La correlación normalizada se calcula utilizando

$$R[t] = \frac{\sum_{n=0}^{39} u3[n]y2'^{(t)}[n]}{\sqrt{\sum_{n=0}^{39}(y2'^{(t)}[n])^2}} \tag{3.13}$$

El valor de t que maximiza $R[t]$ en el intervalo $[T_{mín}, T_{máx}]$ se denota como T, que es el retardo entero buscado.

4.6. Búsqueda del Retardo Fraccionario

Si se trata de la primera sub-trama y el retardo entero satisface $T > 84$, la parte fraccional del período de pitch se establece en cero. De lo contrario, se evalúan las fracciones alrededor de T mediante la interpolación de los valores de correlación (3.13) de la siguiente manera:

$$R_0(t,f) = \sum_{i=0}^{3} R[t-i]w13[3f+3i] + \sum_{i=0}^{3} R[t+i+1]w13[3-3f+3i] \tag{3.14}$$

Para $f = \left\{0, \frac{1}{3}, \frac{2}{3}\right\}$. En la implementación real, se consideran cinco valores fraccionarios: $f = \left\{-\frac{2}{3}, -\frac{1}{3}, 0, \frac{1}{3}, \frac{2}{3}\right\}$. Debido a que (3.14) sólo se aplica para $f \geq 0$, t debe ser disminuido en uno con el fin de evaluar f negativo. Para la transmisión, sólo se utilizan tres valores fraccionarios:

$f = \left\{-\frac{1}{3}, 0, \frac{1}{3}\right\}$. Si $f = -\frac{2}{3}: T \leftarrow T - 1, f \leftarrow \frac{1}{3}$. Si $f = \frac{2}{3}: T \leftarrow T + 1, f \leftarrow \frac{1}{3}$.

Los pesos de interpolación $w13$ se obtienen a través de una Función Sinc Truncada de Hamming en ± 11 y se rellena con ceros en ± 12.

4.7. Generación del Codevector Adaptivo

Dados t, f, y $d2$, el codevector adaptivo actual utilizado para la síntesis de voz se calcula con el siguiente procedimiento:

```
Adaptive_cv(d2,t,frac)
1. n0 ← -t
2. frac ← frac
3. if frac < 0
4.    frac ← frac + 3
5.      n0 ← n0 - 1
6. for n ← 0 to 39
7.      d2[n] ← 0
8.      for i ← 0 to 9
9.      d2[n] ← d2[n] + d2[n0-i+n] . w31[frac+3.i] + d2[n0+1+i+n] - w31[3-frac+3.i]
```

En el código anterior, el parámetro frac es igual a -1, 0 ó 1, que corresponden a los valores fraccionarios de $-\frac{1}{3}$, 0 y $\frac{1}{3}$, respectivamente.

Los codevectores se obtienen con $d2\,[n]$, para $n = 0$ a 39.

Los pesos de interpolación $w31\,[n]$ se obtienen de una Función Sinc Truncada de Hamming en ± 29 y rellena con ceros en ± 30.

Notemos desde el pseudocódigo anterior de que los valores iniciales de la matriz $d2$ para $n \in [0, 39]$, son completamente reemplazados.

Recordemos que, antes de buscar en el Libro de Códigos Adaptivo, estos elementos de la matriz están llenos con muestras de predicción de error.

La interpolación se hace para todos los valores fraccionarios, incluyendo $f = 0$.

Así, codevector adaptivo, en adelante se denotará como $d2_0^{(t+f)}[n] = d2[n], n = 0$ a 39, con $d2[n]$ como el resultado de la aplicación de la rutina de interpolación "Adaptive_cv", que en general, ya no son los mismos codevectores preliminares, definidos en (3.11).

De esta manera, el codificador y el decodificador son capaces de generar exactamente los mismos codevectores a partir de t y f, puesto que las muestras del codevector se forman exclusivamente a partir de excitación pasada.

Por lo tanto, tenemos un procedimiento de búsqueda, donde los retardos son números enteros y fraccionarios basado en un conjunto de codevectores preliminares.

Una vez que los retardos se determinan, se encuentra el codevector, que es diferente de la versión preliminar. ¿Qué tan preciso podría ser? Pues bien, como se muestra en el siguiente ejemplo, la precisión es bastante buena cuando $t \geq 40$. En este caso, utilizando la interpolación (3.14) e identificando el pico de correlación se produce un resultado similar al primero calcular los codevectores adaptivos actuales y luego encontrando la correlación utilizando una ecuación como (16.13). Esto es debido a que los pesos de interpolación $w13$ y $w31$ se seleccionan para lograr una igualdad. Para $t < 40$ sin embargo, el proceso es inexacto, ya que los codevectores preliminares obtenidos a partir de las muestras de predicción de error no son los mismos que los que se encuentran a través de interpolación de las muestras anteriores. Ello limita el potencial del codificador para representar la voz con períodos de paso corto, como por ejemplo las mujeres o los niños.

No obstante, la técnica funciona bastante bien y cuando se combina con el resto de componentes del codificador, la voz sintética es de buena calidad en la mayoría de los casos.

4.8. Resumen

El G.729 adopta una estrategia para la Búsqueda de Libro de Códigos Adaptivo que intenta ser computacionalmente eficiente sobre algún sacrificio en la precisión global,

especialmente durante períodos cortos de pitch $(t + f < 40)$. Podemos decir que la aproximación es subóptima en el sentido de que no se encuentra la mejor solución posible, en beneficio de reducción de carga computacional. Sin embargo, en la práctica, el método en combinación con otros componentes del codificador produce voz sintética de alta calidad.

5. Codificación y Decodificación

El G.729 sigue los esquemas básicos de codificación y decodificación como en el codificador CELP genérico con una complejidad adicional introducida para lograr un mayor rendimiento.

5.1. Señales involucradas en la Codificación y Actualización del Filtro de Estado

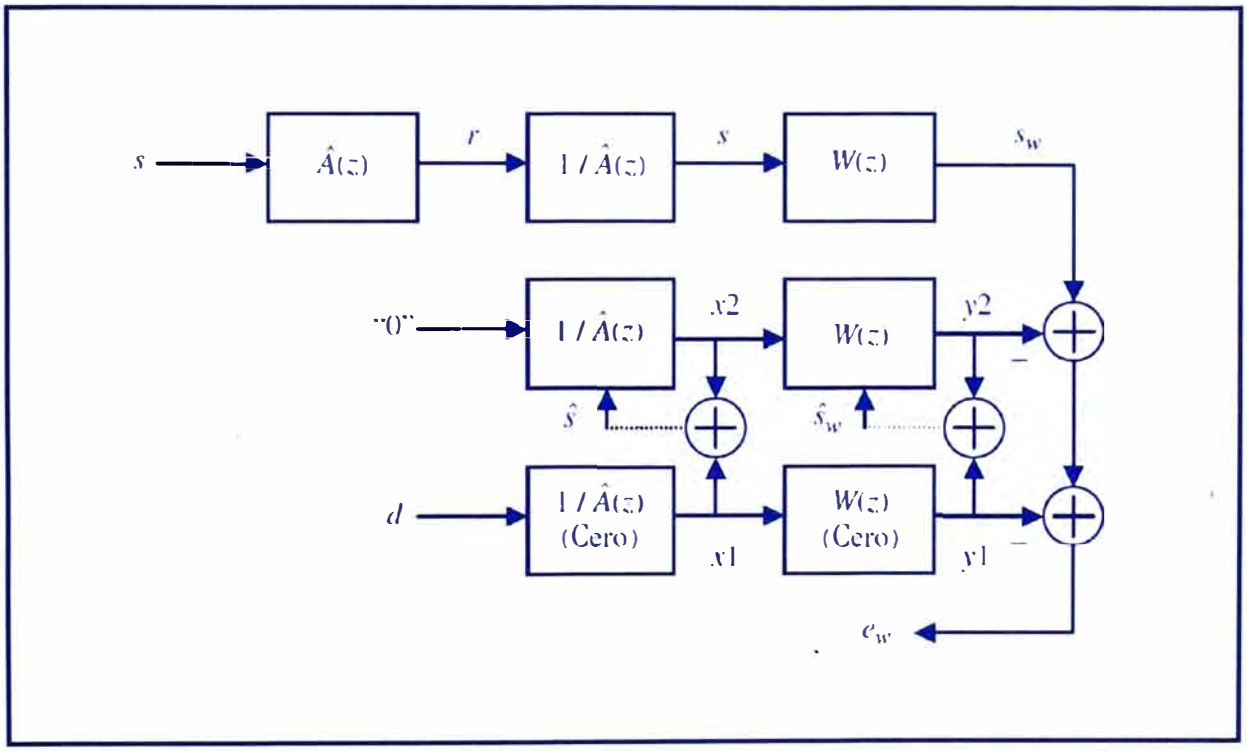

Fig. 3.4: Señales involucradas en la Codificación G.729 (equivalente a la Fig. 3.1)

Para el codificador G.729, el Filtro de Ponderación Perceptual utiliza los LPCs originales (no cuantificados), mientras que el Filtro de Síntesis de Formantes emplea los LPC cuantificados.

La ventaja de esta disposición es que, en teoría, la ponderación perceptual es más precisa, ya que no se introduce distorsión debido a la cuantización. El precio a pagar es la complejidad adicional.

Es posible simplificar aún más el sistema mediante la combinación de las dos capas superiores de los filtros. La Fig. 3.5 resume la idea. Los filtros marcados como #1 y #2 requieren una atención especial.

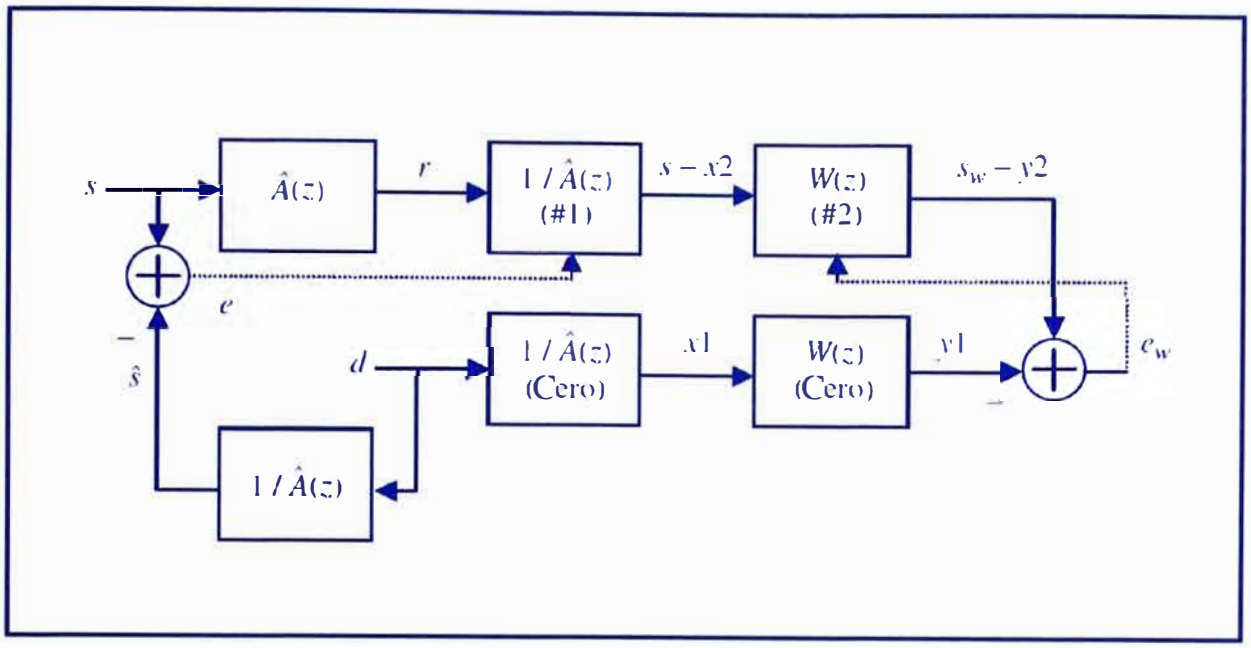

Fig. 3.5: Señales que participan en la codificación G.729: otra realización equivalente.

Para el filtro #1, su estado está dado por el error de síntesis $e = s - \hat{s}$.

Así, la salida del filtro en respuesta a r (error de predicción) con un estado inicial dado por s voz de entrada es igual a s, mientras que la salida en respuesta a la "entrada cero" con estado inicial dado por $\hat{s}$ (voz sintética) es igual a $x2$ (Fig. 3.4). Por lo tanto, la respuesta general es dada por $s - x2$.

Para el filtro #2, su estado viene dado por el error de síntesis ponderado $e_w = s_w - \hat{s}_w$. La salida del filtro en respuesta a s con estado inicial dado por s_w (voz de entrada ponderada) es igual a s_w, mientras que la salida en respuesta a $-x2$ con el estado inicial $-\hat{s}_w$ es igual a $-y2$ (Fig. 3.5). Por lo tanto, la respuesta general es dada por $s_w - y2$.

La ventaja de esta configuración es su mayor rendimiento en un entorno de punto fijo. En el enfoque genérico y en la Fig. 16.9, la respuesta $(x2)$ del filtro con "Entrada Nula" se calcula por separado y después se resta del la voz perceptualmente ponderada. El problema es que las muestras de la respuesta de entrada nula decaen rápidamente hacia cero, debido al hecho de que el filtro es estable.

Con respecto al punto fijo, las muestras de baja magnitud no pueden ser representadas fielmente, dando lugar a errores de cuantificación relativamente altos. El sistema de la Fig. 3.10 evita calcular la respuesta de entrada nula por separado. Como podemos ver, los filtros procesan en todo momento entradas de voz de relativamente alta amplitud. Mediante la inicialización adecuada de los filtros, la funcionalidad equivalente se logra.

Por lo tanto, no hay señales débiles que se calculen de forma aislada, llevando a una mayor precisión de los resultados finales.

5.2. Revisión de las operaciones de codificación

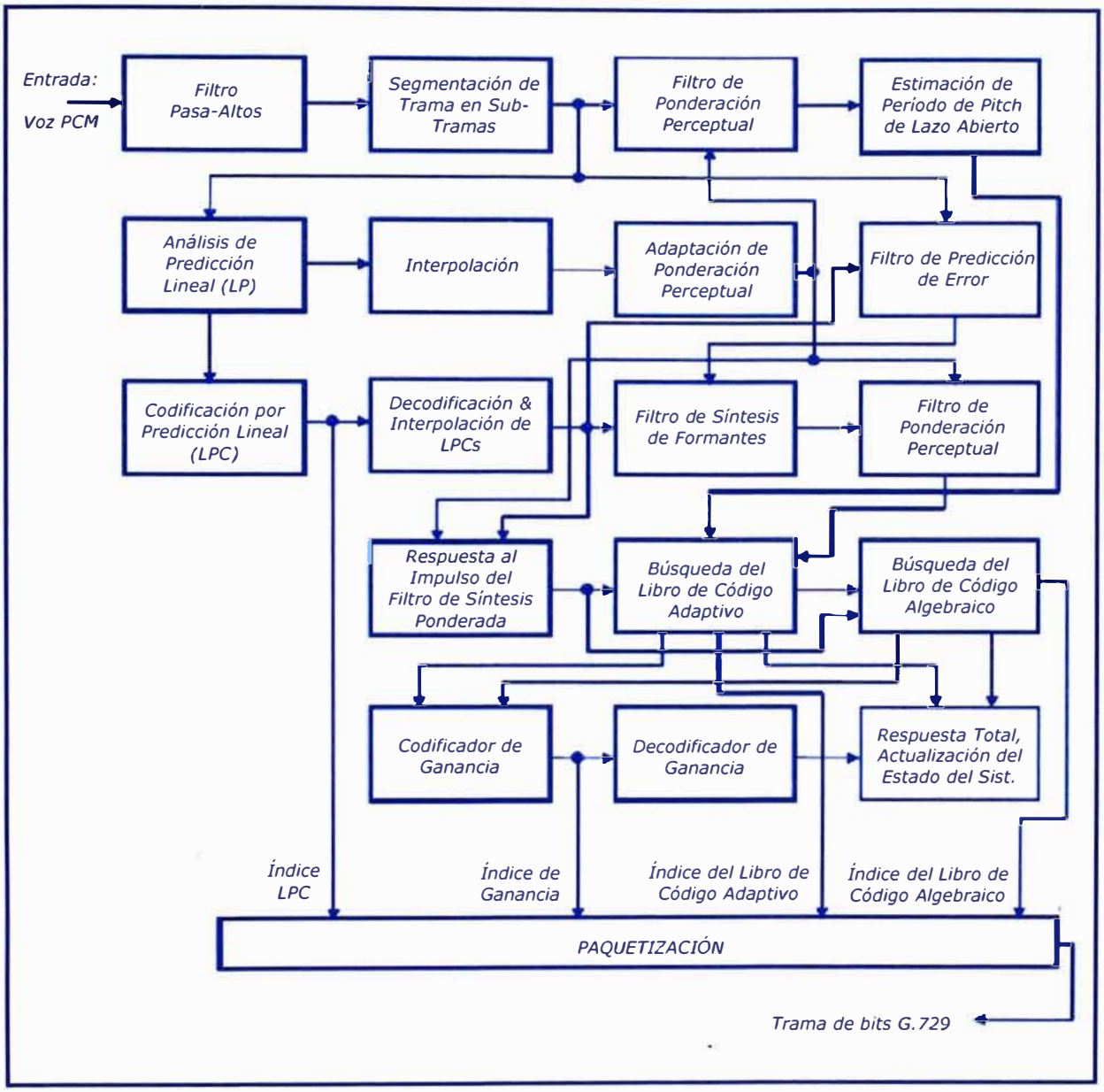

Fig. 3.6: Diagrama de bloques del codificador G.729.

La señal de voz de entrada atraviesa un filtro paso alto y es segmentada en tramas de 80 muestras (10 ms), subdivididos en dos sub-tramas de 40 muestras,

Se realiza el análisis de LP una vez por cuadro.

Se obtienen al final dos conjuntos de LPCs: originales y cuantificados. Ambos conjuntos se interpolan con el fin de aplicar individualmente a cada sub-trama.

Los LPC originales (después de la interpolación) son utilizados por el filtro de ponderación perceptual. La sub-trama de voz de entrada se filtra a continuación por el mismo, con la salida que se utiliza en la estimación del período de pitch en bucle abierto.

Se siguen los procesos de adaptación de ponderación perceptual y la estimación de periodo de pitch

Se calcula la respuesta al impulso de la cascada de filtros entre el filtro de ponderación perceptual y el filtro de síntesis de formante.

El filtro resultante mediante la función del sistema $W(z)/\hat{A}(z)$ se le conoce como el Filtro de Síntesis Ponderado. Tenemos en cuenta que el Filtro de Síntesis de Formantes utiliza LPCs $\hat{a}_i$ cuantizados, con la función del sistema.

$$\frac{1}{\hat{A}(z)} = \frac{1}{1 + \sum_{i=1}^{10} \hat{a}_i z^{-i}} \tag{3.15}$$

El objetivo de la señal $u3$ en la búsqueda de libro de códigos adaptivo es encontrar la inicialización adecuada para los filtros, utilizando el esquema de la Fig. 3.2.

El cálculo de la ganancia del Libro de Códigos Adaptable resultante está delimitado entre 0 y 1.2.

Tabla 3.3: Asignación de Bits en la codificación G.729

Parámetro	**Número por Tramas**	**Resolución**	**Total de Bits por Trama**
Índice LPC	1	18	18
Período de Pitch (Índice de Libro de Código Adaptivo)	2	8, 5	13
Bit de Paridad para el Período de Pitch	1	1	1
Índice de Libro de Código Algebraico	2	17	34
Índice de Ganancia	2	7	14
Total			**80**

5.3. Resumen de Operaciones de Decodificación

La Fig. 3.12 muestra el decodificador. El codevector algebraico se halla a partir del índice recibido, especificando las posiciones y los signos de los cuatro pulsos, atraviesa el Filtro de Síntesis de Pitch en el estado cero.

El codevector adaptativo se recupera de la parte entera del período de pitch y se interpola según la parte fraccionaria del período de pitch.

Los codevectores algebraicos y adaptativos se escalan por las ganancias decodificadas y se adicionan para formar la excitación del Filtro de Síntesis de Formante. Un post-filtro se incorpora para mejorar la calidad subjetiva.

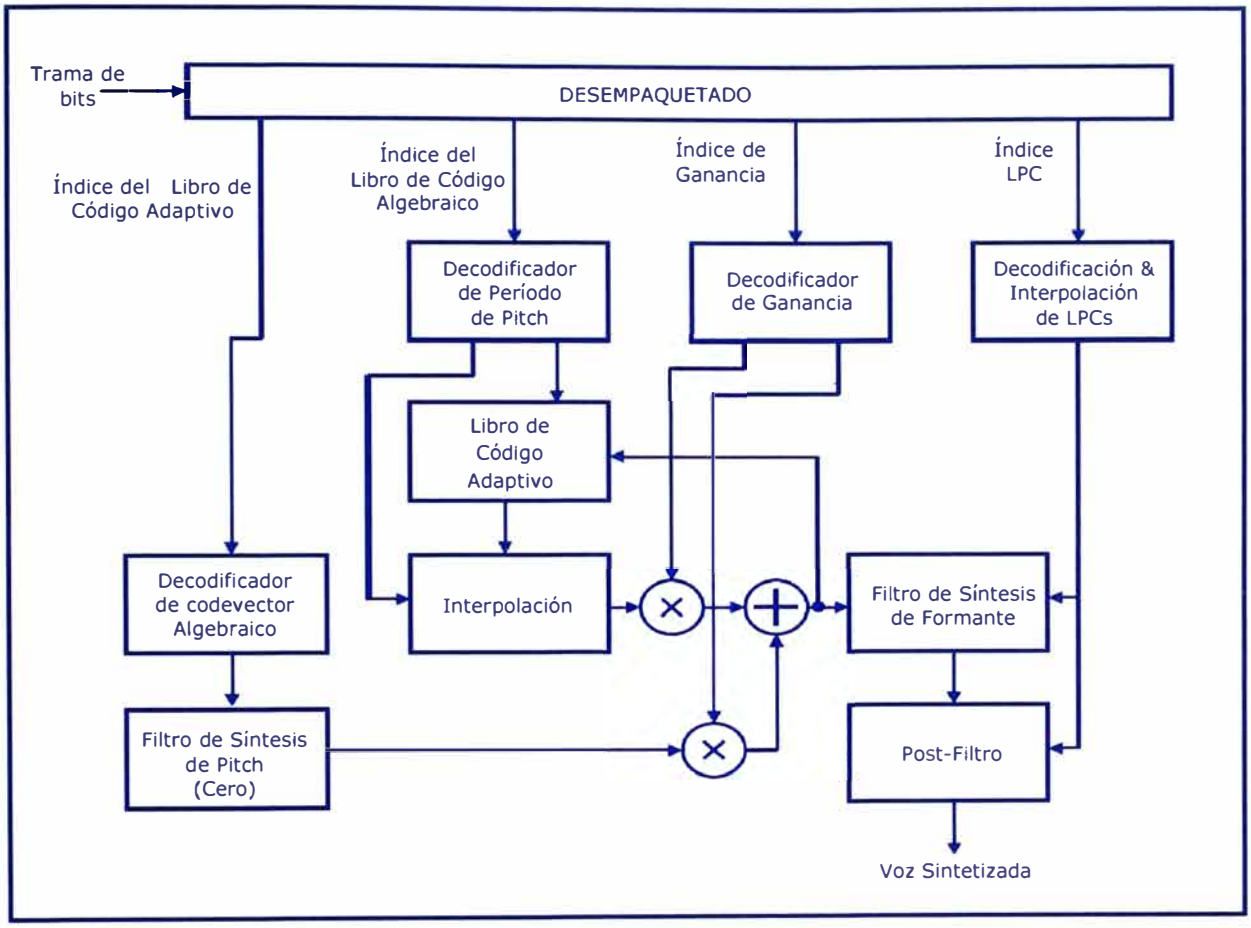

Fig. 3.6: Diagrama de bloques del decodificador G.729.

CAPÍTULI IV
SIMULACIONES

1. Captura y Reproducción de Tramas RTP codificadas con G.711

Para este caso haremos únicamente uso de las herramientas del software Wireshark, con el que capturaremos las tramas de una conversación VoIP, seleccionaremos las tramas codificadas y las reproduciremos con la ayuda del algoritmo G.711 incorporado y la herramienta de reproducción de las tramas decodificadas ambos incorporados en el Software Wireshark.

Se empleará una PC dispuesta de acuerdo al esquema de la Fig. 4.1, con la que podremos analizar las tramas que lleguen y salgan del Teléfono IP N° 2.

1.1. Esquema utilizado:

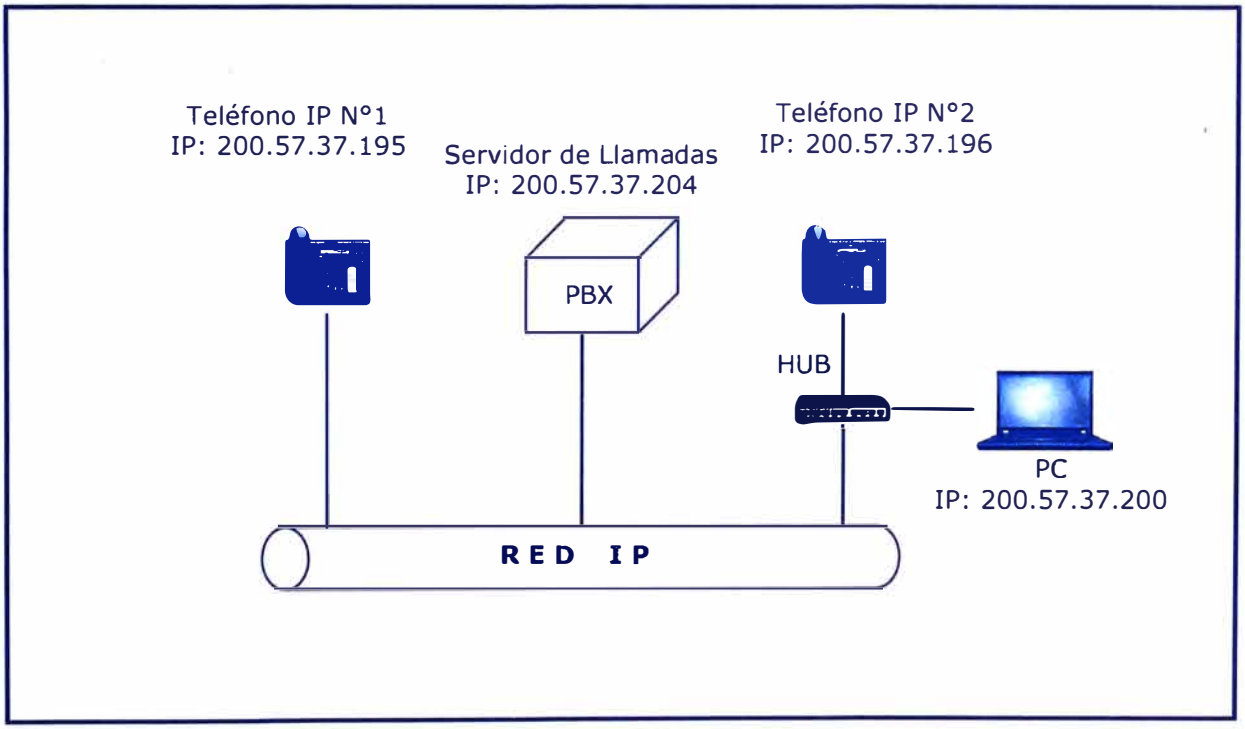

Fig. 4.1: Esquema utilizado para la captura y reproducción de tramas RTP codificadas en G.711

Para capturar las tramas entrantes y salientes al Teléfono IP N°2, colocamos una PC entre este teléfono IP y la RED IP, a través de un HUB.

1.2. Captura de trama con Wireshark

Activamos la captura de tramas con esta herramienta en el momento preciso en el que se va a realizar una llamada VoIP.

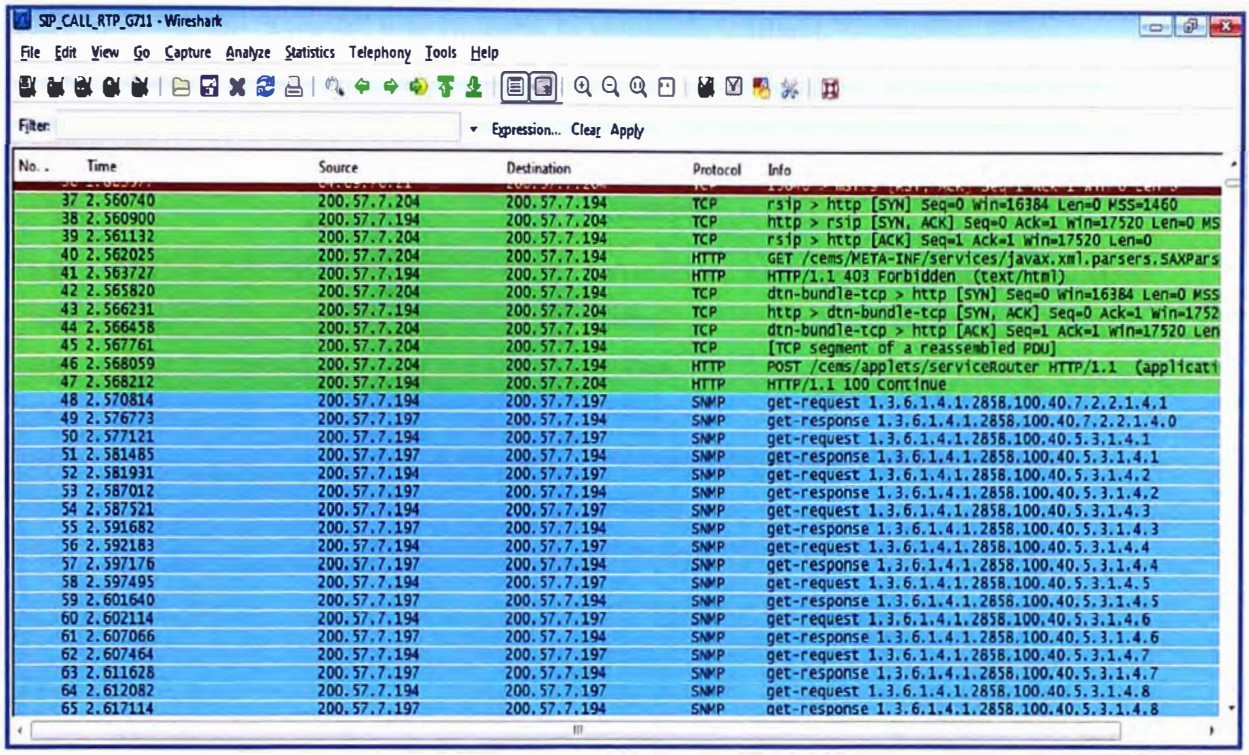

Fig. 4.2: Vista de tramas capturadas por el wireshark

1.3. Filtrado de las llamadas VoIP.

Utilizamos la herramienta para filtrar las llamadas VoIP, con lo que se mostrar{an todas las tramas que correspondan a alguna llamada VoIP que lleguen hasta el teléfono IP N° 1. Usamos la siguiente ruta de opciones en la vista principal del wiereshark:

Telephony / VoIP Calls /

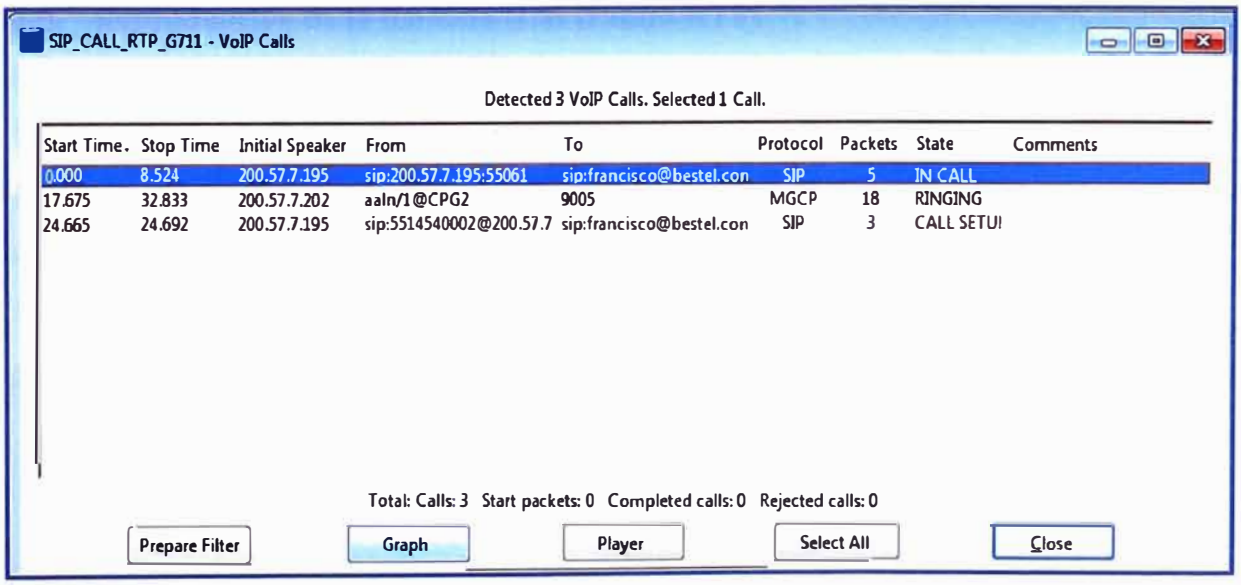

Fig. 4.3: Vista de tramas de llamadas VoIP capturadas por el Wireshark

1.4. Gráfica de análisis de la llamada

Según la Figura 4.3, seleccionamos la primera opción, que muestra la llamada originada por el teléfono IP N° 2 de dirección 200.57.7.195.

Hacemos clic en "Graph" y se muestra la siguiente ventana.

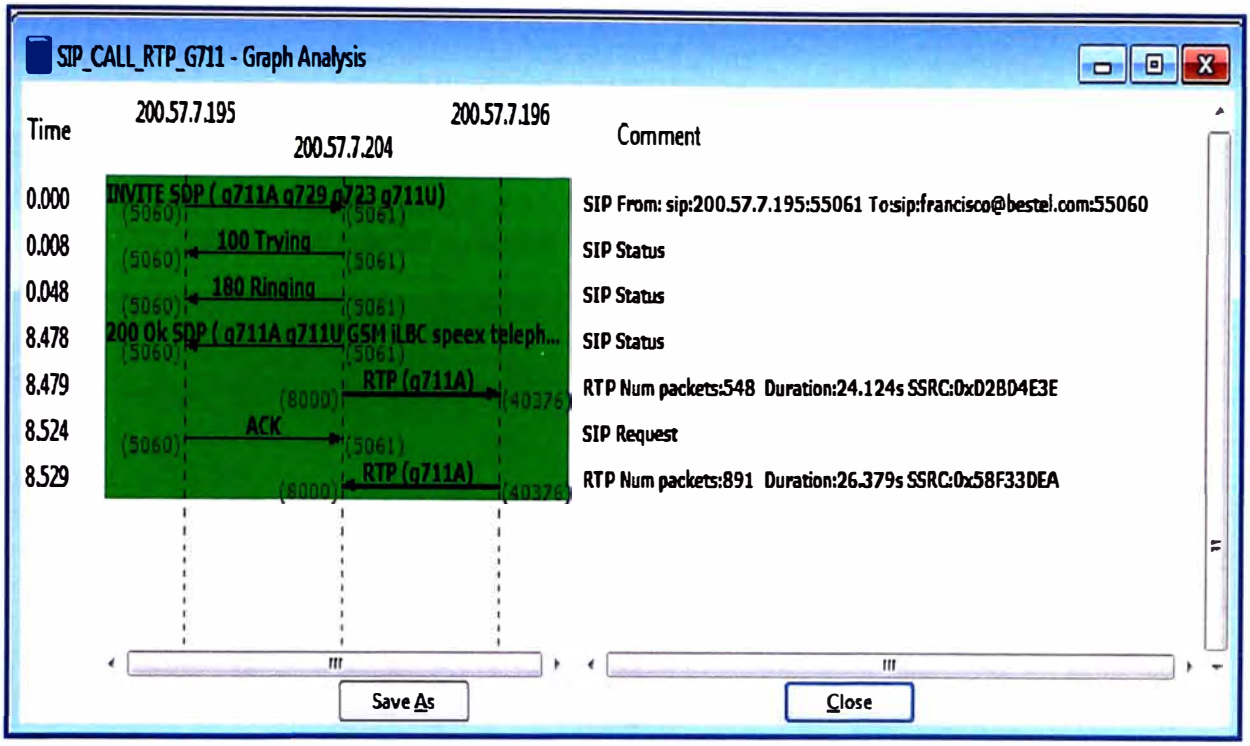

Fig. 4.4: Vista de las secuencias seguidas en la comunicai{on de acuerdo al esquema de la Fig. 4.1

Se observa la gestión SIP desde 200.57.7.195 (Teléfono IP N° 1) al 200.57.7.196 (Teléfono IP N° 2), a través del 200.57.7.204 (Servidor de Llamadas), así como del paso de las tramas entre el Servidor de Llamadas y el Teléfono IP N° 2.

1.5. Reproducción de la llamada (Las tramas RTP)

Para reproducir la llamada que consta de 02 tramas RTP una desde el servidor de llamadas hacia el teléfono IP N° 2 y la otra en sentido contrario, del teléfono IP N°2 hacia el teléfono, utilizamos la herramienta de reproducción de wireshark, la cual cuenta con el algoritmo del codificador G.711 incorporado. Para esto seguimos los siguientes pasos.

Seleccionamos "Player" de acuerdo a la Fig. 4.3

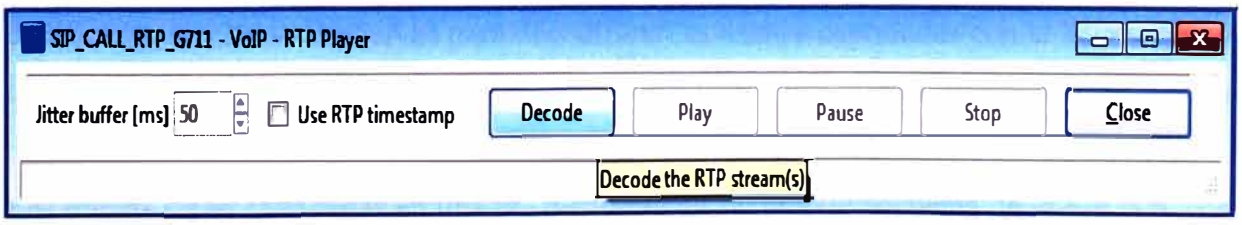

Fig. 4.5: Vista inicial del reproductor de Wireshark

Seleccionamos "Decode", y se empiezan a crear las vistas de las señales de voz decodificadas.

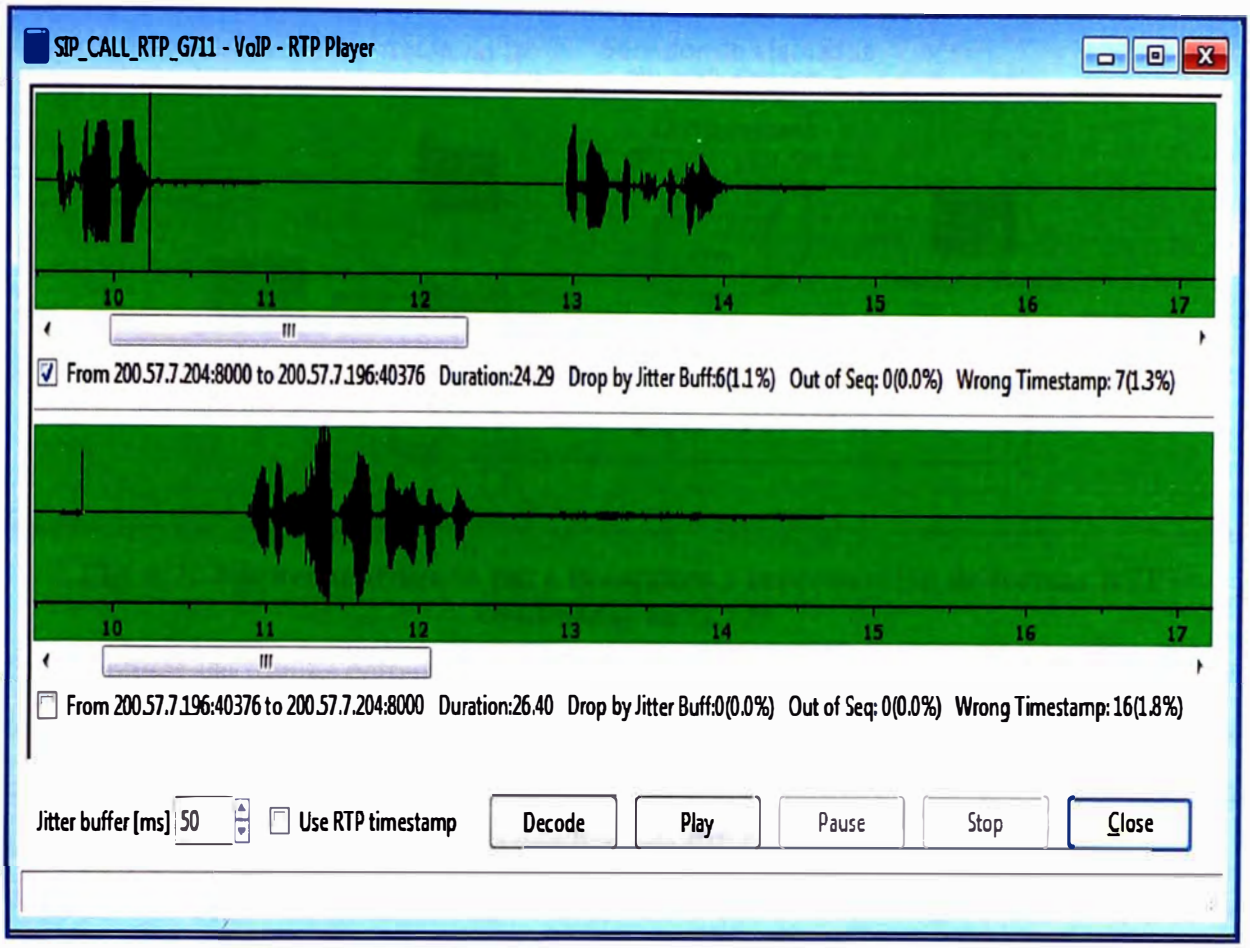

Fig. 4.6: Vista de las señales de voz decodificadas

Al seleccionar "Play", podemos escuchar la conversación entre los 2 puntos.

"¿Se escucha?" ----------"bueno si te escucho bien"----------

---------- "acá se escucha claramente"----------

2. Captura y Reproducción de Tramas RTP codificadas con G.729

Para este caso haremos uso de las herramientas del software Wireshark, como en el caso anterior, pero solo para capturar las tramas codificadas con G.729 de una conversación VoIP, las cuales no podremos reproducir con el reproductor del wireshark, debido a que esta herramienta no cuenta con los algoritmos para comprimir ni descomprimir tramas G.729, por este motivo utilizaremos dos herramientas adicionales, la primera, el codificador G.729 de VoiceAge para decodificar las tramas capturadas con el Wireshark, y la segunda, el reproductor Audacity, para reproducir las tramas decodificadas por el decodificador G.79 de VoiceAge.

Las tramas serán capturadas en una PC dispuesta de acuerdo al esquema de la Fig. 4.6, con la que podremos analizar las tramas que lleguen y salgan del Teléfono IP.

2.1. Esquema:

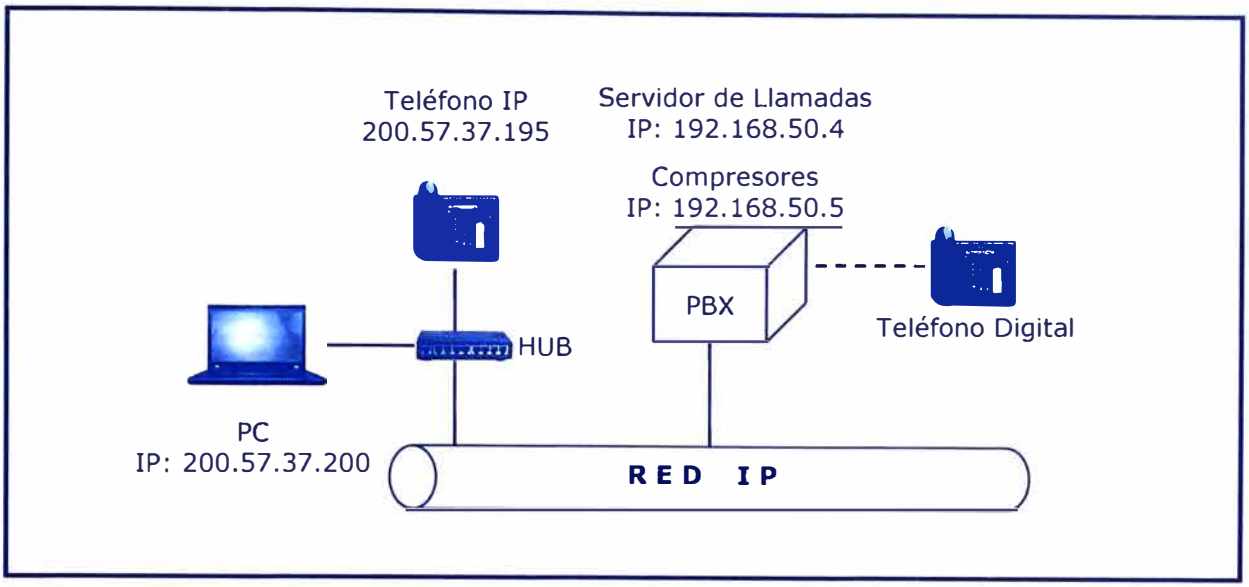

Fig. 4.7: Esquema utilizado para la captura y reproducción de tramas RTP codificadas en G.729

Para capturar las tramas entrantes y salientes al Teléfono IP, colocamos una PC entre este teléfono IP y la RED IP, a través de un HUB. Notemos que este servidor de llamadas cuenta con una dirección IP principal y otra dirección IP para la tarjeta de compresores con la que cuenta. Captura de la trama de una llamada SIP (G729)

Activamos la captura de tramas con esta herramienta en el momento preciso en el que se va a realizar una llamada VoIP.

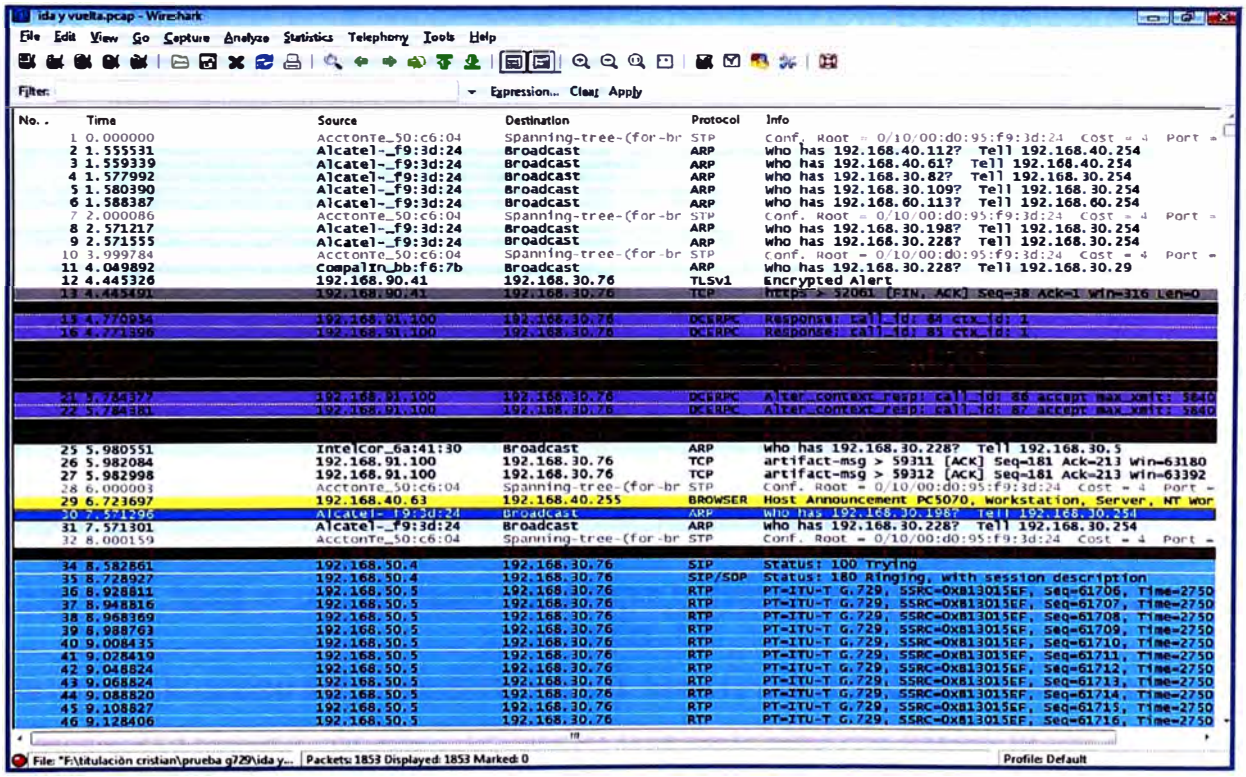

Fig. 4.8: Vista de tramas capturadas por el wireshark

2.2. Filtrado de las llamadas VoIP.

Utilizamos la herramienta para filtrar las llamadas VoIP, con lo que se mostrar{an todas las tramas que correspondan a alguna llamada VoIP que lleguen hasta el teléfono IP N° 1. Usamos la siguiente ruta de opciones en la vista principal del wiereshark:

Telephony / VoIP Calls /

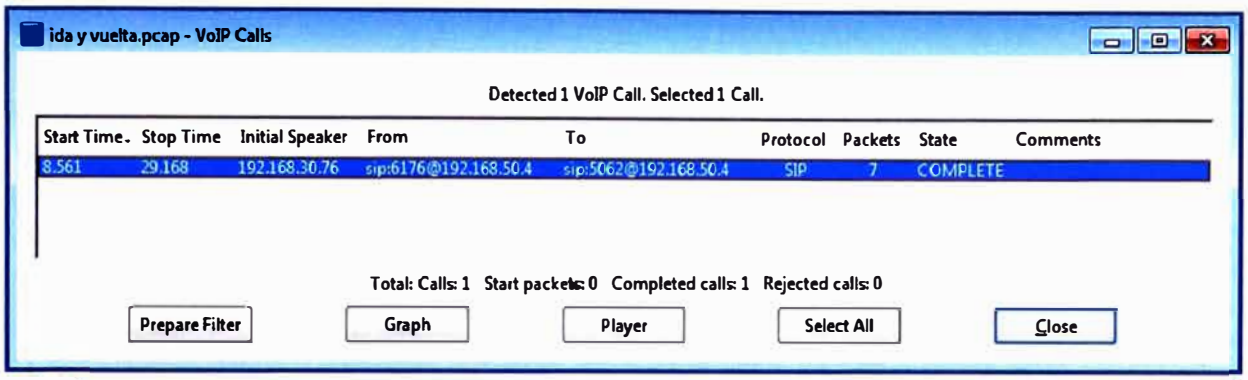

Fig. 4.9: Vista de tramas de llamadas VoIP capturadas por el Wireshark

Nos muestra solo una llamada detectada, la cual se encuentra en estado "Completo". Es decir tiene un inicio de sesión, una sesión propiamente dicha y un fin de sesión.

2.3. Gráfica para analizar la llamada

Según la Figura 4.8, Hacemos clic en "Graph" y se muestra la siguiente ventana.

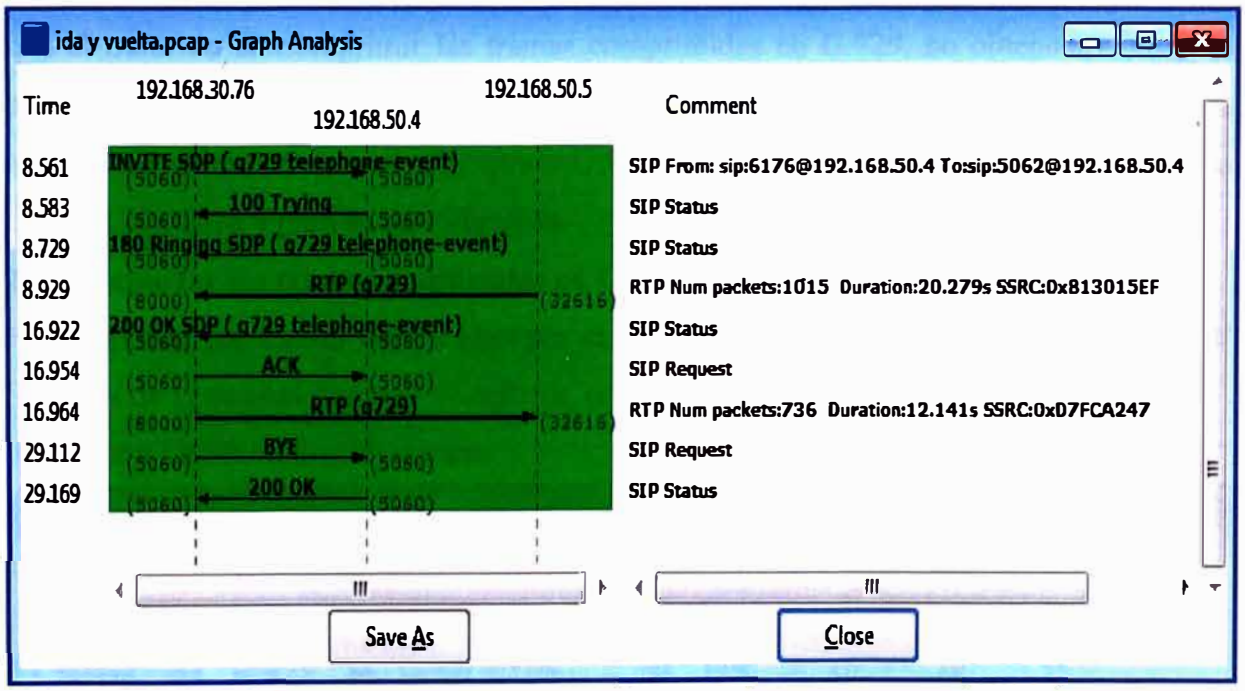

Fig. 4.10: Vista de las secuencias seguidas en la comunicaión de acuerdo al esquema de la Fig. 4.7

Se observa la gestión SIP desde el Teléfono IP hacia la tarjeta de Compresores, a través del Servidor de Llamadas, así como del paso de las tramas entre el Teléfono IP y la tarjeta de Compresores.

2.4. Reproducción de la llamada (Las tramas RTP) con wireshark

Para reproducir la llamada que consta de 02 tramas RTP como en el caso del esquema anterior, utilizando el propio reproductor del wireshark, obtenemos lo siguiente.

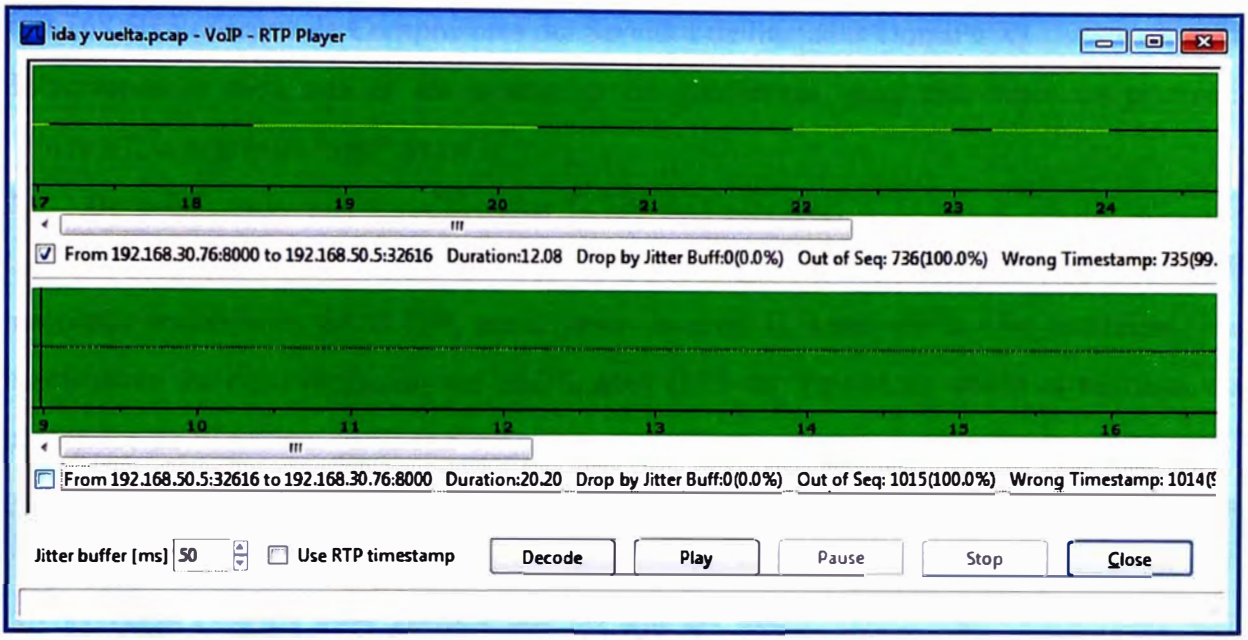

Fig. 4.11: Vista del reproductor del Wiresahrk

En este caso a diferencia del anterior, no obtenemos las señales de voz decodificadas, puesto que la herramienta Wireshark, únicamente cuenta con el algoritmo G.711, por lo que al tratar de descomprimir las tramas comprimidas en G.729, no obtenemos ninguna señal. Por esta razón, requerimos de otro par de herramientas para la descompresión de las tramas comprimidas con el algoritmo G.729 del teléfono y el servidor de llamadas OmniPCX Office y para su reproducción.

2.5. Exportar las tramas codificadas en G.729

Para exportar las tramas codificadas por el teléfono IP y la Tarjeta de Compresores del Servidor de Llamadas desde el Wireshark, seguimos la siguiente ruta:

Telephony / RTP / Show all streams

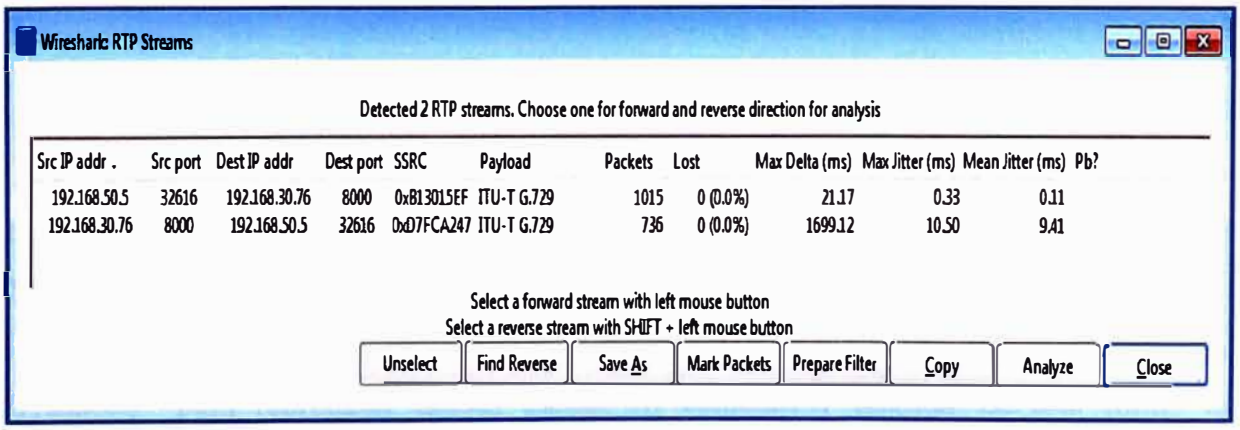

Fig. 4.12: Vista de tramas VoIP encontradas con el Wireshark

Podemos identificar 02 tramas RTP: La primera desde la IP 192.168.50.5 (Tarjeta de Compresores del Servidor de llamadas OmniPCX) hacia la IP 192.168.30.76 (Teléfono IP). La segunda en sentido contrario, desde la IP 192.168.30.76 (Teléfono IP) hacia la IP 192.168.50.5 (Tarjeta de Compresores del Servidor de llamadas OmniPCX)

Seleccionamos cada una de las mismas y las guardamos, para este caso: La primera: "ida.raw", la segunda: "vuelta.raw"

2.6. Decodificando con el codificador G.729 de VoiceAge.

Contamos con las tramas "ida.raw" y "vuelta.raw", las cuales se encuentran hasta el momento codificadas en G.729, para poder rescatar la señal de la voz, utilizamos la herramienta de descompresión del codificador G.79 de VoiceAge, desde el terminal de comandos de Windows "cmd.exe".

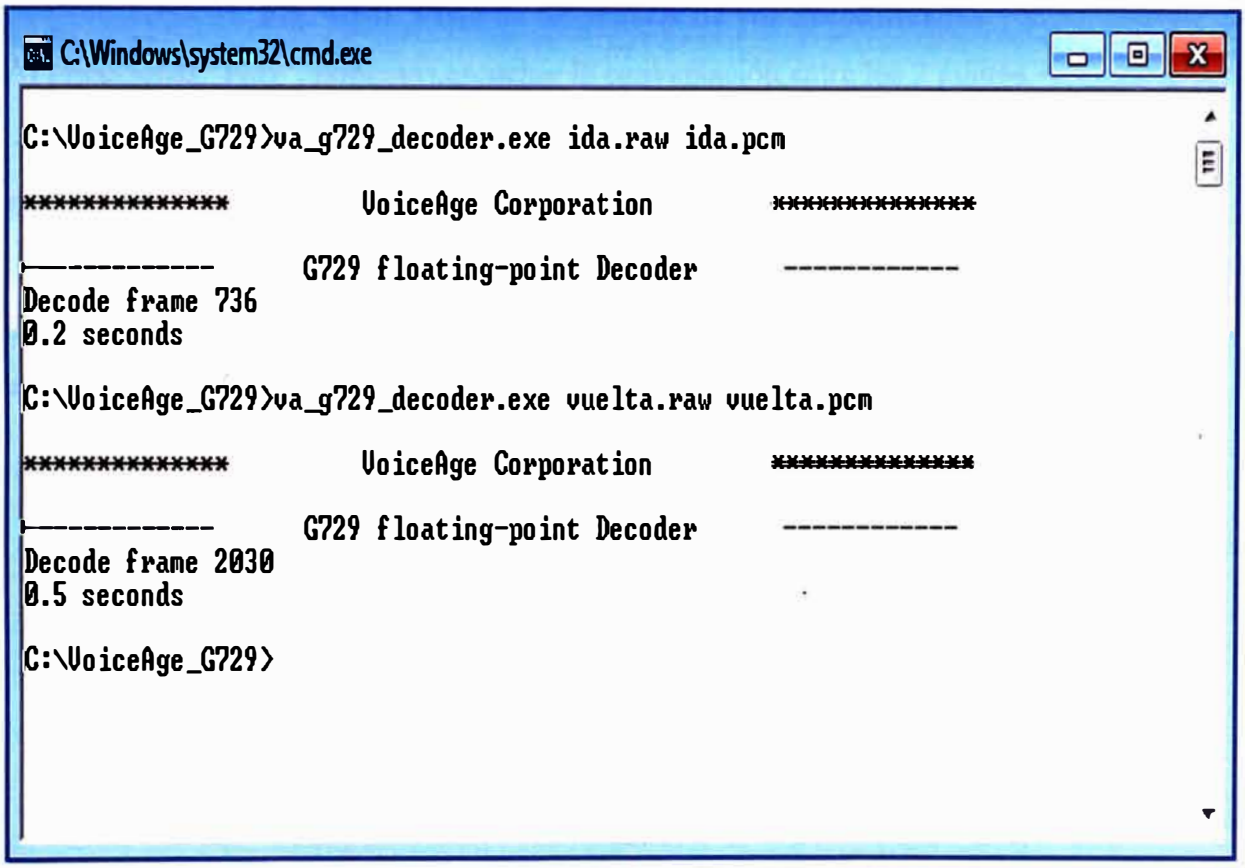

Fig. 4.13: Vista del procedimiento y resultados de decodificar las tramas *.raw

Con este procedimiento hemos decodificado las tramas y las hemos almacenado en sus correspondientes archivos "ida.pcm" y "vuelta.pcm"

2.7. Reproduciendo las señales de voz con el reproductor Audacity.

Ahora que contamos con los archivos decodificados hacemos uso de la herramienta "Audacity" para reproducir ambas tramas en conjunto y rescatar de esta forma la conversación.

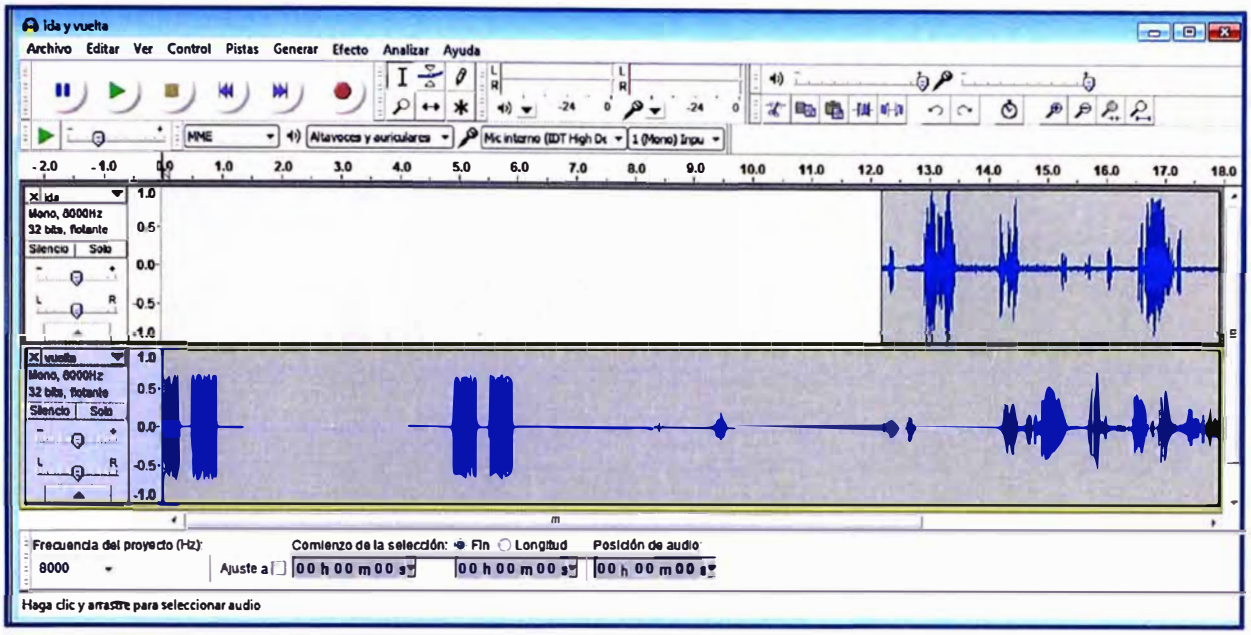

Fig. 4.14: Vista de las señales de voz decodificadas

Al seleccionar "Play", podemos escuchar la conversación entre los 2 puntos.

----------"ey"—"¿q tal se escucha?"------"¿ahora?"----------

Timbrado--- Timbrado-----"aló"----"aló"-------"se escucha ruidoso"----"no se escucha muy bien"

En este caso, notamos que los timbrados también estaban codificados, ya que el servidor de llamadas tiene almacenados los timbrados como archivos de sonido y el llamante desde el teléfono IP escucha esos timbrados luego de que el terminal IP lo decodifica.

ANEXO A
G729 ANEXO A (G.729A)

Codificador de Voz Mediante Predicción Lineal Con Excitación Por Código Algebraico de Estructura Conjugada a 8 kbit/s de Complejidad Reducida:

El Anexo A a la Recomendación UIT-T G.729, ha sido preparado por la Comisión de Estudio 15 (1993-1996) del UIT-T y fue aprobado por el procedimiento de la Resolución N.° 1 de la CMNT el 8 de noviembre de 1996.

Es una versión de complejidad reducida del códec de señales vocales G.729. Esta versión puede interfuncionar en trenes de bits con la versión completa, es decir, puede utilizarse un codificador de complejidad reducida con una realización completa del decodificador y viceversa. No obstante, se debe tener en cuenta que la calidad de funcionamiento de este códec puede no ser tan buena como la realización completa de la Recomendación G.729 en ciertas circunstancias.

La versión de complejidad reducida del códec G.729 ha sido preparada para aplicaciones de voz y datos simultáneos en multimedios, aunque la utilización del códec no se limita a tales aplicaciones.

El codificador incluye la funcionalidad de Detección de Actividad de Voz y Confort Noise Generation (VAD / CNG) y el decodificador es capaz de aceptar marcos de silencio. El rendimiento de este códec puede no ser tan bueno como el G729 en ciertas circunstancias, pero es uno de los más utilizados para la Voz sobre IP (VoIP)

Las características principales del G.729a son las siguientes:

- Tasa de bits 8kbps
- Frecuencia de Muestreo 8kHz
- Duración de la trama 10ms
- Retardo por pre-análisis 5ms
- Muestras por trama 80

- Tamaño de la trama andes de decodificar 160 bytes
- Tamaño de la trama después de decodificar 10 bytes

La complejidad del algoritmo tiene una clasificación de 15, usando una escala relativa donde G.711 es 1 y G.723.1 es 25.

ANEXO B

G729 ANEXO B (G.729B)

ANEXO B

G729 ANEXO B (G.729B)

Esquema de Compresión de Silencios para la Recomendación G.729, Optimizado para Terminales Conformes a la Recomendación V.70

El Anexo B a la Recomendación UIT-T G.729, ha sido preparado por la Comisión de Estudio 15 (1993-1996) del UIT-T y fue aprobado por el procedimiento de la Resolución N.° 1 de la CMNT el 8 de julio de 1996.

Define un detector de actividad vocal y un generador de ruido para la comodidad de la audición, para uso con la Recomendación G.729, o con el Anexo A a la Recomendación G.729, optimizado para aplicaciones DSVD de la Recomendación V.70.

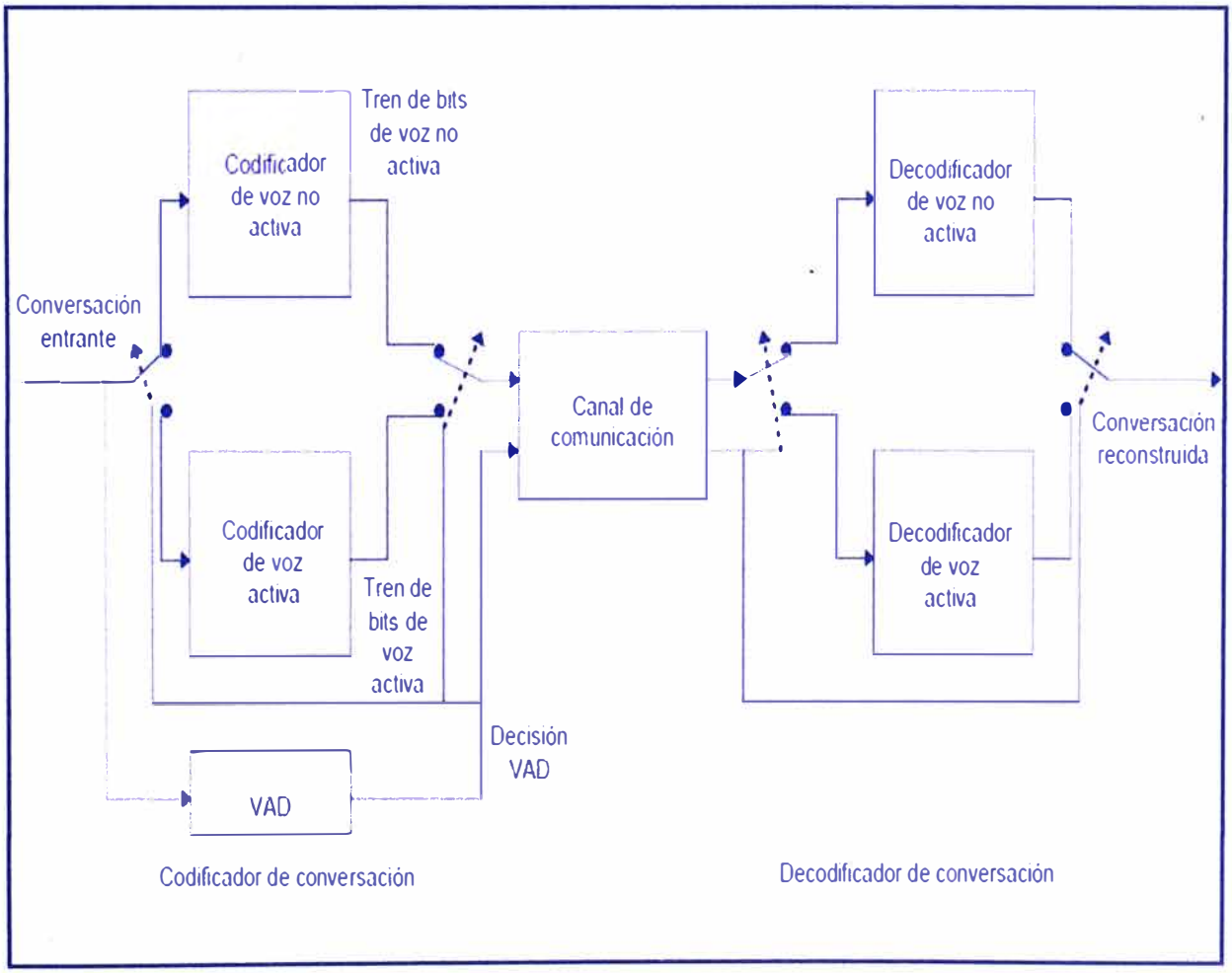

Fig. AnexoB.1: Sistema de codificación de conversación con detección de actividad vocal

En este anexo se añaden a la versión original algoritmos de detección de actividad vocal (VAD, voice activity detection), transmisión discontinua (DTX, discontinuous transmission) y generador de ruido de comodidad de la audición (CNG, comfort noise generator). Estos algoritmos se utilizan para reducir la velocidad de transmisión durante los periodos de silencio en la conversación. Se han concebido y optimizado para funcionar en el contexto de la Recomendación V.70. La Recomendación V.70 prescribe la utilización de los métodos de codificación de la conversación del Anexo A/G. No obstante, cuando sea conveniente, se puede también emplear el texto íntegro de dicha Recomendación para mejorar la calidad de la conversación. Los algoritmos están adaptados para funcionar tanto cuando se emplea el texto íntegro de la Recomendación G.729, como sólo su Anexo A/G.729. En la Figura B.1 se presenta un diagrama de bloques de un sistema de comunicación de la conversación con compresión de silencio.

ANEXO C
ESTÁNDAR G.729.1

ANEXO C
ESTÁNDAR G.729.1

Codificador de velocidad binaria variable incorporado basado en el codificador G.729: Tren de bits de codificador de banda ancha escalable a 8-32 kbit/s interoperable con G.729.

La Recomendación UIT-T G.729.1 fue aprobada el 29 de mayo de 2006 por la Comisión de Estudio 16 (2005-2008) del UIT-T por el procedimiento de la Recomendación UIT-T A.8.

UIT-T estandarizó el códec G.729.1 desarrollado con el fin de migrar sin problemas la telefonía de banda estrecha y de voz sobre IP (VoIP) a la calidad de banda ancha (50 a 7.000 Hz).

La actividad de estandarización se puso en marcha en enero de 2004, en ese momento se conocía como G.729EVG – G.729 basado en codificación de tasa de bits variable embebida. Cinco codificadores fueron parte de la fase de calificación G.729EV que terminó en julio de 2005. Los seleccionados luego se fusionaron en un solo codificador en la fase de optimización / caracterización de G.729EV. El codificador resultante fue aprobado en mayo de 2006 bajo los nombres de la UIT-T G.729.1 y de anexo J de G.729.

Uno de los principales impulsadores del desarrollo G.729.1 ha sido la evolución del mercado de las telecomunicaciones, para favorecer el acceso a Internet de banda ancha con servicios integrados (por ejemplo, se ofrecen servicios 3en1 incluyendo VoIP de banda estrecha, Internet y TV). En este contexto, un factor diferenciador es ofrecer una mejor calidad de audio mediante telefonía de banda ancha y VoIP. Sin embargo, este nuevo servicio implica importantes mejoras de los terminales e infraestructuras existentes.

G.729.1 ha sido diseñado para responder a esta necesidad del mercado al tiempo que permite una transición fluida de banda estrecha de calidad PSTN (300-3400 Hz) a la calidad de banda ancha (50 a 7.000 Hz).

En concreto, el codificador es una extensión integrada del estándar de codificación de voz ampliamente utilizado ITU-T G.729.

G.729.1 es un codificador de voz de ancho de banda escalable que opera a tasas de transmisión de 8 a 32 kbit/s.

Es compatible con señales de entrada y de salida muestreadas a 8 y 16 kHz.

El flujo de bits consta de 12 capas embebidas con una capa de núcleo interoperable con UIT-T G.729.

Se estructura en capas jerarquicas que vienen definidas por una cierta tasa de codificación. La primera capa corresponde a una tasa de 8 kbps y viene dado por la especificación G.729. La segunda capa (12 kbps) realza el rendimiento en banda estrecha de la primera. La tercera capa (14 kbps) se trata de una extensión en ancho de banda acústico (8 khz), mientras que el resto de capas se destinan al realce de ésta.

G.729.1 opera con una "longitud de la trama" de 20 ms, sin embargo, para ser coherentes con G.729 con tramas de 10ms y subtramas de 5ms, se hace referencia a estas tramas como supertramas.

ANEXO D
SOFTWARE WIRESHARK

ANEXO D

SOFTWARE WIRESHARK

Wireshark, antes conocido como Ethereal, es un analizador de protocolos utilizado para realizar análisis y solucionar problemas en redes de comunicaciones, para desarrollo de software y protocolos, y como una herramienta didáctica para educación. Cuenta con todas las características estándar de un analizador de protocolos.

La funcionalidad que provee es similar a la de tcpdump, pero añade una interfaz gráfica y muchas opciones de organización y filtrado de información. Así, permite ver todo el tráfico que pasa a través de una red (usualmente una red Ethernet, aunque es compatible con algunas otras) estableciendo la configuración en modo promiscuo. También incluye una versión basada en texto llamada tshark.

Permite examinar datos de una red viva o de un archivo de captura salvado en disco. Se puede analizar la información capturada, a través de los detalles y sumarios por cada paquete.

Wireshark incluye un completo lenguaje para filtrar lo que queremos ver y la habilidad de mostrar el flujo reconstruido de una sesión de TCP.

Wireshark es software libre, y se ejecuta sobre la mayoría de sistemas operativos Unix y compatibles, incluyendo Linux, Solaris, FreeBSD, NetBSD, OpenBSD, y Mac OS X, así como en Microsoft Windows.

2.7.1. Aspectos importantes de Wireshark

- Mantenido bajo la licencia GPL
- Trabaja tanto en modo promiscuo como en modo no promiscuo.

- Puede capturar datos de la red o leer datos almacenados en un archivo.
- Basado en la librería pcap.
- Tiene una interfaz muy flexible.
- Gran capacidad de filtrado.
- Admite el formato estándar de archivos tcpdump.
- Reconstrucción de sesiones TCP
- Se ejecuta en más de 20 plataformas.
- Es compatible con más de 480 protocolos.
- Puede leer archivos de captura de más de 20 productos.

2.7.2. Seguridad

Para capturar paquetes directamente de la interfaz de red, generalmente se necesitan permisos de ejecución especiales. Es por esta razón que Wireshark es ejecutado con permisos de Superusuario. Tomando en cuenta la gran cantidad de analizadores de protocolo que posee, los cuales son ejecutados cuando un paquete llega a la interfaz, el riesgo de un error en el código del analizador podría poner en riesgo la seguridad del sistema (como por ejemplo permitir la ejecución de código externo). Por ésta razón el equipo de desarrolladores de OpenBSD decidió quitar Ethereal antes del lanzamiento de la versión 3.6.1

Una alternativa es ejecutar tcpdump o dumpcap que viene en la distribución de Wireshark en modo Superusuario, para capturar los paquetes desde la interfaz de red y almacenarlos en el disco, para después analizarlos ejecutando Wireshark con menores privilegios y leyendo el archivo con los paquetes para su posterior análisis.

ANEXO E
CODIFICADOR G.729 DE VOICEAGE

ANEXO E

CODIFICADOR G.729 DE VOICEAGE

VoiceAge G.729 es un codificador implementado por la Corporación canadiense VoiceAge en el año 2004 en su iniciativa Open G.729.

Este codificador, codifica/decodifica señales de voz a 8kbps usando un algoritmo de Predicción Lineal con Excitación por Código Algebraico de Estructura Conjugada (CS-ACELP). Es una versión de complejidad reducida del códec G.729 y es interoperable con tramas de bits de la versión G.79 completa.

El codificador opera en tramas de voz de 10ms, correspondientes a 80 muestras en una tasa de muestreo de 8000 muestras/s. En adición a los 10ms de duración de la trama de voz, también considera un retardo de pre-análisis de 5ms, resultando un retardo algorítmico total de 15ms.

Es una versión completamente abierta y descargable desde la página web de la corporación www.voiceage.com, el paquete de la descarga contiene:

va_td_g729.pdf: Documento de la aplicación.

va_g729a.lib: Biblioteca estática Win32 para Pentium y procesadores compatibles.

va_g729a.h: APIs prototipos y declaraciones de constantes para los *.c.

va_g729a_encoder.c: Aplicación que realiza las llamadas a las APIs del códec para codificar un archivo de voz.

va_g729a_decoder.c: Aplicación que realiza las llamadas a las APIs del códec para decodificar un archivo de voz.

va_g729a_encoder.exe: Programa ejecutable codificador para la plataforma Win32.

va_g729a_decoder.exe: Programa ejecutable decodificador para la plataforma Win32.

ANEXO F

REPRODUCTOR AUDACITY

ANEXO F

REPRODUCTOR AUDACITY

Audacity es una aplicación informática multiplataforma libre, que se puede usar para grabación y edición de audio, fácil de usar, distribuido bajo la licencia GPL.

Es el editor de audio más difundido en los sistemas GNU/Linux.

2.7.3. Historia

Fue creado en otoño de 1999 por Dominic Mazzoni y Roger Dannenberg en la universidad de Carnegie Mellon. Tras lo cual fue publicado en SourceForge.net como software libre en mayo de 2000.

2.7.4. Características

- Grabación de audio en tiempo real.
- Edición archivos de audio tipo Ogg Vorbis, MP3, WAV, AIFF, AU , LOF y WMP.
- Conversión entre formatos de audio tipo.
- Importación de archivos de formato MIDI, RAW y MP3.
- Edición de pistas múltiples.
- Agregar efectos al sonido (eco, inversión, tono, etc).
- Posibilidad de usar plug-ins para aumentar su funcionalidad.

CONCLUSIONES

- El aumento de la necesidad de transmisión de la mayor cantidad de información utilizando la menor cantidad de recursos del medio de transmisión, ha sido el principal motivo para impulsar el desarrollo de las tecnologías que brinden solución a estos.
- El códec G.729, nace del requerimiento de optimizar tecnologías y modelos que ya existían y resolvían la necesidad de enviar voz codificada a través de medios de transmisión digitales, pero de una forma poco eficiente.
- A pesar de haberse concebido el modelo CS-ACELP muchos años antes del lanzamiento de la primera versión del códec G.729, se tuvo que esperar hasta 1996 para que la tecnología de hardware esté a la par con la tecnología de software y pueda soportar la cantidad de procesos que se requieren para poder implementar la compleja estructura de este códec.
- El códec G.729 adopta una estrategia para la Búsqueda de Libro de Códigos Adaptivo que intenta ser computacionalmente eficiente sobre algún sacrificio en la precisión global, especialmente durante períodos cortos de pitch ($t + f < 40$). En la práctica, sin embargo, éste método debido a la compleja combinación de sus componentes, produce voz una sintética de alta calidad.
- Para el diseño eficiente de un sistema de codificación de voz es necesario conocer el funcionamiento de los sistemas fonatorio y auditivo del ser humano. Siendo en la actualidad muchos de estándares de codificación de voz diseñados para tomar ventaja de las propiedades de estos sistemas.

BIBLIOGRAFÍA

[1] La Voz Humana

Federico Miyara – 2002

[2] Así Funciona la Conversión Analógico Digital

José Antonio García Álvarez

[3] Recomendación UIT-T G.729 "Codificación de la voz a 8 kbit/s mediante predicción lineal con excitación por código algebraico de estructura conjugada (CS-ACELP)"

ITU-T – 1996

[4] Recomendación UIT-T G.711 "Modulación por impulsos codificados (PCM) de frecuencias vocales"

ITU-T – 1988

[5] Recomendación UIT-T G.729 – Anexo A "Codificador de la voz mediante predicción lineal con excitación por código algebraico de estructura conjugada a 8 kbit/s de complejidad reducida"

ITU-T – 1996

[6] Recomendación UIT-T G.729 – Anexo B "Esquema de compresión de silencios para la Recomendación G.729, optimizado para terminales conformes a la Recomendación V.70"

ITU-T – 1996

[7] Speech Coding Algorithms - Foundation and Evolution of Standardized Coders

Wai C. Chu – 2003

[8] Ampliación de señales Aleatorias – Temas (1-4)

Javier Ortega García – Universidad Autónoma de Madrid – 2007