UNIVERSIDAD AUTÓNOMA METROPOLITANA

UNIDAD IZTAPALAPA - DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍA

DISEÑO DE UN CODIFICADOR DECODIFICADOR DE VOZ Y AUDIO BAJO UN ESQUEMA UNIFICADO

Idónea Comunicación de Resultados que presenta

Daniel Edgar Sauceda Peña

Para obtener el grado de

Maestro en Ciencias

(Ciencias y Tecnologías de la Información)

Asesor: Dr. Alfonso Prieto Guerrero

Jurado Calificador:

Presidente: Dr. Sergio Suárez Guerra IPN

Secretario: Dr. Alfonso Prieto Guerrero UAM - I

Vocal: Dr. John Goddard-Close UAM - I

México, D.F. Julio de 2013

Agradecimientos

A la Universidad Autónoma Metropolitana A mi Asesor

Contenido

Resumen

0.1. Resumen

Las tecnologías para la codificación de señales acústicas han seguido dos paradigmas fundamentales, para las señales de voz se ha utilizado la codificación basada en la predicción lineal, mientras que para el audio (dominado por la música) se ha utilizado el enfoque basado en transformadas. Por cuestiones de practicidad, y con la finalidad de hacer eficiente la codificación de señales acústicas, se ha planteado la necesidad de un codificador unificado capaz de obtener las tasas de bit más bajas posibles para secuencias de audio con contenido mixto. Dicho codificador debe tener la capacidad de distinguir los segmentos de voz o audio y utilizar el esquema de codificación más adecuado.

Uno de los principales retos en el diseño de un codificador unificado radica en la distinción entre las señales acústicas de entrada. Esta etapa es de muy alta importancia ya que determinará el tipo de tecnología de codificación que se utilizará.

El trabajo que aquí se presenta plantea el desarrollo de un módulo de decisión cuyo objetivo es determinar si una señal de entrada al sistema tiene las características que la definen como una señal acústica de voz o de audio mediante una técnica innovadora basada en la transformada continua wavelet y la identificación de cordilleras.

El presente proyecto consta de tres partes fundamentales: una investigación sobre las tecnologías actuales, el desarrollo del código y una fase de pruebas. En primer término, se realizó una investigación bibliográfica donde se consultaron las publicaciones relacionadas con el tema de clasificación de señales acústicas, reconocimiento de patrones y particularmente sobre los modelos de codificación unificada. Durante la etapa de desarrollo del código, se analizó el comportamiento de varios segmentos de señales acústicas, se incluyeron segmentos de habla, voz cantada a capella, instrumentos solos, conjuntos de instrumentos y sonidos ambientales como lluvia. La etapa de análisis consistió en observar el comportamiento de las señales acústicas al aplicar la transformada continua wavelet. El desarrollo del proyecto mostró que es posible distinguir entre las señales de voz y audio mediante el análisis asistido por la transformada continua wavelet y la detección de cordilleras asociadas a esta transformada. Al identificar características propias a cada tipo de señal acústica fue posible establecer

un algoritmo de programación enfocado a distinguirlas de manera automática.

Motivación

La demanda por servicios de difusión en línea (streaming) tiene un incremento exponencial, esto es una consecuencia de la penetración de los dispositivos tecnológicos en la mayoría de la población. Los proveedores de servicios de Internet establecen cuotas altas y topes en la cantidad de información que cada usuario tiene derecho a descargar. El contenido multimedia representa la aplicación que más ancho de banda demanda. Derivado de las condiciones descritas, se establece como una necesidad fundamental hacer un uso eficiente del ancho de banda disponible. Las tecnologías enfocadas en la codificación multimedia, especialmente el audio, han desarrollado nuevas técnicas para reducir la tasa de bit. Originalmente la voz y el audio (generalmente representado por la música) han seguido dos paradigmas de codificación distintos, ambos con una alta eficiencia en la compresión de los datos. Ante la posibilidad de contar con dos tecnologías que permiten hacer el uso más eficiente del canal de comunicación, se plantea la necesidad de un codificador-decodificador capaz de adaptarse a la señal de entrada de una forma automática y así, garantizar siempre la tasa de bit más baja en contenidos mixtos, es decir, que a veces transmitan solo voz y en otras ocasiones solo audio. El paradigma que se plantea se le conoce como USAC (por sus siglas en inglés).

Objetivo

Desarrollar un algoritmo con base en la transformada wavelet capaz de clasificar las señales acústicas en dos tipos, voz y audio como parte de un codificador unificado.

Justificación

La eficiencia en el uso del ancho de banda disponible para la transmisión de tramas de audio constituye un objetivo primordial en la industria. Se puede utilizar un clasificador como un módulo de decisión previo a la codificación de señales acústicas. Si se conoce el tipo de señal a priori se puede elegir la tecnología de codificación más adecuada para la señal de entrada y así obtener una tasa de bit más baja.

0.2. Abstract

The coding technologies for acoustic signal codification have followed two main paradigms, for voice signals the coding is based on linear prediction, for audio (mainly music) the coding techniques are based on transforms. For practical purposes, and aiming to increase the efficiency in the acoustic signal codification, there is a need for a unified coded capable to obtain the lowest possible bitrates for acoustic sequences with mixed content. This codec must have the ability to distinguish between voice or audio segments in order to define which

coding scheme is the most suitable to use.

One of the main challenges in the design of a unified codec lays in the distinction between the acoustic input signals. This stage has a very high importance because it will determine the type of coding technology that will be used.

The research work, which is presented in this document, establishes the development of a decision module that determines whether an input acoustic signal has the characteristics that define it as voice or audio through an innovative technique based on the continuous wavelet transform and the identification of ridges.

This project has three fundamental parts: an investigation regarding actual technologies, the code development and an evaluation phase. In first place, a bibliographic investigation was done; several publications were consulted with topics related to acoustic signal classification, pattern recognition and particularly the models for unified coding. During the code development stage, the behavior of several acoustic signals was analyzed, includes spoken voice, a capella singed voice, single instruments, sets of instruments and ambient sounds like rain. The analysis phase consisted in the observation of the behavior of the acoustic signals after applying the continuous wavelet transform. The development of the project showed that it is possible to distinguish between voice and audio signals aided by the continuous wavelet transform and the ridge detection associated to this transform. After identifying specific characteristics to each type of acoustic signal it was possible to program an algorithm focused on the automatic distinction.

Motivation

The demand for streaming services has shown an exponential grow as due to the high penetration rates of latest technology devices among population. Internet service provides establish high access tariffs and bit caps. Multimedia content represents the most bandwidth demanding application. A consequence of the described conditions, there's a basic need to make the most of the available bandwidth. The technologies focused in multimedia coding, especially audio, have developed new techniques for the bit rate reduction. Originally voice and audio (generally represented by music) have followed different coding paradigms, both with a high efficiency in the data compression. Since it is possible to use two technologies that allow to make the most efficient use of the communication channel, the design of a codec able to adapt automatically to the input signal is established, by this means, the lowest bit rate in mixed acoustic content is guaranteed. The stated paradigm is known as USAC (Unified Speech Audio Coding).

Objective

Develop an algorithm based on the wavelet transform able to classify the acoustic signals in two categories, voice and audio as part of a unified codec.

Justification

One of the main concerns in the telecom industry is the efficient use of the available bandwidth. A classifier can be used as a decision module previous to the coding of acoustic signals. If the type of signal is known a priori, the most appropriate coding technology can be chosen for the input signal and there for the lowest bit rate can be obtained.

Capítulo 1
Introducción

En la última década se ha incrementado de manera exponencial el uso de dispositivos de comunicación móvil, los cuales han disminuido notablemente sus costos e incrementado su poder de procesamiento. Lo anterior se traduce en una amplia variedad de servicios disponibles, entre los que se pueden mencionar: la difusión inalámbrica de audio y video bajo demanda, el envío de mensajes multimedia así como la difusión en redes celulares. El incremento en el número de usuarios establece la necesidad de utilizar el ancho de banda disponible de la manera más eficiente. La estrategia para lograr dicho objetivo radica en la codificación de las señales, ya que de esta manera se logran reducir las tasas de bit sin sacrificar la calidad.

Las señales acústicas se clasifican en dos grupos principales, señales de voz y señales de audio. Dicha clasificación es esencial para el procesamiento de las mismas, ya que cada tipo involucra un esquema de codificación particular. El paradigma de la codificación de señales de voz radica en modelar la producción de la voz mediante un método predictivo, por otro lado el paradigma de la codificación de señales de audio se enfoca en un modelo psicoacústico. Hasta años recientes la codificación de estos dos tipos de señales se realizaba de manera separada. Actualmente se ha desarrollado un nuevo paradigma que propone un esquema de codificación unificada que iguale las tasas de bit bajas de los codificadores del estado del arte tanto para voz como para audio, es decir, una tecnología que garantice la máxima compresión conservando la calidad independientemente del tipo de señal procesada. La idea central radica en el uso de dos codificadores existentes que provean el mejor desempeño, uno para voz y otro para audio y conmutar entre uno y otro de acuerdo a la señal procesada.

La distinción entre los tipos de señales es el objetivo del trabajo de investigación que aquí se reporta. En específico, el método de distinción propuesto radica en el análisis de un segmento de señal que involucra la identificación de característica propias de la voz o audio mediante el uso de la transformada wavelet. Con base en esta regla de decisión es posible plantear un codificador que determine qué codificador-decodificador (códec) será el más adecuado para la señal analizada.

El uso de un códec unificado provee ventajas para el uso más eficiente del ancho de banda disponible para alguna transmisión en cualquier tipo de red, por ejemplo al transmitir contenidos multimedia se reducirá el ancho de banda requerido para el canal de audio si

se transmite sólo voz, permitiendo hacer uso de una porción excedente para el video. Las aplicaciones son variadas, por ejemplo la identificación de voz sobre audio, o de voz entre segmentos de audio y de sólo audio o sólo voz.

El contenido de este documento está estructurado de la siguiente manera, en el capítulo 2 se describe el estado del arte de los codificadores que actualmente tienen el mejor desempeño para voz y para audio. También se plantea una alternativa para la distinción entre los dos tipos de señales basada en la relación señal a ruido. En el capítulo 3 se profundiza en dicha técnica y se reportan las pruebas realizadas y resultados obtenidos utilizando este esquema. Posteriormente en el capítulo 4 se analiza el uso de un módulo de decisión basado en el análisis e identificación de *cordilleras de wavelets*, detallando el fundamento teórico. En el capítulo 5 se establece la forma de integrar el módulo de decisión basado en *wavelets* en un códec bajo el paradigma de codificación unificada. Finalmente en el capítulo 6 se enuncian las conclusiones y se plantea el trabajo futuro.

Capítulo 2
Codificación unificada de voz y audio: Antecedentes

2.1. Fundamentos de la discriminación de audio/voz

La discriminación de señales acústicas es materia de investigación de distintas ramas del procesamiento digital de señales. Las aplicaciones son diversas, por ejemplo se puede incluir un algoritmo discriminatorio en un software para edición de audio, o para identificar hitos en un editor de video (cuando hay aplausos). En los siguientes párrafos se describen algunas de las técnicas para la clasificación de voz/audio. Asimismo se incluye el estado del arte de los códecs de mayor uso en la industria del entretenimiento y las telecomunicaciones.

2.1.1. Resumen de técnicas para la discriminación de voz y audio

Existe una variedad de técnicas de discriminación de voz/audio basadas en diversas herramientas de análisis. En la tabla 2.1 se resumen algunos de los algoritmos existentes, la técnica de clasificación empleada y el año de su publicación.

Nombre de la publicación	Autores	Técnica de discriminación	Año de publicación
Content-based classification search and retrieval of audio [1]	Wold, Blum, Keislar, Wheaton, Fish	Vectores característicos, distancia vectorial para espacios de N dimensiones	1996
Real-time discrimination of broadcast speech music [2]	Sanders	Clasificador Gaussiano multivariable	1996

Construction and evaluation of a robust multifeature speech music discriminator [3]	Scheirer, Slaney	Estimador Gaussiano máximo a posteriori multidimensional Clasificador de vecino más cercano Esquema de partición basado en árboles k-d	1997
A fast audio classification form mpeg coded data [4]	Nakajima, Lu, Sugano, Yoneyama, Yanagihara, Kurematsu	Discriminador de Bayes para distribución Gaussiana multivariable	1999
A multimode transform predictive coder (MTPC) for speech and audio [5]	Ramprashad	Decisión basada en ganancias de filtros	1999
Wideband speech and audio coding using gammatone filter banks [6]	Ambikairajah, Epps, Lin	No presenta un módulo de decisión o discriminación	2001
Content analysis for audio classification and segmentation [7]	Lu, Zhang, Jiang	K-vecinos más cercanos (KNN) y cuantificación vectorial de pares lineales espectrales (LSP-VQ)	2002
Noise-robust pitch detection method using wavelet transform with aliasing compensation [8]	Chen, Wang	Transformada wavelet para detección de pitch	2002
Audio signal classification history and current techniques [9]	Gerhard	Panorama general de las diversas técnicas existentes: Análisis multiresolución, Cadenas de Markov escondidas, enfatiza la extracción de características	2003
Audio classification and categorization based on wavelets and support vector machine [10]	Lin, Chen, Truong, Chang	Máquinas de soporte vectorial, extracción de características mediante wavelets	2005

Audio signal classification based on optimal wavelet and support vector machine [11]	Shantha, Sugumar, Sadasivam	Máquinas de soporte vectorial, extracción de características mediante wavelets	2007
Technologies for speech and audio coding [12]	Moriya	Estándares para técnicas de codificación	2009
Dual-mode switching used for unified speech and audio codec [13]	Lu, Zhang, Dou	Árboles de decisión binarios (BDT), espacio euclidiano de 4 y 5 dimensiones	2010
Enhanced long-term predictor for unified speech and audio coding [14]	Song, Oh, Kang	Decisión basada en ganancias de filtros	2011

Cuadro 2.1: Desarrollos previos para la discriminación de señales voz/audio.

El proceso de discriminación se puede dividir en dos pasos fundamentales:

1. Extracción de características y
2. Clasificación

En [9] se presenta un análisis de las diversas técnicas que se utilizan para la extracción de características que ayudan a discriminar entre diversos patrones para realizar una clasificación exitosa. La premisa es la clasificación de señales acústicas pero es importante mencionar que la clasificación es algo subjetivo ya que puede realizarse de acuerdo a una infinidad de criterios, por ejemplo, el compositor, la fecha de creación, los instrumentos utilizados, etc. El artículo plantea 5 categorías principales:

1. Ruido
2. Sonidos naturales
3. Sonidos artificiales
4. Voz
5. Música

Señala como antecedentes a la clasificación de señales las técnicas para el análisis y caracterización:

- Transformada de Fourier
- Transformada de Fourier en corto plazo (STFT por sus siglas en inglés)
- Transformada constante - Q
- Vocoder de fase
- Análisis multiresolución

El artículo enuncia los problemas básicos en la clasificación de audio y voz:

- Análisis de escena de audio (ASA por sus siglas en inglés) ; el problema característico es el de la fiesta, en la cual existe mucho ruido y se debe filtrar sólo el sonido de la voz con quien se mantiene una conversación.
- Detección de pitch, no consiste exclusivamente en la extracción de la frecuencia fundamental (f_0) aunque existe una relación psicológica entre el pitch y la frecuencia fundamental.
- Transcripción automática compuesta por estimación espectral, detección de pitch y formación de texto.
- Reconocimiento de voz; existen diversas áreas como son: la detección, la síntesis, la identificacación de signos de puntuación y el reconocimiento del hablante.

Los elementos que caracterizan las señales acústicas pueden agruparse en diversos conjuntos, a lo largo del análisis bibliográfico se identificó que los más utilizados son los siguientes:

- Características espectrales
  - Espectrograma, es la representación tiempo-frecuencia de la potencia de la señal de audio/voz
  - Armonía, relación entre picos en el espectro distribuidos de manera uniforme (útil para diferenciar entre sonidos vocálicos y no vocálicos)
  - Punto de atenuación espectral, es la frecuencia bajo la cual se concentra la mayoría de la potencia de la señal
  - Centroide espectral (frecuencia promedio del espectro de la señal)
  - Flujo espectral (tasa de cambio de la información espectral)
  - Frecuencia fundamental (*pitch*)
  - Ubicación de los formantes
- Características psicoacústicas
  - Prosodia

- Ritmo (pulsos energéticos periódicos)

Para realizar el análisis, es fundamental la elección de un tamaño de trama (número de muestras) adecuado (generalmente basado en la estacionalidad de la señal), factor determinante para la extracción de características y tiempo de procesamiento necesario para el análisis. Se observó que pueden utilizarse un tamaño de trama fijo o variable. En algunos casos se utilizan empalmes (que también pueden ser variables).

2.1.2. Antecedentes de la clasificación de audio y voz

Como se muestra en la tabla 2.1 el desarrollo de técnicas para la discriminación o categorización de muestras de archivos sonoros no es un tema nuevo, los artículos consultados muestran un interés desde 1996 motivado por diversos factores.

En [1] se plantea el desarrollo de un algoritmo con la capacidad de clasificar y buscar en una base de datos segmentos de audio con base en sus características. La estrategia que se plantea se basa en un motor de clasificación que reduce los sonidos a un pequeño conjunto de características acústicas y perceptivas para realizar una comparación y extracción del archivo de una base de datos. Posteriormente se aplican técnicas estadísticas para clasificar los sonidos con base en los siguientes criterios:

- Similitud es el parecido a un sonido o grupo de sonidos
- Características acústicas/perceptivas es la descripción claridad y pitch
- Características subjetivas es la descripción personal de los sonidos
- Onomatopeya es la imitación de un sonido

Con base en las premisas de búsquedas se establecen las técnicas de comparación sobre un vector N dimensional conteniendo las N características extraidas:

- Correlación, es análogo a una búsqueda de texto difusa.
- Búsquedas por características acústicas específicas.
- Comparación de propiedades aurales, realiza el mapeo a diferentes regiones del espacio de N dimensiones.

Las características que se utilizan para hacer el análisis, las comparaciones y la clasificación son las siguientes:

- Potencia de la señal en decibeles
- Pitch obtenido a través de la STFT

- Ancho de banda
- Armonicidad (desviación del la línea fundamental espectral del sonido al compararla contra un espectro perfectamente armónico)
- Autocorrelación

El sistema propuesto se somete a un entrenamiento mediante el cual es posible especificar un sonido por medio de restricciones a los parámetros de los vectores de N dimensiones. La clasificación de sonidos se realiza mediante el cálculo de la distancia euclidiana entre el vector del nuevo sonido y un modelo del sistema. El objetivo principal del método analizado no es la discriminación entre voz y audio sin embargo, plantea la extracción de características que ayudan a categorizar el contenido auditivo de un segmento acústico.

En [2] se establece como objetivo principal el desarrollo de un algoritmo que permite discriminar entre voz y audio en transmisiones de radio para cambiar automáticamente de estación de radio cuando se detecta que se transmiten comerciales; se asume que en los comerciales predomina voz (ya que por lo general se agrega cierto énfasis). Las diferencias particulares entre el habla y la música que se distinguen son:

- Tonalidad; la música con multiplicidad de tonos, el habla alterna entre secuencias tonales y segmentos que asemejan ruido.
- Ancho de banda; la música tiene aproximadamente 20 kHz y el habla 8 kHz.
- Patrones de excitación; el pitch para voz generalmente existe a lo largo de 3 octavas, mientras que la música a lo largo de 6.
- Duración tonal; la duración de sonidos vocálicos en el habla es muy regular, la música exhibe una variación más amplia que no se restringe a procesos de articulación.
- Secuencias energéticas; el habla muestra patrones explosivos de alta energía seguido por condiciones de baja energía, la envolvente de la música tiene menor probabilidad de exhibir dicho comportamiento.

Considerando el hecho de que la voz produce elevaciones marcadas en los periodos fricativos, al inicio y fin de las palabras; la música no presenta incrementos abruptos al ser tonalmente larga (salvo excepciones), uno de los métodos más robustos para detectar el habla es tasa de cruce por cero (ZCR por sus siglas en inglés) de la forma de onda en el dominio del tiempo. Para determinar la categoría de las muestras se realiza lo siguiente: Medición de ZCR en segmentos de 2.4 s, obtención de la media de ZCR, establecimiento de un umbral, contabilización de la diferencia entre el número de puntos por debajo y por encima del umbral establecido. Si se rebasa el umbral estadístico la forma de onda seguramente es habla.

Para la evaluación del algoritmo se utilizó una frecuencia de muestreo de 16 kHz, las tramas de 256 muestras en intervalos de 16 ms, se contabilizó el número de cruces por cero

y se calculó el voltaje rms de cada una de las tramas. Se utilizó un clasificador Gaussiano multivariable para separar las características espaciales y decidir el marcador de categoría (habla o música). El desempeño reportado fue de 90 %. Se considera que el algoritmo descrito es un tanto robusto y susceptible a varios errores, ya que se enfoca en características que pueden estar presentes en ambas categorías, voz y audio.

En [3] se propone el análisis mediante la combinación de características que describen propiedades conceptuales que diferencian las señales de voz y audio bajo un esquema de clasificación multidimensional. El conjunto de características que se extraen para el análisis son:

1. Energía de modulación en 4 kHz. Se obtiene dicha energía a través de los MFCC (Mel-Frequency Cepstrum Coefficients). La voz posee una mayor energía de modulación que la música en 4 kHz.

2. Porcentaje de tramas de baja energía.

3. Punto de atenuación espectral, percentil del 95 de la densidad espectral de potencia, el cual permite la distinción de los sonidos no vocálicos de los vocálicos.

4. Centroide espectral, el punto de balance de la distribución espectral de potencia; la música tiene una media espectral más alta.

5. Flujo espectral, norma-2 de la amplitud espectral de trama a trama del vector de diferencia, la música presenta una tasa de cambio más alta.

6. ZCR, correlación del centroide espectral.

7. Re-síntesis cepstral de magnitud residual, análisis cepstral y se suaviza el espectro, después se re-sintetiza y se realiza la comparación entre suavizado y no suavizado; sonidos no vocálicos se acoplan mejor que los sonidos vocálicos y la música.

8. Métrica de pulso, autocorrelación de largo plazo para determinar el ritmo en ventanas de 5 s.

9. Varianzas de punto de atenuación espectral, de centroide espectral, de flujo espectral, de tasa de cruce por cero, y de re-síntesis cepstral de magnitud residual.

Los esquemas de clasificación que el artículo examina son 4:

1. Estimador multidimiensional Gaussiano MAP.- modela cada categoría de datos por medio de un cluster en el espacio característico, estimación paramétrica (media y covarianza) con cada categoría en la fase de entrenamiento supervisada.

2. Clasificador modelo mixto Gaussiano (GMM por sus siglas en inglés) .- modela cada categoría de datos como la unión de varios clusters Gaussianos en el espacio característico, en contraste con el clasificador máximo a posteriori (MAP por sus siglas en inglés)

los clusters individuales no se representan con matrices de covarianza completas, sólo con las aproximaciones diagonales. Estima desde que categoría es lo más probable que haya partido un punto.

3. Clasificador del vecino más cercano.- coloca los puntos del conjunto de entrenamiento en el espacio característico, examina el vecindario local para determinar qué punto de entrenamiento es más cercano al punto evaluado y asigna la clasificación de acuerdo al vecino más cercano.

4. Esquema de partición basado en árboles k-d.- aproxima el k-vecino más cercano al votar sólo por aquellos puntos de entrenamiento en regiones particulares del espacio, es mucho más rápido que el esquema de vecinos más cercano.

De los resultados se observa que las Gaussianas no son un mal modelo para las características individuales, pero la distribución conjunta no se puede describir fácilmente, de manera particular, existen localidades en el espacio donde la distribución de datos es homogénea y por lo tanto no caen en las regiones de clusters principales. Reportan que la mejor clasificación se realiza con un margen de error del 5.8 % con base en un esquema de trama por trama y un error de 1.4 % al integrar segmentos largos de 2.4 s.

En [4] se realiza un análisis de segmentos de audio codificados en MPEG1 capa II a 112 kbit/s por canal con una frecuencia de muestreo de 44.1 kHz estándar muy utilizado en la industria. La premisa es disminuir el tiempo de análisis al evitar la transcodificación. La motivación del desarrollo se centra en la clasificación con fines de edición de contenido de video utilizando densidad temporal, ancho de banda y frecuencia central de las subbandas de energía. El algoritmo propuesto distingue entre 4 categorías:

1. Segmentos de silencio

2. Música

3. Habla

4. Aplausos

Para discriminar entre las 4 clasificaciones propuestas se utiliza un discriminador de Bayes para distribución Gaussiana multivariable, se utilizan los vectores de 4 dimensiones:

- Densidad de energía de subbanda
- Número promedio de subbandas
- Promedio del centroide de subbanda
- Varianza del centroide de subbanda

En [5], se toma una decisión sobre qué codificador se adapta mejor a la trama analizada con base en las ganancias de los bancos de filtros utilizados. Fue diseñado para aplicaciones de banda ancha de 5 kHz a 11 kHz. Utiliza muestras de 16 kHz y 16 bits, tramas de 20 ms y 4 ms de vista hacia adelante (*lookahead*), la trama también se puede subdividir en 4 subtramas de 5 ms.

- Primer bloque de análisis: obtención de medidas de correlación, definición de filtro residual A(z), se calculan los coeficientes de LPI así como un filtro con coeficientes LPC 1/B(z). Posteriormente se define un filtro ponderado A(z)/B(z), se realiza una interpolación de los dos filtros definidos a lo largo de 5 subtramas de 4ms, también se calculan el retraso de pitch y la ganancia.
- Segundo bloque de análisis: se utiliza el retardo presente y pasado del pitch, ganancias de predictor de largo plazo, parámetros del filtro y mediciones de la señal de entrada.
    1. **Modo 1 habla** segmentos de señales con cambios espectrales rápidos (voz, segmentos de música vocal, segmentos tonales altos de música instrumental).
    2. **Modo 3 audio** señales con un espectro con variaciones lentas y con ganancias de codificación LPC moderadas a altas. Secuencias vocales y música.
    3. **Modo 2 a,b,c transición** señales con bajas ganancias de predictor LPC y cambios rápidos de espectro se compone de tres submodos de operación, se adapta a configuraciones cercanas a los modos 1 y 3. Las señales objetivo son el habla no vocálica (unvoiced), ruido de fondo y segmentos transitorios.
        - **Modo 2a** equivalente a modo 1 predictor de largo término se utiliza sólo primeras 3 subtramas.
        - **Modo 2b** equivalente a modo 1 predictor de largo término se utiliza sólo últimas 3 subtramas.
        - **Modo 2c** no utiliza predictor de largo término

Se incluye un poco de histéresis para prevenir fluctuaciones abruptas, gracias al Modo 2 es posible regresar a los modos 1 o 3 y se evita que el proceso de decisión produzca errores críticos o cree artefactos auditivos. La estructura de múltiples modos de codificación permite una adaptación más suave entre diversas codificaciones.

En [6] no se propone un modelo discriminativo ni un módulo de decisión, realiza la codificación y síntesis de ambas señales, audio y voz, en el dominio auditivo. Se utilizan filtros de tono-gama de fase lineal para obtener una parametrización tiempo-frecuencia que abarca las bandas críticas de los trenes de pulsos, aproxima la generación de patrones neuronales percutores del nervio auditivo y conserva la información temporal presente en ambas señales. Se identifica como ventaja la facilidad en el escalamiento entre diferentes tasas de muestreo, tasa de bit y tipos de señal. Se incluye el uso de modelos de enmascaramiento simultáneo y temporal para eliminar la información redundante en la banda crítica. El banco de filtros es para una señal de 8 kHz se utilizan 21 filtros FIR con una longitud de coeficientes de

2N-1 que se obtuvieron a partir de la convolución del muestreo de la respuesta al impulso del tono-gama de longitud N = 100 con su inverso temporal. La salida de cada filtro se procesa por un rectificador de media onda y se localizan los pulsos positivos de las bandas críticas utilizando un detector de picos simple, proceso correspondiente a cabellos auditivos (movimiento de la membrana basilar en una sola dirección). Los componentes de menor potencia de estas señales de bandas críticas son inaudibles debido a bandas vecinas. El propósito de aplicar el enmascaramiento es producir una parametrización más eficiente (remueve 10 % de los pulsos) el enmascaramiento temporal posterior remueve 55 % de los pulsos. El tren de pulsos enmascarado en cada banda crítica se normaliza utilizando la media de la amplitud de los pulsos diferentes a cero en la trama. Los parámetros que se requiere codificar de las señales de voz/audio son, la ganancia de cada banda crítica, la amplitud y posición de cada pulso. Se pueden acomodar anchos de banda diferentes al incluir u omitir trenes de pulsos de distintas bandas críticas según se requiera.

En [7] se clasifica un fragmento de audio en voz, música, sonido ambiental o silencio. Se propone el uso de ventanas de 1 segundo, se afirma que el algoritmo es poco costoso en términos computacionales. Se capturan 6 características:

1. Tasa alta de cruce en cero (HZCRR por sus siglas en inglés).- más discriminativa que ZCR, tasa de tramas que superan umbral (1.5 veces el promedio de cruce en 1 segundo) de ZCR. HZCRR de voz es más alto que el de la música.

2. Tasa baja de energía en tiempo corto (LSTER por sus siglas en inglés).- utiliza la variación y no valor exacto, número de tramas cuya energía de corto plazo (STE por sus siglas en inglés) es menor que 0.5 veces el promedio en una ventana de 1 segundo. Buen discriminador de música-voz.

3. Flujo espectral (SF por sus siglas en inglés) .- Variación promedio del valor del espectro entre dos tramas adyacentes en una ventana de 1 segundo, los valores de SF en la voz son más altos que en la música. Distingue entre los tres tipos de señales.

4. Distancia divergente de pares espectrales lineales (LSP por sus siglas en inglés) .- Derivado de los coeficientes de predicción lineal, se compone de: covarianza de dos segmentos y la media; efectiva par la discriminación de voz y música, así como para discriminar hablantes.

5. Periodicidad de banda(BP por sus siglas en inglés) .- Análisis de correlación de 4 subbandas: 500-1000 Hz, 1000-2000 Hz, 2000-3000 Hz y 3000-4000 Hz. La periodicidad se representa mediante un máximo local de la función de correlación normalizada. Este parámetro es efectivo para discriminar entre música y sonidos ambientales.

6. Tasa de trama de ruido (NFR por sus siglas en inglés) .- Determina que es ruido al detectar si un pico máximo local de su función de correlación normalizada es menor que un umbral establecido. Parámetro efectivo para diferenciar entre un ambiente ruidoso y música.

El proceso de clasificación se divide en dos partes:

1. Clasificación robusta, se discrimina entre señales de voz y no-voz por algoritmo KNN y LSP-VQ

2. Clasificación de señales en música, sonido ambiental y silencio mediante un esquema basado en reglas

Finalmente se aplica un esquema de post procesamiento para reducir los errores de clasificación. El algoritmo se diseñó para señales a 8 kHz y se segmentan en ventanas de 1 s. Los segmentos de audio se dividen en 40 tramas de 25 ms sin empalme a las cuales se les aplica una expansión de 15 Hz de ancho de banda, se extrae un vector característico basado en las 40 tramas para representar la ventana. Se reporta una buena precisión bajo el esquema propuesto, 97.45 % de las muestras de voz, música y sonidos ambientales clasificados de manera correcta.

La consulta de artículos nos llevó al uso de técnicas de análisis más actuales, como la transformada Wavelet; en [8], se utiliza una función de correlación espacial modificada a un eliminador de ruido de señal basado en wavelets que mejora la detección de pitch en un ambiente ruidoso. Hace uso de un algoritmo de compensación de alias para eliminar la distorsión de alias provocada por las operaciones de diezmado y adición de muestras (downsampling, upsampling) de la transformada wavelet. Consiste en tres pasos básicos:

1. Transformada wavelet con el algoritmo de compensación de alias para descomponer una señal de voz en varias subbandas.

2. Correlación espacial modificada determinada de la aproximación de señales obtenida en la etapa anterior.

3. Extracción de la ubicación de los puntos de cierre glotal de la función de correlación modificada y aplicación de un algoritmo de corrección de pitch para aislar los errores del cierre glotal.

Utiliza señales de voz muestreadas a 8 kHz con una resolución de 8 bits, se utiliza una wavelet Daubechies de longitud 8. La transformada wavelet se implementa por medio de una estructura de banco de filtros, es un sistema lineal variante en el tiempo (LTV por sus siglas en inglés) , por lo tanto presenta efectos de alias e imagen (imaging) debido a los efectos de las operaciones de diezmado y adición de muestras. Los términos de alias no se pueden eliminar por completo debido a la imperfección en la respuesta del filtro.

En el artículo [10] se plantea una combinación de wavelets (para extracción de características) y máquinas de soporte vectorial (para la clasificación). En [10] se propone un patrón de clasificación denominado línea característica más cercana (NFL por sus siglas en inglés) que contrasta con el algoritmo de vecino más cercano, se basa en las siguientes consideraciones:

1. La muestra corresponde a un conjunto en un espacio característico.

2. Cuando un sonido cambia constantemente describe una trayectoria ligando los conjuntos de características en el espacio característico.

3. Las trayectorias que cambian de acuerdo a sonidos prototipo constituyen un subespacio que representa esa clase.

La extracción de características sigue una serie de pasos:

1. Preprocesamiento, muestreo a 8 kHz resolución de 16 bits, tramas de 256 muestras (32ms), empalme de 75 % entre tramas, filtrado de preenfásis para frecuencias altas, detección de tramas silenciosas mediante umbral de potencia empírico.

2. Extracción de características, transformada wavelet (Daubechies ortogonal, longitud 8) y la transformada de Fourier (coeficientes ceptrales y perceptivos):

   a) Potencia de 3 subbandas

   b) Frecuencia de pitch, transformada wavelet eliminación de alias

   c) Claridad, centroide de la transformada de Fourier

   d) Ancho de banda, raíz cuadrada promedio ponderado de potencia

   e) Coeficientes cepstrales de frecuencia (característica L)

3. Formación de vector característico, media y desviación estándar de cada característica, tasa de pitch y tasa de silencios, para obtener un vector característico 14 + 2L.

4. Normalización para entrenamiento y evaluación, división en dos subconjuntos, uno de entrenamiento y otro de evaluación, cada característica extraída tendrá ponderaciones similares tras el proceso de normalización.

Las máquinas de soporte vectorial utilizan una función núcleo (función base de radio exponencial (ERBF por sus siglas en inglés)) para definir el hiperplano que separa los puntos en dos clases predefinidas. El esquema de árbol binario de fondo hacia arriba sólo selecciona una clase para la clasificación de la muestra de audio (multicaso). El método basado en este esquema clasifica clases con respecto a un audio por medio de un procedimiento iterativo. En cada turno remueve la clase ganadora de la raiz del árbol y reconstruye una nueva estructura de árbol. La clase que se remueve primero es la clase con la cual la muestra tiene mayor similitud, la clase que se remueve al último es aquella con la que la muestra tiene menor similitud.

En [11] se utilizan los mismos principios que en [10], propone una técnica mejorada de formación de vectores característicos, utiliza wavelets para la descomposición y extracción de las características de las señales:

- Potencia de subbandas
- Claridad

- Ancho de banda
- Información de pitch

Longitud de trama es de 512 muestras, además se rediseña el tamaño de empalme. El proceso de formación de vectores se realiza después de la extracción de características, se utiliza una máquina de soporte vectorial fondo hacia arriba (buttom-up) iterativa para empatar (match) una selección de audio con subconjuntos o categorías de clases. Los parámetros de entrenamiento para la máquina de soporte vectorial tales como el límite superior y la varianza de la función base radial exponencial se modifican para mejorar la precisión de la clasificación. La extracción de características se realiza mediante las transformadas:

- Wavelet: potencia de 4 subbandas, claridad (punto de balance del espectro), ancho de banda (rango con mayor energía), frecuencia de pitch (periodo fundamental de forma de onda).
- Fourier: coeficientes de frecuencia cepstral, toma en cuentas las propiedades auditivas no lineales del oído humano
- Dominio temporal: razón de cambio de silencios

Los vectores de prueba se utilizan en una SVM de esquema de árbol de fondo hacia arriba (buttom-up) para concretar las evaluaciones, el algoritmo de entrenamiento es Medio vs. Medio (half vs half) con los siguientes pasos:

1. Entrenamiento con conjuntos de datos de manera individual
2. Se proporcionan datos de prueba comunes
3. Los conjuntos de datos completos se dividen en dos grupos iguales
4. Nuevamente se divide en dos grupo de pertenencia de datos de prueba
5. Proceso continúa hasta que los datos de prueba clasifican
6. Se proporcionan los datos en cualquier instante, continúan los pasos 3 a 5 hasta que se proporcionan todos los datos; finalmente se calcula la precisión.

Se concluye afirmando que la precisión en la clasificación se mejora utilizando wavelets, también establece que mientras más larga sea la trama, se incrementa la complejidad computacional.

En [13] se analiza un método de conmutación. Introduce una estrategia de pre-codificación para suavizar las transiciones entre los dos códecs. El módulo de clasificación habla/audio se coloca al inicio de codificador. El algoritmo de clasificación está basado en árboles de decisión binarios (BDT por sus siglas en inglés) :

- Espacio Euclidiano de 4-dimensiones (características de corto-plazo)

- Espacio Euclidiano de 5 dimensiones (características de largo-plazo)
- Buffer para almacenar vectores de caracterización basados en 250 tramas (para cálculo de las características de largo-plazo)

Al alimentar al sistema con una señal, primero se utiliza el árbol de decisión de corto-plazo y el vector característico se determina por trama, una vez que se satura el buffer se activa el árbol de decisión de largo-plazo. La etiqueta de clasificación se exporta al módulo de conmutación dual y se añade a la trama de bit.

- **Conmutación de audio a voz:** La continuidad en las tramas de voz depende de la memoria del filtro de síntesis y la excitación adaptativa en ACELP, al encontrar una discontinuidad el método de codificación realiza la conmutación. La propuesta de pre-codificación considera que si la trama presente se detecta como la última trama de audio se utilizará una nueva ventana MDCT.
- **Conmutación de voz a audio:** MDCT obliga a que cada trama de audio se reconstruya con dos tramas de bits codificados de AAC; cuando se detecta una trama como la última de audio, las muestras codificadas por ACELP se confinan a sólo una segunda media trama y la codificación AAC de la trama también se aplica.

La diferencia con la conmutación de audio a voz radica en que los bits codificados de ACELP y AAC se escriben en la trama de transporte, lo cual tiene como consecuencia saltos en la trama de bits en esta trama particular. De hecho la trama final de voz se codifica y decodifica en modo AAC. Lo anterior introduce un error de codificación que nunca podrá eliminarse debido a la lógica que se ha planteado, por lo tanto no tendrá una precisión de 100 % sin embargo, es imposible determinar cuándo termina una trama de audio e inicia una de voz y viceversa, ni siquiera un humano podría determinar el instante preciso. Las tramas de bit se establecen a 15.85 kbps para AMR-WB y a 18 kbps para HE-AAC con una frecuencia de muestreo de 16 kHz. Se reporta una precisión de 95 %.

En [14] se propone un predictor de largo plazo mejorado (eLTP por sus siglas en inglés) que utiliza de manera eficiente las redundancias periódicas inter e intra tramas temporales. Se propone el uso del algoritmo eLTP como un preproceso para eliminar redundancias armónicas provocadas por la periodicidad de la señal. Las actualizaciones en el desarrollo del códec USAC incluyen algunas nuevas tecnologías como: modelado de ruido en el dominio de la frecuencia (FDNS por sus siglas en inglés) y cancelación de alias adelante (FAC por sus siglas en inglés) , ambas herramientas fueron introducidas con el propósito de suavizar las transiciones entre las dos ramas de codificación, las características de cada rama son diferentes:

- Codificación con predicción lineal en el dominio residual sin empalme de tramas adyacentes
- Codificación MDCT en el dominio frecuencial con empalme de 50 %

Dado lo anterior, una transición suave es un problema crítico para garantizar el mérito de la arquitectura USAC, pero aún así lo óptimo es reducir las transiciones.

Los parámetros que se transmiten al módulo decodificador eLTP son, el retardo de pitch y las ganancias de pitch por cada sub bloque (256 muestras, alrededor de 20 ms dependiendo de la frecuencia de muestreo del codificador). El umbral de decisión para utilizar el algoritmo eLTP está determinado por la ganancia de pitch, si esta excede 0.8, se utilizará. A pesar de que este valor pareciera ser alto, en los experimentos reportados se observó que más del 50 % de los segmentos evaluados seleccionaron el uso de eLTP.
Las discontinuidades representan un problema y pueden ser provocadas por los cambios abruptos en los parámetros de pitch de eLTP, para evitar este problema se debe mantener la ganancia de código del algoritmo propuesto eLTP, se enuncian tres alternativas:

1. Estructura basada en empalme y adición; empalme del 50 % entre sub bloques, elimina discontinuidad, pero el bloque de análisis se duplica y el desempeño del predictor se vuelve sub óptimo.

2. Tamaño de sub bloque adaptable; tamaño variable, basado en puntos óptimos (baja energía de señal en el intervalo de observación).

3. Suavizar parámetros; limita el rango de búsqueda de parámetros en lazo abierto al bloque actual comparado con el sub bloque previo.

En [12] se analizan los paradigmas de codificación que existen actualmente, se plantea la premisa de que el desarrollo de los codificadores para voz y para audio han seguido un desarrollo independiente; señala que el mercado más importante para la voz son las telecomunicaciones de dos vías, se utiliza una trama corta (5-10ms) en el dominio del tiempo mediante codificación predictiva. El mercado más importante para el audio son las transmisiones unidireccionales como la difusión (broadcasting, streaming) y descargas de audio (con mayores retardos) tramas típicamente mayores a 20ms.

2.2. Estructura básica del codificador-decodificador (códec) unificado

La parte medular de un codificador unificado radica en la distinción entre señales de voz y señales de audio. La codificación se realiza con base en los estándares tecnológicos existentes. Particularmente, en este trabajo se utilizaron los siguientes codificadores:

- Para la codificación de señales de voz, la ancho de banda multirango adaptiva (AMRWB pos sus siglas en inglés) basado en un modelo predictivo (predicción lineal algebráica de código excitado (ACELP por sus siglas en inglés)).

- Para la codificación de señales de audio, el codificación avanzada de audio de alta eficiencia versión 2 (HE-AACv2 por sus siglas en inglés) basado en un esquema de banco de filtros polifásicos (codificación avanzada de audio (AAC por sus siglas en inglés)).

La identificación errónea entre las señales afecta el desempeño del sistema general: la calidad auditiva se degrada cuando las señales de voz son codificadas con el códec HE-AACv2 (especialmente en tasas de bit bajas) y el AMR-WB no tiene un buen desempeño al codificar señales de música.

La señal a codificar debe pasar una etapa de preprocesamiento que generalmente involucra un filtrado para eliminar las componentes de altas frecuencias. Posteriormente la señal preprocesada es analizada a través de un módulo de decisión que determina si se trata de una señal de voz o una de audio. En la figura 2.1 se presenta la estructura básica de un codificador unificado.

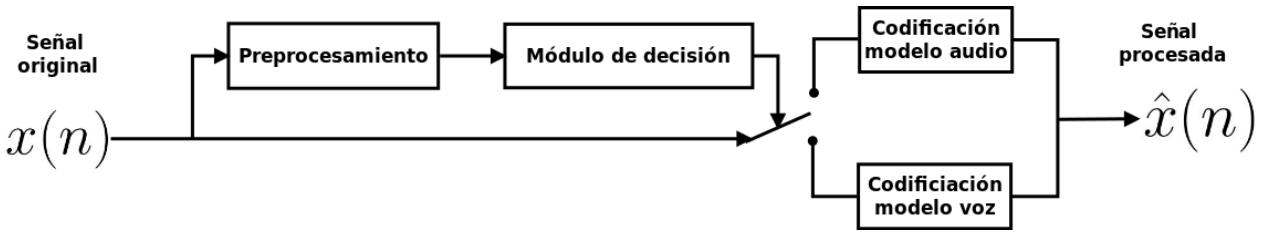


Figura 2.1: Diagrama general de codificador unificado.

El codificador unificado tiene la capacidad de conmutación entre los dos tipos de códecs líderes, de tal forma que se pueden codificar varias tramas utilizando un mismo modelo o conmutar entre ellos. En lo sucesivo nos referiremos al codificador elegido para un tipo de señal en particular como un *modo de codificación*.

El principal objetivo de este trabajo de investigación radica en el diseño de un módulo capaz de distinguir señales de voz y señales de audio. En la siguiente sección se describen los códecs empleados y se establece un estudio del estado del conocimiento de la codificación unificada.

2.3. Estándares seleccionados

Los códecs seleccionados para el desarrollo de este trabajo marcan un hito en las tecnologías de codificación que actualmente se utilizan. Por un lado está el codificador AMR-WB, que es el estándar actual para las comunicaciones por voz, tanto para comunicaciones inalámbricas (celulares, telefonía móvil) como redes alambradas (telefonía fija). El organismo regulatorio proyecto de asociación de 3ra generación (3GPP por sus siglas en inglés) ha establecido

este códec como la tecnología oficial, de ahí el sustento para el uso de dicho códec. El codificador HE-AACv2 es el códec líder en lo que a codificación de música se refiere, los reproductores de música portátiles más populares utilizan esta codificación; su uso no está estandarizado por algún organismo regulatorio internacional, pero está ampliamente difundido.

2.3.1. Codificador HE-AAC v2

Se puede entender al códec HE-AACv2 como una versión evolucionada del códec original AAC. Esta tecnología se volvió muy popular al ser la empleada en el famoso dispositivo iPod. El códec HE-AACv2 es capaz de lograr las tasas de bit más bajas sin reducir significativamente la calidad subjetiva, por lo tanto es común su uso para la distribución de contenido en redes con ancho de banda limitado. Esta nueva versión integra fundamentalmente tres tecnologías que le ayudan a disminuir aún más la tasa de bit al compararlo con la versión original, a saber:

- AAC es el códec núcleo
- La replicación espectral de banda (SBR por sus siglas en inglés) para la optimización en el uso de recursos
- El estéreo paramétrico (PS por sus siglas en inglés) permite la mezcla a monoaural para envío de menos información

A continuación se describen las tres tecnologías innovadoras (AAC, SBR y PS) que utiliza el códec HE-AACv2.

Tecnología AAC

Esta tecnología nace a partir del estándar grupo de expertos de fotografías en movimiento (MPEG por sus siglas en inglés) -2. El códec AAC es el núcleo del códec HE-AACv2. Se basa en el uso de técnicas de compresión por medio de la codificación perceptiva. Los bloques básicos de la codificación AAC se describen a continuación [15]:

1. Banco de filtros
    - Filtrado de la señal en 18 subbandas con la finalidad de identificar y eliminar redundancia. Al tener una mejor resolución en frecuencia es más sencillo rastrear el umbral de enmascaramiento y por ende controlar el error de la señal.
    - transformada modificada de coseno discreto (MDCT por sus siglas en inglés)
2. El modelo perceptivo consiste en obtener los valores del umbral de enmascaramiento o *ruido permitido*. Si el ruido de cuantificación se mantiene por debajo de este umbral, el resultado de la compesión será indistinguible de la señal original.

3. La cuantificación y codificación de los valores por medio de la codificación de Huffman, que trabaja en pares y en algunos casos en cuadruplas, se realiza modelado de ruido para mantener el ruido de cuantización por debajo del umbral. Para encontrar una ganancia y factores de escalamiento óptimos se realizan dos iteraciones anidadas en una forma de análisis por síntesis.

AAC supera la codificación de MPEG capa 3 (conocido comúnmente como mp3) debido a que utiliza nuevas herramientas para mejorar la calidad en tasas de bit bajas:

- Para obtener mayor resolución en frecuencia se utilizan hasta 1024 líneas de resolución en comparación con mp3 que tan solo usa 576.
- Predicción, opcionalmente incluye la predicción en retrospectiva con la que se obtiene mejor eficiencia de codificación especialmente para señales con tonalidades parecidas.
- La codificación estéreo conjunta mejorada presenta mayor flexibilidad.
- La codificación Huffman mejorada utiliza cuadruplas.

Existen mejoras enfocadas en la calidad acústica las cuales se mencionan a continuación:

- Conmutación de bloques mejorada, utiliza un banco de filtros MDCT estándar conmutado con una respuesta al impulso de 5.3 a una frecuencia de muestreo de 48 kHz para bloques pequeños.
- modelado temporal de ruido (TNS por sus siglas en inglés) modelado de ruido en el dominio temporal mediante una predicción de lazo abierto en el dominio de la frecuencia.

La figura 2.2 muestra el diagrama a bloques básico de un códec AAC, que incluye las herramientas previamente descritas.

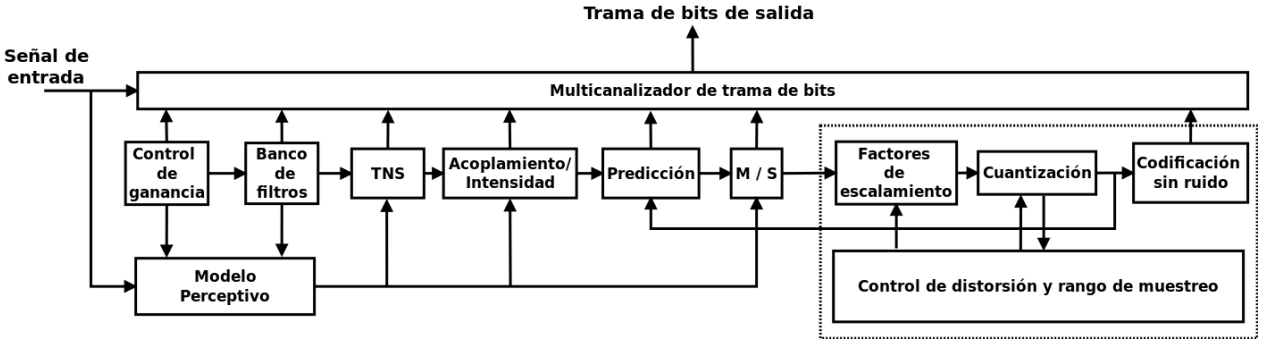
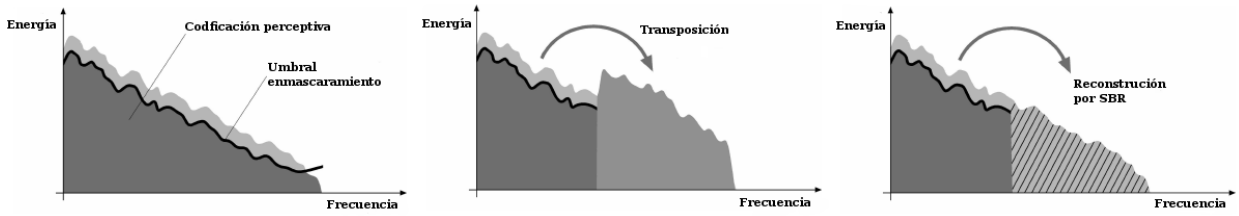


Figura 2.2: Diagrama básico de un codificador AAC [15].

Tecnología SBR

La codificación de audio perceptible fue desarrollada tomando en cuenta al sistema auditivo humano, especialmente el enmascaramiento de distorsiones perceptibles mediante señales más potentes en la vecindad espectral. El desarrollo se basa en el cálculo de dicho umbral en periodos cortos. Mientras más se disminuye la tasa de bit aumenta la potencia del ruido de cuantificación, hasta llegar al punto en el cual se rebasa el umbral establecido. Con el objetivo de reducir la tasa de bit pero evadir la percepción del ruido, se desarrolló la técnica de SBR. Principalmente consiste en limitar el ancho de banda de la señal de audio previo al proceso de codificación. De esta manera se evita la codificación de la energía de la alta frecuencia, por lo tanto la información disponible para el segmento de banda baja es mayor y consecuentemente la señal será más limpia pero "hueca". La tecnología SBR se plantea como un esquema de codificación híbrida por forma de onda y método paramétrico. La premisa de este método radica en el hecho de que existe una alta dependencia entre las partes de alta y baja frecuencia de una señal de audio, por lo tanto las frecuencias altas de una señal de audio pueden ser reconstruidas de manera eficiente a partir de la parte de frecuencias bajas. Por ende no es necesario transmitir la información correspondiente a las frecuencias altas, tan sólo se envían unos cuantos datos de control en la trama de bit para garantizar la reconstrucción óptima de las frecuencias altas. La figura 2.3 muestra de manera gráfica la traslación del contenido de frecuencias bajas a la zona de frecuencias altas.



(a) Espectro original (b) Transposición (c) Reconstrucción

Figura 2.3: Generación de frecuencias altas a partir de frecuencias bajas. Imágen tomada de [16].

La transposición no es suficiente, por lo que es necesario ajustar de acuerdo con la envolvente original, precisamente estos son algunos de los parámetros que se deben enviar al decodificador para realizar la reconstrucción óptima. El codificador SBR trabaja en paralelo con un códec núcleo, la información generada básicamente se utiliza en el post-procesamiento después de la decodificación. Los parámetros obtenidos durante la codificación son estimados a través de un filtro espejo en cuadratura (QMF por sus siglas en inglés) . La resolución en tiempo y frecuencia de las envolventes espectrales pueden ser elegidas con mucha libertad permitiendo la adaptación más adecuada para cada caso particular.

La figura 2.4 muestra el diagrama básico de un codificador SBR se distingue la operación en paralelo con un codificador núcleo, la funcionalidad principal radica en la extracción de parámetros para la reconstrucción óptima de la señal.

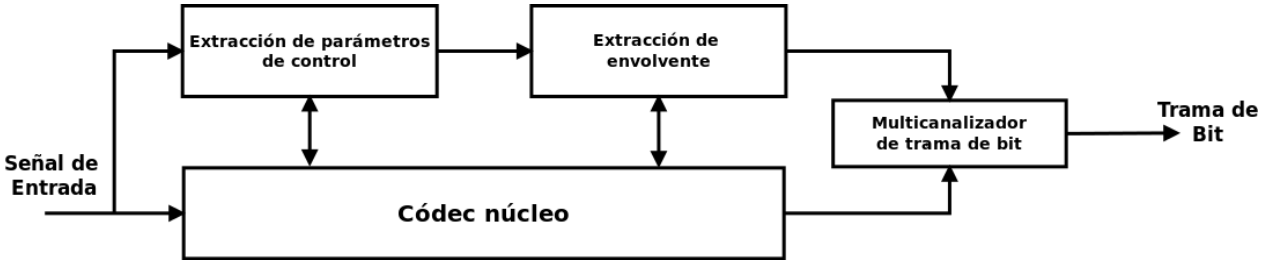


Figura 2.4: Diagrama básico de un codificador SBR [16].

Tecnología PS

La premisa de esta tecnología se enfoca en extraer información paramétrica de una señal estéreo para codificar la señal de forma monoaural y posteriormente en el decodificador reconstruir la señal estéreo a partir de la información extraída durante la etapa de codificación. En el artículo [17] se enuncian tres parámetros clave para describir una señal estéreo:

1. diferencias de intensidad inter-canal (IID por sus siglas en inglés) , descripción de las diferencias de intensidad entre los canales.

2. diferencias de fase inter-canal (IPD por sus siglas en inglés) , descripción de las diferencias de fase entre los canales.

3. coherencia inter-canal (IC por sus siglas en inglés) , descripción de la coherencia (máxima correlación cruzada como función de fase o tiempo) entre los canales.

Pero los parámetros IPD sólo especifican las diferencias de fase relativas entre los canales de la señal de entrada, no representa las distribución de las diferencias de fases sobre los canales izquierdo y derecho, por lo tanto se introduce un cuarto parámetro que describe la diferencia de fase general (OPD por sus siglas en inglés) .

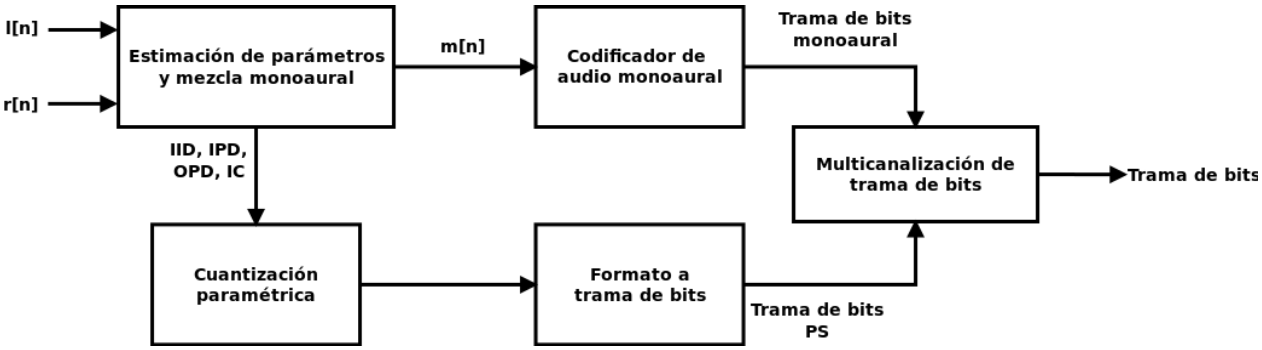


Figura 2.5: Diagrama básico de un codificador PS [17].

En la figura 2.5 se muestra el diagrama básico de un codificador PS, las entradas al sistema $l[n]$ y $r[n]$ representan el canal izquierdo y derecho de una señal estéreo respectivamente.

Las especificaciones técnicas de las tres tecnologías mencionadas están documentadas en el ISO/IEC 14496-3. El estándar del codificador HE-AACv2 es el ISO/IEC 144 ISO/IEC 14496-3:2001/Amd.4.

Finalmente, se resaltan algunas características importantes del HE-AACv2, la señal de entrada pasa por un banco de filtros QMF de 64 bandas, posteriormente existen dos casos particulares:

1. Tasa de bit de señal de entrada <36 kb/s.

   - Se utiliza el codificador PS para extraer información sobre los parámetros de la señal estéreo con base en las muestras que se obtienen del banco de filtros QMF.
   - Se realiza una mezcla de los canales estéreo a uno mono.
   - Con un sintetizador QMF de 32 bandas se transforma la representación mono al dominio del tiempo con la mitad de la frecuencia de muestreo $f_s/2$.
   - Esta señal entra al codificador AAC.

2. Tasa de bit de señal de entrada >36 kb/s.

   - No se utiliza la herramienta PS
   - La señal pasa por una etapa de submuestreo 2:1.
   - Esta señal entra al codificador AAC.

Para ambos casos se extrae la envolvente espectral por medio del uso de SBR. Se destaca el hecho de que este códec funciona en dos frecuencias diferentes, la codificación que se realiza utilizando AAC es a la mitad de la frecuencia de muestreo de aquella necesaria para la extracción de parámetros. La figura 2.6 muestra un diagrama a bloques con la funcionalidad del codificador HE-AACv2, f_s representa la frecuencia de muestreo de la señal.

2.3.2. Codificación AMR-WB

El concepto de análisis-por-síntesis se refiere a la técnica de codificación en la cual la señal de entrada se descompone o se analiza por medio de componentes que son una representación fundamental una señal. Dichos componentes cuantificados se utilizan para sintetizar una reproducción de la señal original [19].

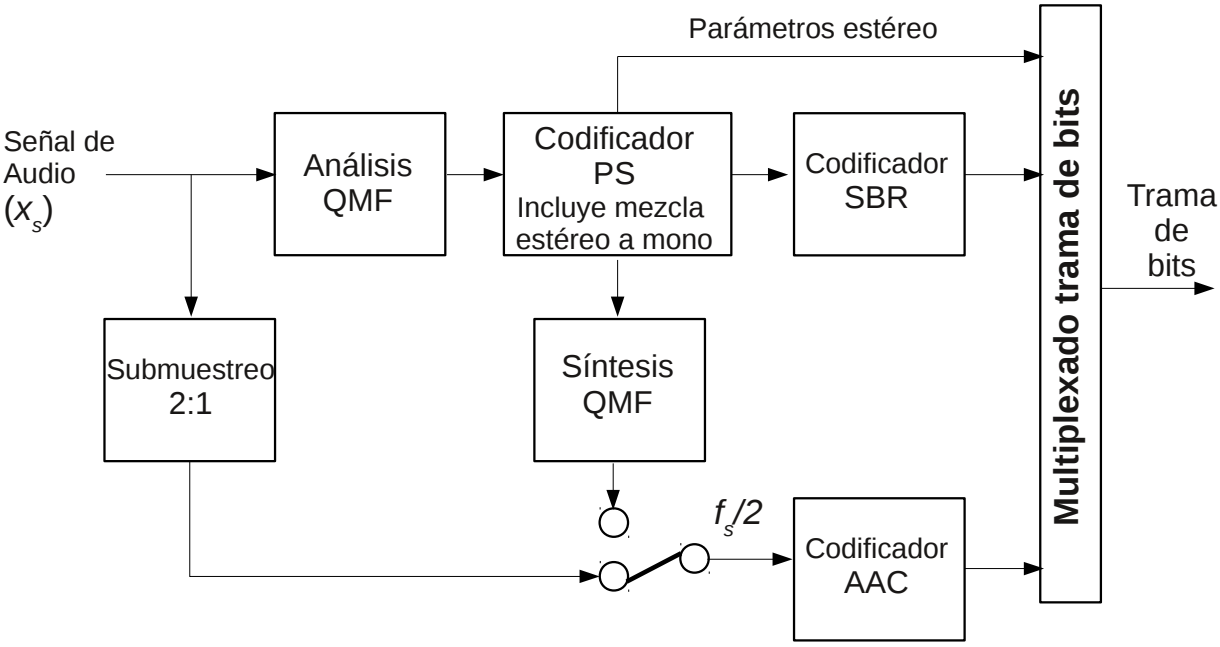


Figura 2.6: Estructura de codificador HE-AACv2 mostrada en [18].

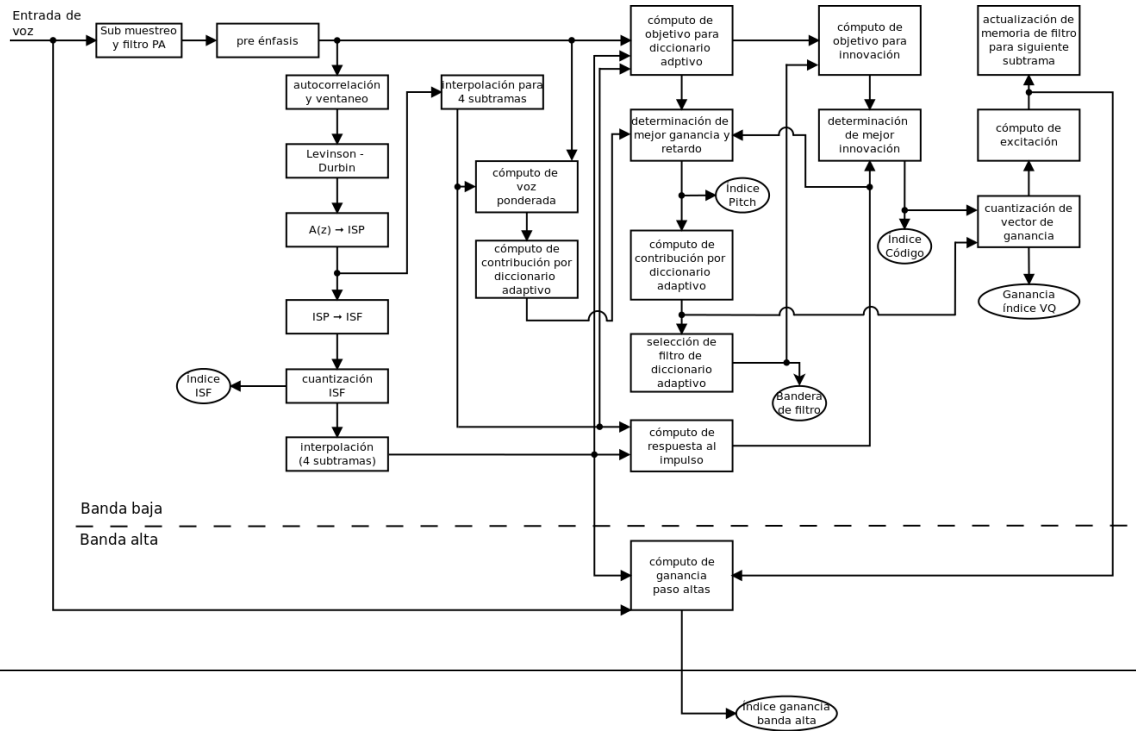


Figura 2.7: Estructura de codificador ACELP para AMR-WB mostrada en [20].

El códec AMR-WB provee un excelente desempeño en tasas de bit muy bajas, inferiores a 10 kbps y hasta 24 kbps. Está constituido por una estructura híbrida, que integra los sistemas de codificación siguientes:

- ACELP
- Excitación de transformada codificada (TCX por sus siglas en inglés) .

El núcleo permite la conmutación etre ambas tecnologías. Este modelo siempre calcula los coeficientes de predicción lineal (LPC por sus siglas en inglés) (LPC). Asumiendo que se tiene una señal de entrada estéreo el procedimiento para realizar la codificación es el siguiente:

- Se calcula una señal que representa la suma de ambos canales y otra con la diferencia de las mismas.
- Se descompone la señal suma en dos bandas S_L (banda baja, frecuencias inferiores a 6.4 kHz) y S_H (banda alta, frecuencias superiores a 6.4 kHz) submuestreada a 12.8 kHz, la frecuencia nominal del AMR-WB como se especifica en [21].
- Posteriormente se aplica el modelo híbrido ACELP/TCX a S_L.
- Finalmente, S_H se codifica utilizando la técnica de extensión de ancho de banda (BWE por sus siglas en inglés) .

Las frecuencias bajas se codifican utilizando cualquiera de los dos modos, para definir cuál utilizar se realiza un análisis de la razón señal a ruido (SNR por sus siglas en inglés) por bloques de muestras. Las frecuencias altas, consideradas arriba de 6.4 kHz, se codifican utilizando la técnica BWE, mediante la cual se extrae la envolvente espectral y las ganancias son cuantificadas y enviadas al decodificador. La señal de excitación que se utiliza para la reproducción de la franja de altas frecuencias es la misma que se utiliza para las frecuencias bajas, la cual está disponible en el decodificador [21]. La selección del modo se puede determinar por medio del uso de lazo cerrado o lazo abierto, lo que permite controlar la complejidad del codificador. TCX es una codificación basada en transformaciones que utiliza una ventana no rectangular con empalme la cual permite mejorar la ganancia de codificación. ACELP utiliza una ventana rectangular.

Codificación de la banda baja para señal monoaural

De acuerdo con la frecuencia de muestreo nominal en el códec AMR-WB (12.8 kHz) se definen una super trama de 80 ms (1024 muestras), tramas de 40 ms (512 muestras) y 20 ms (256 muestras). La señal se procesa en super tramas de 1024 muestras, conformadas por bloques de 256, 512 o 1024 muestras. Cada trama de 256 muestras puede ser codificada utilizando ACELP o TCX, mientras que una super-trama de 512 muestras o una trama de 1024 muestras se codifican utilizando TCX.

Existen 26 combinaciones distintas de modos dentro de una super-trama. En la figura 2.8 se muestran todas las posibles combinaciones.

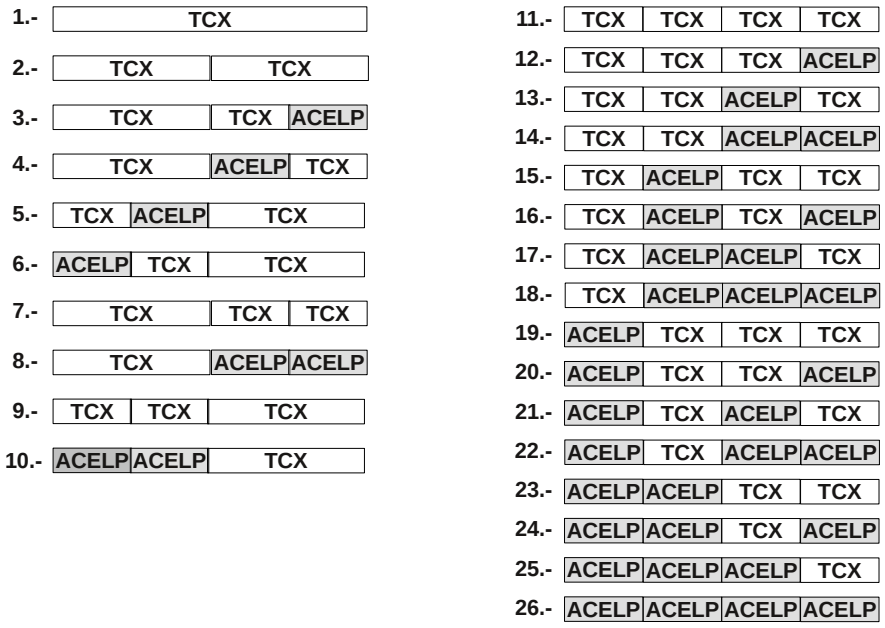


Figura 2.8: Posibles modos de codificación propuestos en [22].

La ventana utilizada por TCX tiene las siguientes características:

- Plana en la parte media (cubre la mayoría de la trama).
- Se extiende a la siguiente ventana en forma de medio coseno decreciente, empalme.
- El inicio de la ventana puede tener dos formas:
  - Plana si la ventana anterior fue ACELP.
  - Medio coseno si la ventana anterior fue TCX.

La transición entre ventanas TCX es aceptable mediante el empalme de las mismas. Para la transición de una ventana ACELP a TCX se maneja la transición calculando respuesta al impulso cero (ZIR por sus siglas en inglés) del filtro.

Codificación de la banda alta para señal monoaural:

Esta etapa de codificación, para la banda de alta frecuencia (arriba de 6.4 kHz), se realiza a través del BWE, mediante el cual se obtiene un modelo paramétrico (envolvente espectral y ganancias), cuyos parámetros son cuantificados y enviados al decodificador.

La envolvente espectral se modela por un filtro LPC de 8º orden se calcula con la versión submuestreada de S_H. Los LPC son transmitidos una vez por ventana. Las correcciones de ganancia son calculadas y transmitidas para cada subventana, esto asegura la continuidad en la unión de 6.4 kHz entre la banda baja y la banda alta. Como sólo se envían unos cuantos parámetros la tasa de bit total para BWE puede ser tan baja como 0.8 kb/s.

Codificación estéreo:

Para el caso de una señal estéreo, se realiza la misma división en dos bandas al igual que en el caso monoaural. Los dos canales son mezclados en una sola señal monoaural que se codifica utilizando el codificador núcleo de AMR-WB, la información de la imagen estéreo se codifica realizando otra descomposición de la banda de baja frecuencia en dos bandas:

- 0 – 1.0 kHz
- 1.0 – 6.4 kHz

Para la banda de muy baja frecuencia se deriva un factor de balance estéreo que representa el cociente entre la señal mono y la señal lateral. La parte de alta frecuencia de la banda baja se codifica de acuerdo a un filtro en el dominio del tiempo con ganancia noble restringida la cual se asemeja a una técnica de predicción inter-canal. La banda alta se codifica utilizando BWE paramétrica en los dos canales.

Escalabilidad de AMR-WB:

El uso de cuantificación Vectorial (VQ por sus siglas en inglés) algebráica hace que AMR-WB sea altamente escalable en términos de la tasa de bit total y la distribución de tasa de bit entre la codificación monoaural y estéreo. La escalabilidad se puede llevar más allá gracias a la capacidad de escalamiento en la frecuencia de muestreo nominal de 12.8 kHz con factores en un rango de 0.5 a 1.5; lo anterior es equivalente a escalar la tasa de bit total del codec y el ancho de banda del audio codificado. Esto permite la operación de AMR-WB en un ancho de banda limitado y también para una operación en tasas altas, hasta 48 kb/s, con un ancho de banda de audio de hasta 20 kHz.

2.4. Antecedentes de la codificación unificada

Las características espectrales de la voz y las de audio difieren principalmente en el ancho de banda. Esta condición tiene como consecuencia la forma de establecer la codificación. Mientras la codificación de voz se basa en un modelo predictivo ideal para las frecuencias bajas (frecuencias de muestreo entre 8 y 16 kHz), no presenta una buena adaptación para la codificación de detalles espectrales precisos, por ejemplo los que están presentes en las

señales de banda completa como música (frecuencias de muestreo $\geq$ 32 kHz). Por lo tanto el esquema, para la codificación de audio se basa en un modelo acústico de percepción humana. Actualmente se han propuesto algunos modelos para unificar la codificación de señales acústicas en un solo códec, a continuación se describen algunos trabajos previos basados en este nuevo paradigma de codificación y cuya base de implementación son los dos códecs previamente descritos.

En el artículo [22] se establece un método de distinción de señales acústicas basado en el cálculo del SNR.
Los codificadores basados en transformadas (TCX o codificación en subbandas) no tienen un buen desempeño a tasas de bit particularmente bajas, i.e. al ser usadas en la codificación de voz, de manera análoga los codificadores basados en modelos de filtros de excitación (codificadores predicción lineal de código excitado (CELP por sus siglas en inglés)) diseñados para señales de voz producen artefactos auditivos molestos cuando se codifican señales de audio i.e. música.

En [23] se propone que la señal de excitación sea codificada utilizando un diccionario ACELP o aplicando cuantización vectorial al la transformada rápida de Fourier (FFT por sus siglas en inglés) de la señal objetivo. En [22] se retoma dicha idea pero se exitenden las tramas de análisis a 80 ms con empalme adaptable. Se utiliza la cuantización vectorial para evitar la saturación del cuantizador en el modo TCX.

En la figura 2.9 se muestra la estructura básica del codificador unificado propuesto en [23]

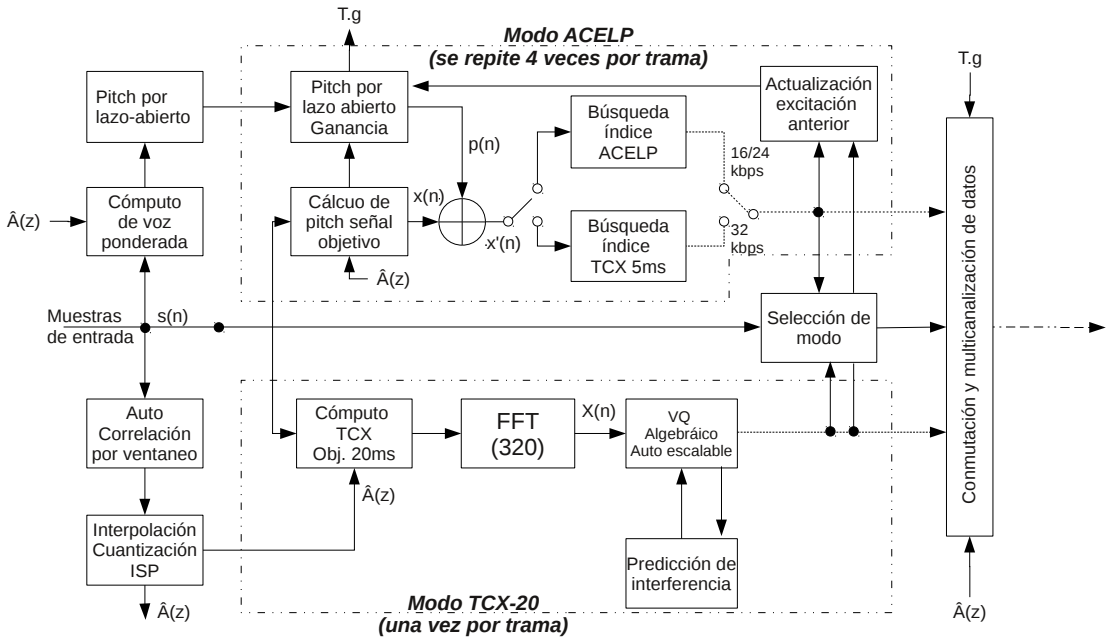


Figura 2.9: Diagrama de codificador híbrido.

La división en subtramas permite una amplia gama de combinaciones en la codificación. En la figura 2.8 se ilustran todos los modos posibles. En el esquema planteado en [22] las tramas de análisis son más largas (80 ms) por ende las subtramas también (40 ms y 20 ms). Es importante resaltar que al codificar (posterior a la decisión) las señales de voz sólo se hará en segmentos de 20 ms, aún cuando toda la trama completa haya sido identificada como una señal de voz.
La selección de modo se puede realizar en lazo abierto o cerrado. Los modos ACELP y TCX están integrados en el sentido de que ambos dependen del análisis LPC y codificación por excitación.

Las subtramas de 20 ms se codifican en uno de los dos modos: ACELP o TCX. La tabla 2.2 es una representación gráfica del proceso de decisión en lazo cerrado, representa 11 pasos en los cuales se realiza la codificación utilizando ambos modos. La letra “A” representa el modo ACELP, “T” representa el modo TCX (el número que precede a la letra representa el tamaño de la subtrama).

Iteración	Subtrama 1	Subtrama 2	Subtrama 3	Subtrama 4
1	A			
2	T20			
3		A		
4		T20		
5	T40	T40		
6		A		
7		T20		
8				A
9				T20
10			T40	T40
11	T80	T40	T80	T80

Cuadro 2.2: Tabla de iteraciones para decisión basada en SNR.

Iteraciones:

1. codificación 1^a subtrama 20 ms modo ACELP
2. codificación 1^a subtrama 20 ms modo TCX
   - se calcula el SNR y se elige el modo más apropiado
3. codificación 2^a subtrama 20 ms modo ACELP
4. codificación 2^a subtrama 20 ms modo TCX
   - se determina el modo más apropiado.
5. codificación primera subtrama de 40ms (compuesta por las subtramas 1 y2) modo TCX
   - se determina si es más conveniente TCX o se conservan las decisiones anteriores
6. codificación 3^a subtrama 20 ms modo ACELP
7. codificación 3^a subtrama 20 ms modo TCX
   - se determina el modo más apropiado.
8. codificación 4^a subtrama 20 ms modo ACELP
9. codificación 4^a subtrama 20 ms modo TCX
   - se determina el modo más apropiado.
10. se codifica la segunda subtrama de 40 ms (compuesta por las subtramas 3 y 4) modo TCX

- se realiza el mismo procedimiento que en el paso 5

11. se codifica toda la trama (80ms) utilizando modo TCX

    - se define qué modo se utilizará en toda la trama o si se utilizará un modo mixto dentro de la subtrama

En [24] se plantea una codificación unificada adoptando un módulo de separación armónica de canal único a manera de pre-procesador; para ello se emplea un método de análisis por modulación en frecuencia. Se adoptan dos estándares internacionales (AMR-WB y HE-AACv2) para proveer una operación interoperativa.

Este modelo enfoca el análisis en la modulación frecuencial el cual consiste en los siguientes pasos esenciales:

- Un banco de filtros STFT.
- Detección de envolvente de subbanda (utilizando la magnitud espectral de la transformada de Fourier).
- Análisis en frecuencia de las envolventes en cada subbanda.

La magnitud del espectro de la envolvente de subbanda se despliega típicamente en una representación de espectrograma modulado.

La alta energía provocada por la región de pitch de la señal de voz es una característica prominente del espectro de modulación. La suma de la energía a lo largo de las frecuencias acústicas también representará niveles altos en la modulación de frecuencias relacionas con la región de pitch.
En los intervalos donde no existan picos, se puede utilizar el valor del promedio en movimiento (moving average) de las ventanas previas para estimar la región de pitch. Es posible considerar que el intervalo para señales no armónicas está fuera de la región de pitch en el intervalo de búsqueda de pitch. La función de supresión de frecuencias está determinada por el cociente entre el objetivo (armónico) y la región restante.

El AMR-WB y el HE-AAC deben operar simultáneamente, i.e. la longitud de las ventanas debe ser la misma. El HE-AAC opera a 16 kHz de muestreo tiene una ventana de 1024 muestras, pero el AMR-WB tiene 320 muestras. Debido a esto se modifica la longitud de la trama a 256 muestras por medio del submuestreo en AMR-WB y se unen cuatro tramas contiguas para obtener tramas de longitud idéntica entre los dos codificadores. La señal armónica se separa de la señal de entrada utilizando mediante el método de decisión previamente descrito. Una vez realizado esto, el cociente de la señal separada y la señal original se calcula, dicha cantidad (potencia) determina los bits que se asignará a ambos codificadores.

$$PowerRatio = \frac{\sum_{frame}[x_h(n)]^2}{\sum_{frame}[x(n)]^2}$$

Las señales objetivo son muestreadas a 16 kHz para tasas de bit de 19.85 kbits/s. La operación del codificador propuesto tiene 4 modos de operación para la asignación de bits, de acuerdo con el criterio de cociente de potencias previamente mencionado:

Modo	Criterio 1	Tasa de bit para AMR-WB (kbps)	Tasa de bit para HE-AAC (kbps)
A	$0 \leq Pow_{ratio} \leq Thr_C$	0	19.85
B	$Thr_C \leq Pow_{ratio} \leq Thr_B$	6.60	13.25
C	$Thr_B \leq Pow_{ratio} \leq Thr_A$	6.60	13.25
D	$Thr_A \leq Pow_{ratio} \leq 1$	19.85	0

Cuadro 2.3: Cuatro modos diferentes para el códec propuesto en [24].

En el diagrama de estados en la figura 2.10 se muestran las transiciones entre los módulos de operación, la transición del estado A al estado D y viceversa está prohibida (cambio abrupto) ya que esto puede generar una discontinuidad perceptible (artefacto auditivo).

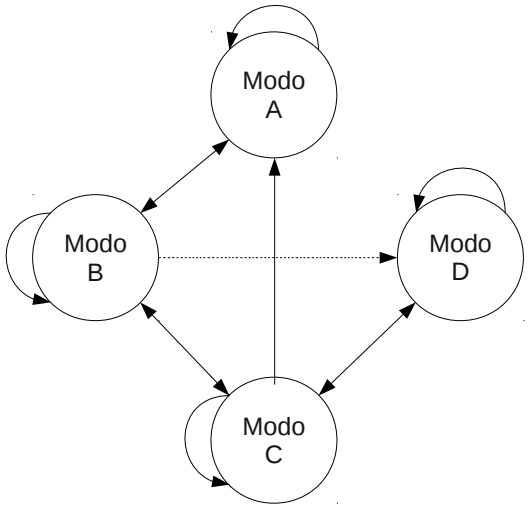


Figura 2.10: Diagrama de estados para modos de codificación.

Parámetro	subventana 1/3	subventana 2/4	Total por ventana
modo			1
ISPs			46
retardo de pitch	9	6	30
ganancia de pitch	6	6	24
ganancia de diccionario	6	6	24
índice de diccionario	45/80/128	45/80/128	180/320/512
bits no usados			15/35/3
Total			320/480/640

Cuadro 2.4: Disposición de bits para el algoritmo de codificación para 16/24/32 kbit/s en modo ACELP.

Parámetro	Total por ventana
modo	1
ISPs	46
VQ algebráico	273/433/593
Total	320/480/640

Cuadro 2.5: Disposición de bits para el algoritmo de codificación para 16/24/32 kbit/s en modo TCX.

Capítulo 3
Módulo de decisión basado en la Razón Señal a Ruido (SNR)

La decisión de codificación basada en el índice SNR presenta, como ventaja principal, el bajo nivel de complejidad en la operaciones necesarias para la determinación del códec específico a utilizar. Para ello primero se obtiene una diferencia entre la señal original y una versión codificada, después se obtiene un cociente entre la señal original y la diferencia previamente obtenida. Sin embargo, tiene como deficiencia principal el hecho de que el proceso de análisis requiere que la señal sea codificada utilizando ambos códecs y una comparación posterior entre los resultados particulares y así definir a qué tipo pertence (voz o audio). A lo largo de este capítulo se describirá el cálculo de SNR, la comparación de las muestras y la selección de tramas para la construcción de una versión codificada.

3.1. Módulo de decisión

El módulo que a continuación se describe está basado en la evaluación del "error de codificación". Sea $x(n)$ la señal original, y $\hat{x}(n)$ la misma señal procesada por alguno de los dos codificadores, el error (o ruido de codificación) está definido por:

$$e_x(n) = x(n) - \hat{x}(n) \tag{3.1}$$

La razón señal a ruido está definida por:

$$SNR = 10 * log_{10}\left(\frac{\Sigma x^2(n)}{\Sigma e_x^2(n)}\right) = 10 * log_{10}\left(\frac{\Sigma x^2(n)}{\Sigma (x(n) - \hat{x}(n))^2}\right) \text{dB} \tag{3.2}$$

Para determinar si la señal acústica que se desea codificar es voz o audio se realiza el siguiente procedimiento:

1. Se codifica la señal original utilizando cada uno de los códecs (AMR-WB y HE-AAC).

2. Una vez que se ha codificado la muestra se obtiene el SNR mediante la comparación de las versiones codificadas contra la señal original.

3. Después de haber obtenido los índices SNR, se determina el tipo de señal a través de la selección del codificador que obtuvo el SNR más alto.

4. Si la señal codificada utilizando el códec AMR-WB presenta el mayor SNR, se asume que el segmento de la señal es de voz.

5. Si es el caso del códec HE-AAC, se asume que el segmento de la señal es audio.

Este análisis se realiza en subtramas de 20 ms sin embargo, puede considerarse una trama mayor. En el caso de codificadores ACELP se tiene un desempeño óptimo con las tasas de compresión más altas en el bloque temporal de 20 ms debido a las características de estacionalidad de las señales de voz. La codificación de las tramas puede realizarse en segmentos de 20, 40 u 80 ms. Para definir el tamaño de la subtrama a codificar se proponen métodos de lazo cerrado.

Con la finalidad de reducir (o eliminar) los artefactos auditivos causados por la conmutación entre códecs, en el artículo [22] se plantea el uso de ventanas para codificar las tramas analizadas previamente, si la trama corresponde a una señal de voz se utiliza una ventana cuadrada para delimitar las muestras, si la trama corresponde a una señal de audio, la ventana que delimita las muestras a utilizar será recta en la parte central y de medio coseno en los extremos, de esta manera se genera cierto empalme entre las tramas a analizar.

En la figura 3.1 se muestran las ventanas utilizadas, cuadrada cuando la trama previa y actual fueron determinadas como voz y de medio coseno cuando la trama previa y actual fueron determinadas como audio. Es posible apreciar el empalme en la ventana de medio coseno con la trama previa y siguiente, la ventana cuadrada no presenta dicho empalme. En el caso de que la trama previa haya sido una señal de voz, y la ventana siguiente una de audio, el inicio de la ventana para codificar la trama actual (audio) sería cuadrada pero el final de la ventana (extremo derecho) sería igual, medio coseno.

El desarrollo del proyecto reportado en este documento incluye la simulación de este método de decisión, sin embargo, el desarrollo no fue exhaustivo dado que no se pretendía profundizar en este esquema, por lo tanto se deben aclarar los siguientes puntos antes analizar los resultados obtenidos:

1. La muestra de audio se codifica en su totalidad utilizando ambos códecs.

2. La codificación se realiza en una sola pasada.

3. Al construir el archivo a partir de la simulación no se realiza el filtrado por ventanas.

4. La elección de códec se realiza para un tamaño fijo de subtrama.

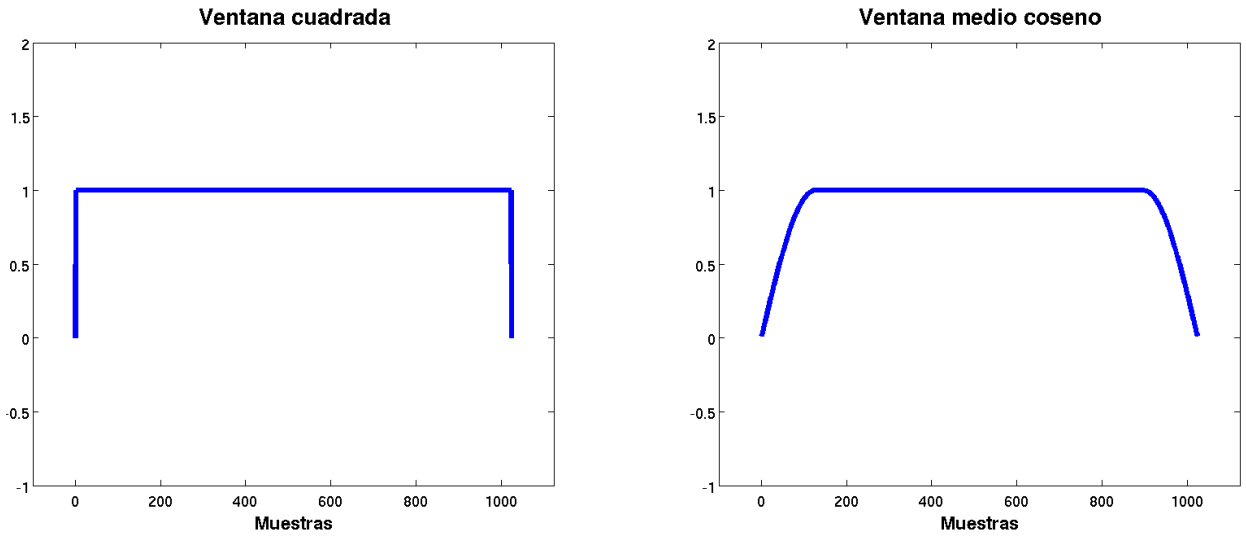


Figura 3.1: Ventanas para selección de muestras.

3.2. Codificación de muestras

Para realizar las simulaciones correspondientes al método de decisión antes descrito fue necesario seleccionar algunas muestras esencialmente de alta calidad con tres características indispensables:

- formato WAV
- modulación de impulsos codificados (PCM por sus siglas en inglés) con 16 bits/muestra
- frecuencia de muestreo de 44.1 kHz

En el caso de señales de audio se utilizaron fragmentos de pistas de CDs de música instrumental, música con voz femenina y música con voz masculina. Para las señales de voz se descargó un archivo de uso libre del sitio Free Sound [25].

Para el desarrollo de la simulación se utilizaron los códecs disponibles para la industria. No se tuvo acceso al código fuente, sólo a los programas ejecutables. Se buscó estar alineados con los estándares internacionales, por lo tanto se descargaron los códecs de los siguientes sitios, en el primer caso desde una institución reguladora y en el segundo caso de una empresa líder en el desarrollo de tecnologías de esta naturaleza, el tercer caso corresponde a una empresa desarrolladora de software ampliamente conocida por sus programas para la creación de CDs y DVDs:

- códec AMR-WB se descargó de 3GPP [26]
- códec AAC se descargó de VoiceAge [27]

- códec HE-AACv2 se descargó de NERO [28]

Se realizaron pruebas con los tres códecs señalados, pero debido a las limitaciones de uso, se optó por descartar al códec obtenido en el sitio de VoiceAge. En su lugar se utilizó el códec desarrollado por NERO, ya que este programa resultó más versátil al ofrecer la opción de realizar codificaciones a tasas de bit constante, variable y promedio, además incluye la variante de perfil, por medio de la cual se puede elegir entre las tres versiones más relevantes de la tecnología AAC:

- codificación avanzada de audio de baja complejidad (LCAAC por sus siglas en inglés) ,
- codificación avanzada de audio de alta eficiencia (HE-AAC por sus siglas en inglés) , y
- HE-AACv2

Al contar con los códecs más convenientes se definieron las tasas de bit con las cuales se realizaría el análisis. Los códecs seleccionados presentan limitaciones distintas: el AMR-WB al ser un códec enfocado en las señales de voz solo permite el uso de tasas de bit relativamente bajas, por el contrario, el códec HE-AACv2 tiene la capacidad de utilizar tasas de bit más altas. Las tasas empleadas en esta simulación se determinaron al encontrar tasas comunes para ambos códecs, por lo tanto solo se emplearon tasas bajas. Los bloques de muestras se codificaron utilizando las siguientes tasas de bit: 8, 10, 12, 24, 48 y 72 kb/s.

La simulación se realizó en el ambiente Matlab, y cada señal fue sometida al proceso de codificación e inmediata decodificación de tal manera que los datos tuvieran el formato adecuado para el procesamiento y análisis. Al codificar las señales de alta calidad el archivo resultante está en un formato diferente, ya que el proceso de decodificación se realiza al momento de la reproducción. La figura 3.2 muestra un diagrama del proceso necesario para el uso de las muestras de alta calidad en la simulación. Otro factor a considerar está relacionado con la decodificación de los archivos procesados por el códec AMR-WB ya que la señal de salida presentaba una frecuencia de muestreo superior 48 kHz, por lo tanto fue necesario utilizar alguna técnica de diezmado (submuestreo). Para realizar el submuestreo se utilizó el código desarrollado por Naoki Shibata [29] el cual está bajo la licencia GNU. La elección de este código está justificada por una investigación realizada por John A. Phillips [30] en la cual evalúa el desempeño de diversos códigos para interpolar o submuestrear.

Figura 3.2: Procesamiento de señales para simulación.

El archivo resultante del proceso antes descrito está en un formato adecuado para el análisis, sin embargo, el proceso de codificación-decodificación presentó un efecto adverso, un

pequeño desplazamiento temporal del proceso de filtrado a través de los bancos de filtros del códec AMR-WB. La figura 3.3 muestra el efecto descrito.

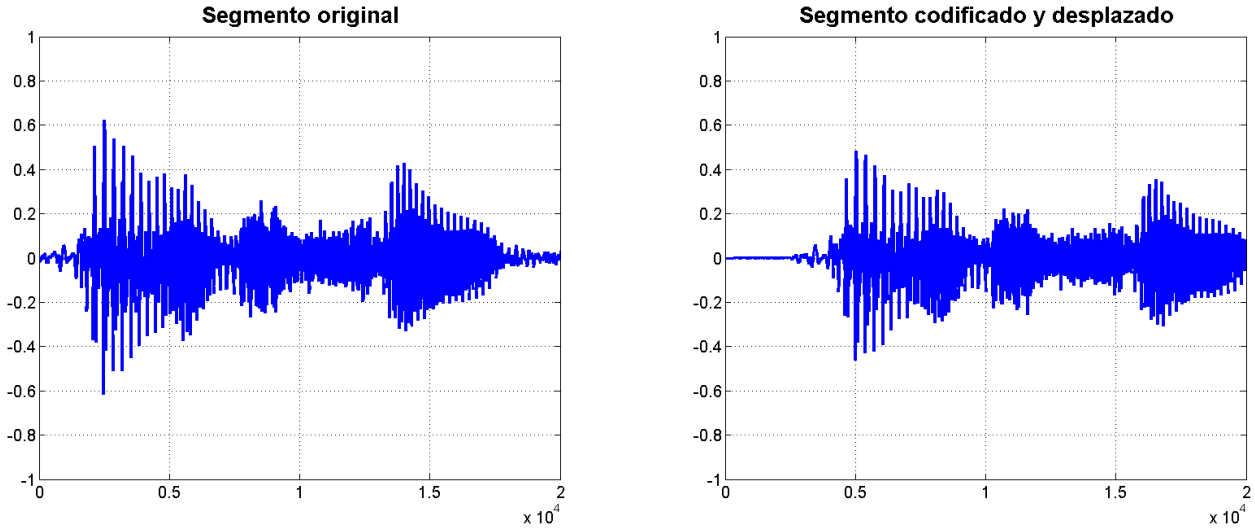


Figura 3.3: Desplazamiento temporal debido al proceso de codificación-decodificación.

Originalmente se detectó el desplazamiento al reproducir los archivos utilizando el programa Audacity[1] el cual cuenta con funciones de edición y reproducción básicas de archivos de audio. Además permite gráficar los niveles de potencia de la señal así como el espectro en frecuencia. Por lo tanto el último ajuste que se realizó a la señal previo al análisis fue la alineación. Para tal efecto se realizó la comparación SNR de la señal original vs. procesada, se hizo un ajuste empírico basado en la observación. Se asumió que las señales estaban alineadas al obtener el valor más alto de evaluación SNR, la comparación se realizó utilizando toda la señal, no por bloques.

3.3. Proceso de Selección del códec óptimo

La selección de un códec óptimo se realiza mediante el análisis de tramas y subtramas. Como se mencionó previamente las tramas de codificación pueden ser de tres tamaños diferentes (20 ms, 40 ms y 80 ms) sin embargo, para la simulación que se desarrolló sólo se utilizaron subtramas de 20 ms para realizar la codificación. Para determinar el códec más apropiado se utiliza el valor promedio del SNR calculado a partir de la comparación de segmentos de audio de 5 ms. Por lo tanto cada subtrama de 20ms se divide en 4 segmentos de 5ms y con cada uno de esos segmentos se calcula el índice SNR, posteriormente se obtiene el

[1]http://audacity.sourceforge.net/

promedio en esa subtrama. La figura 3.4 muestra cómo se divide una señal en subtramas de 20ms (líneas gruesas) y segmentos de 5ms (líneas delgadas).

Figura 3.4: Señal dividida en tramas de 20 ms y segmentos de 5 ms.

La comparación entre segmentos de 5 ms se hace para las dos versiones procesadas, tanto AMR-WB como AAC, el códec más adecuado para la trama de 20 ms se elige de acuerdo al promedio como se señala en la figura 3.4.

Al realizar la simulación en Matlab era necesario contar con las tres versiones de la señal:

- Original
- Procesada por el códec AMR-WB
- Procesada por el códec HE-AACv2

Todas en formato .wav, ya que este tipo de archivos son los que pueden ser leídos y manipulados por la versión de Matlab utilizada (R2010b).

Posterior a la elección del códec óptimo para cada trama de 20 ms se construye una nueva versión de la señal, la cual está compuesta por tramas de 20 ms procesadas por alguno de los dos códecs. En el ambiente de Matlab esto corresponde a crear un nuevo arreglo a partir de elementos tomados de otros arreglos. Gracias a las características de dicho ambiente de trabajo es posible guardar el archivo creado en formato .wav y posteriormente reproducirlo utilizando cualquier otro programa. La figura 3.5 muestra una representación gráfica de la construcción del arreglo resultado de las comparaciones.

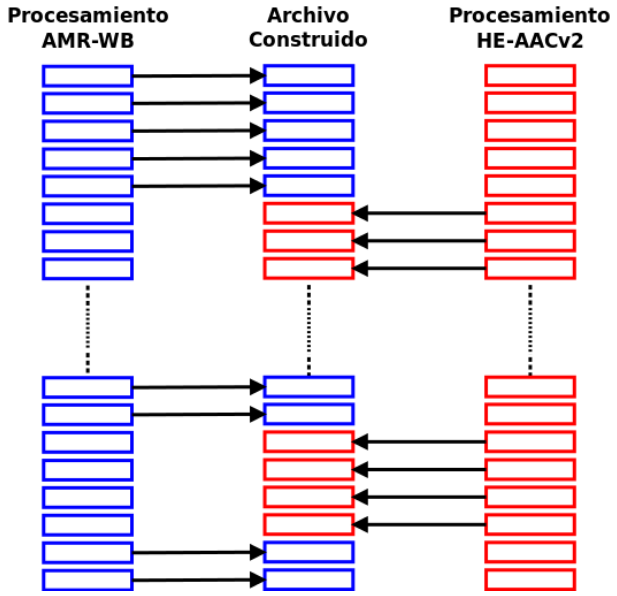


Figura 3.5: Construcción del archivo a partir de tramas procesadas.

3.3.1. Evaluación subjetiva

La evaluación subjetiva es un método mediante el cual se determina el mayor parecido de una señal procesada al compararla con la versión original. Esta evaluación sólo puede ser realizada por medio de escuchas (personas a las cuales se les presenta dicha comparación). La metodología para realizar la evaluación subjetiva fue la siguiente:

1. Se reprodujo el archivo original
2. Inmediatamente después se reprodujo el archivo codificado con AMR-WB
3. Se reprodujo, de nuevo, el archivo original
4. Se reprodujo, de nuevo inmediatamente después, el archivo codificado con HE-AACv2
5. Se repitieron los pasos 1 a 4 en dos ocasiones
6. Se reporta la evaluación subjetiva

Con la finalidad de evaluar la calidad subjetiva de los códecs utilizados en esta simulación, el proceso antes señalado se realizó para todas las calidades empezando por la más baja (8 kb/s) hasta llegar a la más alta (72 kb/s).

3.4. Resultados

Al graficar el comportamiento de los códecs en el sentido del SNR se observó que conforme se aumenta la tasa de bit se incrementa el valor del SNR para las codificaciones realizadas

con el códec AMR-WB, mientras que para el códec HE-AACv2 se mantiene en un nivel casi constante, el incremento en el índice es muy pequeño. El cálculo del SNR se realizó comparando todo el archivo de audio (no por segmentos), de esta manera se obtiene sólo un valor. La figura 3.6 muestra el comportamiento del SNR de ambos códecs conforme se incrementa la tasa de bit, los incrementos del códec HE-AACv2 son sutiles parece que la tendencia es mantener cierto nivel. Los resultados presentados en la figura 3.6 corresponden al análisis de un fragmento de una pieza de música instrumental que involucra una amplia gama de frecuencias.

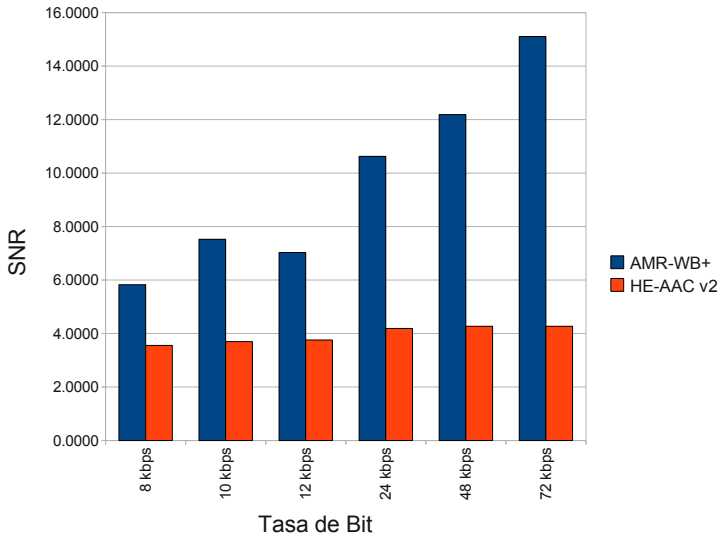
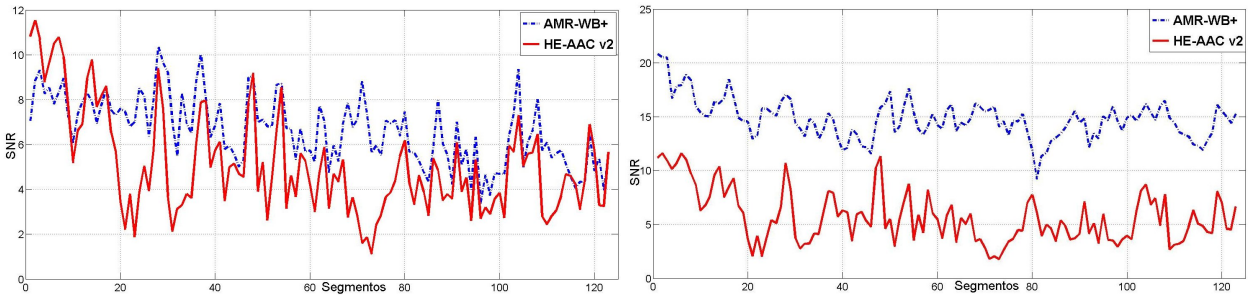


Figura 3.6: Gráfica de barras respuesta con señal de música.

Resulta particularmente útil observar cómo se comporta la gráfica de SNR al realizar el cálculo por segmentos, ya que se aprecia de manera clara en dónde se encuentran los cruces entre ambos códecs, de tal forma que puede proporcionar una idea de cómo será el resultado de la codificación unificada. La figura 3.7 muestra el comportamiento del SNR en bloques de 80 ms en las codificaciones de 8 y 72 kb/s. Vale la pena profundizar en el análisis de este resultado ya que en primer lugar el códec AMR-WB (que originalmente fue diseñado para voz) tiene un mejor desempeño que el códec HE-AACv2 (diseñado específicamente para audio). Conforme la tasa de bit aumenta el SNR de la señal codificada con AMR-WB también, el SNR de la señal codificada con HE-AACv2 se mantiene en relativamente constante. Los resultados obtenidos no concuerdan con lo que se esperaba ya que el resultado lógico parecería ser que conforme aumenta la tasa de bit, la señal de que se obtiene como resultado del procesamiento sería más parecida a la original y por lo tanto el SNR más alto. En el caso del códec AMR-WB si sucede pero en el caso del códec HE-AACv2 diseñado para señales con un amplio contenido espectral no.

La figura 3.8 muestra un resultado esperado, ya que al aumentar la tasa de bit aumenta la eficiencia del códec HE-AACv2 en el sentido de SNR. En este caso llama la atención la similitud en los índices SNR para ambos códecs, hay que destacar que mientras más alta es



(a) Codificación de audio a 8kb/s (b) Codificación de audio a 72kb/s

Figura 3.7: Comportamiento en bloques de señal de audio a diferentes tasas de bit.

la tasa de bit empleada pareciera que muestran una tendencia a estabilizarse. Debido a las limitaciones establecidas por el códec AMR-WB utilizado, no fue posible indagar con tasas de bit más altas, ya que se reporta que el códec HE-AACv2 tiene mejor desempeño en tasas de bit superiores. Se puede inferir que este códec superaría el desempeño de AMR-WB a partir de la tendencia que muestra la figura 3.8 sin embargo, resulta notable el comportamiento de la curva SNR por bloques.

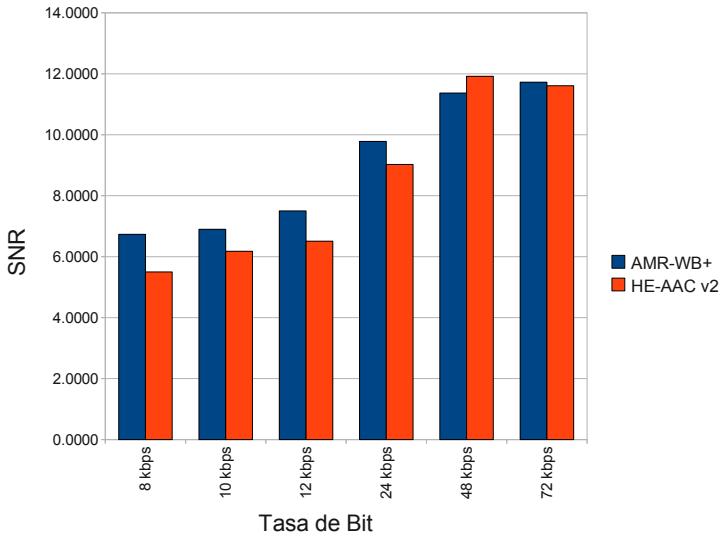
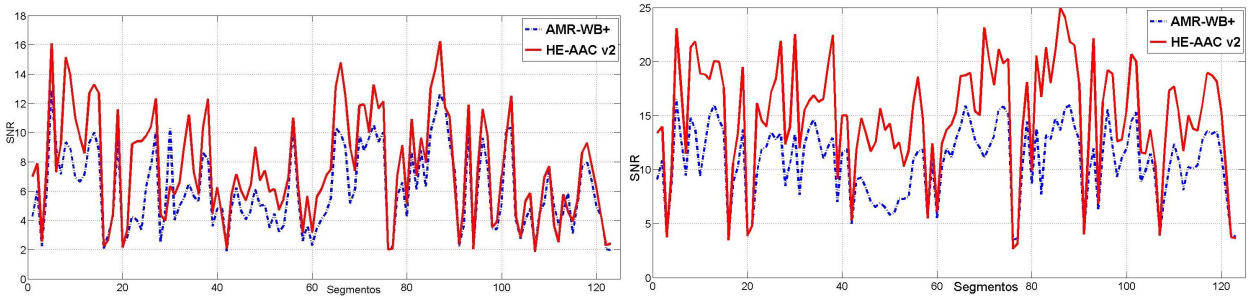


Figura 3.8: Gráfica de barras respuesta con señal de voz.

La figura 3.9 muestra el comportamiento de las curvas de SNR para ambos códecs, el comportamiento es similar a aquel mostrado en la figura 3.7, pero resulta interesante observar que la varianza del códec HE-AACv2 también se incrementa además de que hay regiones con valores similares, particularmente los picos en los valores más pequeños, al proyectar las curvas de SNR sobre la representación de los niveles de potencia de la señal de audio

original, parece que traza una envolvente donde los picos (casi siempre) corresponden a las pausas entre palabras.



(a) Codificación de voz a 8kb/s

(b) Codificación de voz a 72kb/s

Figura 3.9: Comportamiento en bloques de señal de voz a diferentes tasas de bit.

VOZ		
	AMR-WB	HE-AACv2
8 kb/s	☑	☒
10 kb/s	☑	☒
12 kb/s	☑	☒
24 kb/s	☑	☑
48 kb/s	☑	☑
72 kb/s	☒	☑

(a) Voz

AUDIO		
	AMR-WB	HE-AACv2
8 kb/s	☑	☒
10 kb/s	☑	☒
12 kb/s	☑	☑
24 kb/s	☒	☑
48 kb/s	☒	☑
72 kb/s	☒	☑

(b) AUDIO

Cuadro 3.1: Tabla de evaluación subjetiva.

Los resultados producto del análisis objetivo y subjetivo en el caso de la voz son muy similares, en las codificaciones de 24 y 48 kbps no hubo diferencias perceptibles, ambos códecs mostraron una alta fidelidad, los resultados de las gráficas son consistentes con el resultado de la tabla subjetiva. En el caso de la música los resultados de ambas pruebas no son similares, la codificación utilizando HE-AAC v2 muestra una tasa de bit casi constante con variaciones muy pequeñas, en el análisis objetivo siempre tiene mejores resultados el códec AMR-WB+. El segmento de música utilizado es un fragmento de los "Conciertos de Brandenburgo", la selección de esta muestra radicó en el alto contenido espectral.

Capítulo 4
Modelo de decisión basado en Wavelets

La información obtenida mediante la transformada de Fourier presenta algunas limitaciones, el análisis no provee información relativa al comportamiento de las componentes espectrales en el tiempo. Para sobreponerse a esta limitante, existe una técnica mediante la cual se divide la señal en segmentos temporales y a cada uno de estos se les aplica un análisis de Fourier local. Este tipo de análisis se denomina STFT, el cual provee información temporal y espectral por lo que es una herramienta útil.

Asimismo, existe otro tipo de transformada que provee información sobre el comportamiento de una señal en el dominio temporal a través de un parámetro denominado la escala (la cual es posible considerar bajo ciertas circunstancias, como el inverso de la frecuencia), denominada la transformada continua en Wavelets (transformada Wavelet continua (CWT por sus siglas en inglés)). Al igual que el espectrograma para el caso de STFT, existe una representación gráfica de la densidad espectral de potencia de la CWT llamada escalograma. Por medio del análisis de dichas representaciones es posible distinguir patrones específicos para una señal de voz o de audio. Estos patrones se asemejan a las cordilleras de los sistemas montañosos; por ello se propone el diseño de un sistema que se enfoca en la identificación y análisis del comportamiento de dichas cordilleras para distinguir entre una señal de voz y una de audio.

4.1. Fundamento teórico

El contenido espectral de una señal obtenido por el análisis de transformada de Fourier provee información acerca de todas las frecuencias presentes en el segmento temporal de la señal analizada. A pesar de que es una herramienta de gran utilidad tiene una desventaja importante: no provee información respecto a lo que sucede en el tiempo, esto es, no se puede saber en que lapso temporal se presentan dichas características frecuenciales. Por ello es necesario establecer representaciones que permitan observar este contenido en frecuencia a lo largo del tiempo. A continuación se presentan brevemente la STFT y la CWT y se realiza

una comparación entre las mismas con el fin de fundamentar la elección de una de ellas para la discriminación de señales de audio y voz.

4.1.1. STFT

La STFT está dada por:

$$STFT_x(\omega, \tau) = \int_{-\infty}^{\infty} \omega^*(t-\tau)x(t)e^{j\omega t}dt \tag{4.1}$$

donde $x(t)$ es la señal a analizar, $\omega^*(t)$ es la ventana de análisis.

La idea central de la STFT consiste en fragmentar la señal a analizar en pequeños bloques temporales y obtener la transformada de Fourier de cada uno de estos bloques. Evidentemente la señal se considera estacionaria en el intervalo que dura la ventana.

La selección de una función ventana $\omega(t)$, es un factor de suma importancia para un análisis adecuado. Gabor [31] propuso el uso de una ventana Gaussiana para obtener una buena resolución tiempo-frecuencia.

Al utilizar una Gaussiana se obtiene la mejor localización conjunta tiempo-frecuencia ya que cumple con el límite inferior establecido por el principio de incertidumbre (ventana de Gabor):

$$w(t) = \beta e^{-\alpha t^2}, \alpha, \beta > 0 \tag{4.2}$$

Una vez que se ha elegido la frecuencia de análisis $\omega = \omega_0$ para la STFT $\omega = \omega_0$ no podrá ser modificada, es decir, todo el análisis se realizará con la misma frecuencia, debido a lo anterior está claro que el análisis se realizará con una resolución constante.

Un espectrograma es una densidad de energía y está definida por

$$P_{STFT_x} \triangleq |STFT_x|^2 \tag{4.3}$$

y mide la energía de $x(t)$ en la vecindad de (u, ξ) espicificada por los átomos de Heisenberg representados en la figura 4.4.

La figura 4.1 muestra los átomos para la transformada STFT

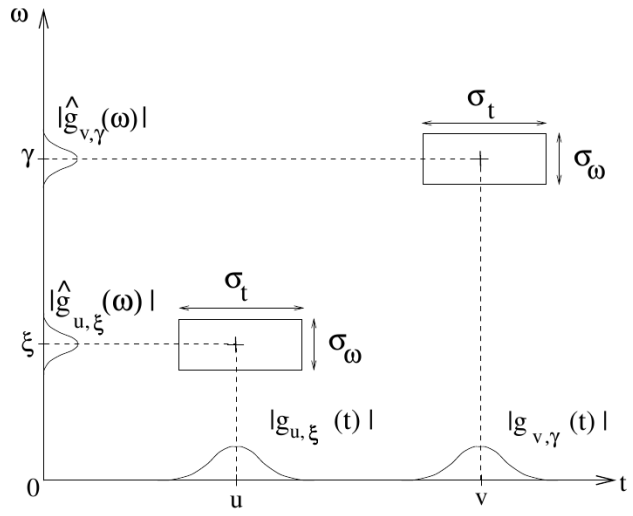


Figura 4.1: Átomos obtenidos con la transformada STFT.

El principio de incertidumbre de Heisenberg establece que la resolución máxima en el plano tiempo-frecuencia está limitada por el producto $\sigma_t \sigma_\omega$, y satisface:

$$\sigma_t \sigma_\omega \geq \frac{1}{2} \tag{4.4}$$

4.1.2. La transformada continua en Wavelets, CWT

Ofrece una alternativa al análisis tiempo frecuencia al proveer una descripción en el plano tiempo-escala. La CWT está definida por la ecuación 4.5

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi^* \left(\frac{t-b}{a}\right) dt \tag{4.5}$$

donde $x(t)$ es la señal a analizar y $\psi_{a,b}(t)$ es una función denominada wavelet madre.

Al normalizar mediante el factor $1/\sqrt{a}$ se asegura que la energía en cada wavelet $\psi_{a,b}(t)$ permanezca independiente de la escala a. La función de análisis para la transformada wavelet se define como:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \left(\frac{t-b}{a}\right) \tag{4.6}$$

La función $\psi_{a,b}(t)$ por lo general es un filtro pasa bandas, si $a \gg 1$ corresponde a funciones de base con gran soporte temporal. La escala se relaciona con el inverso de la frecuencia ya que la escala es proporcional a la duración de la función base utilizada en la expansión de la señal [32].

El análisis por transformada de Fourier abarca un rango infinito, por lo tanto no permite una buena localización en el tiempo. La CWT permite una buena localización tanto en el tiempo como en la frecuencia. El rango temporal permanece finito y es adecuado para localizar singularidades en la señal analizada. Las funciones clasificadas como wavelets deben satisfacer las siguientes condiciones [33]:

1. La wavelet madre debe tener energía finita, i.e.,

$$E_\psi = \int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty \tag{4.7}$$

2. La función debe satisfacer la condición de admisibilidad dada por:

$$C_\psi = \int_{-\infty}^{\infty} \frac{|\hat{\Psi}(\omega)|^2}{\omega} d\omega < \infty \tag{4.8}$$

   donde $\hat{\Psi}(\omega)$ representa la transformada de Fourier de la función, esto implica que la función wavelet no tiene componente en la frecuencia cero, que en el dominio del tiempo se traduce como una función de media cero.

3. Para las wavelets analíticas (complejas), la transformada de Fourier debe ser real y desvanecerse para las frecuencias negativas.

De acuerdo con la fórmula de Plancherel, la CWT conserva la energía total de la señal E_x:

$$E_x \triangleq ||x||_2^2 = \int_{-\infty}^{\infty} |x(t)|^2 dt = C_\psi^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |CWT(a,b)|^2 \frac{dadb}{a} \tag{4.9}$$

Un escalograma es la representación de la magnitud de la densidad de energía local para cierto tiempo-escala está definido por la ecuación 4.10

$$P_W(a,b) \triangleq |W(a,b)|^2 \tag{4.10}$$

La figura 4.2 muestra los átomos de la CWT

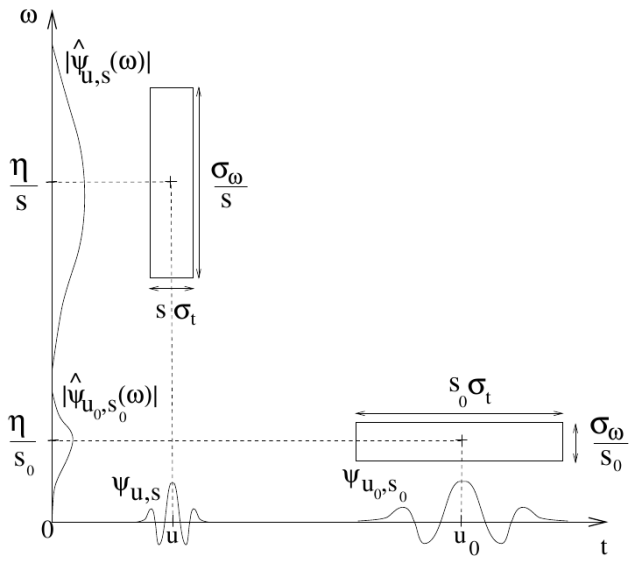


Figura 4.2: Átomos obtenidos con la CWT.

4.1.3. Comparación entre STFT y CWT

La CWT y STFT comparten varias características, ambas son una transformada que mapea la representación de una señal en 1-D a una representación en 2-D, por lo tanto ambas son altamente redundantes.

Cuadro 4.1: Lista de características de la STFT y la CWT.

STFT	**CWT**
Transformada de Fourier local	Transformada wavelet
Modulación y desplazamiento	Escalamiento y desplazamiento
Espectrograma	Escalograma
Resolución constante	Resolución variable
Caracterización de la regularidad	Aislamiento de discontinuidades
No detecta variaciones dentro de una misma ventana	Puede detectar diversos comportamientos en una misma ventana
Comportamiento similar a banco de filtros con anchos de banda constante	Comportamiento similar a banco de filtros paso-bandas con anchos de banda distintos
Misma localización en todo el análisis	Localización temporal precisa en escalas pequeñas (frecuencias altas)
No está restringido por la admisibilidad	Se debe cumplir la admisibilidad

Al comparar la definición de la STFT dada por la ecuación 4.1 con la definición de la CWT en la ecuación 4.5, es evidente que ambas transformadas tienen el mismo principio, una medida de la similitud: tanto la STFT y la CWT pueden ser vistas como una correlación de sus respectivas funciones de base con la señal $x(t)$,

$$CWT_x(a,b) = \langle \psi_{a,b}(t), x(t) \rangle \tag{4.11a}$$
$$STFT_x(\omega,\tau) = \langle g_{\omega,\tau}(t), x(t) \rangle \tag{4.11b}$$

Densidades espectrales de potencia

El alto parecido entre las dos transformadas se extiende a la representación gráfica, en ambos casos es trazada por la densidad energética en una vecindad específica; en el caso de la STFT ((u,ξ) espectrograma) y para la CWT (escalograma en la vecindad $(u,\xi = \eta/s)$).

$$P_{STFT_x}(u,\xi) = |STFT_x(u,s)|^2 \tag{4.12a}$$
$$P_{CWT_x}(u,\xi) = |CWT_x(u,s)|^2 \tag{4.12b}$$

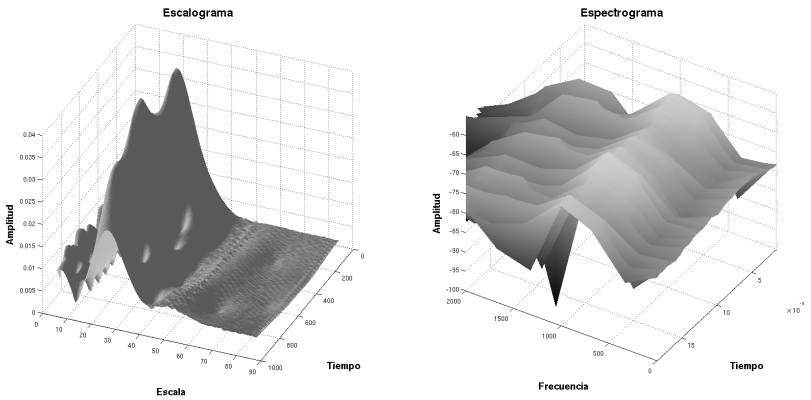


Figura 4.3: Escalograma y espectrograma de una misma señal.

En este trabajo se decidió utilizar la CWT dada el buen desempeño que presentan la wavelets complejas para el seguimiento de frecuencias instantáneas. En los siguientes párrafos se describen los puntos fundamentales del método de decisión desarrollado basado en estas técnicas.

4.2. Wavelets Complejas

Una wavelet compleja (o analítica) es una función cuyo espectro sólo tiene frecuencias positivas. Una wavelet compleja únicamente responde a frecuencias no negativas de una señal, por lo tanto produce una transformada cuyo módulo es menos oscilatorio que en el caso de

una wavelet real. Esta propiedad representa una ventaja al detectar y dar seguimiento a frecuencias instantáneas contenidas en la señal, y será de gran utilidad al analizar señales acústicas.

Existe una familia de wavelets complejas bien conocidas. Dentro de estas funciones se se encuentran la wavelet compleja de Cauchy que se utiliza en mecánica cuántica [34], la wavelet compleja "sombrero Mexicano" [35] que se utiliza en el análisis sónico del eco, y la wavelet compleja de Morlet utilizada de manera vasta en mecánica [36] [37]. El uso en los diversos dominios depende de las particularidades y las restricciones de las aplicaciones. Este trabajo se centra en el uso de la wavelet de Morlet para la identificación de características particulares de la voz o el audio.

4.2.1. Wavelet Compleja de Morlet

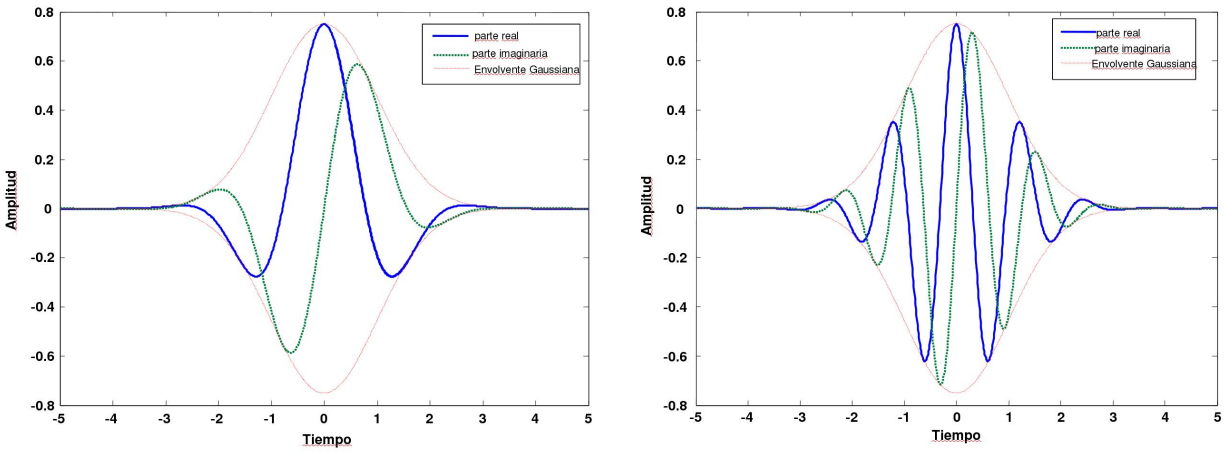
Para analizar la evolución temporal de los tonos frecuenciales se debe utilizar una wavelet analítica para separar la información correspondiente a la fase y amplitud de las señales [33]. La wavelet Morlet es una exponencial compleja modelada por una gaussiana, definida por: 4.13

$$\psi(t) = \frac{1}{\sqrt{2\pi}} e^{-j\omega_0 t} e^{-t^2/2} \tag{4.13}$$

Esta wavelet no cumple con la condición de admisibilidad sin embargo, cuando se escoge ω_0 tal que el segundo máximo de $Re\psi(t)$ (para $t > 0$) sea la mitad del primero ($\omega_0 = 5.336$) se satisface la condición de admisibilidad, y su valor es despreciable en la frecuencia $\omega = 0$:

$$\Psi(\omega)|_{\omega=0} \cong 7 \times 10^{-7} \tag{4.14}$$

Satisfacer la condición de admisibilidad es esencial para demostrar la conservación de la energía. La figura 4.4 muestra la wavelet Morlet compleja con la frecuencia central de $\omega_0 = 2$ y $\omega_0 = 5$. En el primer caso no se cumple con la condición de admisibilidad, para el segundo si.



(a) $\omega_0 = 2$ (b) $\omega_0 = 5$

Figura 4.4: Wavelet compleja Morlet.

La figura 4.5 muestra el espectro en ambos casos y se distingue el cumplimiento de la condición de admisibilidad para la frecuencia central de $\omega_0 = 5$, para el caso de la frecuencia central de $\omega_0 = 2$ no se cumple.

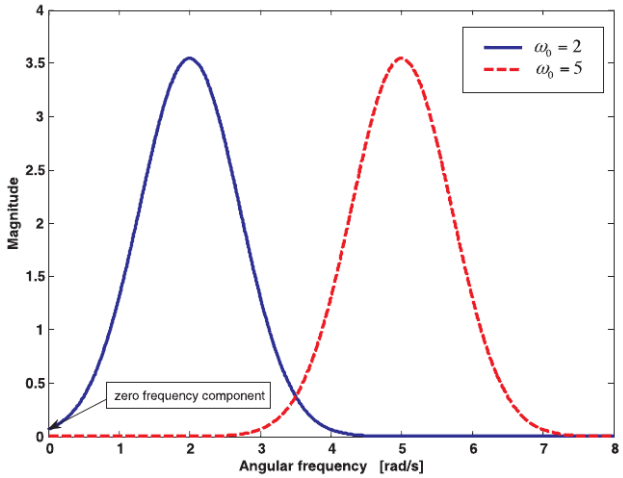


Figura 4.5: Condición de admisibilidad.

La wavelet compleja de Morlet coincide con la ventana de Gabor bajo ciertas condiciones: $a = 1/\sqrt{2\alpha}$ y $\omega = \omega_0\sqrt{2\alpha}$. El parecido se debe a que ambas tienen la forma de una gaussiana, el caso particular de la wavelet Morlet compleja es una función modulada por una gaussiana, en el caso de Gabor es una gaussiana. A partir de las condiciones planteadas resulta interesante contrastar ambas ecuaciones, asumiendo que no hay desplazamiento en el

tiempo:

$$\psi_{a,0}(t) = \frac{1}{\sqrt{2}\pi a} e^{-j\omega_0 t/a} e^{-t^2/2a^2} \tag{4.15a}$$

$$g_{\omega,0}(t) = \beta e^{j\omega_0 t} e^{-\alpha t^2} \tag{4.15b}$$

La condición de igualdad mostrada en las ecuaciones 4.15 solo aplica para las condiciones especificadas previamente, para el resto de los puntos el análisis difiere ya que el análisis por wavelet utiliza ventanas variables, en contraste con el uso de ventanas de tamaño estático del análisis de Fourier local.

4.3. Cordilleras de Wavelets (*Ridges*)

Como la CWT de una señal es la convolución entre la señal y la wavelet madre trasladada y dilatada, se puede expresar como un producto en el dominio de Fourier:

$$CWT(a,b) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{X}(\omega) \hat{\Psi}^*_{a,b}(\omega) d\omega \tag{4.16}$$

donde $\hat{X}(\omega)$ es la transformada de Fourier de la señal analizada $x(t)$ y

$$\hat{\Psi}^*_{a,b} = \sqrt{a}\hat{\Psi}^*(a\omega) e^{j\omega b} \tag{4.17}$$

es la transformada de Fourier de la wavelet madre en la escala a y tiempo b.

Cuando la transformada de Fourier de la wavelet está concentrada de manera importante alrededor de un valor de frecuencia específico, la CWT tenderá a concentrar en los valores de frecuencia asociados con los harmónicos dominantes en la señal, de tal forma que definirá una serie de curvas llamadas cordilleras que evolucionarán en el tiempo. Estas cordilleras están ubicadas en donde la frecuencia de la wavelet escalada coincide con la frecuencia local de la señal. La teoría de cordilleras se establece a continuación. La representación analítica de una señal está dada por:

$$Z_x(t) \triangleq x(t) + j\hat{x}(t) = A(t) e^{j\phi(t)} \tag{4.18}$$

donde $\hat{x}(t)$ es la transformada de Hilbert de la señal $x(t)$, dada por [38]:

$$\hat{x}(t) = \int_{-\infty}^{\infty} \frac{x(t-\tau)}{\pi\tau} d\tau \tag{4.19}$$

$A(t)$ es la envolvente y $\phi(t)$ es la fase instantánea. Una señal compleja se puede caracterizar por medio de una amplitud $A(t)$ y fase $\phi(t)$ (con valores en el intervalo $[0, 2\pi)$ formando un par canónico. $Re[Z_x(t)] = x(t)$

$$x(t) = A(t) cos(\phi(t)) \tag{4.20}$$

La ecuación (4.19) es la convolución de $x(t)$ con $1/t$, también se puede entender como un corrimiento de fase de la señal por $\pi/2$. Se dice que $x(t)$ y $\hat{x}(t)$ están en cuadratura.

Con base en la representación de par canónico señalado en 4.20, Ville [39] propuso calcular la frecuencia instantánea asociada con la fase instantánea $\phi(t)$. Se asume que las amplitudes de la señal y wavelet varían lentamente, i.e. las señales se consideran asintóticas.
De la ecuación (4.18) se define la frecuencia instantánea como:

$$\omega_{inst}(t) = \frac{d\phi(t)}{dt} \tag{4.21}$$

Considerando una wavelet analítica que es asintótica y expresada en forma general $\psi(t) = A_\psi(t)e^{j\phi_\psi(t)}$, la CWT de la señal $x(t)$ se obtiene por [38]

$$CWT(a,b) = \frac{1}{2a}\int_{-\infty}^{\infty} A(t)A_\psi\left(\frac{t-b}{a}\right)e^{\left(j\left(\phi(t)-\phi_\psi\left(\frac{t-b}{a}\right)\right)\right)}dt \tag{4.22}$$

Ya que la integral es asintótica, el teorema de estacionalidad de fase enunciado por Delprat en [38] se puede aplicar. Sea $\Phi_{a,b}(t) \triangleq \phi(t) - \phi_\psi((t-b)/a)$ el argumento de la integral de la ecuación (4.22) y $t = t_0$ el punto estacionario del argumento, entonces

$$\Phi'_{a,b}(t_0) = \phi'(t_0) - \frac{1}{a}\phi'_\psi\left(\frac{t_0-b}{a}\right) = 0 \tag{4.23}$$

y

$$\Phi''_{a,b}(t_0) = \phi''(t_0) - \frac{1}{a}\phi''_\psi\left(\frac{t_0-b}{a}\right) \neq 0 \tag{4.24}$$

Por lo tanto, la cordillera de la CWT de $x(t)$ es el conjunto de puntos (a,b) que son puntos estacionarios del argumento $\Phi_{a,b}(t)$ en la ecuación (4.22) [i.e.,$(t_0(a,b) = b]$. Se puede calcular la cordillera a partir de la ecuación (4.23) como:

$$a(b) = \frac{\phi'_\psi(0)}{\phi'_\psi(b)} \tag{4.25}$$

entonces la frecuencia instantánea junto con su cordillera se convierte en:

$$\omega_{inst}(t)|_{t=b} = \phi'(b) = \frac{\phi'_\psi(0)}{a(b)} \tag{4.26}$$

Las ecuaciones (4.25) y (4.26) muestran que la cordillera es proporcional a la frecuencia instantánea, la ecuación (4.26) se debe calcular después de extraer la cordillera. Todorovska [40] obtiene una aproximación para la CWT dada por:

$$CWT(a,b) \approx \sqrt{\frac{\pi}{2}}\frac{Z_x(t_0)\psi^*\left(\frac{t_0-b}{a}\right)}{corr(a,b)} \tag{4.27}$$

donde

$$corr(a,b) = a|\Phi''_{a,b}(t_0)|^{1/2} e^{-j(\pi/4)sgn[\Phi''_{a,b}(t_0)]} \tag{4.28}$$

Las cordilleras de la wavelet son las cimas de una "montaña" particular analizada, son un conjunto de puntos que dibujan un contorno sobre la montaña, y muestran el cambio lento de la envolvente de la señal. Un acoplamiento máximo no presenta cambios bruscos lo cual marca la estacionalidad del máximo local. Gracias a que los cambios son lentos en la señal analítica, los puntos que lo rodean se mantienen. La representación gráfica se hace en un plano 2D, así que es una "vista superior", la amplitud de cada cordillera (altura) no es perceptible a menos que se utilice un patrón de colores o se haga una representación gráfica en 3D, como se muestra en la figura 4.6.

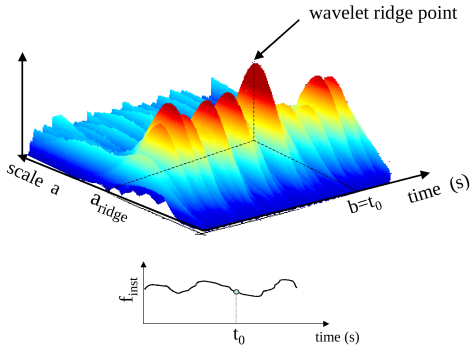


Figura 4.6: Cordilleras de wavelets.

La wavelet Morlet compleja tiene su máximo en $t_0 = b$ en la ecuación (4.27); en el mismo instante la CWT $CWT(a,b)$ también tiene un máximo local despreciando el término de corrección $corr(a,b)$. Esta asunción representa una forma sencilla de obtener las cordilleras a partir de un valor máximo de amplitud local de la CWT, como se muestra en la figura 4.7.

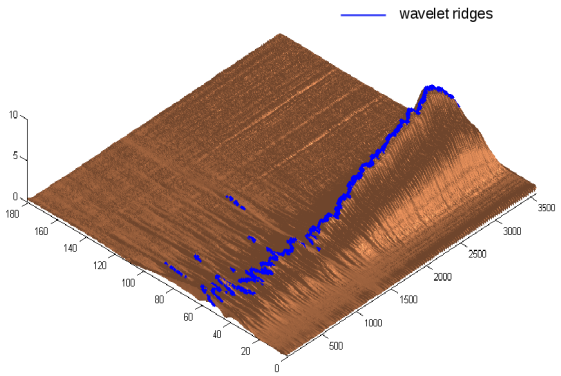


Figura 4.7: Cordilleras de wavelets y frecuencias instantáneas.

4.4. Distinción entre señales de voz/audio

Las diferencias entre las señales de voz y las de audio no son obvias en todos los casos. La forma principal para distinguir entre ellas es buscando armónicos y detectando frecuencias de alta energía en la región del pitch. Debido a la amplia variedad de tipos de música (i.e. instrumental, metal, jazz, clásica) no es posible establecer un solo patrón que se adapte a todos los tipos de música y que además abarque el contenido mixto. Sin embargo, a través del análisis que aquí se propone, es posible afirmar que existen ciertas características particulares cuando se analiza una señal de voz. Las representaciones gráficas de la CWT de los archivos de voz analizados presentan algunas gráficas interesantes en 3D donde se observan comportamientos particulares; que sirven de base para el establecimiento de patrones distintivos entre el audio y la voz.

Al ser más sencillo caracterizar las señales de voz, la base del método presentado es la siguiente: si se identifica una señal de voz, ésta se caracteriza como voz; si no se identifica una señal de voz se caracteriza como audio. La región de pitch se identifica en una banda de baja frecuencia, donde se ha establecido un umbral de 1200 Hz. Para el caso de instrumentos musicales (i.e. piano, violín) la conducta del tono fundamental tiene características similares a las de la voz en la región del pitch sin embargo, en las señales de voz no se identifican armónicos como aquellos presentes en las señales de audio producidas por instrumentos. Particularmente se presenta un caso desafiante cuando un instrumento musical se utiliza para incrementar las características de la voz en una canción, i.e. interferencia constructiva debido a una colisión.

El análisis divide a la señal en dos bandas de frecuencias: 176 Hz $< S_L <$ 1155 Hz y 1156 Hz $< S_H <$ 4479 Hz. El análisis principal se desarrolla en una banda de frecuencia baja, la banda de frecuencia alta puede proveer información adicional. La idea central de este tipo de análisis fue propuesta en el códec AMR-WB el cual analiza la banda de baja frecuencia de la señal (donde se encuentran las características principales de una señal de voz) para determinar si la señal es audio o voz y así realizar una codificación adecuada. Aquí sólo se reporta el análisis de la banda baja. Es importante hacer la observación de que el análisis en la banda alta puede contribuir a una clasificación más certera.

4.4.1. Clasificación de las Señales

La banda de frecuencias bajas (S_L) expresada en términos de la escala corresponde a los siguientes valores: 31 (umbral de frecuencia superior) y 203 (umbral de frecuencia inferior). La wavelet utilizada fue Morlet compleja con una frecuencia central de 0.8 Hz y un ancho de banda de 7 Hz.

Los resultados que se reportan fueron obtenidos mediante el uso de un código para detectar

cordilleras previamente desarrollado por J.M. Lilly [41]. El procedimiento para clasificar las señales acústicas se describe por medio de los siguientes pasos:

1. Se consideran tramas de 100 ms de una señal (voz o audio)
2. Se divide la trama en 5 subtramas, cada una de 20 ms
3. Se realiza la CWT de cada subtrama
4. Se identifican las cordilleras en cada subtrama
5. Se determina en cada subtrama, con base en las cordilleras, el tipo de señal (voz o audio)
6. El resultado mayoritario en las 5 subtramas determina que tipo de señal representa la trama de 100 ms

La división de las muestras para el análisis en tramas de 20 ms se ha establecido de acuerdo a lo propuesto en [22]. En el desarrollo de este proyecto en lugar de utilizar una supertrama de 80 ms (4 subtramas de 20 ms) se utiliza una trama de análisis principal de 100 ms, ya que de esta manera se pretende evitar empates al clasificar una señal. Si el análisis se realizara con un número par de divisiones, se tomaría en cuenta otro criterio para definir cómo se clasificaría la supertrama, bajo el esquema de una división en un número impar de subtramas siempre se obtiene un resultado absoluto. Si 3 o más subtramas arrojan un tipo de de señal específica (audio o voz), la supertrama se clasifica como el resultado mayoritario.

Al aplicar la CWT con la wavelet compleja de Morlet, las muestras de voz masculinas trazan una "montaña" prominente en la región del pitch, por lo tanto la identificación de cordilleras describe una sucesión de puntos los cuales determinan el tipo de señal.

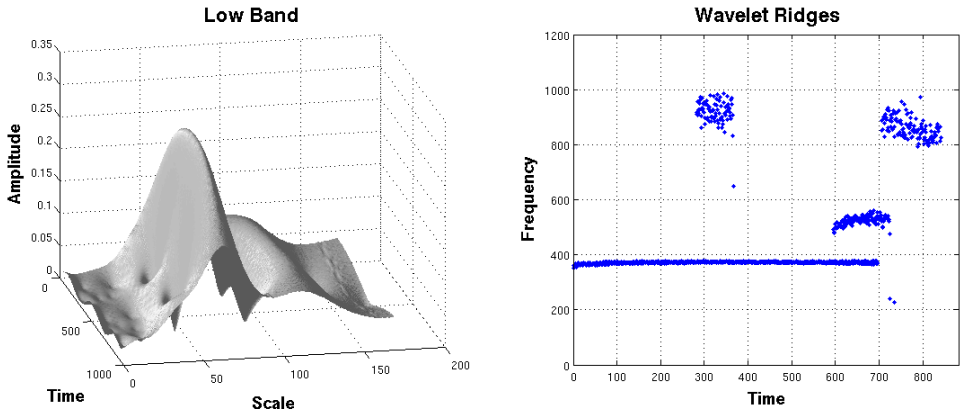


Figura 4.8: Transformada wavelet y cordilleras de muestras de voz masculina.

La figura 4.8 corresponde a la representación gráfica de CWT de una subtrama de 20 ms de una señal de voz masculina grabada sin ruido ambiental. Este es un ejemplo de cómo se

identifican las señales de voz: no se muestran armónicos y tampoco ocurren cambios abruptos en la amplitud, por lo tanto el algoritmo para la detección de cordilleras trabaja de manera adecuada. Las características que se toman en cuenta al clasificar una señal como voz fueron determinadas por medio de la observación y detección de patrones al hacer la pruebas con varias muestras de voz, tanto masculinas como femeninas. Los patrones identificados en la representación gráfica de la CWT deben estar presentes en la detección de cordilleras. Las muestras femeninas tienen cordilleras por encima de la frecuencia promedio de las muestras masculinas (como era de esperarse). También la presencia de varias cordilleras es algo muy común pero siempre es perceptible una cordillera principal de mayor amplitud.

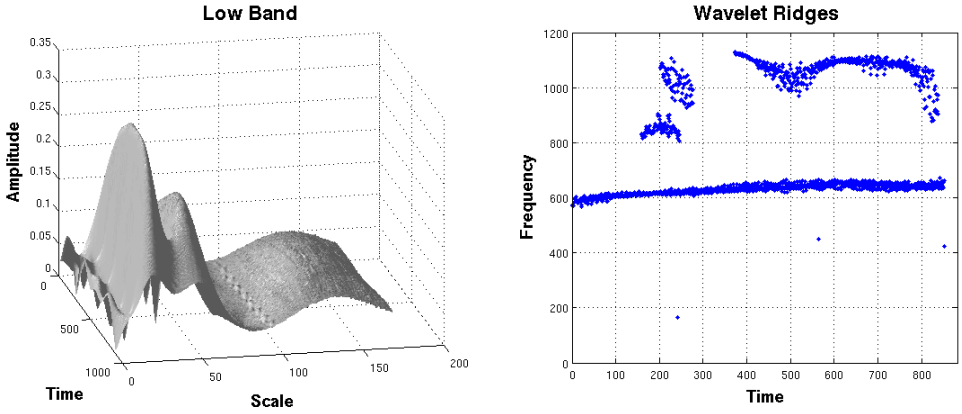


Figura 4.9: Transformada wavelet y cordilleras de muestras de voz femenina.

La figura 4.9 muestra la cordillera principal (también en la región de pitch), está centrada en un valor de frecuencia más alto, pero existen ciertas características en común en ambos géneros. Las características deben ser interpretadas de tal forma que puedan ser detectadas por el algoritmo desarrollado. Las tres condiciones principales para la detección son:

1. Duración temporal mínima
2. Frecuencia máxima
3. Desviación estándar máxima de la frecuencia

La duración temporal de la cordillera detectada debe ser mayor al 44 % de la trama analizada, en el caso de una subtrama de 20 ms debe se mayor a 8.8 ms. Hay que señalar que en la mayoría de las muestras analizadas la duración de las cordilleras detectadas supera dicho umbral. En las figuras 4.8 y 4.9 la duración temporal de la cordillera se muestra a lo largo del eje x.

El umbral de frecuencia se estableció en 800 Hz. Las frecuencias detectadas por encima de este valor en las muestras corresponden a voces femeninas cantadas, pero en términos prácticos no es común establecer una conversación cantando en lugar de hablando, dicha

situación existe en la ópera, pero entonces la señal se clasificaría como audio.

La última condición básica es la desviación estándar, la cual básicamente busca una cordillera que sea "recta", la existencia de curvas puede ser un indicador de presencia de música (tambié aplica para cordilleras muy gruesas). En algunos casos donde hay música, en ambos lados de la cordillera tiende a haber un ensanchamiento de los extremos, la forma puede recordar a un reloj de arena en posición horizontal (más delgado en el centro). La desviación estándar considerando este análisis es de σ =100 Hz.

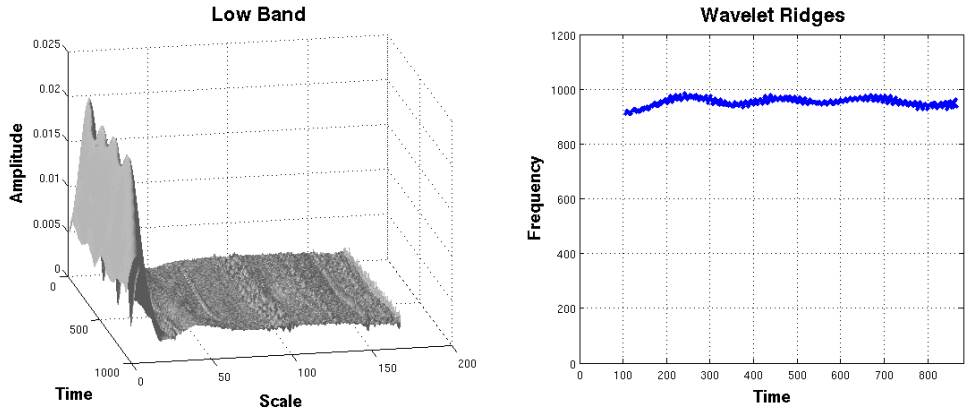


Figura 4.10: Transformada wavelet y cordilleras de muestras de un violín.

La figura 4.10 muestra la CWT y la cordillera del análisis de un solo violín tocando una nota.

Una comparación humana entre las representaciones gráficas de las figuras 4.8, 4.9 y 4.10 pude mostrar diferencias obvias, pero cuando se trata de un algoritmo computacional, la manera de determinar las diferencias puede no ser tan sencilla. Al aplicar las condiciones previamente establecidas surge una clasificación adecuada. Otros tipos de música (i.e. rock, metal) tienen un comportamiento caótico debido a la estridencia, como resultado se detectan varias cordilleras con curvas, lo cual las hace fácilmente distinguibles.

Es posible experimentar con diversos valores para los umbrales de las reglas previamente descritas con la finalidad de encontrar un mejor desempeño. La banda de altas frecuencias tiene un comportamiento más complejo, pero puede ayudar en casos especiales i.e. la detección de cordilleras con una definición tan clara como las de las bandas de frecuencia baja, ya que un indicador como la presencia de harmónicos solamente sucede en música.

4.5. Precisión en la identificación

Cuando el análisis se realiza en fonemas no vocálicos, la detección tiende a fallar. Lo anterior se debe a los rápidos cambios de amplitud, lo que impide que el algoritmo detecte las cordilleras. La representación gráfica muestra un comportamiento caótico y por ende no se detecta un patrón. Durante el análisis de varias muestras surgió un detalle particular: algunas veces el algoritmo de detección encontró muchas cordilleras pequeñas (en algunos casos más de 100 con una duración de 0.6ms) o ninguna cordillera. Debido a esta observación se creó una nueva condición; si 2 o más tramas de 20 ms cumplen con la condición de exceso de cordilleras o ninguna cordillera, la supertrama de 100 ms se consideraría inválida y se elige un nuevo punto inicial para el análisis 100 ms más adelante del anterior. El desplazamiento de supertrama sólo se realiza en 5 ocasiones. Al aplicar esta regla se evita incurrir en extender el análisis de manera indefinida. Se percibió una mejora al incrementarse la eficiencia.
La detección de voz masculina presenta un mejor desempeño cuando el análisis se realiza en segmentos donde ocurre un fonema vocálico. Esta afirmación también aplica para muestras de voz femeninas. Vale la pena señalar que al aplicar el criterio anterior a una señal de voz femenina cantando "a capella", se encontraron patrones muy similares a los del piano, por lo tanto la decisión fue tomada como audio en lugar de voz. Para distinguir entre este tipo de casos específicos, se podría aplicar una detección de patrones orientada a la alta frecuencia. Por otro lado está el caso en el que existe música instrumental en de fondo y voz cantada, el caso particular es el de un hombre cantando y música jazz. La mezcla entre los instrumentos y la voz resaltaba las características de la voz, y como consecuencia el análisis no distingue de manera clara la diferencia entre el instrumento y la voz, por lo tanto la cordillera más prominente se interpreta por el algoritmo como voz.

La tabla 4.2 muestra el resultado de un análisis estadístico para determinar la eficiencia del algoritmo aplicado a un conjunto de 150 muestras, las voces humanas (femeninas o masculinas) siempre son una palabra, algunas de ellas monosilábicas (para este caso particular la mayor parte del tiempo la decisión fue incorrecta). Para las muestras de música la diversidad en estilos fue la característica más importante. La tabla 4.3 muestra el número

Cuadro 4.2: Eficiencia en la clasificación de señales acústicas.

Tipo	**Muestras**	**Eficiencia**
Voz Masculina	51	58 %
Voz Femenina	34	70 %
Música	65	53 %
Eficiencia total 60 %		

de muestras elegidas para cada género musical. Las características son muy diferentes, en la sección clásica se incluye un conjunto de violines o sólo uno. El alto parecido entre un violín y la voz humana es impresionante, pero la identificación de esta similitud no es nueva, como se señaló en 1636 por Marin Mersenne en *Harmonie Universelle* "El violín es capaz de imitar

Cuadro 4.3: Géneros musicales utilizados para la identificación de muestras.

Géneros	**Muestras**
Black metal	9
Doom metal	8
Gothic metal	2
Jazz femenino	3
Jazz masculino	4
Barroco	3
Clásico	13
Ópera	7
Piano	7
Hip-hop	8
Lluvia	1

varios instrumentos y voces...” [42].

El conjunto de muestras mediante las cuales se diseñó el módulo de decisión fue el mismo que se empleó al realizar la evaluación de eficiencia, esta puede ser una de las razones por las cuales la eficiencia se percibe alta. La afirmación anterior fue una observación realizada durante la presentación de este trabajo en [43]. El modelo se diseñó a partir de observaciones generales, idealmente se debería probar la eficiencia con un conjunto de muestras ajento al empleado previamente, pero deberían tener las mismas características estructurales previamente señaladas (sin silencios al inicio de la muestra).

La detección por medio de un análisis basado en Wavelets representa un proceso robusto con un alto consumo de recursos sin embargo, la tendencia tecnológica apunta hacia dispositivos con capacidades superiores (en velocidad de procesamiento y almacenamiento), más compactos y económicos (en términos de consumo de energía), así que el modelo propuesto sólo marca una tendencia.

Capítulo 5
Codificación unificada

La simulación de codificación unificada se realizó con base en los módulos de decisión descritos en los capítulos 3 y 4. Los resultados más sobresalientes se obtuvieron al aplicar el método de decisión basado en las cordilleras de wavelets.

Como se estableció en el capítulo 4, el análisis de la señal se realiza dividiéndola en subtramas de 20 ms, la decisión se realiza por cada super trama de 100 ms. Dicha propuesta resulta ser suficiente al identificar archivos cuyo contenido es de un solo tipo, pero cuando se trata de un archivo que puede incluir contenido tanto de voz como audio o una mezcla de ambos, el análisis se debe realizar a toda la señal (archivo). Para efectos de la simulación realizada se debe considerar que un fragmento de 100 ms no es distinguible en una secuencia acústica por lo tanto se ha decidido utilizar como unidad mínima de codificación 500 ms. Nuevamente la decisión se realiza de acuerdo al criterio de mayoría, es decir si tres super tramas de 100 ms dentro de un segmento de 500 ms son de voz, el segmento se codificará utilizando el códec AMR-WB, si son de audio se codificará utilizando el códec HE-AACv2.

La figura 5.1 muestra de manera gráfica cómo se codifica un segmento de 500 ms. Para ilustrar la decisión que se toma en una unidad mínima de codificación se señalan tres pasos:

1. Análisis, representa el proceso de identificación de subtrama.

2. Selección, tipo de subtrama predominante será asignado a la super trama.

3. Codificación, de manera similar a la selección, se asignará al segmento un códec específico de acuerdo a la mayoría de super tramas.

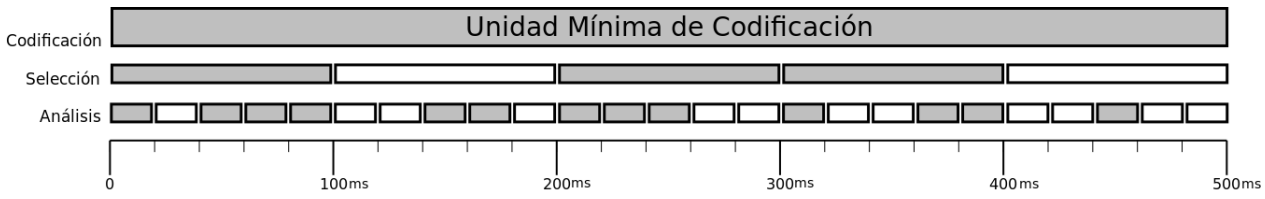


Figura 5.1: Codificación segmento de 500 ms

El objetivo principal de la codificación es reducir el "tamaño" de un archivo, ya sea con fines de almacenamiento o transmisión y difusión. Actualmente el interés en el desarrollo de códecs más eficientes está enfocado a la industria de la distribución de contenidos. Por citar un ejemplo se considera el deseo de alguna persona por escuchar una canción en un dispositivo multimedia portátil mientras se desplaza en el transporte público. El esquema tradicional asumía que se poseía el archivo de música digital, actualmente la tendencia indica que el contenido deseado se consulta desde algún proveedor de servicios remoto (acceso bajo demanda). Debido a este esquema de acceso a la información y con la intención de no saturar las redes de comunicaciones, este tipo de servicios no se rigen bajo el paradigma de "*packet oriented*", por el contrario si se desea descargar el contenidio y almacenarlo el esquema cambia. Retomando la idea de acceso bajo demanda, dependiendo de la tecnología de acceso al medio se asignan anchos de banda específicos para cada bloque multimedia (audio, video).

En el artículo [21] se especifican las tasas de bit permitidas para cada tipo de tecnología, esa información resulta particularmente interesante ya que sirve de guía para realizar las codificaciones, ya que toma como referencia las tasas de bit disponibles. La tabla 5.1 muestra las tasas de bit disponibles para cada tecnología particular. Una ventaja al utilizar el esquema de codificación unificada se deriva de los cambios en la tasa de bit empleada instantáneamente, se tiene un ancho de banda determinado, pero de acuerdo al esquema de codificación unificada planteado existirán espacios temporales en los que no se utilice por completo el ancho de banda asignado, por ejemplo cuando el segmento acústico del contenidio multimedia corresponda únicamente a voz. En ese esquema la codificación de voz tendría una tasa de bit menor a la asignada, por lo tanto se desperdiciaría una fracción del ancho de banda total disponible. Se plantea la posibilidad del uso de este códec en una tecnología de acceso adaptable, que modifique el ancho de banda asignado a cada segmento de acuerdo a la demanda y disponibilidad, por ejemplo cuando se transmita voz, agregar la porción no utilizada del ancho de banda designado a la transmisión de video y de esta forma incrementar el ancho de banda.

5.1. Reducción en la tasa de bit

Al realizar la segunda parte de la simulación, los códecs empleados fueron diferentes, la codificación AMR-WB se realizó utilizando un códec provisto por Nokia [44], la codificación HE-AACv2 se realizó con el mismo cócec, el de Nero. El cambio se debió a que la versión de Nokia permite codificar los archivos y almacenarlos en ese mismo formato, contrario a lo que sucedió en la primera parte de la simulación, el procesamiento del archivo consistía en la codificación y decodificación entregando los archivos en formato .wav, para efectos del primer análisis resultó conveniente para la simulación de codificación es necesario mantener el archivo en el formato comprimido.

Debido a características específicas del nuevo códec empleado las tasas de bit empleadas

Cuadro 5.1: Disponibilidad de tasas de bit por tecnología

		Audio		Audio-visual	
Transporte	RAT	Total	Audio (tasas netas)	Total	Audio (tasas netas)
PSS	UTRAN	64 kb/s	48 kb/s	64 kb/s	∼14 kb/s
	GPRS	36 kb/s	24 kb/s	36 kb/s	<∼10 kb/s
MBMS Difusión	UTRAN	64 kb/s	48 kb/s	64/(128) kb/s	12-16/(24) kb/s
	GPRS	36 kb/s	24 kb/s	36 kb/s	<∼10 kb/s
MMS	UTRAN	100 kB	0.5 min * 24 kb/s	75 kB video + 25 kB audio	20 seg * 10 kb/s
	GPRS		1 min * 14 kb/s		
MBMS Descarga	UTRAN	300 kB	1.5 min * 24 kb/s	225 kB video + 75 kB audio	60 seg * 10 kb/s
	GPRS		3 min * 14 kb/s		

difieren respecto a las utilizadas en el capítulo 3, además el objetivo de esta simulación no es el análisis, es la codificación, por lo tanto la asignación fue diferente, ahora con miras al aprovechamiento de segmentos de ancho de banda específicos. La tabla 5.2 muestra las tasas de bit que se emplearon en la simulación de codificación, la tasa de bit de una codificación de audio de alta calidad (muestreo a 44.1 kHz y 16 bits/muestra) es de aproximadamente 689.1 kb/s.

Cuadro 5.2: Tasas de bit empleadas para simulación de codificación

Códec	Tasas de bit (kb/s)			
AMR-WB	8.85	15.85	19.85	23.85
HE-AACv2	16	24	72	128

El desarrollo de esta simulación se realizó utilizando un conjunto de archivos específicos la tabla 5.3 muestra el tipo de archivos seleccionados, esatas muestras son un subconjunto de aquellas clasificadas de manera exitosas (correcta) por el método descrito en el capítulo 4.

Cuadro 5.3: Muestras empleadas para simulación de codificación

Voz	masculina (7)
Voz	femenina (2)
Audio	música hip-hop (3)
Audio	música metal (3)

5.1.1. Compresión

Con fines ilustrativos se examinan las tasas de compresión al codificar algunas de las muestras del conjunto de archivos seleccionados, todos se codificaron utilizando ambos códecs con las tasas de bit antes mencionadas (independiente del tipo de señal acústica, voz o audio). La ecuación 5.1 indica cómo se calculó el nivel de compresión mediante el uso de cada codificador.

$$compresión\,\% = 100 - \left(\frac{Tamaño\,comprimido}{Tamaño\,original} * 100\right) \tag{5.1}$$

La tabla 5.4 muestra los rangos de compresión para los archivos de voz.

Cuadro 5.4: Tasas de compresión para archivos de voz

VOZ				
AMR-WB	*Tasas de bit (kb/s)*			
	8.85	*15.85*	*19.85*	*23.85*
Señal	*compresión (%)*			
masc	99.32	98.85	98.57	98.29
fem	98.63	97.70	97.10	96.54
HE-AACv2	*Tasas de bit (kb/s)*			
	16	*24*	*72*	*128*
Señal	*compresión (%)*			
masc	97.28	96.69	92.89	88.81
fem	95.54	94.44	87.28	78.99

La tabla 5.5 muestra los rangos de compresión para los archivos de audio.

Cuadro 5.5: Tasas de compresión para archivos de audio

AUDIO				
AMR-WB	*Tasas de bit (kb/s)*			
	8.85	*15.85*	*19.85*	*23.85*
Señal	*compresión (%)*			
hip-hop	99.32	98.85	98.57	98.29
metal	99.32	98.84	98.56	98.28
HE-AACv2	*Tasas de bit (kb/s)*			
	16	*24*	*72*	*128*
Señal	*compresión (%)*			
hip-hop	97.44	96.88	93.09	89.02
metal	97.70	97.17	93.37	89.40

5.2. Simulación

La prueba fundamental consiste en identificar los tipos de señal en una secuencia sonora. Un archivo ideal para medir el desempeño del códec sería aquel que combinara fragmentos de voz con audio en intervalos aleatorios, además de incluir algún fragmento mixto. Se construyó un archivo combinando los archivos de audio previamente descritos, existe empalme entre cada una de las secciones, no se introdujeron cambios abruptos (generalmente identificados como un salto o un silencio).

La simulación de codificación está constituída por dos pasos esenciales, la identificación y separación en secciones y la codificación de las secciones. Al iniciar el procesamiento de una señal se realiza el análisis de un archivo .wav en Matlab. Como resultado del análisis, el archivo original se segmenta en varios archivos .wav, cada uno está identificado por medio de un número de secuencia y el tipo de señal mediante el cual fue catalogado. La estructura del nombre que se le otroga a cada uno de los segmentos es de gran importanica ya que inidica la secuencia el tipo de códec que se utilizará. Por ejemplo, si el 5º segmento de un archivo mixto se catalogó como voz el archivo tendrá el siguiente nombre: 5voz.wav. Un proceso batch generado para línea de comando identifica todos los archivos dentro de un directorio y posterirormente codifica cada uno de ellos de acuerdo al nombre utilizando el códec adecuado (AMR-WB para voz o HE-AACv2para audio). Después un programa desarrollado en C recopila todos los archivos codificados en uno solo, algo asi como un envolvente que tendrá le extensión .usac. Los nombres de cada archivo se conservarán ya que proveen la información necesaria para reagrupar los segmentos en orden (previa decodificación) y generar la nueva versión del archivo original.

La segmentación la realiza Matlab ya que las versiones más recientes tienen la capacidad de escribir archivos .wav a partir de arreglos, entonces la agrupación en segmentos será en múltiplos de 500 ms, de tal forma que si existen dos segmentos consecutivos o más prete-

necientes a la misma clasificación (voz o audio) serán agrupados en un mismo archivo .wav, lo anterior tiene como objetivo evitar la generación de información extra a partir de los encabezados de los archivos .wav, ya que la simulación de codificación es un proceso un tanto burdo que sólo agrupa archivos. En la figura 5.2 se ilustra el procesamiento de un archivo sonoro para generar un archivo codificador unificado de voz y audio (USAC por sus siglas en inglés) . Las tasas de bit empleadas al generar el archivo USAC deben especificarse antes de codificar los segmentos, se proponen 4 tipos de codificación de acuerdo a la calidad deseada:

- Alta.- segmentos de voz a 23.85 kb/s y de audio a 128 kb/s
- Media/Alta.- segmentos de voz a 19.85 kb/s y de audio a 72 kb/s
- Media/Baja.- segmentos de voz a 15.85 kb/s y de audio a 24 kb/s
- Baja.- segmentos de voz a 8.85 kb/s y de audio a 16 kb/s

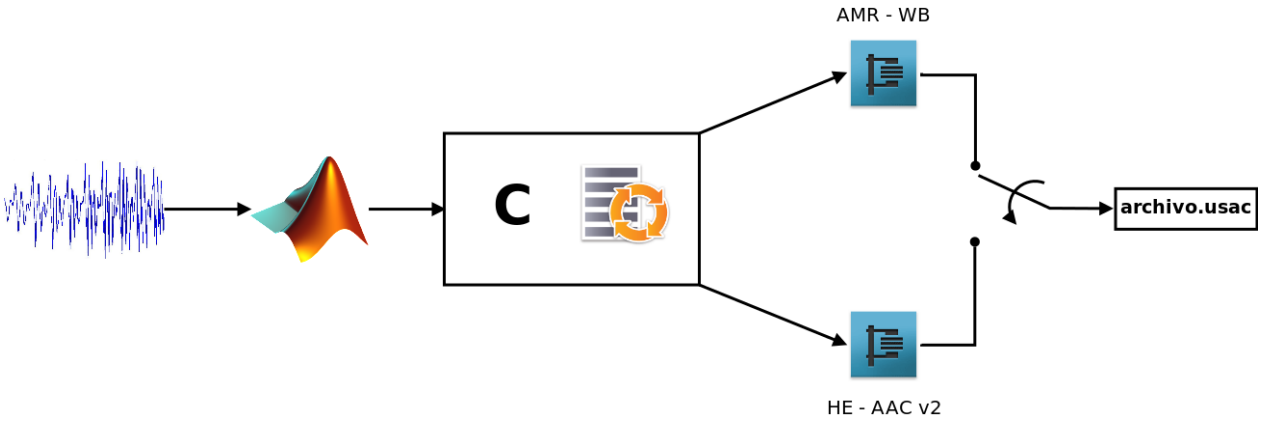


Figura 5.2: Diagrama de procesamiento USAC

5.3. Resultados

Como se mencionó previamente, el archivo de audio con contenido mixto está conformado por segmentos de audio que se habían catalogado exitosamente en un análisis previo. Es importante recordar que la distinción entre los dos tipos de señales depende de un comportamiento estadístico, al ser un solo archivo de contenidio mixto es posible que dos segmentos consecutivos de una misma categoría tengan en sus extremos (uno en el extremo final y el siguiente segmento en el extremo inicial) características del tipo de señal contraria, es decir, al que no corresponden, pero al unirse estos dos extremos se genera un segmento de 500 ms que será catalogado de otra forma y por ende habrá una discontinuidad en un segmento que de otra manera se consideraría uniforme.

Los resultados porducto de la simulación que aquí se reporta, se obtuvieron utilizando un archivo .wav con un tamaño original de 3,703,306 Bytes. Después de haber sido segmentado se crearon 14 archivos .wav, el tamaño total fue de 3,703,784, la diferencia entre el orginal y la versión segmentada es despreciable, el incremento de información debido a los encabezados para cada archivo es mínimo. El archivo original tiene una duración de 21 segundos, el análisis y segmentación se realizó en 7619.67 segundos.

En la tabla 5.6 se muestran las tasas de compresión de acuerdo a las calidades definidas previamente, la compresión se realiza comparando el tamaño de la versión segmentada original vs. el tamaño total de los segmentos codificados. El cálulo se realizó de acuerdo a la ecuación 5.1.

Cuadro 5.6: Compresión bajo esquema híbrido

Compresión USAC				
Calidad	*Baja*	*Media/Baja*	*Media/Alta*	*Alta*
compresión (%)	94.84	95.19	94.11	92.78

La tabla 5.7 muestra los rangos de compresión utilizando un solo códec y cada una de las tasas de bit correspondientes.

Cuadro 5.7: Tasas de compresión para archivo mixto

MIXTO				
AMR-WB	*Tasas de bit (kb/s)*			
	8.85	*15.85*	*19.85*	*23.85*
Señal	*compresión (%)*			
mixta	96.41	93.59	92.03	90.47
HE-AACv2	*Tasas de bit (kb/s)*			
	16	*24*	*72*	*128*
Señal	*compresión (%)*			
mixta	98.75	98.20	95.34	95.34

Al comparar ambas tablas resaltan algunos detalles: al parecer el rango de compresión es un promedio entre el uso de ambos códecs con las tasas de bit correspondientes, lo cual era de esperarse ya que prácticamente la mitad del archivo se codificó como voz y la otra como auido, debido a que así está constituido el archivo de prueba.

Para determinar el umbral de frecuencia superior para el análisis se tomó en cuenta el ancho de banda del espectro de AMR-WB, por lo generarl es de hasta 7000 Hz, pero el análisis no mostró ningúna ventaja al utilizar un valor más alto sin embargo, tuvo un impacto importante en el tiempo necesario para realizar la clasificación de alguna señal, ya que fue necesario realizar más cálculos. A medida de que se incrementaba la frecuencia, la detección de coordilleras se vueve más caótica; no se detectan patrones evidentes.

Capítulo 6
Conclusiones y trabajo futuro

El trabajo presentado plantea un modelo para desarrollar un codificador-decodificador basado en la elección del códec más adecuado dependiendo del tipo de señal acústica de entrada al sistema. La parte más importante del trabajo radica en el módulo de decisión que determina si la señal de entrada al sistema es voz o audio. Para ello se desarrollaron dos técnicas, una basada en la razón señal a ruido y otra basada en la detección de cordilleras obtenidas a través de la transformada continua wavelet.

Los resultados de la técnica basada en la razón señal a ruido muestran que: al utilizar tasas de bit bajas en la codificación de señales de audio, el desempeño de ambos códecs es muy similar, y al aumentar la tasa de bit muestra un mejor desempeño el códec diseñado para señales de voz AMR-WB. Las gráficas, que son el resultado de la evaluación SNR, no corresponden a lo que se esperaría, que el códec AMR-WB mostrara un mejor desempeño que el códec HE-AACv2; dicho resultado se puede justificar debido a la estructura del segundo códec, ya que en realidad se repite la operación de separación por bandas de frecuencias para acotar las señales que se encuentran dentro de los umbrales establecidos para voz, por lo tanto siempre obtendrá mejores resultados AMR-WB, puesto que la señal ya había sido sometida a un proceso de selección, en cierta forma el resultado está optimizado para esta prueba ya que el códec establece los criterios de selección con base en resultados de un análisis SNR. Se debió haber buscado un codificador ACELP que no incluyera la extensión de frecuencia.

La técnica basada en SNR muestra una deficiencia importante, ya que el códec AMR-WB está basado en el cálculo de índices SNR por lo tanto es obvio que el resultado se inclinará por el uso de este códec ya que en cierta forma está optimizado para ello. A pesar de que el método mostró una eficiencia baja al compararlo con los métodos analizados en la literatura, existe mucho trabajo que se puede realizar para incrementar su eficiencia, por ejemplo se pueden aplicar más reglas así como extender el análisis a las bandas de frecuencias superiores.

Los resultados de la CWT muestran diferentes patrones además de la frecuencia central de la cordillera principal, el "ancho" de la base también resulta una característica importante que debe ser tomada en cuenta y analizada con mayor detenimiento. Las muestras de voz por lo general muestran una área de cobertura mayor en la base, los instrumentos musicales muestran una tendencia a ser más angostos. En algunas muestras de voz masculina aparece

una base que se bifurca dando origen a dos cordilleras. Una desventaja del algoritmo que se utilizó para encontrar las cordilleras se identifica como el hecho de que sólo puede detectar las cimas y no provee información referente a la base. Al modificar el algoritmo de búsqueda se podría obtener una mejora significativa. El método presentado no es rápido, pero se pueden realizar varias mejoras ya que en esta etapa se realizaron simulaciones utilizando Matlab, en términos de eficiencia del algoritmo hay mucho por hacer.

6.1. Optimización de código

Para el desarrollo del proyecto descrito en los capítulos anteriores, se utilizaron códecs comerciales para cumplir con los estándares industriales. Algunos códecs son de código abierto y otros no. Actualmente el algoritmo de distinción de tramas es independiente del proceso de codificación, debido a esta razón se incrementa sustancialmente el tiempo para la identificación y construcción del archivo codificado bajo el esquema unificado. Se identificó que una de las optimizaciones de código más significativas radica en la inclusión del código de codificación en el código del algoritmo de análisis.

El proceso ideal contempla que todo el código esté desarrollado en el lenguaje de programación C, por una parte se exporta el código desarrollado en Matlab (junto con las bibliotecas utilizadas) en forma de funciones, en el proceso final la identificación sería el llamado de dichas funciones. La forma en la que fue estructurado el código en Matlab permite contar con cierta modularidad.

6.2. Mejoras en la identificación

El algoritmo de identificación basado en la detección de cordilleras de wavelets es susceptible a mejoras: durante el desarrollo de la investigación se identificó que no sólo las cimas aportan elementos para la distinción entre voz y audio, la presencia de valles también puede proporcionar parámetros de identificación, así como la "distancia.^{en}tre la cima y el valle en un fragmento sonoro.

Las cordilleras identificadas pueden ser la cima de una montaña cuya base sea ancha o delgada este factor puede ayudar a la identificación de sonidos de instrumentos específicos, ya que mientras menor sea el ancho de banda de un sonido, es más probable que pertenezca a un instrumento. La caracterización de dichas bases puede incrementar la eficiencia en la identificación, ya que el algoritmo de las cordilleras sólo traza la cima de éstas. Durante el desarrollo de la investigación se observó que las señales de voz tienden a formar montañas con una base más amplia.

Existen bandas de frecuencias que no fueron analizadas ya que incrementan el tiempo del proceso de identificación. Mediante el uso de cómputo en paralelo (vgr. GPU), se pueden asignar distintos hilos al análisis de bandas específicas y disminuir el tiempo de cómputo. Un procesador (o varios procesadores) con núcleos múltiples analizarían bandas específicas, al realizar un análisis centrado en bandadas específicas se podrían determinar características particulares, mediante este análisis se puede determinar si el contenido espectral en una banda es característico de señales de voz o de audio. La presencia de cordilleras en bandas específicas también podría contribuir sustancialmente a la identificación de instrumentos. Cada procesador (o núcleo) que ejecute el código contribuiría con una decisión para determinar si el segmento analizado corresponde a una señal de voz o audio, este esquema representaría un análisis exhaustivo que ayudaría a codificar con mayor precisión.

La evaluación de la precisión en la identificación de segmentos de audio se realizó utilizando el conjunto de muestras con el que se entrenó al sistema, para un proceso de mejora continua, se deberían incorporar más muestras además de utilizar un conjunto diferente para la evaluación de la precisión en la discriminación.

Apéndice A
Acrónimos

AAC codificador de audio avanzado (AAC por sus siglas en inglés) Advanced Audio Coding

AMR-WB ancho de banda multirango adaptiva (AMRWB pos sus siglas en inglés) Adaptative Multi-Rate Wideband

PCM modulación de impulsos codificados (PCM por sus siglas en inglés) Pulse Code Modultation

3GPP proyecto de asociación de 3ra generación (3GPP por sus siglas en inglés) 3rd Generation Partnership Project

LPC coeficientes de predicción lineal (LPC por sus siglas en inglés) Linear Prediction Coeficients

ACELP predicción lineal algebráica de código excitado (ACELP por sus siglas en inglés) Algebraic Code-Excited Linear Prediction

CELP predicción lineal de código excitado (CELP por sus siglas en inglés) Code-Excited Linear Prediction

TPC codificación por transformada predictiva (TPC por sus siglas en inglés) Transform Predictive Coding

TCX Excitación de transformada codificada (TCX por sus siglas en inglés) Transform Coded Excitation

MDCT transformada modificada de coseno discreto (MDCT por sus siglas en inglés) Modified Discrete Cosine Transform

FFT transformada rápida de Fourier (FFT por sus siglas en inglés) Fast Fourier Transform

OTT uno a dos (OTT por sus siglas en inglés) One To Two

TDA alias en dominio de tiempo (TDA por sus siglas en inglés) Time Domain Aliasing

MUSHRA estímulo múltiple con ancla y referencia escondida (MUSHRA por sus siglas en inglés) MUltiple Stimuli with Hidden Reference and Anchor

ZIR respuesta al impulso cero (ZIR por sus siglas en inglés) Zero Impulse Response

BWE extensión de ancho de banda (BWE por sus siglas en inglés) Bandwidth Extension

ITU-T Unión internacional de telecomunicaciones sector de estandarización de telecomunicaciones (ITU-T por sus siglas en inglés) International Telecommunications Union Standardization Sector

MPEG grupo de expertos de fotografías en movimiento (MPEG por sus siglas en inglés) Moving Picture Experts Group

HE-AACv2 codificación avanzada de audio de alta eficiencia versión 2 (HE-AACv2 por sus siglas en inglés) High-Efficiency Advanced Audio Coding version 2

LC-AAC codificación avanzada de audio de baja complejidad (LCAAC por sus siglas en inglés) Low Complexity Advanced Audio Coding

HE-AAC codificación avanzada de audio de alta eficiencia (HE-AAC por sus siglas en inglés) High-Efficiency Advanced Audio Coding

AAC codificación avanzada de audio (AAC por sus siglas en inglés) Advanced Audio Coding

SBR replicación espectral de banda (SBR por sus siglas en inglés) Spectral Band Replication

eSBR replicación espectral de banda mejorado (eSBR por sus siglas en inglés) Enhanced Spectral Band Replication

PS estéreo paramétrico (PS por sus siglas en inglés) Parametric Stereo

STFT Transformada de Fourier en corto plazo (STFT por sus siglas en inglés) Short Time Fourier Transform

VQ cuantificación Vectorial (VQ por sus siglas en inglés) Vector Quantization

ISP par espectral de immittancia (ISP por sus siglas en inglés) Immittance Spectral Pairs

QMF filtro espejo en cuadratura (QMF por sus siglas en inglés) Quadrature Mirror Filter

SNR razón señal a ruido (SNR por sus siglas en inglés) Signal to Noise Ratio

CWT transformada Wavelet continua (CWT por sus siglas en inglés) Continuous Wavelet Transform

USAC codificador unificado de voz y audio (USAC por sus siglas en inglés) Unified Speech and Audio Coding

IID diferencias de intensidad inter-canal (IID por sus siglas en inglés) Interchannel Intensity Difference

IPD diferencias de fase inter-canal (IPD por sus siglas en inglés) Interchannel Phase Difference

IC coherencia inter-canal (IC por sus siglas en inglés) Interchannel Coherence

OPD diferencia de fase general (OPD por sus siglas en inglés) Overall Phase Difference

TNS modelado temporal de ruido (TNS por sus siglas en inglés) Temporal Noise Shaping

PSS servicio de transmisión por conmutación de paquetes (PSS por sus siglas en inglés) Packet-Switched Streaming

MBMS servicio de difusión múltiple multimedia (MBMS por sus siglas en inglés) Multimedia Broadcast/Multicast Service

FAC cancelación de alias adelante (FAC por sus siglas en inglés) Forward Alias Cancellation

FDNS modelado de ruido en el dominio de la frecuencia (FDNS por sus siglas en inglés) Frecuency Domain Noise Shaping

BDT árboles de decisión binarios (BDT por sus siglas en inglés) Binary Decision Tree

ERBF función base de radio exponencial (ERBF por sus siglas en inglés) Exponential Radial Basis Function

NFL línea característica más cercana (NFL por sus siglas en inglés) Nearest Feature Line

LTV lineal variante en el tiempo (LTV por sus siglas en inglés) Linear Time-Variat

GMM modelo mixto Gaussiano (GMM por sus siglas en inglés) Gaussian Mixed Model

MAP máximo a posteriori (MAP por sus siglas en inglés) Maximum A-Posteriori

STE energía de corto plazo (STE por sus siglas en inglés) Short Time Energy

SF Flujo espectral (SF por sus siglas en inglés) Spectral Flux

LSP DD Distancia divergente de pares espectrales lineales (LSP por sus siglas en inglés) Line Spectrum Pairs Divergence Distance

BP Periodicidad de banda(BP por sus siglas en inglés) Band Periodicity

NFR Tasa de trama de ruido (NFR por sus siglas en inglés) Noise Frame Rate

ZCR tasa de cruce por cero (ZCR por sus siglas en inglés) Zero Cross Rate

ASA Análisis de escena de audio (ASA por sus siglas en inglés) Audio Scene Analysis

eLTP predictor de largo plazo mejorado (eLTP por sus siglas en inglés) Enhanced Long Term Prediction

Lista de Figuras

Lista de Tablas

Referencias

[1] E. Wold, T. Blum, D. Keislar, y J. Wheaten, "Content-based classification, search, and retrieval of audio," *IEEE MultiMedia*, vol. 3, pp. 27 –36, fall 1996.

[2] J. Saunders, "Real-time discrimination of broadcast speech/music," vol. 2, pp. 993–996 vol. 2, 1996.

[3] E. Scheirer y M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," vol. 2, pp. 1331 –1334, apr 1997.

[4] Y. Nakajima, Y. Lu, M. Sugano, A. Yoneyama, H. Yamagihara, y A. Kurematsu, "A fast audio classification from mpeg coded data," vol. 6, pp. 3005 –3008 vol.6, mar 1999.

[5] S. Ramprashad, "A multimode transform predictive coder (mtpc) for speech and audio," pp. 10 –12, 1999.

[6] E. Ambikairajah, J. Epps, y L. Lin, "Wideband speech and audio coding using gammatone filter banks," vol. 2, pp. 773 –776, 2001.

[7] L. Lu, H.-J. Zhang, y H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 504 – 516, oct 2002.

[8] S.-H. Chen y J.-F. Wang, "Noise-robust pitch detection method using wavelet transform with aliasing compensation," *IEEE Proceedings on Vision, Image and Signal Processing*, vol. 149, pp. 327 – 334, dec 2002.

[9] D. Gerhard, "Audio signal classification: History and current techniques," *IEEE Transactions on Speech and Audio Processing*, nov 2003.

[10] C.-C. Lin, S.-H. Chen, T.-K. Truong, y Y. Chang, "Audio classification and categorization based on wavelets and support vector machine," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 644 – 651, sept. 2005.

[11] R. Shantha Selva Kumari, D. Sugumar, y V. Sadasivam, "Audio signal classification based on optimal wavelet and support vector machine," vol. 2, pp. 544 –548, dec. 2007.

[12] T. Moriya, "Technologies for speech and audio coding," pp. 148 –149, may 2009.

[13] M. Lu, S. Zhang, y W. Dou, "Dual-mode switching used for unified speech and audio codec," pp. 700 –704, nov. 2010.

[14] J. Song, H. o Oh, y H.-G. Kong, "Enhanced long-term predictor for unified speech and audio coding," pp. 505 –508, may 2011.

[15] "Mp3 and aac explained," *AES 17 International Conference on High Quality Audio Coding*, 1999.

[16] "Spectral band replication, a novel approach in audio coding," *AES 112 convention*, vol. 5553, 2002.

[17] "Low complexity parametric stereo coding," *AES 116 convention*, 2004.

[18] M. Stefa y M. Geral, "Mpeg-4 he-aac v2 audio coding for today's digital media world," *EBU Technical Review*, 2006.

[19] A. Gersho y R. G. M., *Vector Quantizadion and Signal Compression*. USA: Kluwer Academic Publishers, 2001.

[20] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, y K. Jarvinen, "The adaptive multirate wideband speech codec (amr-wb)," *IEEE Transactions on Speech and Audio Processing,*, vol. 10, pp. 620 – 636, nov 2002.

[21] J. Makinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, y A. Taleb, "Amr-wb+: a new audio coding standard for 3rd generation mobile audio services," vol. 2, pp. ii/1109 – ii/1112, march 2005.

[22] B. Bessette, R. Lefebvre, y R. Salami, "Universal speech/audio coding using hybrid acelp/tcx techniques," vol. 3, pp. iii/301 – iii/304, march 2005.

[23] B. Bessette, R. Salami, C. Laflamme, y R. Lefebvre, "A wideband speech and audio codec at 16/24/32 kbit/s using hybrid acelp/tcx techniques," pp. 7 –9, 1999.

[24] S.-W. Shin, C.-H. Lee, H. o Oh, y H.-G. Kang, "Designing a unified speech/audio codec by adopting a single channel harmonic source separation module," pp. 185 –188, april 2008.

[25] "http://www.freesound.org."

[26] "http://www.3gpp.org/ftp/specs/html-info/26304.htm."

[27] "http://www.voiceage.com/freecodecs.php."

[28] "http://www.nero.com/enu/downloads-nerodigital-nero-aac-codec.php."

[29] N. Shibata, "http://shibatch.sourceforge.net/."

[30] J. A. Phillips, "http://www.mainly.me.uk/resampling/index.html."

[31] D. Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, vol. 93, pp. 429 –441, november 1946.

[32] M. Vetterli y J. Kovačević, *Wavelets and Subband Coding*. Prentice Hall, 1995.

[33] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, second ed., 1996.

[34] "Two-dimensional directional wavelets and the scale-angle representation," *Signal Processing, 52, 3, 259*, 1996.

[35] J. W. P. S. Addison y T. Feng, "Low-oscillation complex wavelets," *Sound Vibration, 254, 4, 733*, 2002.

[36] W. J. Staszewski, "Identification of non-linear systems using multi-scale ridges and skeletons of the wavelet transform," *Sound Vibration, 214, 4, 639*, 1998.

[37] L. G. M. Ruzzene, A. Fasana y B. Piombo, "Natural frequencies and dampings identification using wavelet transform: Application to real data," *Mechanical Systems Signal Processing, 11, 2, 207*, 1997.

[38] N. Delprat, B. Escudie, P. Guillemain, R. Kronland-Martinet, P. Tchamitchian, y B. Torresani, "Asymptotic wavelet and gabor analysis: extraction of instantaneous frequencies," *IEEE Transactions on Information Theory*, vol. 38, pp. 644 –664, mar 1992.

[39] V. J., *Theorie et application de la notion de signal analytique*. Cables Trans, 1984.

[40] "Estimation of instantaneous frequency of signals using the continuous wavelet transform," *University of Southern California, Department of Civil engineering*, 2001.

[41] J. M. Lilly, "http://www.jmlilly.net/."

[42] A. Basso, *Historia de la Música: La*. Academic Press, second ed., 1996.

[43] D. S. y A. Prieto, "Discrimination module for voice/audio signals based on wavelet ridges analysis," *AES 43rd International Congress*, 2010.

[44] Nokia, "http://www.developer.nokia.com/."

Casa abierta al tiempo

UNIVERSIDAD AUTÓNOMA METROPOLITANA

UNIDAD IZTAPALAPA - DIVISIÓN DE CIENCIAS BÁSICAS E INGENIERÍA

DISEÑO DE UN CODIFICADOR DECODIFICADOR DE VOZ Y AUDIO BAJO UN ESQUEMA UNIFICADO

Idónea Comunicación de Resultados que presenta

Daniel Edgar Sauceda Peña

Para obtener el grado de

Maestro en Ciencias

(Ciencias y Tecnologías de la Información)

Asesor: Dr. Alfonso Prieto Guerrero

Jurado Calificador:

Presidente: Dr. Sergio Suárez Guerra IPN

Secretario: Dr. Alfonso Prieto Guerrero UAM - I

Vocal: Dr. John Goddard-Close UAM - I

México, D.F. Julio de 2013