UNIVERSIDAD NACIONAL DEL LITORAL

Medidas de Información Multiresolución Aplicadas al Procesamiento de Señales de Habla

Analía Soledad CHERNIZ

Tesis de Maestría **2017**

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

MEDIDAS DE INFORMACIÓN MULTIRESOLUCIÓN APLICADAS AL PROCESAMIENTO DE SEÑALES DE HABLA

Bioing. Analía Soledad Cherniz

Tesis remitida al Comité Académico de Maestría
como parte de los requisitos para la obtención del grado de
MAGÍSTER EN COMPUTACIÓN
APLICADA A LA CIENCIA Y LA INGENIERÍA
de la
UNIVERSIDAD NACIONAL DEL LITORAL

2017

Comisión de Posgrado, Facultad de Ingeniería y Ciencias Hídricas, Ciudad Universitaria, Paraje "El Pozo", S3000, Santa Fe, Argentina.

MAESTRÍA EN COMPUTACIÓN APLICADA A LA CIENCIA Y LA INGENIERÍA

Título de la obra:

Medidas de Información Multiresolución Aplicadas al Procesamiento de Señales de Habla

Multiresolution Information Measures Applied to Speech Signal Proccessing

Autor: Bioing. Analía Soledad Cherniz

Director: Dr. Hugo Leonardo Rufiner (UNL-UNER-CONICET)

Lugar: Santa Fe, Argentina

Palabras Claves:

Medidas de información, Entropías, Divergencias,
Análisis multiresolución, Medidas de complejidad,
Cambios de dinámica, Parametrización de la señal de voz,
Reconocimiento robusto del habla, Segmentación automática de fonemas.

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

Santa Fe, 26 de julio de 2017

Como miembros del Jurado Evaluador de la Tesis de Maestría titulada ***"Medidas de información multiresolución aplicadas al procesamiento de señales de habla"***, desarrollada por la Bioing. Analía Soledad CHERNIZ, certificamos que hemos evaluado la Tesis y recomendamos que sea aceptada como parte de los requisitos para la obtención del título de Magíster en Computación Aplicada a la Ciencia y la Ingeniería. La aprobación final de esta disertación está condicionada a la presentación de dos copias encuadernadas de la versión final de la Tesis ante el Comité Académico de la Maestría en Computación Aplicada a la Ciencia y la Ingeniería.

.....
Dr. Rubén Spies

.....
Dr. Juan Carlos Gómez

.....
Dra. Patricia Pelle

.....
Dr. Marcelo Risk

Santa Fe, 26 de julio de 2017

Certifico haber leído esta Tesis preparada bajo mi dirección y recomiendo que sea aceptada como parte de los requisitos para la obtención del título de Magíster en Computación Aplicada a la Ciencia y la Ingeniería.

.....
Dr. Leonardo Rufiner
Director de Tesis

Universidad Nacional del Litoral
Facultad de Ingeniería y
Ciencias Hídricas

Secretaría de Posgrado

Ciudad Universitaria
C.C. 217
Ruta Nacional Nº 168 – Km. 472,4
(3000) Santa Fe
Tel: (54) (0342) 4575 229
Fax: (54) (0342) 4575 224
E-mail: posgrado@fich.unl.edu.ar

Dedicado a

Alexis: que siempre me sorprende con su capacidad de amor y genialidad

Lara: cuya dulzura en una caricia al alma y su sonrisa, la felicidad

Elián: porque cada día me enseña que aun tengo mucho que aprender

Carlos: mi compañero en esta aventura que es vivir

Mis padres: María del Carmen y Carlos

Y a los amigos y amigas que me ayudaron y permitieron llegar hasta acá.

Agradecimientos

"La esperanza es paradójica. Tener esperanza significa estar listo en todo momento para lo que todavía no nace, pero sin llegar a desesperarse si el nacimiento no ocurre en el lapso de nuestra vida"
Erich Fromm

Como ya sabemos, la elaboración de una tesis es un trabajo largo, que requiere esfuerzo, tiempo y dedicación. En ocasiones, conforme vamos desarrollando el trabajo se van presentando inconvenientes y problemas, que, muchas veces, son ajenos al trabajo de investigación. O bien, aparecen cambios que obligan a reorientar prioridades, posponer objetivos y desplegar otras estrategias.

Afortunadamente, cuando emprendemos este tipo de proyectos, no estamos solos. Incluso cuando aparecen situaciones donde las cosas no salen como esperamos y el desánimo se hace presente, hay personas que, de una u otra forma, nos dan una mano o una palabra de aliento para seguir adelante. Y es este espacio, al momento de redactar los agradecimientos, cuando se nos hacen presentes nuevamente. Y es a todos ellos, a quienes van dirigidas estas palabras, aunque a veces (por una cuestión de extensión) solo podamos nombrar a algunos y otros queden, sólo en el documento, en el anonimato.

Así, en primer lugar, quiero agradecer a mis directores: el Dr. Leonardo Rufiner y la Dra. María Eugenia Torres. Ellos me dieron la oportunidad de emprender esta tarea y gracias a su orientación, responsabilidad y rigor académico y científico fue posible emprender una formación completa y enriquecedora en investigación. En particular, deseo reconocer especialmente al Dr. Rufiner, pues su compromiso, paciencia y motivación fueron fundamentales para que hoy pueda cerrar esta etapa y porque excedieron considerablemente sus responsabilidades como director.

También, me gustaría agradecer el acompañamiento de aquellos integrantes del Laboratorio de Cibernética y el Laboratorio de Señales y Dinámicas no lineales, de la Facultad de Ingeniería de la UNER, y el Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional de la Facultad de Ingeniería y Ciencias Hídricas de la UNL, con quienes compartimos experiencias de aprendizaje en las diferentes etapas de este proceso. De todos ellos, quiero mencionar al Dr. César Martínez y agradecerle, además de su amistad, sus múltiples consejos y horas de escucha y ayuda, sobre todo, en los momentos donde se presentaron inconvenientes y problemas.

Por otro lado, fueron muchos los compañeros de trabajo, alguno de los cuales tengo la satisfacción de poder llamar amigos, quienes directa o indirectamente, ayudaron a que pudiera finalizar la tesis, acompañando no tanto desde lo técnico, sino

desde lo espiritual. Y porque sería una lista demasiado larga nombrar a todos y no sabría a quién dejar fuera de la misma, voy a mencionar sólo un nombre, que no puedo omitir debido a la relevancia que tuvo para este proceso, y agradecer a la Mg. Claudia Bonell por sus palabras de aliento y creer en mí.

Por último, agradezco a mi familia toda, y a quienes ya he dedicado este trabajo, y a mis otros amigos, esos que no entienden muy bien de que se trata todo esto, pero de todas formas alientan y acompañan para que tenga éxito.

¡Muchas gracias a todos!

Resumen

Las actividades realizadas en esta tesis se centraron en el tratamiento de la señal de habla mediante procesamientos basados en medidas de información multirresolución. La finalidad pretendida con el uso de estas técnicas fue obtener representaciones de dicha señal que permitieran resaltar características relacionadas con los cambios que se producen en la dinámica de la señal, debidos a la naturaleza de su generación. También se indagó acerca de la robustez de este tipo de medidas a la hora de caracterizar la señal de habla. El propósito de esta representaciones es preservar y resaltar las pistas acústicas significativas vinculadas a los cambios fonéticos y reducir o eliminar las posibles perturbaciones causadas por el ruido que pueda contaminar la señal. Para ello se plantearon codificaciones de la señal de voz, que luego se utilizaron en dos tipos de tareas: reconocimiento automático del habla y segmentación automática de fonemas.

Para obtener las representaciones de la señal de habla se utilizaron medidas de información multirresolución basadas en entropías y divergencias, como la entropía de Shannon, distancia de Kullback-Leibler, la entropía de Tsallis y su divergencia asociada y divergencia de Jensen-Shannon, a fin de evaluar la evolución temporal de la complejidad de los coeficientes de la transformada ondita continua de la señal de voz. Mediante este procedimiento se obtuvieron coeficientes basados en la entropía multiresolución continua (CME) y la divergencia multiresolución continua (CMD).

Para la tarea de reconocimiento automático del habla, se seleccionaron uno y dos coeficientes a partir de la CME y CMD (usando diferentes entropías y divergencias), que fueron concatenados a la parametrización clásica de coeficientes cepstrales en frecuencias de mel (MFCC). Se analizó el comportamiento del reconocedor utilizando las nuevas representaciones propuestas con señales contaminadas con ruido aditivo blanco y de murmullo; y se comparó su desempeño usando como referencia una codificación sólo en MFCC. Los resultados obtenidos mostraron que estas medidas de información, calculadas en el plano tiempo-escala, proveen una mejora significativa en el desempeño del sistema. Por lo que se puede considerar que el procesamiento propuesto proporciona información que permite mejorar la tarea del reconocedor bajo condiciones de ruido.

La tarea de segmentación de fonemas se llevó a cabo utilizando un algoritmo de segmentación automática independiente del texto. Para dicha aplicación se utilizaron, como entrada del sistema, representaciones del habla basadas en la CME, con entropía de Shannon, y la CMD, usando distancia de Kullback-Leibler. Se utilizó la codificación Melbank como representación de referencia para comparar el desempeño de las codificaciones propuestas. Los resultados evidenciaron que el método basado en la CMD incrementa la habilidad del algoritmo de segmentación para realizar su tarea.

Los resultados alcanzados en este trabajo muestran que la CME y CMD facilitan la detección de cambios dinámicos del tracto vocal, lo cual es clave para realizar la tarea de reconocimiento. El procesamiento antes mencionado ha presentado un comportamiento robusto ante condiciones de ruido aditivo. Por otro lado, la CMD provee información relacionada a características acústicas que tienen en cuenta las transiciones fonéticas, lo cual es de vital importancia al momento de realizar la tarea de segmentación.

Índice general

Índice de figuras

Índice de tablas

Listado de Acrónimos

A continuación, se indica el significado de los acrónimos más relevantes utilizados en el presente trabajo. Los mismos se denotan conforme a las denominaciones correspondientes en idioma inglés.

- ASR (automatic speech recognition): reconocimiento automático del habla
- CC (complex cepstrum): cepstrum complejo
- CMD (continuous multiresolution divergence): divergencia multiresolución continua
- CME (continuous multiresolution entropy): entropía multiresolución continua
- CWT (continuous wavelet transform): transformada ondita continua
- DFT (dicrete Fourier transform): transformada discreta de Fourier
- EM (expectation-maximization): maximización de la esperanza
- HMM (hidden Markov model): modelo oculto de Markov
- HMMs (hidden Markov models): modelos ocultos de Markov
- LPC (linear prediction coding): coeficientes de predicción lineal
- LTI (linear time invariant): lineal invariante en el tiempo
- MFCC (mel frequency cepstral coefficients): coeficientes cepstrales en frecuencias de mel
- ML (maximum likelihood): máxima verosimilitud
- PC (principal component): componente principal
- PCA (principal component analysis): análisis de componentes principales
- PLP (perceptual linear prediction): predicción lineal perceptual
- RASTA-PLP (RelAtive SpecTrAl perceptual linear prediction): predicción lineal perceptual espectral relativa
- RC (real cepstrum): cepstrum real
- ROC (receiver operating characteristic): característica operativa del receptor

- rv (random variable): variable aleatoria
- SNR (signal-to-noise ratio): relación señal a ruido
- WER (word error rate): tasa de error de palabra

Capítulo 1

Introducción

1.1. Objetivos

El objetivo principal de esta tesis consiste en investigar y desarrollar técnicas basadas en medidas de información multiresolución para el tratamiento de la señal de voz. La finalidad es obtener una representación óptima de dichas señales, de modo tal de enfatizar características específicas relevantes para distintas aplicaciones, tales como reconocimiento automático del habla y segmentación. El propósito de esta representación es preservar y resaltar las pistas acústicas significativas vinculadas a dichas tareas y reducir o eliminar las posibles perturbaciones causadas por el ruido que pueda contaminar la señal.

1.2. Antecedentes

A pesar de los importantes avances alcanzados en el campo del análisis, representación y modelado del habla [1, 2], aún existen problemas que no han sido satisfactoriamente resueltos. Entre estos se cuentan la degradación en el desempeño que sufren los sistemas de reconocimiento automático cuando las señales de habla están contaminadas con ruido o presentan perturbaciones externas [3, 4, 5], la segmentación automática de la señal [6, 7, 8], la correcta clasificación de unidades fonéticas altamente confundibles [9, 10], entre otros [1, 11, 12, 13, 14]. Si comparamos los sistemas artificiales de procesamiento del habla con la capacidad innata de las personas en lo que concierne a estas tareas, aun bajo condiciones adversas, está claro que todavía queda mucho por hacer en este campo [5, 15, 16, 17].

Desde hace algunos años se ha planteado la necesidad de recurrir a nuevos métodos y herramientas para solucionar aquellos problemas que, debido a las restricciones de las técnicas clásicas, no se han podido resolver satisfactoriamente [5, 10, 17, 18, 19, 20, 21].

El enfoque tradicional para el procesamiento de la señal de habla supone a ésta como generada a partir de sistemas lineales, invariantes en el tiempo y con componentes aleatorias caracterizadas por estadísticas de hasta segundo orden [22]. Estas simplificaciones constituyen una hipótesis de trabajo válida sólo como una aproximación, lo cual hace posible la utilización de herramientas matemáticas "clásicas" para el tratamiento y análisis de estas señales. Este enfoque se emplea actualmente y resulta útil para una variedad de situaciones controladas. Sin embargo, para ciertos casos, como por ejemplo el tratamiento de señales contaminadas con ruido, las

técnicas lineales presentan limitaciones [1, 3, 5].

Como sucede con la mayoría de los sistemas biológicos, las condiciones de linealidad e invarianza temporal no se cumplen en la realidad [23]. Esto lleva a suponer que es necesario disponer de otro tipo de herramientas de análisis que contemplen las características no lineales, no estacionarias y de estadística significativa de alto orden de estos sistemas. Por otro lado, desde el campo del reconocimiento robusto del habla han surgido una serie de técnicas relacionadas con modelos estadísticos, modelos fisiológicos e inteligencia artificial que proveen nuevas formas de pensar y diseñar métodos y algoritmos de procesamiento "no convencional" que permitirían atacar distintos problemas [10, 17, 19, 24].

Técnicas basadas en la teoría de onditas, la estadística de alto orden, teoría de la información, el análisis de componentes independientes y la obtención de representaciones ralas han sido utilizadas en investigaciones en el campo del procesamiento de señales derivadas de sistemas no lineales, con muy buenos resultados [25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38]. Numerosas son las aplicaciones de estas técnicas al procesamiento de imágenes [39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49]. Estos nuevos paradigmas también se han utilizado para el análisis, modelado y representación del habla [14, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60] y para la reducción de ruido en estas señales [4, 13, 53, 61, 62, 63, 64, 65, 66]. Estas técnicas aparecen como posibles soluciones a los problemas que no pueden ser abordados satisfactoriamente por los métodos clásicos debido a sus restricciones.

En la mayoría de los sistemas de procesamiento, un primer paso muy importante es la representación o parametrización de la señal de voz. Esto permite realzar características importantes de la señal y disminuir la dimensionalidad de los datos a procesar, haciendo más eficiente el tratamiento posterior de la misma.

El uso de medidas de información como las entropías de Shannon y Tsallis y sus correspondientes divergencias, se han utilizado para caracterizar la evolución temporal del grado de complejidad de la señal de habla [67, 68]. Y la combinación de dichas medidas con parámetros tradicionales, para alimentar un sistema de reconocimiento automático, han permitido mejorar el desempeño del sistema ante condiciones de ruido [68]. Estas medidas de información y otras, tales como la divergencia Jensen-Shannon, se han utilizado en diferentes aplicaciones relacionadas con el procesamiento del habla [69, 70, 71, 72, 73].

La utilización de estas medidas de complejidad se ha extendido para diferentes distribuciones espacio–temporales. En el campo del procesamiento del habla, la entropía espectral se ha utilizado en tareas relativamente simples como la segmentación y detección de silencios [74, 75] y para compresión [76]. La entropía multiresolución es una herramienta basada en la trasformada ondita, la cual da cuenta de la evolución temporal de la entropía de Shannon de los coeficientes de la transformada [77]. La misma se ha extendido, también, a la entropía generalizada de Tsallis [78]. Su combinación con la transformada ondita continua [38], lo que se conoce como entropía multiresolución continua (CME), ha permitido explorar teóricamente sus propiedades. Asimismo, la CME combinada con redes auto-organizativas de Kohonen ha sido aplicada a la segmentación de frases en español rioplatense en condiciones de ruido, obteniéndose resultados satisfactorios, en particular, en la segmentación de diptongos [79].

1.3. Vinculación del estado del arte con el tema propuesto

Como se mencionó anteriormente, existen nuevas técnicas basadas en la teoría de sistemas no lineales, no estacionarios y con estadística de alto orden que han planteando un nuevo enfoque en áreas del procesamiento de la señal de voz. Diferentes investigaciones han encontrado importantes conexiones entre la manera en la que el cerebro procesa las señales sensoriales y algunos de los principios que sustentan estas nuevas técnicas, entre los que se pueden destacar la existencia de muy pocos elementos activos para lograr la representación de cualquier señal y la independencia estadística entre estos elementos [15, 80, 81, 82, 83, 84]. De esta manera, los sistemas fisiológicos pueden eliminar o disminuir las perturbaciones ajenas a la señal de voz, pero preservando las características significativas que permiten la discriminación fonética. Esto hace pensar que dichas propuestas pueden ser altamente pertinentes para el tratamiento de esta señal [85, 86]. Esto último es coherente con la tendencia empírica a incluir las características de percepción y producción del habla en la modelización de dichos sistemas [87, 88, 89, 90, 91, 92, 93], como en el caso de la parametrización mel cepstra que ha demostrado muy buenos resultados durante mucho tiempo [22].

1.4. Descripción del Trabajo

En esta tesis se ha avanzado en el planteo de nuevas parametrizaciones basadas en la utilización de la entropía multiresolución continua, a fin de caracterizar la complejidad de la señal de habla en el plano espacio–temporal. La CME es una técnica basada en la trasformada ondita continua (CWT) de la señal, la cual da cuenta de la evolución temporal de la entropía de los coeficientes de la trasformada. La CME ha mostrado poseer un comportamiento robusto en la detección de cambios suaves en la dinámica no lineal de señales fisiológicas ante condiciones de ruido aditivo [68, 79, 94].

Las medidas de información y sus divergencias asociadas se han utilizado para obtener información acerca del evolución temporal de la complejidad de la señal de voz en el marco de un sistema de reconocimiento automático, con buenos resultados [68]. Esto motivó, no sólo la utilización de la CME para aplicaciones en el área del procesamiento del habla, sino además, la propuesta de utilizar medidas de divergencia para evaluar la complejidad de los coeficientes de la CWT; dando lugar a una nueva técnica, denominada, en este caso, como divergencia multiresolución continua (CMD) [95].

Este tipo de procesamiento se utilizó en el marco de un sistema de reconocimiento automático del habla. La información proveniente de la entropía multiresolución continua y las diferentes divergencias se incorporó a la codificación clásica en coeficientes cepstrales en frecuencias de mel (MFCC) de la señal de habla y se estudió el comportamiento del reconocedor utilizando las nuevas representaciones propuestas bajo distintas condiciones de ruido. Los resultados obtenidos mostraron que estas medidas de información, calculadas en el plano tiempo-escala, proveen una mejora significativa en el desempeño del sistema, proporcionando información valiosa que mejora la tarea del reconocedor bajo condiciones ruidosas. Esto puede estar vincu-

lado con el hecho de que la detección de cambios dinámicos del tracto vocal, que es clave para realizar el reconocimiento, se facilita mediante el uso de estas herramientas. Además, para decodificar el mensaje de la señal de habla, el sistema auditivo humano usa, simultáneamente, información de diferentes escalas temporales. Las medidas de información basadas en el análisis ondita semejan estas característica biológicas y aportan mayor información al sistema [95, 96].

También se ha propuesto la utilización de medidas de información multiresolución para realizar la segmentación de fonemas. En este contexto, se han usado para codificar la señal de habla utilizada como entrada de un segmentador automático independiente del texto. Los resultados evidenciaron que la técnica propuesta incrementa la habilidad del algoritmo de segmentación. Esto sugiere que estas medidas proveen información relacionada a características acústicas que tienen en cuenta las transiciones fonéticas [97, 98].

1.5. Organización del documento

Este documento se organiza de la siguiente manera. En el capítulo 2 se describen brevemente las bases fisiológicas de la comunicación humana. Allí, se analiza cómo se produce el habla y se estudia el aparato fonador como un modelo cuyos parámetros son variables en el tiempo. Se analiza, además, cómo se lleva a cabo la percepción del sonido y cómo es el proceso de la audición desde el oído externo hasta la corteza. El procesamiento cognitivo superior que se realiza sobre el mensaje escapa al alcance de esta tesis, por lo tanto, no se tratará el mismo en este documento. Por otra parte, además, se introducen las características principales de los diferentes métodos y técnicas utilizados para procesar la señal de habla, a fin de reducir la dimensionalidad de los datos y destacar sólo las características más relevantes de acuerdo al tipo de aplicación en el que se estén utilizando las señales.

El procesamiento de la señal de habla mediante medidas de información multiresolución se describe en detalle en el capítulo 3. En las diferentes secciones se indican los pasos necesarios y las técnicas utilizadas para implementar la CME y CMD, usando diferentes medidas de información.

La utilización de las medidas de información multiresolución para la extracción de características en un sistema de reconocimiento automático del habla se describe en el capítulo 4. Y en el capítulo 5 se muestran y discuten los resultados obtenidos luego de aplicar estas medidas para realizar la segmentación automática de fonemas.

Finalmente, el documento cierra con las conclusiones generales de esta tesis y las propuestas de posibles trabajos a futuro sobre los temas tratados.

Capítulo 2

Comunicación Humana: Bases Fisiológicas y Análisis del Habla

2.1. Introducción

El habla es el acto individual por medio del cual una persona elige códigos, signos y reglas del lenguaje para emitir un mensaje a través de la emisión de sonidos (fonación) [99]. Mediante este mecanismo, un hablante utiliza el lenguaje para establecer un acto de comunicación con un receptor (oyente) [100]. Se denomina comunicación al proceso de transmisión y recepción de la información.

En la figura 2.1 se aprecia un diagrama simplificado del proceso de comunicación oral humano; y, a modo de comparación, se muestra también un diagrama en bloques de un sistema genérico de comunicación, donde se destacan los elementos constitutivos básicos. La fuente (o emisor) es la encargada de seleccionar los signos que formarán el mensaje, para lo cual se parte de una idea o pensamiento que el hablante desea transmitir al oyente. El transmisor es el responsable de codificar adecuadamente el mensaje en una señal, es decir, el hablante traduce este pensamiento a través de una serie de procesos neurológicos y movimientos musculares para producir una onda de presión sonora. Esta señal se transporta a través del canal (en el ejemplo, el aire) y se recibe y decodifica en el receptor. La señal recibida por el sistema auditivo del oyente se procesa y se convierte nuevamente en una señal neurológica. Finalmente, el significado del mensaje es interpretado en el destino, donde el oyente forma una idea del mensaje recibido. Hay que tener en cuenta que, para la adecuada transmisión de un mensaje, ambas partes deben compartir un código y un conjunto de signos comunes. Toda interferencia en este proceso se considera como ruido. Este complejo proceso se lleva a cabo en los órganos del aparato fonador, el sistema auditivo y el cerebro. Este último, es el encargado de controlar, coordinar y procesar la información, para lo cual es necesario, además, contar con una perspectiva lingüística [100, 102].

El aparato fonador y el sistema auditivo no pueden tratarse de manera aislada. El sistema auditivo impone un conjunto de restricciones sobre la naturaleza acústica del habla que son cruciales para entender la forma en que la información se integra en la señal de voz. Por ejemplo, el espectro del habla está sesgado hacia las bajas frecuencias, las cuales son particularmente resistentes a alteraciones debidas al ruido

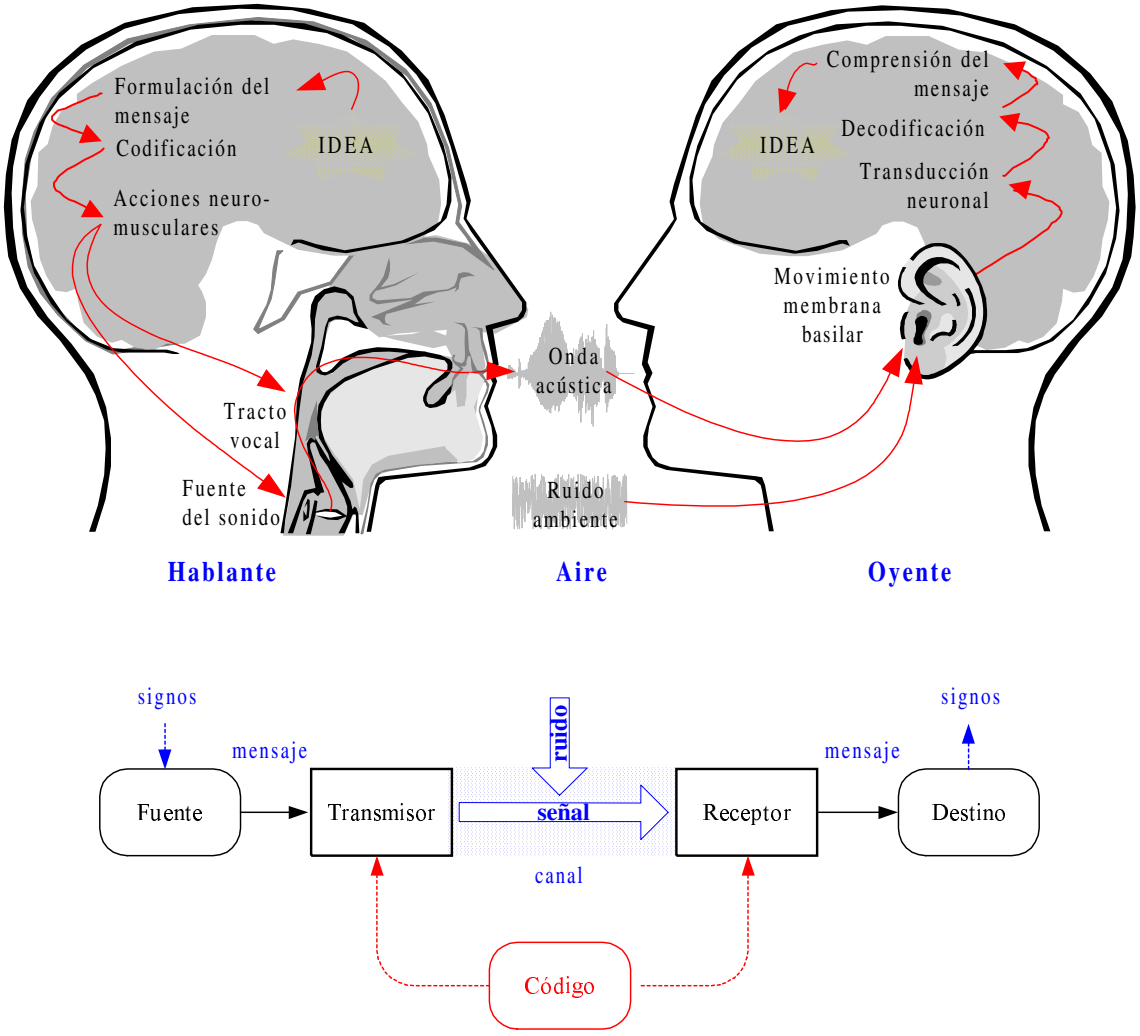

Figura 2.1: Diagrama simplificado del proceso de comunicación oral de un mensaje en el hombre (arriba), donde se muestran solo las etapas y órganos intervinientes más importantes, en un único sentido. Y diagrama en bloques de un sistema de comunicación genérico (abajo). Imagen cortesía de [101].

de fondo. Además, el nivel de presión sonoro de la mayor parte del habla es suficientemente alto como para asegurar que esa información espectral de baja frecuencia se extienda por una amplia serie de canales de frecuencia auditiva. Por otro lado, la periodicidad glótica asegura que el sistema pueda seguir o rescatar el habla en condiciones acústicamente adversas o de ruido y la modulación de la longitud de las sílabas ayuda al cerebro a juntar entidades espectrales dispares en unidades más significativas. Estas propiedades son esenciales para la codificación robusta (tolerante a fallas) y fiable de la información en condiciones acústicas ruidosas y en circunstancias en que pueden ocurrir múltiples interacciones vocales (charla entre varias personas). Dentro de este marco, el sistema auditivo precondiciona la representación neural para maximizar la fiabilidad y la tasa de transmisión de información, mediante mecanismos auditivos que operan bajo una amplia gama de condiciones acústicas potencialmente adversas [103].

A fin de poder emular las características de este sistema natural, es necesario comprender la naturaleza del habla y su forma de producción, como así también, entender los aspectos fundamentales del procesamiento que lleva a cabo en el sistema auditivo para poder extraer las características significativas de la señal de voz y decodificar el mensaje. Para ello, es necesario estudiar los fundamentos anatómicos y fisiológicos involucrados en el proceso de la comunicación oral humana.

Sin embargo, surge el interrogante acerca de cuales son las características que pueden ser capturadas por los dispositivos artificiales, que permitan asegurar capacidades de utilidad práctica, de acuerdo al problema que se desea resolver. Por ejemplo, en el caso de un sistema de reconocimiento automático del habla (ASR), es deseable poder lograr la independencia de su desempeño bajo diferentes condiciones, como ser cambios en el volumen y la velocidad de pronunciación, en la identidad del hablante (por variaciones de rasgos particulares o cuestiones regionales) o en las interferencias del ambiente acústico circundante. Este tipo de análisis permitirá poder encontrar una representación de la señal que sea óptima para el diseño de nuevos dispositivos tecnológicos.

Durante el desarrollo de este capítulo se explicarán con mayores detalles los aspectos fundamentales involucrados en la producción y percepción del habla, destacando aquellos que deberían tenerse en cuenta para la representación o codificación de la señal de voz. Para mayor información, se remite al lector a indagar en la extensa bibliografía específica disponible para cada área (por ejemplo [104, 105, 106, 107]).

Este capítulo se organiza como se indica a continuación. En la sección 2.2 se describirá el mecanismo de producción del habla y los órganos involucrados en el mismo. Si bien los aspectos funcionales de este proceso son relativamente independientes del idioma considerado, en esta tesis el análisis se limitará al idioma español (principalmente en su versión *argentina rioplatense* [107]). Luego, en la sección 2.3, se esbozarán los principios y elementos que intervienen en la percepción de los sonidos del habla y la audición, enfatizando sobre aquellos fundamentos relacionados con la codificación de la señal de voz a nivel neurosensorial. Posteriormente (sección 2.4), se discutirán aspectos relacionados con el procesamiento de la señal de voz, presentando algunos ejemplos típicos. Finalmente, en la sección 2.5 se sintetizan y retoman los aspectos fundamentales sobre los que se sustentan las actividades planteadas en los capítulos siguientes.

2.2. Sistema fonatorio

El proceso de comunicación oral comienza cuando el hablante traduce una idea a patrones de variación de la presión sonora en la señal de voz. Esto se lleva a cabo principalmente en la corteza cerebral e involucra varias áreas, de manera simultánea o alternada. Este es un proceso bastante complejo, ya que el cerebro debe enviar las órdenes adecuadas al aparato fonador para codificar la información acústica a transmitir por medio de una serie de reglas lingüísticas que se manifiestan en diferentes niveles, a fin de proveer la redundancia necesaria para aumentar la robustez de la comunicación. Cada uno de estos niveles impone ciertas restricciones y "estructuras" que forman parte del "código" (fonológico, fonético, léxico, sintáctico, etc.) compartido entre el hablante y el oyente [100, 105, 108, 109].

El habla, como medio oral de comunicación, tiene asociados dos componentes importantes: la articulación y la voz [110]. La articulación es la manera en que se producen los sonidos (por ejemplo, en el caso de los niños, estos tienen que aprender a producir el sonido de la "s"para poder decir "sol"en lugar de "tol"). Por otro lado, la voz es el uso de las cuerdas vocales y la respiración para producir sonidos. Estos dos elementos se estudian a partir de una rama de la fonética que es la fonética articulatoria, la cual se ocupa de describir la producción física del habla en el aparato fonador.

En la figura 2.2 se observa un esquema simplificado del aparato fonador en conjunto con una sección sagital del mismo (donde no se incluyen los pulmones). La zona comprendida entre la laringe (glotis) y los labios constituye el *tracto vocal* propiamente dicho y está formado por las cavidades supraglóticas, faríngeas, oral y nasal. En el diagrama se ejemplifican las señales temporales, sus correspondientes espectros y sus funciones de transferencia espectrales para el caso de producción de un fonema sonoro.

Los sonidos se generan en el aparato fonador a partir del aire procedente de los pulmones, que se sonoriza al hacer vibrar las cuerdas vocales. La tensión, elasticidad, altura, anchura, longitud y grosor de las cuerdas vocales pueden variar, dando lugar a diferentes efectos sonoros (el efecto más importante de las cuerdas vocales es la producción de una vibración audible en los llamados sonidos sonoros). Las cuerdas vocales se encuentran en la laringe y es allí donde se produce la voz en su tono fundamental y sus armónicos. Luego, la voz sufre una modificación en la caja de resonancia que conforman la nariz, la boca y la garganta, en la que se amplifica y se forma el timbre de la voz. Los órganos articuladores (labios, dientes, paladar duro, velo del paladar, mandíbula) van a moldear esta columna sonora transformándola en fonemas, sílabas y palabras.

La teoría acústica de la producción del habla describe este proceso como una secuencia de una o más fuentes de sonidos, un sistema de filtros para el tracto vocal y características de radiación. Usando representación simbólica, y suponiendo linealidad, si $H(f)$ es la función de transferencia del filtro que representa el tracto vocal en un instante dado y $X(f)$ la fuente de excitación, el producto $Y(f) = H(f) \cdot X(f)$ representa el sonido resultante. La fuente $X(f)$ indica la perturbación acústica de la corriente de aire proveniente de los pulmones. A veces suele agregarse a este modelo la función transferencia $L(f)$ que modela el fenómeno de radiación a la salida de los labios. Por lo tanto, los sonidos del habla son el resultado de la excitación acústica del tracto vocal, el cual varía constantemente sus características. En este proceso

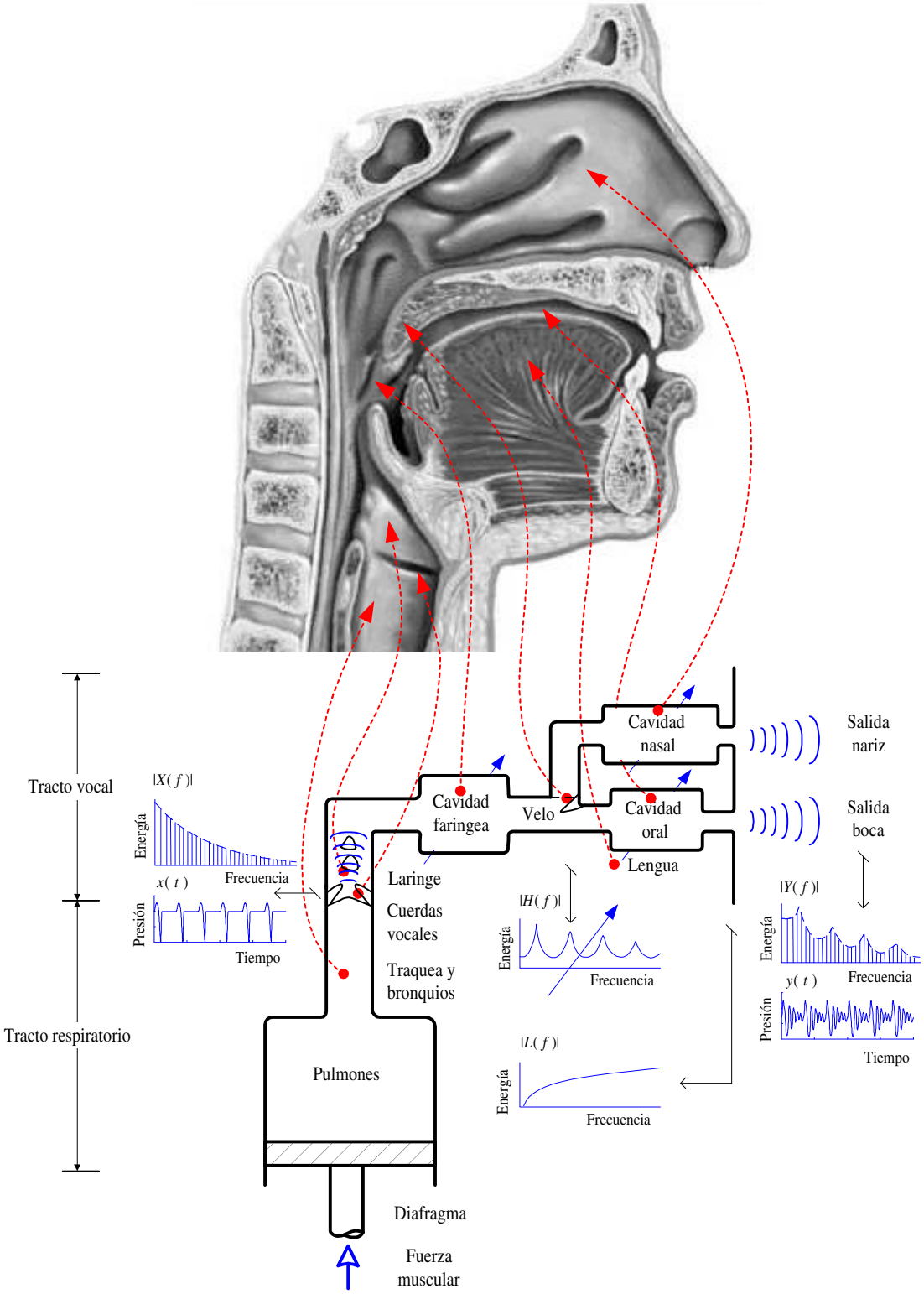

Figura 2.2: Corte sagital anatómico del aparato fonador (arriba) y diagrama esquemático del mismo que ilustra su funcionamiento (abajo). Imagen cortesía de [101].

los órganos fonatorios desarrollan distintos tipos de actividades, como por ejemplo movimientos de pistón que inician una corriente de aire o movimientos o posiciones de válvula que regulan el flujo de aire y, al hacerlo, generan sonidos o, en algunos casos, simplemente modulan las ondas generadas por otros movimientos [110]. Una manera de representar la forma en que el tracto vocal varía sus características es mediante un un modelo sencillo de dos tubos uniformes sin pérdida, los cuales varían su ancho o su longitud [111]. En la figura 2.3 se puede apreciar un modelo de este tipo, donde se observa cómo las diferentes configuraciones del modelo modifican la respuesta en frecuencia del mismo (debido al cambio en las frecuencias de resonancia de los tubos), constituyendo un filtro acústico variante en el tiempo. Esto permite explicar no solo las diferencias entre los sonidos producidos por un mismo hablante, sino también las existentes entre los sonidos de diferentes hablantes, debido a sus variaciones anatómicas [112].

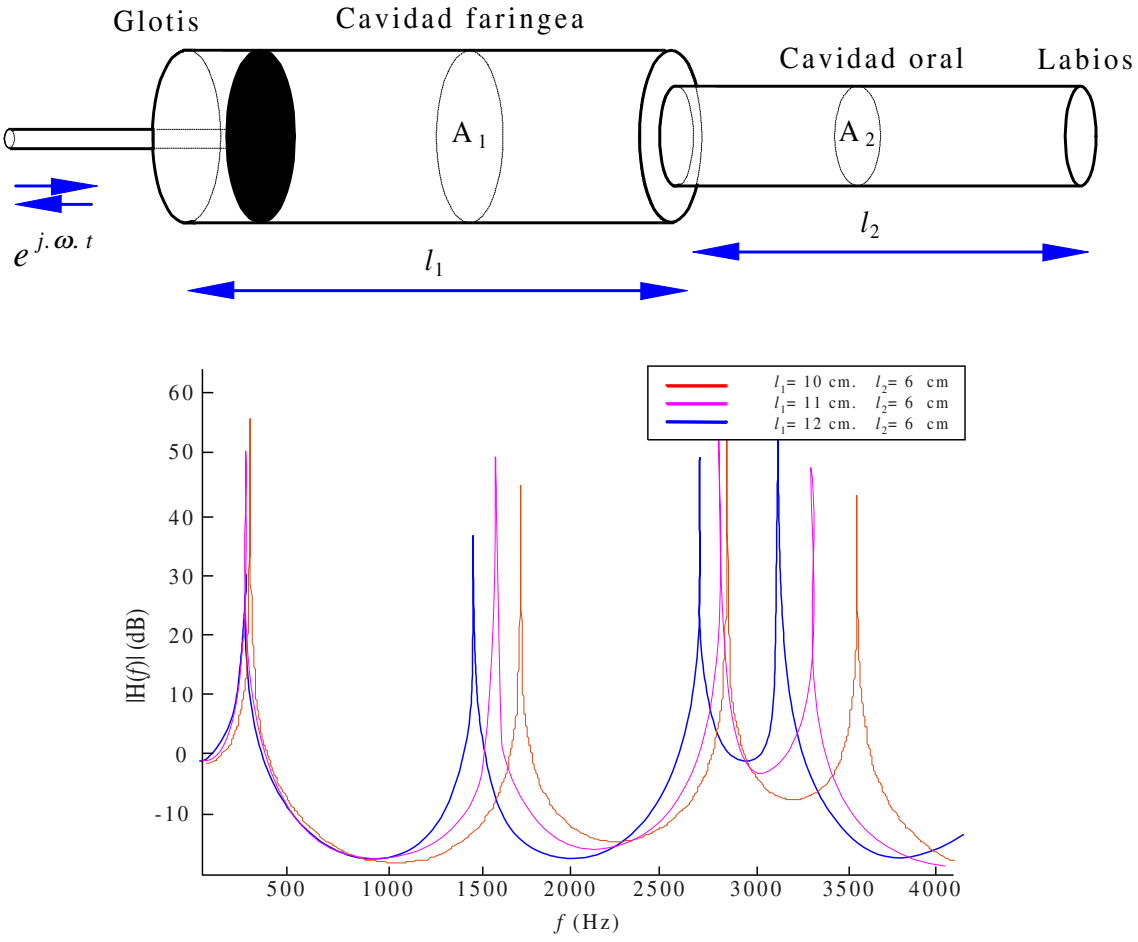

Figura 2.3: Modelo de dos tubos sin pérdida para el tracto vocal (arriba) y respuesta en frecuencia del mismo para diferentes longitudes de la cavidad faríngea (abajo). Imagen cortesía de [101].

Durante la secuencia de apertura y cierre de las cuerdas vocales se producen variaciones bruscas en la presión sonora a la salida de las mismas, las cuales pueden representarse en forma simplificada como una onda triangular periódica de período T. En el hombre, la frecuencia de vibración de las cuerdas vocales varía entre 100 y 170 Hz, en las mujeres entre 180 y 280 Hz y en los niños puede superar los 300 Hz. Los valores de esta vibración glótica (o frecuencia glótica) se modifican en forma

voluntaria y son los responsables de la frecuencia fundamental (denominada F_0) producida al hablar [110]. El tracto vocal actúa como un filtro acústico, modulando el tono glótico o cerrando el paso del aire [113].

El tracto vocal modifica sus parámetros en forma continua y los cambios resultantes de su configuración producen los diferentes sonidos y fonemas. Estos cambios pueden observarse directamente en el espectro del sonido vocálico, el cual proporciona información acerca de los aspectos relevantes de la configuración del tracto en ese instante. A modo de ejemplo, en la figura 2.4 pueden observarse los sonogramas (gráficas de variación de la presión sonora en función del tiempo) de las vocales /a/ e /i/ del español, junto con sus respectivas envolventes espectrales, donde estas resonancias se manifiestan a través de los picos (*formantes*) presentes en la respuesta en frecuencia [110].

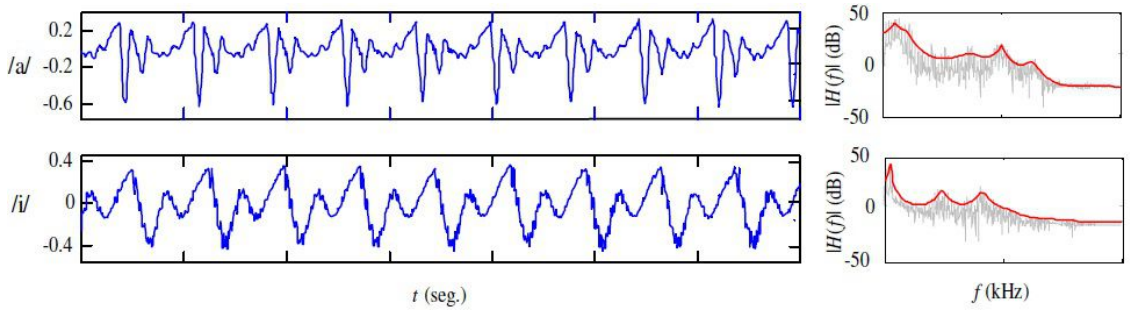

Figura 2.4: Ejemplos de sonogramas (izquierda) y espectros (derecha) de las vocales /a/ e /i/ del español, pronunciadas en forma sostenida y aislada por un hablante masculino nativo. Imagen modificada de [101].

2.2.1. Sonidos y fonemas

El habla se puede dividir en unidades lingüísticas básicas denominadas fonemas. Los fonemas son el conjunto mínimo de unidades que permite decir cualquier palabra en un idioma determinado [105]. Dos fonemas son distintos si el cambio de uno por otro cambia la palabra (por ejemplo *boda* vs. *moda*). El fonema, como modelo de los sonidos, puede diferir en su expresión acústica o realización, dando lugar a lo que se conoce como variantes o alófonos (también se utiliza el término *fono* como sinónimo de alófono). Hay fonemas que se articulan de forma distinta según su posición en la palabra o el carácter de los fonemas vecinos. Por ejemplo, en las palabras "donde" e "idioma" aparece el fonema /d/. En "donde" el fonema se articula tocando la parte trasera de los dientes con la lengua. En cambio, para pronunciar la palabra "idioma" se coloca la lengua entre los dientes. Si bien el fonema es el mismo, /d/, se tienen dos sonidos diferentes (uno dental y otro interdental), los cuales son similares en una serie de rasgos (los propios del fonema)[1] [100].

[1]Existen alfabetos fonéticos para aplicaciones tecnológicas con adaptaciones particulares para el español rioplatense [114], tales como:

- SAMPA: `http://www.phon.ucl.ac.uk/home/sampa/spanish.htm`
- Worldbet: `http://www.ling.gu.se/~jimh/courses/ipa.ps`

Los fonemas no son sonidos con entidad física, sino abstracciones mentales o abstracciones formales de los sonidos del habla. Por lo tanto, mientras el sonido pertenece al habla, el fonema pertenece a la lengua. En este sentido, un fonema puede ser representado por una familia o clase equivalente de sonidos (los alófonos o fonos), que los hablantes asocian a un sonido específico durante la producción o la percepción del habla. El número de fonemas de una lengua es finito y limitado en cada lengua y el número de alófonos potencialmente definibles, especialmente si especificamos rasgos fonéticos muy sutiles, es potencialmente ilimitado y varía según el contexto fonético y la articulación lingüística individual de los hablantes. En la figura 2.5 se muestra un cuadro con los fonemas de uso corriente en nuestro idioma, clasificados de acuerdo con las características acústicas y los gestos articulatorios que dan lugar a cada tipo de sonido. La principal división que se da en este sentido es la diferencia entre vocales y consonantes [105].

vocales:	/a/ /e/ /i/ /o/ /u/	
fricativos:	/f/ /s/ /j/ /y/	consonantes
africados:	/ch/	
oclusivos:	/b/ /b/ /g/ /p/ /t/ /k/	
nasales:	/m/ /n/ /ñ/	
vibrantes:	/r/ /rr/	
laterales:	/l/ /ll/	

Figura 2.5: Cuadro simplificado de clasificación de los fonemas del español *rioplatense* [107].

El tracto vocal actúa como sistema resonador, modificando su configuración (y con ello sus frecuencias de resonancia) como una especie de filtro acústico adaptativo que, junto con la fuente de excitación actuante, dan al sonido su peculiar cualidad fonética. La división entre fonemas vocálicos y consonánticos se sustenta tanto en las características acústicas como en los gestos articulatorios que dan lugar a cada tipo de sonido [100].

Durante la articulación de vocales y sonidos tipo vocálicos, el tracto presenta una configuración relativamente abierta y la fuente de excitación es siempre glótica. Las propiedades de estos sonidos persisten por un tiempo apreciable o cambian muy lentamente mientras se mantenga la configuración del tracto. La forma del tracto en la producción de las vocales esta controlada, principalmente, por la posición de la lengua, de la mandíbula y de los labios [108].

Los sonidos consonánticos se producen con una configuración relativamente cerrada del tracto vocal. Entre los factores que determinan la cualidad del sonido resultante, deben distinguirse aquellos que hacen al modo de articulación (cierre o estrechamiento) de los que señalan la zona o lugar de articulación (lugar donde se produce cierre o estrechamiento). La participación o no de la fuente glótica, la naturaleza del cierre o estrechamiento y la transmisión a través de la cavidad oral y/o nasal, constituyen los principales factores del modo de articulación.

Sin embargo, por razones de sencillez y salvo que se indique lo contrario, para hacer referencia a los fonemas se utilizará la grafía más cercana (a su pronunciación) encerrada entre /•/.

Además de las diferencias acústicas debidas a las distintas configuraciones del tracto vocal, existen diferencias en la duración temporal de los fonemas, la cual no es uniforme. Para dar una idea general, se puede decir que las vocales son más largas (en el orden de los 100 mseg promedio) que las consonantes (en el orden de los 20 mseg promedio). Por otro lado, a la duración intrínseca variable que tienen los fonemas, hay que sumarle la variación que se presenta debido a los elementos que integran la sílaba (fonemas vecinos) o las características suprasegmentales, las cuales se describen en la siguiente subsección[100, 115].

2.2.2. Segmentos, suprasegmentos y sílabas

El habla se compone una sucesión de fonemas. Esto resulta, de alguna manera, en un fenómeno secuencial "discreto" y es posible, por lo tanto, asignar *etiquetas* a los diferentes trozos de señal (segmentos) asociados con estos fonemas. Sin embargo, si se observa la representación acústica de una frase (señal de la voz), ésta constituye un continuo acústico, producido por un movimiento ininterrumpido de los órganos del aparato fonador, donde se ven muy pocas pausas o intervalos entre los sonidos. A pesar de la naturaleza continua de la voz, los oyentes pueden segmentarla en sonidos.

Por otro lado, el habla implica más que la mera concatenación de sonido individuales (fonemas o segmentos). Existen elementos, relacionados con la pronunciación, que afectan a más de un fonema a la vez, tales como el acento, la entonación, el ritmo y la duración. Estos elementos se denominan características suprasegmentales (o prosodemas) y resultan de una utilización particular de recursos del aparato fonatorio [100]. Por ejemplo, las variables que intervienen en la entonación, la cual define la prosodia, son las variaciones de frecuencia fundamental (F_0), la duración y las variaciones de energía y sonoridad. Así, un hablante controla la entonación aplicando mayor o menor tensión a las cuerdas vocálicas, lo cual le permite enfatizar más unas partes de la oración u otras o darle un tono de sorpresa o de interrogación. Los elementos suprasegmentales transmiten información contenida en el habla, que no está contenida en los fonemas [109].

La sílaba constituye una unidad lingüística de escala temporal mayor que la del fonema. Si bien la mayoría de las personas puede identificar fácilmente las sílabas, es relativamente difícil dar una definición exacta de este término. Aunque para una lengua la cantidad de sílabas es muy superior a la de fonemas, en general la variabilidad acústica de estas unidades es también mucho menor. Por ello, algunos investigadores prefieren su utilización como unidad de modelado del habla [113].

2.3. Percepción del Lenguaje

En el proceso de comunicación oral, el emisor codifica en una señal sonora la información lingüística que desea transmitir y el receptor debe decodificar la información transmitida por la onda sonora para comprender el mensaje. En esta sección se describe la forma en que el sistema auditivo realiza el procesamiento de la señal de habla para poder decodificar el contenido del mensaje contenido en la misma.

A diferencia de lo que sucede en la lengua escrita, el habla es un proceso sonoro continuo, donde no se encuentran "separaciones" entre sonidos o palabras. La decodificación, por lo tanto, requiere segmentar el continuo para identificar las unidades lingüísticas que contiene. Esta operación depende del contexto en el que se emite el

enunciado y de la información acústica presente en la señal sonora [116]. Esta última característica es la que particularmente interesa en esta tesis: la percepción del habla que se basa en la extracción de indicios acústicos (*pistas acústicas*) presentes en la señal sonora.

2.3.1. El oído

El procesamiento de la señal de habla comienza en el oído, para terminar finalmente en el cerebro, que es donde realmente se produce el fenómeno de la audición. El oído es el encargado de la recepción y adecuación del sonido y de su transducción a impulsos nerviosos. En la figura 2.6 se aprecia un corte transversal del oído, junto con un diagrama esquemático que ilustra su funcionamiento. En el mismo se observan sus tres secciones principales: el oído externo, el medio y el interno. Se podría decir que las dos primeras partes se encargan de la recepción y adecuación del sonido para su posterior procesamiento en el oído interno, donde se lleva a cabo la función más importante: la transducción del sonido a impulsos nerviosos.

La parte más externa del oído es el pabellón auditivo, el cual se encarga de captar el sonido y enfocarlo hacia el conducto auditivo. Las ondas de presión viajan por el conducto auditivo hasta el tímpano, que es una membrana elástica que separa el oído externo del oído medio. El oído medio es una cámara ocupada por aire, que se comunica con la faringe a través de la trompa de Estaquio, y contiene un conjunto de huesecillos: el martillo, el yunque y el estribo, cuya función principal es la de adaptación de impedancias acústicas. La membrana timpánica se mueve como consecuencia de las vibraciones del aire que llega a través del canal auditivo externo y ese movimiento se transmite a través de la cadena de huesecillos en el oído medio, transformando las variaciones de presión sonora en movimiento mecánico. El estribo, el más interno de estos huesecillos, establece contacto con la ventana oval que está ubicada en la base de la cóclea, en lo que constituye el oído interno. La amplificación de las vibraciones producidas en el tímpano está limitada, en condiciones de cambios abruptos, por el reflejo estapedial (también llamado timpánico o auditivo), cuya función es proteger al oído interno (esto funciona en la práctica como un control automático de ganancia mecánica).

El órgano principal del oído interno es la cóclea. La cóclea tiene forma de tubo cónico enrollado en forma de caracol y está formada por tres cámaras longitudinales llenas de fluidos: la rampa timpánica y la rampa vestibular, que contienen perilinfa, y la rampa media o coclear, que contiene endolinfa. Estas tres cámaras están separadas por dos membranas: la membrana de Reissner, entre la rampa vestibular y la rampa media o coclear; y la membrana basilar, entre la rampa media o coclear y la rampa timpánica. En la membrana basilar descansa el *órgano de Corti* (órgano del sentido de la audición) con sus *células ciliadas* (estereocilios). Los estereocilios son células con microvellosidades y son los receptores auditivos encargados de transformar las vibraciones del sonido en impulsos nerviosos que son enviados hasta el cerebro.

Una vez excitada la ventana oval (presionada por las vibraciones del estribo en respuesta al sonido), el sonido se transmite a través del líquido de la rampa vestibular, atraviesa el helicotrema (el vértice del "caracol") y sigue su recorrido en la rampa timpánica hasta la ventana redonda. La ventana oval y la redonda trabajan de forma coordinada, de modo tal que cuando una se comba hacia adentro la otra se comba hacia afuera y viceversa. El movimiento hacia adentro y afuera se repite a

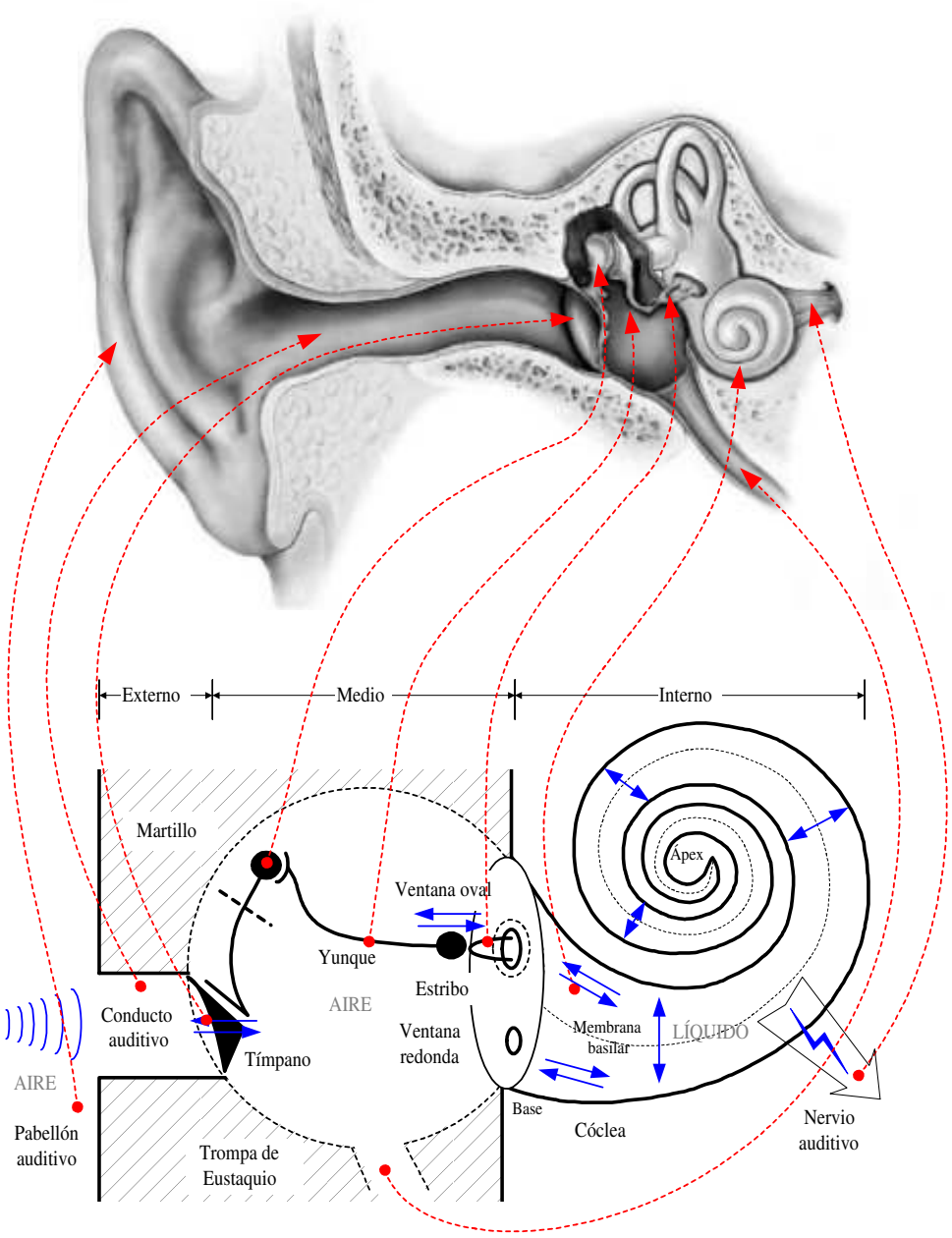

Figura 2.6: Corte sagital anatómico del oído (arriba) y diagrama esquemático que ilustra su funcionamiento (abajo). En el diagrama se resaltan sus secciones principales: el oído externo, el medio y el interno. Imagen cortesía de [101]

la misma frecuencia del estímulo sonoro. Esta onda de presión sonora deformará la membrana basilar en una zona concreta en función de la frecuencia de dicho sonido. Las frecuencias altas actuarán sobre la base de la membrana basilar y las bajas frecuencias sobre el ápice, dando lugar, de manera selectiva, a la transducción auditiva. Esto se debe a que la membrana basilar varía en masa y rigidez a lo largo de toda su longitud, con lo que su frecuencia de resonancia no es la misma en todos los puntos. En el extremo más próximo a la ventana oval, la membrana es rígida y ligera, por lo que su frecuencia de resonancia es alta. Por el contrario, en el extremo más distante, la membrana basilar es pesada y suave, con lo que su resonancia es a baja frecuencia. A esta característica se la denomina *tonotopía de la membrana*. En la figura 2.7 se muestra una versión aislada y desenrollada de la cóclea, donde se observa cómo la amplitud de la onda sobre la membrana basilar depende de la frecuencia de estimulación.

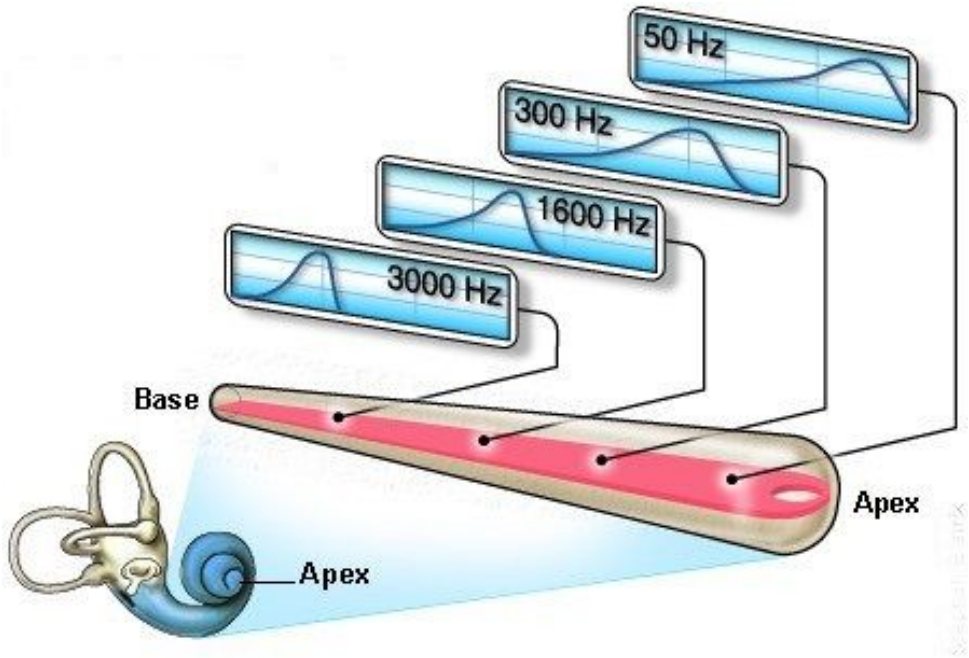

Figura 2.7: Versión desenrollada de la cóclea donde se muestra cómo la amplitud de la onda aumenta a medida que se propaga, pasando por un máximo y luego decrece rápidamente, dependiendo de la frecuencia de estimulación. Imagen adaptada de [117].

De esta forma, las vibraciones de frecuencias altas tienen su máxima amplitud cerca del lugar donde las ondas comienzan a desplazarse, luego disipan la mayor parte de su energía y se desvanecen en el camino, no alcanzando nunca el ápice. Las vibraciones de baja frecuencia, por el contrario, comienzan con una amplitud pequeña cerca de la base y la aumentan a medida que se acercan al ápice. De esta manera están representadas todas las frecuencias audibles a lo largo de toda la cóclea [118, 119, 120]. En este mapa tonotópico existe una relación entre la frecuencia y la posición característica sobre la membrana basilar, pero la ubicación de las frecuencias no es lineal, sino más bien de tipo logarítmica. La membrana basilar se dividide en 3 porciones, cada una con un rango de frecuencia, la porción apical (20 a 200 Hz), la porción media (200 Hz a 2kHz) y la porción basilar (2kHz a 20kHz). Esta es una de las causas por las que la resolución frecuencial y la percepción de las frecuencias no

es uniforme en toda la cóclea. A la escala psicoacústica que da cuenta de la relación entre la frecuencia física del sonido y la percibida se la denomina *escala de mel* (ver figura 2.13 más adelante)[110].

La transducción mecánico-eléctrica del sonido se produce en el órgano de Corti, el cual se ubica a lo largo de toda la membrana basilar. Dicha transducción tiene lugar como respuesta a una curvatura de las cilias (sensibles a los estímulos mecánicos) de las células ciliadas, debido a las vibraciones de la membrana basilar. El desplazamiento mecánico de las cilias produce una variación en el potencial de membrana de las células; si las cilias se curvan hacia el cuerpo basal se produce una despolarización, mientras que si se curvan en el otro sentido se produce una hiper-polarización. En función de estos patrones, al ser estimuladas, las células ciliadas producen un componente químico que genera los impulsos eléctricos que son trasmitidos primero al nervio acústico y posteriormente al nervio auditivo. En la figura 2.8 se muestra un dibujo de cómo está constituído el órgano de Corti. La excitación de las células

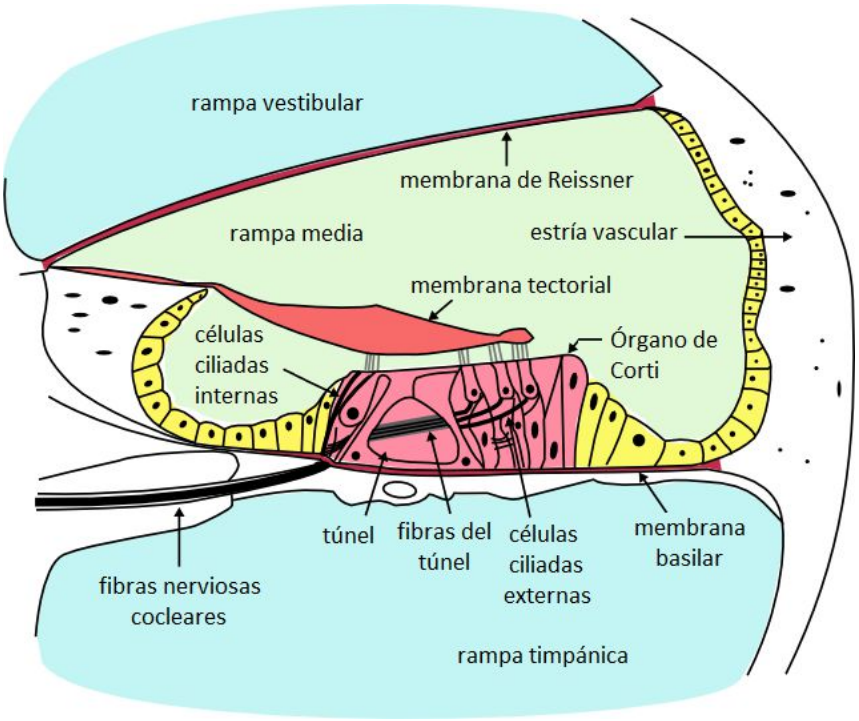

Figura 2.8: Detalle del órgano de Corti, donde se observan las células ciliadas en relación con la membrana tectoria, la membrana basilar y el nervio auditivo.

ciliadas está determinada, en gran medida, por las excursiones de la membrana basilar. De esta manera, dado que la amplitud de las vibraciones en distintos puntos de la cóclea varía con la frecuencia del estímulo, el grado en el cual es excitada una determinada célula ciliada es una función conjunta de su posición en la membrana basilar y de la amplitud del estímulo. Por lo tanto, estas células generan patrones diferenciados, característicos de cada tono (o frecuencia). La curva de resonancia o sintonía mecánica (amplitudes relativas de las excursiones para los distintos puntos sobre la membrana basilar en función de la frecuencia del estímulo) de la membrana basilar describiría con precisión la excitación de las células ciliadas en función de la frecuencia si éste fuera el único factor que influyera en la vibración de las células ciliadas. Sin embargo, las propiedades mecánicas de las cilias y de la membrana

tectoria que las cubre influyen en la vibración de estas células. Esto se debe a que tanto la rigidez de las cilias, como la masa y la elasticidad de la membrana tectoria, también varían de un extremo al otro de la cóclea [121].

Las células ciliadas pueden agruparse en internas (CCIs) y externas (CCEs), con marcadas diferencias como la posición en el órgano de Corti, la ultraestructura, la implantación de las estereocilias y la inervación [122]. Estas diferencias están al servicio de dos acciones diferentes: las CCIs y las CCEs actúan como transductores mecano-eléctricos del sistema auditivo mientras que las CCEs actúan, además, como transductores electro-mecánicos, es decir, como células motoras. Las CCEs responden a cambios de potencial cambiando su longitud. La fuerza generada por ellas es capaz de alterar los delicados mecanismos cocleares incrementando la sensibilidad auditiva y la selectividad de las frecuencias. Cada frecuencia seleccionada se resalta con un efecto mecánico equivalente a las inhibiciones laterales de las estructuras neurales. Estos procesos son explicados por las interacciones entre las diferentes particiones cocleares y el efectivo comportamiento no-lineal de estas células motoras [122]. Las CCIs son las principales encargadas de codificar la frecuencia y la intensidad de la estimulación sonora. Y si bien las CCEs están prepradas para realizar la transducción mecánico-eléctrica, como las CCIs, no trasmiten ninguna característica del estímulo sonoro al cerebro. Al contrario, un mecanismo activo de transducción inversa (electro-mecánico) les permite reenviar la energía para aumentar la sensibilidad y la selectividad frecuencial. Estas características del complejo célula-membrana tectoria tiene el efecto de limitar la sintonía de las células ciliadas a un ancho de banda de frecuencias más estrecho que el del punto de la membrana basilar donde se encuentra la célula. Se debe mencionar, también, que las células ciliadas se despolarizan solo durante la fase positiva de los estímulos sonoros produciendo un efecto de *rectificación de media onda* sobre las respuestas del nervio auditivo [123].

2.3.2. Codificación de los sonidos en el nervio auditivo

La colección de axones periféricos correspondientes a las neuronas aferentes y eferentes que inervan a las células ciliadas se conectan con las fibras nerviosas que conforman el nervio auditivo (aproximadamente 30.000 fibras en el nervio auditivo en el hombre). A través del nervio auditivo los estímulos se transmiten a la corteza auditiva. Los impulsos nerviosos generados en el oído interno contienen (codificados en trenes de pulsos) información acerca de la amplitud y el contenido espectral de la señal sonora. Estos dos parámetros están representados por la tasa de impulsos y la distribución de los mismos en las distintas fibras del nervio auditivo [124].

La respuesta de una fibra aislada puede describirse en términos de la frecuencia del correspondiente tren de pulsos, su fase y su patrón temporal de activación. Se considera que la respuesta de una fibra es estocástica, en el sentido que el patrón de disparo está relacionado de manera probabilística con las características del estímulo [111, 125]. Aún sin estimulación acústica muchas fibras poseen respuesta espontánea, y ésta varía de fibra a fibra.

Cada fibra tiene una *frecuencia característica* (FC) a la cual es más sensible (frecuencia para la cual la intensidad de estímulo necesaria para excitar la fibra es la mínima). Para estímulos de tonos puros, cada fibra responde en una gama limitada de frecuencias (rango dinámico de la fibra nerviosa auditiva) para un nivel de sonido dado [126]. Si se estimula una fibra a su FC la intensidad del estímulo se codifica en

la frecuencia o tasa de disparo (siempre por encima de su frecuencia espontánea). Esta selectividad de frecuencia se debe a la sintonización mecánica de la membrana basilar y de las células ciliadas. Existe una correspondencia precisa entre la frecuencia característica de una fibra y su lugar de inervación a lo largo de la cóclea [125]. Hay que tener presente que las fibras no responden a una única frecuencia, aunque requieren una mayor intensidad para ser excitadas fuera de su FC. Ésto último sirve también para codificar información acerca de la intensidad del estímulo, de acuerdo a la cantidad de fibras que responden. A esta forma de codificación de la frecuencia del estímulo se la denomina *mecanismo de la localización*. Este mapeo entre frecuencia característica y posición espacial se replica a lo largo de los diferentes niveles de la vía auditiva hasta la corteza [127, 128].

Debido a que las descargas de las fibras nerviosas auditivas pueden "seguir" en frecuencia a los estímulos cuando son tonos puros de baja frecuencia (< 5 KHz) e intensidad moderada, la información de dicha frecuencia se puede codificar también en la tasa de disparos [129, 130]. Los intervalos entre estas descargas tienden a ocurrir a múltiplos enteros del período de estímulo, lo cual puede, en principio, ser utilizado para obtener estimaciones muy precisas de la frecuencia de tono [131, 132]. Sin embargo, para tonos de frecuencias mayores ya no es posible seguir el "ritmo" tan de cerca. Entonces se recurre al fenómeno de excitación de varias fibras simultáneas, cada una con una fase diferente pero invariante. Este fenómeno, denominado respuesta *enganchada en fase*, permite la codificación de la frecuencia del estímulo en forma "distribuida" entre varias fibras. Este mecanismo funciona de manera confiable aproximadamente hasta los 3 KHz [123]. Este modelo para la codificación de la frecuencia del estímulo se denomina *mecanismo temporal*. Este fenómeno de respuesta enganchada en fase no está limitado sólo a tonos puros, también se produce para tonos periódicos complejos, como el sonido sostenido de vocales [133, 134, 135]. En este caso, los patrones temporales de descarga de las fibras individuales brindan información sobre el contenido frecuencial del estímulo.

Debido a lo expuesto anteriormente, se podría decir que existe acuerdo de que para la codificación de la frecuencia coexisten los dos mecanismos expuestos. Para las bajas frecuencias se utiliza principalmente el mecanismo temporal y para altas frecuencias principalmente el de localización. No obstante, hay discrepancias acerca de la frecuencia a la cual comienza a reemplazarse uno por el otro [136]. Para la codificación de la intensidad también existe coincidencia acerca de un mecanismo mixto entre las tasas de disparo individuales y la cantidad de fibras que responden. La descripción de estos fenómenos se ha llevado a cabo en diferentes experimentos en animales, utilizando como estímulos principalemente tonos puros. Cabe preguntarse si este comportamiento se mantiene cuando los estímulos son señales más "complejas" (señales cuyo espectro contiene más de un tono puro, no periódicas o aleatorias). Pues dado el conjunto de procesos no lineales que se llevan a cabo en la cóclea durante la transducción y codificación del sonido, es lógico suponer que no es posible comprender el comportamiento frente a estímulos complejos sólo por la simple adición de los efectos producidos por sus componentes sinusoidales.

Para comprender la forma en que responde la cóclea a estímulos sonoros similares al habla humana se realizaron estudios utilizando tonos múltiples y señales de voz sintéticas basadas en modelos de producción del habla [137], y con posterioridad se comenzó a trabajar con señales de voz reales [138]. Sin embargo, los estudios se continuaron realizando en animales, y aunque es posible realizar extrapolaciones

para el caso del hombre, debe tenerse en cuenta que el procesamiento de sonidos como el habla puede ser diferente, ya que se trata de criaturas que no poseen un lenguaje hablado.

Registros sobre neuronas auditivas individuales en respuesta al habla, y estímulos similares a los provenientes del lenguaje, proporcionan descripciones detalladas del procesamiento neural que se realiza en las primeras etapas de la percepción del habla. Puede considerarse que las fibras en el nervio auditivo siguen una disposición ordenada de acuerdo a la FC y que responderán incrementando su probabilidad de descarga cuando el nivel del estimulo supera el umbral. La información experimental de la estimulación del nervio auditivo, ordenada de acuerdo con la FC de las fibras individuales, puede representarse por medio del neurograma [139, 140]. En la figura 2.9 se muestra un neurograma basado en las respuestas fisiológicas a la estimulación acústica de un hablante femenino. Cada línea del neurograma representa la tasa de disparo instantánea promedio de un conjunto pequeño de fibras nerviosas sobre un intervalo de tiempo corto luego del inicio del estímulo (histograma temporal postestímulo) [139].

Estos estudios muestran que una gran cantidad de información, sobre el patrón de formantes y el tono de voz, está presente tanto en las tasas de descarga promedio como en los patrones de descarga temporales de las fibras del nervio auditivo y las células de la cóclea. Las fibras nerviosas respondían como detectores de características sencillas, como ser la ubicación y seguimiento de las frecuencias formantes, así como la codificación de otras características espectrales simples, como por ejemplo, la representación directa de F_0 [138]. Por otra parte, muchas neuronas auditivas muestran respuestas superiores a los cambios rápidos que aparecen en la amplitud y el espectro, los cuales son fonéticamente importantes. La redundancia asociada al efecto de enganche de fase provee cierta robustez en la codificación y ésta es una de las razones por las cuales la información más importante del habla se concentra en las bajas frecuencias [103]. También se corroboraron algunos efectos de enmascaramiento de frecuencias en la presencia de estímulos simultáneos y no simultáneos. Esto conduce a que las características fonéticas relevantes sean destacadas y robustamente codificadas en las respuestas neuronales, lo que sugiere que el sistema auditivo muestra una tendencia a resaltar el conjunto particular de características acústicas utilizadas en la diferenciación fonética [141].

2.3.3. Corteza auditiva

La audición, como cualquier otra modalidad sensorial, posee una vía y unos centros primarios (centros completamente dedicados a esta función) y otras vías no primarias sobre las que convergen el conjunto de otras modalidades sensoriales. Los primeros estudios que relacionaron la estructura y la función de la corteza cerebral del lóbulo temporal con la percepción auditiva y el lenguaje los llevaron a cabo Paul Broca y Carl Wernicke, y han permitido localizar en la corteza cerebral los procesos básicos de la audición y el lenguaje [118].

El nervio auditivo constituye la primera parte de la denominada *vía auditiva*. La información proveniente de los patrones de activación del nervio auditivo viaja a través de diferentes estructuras hasta llegar a la corteza auditiva primaria, que es la región del cerebro responsable del procesamiento de esta información. En la figura 2.10 se muestra un esquema de las estructuras que componen esta vía y el

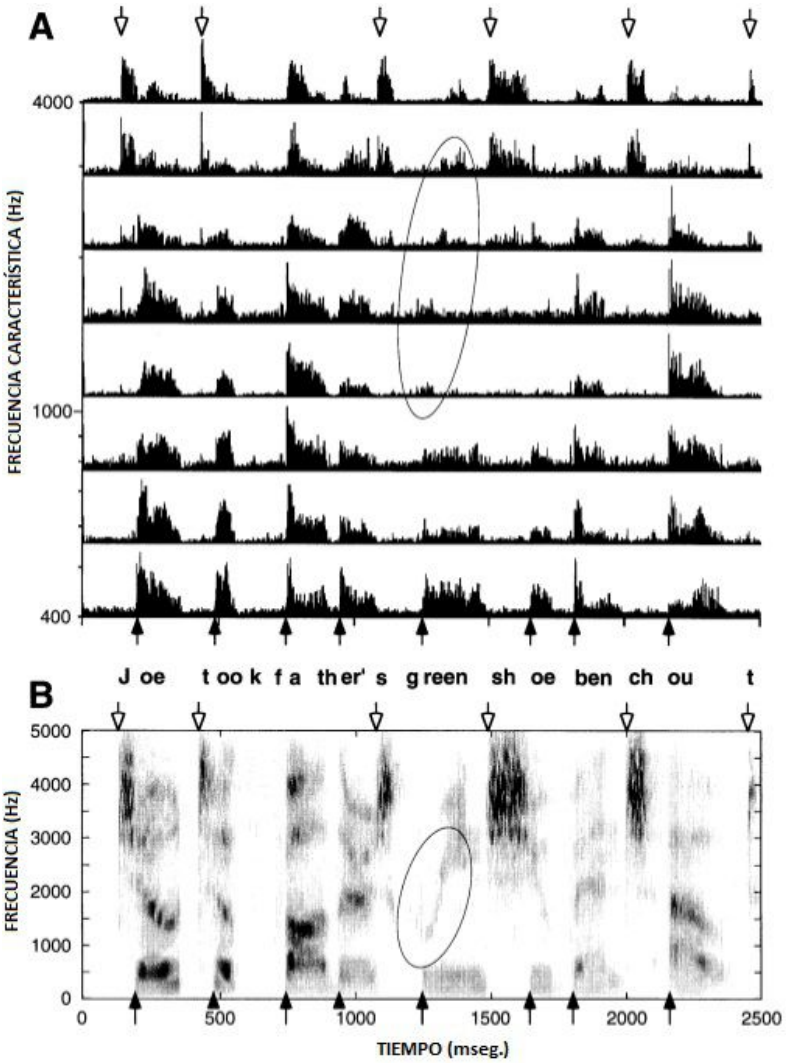

Figura 2.9: (A) Neurograma (arriba) en respuesta a la estimulación acústica de la elocución "Joe took father's green shoe bench out"de un hablante femenino. Histogramas temporales postestímulo de las tasas de disparo de las fibras nerviosas auditivas, ordenados de acuerdo a la frecuencia característica de cada fibra nerviosa. (B) Espectrograma (abajo) de la misma sentencia, que se muestra a modo de comparación. Las elipses se usan para resaltar que incluso detalles finos, como la transición de formantes rápidas como "green", están representadas en los cambios dinámicos de las tasas de disparo del nervio auditivo. Figura modificada de [139].

trayecto que recorre la información. En cada una de estas estructuras se realiza una actividad específica de decodificación e interpretación, que se transmite a los niveles superiores [119].

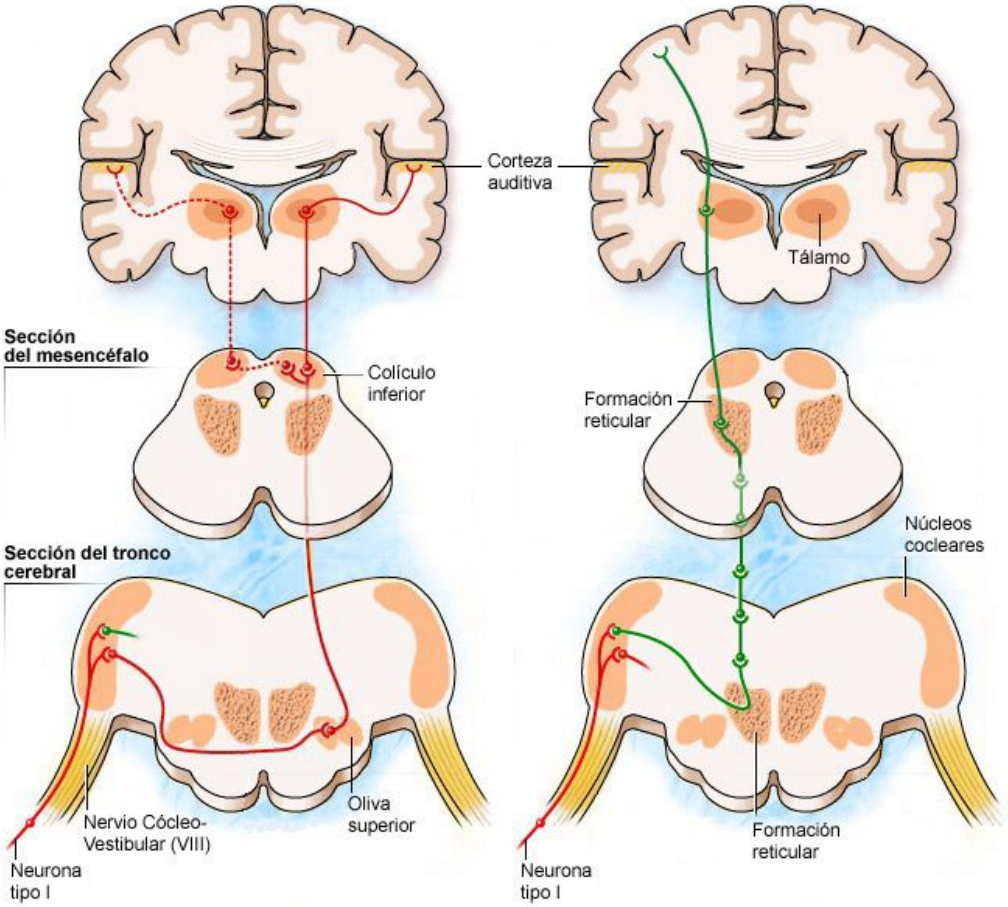

Figura 2.10: Diferentes secciones de la vía auditiva, donde se ilustran las conexiones y el trayecto seguido por la información desde el nervio auditivo hasta la corteza. A la izquierda la vía auditiva primaria y a la derecha la vía auditiva no primaria. Imagen modificada de [117].

Desde el nervio auditivo, la información viaja por el tronco cerebral a través del núcleo coclear, donde se realiza una importante labor de decodificación básica relacionada con la duración, intensidad y frecuencia del mensaje auditivo. Luego, la mayoría de las fibras auditivas hacen sinapsis en el núcleo olivar superior. Al inicio de este nivel, la tercera neurona permite que el mensaje ascienda hacia el mesencéfalo (colículo inferior). Estos dos niveles desempeñan un papel esencial para la localización del sonido. Un último paso, antes de la corteza auditiva, se lleva a cabo en el tálamo (en el núcleo geniculado medio), donde se lleva a cabo un importante trabajo de integración y la preparación de una respuesta motora (por ejemplo, de tipo vocal). La última neurona de la vía auditiva une el tálamo a la corteza auditiva primaria, donde el mensaje auditivo, que ya ha sido decodificado por las neuronas subyacentes, reconocido y memorizado, puede ser integrado en una respuesta voluntaria [142, 143].

En la vía auditiva no primaria, después del primer nivel (núcleos cocleares), una serie de fibras pequeñas se unen a la vía reticular ascendente, que es común

a todas las modalidades sensoriales. Luego de varios pasos dentro de la formación reticular, y después en la parte inespecífica del tálamo, esta vía conduce a la corteza multisensorial. El papel de esta vía, que reagrupa diferentes mensajes sensoriales enviados simultáneamente al cerebro, es hacer una selección del tipo de información que debe ser procesada con prioridad.

Si bien a lo largo de la vía auditiva se procesan y analizan los patrones de información, es en la corteza donde se lleva a cabo la transducción de los estímulos nerviosos para convertirlos en diferentes representaciones internas. Gracias a las técnicas de generación de imágenes funcionales, como la resonancia magnética funcional [144] o la localización de dipolos mediante potenciales evocados auditivos (PEA) [145], es posible el estudio no invasivo de algunas funciones corticales en el hombre. Ésto ha permitido la identificación de las zonas que intervienen en el procesamiento del habla y se ha encontrado que las neuronas de la corteza auditiva están organizadas según la frecuencia de los sonidos a los que responden con mayor eficacia. Las neuronas situadas en un extremo de la corteza auditiva responden mejor a las frecuencias bajas y las ubicadas en el otro extremo responden mejor a las frecuencias más altas, por lo que la organización tonotópica de la cóclea se mantiene en las diversas etapas de la vía auditiva, incluyendo la propia corteza.

2.4. Señal de Habla

En las secciones anteriores se han descrito principalmente los aspectos relacionados con la fonética articulatoria y auditiva, describiendo la forma en la que se originan los sonidos en el aparato fonador y la manera en que el oído reacciona ante las ondas sonoras. En la presente sección se abordarán los aspectos relacionados con la fonética acústica, es decir, el estudio de la señal de voz propiamente dicha. Sobre esta señal se llevarán a cabo los diferentes procesamientos que permitirán obtener las representaciones necesarias para cada aplicación particular.

Así como a los fonemas se les pueden atribuir rasgos articulatorios (ver sección 2.2.1), a los sonidos se les pueden atribuir rasgos acústicos, cuyas mediciones se reflejan en espectrogramas, donde se ven reflejados los distintos formantes en que se descomponen los sonidos [111]. En la figura 2.4 pueden observarse los sonogramas de dos vocales del español, pronunciadas en forma sostenida y aislada, junto con sus respectivos espectros, donde las resonancias del tracto se manifiestan en los picos de los mismos. Estos picos se denominan formantes y se numeran a partir del 1 (F_1, F_2, F_3, etc.), y constituyen un medio para caracterizar a las vocales. Esto último es posible porque cada sonido del habla humana tiene un patrón característico de formantes, es decir, tiene una distribución diferente de la energía sonora entre los diferentes formantes, lo cual permite clasificarlos o categorizarlos. En muchas lenguas, los dos formantes principales son F_1 y F_2, los cuales permiten distinguir la mayoría de sonidos vocálicos del habla. Si bien es imposible hacer una correspondencia uno a uno entre las configuraciones del tracto vocal (ver figura 2.2) y las características de los formantes, ya que son muchas las variables que determinan el comportamiento de estos últimos, se pueden encontrar relaciones entre el grado de apertura de la cavidad oral y el valor de F_1 y el grado de desplazamiento de la lengua y los valores de F_2. Típicamente, se encuentra que la primera formante, de frecuencia más baja, está relacionada con la abertura de la vocal, que en última instancia se relaciona con la frecuencia de las ondas estacionarias que vibran verticalmente en la cavidad oral.

La segunda formante está relacionada con la vibración en la dirección horizontal y también se relaciona con si la vocal es anterior, central o posterior [113, 146].

La presencia de formantes, y en particular de F_0, evidencia si se trata de un segmento sonoro o sordo (con o sin componente glótica). A pesar de la notación, F_0 no constituye estrictamente una formante sino, como ya se ha indicado, la frecuencia fundamental, la cual está directamente relacionada con la entonación de una frase. De acuerdo al modelo lineal de producción de la voz, discutido en la sección 2.2, F_0 es una caracterítica de la fuente mientras que F_1 y F_2 corresponden a características del filtro.

Las formantes de la figura 2.4 han sido obtenidas a partir de vocales aisladas pronunciadas en forma sostenida. No obstante, en el caso del discurso continuo, las formantes siguen siendo un rasgo distintivo importante para las vocales [147]. En este caso es necesario seguir también la evolución de los patrones formánticos, debido a que a lo largo de una frase las variaciones en la morfología del tracto vocal y las características de la excitación dan como resultado un cambio permanente del espectro de la señal resultante. Estos patrones espectrales permiten caracterizar a los distintos fonemas mediante la identificación de determinadas *pistas acústicas*, requeridas para poder diferenciarlos.

En la figura 2.11 se muestra la variación en los patrones espectrales para la frase del español "¿Cómo se llama el mar que baña Valencia?". Esta frase corresponde a la base de datos de habla española Albayzin y ha sido segmentada en fonemas y etiquetada [148]. A partir del espectrograma de esta figura se pueden destacar algunas pistas acústicas, como por ejemplo, se puede observar la corta duración y la "explosión" del fonema oclusivo /k/. La estructura formántica de las vocales se evidencia en las regiones más oscuras, representadas por conjuntos equiespaciados de lineas paralelas en dirección horizontal, producto de su carácter sonoro cuasiperiódico. Se puede observar también el contenido de alta frecuencia de la /s/ y la ausencia de sonoridad.

En muchas aplicaciones, como por ejemplo el reconocimiento automático del habla, la segmentación automática de la señal o en tareas de clasificación, entre otras, normalmente se utilizan pre-procesamientos sobre la señal de voz (denominados *front-end* en inglés) para extraer diferentes características o pistas acústicas. Este pre-procesamiento o parametrización transforma la señal de habla en un vector de características, para su posterior procesamiento por parte de la aplicación. Esta forma de representar o codificar la señal tiene como objetivo no sólo reducir la dimensión de los datos a procesar, sino también captar los aspectos más destacados de la señal de voz, aquellos perceptualmente significativos o los cambios del espectro de la señal a lo largo del tiempo. Por otro lado, se espera que los parámetros o características obtenidas sean robustas, es decir, la tarea a realizar no se vea afectada por las distorsiones que puedan aparecer debido a efectos ambientales, diferencias entre hablantes o a los medios de transmisión utilizados. Los procesamientos empleados para obtener este vector o la selección de la información a codificar dependerá de la tarea que se desea resolver. En algunos casos, como por ejemplo en tareas de reconocimiento de habla, además de la información derivada del sonido físico de la señal, este vector puede contener características provenientes de otras fuentes como expresiones faciales y gestos articulatorios. A continuación, se mencionarán algunos de los métodos más frecuentemente utilizados para procesar la señal de habla.

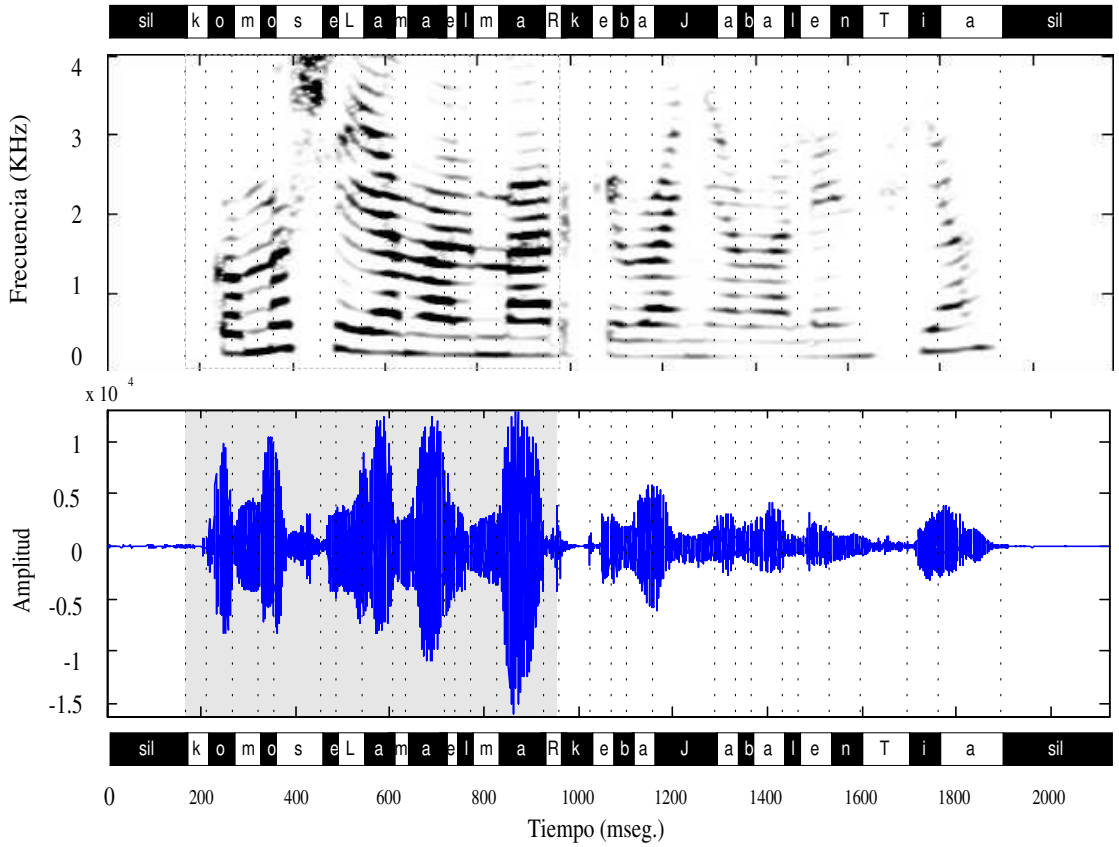

Figura 2.11: Sonograma y espectrograma de la oración "¿Cómo se llama el mar que baña Valencia?", segmentada y etiquetada (etiquetas de acuerdo al alfabeto fonético *Worldbet*).

2.4.1. Métodos de análisis y representación del habla

Existen algunas características de la señal de voz que se pueden identificar mediante análisis relativamente simples como ser la *energía de corta duración* y la *cantidad de cruces por cero* (Cx0) [149]. Estos análisis tienen la ventaja de ser sencillos en su implementación digital y muy rápidos. La energía da información acerca de la intensidad de la señal en función del tiempo y constituye un parámetro de suma importancia, ya que permite diferenciar entre varios tipos de fonemas. Además, también es una parte esencial de la entonación (junto con la F_0). Los cruces por cero brindan una medida indirecta del contenido frecuencial de la señal. La combinación simultánea de estos análisis permite una rápida caracterización de los diferentes fonemas. Por ejemplo, se puede hacer una distinción entre fonemas sonoros, que poseen una menor cantidad relativa de Cx0 que de energía, y fonemas sordos (no sonoros), que poseen poca energía, distribuida en las frecuencias altas.

Para llevar a cabo los procesos mencionados, hay que tener presente que cuando se utilizan señales de voz de habla continua, es necesario dividir la misma en intervalos o tramos de análisis. La selección del tamaño de este intervalo está relacionado con la máxima velocidad de modificación significativa en la morfología del tracto vocal [102]. Normalmente, la longitud del intervalo está entre 20 y 25 ms, lo cual incluye algunos períodos de la glotis. Por lo general, los tramos de análisis están solapados, de forma tal que sus centros se encuentran a unos 10 ms de diferencia. Cada intervalo se multiplica por una ventana fusiforme, de manera que los valores

cercanos a los bordes se vuelven cero y evita que las discontinuidades en los extremos afecten el resultado de la transformaciones posteriores.

Existen otros tipos de análisis que surgen a partir del estudio de las características perceptuales del oído o de los modelos de producción del habla, o bien, se emplean conceptos derivados de ambos esquemas (percepción-producción) en las representaciones finales [110, 149]. A continuación, se describirán algunos de estos análisis, los cuales, debido a su amplio estudio y utilización, se pueden considerar como convencionales en el área del procesamiento del habla.

2.4.2. Análisis cepstral

Un análisis comúnmente empleado para procesar la señal de voz es el denominado *cepstrum* [150]. Este método se basa en el modelo de producción de la voz, presentado en la sección 2.2, en el cual se plantea que la señal de habla corresponde a la salida de un sistema lineal ante una excitación de entrada. Es decir, la señal de voz está compuesta por una señal de excitación, $x(t)$ (fuente), convolucionada con la respuesta al impulso del modelo del tracto vocal, $h(t)$:

$$y(t) = x(t) * h(t). \tag{2.1}$$

Con el fin de poder analizar y modelar estos componentes en forma independiente, para que puedan ser usados en diversas aplicaciones de procesamiento de voz, es deseable eliminar una de las componentes, $x(t)$ o $h(t)$, de forma tal de poder examinar la restante. Esto es, en general, un problema difícil. Sin embargo, existen métodos para resolver este tipo de problemas cuando las señales están combinadas, como en este caso, mediante la convolución. Uno de estos métodos es el *análisis cepstral*, cuyo objetivo es separar el habla en sus componentes, sin ningún conocimiento previo acerca de la fuente y/o el sistema. Para poder realizar esta tarea se aplica la transformada de Fourier. De esta manera, en el dominio de la frecuencia, la ecuación (2.1) se convierte en un producto:

$$Y(f) = X(f)H(f). \tag{2.2}$$

Aplicando logaritmos en ambos miembros de la ecuación (2.2), el producto se convierte en una suma. Finalmente, es posible volver al dominio "temporal" (que se denomina *cuefrencia*) calculando la transformada de Fourier inversa de este último paso.

Así, mediante el cálculo del cepstrum de $y(t)$, es posible convertir la operación convolutiva en una adición. Por lo tanto, se define al cepstrum $C_y(t)$ de la señal $y(t)$ como:

$$C_y(t) = \mathcal{F}^{-1}\left\{\log\left(\mathcal{F}\left\{y(t)\right\}\right)\right\}, \tag{2.3}$$

donde $\mathcal{F}\{\cdot\}$ es el operador de la transformada de Fourier y se supone que $y(t)$ es generada por un sistema lineal invariante en el tiempo (LTI).

El cepstrum representa una transformación sobre la señal de voz con dos propiedades importantes sobre sus componentes: que las mismas se combinan linealmente y que pueden, además, aparecer separadas en el cepstrum. Para que ésto último ocurra es necesario que los espectros, $X(f)$ y $H(f)$, presenten diferencias en sus velocidades de cambio, de manera tal que sus componentes cepstrales aparezcan en cuefrencias diferentes. Este fenómeno se produce en las señales de voz, especialmente

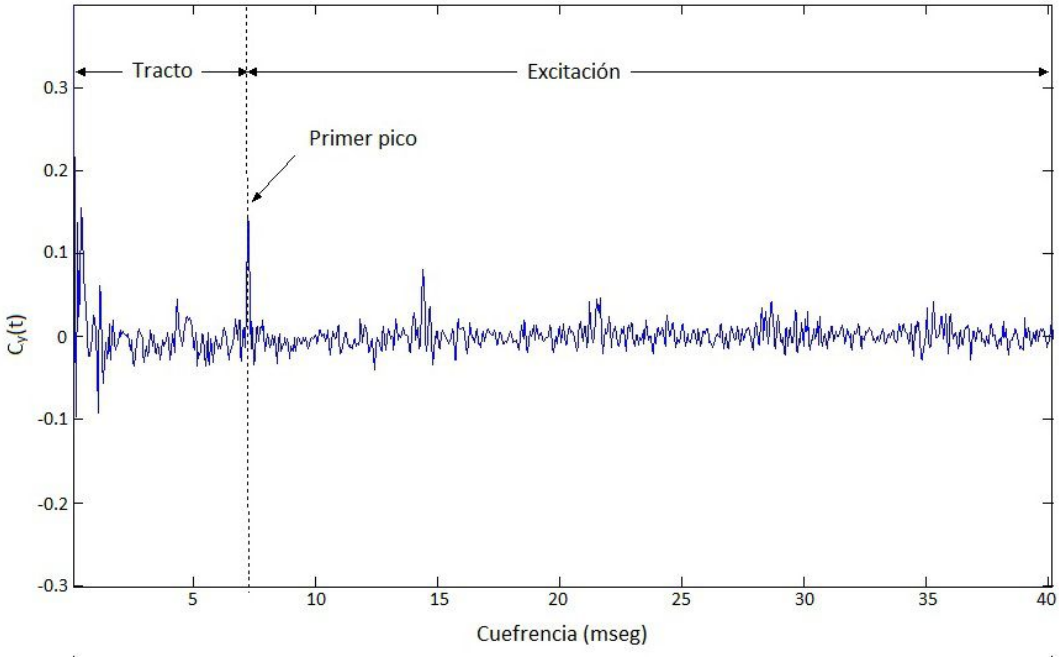

Figura 2.12: Cepstrum real correspondiente a un trozo de una vocal /e/ sostenida de un hablante masculino. Se puede apreciar que la parte de bajas cuefrencias (antes del primer pico) corresponde a la componente de la respuesta del tracto vocal, mientras que las altas cuefrencias corresponden a la componente de la excitación. Figura modificada de [101].

para los fonemas sonoros, donde el espectro de la excitación $X(f)$ se asemeja a un tren de pulsos decreciente, mientras que la respuesta en frecuencia del tracto vocal $H(f)$ es cuasi-continua con sólo algunos picos. Por lo tanto, en el dominio de la cuefrencia, los elementos del tracto vocal están representados por los componentes que se encuentran concentrados en la región más baja y corresponden a aquellos que varían lentamente. Por el contrario, los elementos de exitación se encuentran en la región superior de la cuefrencia y son los componentes de rápida variación [110]. Esta relación se utiliza en el campo del análisis del habla para, por ejemplo, separar la fuente de voz (sonora o sorda) del efecto de transmisión del tracto vocal (las resonancias que determinan los formantes) [149].

El análisis cepstral es un caso especial de una clase general de métodos conocidos como *procesamientos homomórficos*, los cuales son técnicas que involucran una transformación no lineal a un dominio diferente, donde se aplican filtros lineales para, luego, mapear nuevamente al dominio original. El cepstrum derivado del procesamiento homomórfico es comúnmente llamado *cepstrum complejo* (CC), mientras que, generalmente, el más utilizado para el habla es el*cepstrum real* (RC) [102]. El cepstrum real se calcula a partir del espectro de magnitud (parte real del CC), ignorando la fase. La diferencia básica entre el RC y el CC, es que el primero descarta información acerca de la fase de la señal, mientras que el CC la retiene. En la práctica, el CR es el más ampliamente utilizado debido a que el CC es difícil de aplicar y no es posible usarlo sobre señales estacionarias, lo cual excluye un gran número de aplicaciones [151]. Sin embargo, para poder realizar la reconstrucción de la secuencia es necesario preservar la fase, por lo que, en este caso se utiliza el CC.

Una de las aplicaciones más importantes del análisis cepstral en el procesamiento de la voz es su uso para la representación de un modelo de predicción lineal a partir de parámetros cepstrales, el cual se presenta en la sección 2.4.5. En este caso, la señal parametrizada es de fase mínima, una condición bajo la cual el RC y el CC son esencialmente equivalentes [151].

2.4.3. Bancos de filtros

Debido a que la discriminación de las frecuencias en nuestro oído no es lineal (ver sección 2.3.1), cuando se procesan señales de voz generalmente se utilizan bancos de filtros para las denominadas *bandas críticas* [152]. Las escalas "perceptuales" de frecuencia ofrecen normalmente las bases para la implementation de los bancos de filtros [102].

Se han propuesto varios tipos de filtros para las bandas críticas, siendo una de las configuraciones más usadas el de ventana triangular en la *escala psicoacústica de mel* [110]. En la escala de mel, la subdivisión de la frecuencia no es uniforme, y reproduce estrechamente la resolución espectral particular del oído humano. La escala de mel intenta mapear la frecuencia percibida de un tono en una escala que, frecuentemente, es aproximadamente lineal entre 0 y 1000 Hz y logarítmica más alla de 1000 Hz [153]. Esto conduce a diferentes aproximaciones, entre las cuales la comúnmente utilizada es [102]:

$$f_{mel} = 1000 \log \left[1 + \frac{f_{Hz}}{1000} \right], \tag{2.4}$$

donde f_{mel} (f_{Hz}) es la frecuencia percibida (real) en mels (Hz). La relación entre la escala lineal en Hz y la escala de mel se muestra en la figura 2.13. El uso de bancos de filtros para realizar la extracción de características de la señal de voz en escalas perceptuales como la mel conduce a representaciones como la *melbank* (este tipo de representación fue utilizada en los experimentos, por lo que para mayores detalles en cuanto a su implementación se remite al lector a la sección 5.3.3 del capítulo 5). Cuando estas escalas se combinan con el análisis cepstral, se obtiene una representación denominada *coeficientes cepstrales en escala de mel* (MFCC), la cual se describe a continuación.

2.4.4. Coeficientes cepstrales en frecuencias de mel

El análisis basado en coeficientes cepstrales en frecuencias de mel permite obtener una representación de la señal de voz emulando el análisis frecuencial que realiza el sistema auditivo.

Siguiendo el proceso descripto en la sección 2.4.2, se obtiene la transformada de Fourier de la señal de habla $X(k) = \mathcal{F}\{y(t)\}$. El espectro de magnitud $|X(k)|$ se escala en frecuencia y magnitud aplicando los bancos de filtros en escala de mel $H(k, m)$ y se obtiene luego el logaritmo de la siguiente manera:

$$Y(m) = \log \left(\sum_{k=0}^{N-1} |X(k)| \, H(k, m) \right), \tag{2.5}$$

para $m = 1, 2, ..., M$, donde M es el número de bancos de filtros.

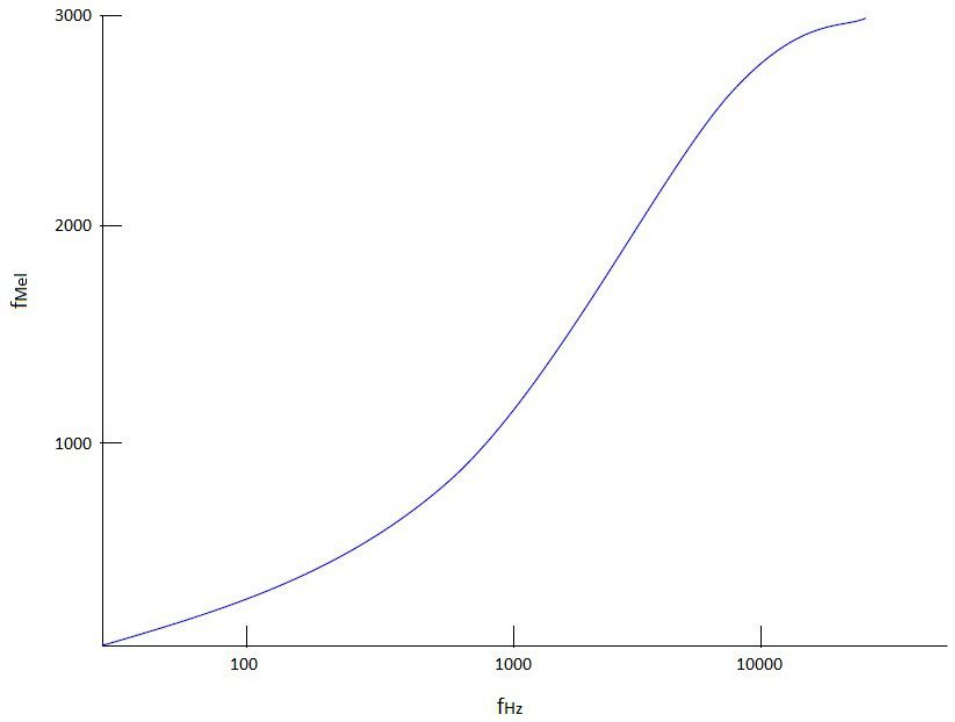

Figura 2.13: Relación entre la escala frecuencial lineal en Hz y la escala frecuencial de mel. Esta escala esta dada por la relación entre la altura tonal percibida y la frecuencia "real" obtenida a partir de experimentos de proporcionalidad entre sensaciones.

El banco de filtros es una colección de ventanas triangulares definidas por su frecuencia central $f_c(m)$. En la figura 2.14 se muestra la forma de un banco de filtros en escala de mel. Se observa que el primer filtro es estrecho y brinda información sobre la cantidad de energía en las frecuencias cercanas a 0 Hz. A medida que las frecuencias aumentan, los filtros se ensanchan, provocando que en las altas frecuencias se tengan menos en cuenta las variaciones del espectro. La expresión para $H(k,m)$ se escribe como:

$$H(k,m) = \begin{cases} 0 & \text{para} \quad f(k) < f_c(m-1) \\ \dfrac{f(k) - f_c(m-1)}{f_c(m) - f_c(m-1)} & \text{para} \quad f_c(m-1) \leq f(k) < f_c(m) \\ \dfrac{f(k) - f_c(m+1)}{f_c(m) - f_c(m+1)} & \text{para} \quad f_c(m) \leq f(k) < f_c(m+1) \\ 0 & \text{para} \quad f(k) \geq f_c(m+1) \end{cases} \tag{2.6}$$

La frecuencia central $f_c(m)$ del banco de filtros se obtiene utilizando la escala de mel de la ecuación (2.4) o una aproximación de la misma:

$$\phi = 2595 \log \left[1 + \frac{f_{Hz}}{700} \right]. \tag{2.7}$$

Se calcula una resolución de frecuencia determinada en escala de mel, correspondiente a una escala logarítmica de la frecuencia de repetición, usando $\Delta f_{mel} = (\phi_{max} - \phi_{min}) / (M+1)$, donde ϕ_{max} es la frecuencia más alta del banco de filtros

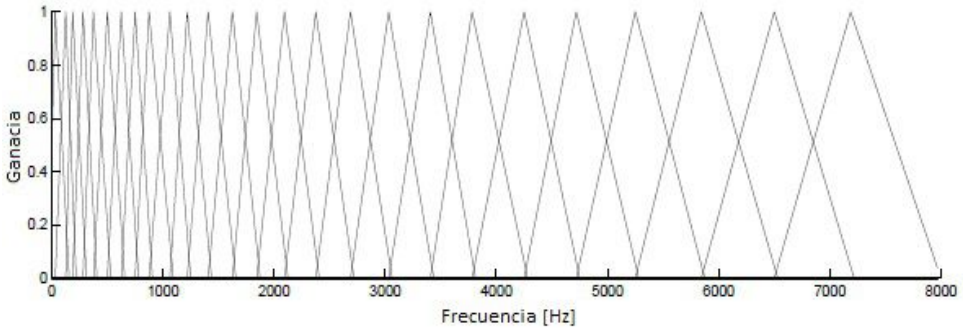

Figura 2.14: Colección de ventanas triangulares que determinan el banco de filtros en escala de mel.

en la escala mel, calculado a partir de f_{max} usando la ecuación (2.4) y ϕ_{min} es la frecuencia más baja en la escala de mel, con su correspondiente f_{min}.

La frecuencia central de la escala mel está dada por $\phi_c(m) = m\Delta\phi$ para $m = 1, 2, \cdots, M$. Para obtener la frecuencia central en Hertz, se aplica la inversa de la ecuación (2.4):

$$f_c(m) = 700\left(10^{\phi_c(m)/2595} - 1\right), \tag{2.8}$$

la cual se utiliza en la ecuación (2.5) para obtener el banco de filtros.

Finalmente, el conjunto de coeficientes MFCCs se obtienen calculando la transformada inversa ($\mathcal{F}^{-1}$) de $|X(k)|$.

2.4.5. Coeficientes de predicción lineal

El método de *análisis predictivo lineal* es una de las técnicas paramétricas de análisis del habla más potentes y se utiliza predominantemente para estimar parámetros básicos, como la frecuencia fundamental, las formantes, el espectro, funciones del área del tracto vocal y para representar el habla para transmisiones de baja velocidad o almacenamiento [149]. La importancia de este método radica en su habilidad de proveer estimaciones extremadamente precisas de los parámetros del habla y en su relativa rapidez de cálculo.

La idea básica del método de análisis predictivo lineal es que las muestras actuales de la señal de voz pueden ser aproximadas por una combinación lineal de sus muestras anteriores. Es decir, dada la señal de voz $y[n]$ de tiempo discreto, esta puede aproximarse mediante la salida de un sistema lineal de tiempo discreto[2], frente a una excitación o entrada $x[n]$, de acuerdo a la siguiente expresión:

$$\hat{y}[n] = -\sum_{q=1}^{Q} c_q\, y[n-q] + g\, x[n], \tag{2.9}$$

donde $\hat{y}[n]$ es la versión estimada de $y[n]$, c_q son los *coeficientes de predicción lineal* (LPC) que pesan las muestras sucesivas y dan cuenta de la relación entre ellas y $g \in \mathbb{R}$ es la ganancia de la excitación $x[n]$. Esto resulta compatible con un modelo lineal discreto *auto-regresivo* (AR) de producción de la voz, como el de la figura 2.15.

[2]Esto resulta similar al planteo de la sección 2.4.2, salvo por el hecho de que ahora el modelo es de tiempo discreto

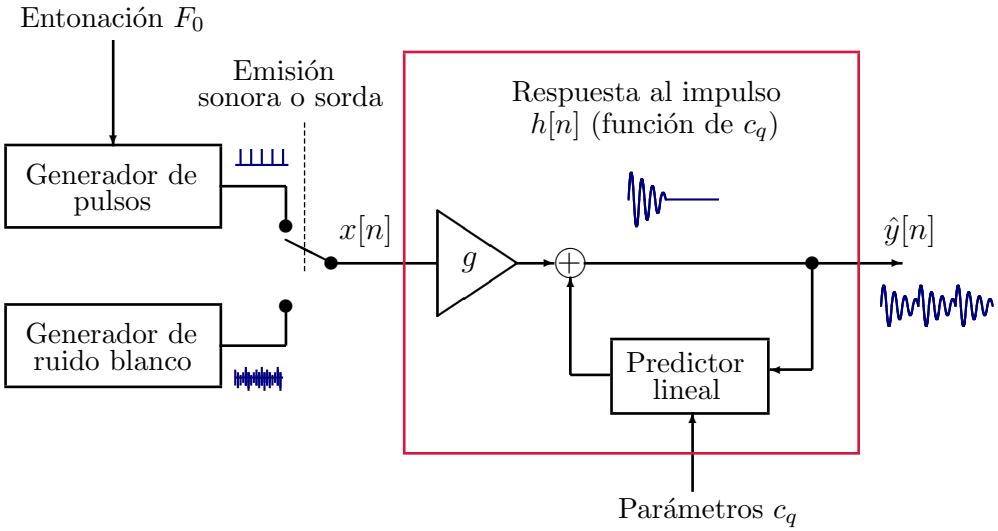

Figura 2.15: Diagrama para el modelo AR del aparato fonador, donde la señal de voz $y[n]$ se aproxima mediante la salida $\hat{y}[n]$ de un sistema lineal de tiempo discreto, frente a una excitación o entrada $x[n]$. Esta señal de excitación puede ser un tren de pulsos o ruido blanco dependiendo de si el fonema considerado es sonoro o sordo, respectivamente.

Es posible obtener un único conjunto de coeficientes de predicción c_q, mediante la minimización del valor esperado de la suma de las diferencias cuadradas entre las muestras reales del habla y las aproximadas linealmente, suponiendo que tanto x como y son variables aleatorias (rv).

De esta forma, para tramos relativamente estacionarios, el habla puede ser modelada mediante un sistema lineal que puede ser excitado por pulsos cuasi periódicos (durante habla sonora) o ruido aleatorio (durante habla sorda) (ver figura 2.15). Los métodos de predicción lineal proveen una forma precisa, confiable y robusta para la estimación de los parámetros que caracterizan este sistema lineal. Por otro lado, el modelo captura tendencias en la señal, como las formantes, dando un espectro suavizado [153]. A medida que se incrementa el orden del modelo utilizado, se mejora la predicción de la señal. Sin embargo, si bien el agregado de nuevos parámetros al modelo permite lograr una mejor aproximación a la señal, también lo hacen más sensible al ruido que puede estar presente en los datos.

El método de predicción lineal, aplicado al procesamiento del habla, admite una variedad de formulaciones para la modelización de la señal de voz que son esencialmente equivalentes. Las diferencias entre estas formulaciones están relacionadas con el enfoque propuesto o con los detalles de los cálculos usados para obtener los coeficientes de predicción. Se han desarrollado una gran variedad de aplicaciones basadas en el análisis por LPC para el procesamiento del habla. Este método ha sido usado en muchos sistemas de análisis y procesamiento para tareas como verificación e identificación de hablantes, ASR, clasificación, derreverberación, entre otras [102].

Así como están planteados, los LPC ponderan uniformemente todo el espectro, lo cual no es consistente con la forma en que se procesa el sonido en el sistema auditivo. El análisis por predicción lineal perceptual, que se describe a continuación (sección 2.4.6), incorpora una ponderación de las frecuencias motivada fisiológicamente [154].

2.4.6. Análisis predictivo lineal perceptual

El análisis por *predicción lineal perceptual* (PLP), introducido por Hermansky [154], incorpora un pesado de las frecuencias basado en aspectos fisiológicos. En este sentido, se tiene en cuenta una resolución frecuencial no lineal en las bandas críticas, en forma similar al mel cepstrum, pero en la escala denominada de *Bark*. Esta escala presente una asimetría de los filtros auditivos e integración más ancha que la de las bandas críticas. Esto provee una sensibilidad desigual a diferentes frecuencias y una relación no-lineal entre la intensidad física del sonido y la sensación correspondiente.

El objetivo de esta técnica es alterar el espectro para minimizar las diferencias entre hablantes, pero preservando la información importante.

Además de que el análisis por PLP es más consistente con la audición humana que el análisis por LPC convencional, este método es eficiente computacionalmente y brinda una representación del habla dimensionalmente baja, y sus características han sido de utilidad en tareas de reconocimiento automático del habla independiente del hablante.

A este enfoque se le añadieron posteriormente una serie de filtros temporales para mejorar el comportamiento del método frente a distorsiones debidas a fenómenos de variación lenta. Esta técnica se denominó *transformación espectral relativa por PLP* (RASTA-PLP) [155].

2.4.7. Modelos auditivos

Es posible aprovechar los conocimientos acerca de la anatomía y fisiología del sistema auditivo para elaborar un modelo de oído que rescate las pistas acústicas más significativas para el análisis [10, 156, 157, 158, 159]. Generalmente este enfoque requiere un mayor tiempo de cálculo, aunque se han reportado modelos bastante "exactos" que se han optimizado en este sentido [158]. Mayoritariamente estos modelos contemplan hasta las denominadas representaciones auditivas tempranas y tienen en cuenta las siguientes consideraciones:

1. El meato auditivo no afecta substancialmente a la señal sonora y es por ello que se considera con transferencia igual a la unidad.

2. La cadena de huesecillos junto con los músculos correspondientes se suele asimilar a un amplificador de ganancia controlada.

3. La membrana basilar se asimila a un banco de filtros de bandas críticas (esta etapa se considera muy importante).

4. La codificación eléctrica llevada a cabo en las células ciliadas se incorpora como una "rectificación".

5. Los nervios y los núcleos se asimilan a un mecanismo sencillo de inhibición lateral.

Con respecto al procesamiento en la corteza, se trata de un análisis de nivel superior que, por lo tanto, no forma parte de los modelos clásicos utilizados en la etapa de extracción de características o análisis sino más bien de las etapas siguientes.

2.5. Comentarios de cierre del capítulo

En este capítulo se han presentado los diferentes aspectos del proceso de la comunicación humana, tanto de la emisión del sonido como de su percepción, que determinan las características principales del mismo. Dichas características podrían explicar, total o parcialmente, las diferencias de robustez y adaptación entre los humanos y los sistemas artificiales a la hora de realizar actividades que involucren procesamiento de habla. El desafío consiste en poder integrar adecuadamente estas particularidades en las diferentes etapas de los sistemas artificiales, comenzando por el procesamiento y la representación de la señal de voz. En este sentido, se han comentado los diferentes tipos de procesamiento de habla, que pueden ser considerados como clásicos, que tienen en cuenta aspectos fisiológicos, basados en la forma en que se produce o procesa la señal de voz.

Cabe preguntarse qué otros aspectos del proceso de la comunicación humana pueden tenerse en cuenta para representar la señal de voz. Estos interrogantes son los que se desarrollan en los capítulos subsiguientes, donde se plantea una nueva forma de obtener coeficientes para codificar la señal de voz y los resultados obtenidos cuando los mismos, solos o en conjunto, se utilizan en tareas como el reconocimiento automático del habla o la segmentación automática de fonemas.

Capítulo 3

Parametrización del Habla mediante Medidas de Información Multiresolución

3.1. Introducción

En el capítulo anterior se describió la forma en que la señal de voz se produce en el aparato fonador y se procesa en el sistema auditivo. Se mencionaron, también, diferentes formas de representar (parametrizar o codificar) la señal de habla para su posterior utilización en diferentes tareas. Estas parametrizaciones pueden considerarse como clásicas y muchas de ellas tienen en cuenta los aspectos biológicos de producción o percepción de la señal de voz. Estos métodos buscan caracterizar el estado del sistema que produjo la señal. Sin embargo, en diversas aplicaciones es también de interés poder obtener información acerca de los cambios que se producen en la dinámica del mismo, para lo cual se propone en esta tesis la utilización de medidas de información en el dominio tiempo–escala.

En este capítulo se presentarán en detalle los procedimientos utilizados para obtener la parametrización de la señal de habla a partir de la utilización de medidas de información multiresolución como la entropía multiresolución continua (CME), introducida por Torres y col. [38] para la detección de variaciones suaves en los parámetros de sistemas con dinámicas no lineales, y la divergencia multiresolución continua (CMD), propuesta en el marco de esta tesis [95].

Para este tipo de procesamiento se parte de una representación en tiempo–escala de la señal de habla, sobre la cual se calculan, luego, las medidas de información. Esta representación tiempo–escala se obtiene a partir de la transformada ondita continua (CWT), que se explica en la sección 3.2. Este tipo de representación permite descomponer la señal de interés en un conjunto de señales continuas en diferentes escalas de observación, lo cual permite estudiar cada componente con distinta resolución. Asimismo, la CWT brinda una representación que es que es redundante respecto a su versión discreta. lo cual tiende a reforzar los rasgos particulares de la señal, especialmente aquellos con información muy sutil, por lo cual permite, además, preservarla mejor en presencia de ruido. Las transformadas onditas redundantes se utilizan frecuentemente para el análisis de señales, la extracción de características o tareas de detección, ya que proporcionan una descripción invariante al cambio o traslación temporal que resulta de interés para las aplicaciones propuestas en esta

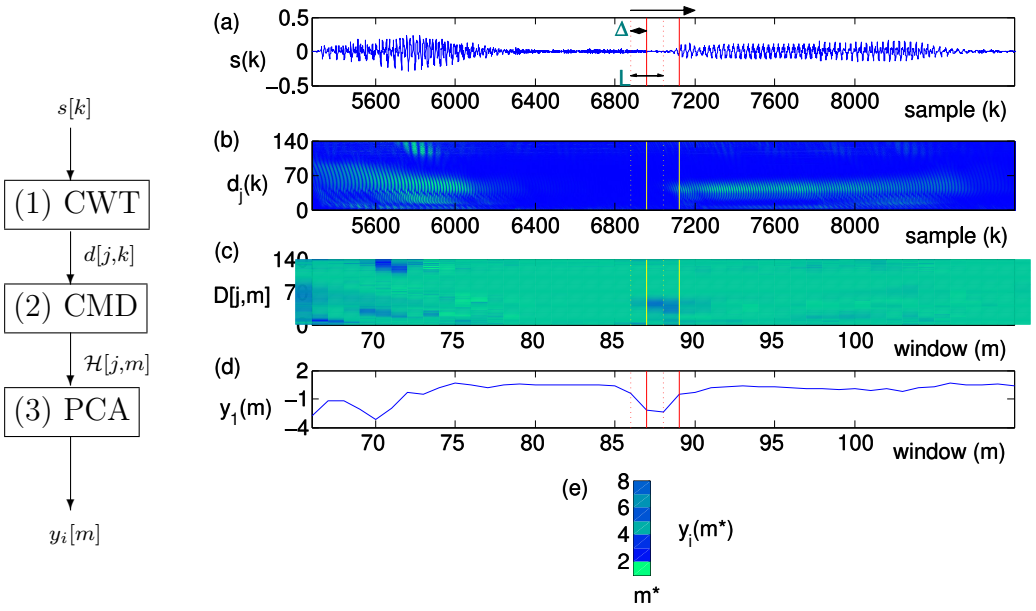

Figura 3.1: Diagrama (izquierda) y figuras (derecha) de las etapas que comprende el método propuesto de la CMD, las cuales se explican en las Sec. 3.4.2. (a) Señal de habla. (b) Escalograma correspondiente a la señal mostrada en (a). (c) CMD del escalograma presentado en (b). (d) Componente principal de la CMD de la imagen (c), obtenida a partir de la aplicación del método PCA. (e) Conjunto de características correspondientes a la ventana m^*, obtenido a partir de los dos períodos de análisis indicados con las líneas punteadas verticales.

tesis [160]. Por otro lado, este tipo de representaciones está en concordancia con la manera en que el cerebro procesa las señales sensoriales, dado que está comprobado que este es un proceso altamente redundante [103].

Como medidas de información pueden utilizarse entropías o divergencias, las cuales se describen en la sección 3.3. Dichas medidas de información son evaluadas para cada escala de la CWT. Los detalles acerca de las formas de obtener la CME y la CMD se indican en las secciones 3.4.1 y 3.4.2, respectivamente.

A fin de lograr un conjunto reducido de coeficientes para representar la señal, se utiliza el análisis de componentes principales (PCA) para extraer las características temporales de mayor varianza. Dicha técnica se describe en la sección 3.5. Mediante este mecanismo de selección de componentes se obtiene una nueva parametrización de la señal de habla, que puede utilizarse sola o en conjunto con otras codificaciones como entrada a los sistemas de reconocimiento de habla o para realizar la segmentación de fonemas. El procedimiento para obtener los coeficientes a partir de la CME o CMD se indica en la sección 3.6.

A modo de ejemplo, en la figura 3.1 se muestra un diagrama que resume las etapas que comprende el método propuesto para realizar la parametrización de la señal de habla a partir de los procesamientos explicados anteriormente.

Finalmente en las sección 3.7 se mencionan las diferentes aplicaciones en las que se ha utilizado esta nueva parametrización, que luego se desarrollarán en los capítulos subsiguientes.

3.2. La Transformada Ondita Continua

La teoría de onditas se ha desarrollado a partir de distintas líneas de pensamiento y desde diferentes enfoques. Este tipo de procesamiento ha resultado de utilidad para extraer información de diferentes tipos de señales, en particular cuando éstas presentan características que complejizan su análisis, como discontinuidades o picos abruptos, o tienen la propiedad de ser finitas, aperiódicas o no estacionarias. En este sentido, la trasformada ondita presenta ventajas sobre los métodos tradicionales y se ha utilizado en diferentes ámbitos, como por ejemplo, para el estudio de señales de electroencefalografía y electromiografía, tacogramas (series RR para análisis de la variabilidad del ritmo cardíaco), secuencias biológicas, señales de audio e imágenes, por mencionar sólo algunos ejemplos [25, 161, 162, 163, 164, 165, 166, 167, 168].

La transformada ondita continua (CWT) es un método de análisis tiempo–frecuencia que permite localizar temporalmente características particulares de la señal analizada. Para ello, se descompone la señal de interés en un conjunto de señales continuas en diferentes escalas de observación, lo cual permite estudiar cada componente con distinta resolución [169].

Para obtener esta transformada, la señal a analizar se proyecta sobre un conjunto de funciones, las cuales se denominan onditas. Esta familia de funciones está formada por la ondita madre y por versiones trasladadas y escaladas de la misma. Así, dada una señal continua $s(t)$, su transformada ondita continua queda definida de la siguiente manera [170]:

$$\Psi_s(a,b) \triangleq \int_{-\infty}^{\infty} |a|^{-1/2} s(t)\ \bar{\psi}\left(\frac{t-b}{a}\right) dt, \tag{3.1}$$

donde $\bar{\psi}(t)$ es el complejo conjugado de la ondita madre ψ.

La ondita puede ser cualquier función oscilatoria con media cero y energía finita, con lo cual su trasformada de Fourier $\hat{\psi}(\omega)$ debe satisfacer la siguiente condición:

$$C_\psi = 2\pi \int_{-\infty}^{\infty} |w|^{-1} |\hat{\psi}(w)|^2 dw < \infty, \tag{3.2}$$

conocida como condición de admisibilidad.

El parámetro a de la ecuación (3.1) corresponde al factor de escala y b determina la localización temporal. Por lo tanto, las funciones $\psi_{a,b}(t) = |a|^{-1/2}\psi\left((t-b)/a\right)$ corresponden a las versiones dilatadas y trasladadas de la ondita madre $\psi(t)$.

En el caso de la tranformación de tipo continua se opera sobre todas las posibles escalas y traslaciones de $\psi(t)$, a diferencia de su par discreto que sólo lo hace en un conjunto específico de escalas y traslaciones. Esto permite analizar la señal en diferentes tamaños de estructura. Esta es la principal ventaja de este tipo de transformación con respecto a otras representaciones, pues permite analizar simultáneamente eventos de corta duración y alta frecuencia, junto con características de larga duración y baja frecuencia que pueden estar presentes en la señal. Otra ventaja es la diversidad de tipos de onditas disponibles para realizar la transformación, lo cual permite elegir la más apropiada según el tipo de señal que se esté analizando.

Debido a que el conjunto de funciones bases de la transformada ondita son generadas a partir de traslaciones y dilataciones de una función base, $\psi(t)$, el resultado de la transformación no está ligado a frecuencias de modulación, como es el caso de la transformada de Fourier, sino que está asociada con un esquema de escalas.

Mediante el uso de la CWT se obtiene un conjunto de coeficientes que dan una medida de cuán similar es la señal que se desea analizar, $s(t)$, de una función particular de la base, $\psi_{a,b}(t)$. Así, con la CWT se obtiene una forma de representar una señal continua en un plano tiempo–escala, al que se denomina escalograma.

Como en la práctica normalmente las señales se procesan computacionalmente, $s(t)$ es una señal discreta, i.e. $s(t) = s[k]$ si $t \in [k, k+1]$ para $k = 1, \ldots, K$. Por lo tanto, para calcular numéricamente la CWT definida en (3.1) se utiliza una interpolación constante por tramos [171], la cual tiene la siguiente forma:

$$\Psi_s(a,b) = |a|^{-1/2} \sum_k s[k] \int_k^{k+1} \bar{\psi}\left(\frac{t-b}{a}\right) dt. \tag{3.3}$$

Dado que la ecuación (3.3) se puede reescribir como:

$$\Psi_s(a,b) = |a|^{-1/2} \sum_k s[k] \left(\int_{-\infty}^{k+1} \bar{\psi}\left(\frac{t-b}{a}\right) - \int_{-\infty}^{k} \bar{\psi}\left(\frac{t-b}{a}\right) dt \right), \tag{3.4}$$

la evolución de los coeficientes $\Psi_s(a,b)$, para $b = k$ (con $k = 1, \ldots, K$) y cualquier escala $a = j\,\delta$ (donde $j = 1, \ldots, J$, con $J \in \mathbb{Z}$ y $\delta \in \mathbb{R}^+$), se puede obtener haciendo la convolución de la señal s con la versión dilatada y trasladada de la integral

$$\int_{-\infty}^{k} \bar{\psi}(t)\, dt, \tag{3.5}$$

y tomando diferencias finitas. Esto produce una versión discretizada de $\Psi_s(a,b)$ que corresponde a la denominada transformada ondita "cuasi-continua". De esta manera, se obtiene una descomposición discreta $\{d[j,k]\} = \{\Psi_{\mathbf{s}}(a = j\delta, b = k)\}$ en el plano tiempo–escala.

En este trabajo, con el objetivo de simplificar la notación, a la evolución temporal de los coeficientes de la CWT para cada escala fija j se la denotará como $\mathbf{d}_j = \{d_j[k]\}$.

3.3. Medidas de Información

La teoría de la información se considera que fue desarrollada por Claude E. Shannon en 1948, con el fin de poder cuantificar la información para determinar las limitaciones de operaciones matemáticas como la compresión, reserva de recursos y comunicación [172, 173]. Desde sus comienzos, sus aplicaciones se han expandido hacia otras áreas de aplicación, como el procesamiento de señales elecrofisiológicas y señales de habla, selección de modelos en ecología, neurobiología, estadística y genética, entre otras [173, 174, 175, 176, 177].

En el marco de esta teoría se establece que la información, tratada como una magnitud física, puede ser medida. Es decir, se busca caracterizar la secuencia de símbolos que componen dicha información. Una de la medidas que se utiliza para esto es la que se conoce como entropía. Diversas medidas de información se han introducido a partir de la teoría de Shannon y otras fuentes [178]. Entre ellas se han propuesto medidas de información relativas, las cuales están asociadas a dos o más distribuciones y expresan la cantidad de información suministrada por los datos

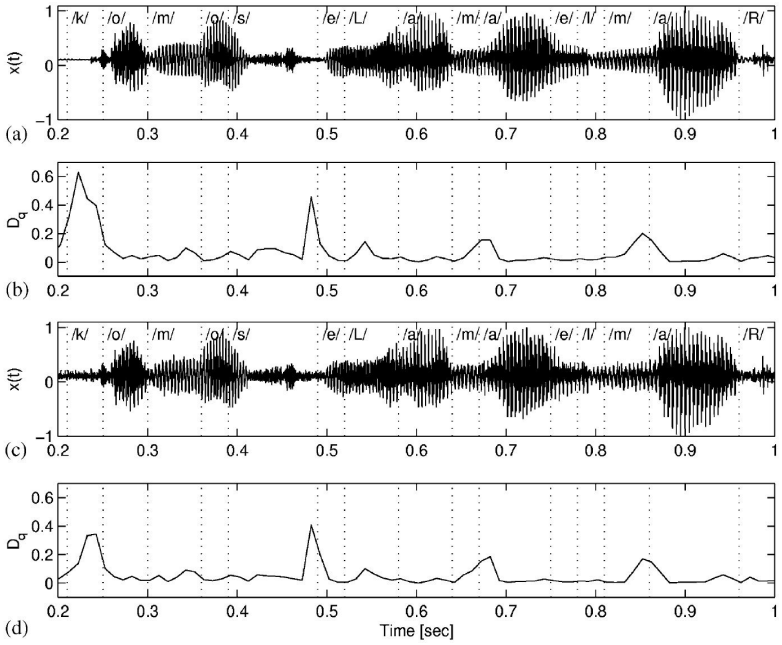

Figura 3.2: (a) Señal de habla limpia. (b) Evolución de la divergencia de Tsallis de la señal mostrada en (a). (c) La misma señal de (a) contaminada con ruido aditivo blanco (20dB). (d) Evolución de la divergencia de Tsallis de la señal (c). Figura tomada de [68].

para discriminar las mismas. La importancia de obtener medidas de "distancia" entre probabilidades permite resolver problemas de inferencia y discriminación [179].

La noción de entropía se ha utilizado, también, para caracterizar el grado de complejidad de señales fisiológicas [180, 181]. La aplicación de estas medidas cuantitativas proporciona información acerca de la dinámica no lineal subyacente en los sistemas que generan estas señales y ayuda a obtener una mejor comprensión de las mismas. Su utilización se ha extendido, además, sobre diferentes distribuciones tiempo–escala [38, 77, 78, 182].

En el campo del procesamiento del habla, el concepto de entropía se ha aplicado de diversas maneras. Las entropías de Shannon y Tsallis, y sus correspondientes divergencias, y la divergencia de Jensen-Shannon, se han utilizado en numerosas aplicaciones para el análisis del habla [67, 69, 70, 71, 72, 73]. En particular, Rufiner y col. [68] encontraron que las entropías y divergencias aplicadas al habla brindan información acerca de los cambios de dinámica de la señal que se mantienen aun en presencia de ruido. En la figura 3.2 se muestra parte de una señal de voz limpia (3.2 (a)) y la misma señal contaminada con ruido aditivo blanco a 20 dB SNR (fig. 3.2 (c)), sobre las cuales se ha calculado la divergencia de Tsallis (3.2 (b) y (d)). A partir de la misma se puede apreciar que existe una correspondencia entre las variaciones que presenta la divergencia y los cambios fonéticos, y que dichas variaciones se localizan en posiciones similares en la señal de habla limpia y en la que está contaminada con ruido, lo cual sugiere una cierta robustez al ruido por parte de esta medida. Estos investigadores han encontrado que la utilización complementaria de información obtenida a partir de entropías y divergencias aplicadas sobre la señal de voz en el dominio temporal, permite mejorar el desempeño de un sistema de reconocimiento automático del habla.

La entropía espectral, por su parte, ha sido utilizada para tareas relativamente simples, como la segmentación de palabras y oraciones y en la detección de silencios [74, 75, 183, 184, 185, 186]. Y se han obtenido buenos resultados en la aplicación de la CME a señales de voz corrompidas con ruido aditivo en experimentos de agrupamiento de mapas auto-organizativos [79].

En esta tesis se consideran como medidas de información, para la obtención de la CME o la CMD, las entropías de Shannon y Tsallis, y sus correspondientes divergencias, así como también la divergencia de Jensen-Shannon. A continuación se indica la forma de obtener dichas medidas de información.

3.3.1. Entropía de Shannon

Dado una variable aleatoria discreta (rv) $\mathbf{x} \in I = [x_{min}, x_{max}]$, su entropía de Shannon $\mathcal{H}_{\mathbf{x}}$ se obtiene de la siguiente manera [172]:

$$\mathcal{H}_{\mathbf{x}}(P) \triangleq -\sum_{n=1}^{N} p_n \log(p_n), \tag{3.6}$$

donde $P = \{p_n,\, n = 1, \ldots, N\}$, p_n es la probabilidad que $\mathbf{x}$ pertenezca a un dado intervalo, N es el número de particiones en los cuales en rango I es uniformemente dividido. La notación $\mathcal{H}_{\mathbf{x}}(P)$ indica que la entropía depende de la distribución de probabilidad de $\mathbf{x}$, en lugar de sus valores.

La entropía de Shannon es una medida relacionada con la cantidad de información necesaria para localizar un sistema en un cierto estado, en donde $\mathcal{H}$ indica nuestra ignorancia acerca del sistema. En caso que exista la certeza de que $\mathbf{x}$ pertenece a determinado intervalo, $p_n = 1$ y $p_n \log(p_n) = 0$, lo cual implica pérdida de información. Del mismo modo, $p_n \log(p_n) = 0$ si $p_n = 0$.

3.3.2. Entropía de Tsallis

La entropía de Tsallis depende de un parámetro real $q \neq 1$ y la misma se define como [187]:

$$\mathcal{H}^q{}_{\mathbf{x}}(P) \triangleq (q-1)^{-1} \sum_{n=1}^{N} (p_n - (p_n)^q). \tag{3.7}$$

Nuevamente, p_n es la probabilidad que $\mathbf{x}$ pertenezca a un dado intervalo y el parámetro q es una medida de cuán correlacionados están estos estados. En el límite donde $q \to 1$ esta entropía se reduce a la entropía de Shannon.

Las principales aplicaciones de la entropía de Tsallis se han llevado a cabo en el contexto de sistemas dinámicos no lineales para estudiar fenómenos caóticos [188, 189]. Diversas investigaciones han utilizado esta entropía en una variedad de campos, para capturar comportamientos que siguen la ley de potencia [190]. Asimismo, este tipo de medida se ha aplicado en el análisis de señales biológicas complejas para detección de cambios suaves en parámetros en el contexto de sistemas dinámicos no lineales [38, 191].

3.3.3. Divergencia de Kullback-Leibler

En el caso de la entropía de Shannon, su entropía relativa asociada se obtiene a partir de dos distribuciones de probabilidad P y R, correspondientes a dos rv's $\mathbf{x}$ y

$\mathbf{y}$, la cual se expresa como [173]:

$$D_{\mathbf{x},\mathbf{y}}(P|R) \triangleq \sum_{n=1}^{N} p_n \log\left(\frac{p_n}{r_n}\right). \tag{3.8}$$

$D_{\mathbf{x},\mathbf{y}}(P|R)$ también se conoce como divergencia de Kullback-Leibler. A diferencia de las entropías antes presentadas, aquí intervienen dos probabilidades, cuya forma de obtención se describe en la sección 3.4.2. Esta entropía relativa mide la "distancia"[1] entre dos distribuciones de probabilidad. En este caso, $p_n \log(p_n/r_n) = 0$ si $p_n = 0$ y si $p_n = r_n$.

La entropía relativa permite medir la disimilitud entre dos grupos de datos en base a la distancia o divergencia entre sus distribuciones de probabilidad P y R. Esto tiene utilidades tanto en los problemas de inferencia estadística, como en las aplicaciones que estudian afinidades entre conjuntos dados de poblaciones. La divergencia mide la diferencia esperada, desde la perspectiva de P, entre la información contenida en P y la información contenida en R. En términos de la relación p_n/r_n, la diferencia en las distribuciones es mayor cuanto más se aleja p_n/r_n de 1.

3.3.4. Divergencia de Tsallis

De forma similar, la divergencia correspondiente a la entropía–q [68, 192] está dada por:

$$D^{q}{}_{\mathbf{x}}(P|R) \triangleq \frac{1}{1-q}\sum_{n=1}^{N} p_n \left[1 - \left(\frac{p_n}{r_n}\right)^{q-1}\right]. \tag{3.9}$$

3.3.5. Divergencia de Jensen-Shannon

Finalmente, se considera la divergencia de Jensen-Shannon [193], la cual comparte propiedades similares a las mencionadas en las secciones anteriores. La misma se define como:

$$D^{JS}{}_{\mathbf{x}}(P|R) \triangleq \mathcal{H}_{\mathbf{x}}(\pi_P\, P + \pi_R\, R) - (\pi_P\, \mathcal{H}_{\mathbf{x}}(P) + \pi_R\, \mathcal{H}_{\mathbf{x}}(R)), \tag{3.10}$$

donde $\mathcal{H}_{\mathbf{x}}(\cdot)$ es la entropía de Shannon y π representa el peso asignado a cada distribución. Además, $\pi_P, \pi_R \geq 0$ y $\pi_P + \pi_R = 1$. Dado que $\mathcal{H}_{\mathbf{x}}(\cdot)$ es una función cóncava, de acuerdo a la desigualdad de Jensen, D^{JS} es no negativa e igual a cero cuando $P = R$.

Esta medida de información relativa también permite cuantificar las diferencias entre distribuciones de probabilidad. Su principal característica radica en la posibilidad de poder asignar diferentes pesos a cada una de las distribuciones de probabilidad involucradas, de acuerdo a su importancia. Asimismo, la divergencia de Jensen-Shannon no requiere la condición de absoluta continuidad para las distribuciones de probabilidad involucradas y puede extenderse a una cantidad arbitraria de distribuciones, no solamente dos [194].

[1] Si bien no es una verdadera distancia, puesto que no es simétrica y tampoco cumple con la desigualdad triangular, comparte algunas propiedades como ser que es siempre positiva y es cero cuando $p_n = r_n$. En algunos casos se hace referencia a este tipo de medidas como pseudo-distancias. No obstante, en la bibliografía también se puede encontrar a esta medida de información con el nombre de *distancia de Kullback-Leibler*.

3.4. Representación del Habla con Medidas de Información Tiempo-Escala

En esta sección se detalla la forma de obtener la representación del habla basadas en medidas de información multiresolución. Estas representaciones se denominan Entropía Multiresolución Continua (CME) y Divergencia Multiresolución Continua (CMD), de acuerdo al tipo de medida de información utilizada, entropías o divergencias, respectivamente.

Esto consiste básicamente en obtener las diferentes entropías sobre la representación cuasi-continua en tiempo-escala de la señal de habla. nstationary phenomena [19] ; [20].

3.4.1. Entropía Multiresolución Continua

Dada la evolución temporal de los coeficientes de la CWT, $\{d_j[k]\}$ (con $d_j[k] = d[j,k]$), que se obtienen como se describe en la sección 3.2, para obtener la CME, cada escala fija j se divide en un conjunto de $\mathcal{W}^j = \{W^j(m,L,\Delta), m = 0,1,2,...,M\}$ ventanas rectangulares deslizantes, donde

$$W^j(m,L,\Delta) = \{d_j[k], k = l + m\Delta,\ l = 1,...,L\}\,, \tag{3.11}$$

las cuales determinan los intervalos de análisis. Estos intervalos dependen de dos parámetros, el ancho $L \in \mathbb{N}$ y el desplazamiento $\Delta \in \mathbb{N}$, donde L y Δ se eligen de manera tal que $L \leq K$ (la longitud de la señal) y $(K-L)/\Delta = M \in \mathbb{Z}$. La selección de estos valores se realiza en concordancia con el ventaneo que se lleva a cabo para obtener la parametrizaciones clásicas de la señal de habla, como la Melbank, mencionadas en la sección 2.4.1 a 2.4.6, que luego son utilizadas para evaluar la propuesta presentada. Es decir, se usan las mismas ventanas de análisis para obtener la CME y las parametrizaciones de referencia.

Sobre cada una de las ventanas $W^j(m,L,\Delta)$, se obtiene un histograma de amplitudes mediante la equipartición $\min_k\{d_j[k]\} = d_j^0 < d_j^1 < \cdots < d_j^{N-1} < d_j^N = \max_k\{d_j[k]\}$, la cual provee de un subconjunto de N subintervalos disjuntos

$$I_n^j = \left\{ \left[d_j^{n-1}, d_j^n\right), n = 1,...,N \right\}, \tag{3.12}$$

tales que $W^j(m,L,\Delta) = \overline{\bigcup_{n=1}^{N} I_n^j}$.

Se denota con $p_m^j(I_n^j)$ la probabilidad de que un dado $d_j[k] \in W^j(m,L,\Delta)$ pertenezca al intervalo I_n^j, que será estimada mediante el histograma de los valores de $d_j[k]$ en cada conjunto $W^j(m)$. Por lo tanto, para cada ventana $W^j(m,L,\Delta)$ se obtiene, mediante el valor del histograma normalizado, un conjunto $P^j[m]$ de probabilidades $p_m^j(I_n^j)$ de dimensión finita N:

$$P^j[m] = \{p_m^j(I_n^j),\ n = 1,\ldots,N\}, \tag{3.13}$$

Debe tenerse presente que m representa aquí la evolución en el eje temporal sobre la escala j considerada.

La entropía Shannon (3.6) del conjunto de probabilidades de cada ventanas $W^j(m, L, \Delta)$ se escribe como:

$$\mathcal{H}_{\mathbf{d}}[j, m] = -\sum_{n=1}^{N} p_m^j(I_n^j) \log\left(p_m^j(I_n^j)\right) \qquad m = 0, 1, \ldots, M. \tag{3.14}$$

Aquí $\mathcal{H}_{\mathbf{d}}$ representa $\mathcal{H}_{\mathbf{d}}(P)$ y en las notaciones que siguen se omitirá la referencia a la probabilidad en todas las medidas de información, a fin de que la notación sea más legible.

Para cada escala fija j y por cada m se obtiene el valor de la entropía correspondiente a los coeficientes ondita de la ventana $W^j(m, L, \Delta)$. Por lo que $\{\mathcal{H}_{\mathbf{d}}[j, m],\ m = 0, 1, \ldots, M\}$ representa la evolución de la entropía Shannon en el instante m. De este modo, con la matriz $\{\mathcal{H}_{\mathbf{d}}[j, m],\ j = 1, \ldots, J,\ m = 0, \ldots, M\}$ se obtiene finalmente la entropía multiresolución continua, que se denota como **CME**, donde $CME(j, m) = \mathcal{H}_{\mathbf{d}}[j, m]$.

Este procedimiento puede aplicarse utilizando otras medidas de información como la entropía de Tsallis (3.3.2). En este caso, la fórmula planteada en 3.14 debe reescribirse de acuerdo a la ecuación planteada en 3.7. Así, la evolución de la entropía de Tsallis o q–entropía calculada sobre cada ventana de $d_j[k]$ es:

$$\mathcal{H}^q{}_{\mathbf{d}}[j, m] = (q-1)^{-1} \sum_{n=1}^{N} (p_m^j(I_n^j) - \left(p_m^j(I_n^j)\right)^q). \tag{3.15}$$

$CME_q(a = j\delta, m) = \mathcal{H}^q{}_d[j, m]$ es la correspondiente matriz de q-entropía multiresolución continua.

3.4.2. Divergencia Multiresolución Continua

En esta sección la idea de entropía multiresolución se extiende a las medidas de información relativas, utilizando la divergencia de Kullback-Leibler.

A partir del conjunto de probabilidades $P^j[m]$ mencionadas anteriormente (3.13), correspondientes a una ventana $W^j(m, L, \Delta)$ de valores de $d_j[k]$, se considera ahora un segundo conjunto de probabilidades, $R^j[m] = \{r_m^j(I_n^j),\ n = 1, \ldots, N\}$ para la siguiente ventana $W^j(m+1, L, \Delta)$. Para ello, a partir de $W^j(m+1, L, \Delta)$ se obtiene nuevamente el histograma de amplitudes a partir del cual se calcula $r_m^j(I_n^j)$, que representa la probabilidad que un $d_j[k]$ dado, para la ventana $W^j(m+1, L, \Delta)$, pertenezca al intervalo I_n^j. A partir de los conjuntos de probabilidades $P^j[m]$ y $R^j[m]$, la divergencia de Kullback–Leibler (ecuación 3.8) de dos ventanas consecutivas se calcula como:

$$\mathcal{D}_{\mathbf{d}}[j, m] = \sum_{n=1}^{N} p_m^j(I_n^j) \log\left(\frac{p_m^j(I_n^j)}{r_m^j(I_n^j)}\right). \tag{3.16}$$

Si se realiza este procedimiento sobre todas las escalas se obtiene la divergencia multiresolución continua **CMD**, donde $CMD(j, m) = \mathcal{D}_{\mathbf{d}}[j, m]$.

La elección de P y R a partir de ventanas consecutivas responde a la posibilidad de poder obtener una medida de la diferencia entre las distribuciones de probabilidad de las secuencias en ambos tramos. Diferencias en las distribuciones podrían indicar cambios en el modelo de generación de las señales.

De forma similar se puede calcular la matriz de divergencia multiresolución continua utilizando otras medidas de información relativas, como las descritas en las secciones 3.3.5 y 3.3.4. En el caso de la q–entropía relativa, la ecuación queda como:

$$\mathcal{D}^{q}{}_{\mathbf{d}}[j,m] = \frac{1}{1-q}\sum_{n=1}^{N} p_m^j(I_n)\left[1-\left(\frac{p_m^j(I_n^j)}{r_m^j(I_n^j)}\right)^{q-1}\right], \tag{3.17}$$

y la q–Divergencia Multiresolución Continua ($\mathbf{CMD}_q$) es $CMD_q(j,m) = D^q{}_{\mathbf{d}}[j,m]$. Para la divergencia de Jensen–Shannon se tiene:

$$\begin{aligned}\mathcal{D}^{JS}{}_{\mathbf{d}}[j,m] = \;& \mathcal{H}_{\mathbf{d}}(\pi_P\, P^j[m] + \pi_R\, R^j[m]) \\ & -\Big(\pi_P\, \mathcal{H}_{\mathbf{d}}(P^j[m]) + \\ & +\pi_R\, \mathcal{H}_{\mathbf{d}}(R^j[m])\Big).\end{aligned} \tag{3.18}$$

Y la divergencia Jensen–Shannon multiresolución Continua, $\mathbf{CMD}_{JS}$, se obtiene de forma similar a las anteriores.

3.5. Análisis de Componentes Principales

El análisis de componentes principales (PCA) es un método estadístico utilizado para el análisis de datos, extracción de características y compresión [195]. Dado los datos $\mathbf{H} = \{h[j,m]\} \in \mathbb{R}^{J\times M}$, se suponen que son generados por un modelo estadístico $\mathbf{\Phi}^{-1}$, con $\mathbf{\Phi} \in \mathbb{R}^{J\times J}$, a partir de una combinación lineal de fuentes ocultas no correlacionadas $\mathbf{Y} = \{y[j,m]\} \in \mathbb{R}^{J\times M}$:

$$\mathbf{H} = \mathbf{\Phi}^{-1}\mathbf{Y}. \tag{3.19}$$

El objetivo del PCA es obtener los componentes $y[j,m]$ que mejor describen los datos $h[j,m]$ en el sentido de la dirección de la máxima varianza. De esta manera, el PCA permite reducir la dimensionalidad de los datos disponibles, resaltando sus componentes más relevantes.

Para una fila fija j, la rv $\mathbf{y}_j = \{y[j,1],\ldots,y[j,M]\}$, obtenida como $\mathbf{y}_j = \phi_j\ \mathbf{H}$, con $\phi_j = \phi[j,1],\ldots,\phi[j,J]$, se denomina la j^n componente principal de $\mathbf{H}$ si:

$$\tilde{\phi}_j = \arg\max_i \mathcal{E}\left[(\phi_i\,\mathbf{H})^2\right], \tag{3.20}$$

donde $\mathcal{E}[\cdot]$ es el valor esperado de la variable correspondiente.

En la practica, esto se resuelve eligiendo las filas de $\mathbf{\Phi}$ como los autovectores de $\mathbf{H}\,\mathbf{H}^T$, suponiendo una distribución gaussiana para $\mathbf{Y}$. Esto diagonaliza la matriz de covarianza de $\mathbf{Y}$,

$$\tilde{\sigma}_{\mathbf{Y}} = \frac{1}{M-1}\mathbf{Y}\,\mathbf{Y}^T, \tag{3.21}$$

y conlleva una reducción de la redundancia, asegurando la independencia estadística de las fuentes $\mathbf{Y}$ para estadística de segundo órden. El término $(M-1)^{-1}$ es una constante de normalización para que la (3.21) sea una estimación no desviada de la covarianza de $\mathbf{Y}$.

La matriz de covarianza de $\mathbf{H}$ puede ser calculada usando también la ecuación (3.21) y satisface $\sigma_{\mathbf{H}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$, donde $\mathbf{Q}$ es la matriz de sus autovectores, $\mathbf{Q}^T \equiv \mathbf{\Phi}$

y $\mathbf{\Lambda} = diag\{\lambda_1, \ldots, \lambda_J\}$ es la matriz diagonal de autovalores asociados a $\mathbf{Q}$, con $\lambda_1 > \cdots > \lambda_J$. Así, la ecuación (3.19) puede escribirse como:

$$\mathbf{Y} = \mathbf{Q}^T \mathbf{H}^*, \tag{3.22}$$

donde $\mathbf{H}^*$ es la versión normalizada (media cero) de $\mathbf{H}$. Para resolver (3.22) se efectúan dos procedimientos: sustracción de la media de $\mathbf{H}$ y cálculo de los autovectores de $\mathbf{H}\mathbf{H}^T$. El *componente principal* (PC) corresponde a aquel que tiene el máximo autovalor.

3.6. Parametrización del habla basada en CME

El método del PCA se utiliza para extraer las características que han de integrar las parametrizaciones propuestas. Este procedimiento se explicará en detalle para la **CME** y puede, fácilmente, extenderse para la **CMD**.

Mediante $\mathbf{U} = \mathbf{CME}^*$ se denota la matriz estadísticamente normalizada asociada a **CME**. La normalización se obtiene escalado cada fila (escalas j) a media cero y varianza unitaria. Se define la matriz de correlación de $\sigma_{\mathbf{CME}} = \mathbf{U}\mathbf{U}^T$ y se obtiene la matriz de autovectores $\mathbf{Q}$ y la matriz diagonal de autovalores Λ asociados a $\mathbf{Q}$, tal que $\sigma_{\mathbf{CME}} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ y

$$\mathbf{Y} = \mathbf{Q}^T \mathbf{CME}^*. \tag{3.23}$$

Las filas de $\mathbf{Y}$ son las proyecciones de los datos **CME** en el sentido de la máxima variabilidad. El componente principal de $\mathbf{Y}$ es la fila $\mathbf{y}_1$ correspondiente al máximo valor de $\mathbf{\Lambda}$, que evoluciona sobre el eje temporal de m.

Como nueva parametrización, basada en la CME, se puede utilizar un conjunto de valores de $\mathbf{Y}$: $\mathbf{y}_i = \{y_i[m], m = 1, ..., M\}$, con $i = 1, .., \mathcal{J}$, cuya selección dependerá de la tarea que se quiere realizar.

Por ejemplo, una de las parametrizaciones propuestas en esta tesis utiliza las primeras ocho filas de $\mathbf{Y}$, asociadas con los ocho mayores valores de $\mathbf{\Lambda}$, por lo que $\mathbf{y}_i = \{y_i[m], m = 1, ..., M\}$, con $i = 1, .., 8$ ($\mathcal{J} = 8$ en este caso). Los elementos $y_i[m]$ de cada una de las componentes $\mathbf{y}_i$ son las nuevas características que representan la señal de habla en el intervalo m. A este conjunto de características, correspondientes a un intervalo m particular, los denotaremos como $\mathbf{y}^m$. El detalle de las codificaciones planteadas y las tareas realizadas se aborda en los capítulos siguientes.

El procedimiento para obtener la parametrización basada en CMD se realiza de manera similar, calculando los componentes principales usando ahora **CMD**. Nuevamente, se obtiene un conjunto de $\mathcal{J}$ características $\mathbf{y}^m$, para cada intervalo m de la señal de habla.

3.7. Comentarios de cierre del capítulo

En el presente capítulo se describió la forma de obtener nuevos coeficientes para representar la señal de habla a partir de la utilización de medidas de información aplicadas sobre la trasformada ondita continua de la señal. Como medidas de información se consideraron las entropías de Shannon y Tsallis, y sus correspondientes

divergencias, y la divergencia de Jensen-Shannon. Se utilizó el método del PCA para extraer las características más relevantes, que luego se utilizarán para proponer nuevas parametrizaciones de la señal de habla.

En el siguiente capítulo las nuevas codificaciones propuestas se utilizarán en el marco de un sistema de reconocimiento automático del habla. Los parámetros basados en la CME y la CMD, utilizando diferentes entropías y divergencias, se incorporarán a la codificación clásica MFCC de la señal de habla y se estudiará el comportamiento del reconocedor bajo distintas condiciones de ruido. Los detalles de los experimentos realizados se muestran en el capítulo 4.

También se propone la utilización de medidas de información multiresolución para realizar segmentación de fonemas. En este contexto, se utilizarán las parametrizaciones basadas en la CME y la CMD como entrada de un segmentador automático independiente del texto y se analizará el desempeño del mismo comparándolo contra la parametrización Melbank. Estos experimentos serán descriptos en el capítulo 5.

Capítulo 4

Análisis Multiresolución Aplicado al Reconocimiento del Habla

4.1. Introducción

En el capítulo anterior se describieron los procedimientos utilizados para obtener una nueva parametrización de la señal de habla a partir de la utilización de medidas de información multiresolución. Para ello se partió de una representación de la señal de habla mediante la CWT, sobre la cual se aplicaron diferentes medidas de información. Haciendo uso de la técnica de análisis de componentes principales se obtienen los coeficientes que conforman la nueva codificación.

En este capítulo se utilizarán las parametrizaciones basadas en CME y CMD para complementar la información brindada por los coeficientes de la parametrización clásica MFCC, a fin de utilizar dicha información en un sistema de reconocimiento automático de la señal de habla (ASR).

El ASR ha sido un activo campo de investigación durante las últimas décadas, dentro del cual se pueden mencionar tres grandes áreas: análisis de la señal de habla, acústica y modelado del lenguaje. Los métodos estadísticos de fuentes de Markov o modelos ocultos de Markov (HMM) han dado lugar a altos niveles de desempeño en ASR y en el campo del análisis de la señal de habla se han aplicado diferentes técnicas [149]. En aquellos experimentos donde se han tenido en cuenta las características de producción y percepción del habla humana, tales como el análisis cepstral y las codificaciones basadas en coeficientes MFCC y LPC, descriptos en las secciones 2.4.2, 2.4.4 y 2.4.5, respectivamente, se han obtenido muy buenos resultados [196].

Pero cuando un sistema de ASR se entrena con señales de habla limpia y luego se prueba con señales contaminadas con ruido se observa un deterioro importante en el desempeño del reconocedor, que puede conducir a incrementar los errores de reconocimiento en más del 80 % [197, 198]. En forma similar, cuando el sistema de ASR se entrena con señales registradas utilizando sistemas de audio de alta calidad y luego se prueba con señales registradas utilizando micrófonos caseros, los errores aumentan en un 50 % [197, 198]. Este es el alcance del reconocimiento "robusto" del habla, que se ha convertido en un campo de investigación importante.

El objetivo del reconocimiento robusto del habla es obtener sistemas de ASR que puedan ser utilizados en ambientes reales, con ruido, reverberación, pérdidas en los canales transmisión, sistemas de audio hogareños, etc. La investigación se orienta en dos áreas principales: (1) técnicas basadas en la transformación de la señal de

habla en el espacio de características (pre-procesamiento) y (2) la adaptación de los modelos al ruido o condiciones ambientales particulares [199].

Si comparamos el desempeño de las personas y las máquinas en términos de la percepción del habla, es evidente que los oyentes humanos son mucho más robustos que los sistemas artificiales de ASR [200]. Esto conduce a diferentes ideas acerca de sus similitudes y diferencias, que promueve la introducción selectiva de nuevos métodos en el dominio del ASR [201]. En este contexto, hay algunas investigaciones que se centran en la mejora de los sistemas ASR mediante el uso de modelos de procesamiento auditivo humanos para el análisis y procesamiento del habla [102] y se diseñan nuevos procedimientos de reducción de ruido, basados en conocimientos fisiológicos y psicoacústicos.

Hay diversos métodos de pre-procesamiento de la señal de habla para mejorar el rendimiento del sistema de ASR [197]. A menudo se supone que ambos, la señal y el ruido, son generados por sistemas lineales y que el ruido tiene características especiales que permiten que sea modelado fácilmente. En la práctica, ninguno de estos es un supuesto real y con frecuencia el ruido consiste en otras voces en una conversación o la señal es generada por un sistema que no es lineal. Por lo tanto, el problema de la robustez de los sistemas de ASR es todavía un tema de investigación "abierto", especialmente para tasas de relación señal-ruido (SNR) bajas.

Lo expuesto anteriormente ha motivado a explorar el uso de nociones de entropía como parte del pre-procesamiento de la señal de habla en un sistema de ASR. Como antecedentes, se puede mencionar que en [68] las entropías de Shannon y Tsallis, y sus correspondientes divergencias, han sido utilizadas en un sistema de ASR, suministrando información acerca de la evolución temporal del grado de complejidad de las señales de habla, mejorando su desempeño.

En esta tesis se propone incorporar características de la señal de habla obtenidas mediante el uso de la CME y la CMD, que provean información acerca de los cambios en la dinámica de la señal de habla para diferentes escalas, en un sistema de ASR como el que se describe en la sección 4.2. Las características propuestas se añadirán (concatenarán) a la parametrización clásica MFCC, descrita en la sección 2.4.4. El detalle de este procedimiento se describe en la sección 4.3.

El desempeño del sistema de ASR con estos nuevos parámetros se comparó con los resultados obtenidos con la parametrización clásica MFCC, bajo diferentes condiciones de ruido. Las señales utilizadas y los índices propuestos para cuantificar el desempeño del sistema de ASR se describen en la sección 4.4.

En la sección 4.5 se muestran los resultados obtenidos con el método propuesto y se discuten los mismos. Finalmente, en la sección 4.6 se presentan las conclusiones sobre el comportamiento de las nuevas parametrizaciones propuestas.

4.2. Sistema de reconocimiento automático del habla

En esta sección se mencionan las principales características de un sistema ASR basado en modelos ocultos de Markov (HMMs). Este tipo de sistema se utilizó para probar el desempeño de las parametrizaciones propuestas en la tarea de reconocimiento. Para mayores detalles sobre este tipo de sistemas ver [22, 202, 203, 204].

Los HMMs proporcionan un enfoque que permite modelizar secuencias de paráme-

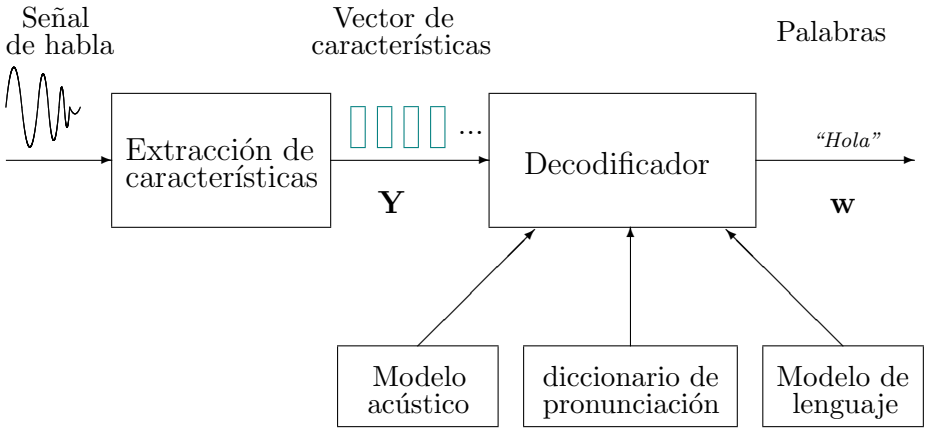

Figura 4.1: Diagrama de las etapas que comprende un sistema de reconocimiento automático del habla basado en modelos ocultos de Markov.

tros espectrales variantes en el tiempo [202]. Debido a que el habla presenta variaciones temporales y, además, puede ser codificada mediante secuencias de características espectrales, los HMMs ofrecen un marco de trabajo que se adapta muy bien para la construcción de estos modelos. Como consecuencia, actualmente la mayoría de los sistemas de ASR se implementan en base a HMMs [22, 203].

Considerando que los principios básicos que subyacen los sistemas de ASR basados en HMM son bastante sencillos, las aproximaciones y suposiciones de simplificación que intervienen para la ejecución directa de los mismos darían lugar a un sistema con baja precisión y sensibilidad. Por lo tanto, la aplicación práctica de los HMMs en los sistemas modernos implica una considerable serie de refinamientos [203].

Los componentes principales de un sistema de ASR basado en HMMs se muestran en la figura 4.1, donde se observa que, a partir del las caraterísticas de entrada (extraídas de la señal de habla mediante métodos convencionales o los propuestos en esta tesis), el decodificador encuentra la secuencia de palabras más probable que haya sido generada por la señal acústica. Para poder llevar a cabo esta tarea, el decodificador utiliza modelos probabilísticos acústicos y del lenguaje, así como diccionarios.

Una vez que la señal de audio es captada por el micrófono se parametriza, convirtiéndola en una secuencia $\mathbf{Y} = \{\mathbf{y}^m, m = 1, ..., M\}$ de características acústicas, donde $\mathbf{y}^m$ es el vector de características correspondiente a la ventana m de análisis. La dimensión de cada vector de características es fija y se elige de forma tal de rescatar los aspectos más destacados de la señal pero con la menor cantidad de características. A partir de $\mathbf{Y}$ el decodificador intenta encontrar la secuencia de palabras $\mathbf{w} = \{w_l, l = 1, ..., L\}$, tal que:

$$\hat{\mathbf{w}} = \underset{w}{\text{argmáx}} \{P(\mathbf{w}|\mathbf{Y})\}. \tag{4.1}$$

Sin embargo, dada la dificultad de modelar $P(\mathbf{w}|\mathbf{Y})$ directamente, se utiliza la regla de Bayes para transformar (4.1) en el problema equivalente de encontrar:

$$\hat{\mathbf{w}} = \underset{w}{\text{argmáx}} \{p(\mathbf{Y}|\mathbf{w})P(\mathbf{w})\}. \tag{4.2}$$

La verosimilitud $p(\mathbf{Y}|\mathbf{w})$ se determina mediante un *modelo acústico* y la probabilidad *a priori* de $P(\mathbf{w})$ se determina mediante un *modelo de lenguaje*.

La unidad básica de sonido representado por el modelo acústico es el fonema (ver sección 2.2.1). Para cualquier $\mathbf{w}$ dado, el correspondiente modelo acústico se sintetiza concatenando modelos fonéticos para generar palabras, definidas a partir de un *diccionario de pronunciación*. Los parámetros de estos modelos fonéticos se estiman a partir de datos de entrenamiento que consisten de señales habla y sus correspondientes transcripciones ortográficas.

El modelo de lenguaje es típicamente un modelo de N-gramas en el cual la probabilidad de cada palabra está condicionada sólo por sus $N-1$ predecesoras. Los parámetros del modelo de N-gramas se estiman a partir de N-tuplas de corpus de texto apropiados.

El decodificador opera buscando a través de todas las posibles secuencias de palabras utilizando un procedimiento de poda para eliminar hipótesis improbables, manteniendo así una búsqueda manejable [204]. Cuando se alcanza el final de la sentencia, el sistema obtiene la secuencia más probable de palabras como salida.

Para obtener el modelo acústico basado en HMMs (componentes básicos), las emisiones de cada palabra w se descomponen en una secuencia de K_w sonidos básicos llamados *fonemas base*. Esta secuencia se denomina pronunciación: $q^{(w)} = \{q_1, ..., q_{Kw}\}$. A fin de permitir la posibilidad de múltiples pronunciaciones, se calcula la verosimilitud $p(\mathbf{Y}|\mathbf{w})$ sobre cada una de ellas:

$$p(\mathbf{Y}|\mathbf{w}) = \sum_{\mathbf{Q}} p(\mathbf{Y}|\mathbf{Q})P(\mathbf{Q}|\mathbf{w}), \tag{4.3}$$

y se realiza la sumatoria sobre todas las secuencias válidas de pronunciación de w. $\mathbf{Q}$ es una secuencia de pronunciación particular

$$P(\mathbf{Q}|\mathbf{w}) = \prod_{l=1}^{L} P(\mathbf{q}^{w_l}|w_l), \tag{4.4}$$

donde cada $\mathbf{q}^{w_l}$ es una pronunciación válida de la palabra w_l. En la práctica, sólo habrá un número muy pequeño de pronunciaciones alternativas para cada w_l, de forma de simplificar la suma (4.3).

Cada fonema base q se representa por un HMM de densidad de probabilidad continua, de la forma que se ilustra en la figura 4.2, con parámetros de probabilidad de transiciones $\{a_{ij}\}$ y distribución de observaciones de salida $\{b_j()\}$. En cada paso de tiempo, un HMM hace una transición desde su estado actual a uno de sus estados conectados. La probabilidad de realizar una transición particular desde el estado $\mathbf{s}_i$ al estado $\mathbf{s}_j$ está dada por la probabilidad de transición $\{a_{ij}\}$. Al entrar en un estado, se genera como salida un vector de características utilizando la distribución $\{b_j()\}$, asociada con el estado en que se ingresa. Esta forma de proceder proporciona los supuestos de independencia condicional estándar para un HMM: los estados son condicionalmente independientes de los otros estados, dado el estado previo; las observaciones son condicionalmente independientes de todas las demás observaciones y de todos los estados, excepto el estado que las generó. Para una discusión más detallada de la operación de un HMM ver [202].

La distribución de salida $b_j()$ se modelará a partir de Gaussianas multivariantes individuales:

$$b_j(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mu^{(j)}, \mathbf{\Sigma}^{(j)}), \tag{4.5}$$

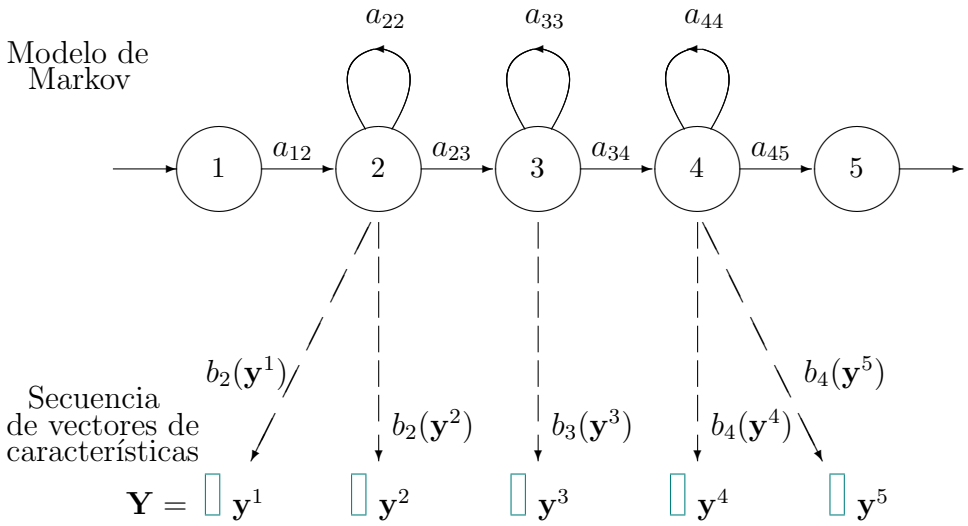

Figura 4.2: Diagrama del modelo del fonema base q, representado por un HMM de densidad continua con parámetros de probabilidad de transiciones $\{a_{ij}\}$ y distribución de observaciones de salida $\{b_j()\}$.

donde $\mu^{(j)}$ es la media del estado $\mathbf{s}_j$ y $\mathbf{\Sigma}^{(j)}$ su covarianza. Debido a que la dimensionalidad del vector de características $\mathbf{y}$ es relativamente alto, las covarianzas normalmente se restringen a una matriz diagonal. El uso de mezcla de Gaussianas proporciona una distribución altamente flexible, capaz de modelar la distribución de datos asimétricos y multi-modales, como los que se generan, por ejemplo, por diferencias entre hablantes, de acentos, de género, etc.

Dado el HMM compuesto de $\mathbf{Q}$, formado mediante la concatenación de todos los fonemas de base $\mathbf{q}^{(w1)}, ..., \mathbf{q}^{(wL)}$, la verosimilitud acústica queda determinada por:

$$p(\mathbf{Y}|\mathbf{Q}) = \sum_{\theta} p(\theta, \mathbf{Y}|\mathbf{Q}), \tag{4.6}$$

donde $\theta = \theta_0, ..., \theta_{M+1}$ es una secuencia de estados a partir del modelo compuesto y

$$p(\theta, \mathbf{Y}|\mathbf{Q}) = a_{\theta_0\theta_1} \prod_{m=1}^{M} b_{\theta_m}(\mathbf{y}^m) a_{\theta_m\theta_{m+1}}. \tag{4.7}$$

En la ecuación (4.7), θ_0 y θ_{M+1} corresponden a los estados de entrada y salida, que no emiten, de la figura 4.2. Estos últimos se incluyen para simplificar el proceso de concatenar los modelos de fonemas para formar palabras. Por simplicidad, en lo que sigue, los estados que no emiten serán ignorados y se focalizará sólo en la secuencia de estados $\theta_1, ..., \theta_M$.

Los parámetros del modelo acústico $\lambda = [\{a_{ij}\}, \{b_j()\}]$ pueden ser eficientemente estimados a partir de un corpus de habla de entrenamiento, usando el algoritmo de avance-retroceso [205], el cual es un tipo de algoritmo de maximización de la esperanza (EM). Para cada emisión $\mathbf{Y}^{(r)}, r = 1, ..., R$, de longitud $M^{(r)}$, se encuentra la secuencia de modelos base (los HMMs que corresponden a la secuencia de palabras de la emisión) y se construye el correspondiente HMM compuesto.

Para una dada emisión, en la primera etapa del algoritmo EM se obtienen, recursivamente, las probabilidades hacia adelante $\alpha_m^j = p(\mathbf{Y}_{1:m}, \theta_m = \mathbf{s}_j; \lambda)$ y hacia

atrás $\beta_m^i = p(\mathbf{Y}_{m+1:M}|\theta_m = \mathbf{s}_i; \lambda)$ de la secuencia de observaciones. A partir de esto se calcula la probabilidad de que el modelo esté en el estado $\mathbf{s}_j$ en el intervalo m para una dada emisión r, que como puede demostrarse es igual a

$$\gamma_m^j = P(\theta_m = \mathbf{s}_j|\mathbf{Y}; \lambda) = \frac{1}{p(\mathbf{Y}; \lambda)} \alpha_m^j \beta_m^j. \tag{4.8}$$

Estas probabilidades de ocupación de estados representan una alineación suave de los estados del modelo a los datos, a partir de donde se calcula el nuevo conjunto de parámetros Gaussianos $\hat{\mu}^{(j)}$ y $\hat{\mathbf{\Sigma}}^{(j)}$, así como las probabilidades de transición $\hat{a}_{ij}$, que maximizan la verosimilitud de los datos dados estos alineamientos. Partiendo entonces de una estimación inicial de los parámetros $\lambda^{(0)}$, se realizan sucesivas iteraciones del algoritmo EM para obtener el conjunto $\lambda^{(1)}$, $\lambda^{(2)}$,..., que garantizan mejorar la verosimilitud en un cierto máximo local. Una elección frecuente para los parámetros iniciales $\lambda^{(0)}$ es asignar la media y covarianza globales de los datos a las distribuciones Gaussianas de salida e igualar todas las probabilidades de transición. Esto se denomina el modelo de inicio plano (*flat start model*). Los detalles acerca de cómo se llevan a cabo los procedimientos antes descritos se pueden consultar en [203, 204].

La probabilidad *a priori* de una secuencia de palabras $\mathbf{w} = w_1, ..., w_K$, requerida en la ecuación (4.2), esta dada por:

$$P(\mathbf{w}) = \prod_{k=1}^{K} P(w_k|w_{k-1}, ..., w_1). \tag{4.9}$$

Para el reconocimiento con vocabularios extensos, el condicionamiento del historial de palabras de (4.10) se trunca a $N-1$ palabras para formar un modelo de lenguaje de N-gramas. N típicamente se encuentra en el rango de 2 a 4. Los modelos de lenguaje son evaluados en términos de su perplejidad, que mide el número medio de palabras que siguen a otra determinada. En otras palabras, la perplejidad tiene en cuenta el grado de restricciones que introduce el modelo de lenguaje en el módulo de reconocimiento acústico, limitando en cada momento el número de modelos o patrones con los que comparar la señal acústica de entrada.

Las probabilidades de los N-gramas se estiman a partir de textos de entrenamiento, contando N-grama ocurrencias para formar estimadores de máxima verosimilitud de parámetros (ML). Por ejemplo, si $C(w_{k-2}w_{k-1}w_k)$ representa el número de ocurrencias de tres palabras w_{k-2} w_{k-1} y w_k y de forma similar para $C(w_{k-2}w_{k-1})$, entonces

$$P(w_k|w_{k-1}, w_{k-2}) \approx \frac{C(w_{k-2}w_{k-1}w_k)}{C(w_{k-2}w_{k-1})}. \tag{4.10}$$

El principal problema con este esquema simple de estimación ML es la raleza de los datos, lo cual puede ser mitigado mediante técnicas de suavizado (veáse [206, 207, 208]).

Una vez definidos los modelos acústicos y del lenguaje, la secuencia de palabras $\hat{\mathbf{w}}$ más probable a partir de los vectores de características $\mathbf{Y}$ se encuentra buscando todas las posibles secuencias de estados que surgen de todas las posibles secuencias de palabras, para la secuencia que es más probable que haya generado los datos observados $\mathbf{Y}$. Una forma eficiente de resolver este problema es usando el algoritmo de Viterbi, que es una aplicación de la programación dinámica. Sea

$\phi_m^{(j)} = \max_\theta\{p(\mathbf{Y}, \theta_m = \mathbf{s}_j; \lambda)\}$ la máxima probabilidad de observar la secuencia parcial $\mathbf{Y}_{1:m}$ y estar luego en el estado $\mathbf{s}_j$ en el intervalo m, dado el modelo de parámetros λ. Esta probabilidad puede ser eficientemente calculada usando el algoritmo de Viterbi

$$\phi_m^{(j)} = \max_i\{\phi_{m-1}^{(i)} a_{ij}\} b_j(\mathbf{y}^m). \tag{4.11}$$

La ecuación (4.11) es inicializada en $\phi_0^{(j)}$ igual a 1 para el estado de entrada, no emisor, y 0 para todos los otros estados. La probabilidad de la secuencia de palabras más posible está dada por $\max_j\{\phi_M^{(j)}\}$ y registrando cada decisión de maximización se tiene un trayecto que brindará la secuencia estado/palabra que mejor coincide con la requerida. En la práctica, la implementación directa del algoritmo de Viterbi se vuelve compleja para habla continua, donde deben tenerse en cuenta la topología de los modelos, las restricciones del modelo de lenguaje y la necesidad de integrar el cálculo. Para afrontar esto, han evolucionado diferentes enfoques. Para decodificación basada en el algoritmo de Viterbi, el espacio de búsqueda se puede restringir, manteniendo múltiples hipótesis en paralelo [209, 210] o se puede expandir dinámicamente a medida que la búsqueda progresa [211, 212, 213]. Alternativamente, se han propuesto otros tipos de enfoques, los cuales pueden consultarse en [203, 214, 215, 216].

Para esta tesis se utilizó un HMMs de tres estados para modelar fonemas independientes del contexto y silencios [22]. Las funciones de densidad de probabilidades para las observaciones se modelaron con mezclas de Gaussianas. Se construyó un modelo completo de todas las frases del corpus y se llevaron a cabo cuatro re-estimaciones de parámetros, usando el algoritmo de Baum-Welch, como algoritmo EM [214]. El enlazado de parámetros se realizó usando un conjunto de 200 Gaussianas para cada estado del modelo. En el mismo modelo de fonemas, las mezclas enlazadas reducen la cantidad total efectiva de parámetros de 855000 a 26200. Esta etapa es necesaria a fin de mejorar la robustez de la estimación debido al reducido conjunto de entrenamiento utilizado. Finalmente, se obtienen las re-estimaciones restantes, hasta completar un total de dieciséis. Para el modelado del lenguaje se estimaron bigramas suavizados mediante la técnica de *backing-off*[1], a partir de transcripciones de la base de datos de entrenamiento [204].

4.3. Nuevos parámetros basados en la CME y CMD

En esta sección se describe la forma en que se obtuvieron los coeficientes que luego se utilizarán para modificar la etapa de pre-procesamiento clásica de la señal de habla, como así también los supuestos sobre los cuales se plantearon las diferentes propuestas. Estos coeficientes basados en la CME y CMD fueron concatenados como una nueva dimensión en la parametrización MFCC de la señal de habla.

En el diagrama en bloques que se muestra en la figura 4.3 se representa cada etapa del algoritmo propuesto en esta tesis, para esta aplicación. A partir de la señal de voz muestreada, se abren dos ramas de procesamiento. En una de ellas se calcula

[1]El objetivo de la técnica de suavizado *backing-off* es mejorar las estimaciones cuando la cantidad total de bigramas en un texto de entrenamiento no es suficiente y, por lo tanto, no posible obtener el estimador de la probabilidad del bigrama como el cociente entre la frecuencia de ocurrencia y el número total de bigramas en un texto de entrenamiento. Para ello tiene en cuenta que si el bigrama no se observa frecuentemente en el texto de entrenamiento, entonces se utiliza una probabilidad basada en el conteo de la ocurrencia dentro de un contexto más pequeño.

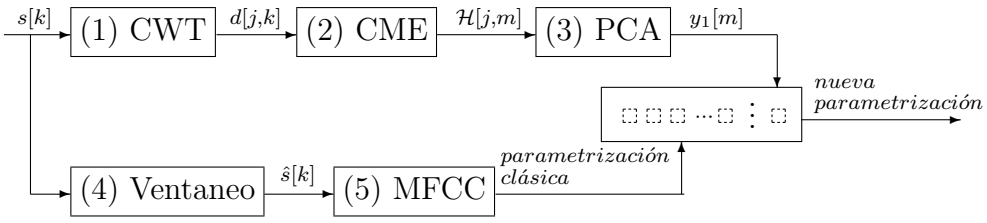

Figura 4.3: Diagrama de las etapas del método propuesto. Ejemplo para una característica correspondiente al método 1 (PC_1), que se describe en la subsección 4.3.1.

la CME (Fig. 4.3: pasos 1–2). En la otra rama se obtiene la parametrización clásica MFCC (figura 4.3: pasos 4–5), de acuerdo a lo descrito en la sección 2.4.4. Esto se lleva a cabo para cada una de las frases del corpus.

A fin de calcular la CME, se evalúan las medidas de información para cada escala de la matriz CWT. La CWT se obtiene utilizando la ondita Daubechies de orden 16. Como medidas de información se utilizan diferentes entropías y divergencias: entropías de Shannon y Tsallis, con sus correspondientes entropías relativas y la divergencia de Jensen-Shannon. La CME y la CMD se obtienen mediante el procedimiento que se explicó en la sección 3.4 del capítulo 3 y se utiliza el procedimiento de PCA (seción 3.5) para extraer los componentes temporales de mayor varianza (figura 4.3, paso 3). Los valores obtenidos se añaden (concatenan) como nuevas dimensiones a la parametrización MFCC. El desempeño de esta nueva propuesta se comparará con al análisis clásico bajo diferentes condiciones de ruido.

A modo de ejemplo, en la figura 4.4 se muestra el comportamiento de una de las divergencias multirresolución cuando se aplica a una señal de voz con y sin ruido. En la figura 4.4(a) se muestra una parte de la señal de habla etiquetada de la sentencia: "¿Cómo se llama el mar que baña Valencia?". La figura 4.4(b) muestra el escalograma ($|d[j,k]|^2$) correspondiente a la señal presentada en (a), obtenida con la ondita Daubechies de orden 16. En la figura 4.4(c) se muestra la correspondiente $\mathbf{CMD}_q$, para $q = 0{,}2$. Las figuras 4.4(d), 4.4(e) y 4.4(f) muestran los resultados obtenidos para la misma señal, pero corrompida con ruido aditivo de conversación de fondo a 10dB SNR. De las figuras 4.4(c) y (f) se puede observar que, en este caso, $\mathbf{CMD}_q$ tiene valores más altos en los puntos marcados como transiciones de un fonema a otro, tanto en la señal limpia como en la corrrompida. Este resultado sugiere que una apropiada inclusión en el modelo de la información provista por estas herramientas podría mejorar el desempeño del sistema de ASR en presencia de ruido, haciéndolo más robusto.

A continuación, se describe cómo obtener las características que luego se concatenarán a la parametrización MFCC. El análisis de componentes principales (ver sección 3.5), que se lleva a cabo con el fin de disminuir la dimensión relativa en el vector final de coeficientes, se utiliza de tres manera diferentes. Estas tres propuestas se describen a continuación, bajo la denominación de método 1, método 2 y método 3, utilizando para ello la matriz **CME**. Cabe destacar que el mismo procedimiento se lleva a cabo utilizando la matriz **CMD**.

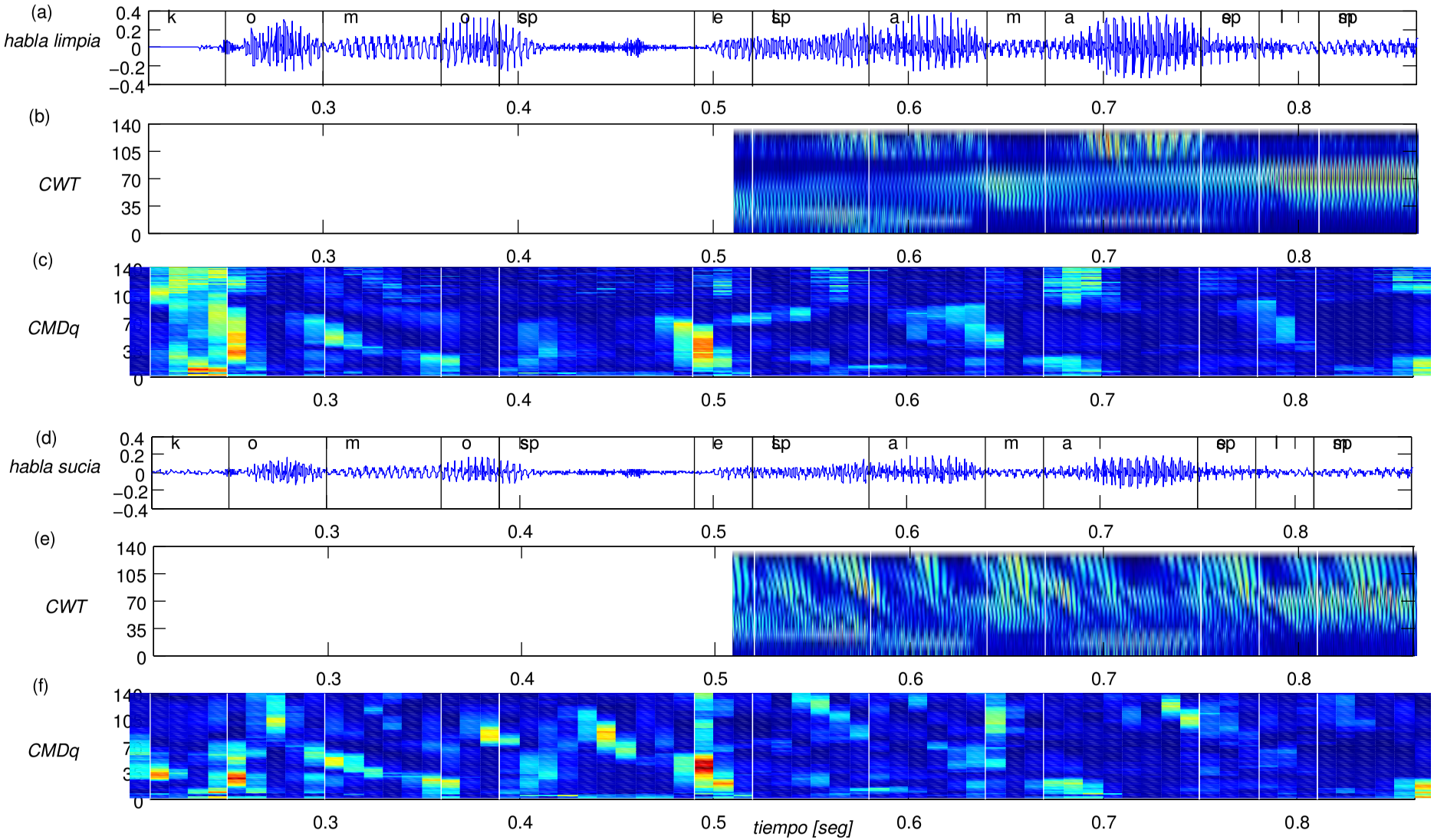

Figura 4.4: (a) Señal de habla etiquetada. (b) Escalograma correspondiente a la señal mostrada en (a). (c) $\mathbf{CMD}_q$ ($q = 0{,}2$) del escalograma mostrado en (b). (d) La misma señal que se muestra en (a) con ruido aditivo de murmullo (10 dB SNR). (e) Escalograma correspondiente a la señal mostrada en (d). (f) $\mathbf{CMD}_q$ ($q = 0{,}2$) del escalograma mostrado en (e).

4.3.1. Método 1: Primer PC ($\mathbf{PC}_1$)

Este método utiliza un componente principal, que se obtiene a partir de la matriz de componentes principales:

$$\mathbf{Y} = \mathbf{Q}^T \mathbf{CME}^*, \tag{4.12}$$

donde $\mathbf{CME}^*$ es la matriz estadísticamente normalizada asociada con $\mathbf{CME}$ y $\mathbf{Q}$ los eigenvectores de la matriz de correlación de datos $\sigma_{\mathbf{CME}} = \mathbf{U}\mathbf{U}^T$, con $\mathbf{U} = \mathbf{CME}^*$ (véase sección 3.6 del capítulo 3, para mayores detalles).

Para ello, se obtiene el vector fila de $\mathbf{Y}$ correspondiente al máximo valor de $\mathbf{\Lambda}$, matriz diagonal de los eigenvalores asociados con $\mathbf{Q}$. Este vector es el componente principal, que se denota como $\mathbf{y}_1$. El componente $y_1[m]$ evoluciona a lo largo de la variable m, correspondiente al eje temporal y se añade (concatena) a la MFCC clásica para obtener una nueva parametrización.

4.3.2. Método 2: Primer y segundo PC ($\mathbf{PC}_{12}$)

En el método 1 se obtiene el vector $\mathbf{y}_1$ a partir de $\mathbf{Y}$. En este método se obtiene también el segundo componente de $\mathbf{Y}$, asociado con el segundo mayor valor de $\mathbf{\Lambda}$, el vector $\mathbf{y}_2$.

Así, ambos elementos $y_1[m]$ y $y_2[m]$ se concatenan a la MFCC para generar el nuevo vector para la parametrización.

El motivo de esta propuesta se debe a las características que mostró la CMD de la señal con ruido en la figura 4.4(b). Comparando dicha figura con la CMD de la señal limpia, 4.4(a), se puede observar que aparecen nuevas estructuras, las cuales pueden relacionarse con el ruido.

4.3.3. Método 3: PC dependiente de la escala ($\mathbf{PC}_{SD}$)

Para este método se consideran dos submatrices para aplicar PCA, una correspondiente a las escalas inferiores de $\mathbf{CME}$ y otra a las escalas superiores:

$$\begin{aligned} \mathbf{U}^{(1)} &= (CME(j\delta, m))^*, \quad \text{donde } 1 \leq j \leq J/2, \text{ y} \\ \mathbf{U}^{(2)} &= (CME(j\delta, m))^*, \quad \text{con } J/2 < j \leq J, \end{aligned}$$

sobre las que se calculan ambas matrices de correlación: ${\sigma_{\mathbf{CME}}}^{(i)} = \mathbf{U}^{(i)}(\mathbf{U}^{(i)})^T$, para $i = 1, 2$.

Las columnas de $\mathbf{Q}^{(i)}$ contienen los eigenvectores de ${\sigma_{\mathbf{CME}}}^{(i)}$ y la matriz diagonal $\mathbf{\Lambda}^{(i)}$ provee los eigenvalores asociados, para $i = 1, 2$. Se obtienen, así, dos expresiones equivalentes a (4.12), pero sobre cada mitad en la que se subdividió la matriz $\mathbf{CME}$. De esta manera, se tiene:

$$\mathbf{Y}^{(i)} = (\mathbf{Q}^{(i)})^T \, \mathbf{U}^{(i)}, \text{para } \; i = 1, 2. \tag{4.13}$$

Sobre cada una de estas ecuaciones se obtienen los correspondientes componentes principales $\mathbf{y}_1^{(1)}$ y $\mathbf{y}_1^{(2)}$. Ambos PCs se concatenan a la parametrización MFCC clásica.

Bajo las mismas ideas consideradas para el método 2, a partir de 4.4(b) se observa que las estructuras relacionadas con el ruido aparecen principalmente en escalas altas, mientras que las correspondientes a características propias de la señal se manifiestan en las escalas inferiores. Esto sugiere que los dos componentes, obtenidos

como se menciona anteriormente, podrían proporcionar información acerca de la señal y del ruido de forma relativamente separada.

Los tres métodos indicados se llevan a cabo de manera similar utilizando las otras medidas multirresolución. Y mediante estos nuevos coeficientes se incorpora información sobre los cambios dinámicos de la señal de voz en el plano tiempo–escala.

4.4. Aspectos Principales de la Implementación

En esta sección se describen los principales aspectos que se tuvieron en cuenta para el entrenamiento y prueba del sistema de ASR utilizado representaciones del habla basadas en medidas de información multiresolución.

En la subsección 4.4.1 se detallan las características del corpus de habla utilizado y cómo se implementaron los esquemas de codificación basados en la CME y la CMD, descriptos en 4.3. El desempeño del sistema de ASR usando las representaciones propuestas se comparó con los resultados alcanzados mediante una parametrización clásica MFCC, descripta en 2.4.4. En la subsección 4.4.4 se indican los índices calculados para evaluar el sistema.

4.4.1. Señales y base de datos

Las señales utilizadas para el entrenamiento y prueba del sistema de ASR se obtuvieron a partir de un subconjunto del corpus de habla en español Albayzin [148]. Este subconjunto consiste de 600 sentencias, con un vocabulario de 200 palabras, relacionadas con geografía española. Cada frase tiene una duración promedio de 3.35 segundos y fueron pronunciadas por seis hombres y seis mujeres de la zona central de España, cuyo promedio de edad es de 31.8 años. Las señales fueron registradas en un estudio de grabación y han sido re–muestreadas a 8 kHz con 16 bits de resolución.

Para probar la robustez del sistema de ASR se utilizaron señales de habla que fueron corrompidas con ruido blanco y ruido de murmullo, mediante el uso de la base de datos NOISEX-92 [217]. El ruido blanco fue digitalizado a partir de un generador de ruido analógico de alta calidad. La fuente de ruido de murmullo fue la conversación de fondo de 100 personas hablando en un bar. Ambos tipos de ruidos fueron re–muestreados a 8 kHz y mezclados aditivamente con las señales de habla limpias a diferentes niveles de SNR.

4.4.2. Codificación MFCC

Para comparar el desempeño del sistema de ASR con las nuevas medidas de información, se utilizó como vector de características de referencia la codificación MFCC, descripta en la sección 2.4.4, ampliamente utilizada en este tipo de aplicaciones [218].

Para ello, cada frase se ha normalizado en media, pre-enfatizado y segmentado mediante una ventana de Hamming en tramos de 25 ms de longitud, con un desplazamiento de 10 ms y 24 canales para el banco de filtros. Para cada segmento se obtienen 13 coeficientes espectrales y su energía[102].

Además, a fin de compensar el supuesto de independencia condicional planteado por los modelos acústicos basados en HMM, se añaden al vector de características

los coeficientes de velocidad de primer orden (Δ) de los MFCC y la energía [219]. Los coeficientes Δ se obtienen de la siguiente manera:

$$\Delta \mathbf{y}^m = \frac{\sum_{i=1}^{N} c_i \{\mathbf{y}^{m+i} - \mathbf{y}^{m-i}\}}{2 \sum_{i=1}^{N} c_i^2}, \tag{4.14}$$

donde $\mathbf{y}^m$ el vector original de características correspondiente a la ventana m, $m = 1, ..., M$ las ventanas de análisis, N el ancho en ventanas utilizado para el cómputo y c_i los coeficientes de regresión[2]. Para este trabajo se utilizó $N = 2$.

Mediante este procedimiento, se obtiene una parametrización de 28 coeficientes para cada uno de los segmentos de la señal de habla, compuesto por 13 MFCC, 1 coeficiente de energía (E) y sus respectivas derivadas temporales ($\Delta MFCC$ y ΔE). Esta fue la parametrización utilizada como referencia.

4.4.3. Codificaciones evaluadas

Para cada uno de los métodos explicados en la sección 4.3 se generaron nuevos vectores de caracterísiticas que fueron concatenados a la parametrización clásica, descripta en la sección anterior, a fin de incorporar información acerca de los cambios en la dinámica de la señal de voz en el plano tiempo–escala. De esta manera, se lograron las siguientes codificaciones:

- Método PC_1: 12 MFCC, lo cual permite mantener el número de coeficientes igual al de la parametrización de referencia, 1 coeficiente de energía y el coeficiente obtenido a partir del PCA sobre cada segmento, $\mathbf{y}_1$. Se obtienen, además, las correspondientes derivadas temporales. Así, esta propuesta queda como:

$$[\, MFCC_{1,..,12} \mid E \mid \mathbf{y}_1 \mid \Delta MFCC_{1,..,12} \mid \Delta E \mid \Delta \mathbf{y}_1 \,]. \tag{4.15}$$

- Método PC_{12}: 11 MFCC, el coeficiente de energía y ambos coeficientes basados en medidas de información, con sus correspondientes derivadas. Esto es:

$$[\, MFCC_{1,..,11} \mid E \mid \mathbf{y}_1 \mid \mathbf{y}_2 \mid \Delta MFCC_{1,..,11} \mid \Delta E \mid \Delta \mathbf{y}_1 \mid \Delta \mathbf{y}_2 \,]. \tag{4.16}$$

- Método PC_{SD}: 11 MFCC, el coeficiente de energía, los dos valores correspondientes a las medidas de información en las escalas bajas y altas y sus derivadas temporales:

$$\left[MFCC_{1,..,11} \mid E \mid \mathbf{y}_1^{(1)} \mid \mathbf{y}_1^{(2)} \mid \Delta MFCC_{1,..,11} \mid \Delta E \mid \Delta \mathbf{y}_1^{(1)} \mid \Delta \mathbf{y}_1^{(2)} \right]. \tag{4.17}$$

Debe tenerse presente que la barra "|" se utiliza aquí para indicar la concatenación de los diferentes vectores, cuyos elementos se utilizan para conformar los nuevos patrones.

[2]Para asegurar que se mantiene la cantidad de ventanas de análisis luego de concatenar los coeficientes delta, los elementos iniciales y finales se replican para completar la ventana de regresión.

4.4.4. Índices para evaluar el desempeño del reconocimiento

A fin de evaluar el desempeño de las parametrizaciones propuestas en esta tesis, se ha utilizado un método de validación cruzada de k iteraciones [220]. Para ello, se implementaron y entrenaron diez modelos. Por cada modelo, se seleccionaron aleatoriamente diferentes particiones para entrenamiento y para prueba. Para el entrenamiento del sistema, cada partición se elaboró utilizando 80 % del subconjunto seleccionado del corpus Albayzin, y el restante 20 % se utilizó para realizar la prueba.

El reconocimiento se evaluó mediante la tasa de error de palabra (WER), considerando como errores los borrados y las sustituciones de palabras [204]. A fin de resaltar las diferencias entre las parametrizaciones propuestas, en comparación con la de referencia, se ha calculado el porcentaje de mejora del error relativo como:

$$\Delta\varepsilon_{\,\%} = \frac{\varepsilon_{ref} - \varepsilon}{\varepsilon_{ref}} 100, \tag{4.18}$$

donde ε es el valor de WER correspondiente a la propuesta y ε_{ref} es el WER de la referencia.

4.5. Resultados y discusión

A continuación se mostrarán y discutirán los resultados obtenidos al utilizar los métodos propuestos en un sistema de ASR entrenado con habla limpia y probado con señales de habla corrompidas con ruido de murmullo y ruido blanco. Estos métodos se compararán con los resultados obtenidos usando la parametrización clásica MFCC.

En la figura 4.5 se compara el WER obtenido con la parametrización clásica y los métodos propuestos en esta tesis, para diferentes SNR, utilizando ruido de murmullo sobre las señales. La figura 4.5(a) muestra el WER obtenido con los métodos PC_1, PC_{12} y PC_{SD}, cuando se concatenan coeficientes basados en CME usando entropía de Shannon al vector de MFCC. En la figura 4.5(b) se utiliza la CME con q–entropía. Trabajos previos sugieren como valor óptimo para este tipo de experimentos $q = 0{,}2$ [68]. Las figuras 4.5(c), 4.5(d) y 4.5(e) muestran el WER obtenido usando divergencia de Kullback-Leibler, q–divergencia y divergencia de Jensen–Shannon, respectivamente.

A partir de las figuras, se puede observar que el método PC_{SD} es el que presenta los mejores resultados, puesto que su curva de WER se encuentra, en la mayoría de los casos, por debajo del resto de las curvas. Cuando se utilizan entropías de Shannon y Tsallis, este método tuvo un comportamiento similar a la parametrización de referencia, mejorando el reconocimiento cuando la SNR es de 15 dB, pero con menor desempeño para SNRs por debajo de 10 dB.

La utilización de medidas de información relativas para obtener la CME mejoraron notablemente los resultados. Esto se observa en las figuras (c), (d) y (e). En particular, las curvas correspondientes al método PC_{12} presentaron comportamientos que mejoraron el WER obtenido con la referencia para SNRs menores a 15 dB. Nuevamente, el método PC_{SD} presentó el mejor desempeño. Para el caso (c), de la divergencia de Kullback-Leibler, se observa que, cuando se aplica la parametrización del método PC_{SD}, el error de reconocimiento de palabras está por debajo del correspondiente a la referencia para SNRs menores e iguales a 15 dB.

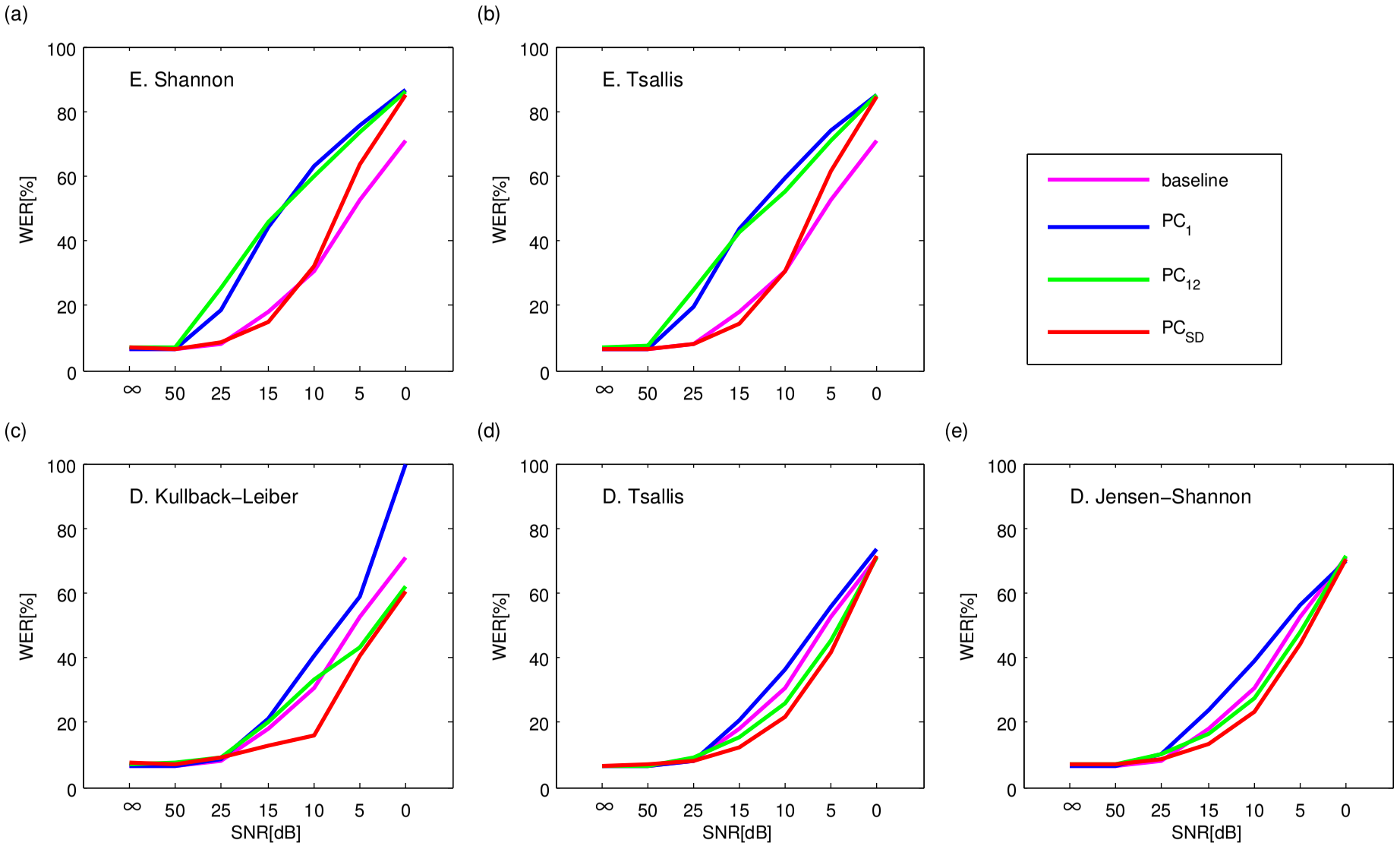

Figura 4.5: Tasa de error de palabra del sistema de ASR *vs.* SNR usando señales corrompidas con ruido aditivo de murmullo. Comparación entre el pre–procesamiento clásico (*baseline*) y los métodos propuestos: PC_1, PC_{12}, PC_{SD}, obtenidos con (a) entropía Shannon, (b) q–entropía, con $q = 0{,}2$, (c) divergencia de Kullback-Leibler, (d) q–divergencia, con $q = 0{,}2$, y (e) divergencia de Jensen–Shannon.

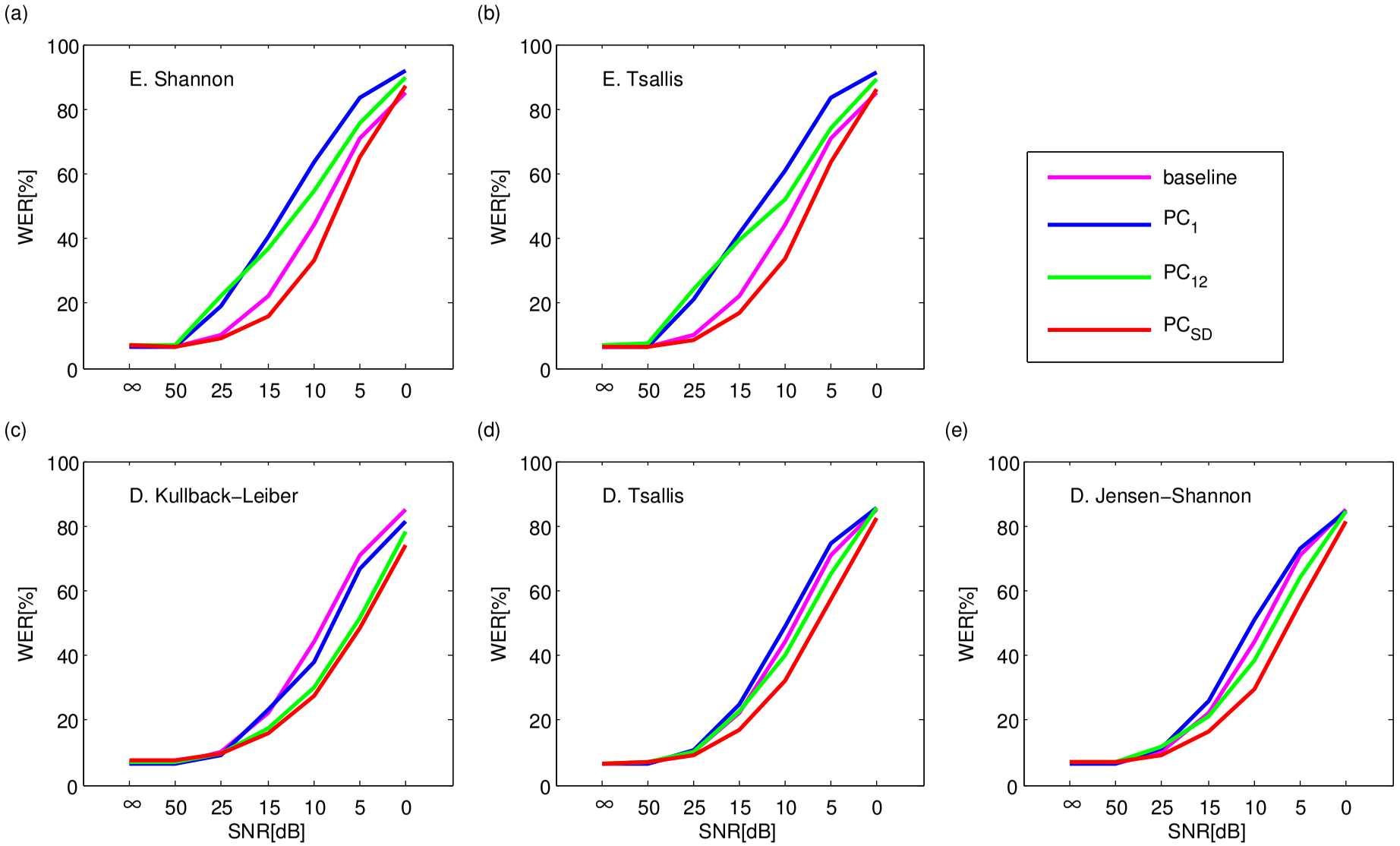

Figura 4.6: Tasa de error de palabra del sistema de ASR *vs.* SNR usando señales corrompidas con ruido aditivo blanco. Comparación entre el pre–procesamiento clásico (*baseline*) y los métodos propuestos: PC_1, PC_{12}, PC_{SD}, obtenidos con (a) entropía Shannon, (b) q–entropía, con $q = 0{,}2$, (c) divergencia de Kullback-Leibler, (d) q–divergencia, con $q = 0{,}2$, y (e) divergencia de Jensen–Shannon.

Tabla 4.1: Mejora del error relativo ($\Delta\epsilon_{\%}$) de las diferentes medidas obtenidas con los métodos PC_1, PC_{12} y PC_{SD} comparadas con el pre–procesamiento clásico para señales de habla corrompidas con ruido aditivo de *murmullo*. Los valores en negrita indican $\Pr(\epsilon < \epsilon_{ref}) > 99{,}99\,\%$.

	SNR_{dB}	∞	50	25	15	10	5	0
	PC_1	1.00	-2.31	-57.24	-59.67	-51.04	-30.82	-18.19
CME	PC_{12}	-7.14	-7.29	-68.55	-61.25	-48.52	-28.74	-17.74
	PC_{SD}	-6.28	-3.57	-5.81	**20.63**	-4.46	-17.60	-16.97
	PC_1	-0.62	-2.09	-59.42	-59.10	-48.23	-29.11	-16.52
$\mathbf{CME}_{q=0,2}$	PC_{12}	-7.46	-13.59	-68.37	-58.34	-44.13	-25.87	-16.72
	PC_{SD}	-2.20	0.78	-2.28	**24.27**	-0.23	-14.65	-15.97
	PC_1	-5.71	-2.71	-6.01	-16.62	-24.08	-10.99	-29.16
CMD	PC_{12}	-10.77	-11.10	-15.07	-10.83	-7.50	**21.77**	**13.74**
	PC_{SD}	-14.67	-9.99	-14.78	**40.86**	**94.52**	**29.07**	**17.60**
	PC_1	-2.70	-1.87	-4.50	-13.53	-15.40	-5.85	-3.90
$\mathbf{CMD}_{q=0,2}$	PC_{12}	-5.43	-2.58	-12.52	**14.63**	**18.01**	**15.76**	0.16
	PC_{SD}	-5.43	-7.76	-1.11	**44.75**	**41.39**	**26.18**	-1.02
	PC_1	-0.06	-0.65	-22.61	-25.60	-21.22	-6.51	0.97
$\mathbf{CMD}_J$	PC_{12}	-7.59	-5.10	-22.12	8.12	**11.67**	**9.35**	-0.83
	PC_{SD}	-9.21	-6.87	-5.65	**36.23**	**33.19**	**17.93**	0.62

En la figura 4.6 se muestra el WER obtenido con la parametrización clásica *vs.* los métodos propuestos para ruido blanco mezclado aditivamente a diferentes SNRs, organizado de forma similar a lo que se muestra en la figura anterior.

En este caso, las figuras (a) y (b), correspondientes a las entropías, muestran que el método PC_{SD} presenta un tasa de error menor que la que se obtiene para el MFCC, en particular para 5, 10 y 15 dB de SNRs. Se puede apreciar, asimismo, que el uso de divergencias mejoran aún más los resultados para este método.

El método PC_{12} también mejora su desempeño cuando se utilizan medidas de informaón relativas, especialmente, cuando se utiliza la divergencia de Kullback-Leibler, figura 4.6 (c). Este último, es el único caso donde el método PC_1 mejora la referencia.

Para el caso de los experimentos con ruido aditivo blanco, la divergencia de Kullback-Leibler presenta el mejor comportamiento, especialmente para bajas SNRs (menores que 10 dB para el método PC_1 y menores que 15 dB para los demás métodos). Para SNRs altas, las tasas de reconocimiento son cercanas a las que presenta la referencia.

Comparando las figuras 4.5 y 4.6, (d) y (e), se observa que el sistema presenta un desempeño similar con cualquiera de los tres métodos propuestos, para ambos tipos de ruido. Sin embargo, dado que el ruido de murmullo es menos estacionario que el ruido blanco, y teniendo presente que se calculan entropías relativas entre ventanas temporales consecutivas, es de esperar que las medidas relativas tengan un comportamiento mejor que las entropías de Shannon y Tsallis cuando se añade ruido de murmullo a la señal.

Los resultados muestran que el método PC_{SD} presenta el mejor desempeño cuando se utilizan las medidas de información relativas. Esto podría estar relacionado con las características de la CMD, que se observan en la figura 4.4, donde las estructuras pertenecientes a la señal de habla limpia se manifiestan principalmente en

Tabla 4.2: Mejora del error relativo ($\Delta\epsilon_{\%}$) de las diferentes medidas obtenidas con los métodos PC_1, PC_{12} y PC_{SD} comparadas con el pre–procesamiento clásico para señales de habla corrompidas con ruido aditivo *blanco*. Los valores en negrita indican $\Pr(\epsilon < \epsilon_{ref}) > 99{,}99\,\%$.

	SNR_{dB}	∞	50	25	15	10	5	0
	PC_1	1.00	-1.26	-47.16	-45.17	-30.91	-14.89	-7.44
CME	PC_{12}	-7.14	-9.78	-55.35	-39.67	-19.17	-6.18	-5.10
	PC_{SD}	-6.28	-5.15	10.61	**40.68**	**32.51**	**9.55**	-2.39
	PC_1	-0.62	-1.91	-52.48	-46.62	-27.69	-14.65	-6.91
$CME_{q=0,2}$	PC_{12}	-7.46	-18.01	-59.04	-43.70	-15.51	-4.31	-4.87
	PC_{SD}	-2.20	-1.68	**18.02**	**30.21**	**31.00**	**11.77**	-1.25
	PC_1	-5.71	-4.30	7.58	-4.23	**16.31**	**6.94**	**4.62**
CMD	PC_{12}	-10.77	-12.37	2.29	**27.57**	**47.07**	**38.50**	**8.58**
	PC_{SD}	-14.67	-13.65	3.92	**39.16**	**60.27**	**47.60**	**14.92**
	PC_1	-2.70	-1.85	-7.56	-9.95	-9.98	-4.87	-0.60
$CMD_{q=0,2}$	PC_{12}	-5.43	-8.66	-1.69	-2.68	**10.55**	**8.78**	-0.47
	PC_{SD}	-5.43	-10.65	7.22	**30.02**	**36.76**	**24.51**	**3.13**
	PC_1	-0.06	0.14	-8.88	-14.47	-13.90	-2.65	0.81
CMD_J	PC_{12}	-7.59	-7.41	-13.61	5.10	**13.89**	**11.31**	0.33
	PC_{SD}	-9.21	-7.84	8.87	**33.98**	**49.03**	**26.62**	**4.66**

las escalas más bajas. En presencia de ruido, las estructuras en las escalas más altas presentan mayores modificaciones, lo que sugiere que la información correspondiente al ruido de murmullo se encuentra más concentrada en estas escalas.

Desde el punto de vista del PCA, en el método PC_1, cuando sólo se tiene en cuenta uno de los componentes principales, la información de la señal de habla cruda y la del ruido son simultáneamente combinadas e incluidas en el vector de coeficientes, por lo que el sistema no puede discriminar entre ellas. En el método PC_{12}, donde se utilizan el primer y segundo componente principal, podría esperarse que la información no proporcionada por el primer PC se manifieste en el segundo, brindando información adicional. Sin embargo, aún no está bien establecido que exista una correspondencia entre las componentes y la señal de voz y ruido. Esta ambigüedad parece ser resuelta por el tercer método aquí propuesto, PC_{SD}.

A fin de evaluar la significancia estadística de estos resultados, se ha estimado la probabilidad de que un reconocedor dado sea mejor que el sistema de referencia ($\Pr(\epsilon < \epsilon_{ref})$). Para realizar esta prueba se ha supuesto la independencia estadística de los errores de reconocimiento para cada palabra y se ha aproximado la distribución binomial de los errores por medio de una distribución Gaussiana. Esto es posible debido a que se cuenta con un gran número de palabras (11077 palabras, si se tienen en cuenta todas las particiones de prueba).

En la tabla 4.1 se muestran los errores relativos $\Delta\varepsilon_{\%}$ obtenidos con cada método, utilizando señales de habla corrompidas con ruido de murmullo a diferentes SNRs (tabla que se corresponde con la figura 4.5). En la tabla 4.2 se presentan los resultados correspondientes a los experimentos realizados utilizando ruido aditivo blanco. Valores positivos significan que los resultados proporcionan errores inferiores a la referencia. En letras negritas se han resaltado los resultados con significancia estadística mayor al 99.99 %. Se puede observar que, para el ruido de murmullo, el

método de PC_{SD} con la divergencia de Kullback-Leibler y la divergencia paramétrica presenta $\Pr(\epsilon < \epsilon_{ref}) > 99{,}99\,\%$, para SNRs entre 5 y 15 dB. Para el ruido blanco, este desempeño se obtiene para SNRs de 0 a 15 dB, con los mismos métodos.

Resultados de experimentos similares, donde se analizó el comportamiento de medidas de información en el dominio temporal, obtenidos en [68], mostraron mejoras en el error relativo del 21.25 % para el ruido blanco y 24.31 % para el ruido de murmullo, en el caso de 15 dB de SNR. Comparando estos resultados con los métodos propuestos, se puede observar que el método PC_{SD} proporciona un gran porcentaje de mejora del error relativo.

4.6. Conclusiones

Los resultados obtenidos con los experimentos realizados muestran que los métodos propuestos presentan un desempeño satisfactorio en el sistema de ASR. El método PC_{SD} proporciona un incremento significativo en los índices de reconocimiento comparado con la parametrización MFCC. Este comportamiento se observó tanto en los experimentos utilizando ruido de murmullo como en aquellos donde se usó ruido blanco, especialmente para las SNRs de 15, 10 y 5 dB y para las medidas de información relativas.

Los resultados obtenidos no sólo lograron superar los valores de referencia, sino que también los alcanzados en [68], donde se utilizaron medidas de información similares a las planteadas en esta tesis, pero en el dominio del tiempo. Esto demuestra que las parametrizaciones basadas en CME y CMD proporcionan información valiosa para que el sistema de ASR pueda llevar a cabo el reconocimiento, incluso en presencia de ruido aditivo. Esto podría estar relacionado con el hecho de que la detección de los cambios dinámicos del tracto vocal, a partir de la señal de habla, es un dato importante para el reconocimiento. Por otra parte, con el fin de decodificar el mensaje transportado por una señal de voz, el sistema auditivo humano utiliza, simultáneamente, información de diferentes escalas temporales. La estrategia de análisis basada en onditas que se presenta aquí, asemeja a estas características biológicas, proporcionando un nuevo enfoque para incluir esta información en la etapa de pre–procesamiento del sistema de ASR.

4.7. Comentarios de cierre del capítulo

En este capítulo se han utilizado las medidas de información multiresolución presentadas en el capítulo 3 como parte de pre–procesamiento de la señal de habla para un sistema de ASR.

Para esta aplicación se propusieron tres métodos, los cuales se probaron utilizando señales de habla corrompidas con ruido aditivo blanco y de murmullo. El desempeño de estas propuestas se comparó con el rendimiento del sistema utilizando la parametrización clásica MFCC.

En el capítulo siguiente, se utilizarán las medidas de información multiresolución para llevar a cabo otra tarea, la segmentación automática de fonemas. Para dicha aplicación se propondrán otras formas de codificar la señal.

Capítulo 5

Análisis Multiresolución Aplicado a la Segmentación Automática del Fonemas

5.1. Introducción

En el capítulo anterior se describieron y discutieron los resultados obtenidos utilizando la parametrización de la señal de habla basada en medidas de información multiresolución en el marco de un sistema de reconocimiento automático del habla. Se encontró que dicha codificación brinda información útil al sistema de reconocimiento, pues se resaltan características relacionadas con cambios en la señal de voz que aportan información para el reconocimiento. Debido a esto, se propuso investigar si este tipo de características podrían ofrecer información relevante para realizar la segmentación de la señal de habla a nivel de fonemas.

Contar con un conjunto grande de señales de voz etiquetadas es un punto fundamental en el área de investigación del habla. Pero obtener un corpus de habla manualmente etiquetado, a nivel de fonemas o de palabras, es una tarea compleja y costosa. Tradicionalmente, la tarea de segmentar y etiquetar los datos de habla ha sido realizada manualmente a través de fonetistas entrenados, que utilizan el audio y pistas visuales para llevar a cabo esta labor. Sin embargo, este procedimiento manual puede ser tedioso, subjetivo, tomar demasiado tiempo y ser propenso a errores, especialmente para aquellos registros de habla espontánea [221, 222, 223]. Como consecuencia, sólo se cuenta con cantidades confiables de datos de habla etiquetada para muy pocos lenguajes [224]. Para salvar estas dificultades se han desarrollado diferentes técnicas para transcripción/etiquetado fonético, las cuales han sido encaradas a través diferentes estrategias [225, 226, 227, 228, 229]. Algunos incorporan conocimiento lingüistico, como aquellos basados en modelos ocultos de Markov (HMM) [226, 230, 231]. Estos procedimientos utilizan un modelo ASR top-down, como el alineamiento de Viterbi dependiente del texto con modelización de variantes de pronunciación.

En el caso de los métodos de segmentación independientes del texto, no es necesario contar con conocimiento lingüistico previo de la señal de voz [223, 225, 232]. Estos procedimientos pueden ser útiles para realizar la segmentación cuando no se tiene la trascripción o esta es inaccesible o inexacta, por ejemplo, en sistemas de identificación del hablante o del lenguaje, síntesis de habla concatenada, entre otros.

El objetivo de la segmentación automática es organizar la señal de voz en una secuencia de segmentos que estén asociados con un conjunto de símbolos, que pueden representar fonemas, palabras, sílabas u otras unidades acústicas, con una mínima o ninguna intervención humana [233, 234].

Al igual que otras aplicaciones, la parametrización de la señal de habla es un paso previo importante en la tarea de segmentación. El objetivo de este pre-procesamiento es disminuir la cantidad de datos a procesar posteriormente, a fin de representar la señal con unos pocos coeficientes en los cuales se remarquen las características más importantes de la señal [102]. Uno de los desafíos que aparece al momento de realizar la segmentación automática es encontrar la codificación que mejor resalte la información relacionada a las transiciones fonéticas de la señal de voz. Por lo que el esquema de codificación seleccionado debería ser capaz de proveer características acústicas cuyo contenido de información esté íntimamente relacionado con las transiciones entre fonemas [235].

En esta tesis se utilizó el algoritmo de segmentación propuesto por Esposito y Aversano [235], el cual emplea una representación multibanda del habla. Además, este algoritmo permite realizar la segmentación sin necesidad de entrenar el sistema o contar con la transcripción fonética de la señal. El término multibanda designa aquellas parametrizaciones para las cuales cada componente del vector de características representa una banda de frecuencias particular. Se han registrado importantes avances en lo referido a representaciones multi-banda del habla para diferentes aplicaciones [95, 236, 237, 238, 239, 240]. Una de las motivaciones para abordar este tipo de representación es la evidencia de que las transiciones fonéticas no ocurren simultáneamente en las diferentes sub-bandas. Experimentos de alineación de fonemas en bandas selectivas mostraron desviaciones temporales significativas entre transiciones determinadas en alineaciones sub-bandas y a banda completa[241]. Esta propiedad, la cual es relevante para la segmentación del habla a nivel de fonemas, sugiere que realizar la segmentación fonética separadamente en diferentes intervalos de frecuencia da mejores resultados que operar en el rango completo de frecuencias.

En este capítulo se describen los experimentos realizados utilizando las parametrizaciones de la señal de habla basadas en medidas de información multiresolución, introducidas en el capítulo 3, en el marco de un sistema de segmentación automática de fonemas. Para realizar esta tarea se utilizó el algoritmo de segmentación propuesto por Esposito y Aversano [235], el cual se detalla en la sección 5.2. En la sección 5.3 se describen la base de datos utilizada en los experimentos, los diferentes esquemas de codificación utilizados, los índices calculados para evaluar el desempeño de las codificaciones propuestas en la tarea de segmentación y las pruebas de validación cruzada aplicadas. En la seción 5.4 se analizan y discuten los resultados obtenidos y en la seción 5.5 se presentan las conclusiones y los posibles trabajos a futuro, derivados de estos experimentos. Por último, en la sección 5.6 se presenta una propuesta de modificación del algoritmo de segmentación utilizado, surgida a partir de los resultados obtenidos en los experimentos anteriores, donde se reemplaza la primera etapa del algoritmo por la codificación basada en la CMD. Asimismo, se muestran los resultados alcanzados con esta modificación, comparándolos con el algoritmo original.

5.2. Algoritmo de segmentación automática de fonemas

El algoritmo de segmentación independiente del texto, propuesto por Esposito y Aversano [235], es un método que realiza la segmentación basada en la detección de inestabilidades espectrales en múltiples bandas de frecuencias y trabaja sobre un número arbitrario de características que varían a lo largo del tiempo, obtenidas a través de un análisis de tiempo corto de la señal. Este procedimiento automático trata de detectar transiciones abruptas en la evolución de estos parámetros, i.e. tramos del habla donde los valores de los parámetros cambian significativamente y de manera rápida. Se ha encontrado que las transiciones abruptas no ocurren simultáneamente para cada parámetro, aunque ocurren en un intervalo corto de tiempo [241]. Por esta razón, este algoritmo combina los eventos de transiciones abruptas, detectados por diferentes características, en una única indicación de límite fonético.

Este método es capaz de realizar la segmentación fonética del habla sin ningún conocimiento previo de la secuencia de fonemas contenida en la señal y sin necesidad de un entrenamiento previo. Los autores han probado que este algoritmo es capaz de obtener un mejor desempeño que otros métodos de segmentación, como los basados en modelos LPC, con funciones de cambio Delta cepstrum [242] y variación espectral [231]. Y, también, mejores resultados que métodos que no utilizan modelos y operan utilizando filtrados paramétricos, identificando los cambios espectrales en la señal de habla mediante medidas de distorsión espectral, como la divergencia espectral de Kullback–Leibler [243] y métodos de descomposición temporal [244].

Las etapas del método de segmentación utilizado se muestran en la Figura 5.1. Este algoritmo es regulado por tres parámetros operacionales: α, β y γ.

El parámetro α identifica cuantos tramos consecutivos, valores de las características de la señal, se utilizan para estimar la intensidad de un cambio abrupto. De esta manera, dado un conjunto de i secuencias de características temporales de la señal de habla $\{y_i[m], m = 0, 1, ..., M\}$, para $i = 1, \ldots, \mathcal{J}$, la función $\mathcal{F}_i^\alpha$ se calcula de la siguiente forma:

$$\mathcal{F}_i^\alpha[m] = \left| \sum_{\mu=m-\alpha}^{m-1} \frac{y_i[\mu]}{\alpha} - \sum_{\mu=m+1}^{m+\alpha} \frac{y_i[\mu]}{\alpha} \right|. \tag{5.1}$$

Se utiliza un procedimiento de umbralado relativo, a través del parámetro β, para identificar el tramo m^* donde una posible transición de un fonema a otro ha sido localizada. El método de umbralamiento relativo se realiza de la siguiente manera: dado un intervalo $[u, v] \subset [\alpha, M - \alpha]$, donde $\mathcal{F}_i^\alpha[u]$ y $\mathcal{F}_i^\alpha[v]$ son dos valles de la función $\mathcal{F}_i^\alpha$, el tramo $m^* \in [u, v]$ se selecciona tal que $\mathcal{F}_i^\alpha[m^*]$ tiene su maximo valor en este intervalo. Esto es:

$$\mathcal{F}_i^\alpha[m^*] > \mathcal{F}_i^\alpha[m^* - 1] > ... > \mathcal{F}_i^\alpha[u], \tag{5.2}$$

con $\mathcal{F}_i^\alpha[u] < \mathcal{F}_i^\alpha[u - 1]$ y

$$\mathcal{F}_i^\alpha[m^*] \geq \mathcal{F}_i^\alpha[m^* + 1] \geq ... \geq \mathcal{F}_i^\alpha[v], \tag{5.3}$$

donde $\mathcal{F}_i^\alpha[v] < \mathcal{F}_i^\alpha[v + 1]$.

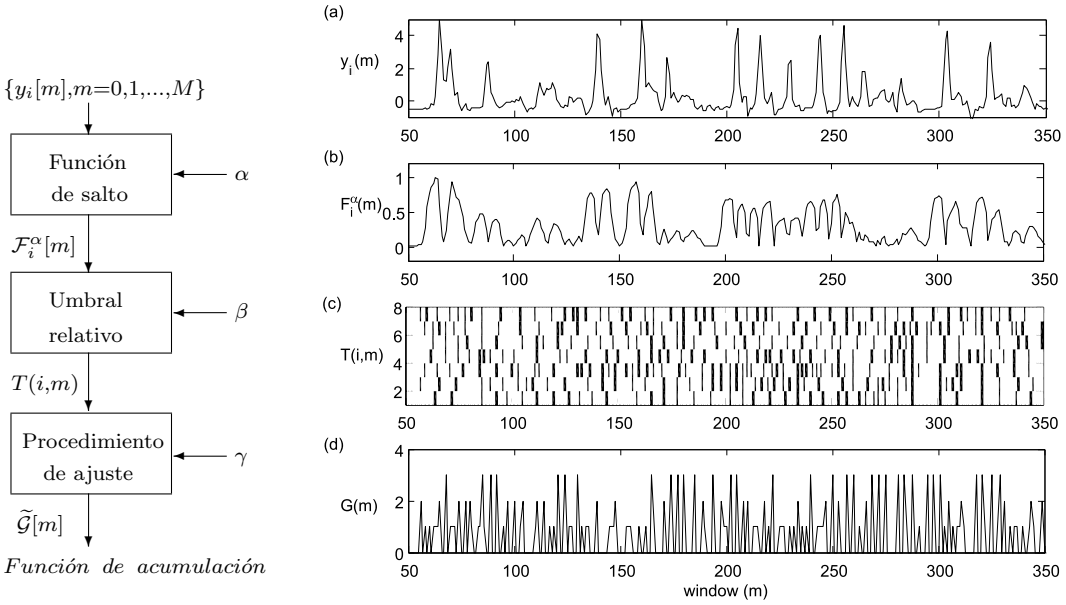

Figura 5.1: Esquema (izquierda) y figuras (derecha) correspondientes a las etapas del algoritmo de segmentación automática de fonemas independiente del texto, propuesto por [235]. (a) Evolución temporal de una de las características de la parameterización de la señal ($\mathbf{y}_1$), obtenida mediante la CMD. (b) Función de salto, $\mathcal{F}_1^\alpha$, de la secuencia temporal que se muestra en (a). (c) Matriz binaria $\mathbf{T} = \{T(i,m)\}$ obtenida a partir de la función descripta en (b). (d) Función de acumulación obtenida de $\widetilde{\mathcal{G}}[m]$, cuyos picos indican una posible transición fonética.

A partir de la la función anterior se obtiene un valor de altura relativa η, que se calcula como:

$$\eta = \text{mín}\left[\mathcal{F}_i^\alpha[m^*] - \mathcal{F}_i^\alpha[u], \mathcal{F}_i^\alpha[m^*] - \mathcal{F}_i^\alpha[v]\right]. \tag{5.4}$$

El tramo m^* donde η excede el umbral β, correspondiente a un pico de la ecuación (5.1), es considerado como una posible transición fonética y se guarda en una matriz binaria $\mathbf{T} = \{T(i,m)\}$. De esta manera, $T(i,m)$ es igual a 1 si una transición válida ha sido detectada en la secuencia temporal i en el tramo m. De lo contrario $T(i,m)$ es igual a 0.

Un procedimiento de ajuste tiene en cuenta la combinación de diferentes eventos de transición detectados en distintas características i, previamente guardados en la matriz $\mathbf{T}$, en una única indicación de limite fonético. Esto se debe a que las transiciones rápidas no ocurren simultaneamente para cada uno de los componente de las representación de la señal de habla, si bien estas ocurren en un intervalo corto de tiempo [241]. Por esta razón, el algoritmo de segmentación usa el procedimiento de ajuste para ubicar el punto de transición dentro de lo que se denomina como el baricentro de cada grupo de transiciones abruptas cuasi-simultáneas. En este sentido, las transiciones detectadas en el vecindario de un tramo m^* se combinan en una única indicación de límite fonético. El parámetro γ es utilizado en el procedimiento de ajuste para identificar el ancho del vecindario donde el baricentro es individualizado. Esto se realiza de la siguiente manera: para cada $m = 1, ..., M - \gamma + 1$ se considera

un intervalo $V = [m, m+1, ..., m+\gamma-1]$, donde se calcula la siguiente función:

$$\mathcal{G}[c] = \sum_{\mu=c}^{c+\gamma-1} \sum_{i=1}^{\mathcal{J}} T(i,\mu)\,|\mu - c|\,, \quad c \in V. \tag{5.5}$$

El posible baricentro del intervalo V es el tramo $\widetilde{c}$ donde:

$$\mathcal{G}[\widetilde{c}] = \min_{c \in V} \mathcal{G}[c], \quad m \leq \widetilde{c} \leq m+\gamma-1. \tag{5.6}$$

Para cada tramo m, el valor $\widetilde{\mathcal{G}}[m]$ indica cuantos baricentros $\widetilde{c}$ se han encontrado en el mismo. Esto conduce a una función de acumulación donde los picos corresponden a las indicaciones de posibles transiciones fonéticas.

En [235], el desempeño de este algoritmo de segmentación se analizó para diferentes representaciones multi-banda de la señal de habla, tales como MFCC, Log-area Ratios, LPC, PLP, entre otros. Los autores encontraron que la codificación Melbank fue la que mostró los mejores resultados para el algoritmo propuesto.

Para este trabajo, se utilizaron parametrizaciones basadas en medidas de información multiresolución como entrada al algoritmo presentado, ya que este tipo de procesamiento puede brindar información acerca de los cambios en la dinámica de la señal de voz, evidenciados en diferentes características o escalas.

5.3. Aspectos Principales de la Implementación

En esta sección se describen los principales aspectos que se tuvieron en cuenta para realizar la segmentación automática de fonemas utilizado una representación de habla basada en medidas de información multiresolución.

En la subsección 5.3.1 se explican las características del corpus de habla utilizado. Los esquemas de codificación basados en la CME y la CMD que se utilizaron como entrada en el algoritmo de segmentación automática descripto anteriormente, se describen en la subsección 5.3.2. Como medidas de información para obtener la CME y la CMD se usaron entropía de Shannon y divergencia de Kullback-Leibler, respectivamente. Los resultados de la segmentación lograda con las representaciones propuestas se compararon con los alcanzados mediante una parametrización Melbank clásica, la cual se detalla en la subsección 5.3.3. Finalmente en la subsección 5.3.4 se describen los índices calculados para evaluar el desempeño de las codificaciones usadas y las pruebas de validación cruzada aplicadas.

5.3.1. Señales y base de datos

Las señales utilizadas para los experimentos de segmentación fueron, también, las procedentes del subconjunto del corpus de habla española Albayzin [148]. Los detalles de este subconjunto del corpus se describieron en la sección 4.4.1 del capítulo anterior.

Este subconjunto fue etiquetado usando una pre-segmentación automática con un modelo de alineamiento basado en modelos ocultos de Markov, asistido por un experto. Esta información fue utilizada como segmentación patrón y consiste en un archivo que registra la posición de todos los límites fonéticos, expresados en milisegundos, y la etiqueta correspondiente.

5.3.2. Codificaciones evaluadas

Para estos experimentos se utilizaron tres diferentes esquemas de codificación: características Melbank, parametrización basada en CME utilizando la entropía de Shannon y parametrización basada en CMD usando la divergencia de Kullback-Leibler. Las características basadas en CME y CMD se obtuvieron de acuerdo al procedimiento descripto en el capítulo 3 sección 3.6, donde:

$$\mathbf{y}_i = \{y_i[m], m = 1, ..., M\}, i = 1, .., 8 \tag{5.7}$$

corresponden a los coeficientes que representan la señal de habla para cada intervalo m, provenientes de las primeras ocho filas de $\mathbf{Y}$ (asociadas con los ocho mayores valores de $\mathbf{\Lambda}$, la matriz diagonal de eigenvalores).

La elección de $\mathcal{J} = 8$ componentes para la nueva parametrización se realizó en acuerdo con la cantidad de características del esquema de codificación Melbank utilizado para realizar la comparación de los resultados. En [235] se evaluaron otras representaciones multi-bandas de la señal de habla utilizando el algoritmo de segmentación usado en esta tesis. Los autores encontraron que la codificación Melbank, utilizando ocho coeficientes, fue la que mostró el mejor desempeño.

Por otro lado, desde el punto de vista del PCA, los primeros ocho componentes contienen más del 95 % de la varianza total de la señal.

Cada frase en el corpus ha sido normalizada en media, pre-enfatizada y segmentada con una ventana Hamming en tramos de 20 ms, con un desplazamiento de 10 ms [102]. Para las tres parametrizaciones evaluadas se utilizó una representación de ocho coeficientes.

5.3.3. Parametrización Melbank

La parametrización Melbank es un procesamiento estándar de tiempo corto, basado en un modelo de bancos de filtros, como el mencionado en la sección 2.4.3 del capítulo 2. Esta parametrización utiliza la señal de habla discretizada y ventaneada y procesa cada uno de estos tramos a través de un banco de filtros pasa–banda, de manera tal de obtener una medida de la energía de la señal en una determinada banda de frecuencias [22]. La codificación Melbank correspondiente a cada tramo de la señal proporciona un conjunto de características acústicas que pesan su espectro de Fourier en el tramo correspondiente y modela las características espectrales del habla.

Para la implementación de la parametrización Melbank se utilizó la siguiente aproximación para la escala de mel [204]:

$$Mel(f) = 2595 \log_{10} \left(1 + \frac{f}{700}\right), \tag{5.8}$$

Antes de realizar la parametrización Melbank, la señal de habla discretizada $\mathbf{s} = \{s[k]\}$, de longitud K, es ventaneada en M tramos de ancho L, superpuestos en Δ. Los parámetros $L \in \mathbb{N}$ y $\Delta \in \mathbb{N}$ se eligen de manera tal que $L \leq K$ y $(K - L)/\Delta = M \in \mathbb{Z}$. De esta manera, cada tramo m es una partición de la señal de habla, que se denota como $\mathbf{s}_m = \{s_m[l]\}$, $l = 1, ..., L$, i.e. la señal $\mathbf{s}$ correspondiente al tramo arbitrario m, con $m = 1, ..., M$. Una parametrización basada en bancos de filtros se obtiene haciendo pasar $\mathbf{s}_m$ a través de un banco de $\mathcal{J}$ filtros

pasabandas, los cuales cubren el rango de frecuencias de **s**, desde f_{min} a f_{max}. Esto se realiza calculando primero el espectro de potencia de tiempo corto $S_m(f)$, el cual se obtiene multiplicando el complejo de la transformada discreta de Fourier (DFT) del segmento de habla en el tramo m por su conjugado. Por medio de la ecuación (5.8) se realiza un cambio de variables en el espectro de potencia $S_m(f)$ y se obtiene el espectro de potencia en escala de mel $S_m(\nu)$, con $\nu = Mel(f)$. $S_m(\nu)$ se convoluciona luego con una curva de enmascaramiento triangular $B(\nu)$ de ancho τ dada por:

$$B(\nu) = \begin{cases} 2\nu/\tau & \text{if} \quad 0 < \nu \leq \tau/2 \\ 2 - 2\nu/\tau & \text{if} \quad \tau/2 < \nu \leq \tau \end{cases}, \tag{5.9}$$

$$\tau = 2\,\frac{Mel(f_{max}) - Mel(f_{min})}{\mathcal{J}}, \tag{5.10}$$

donde $\mathcal{J}$ es la cantidad deseada de filtros y f_{min} y f_{max} son, respectivamente, las frecuencias de corte inferior y superior del rango completo del banco de filtros. Se obtiene así un conjunto de filtros triangulares igualmente espaciados a lo largo de la escala de mel. Esta convolución se describe matemáticamente con la ecuación:

$$\hat{S}_m(\nu_i) = \sum_{\nu} S_m(\nu - \nu_i)B(\nu), \quad i = 1, ..., \mathcal{J}, \tag{5.11}$$

donde $\nu_i = \frac{\tau}{2}i$ son las frecuencias centrales (en mels) de los filtros y $\hat{S}_m(\nu_i)$ son los muestras resultantes del espectro. Esto significa que cada coeficiente de magnitud de la DFT se multiplica por la correspondiente ganancia del filtro de $B(\nu)$ y, luego, el resultado es acumulado. Así, cada $\hat{S}_m$ mantiene una suma ponderada que representa la magnitud espectral en el correspondiente canal del banco de filtros.

Finalmente, el espectro $\hat{S}_m(\nu_i)$ se comprime logarítmicamente. De esta manera, se obtienen los coeficientes $y_i[m] = \ln \hat{S}_m(\nu_i)$, $i = 1, ..., \mathcal{J}$, que son las características Melbank correspondientes al tramo m de la señal de habla.

5.3.4. Índices para evaluar el desempeño de la segmentación

A fin de poder evaluar la segmentación que se obtiene con las parametrizaciones basadas en la CME y la CMD, se consideraron dos índices de desempeño: el porcentaje de límites fonéticos correctamente detectados (PC) y el porcentaje de puntos erróneamente insertados (PI), también llamada tasa de sobre-segmentación [245].

Para estudiar la eficacia de la segmentación automática, los resultados obtenidos fueron comparados con la segmentación patrón del corpus de Albayzin. Los índices PC y PI se calculan en relación con la información de etiquetado de la base de datos.

Con el objeto de contrastar las diferencias, las segmentaciones obtenidas con las codificaciones propuestas en esta tesis se compararon con la segmentación que se consigue utilizando la codificación Melbank como sistema de referencia.

El índice PC relaciona el número total de límites fonéticos que figuran en la base de datos, B_T, con la cantidad de límites correctamente detectados, B_C, y se define como:

$$PC = \frac{B_C}{B_T}100. \tag{5.12}$$

Un punto de segmentación se considera como un límite correcto si corresponde a una transición fonética detectada por el algoritmo con una tolerancia de ± 20 ms.

Este tiempo es equivalente al umbral de detección de límites fonéticos que puede alcanzar un experto humano [79].

El número de límites erróneamente insertados, B_I, se cuantifica por la diferencia entre el total de puntos de segmentación detectados por el algoritmo B_D y B_C: $B_I = B_D - B_C$. Así, el índice PI relaciona los límites detectados erróneamente con la cantidad total de tramos F_T de la señal de voz. Esto se expresa como:

$$PI = \frac{B_I}{F_T} 100. \tag{5.13}$$

Se evaluaron otros índices derivados de la teoría de detección de señales, similares a los definidos anteriormente aunque matemáticamente más precisos. Teniendo en cuenta la segmentación patrón como el objetivo para la tarea de detección; un punto de segmentación detectado automáticamente será aceptado como un límite fonético si coincide con un punto de la segmentación patrón, de otro modo se considerará como una transición errónea. En este marco, un punto identificado incorrectamente como un límite fonético da un error conocido como una falsa alarma. Cuando el algoritmo de segmentación no puede detectar un punto que efectivamente corresponde a una transición entre fonemas, el tipo de error se conoce como una detección fallida.

Esto permite definir dos índices de desempeño, denominados como la tasa de falsas alarmas P_{fa} y la tasa de detección perdida P_{md}, que se obtienen a través de las siguientes fórmulas matemáticas:

$$P_{fa} = \frac{B_I}{F_T - B_T} 100 \tag{5.14}$$

$$P_{md} = \frac{B_T - B_C}{B_T} 100. \tag{5.15}$$

Con los valores P_{fa} y P_{md} se construyen las curvas de características operativas del receptor (ROC) para cada esquema de codificación evaluado.

Nuevamente, con el fin de evaluar la significación estadística de estos resultados, se estima la probabilidad de que un determinado esquema de codificación sea mejor que una parametrización de referencia ($\Pr(\epsilon < \epsilon_{ref})$). Para lo cual, se supone la independencia estadística de los errores de detección de las transiciones para cada fonema y se aproxima la distribución binomial de los errores por medio de una distribución gaussiana.

5.4. Resultados y discusión

En esta sección se presentan y discuten los resultados obtenidos utilizando las parametrizaciones basadas en CME y CMD, introducidas en el capítulo 5.3.2, para realizar la segmentación automática de fonemas con el algoritmo descrito en la sección 5.2. El desempeño de las codificaciones propuestas se comparó con el obtenido mediante la parametrización Melbank.

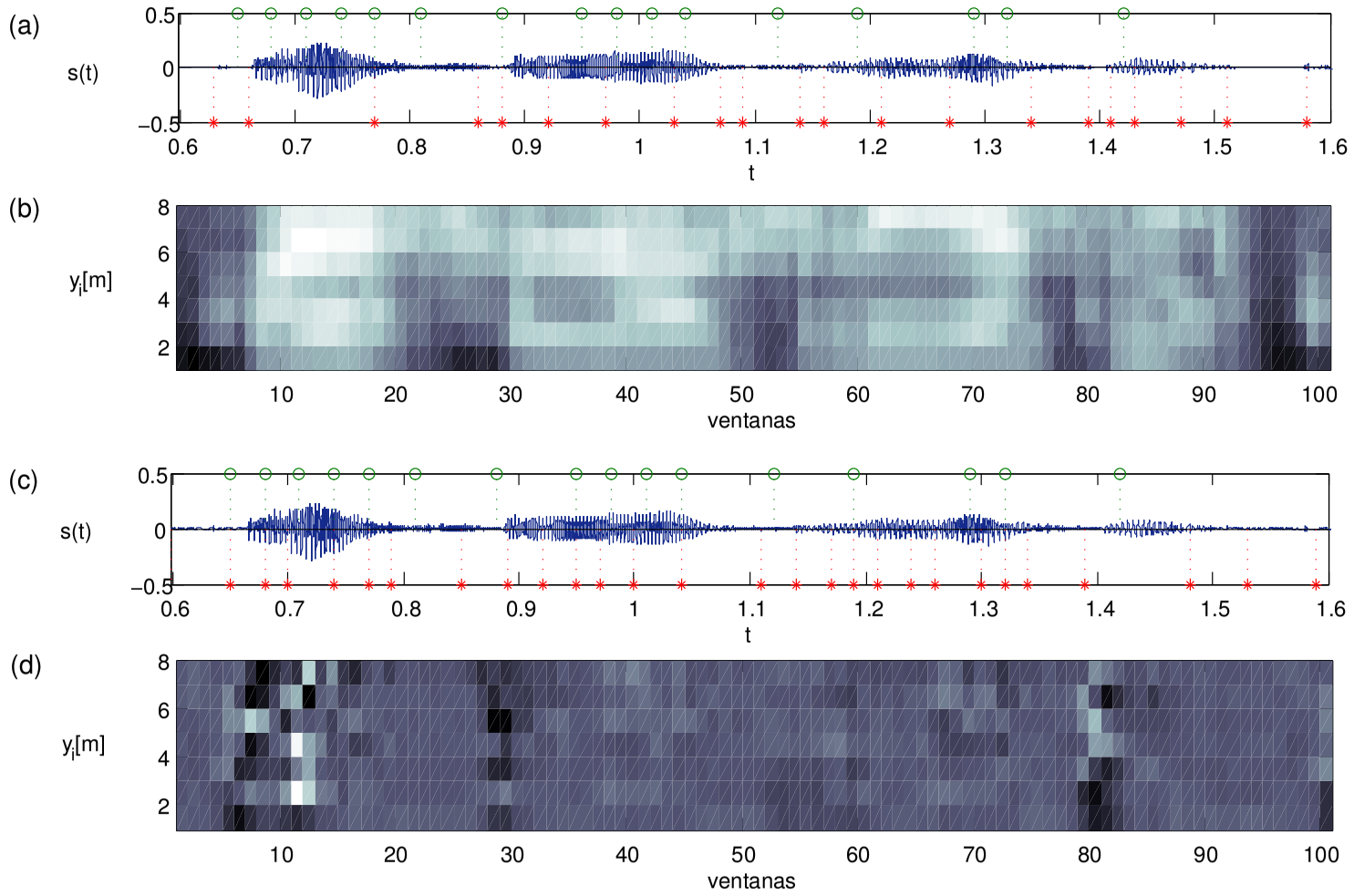

Figura 5.2: (a) Señal de voz con etiquetado Albayzin (líneas punteadas superiores con marcadores de círculos) y segmentación automática realizada utilizando parametrización Melbank (líneas punteadas inferiores con marcadores estrella), con $\alpha = 6$, $\beta = 0{,}01$ y $\gamma = 3$. (b) Representación Melbank correspondiente a la señal representada en (a). (c) Señal de voz con etiquetado Albayzin (líneas punteadas superiores con marcadores de círculos) y segmentación automática realizada con la parametrización basada en CMD (líneas punteadas inferiores con marcadores estrella), para los mismos parámetros operacionales utilizados en (a). (d) Parametrización basada en CMD de la señal representada en (c).

A modo de ilustración, en la Figura 5.2(a) se presenta una parte de la señal de habla de la sentencia del corpus Albayzin: "¿Cuáles son los ríos cuya longitud es superior a cien kilómetros?". Las líneas punteadas superiores (con marcadores con círculos) indican la segmentación patrón, obtenida a partir de los archivos de voz etiquetada. Las líneas punteadas inferiores (con asteriscos) representan los puntos de segmentación detectados por el algoritmo usando la representación Melbank. Esta parametrización se muestra en la Figura 5.2(b). La Figura 5.2(c) muestra la misma señal de habla desplegada en (a), pero ahora las líneas de puntos con asteriscos corresponden a los puntos de segmentación detectados por el algoritmo usando la parametrización basada en CMD. La Figura 5.2(d) presenta la codificación basada en CMD de la señal de habla.

Tabla 5.1: Porcentaje de límites fonéticos detectados correctamente (PC) y porcentaje de puntos erróneamente insertados (PI) obtenidos para los tres esquemas de codificación evaluados utilizando el algoritmo de segmentación para diferentes parámetros operacionales. Los números en negrita indican los mejores resultados obtenidos para cada parametrización.

Parámetros			Codificaciones					
γ	α	β	Melbank		CME		CMD	
			PC	PI	PC	PI	PC	PI
3	2	0.01	**86.52**	16.63	81.64	16.79	84.65	15.13
		0.05	83.97	15.37	84.40	17.44	81.80	8.36
		0.1	79.97	10.49	85.48	17.51	77.20	6.96
	4	0.01	81.11	13.56	83.79	14.90	83.09	12.59
		0.05	78.78	11.96	83.65	15.00	79.31	7.01
		0.1	75.92	7.24	83.39	14.83	74.98	6.08
	6	0.01	75.41	13.31	86.25	17.43	**86.75**	13.87
		0.05	72.02	8.77	**86.91**	17.16	83.51	9.04
		0.1	68.05	5.22	86.09	16.61	79.94	8.05

La tabla 5.1 muestra los índices PC y PI para los tres esquemas de codificación evaluados, usando los siguientes parámetros operacionales para el algoritmo de segmentación: $\gamma = 3$, $\alpha = 2$, 4, 6 y $\beta = 0{,}01$; 0,05; 0,1. Los números en negrita indican el mejor PC obtenido para cada codificación y su valor asociado de PI. Se puede observar que las dos parametrizaciones propuestas tienen índices PC mayor que Melbank. Para casi todos los parámetros operacionales que se presentan en la tabla 5.1, los índices PC obtenidos con las parametrizaciones basadas en CME y CMD son mejores que los correspondientes a Melbank. La parametrización basada en CME muestra índices PC aceptables, mayores al 83 % para los parámetros operacionales listados en la tabla, aunque los valores PI no presentan un buen desempeño. En cambio, la parametrización basada en CMD muestra un mejor comportamiento, ya que los índices PC son altos, con valores bajos de PI. Esto se evidencia comparando los índices en negrita de la Tabla 5.1. Vale la pena aclara que los índices PC y PI óptimos son 100 % y 0 % respectivamente.

Tabla 5.2: Evaluación de la significancia estadística de los mejores valores de PC de la Tabla 5.1 (en negrita) y de sus correspondientes valores PI, comparando cada una de las parametrizaciones propuestas con respecto a la referencia (Melbank). Las flechas hacia arriba y abajo indican que la codificación incrementa o decrementa los correspondientes índices.

$\Pr(\epsilon < \epsilon_{ref})$	CME	CMD
PC	$\uparrow$ 93,02 %	$\uparrow$ 80,57 %
PI	$\uparrow$ 96,11 %	$\downarrow$ 99,99 %

La tabla 5.2 muestra la evaluación de la significancia estadística de los mejores valores PC (en negrita) de la tabla 5.1, y los correspondientes valores PI. Se puede observar que la codificación basada en CME aumenta significativamente la tasa de detección, la probabilidad de que mejore el índice PC es de 93,02 % ($\Pr(\epsilon < \epsilon_{ref})$). En cambio, la probabilidad de incrementar este valor usando la codificación basada en CMD es 80,57 %. Sin embargo, la CME aumenta también la probabilidad de la tasa de sobre-segmentación en un 96.11 %, mientras que la CMD la decrementa en un 99.99 %.

Por simplicidad, en la tabla 5.1 sólo se muestran los parámetros operacionales que proveen los mejores resultados para los esquemas de codificación evaluados. Otros valores de parámetros operacionales han sido utilizados y su desempeño se muestra en las figuras 5.3, 5.4 y 5.5.

En la figura 5.3 se compara el desempeño del algoritmo de segmentación para las codificaciones Melbank (círculos), basada en CME (cuadrados) y basada en CMD (rombos) cuando se utilizan valores de $\alpha = 2, 3, 4, 5, 6$, con $\gamma = 3$ y $\beta = 0{,}01$. La Figura 5.3(a) muestra el PC obtenido con estas parametrizaciones para los diferentes valores del parámetro α. Los valores más altos corresponden al mejor rendimiento. En la Fig. 5.3(b) se muestra el índice PI para estos experimentos. Aquí, los valores más bajos presentan el mejor rendimiento.

Se observa que Melbank decae de manera sostenida a medida que α se incrementa, mientras que las codificaciones basadas en CME y CMD son menos afectadas por los cambios en este parámetro. Para $\alpha \geq 3$, las parametrizaciones propuestas muestras mejores valores de PC. En el caso del índice PI, la parametrización basada en CMD proporciona los mejores resultados para $\alpha \leq 4$ y valores similares a los correspondientes a Melbank para $\alpha = 5$ y 6. La parametrización basada en CMD ofrece el mejor balance entre PI y PC en el rango considerado para el parámetro α. Estos resultados son consistentes con los observados en la tabla 5.1.

Los resultados de la segmentación utilizando las codificaciones basadas en CME y CMD parecen ser más estables que Melbank cuando el parámetro α crece. Esto podría estar relacionado con las características de la evolución de los coeficientes de estas parametrizaciones, que se pueden ver en la figura 5.2. A partir de las figuras 5.2(b) y (d) se observa que la evolución de los parámetros de Melbank es más suave que aquellos basados en CMD. Como ya se ha mencionado, cuando aparece un cambio abrupto en la señal de voz, esto se evidencia no sólo en los parámetros correspondientes al tramo donde se produce el evento, sino también en un vecindario de este tramo. Este vecindario es más ancho para Melbank que para la codificación

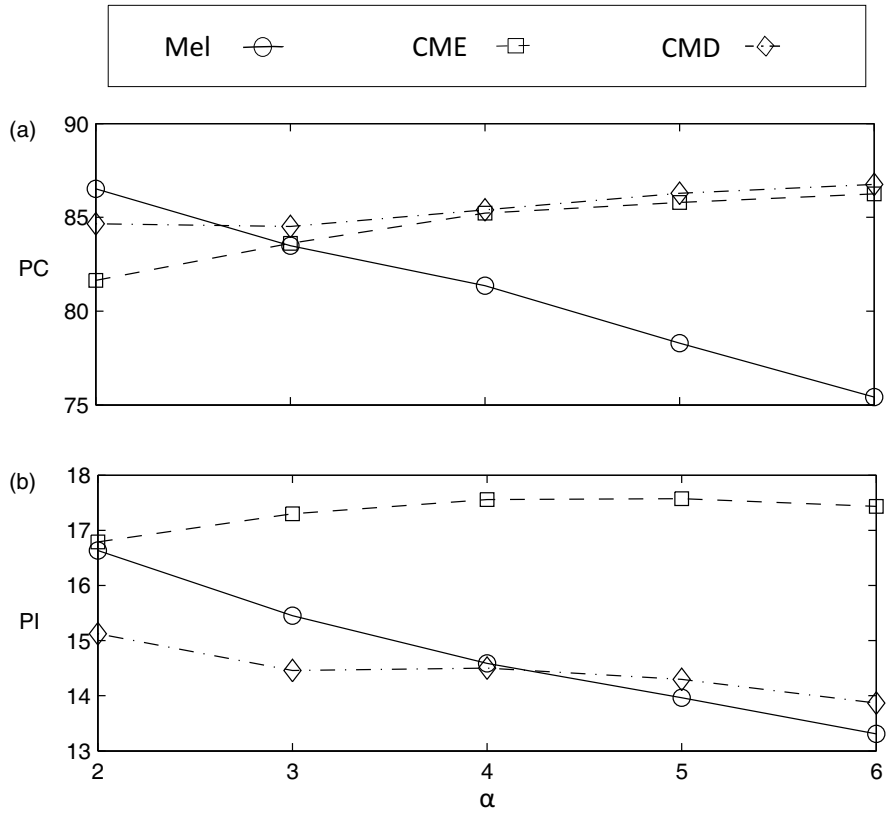

Figura 5.3: (a) Porcentaje de límites fonéticos detectados correctamente (PC) y (b) porcentaje de puntos erróneamente insertados (PI), obtenidos con el algoritmo de segmentación automática de fonemas utilizando la parametrización Melbank clásica (línea continua con círculos), la codificación basada en CME (línea discontinua con cuadrados) y la codificación basada en CMD (línea de puntos y rayas con rombos) para $\alpha = 2, 3, 4, 5, 6$, $\beta = 0{,}01$ y $\gamma = 3$.

basada en CME o CMD. El parámetro α se relaciona con el número de tramos que se utilizan para calcular la función de salto (5.1). Debido a esto, cuando se cambia el valor del parámetro α, la parametrización basada en CMD todavía tiene un buen desempeño. Los cambios en la señal de voz se concentran en un pequeño grupo de tramos. Como resultado de esto, el valor de α no afecta al rendimiento del algoritmo cuando se usa este esquema de codificación.

De la comparación de las figuras 5.2(a) y (c) se puede apreciar que la CMD hace más detecciones del tipo verdaderos positivos en la primera parte de la señal mostrada, mientras que en la segunda parte aparecen algunas detecciones falsas positivas. En contraste, en (a) hay casi el mismo número de falsos positivos, pero también hay varios falsos negativos. Esto sugiere que, para el propósito de la segementación de fonemas, la CMD podría ofrecer un mejor rendimiento.

En la figura 5.4 se compara el desempeño del algoritmo de segmentación, usando Melbank y las parametrizaciones propuestas, para $\beta = 0{,}01; 0{,}05; 0{,}1; 0{,}2$, donde $\alpha = 6$ y $\gamma = 3$. La elección de los valores para α y γ se debe a que estos brindan segmentaciones similares para las tres codificaciones, lo cual permite realizar una mejor comparación de la influencia de β para cada caso (ver Tabla 5.1). En dicha figura se puede observar que los índices PC y PI caen sostenidamente a medida que β aumenta. Esto está relacionado con el hecho de que este parámetro determina el

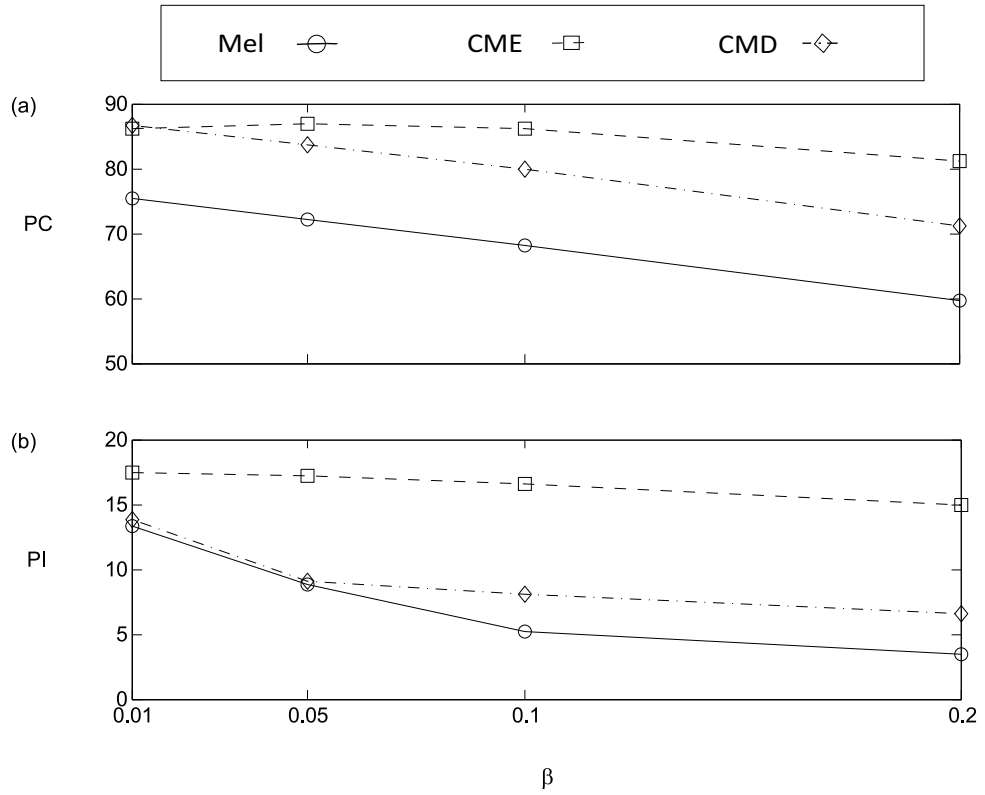

Figura 5.4: (a) Porcentaje de límites fonéticos detectados correctamente PC y (b) porcentaje de puntos erróneamente insertados PI, obtenidos para el algoritmo de segmentación automática de fonemas usando la parametrización Melbank clásica (línea continua con círculos), la codificación basada en CME (línea discontinua con cuadrados) y la codificación basada en CMD (línea de puntos y rayas con rombos) para $\alpha = 6$, $\gamma = 3$ and $\beta = 0{,}01; 0{,}05; 0{,}1; 0{,}2$.

umbral aplicado a la función $\mathcal{F}_i^{\alpha}[m]$, dada por (5.1), para detectar sus picos. Por lo tanto, cuando β es alto, la sensibilidad del algoritmo es baja y algunos cambios no pueden ser detectados. El parámetro β se selecciona $\leq 0{,}1$.

Como se puede observar en la figura 5.4, los índices de rendimiento para la segmentación obtenida mediante la parametrización basada en CME sólo presentan un ligero deterioro. Esto se debe a que esta parametrización tiene más variaciones (picos más aislados) que las otras. La razón es que la codificación basada en CME tiene en cuenta la información de un tramo a la vez, sin conocimiento de los tramos de los alrededores. La parametrización basada en CMD, en cambio, calcula la divergencia de Kullback-Leibler de tramos consecutivos, lo cual suministra información relativa tramo a tramo. La falta de dependencia entre tramos en el cálculo de la CME produce muchos puntos en la segmentación, lo cual se traduce no sólo en índices PC altos, sino también más límites erróneamente insertados (alto PI). Esta característica de la CME es lo que determina, también, las formas de las curvas PC y PI de la Figura 5.3.

La figura 5.5 muestra el desempeño del algoritmo de segmentación para diferentes valores del parámetro $\gamma = 3, 4, 5, 6, 7$. Se puede apreciar que las parametrizaciones basadas en CME y CMD presentan índices PC buenos para todos los valores de γ. La curva de PI para la codificación basada en CMD es similar a los que se obtienen para Melbank (la más baja) y se mantiene casi constante para $\gamma \geq 4$.

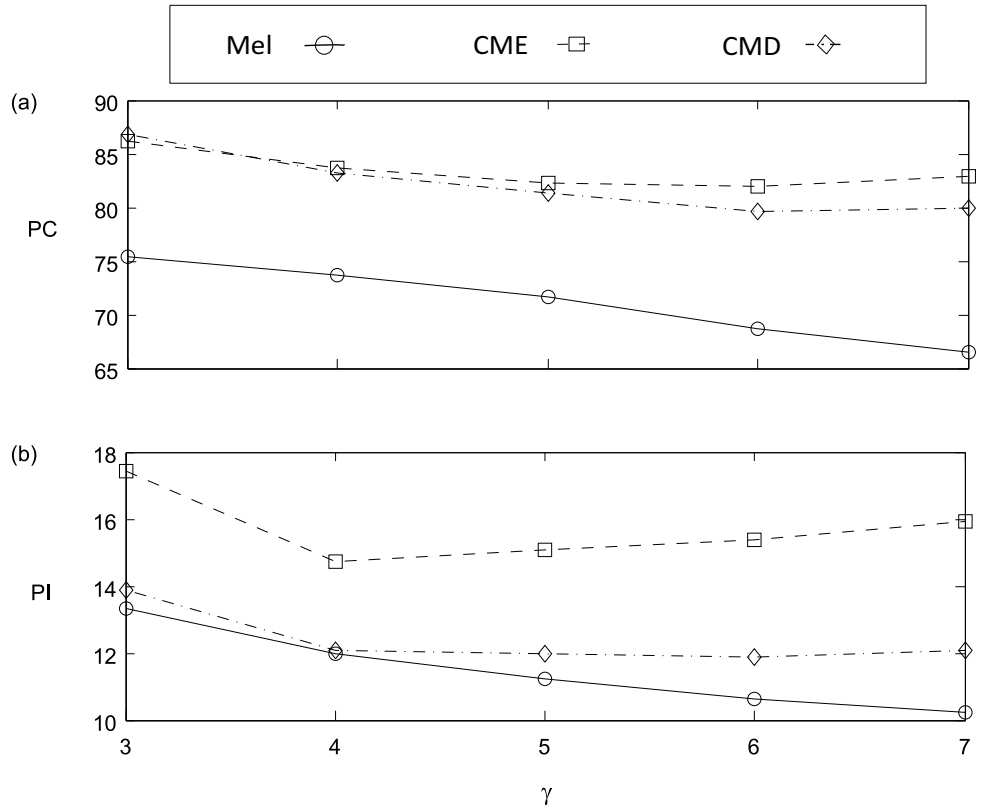

Figura 5.5: (a) Porcentaje de límites fonéticos detectados correctamente PC y (b) porcentaje de puntos erróneamente insertados PI, obtenidos para el algoritmo de segmentación automática de fonemas usando la parametrización Melbank clásica (línea continua con círculos), la codificación basada en CME (línea discontinua con cuadrados) y la codificación basada en CMD (línea de puntos y rayas con rombos) para $\gamma = 3, 4, 5, 6, 7$, $\alpha = 6$ y $\beta = 0{,}01$.

El parámetro evaluado en la figura 5.5 está relacionado con el vecindario utilizado para realizar el procedimiento de ajuste en el algoritmo de segmentación. Por lo tanto, valores bajos de γ incrementan el índice PI debido a que un vecindario estrecho en el procedimiento de ajuste no cubre por completo la cantidad de tramos donde se produce una transición fonética. Esto podría producir que cierta transición se compute como dos puntos de segmentación en lugar de uno. Por otro lado, cuando γ aumenta, el vecindario a procesar es más ancho y tramos adyacentes que corresponden a diferentes límites fonéticos se procesan como uno sólo. Esto resulta en una disminución en el rendimiento del algoritmo que se puede observar, también, en el índice PC de la parametrización Melbank y, de alguna manera, en la codificación basada en CMD (Figura 5.5(a)).

Los resultados obtenidos sugieren que el algoritmo de segmentación de fonemas, utilizando los dos esquemas de codificación que aquí se proponen, muestra valores más altos de límites fonéticos detectados correctamente (mayor PC) que Melbank. En este sentido, la parametrización basada en CMD es la que tiene el mejor rendimiento, ya que no sólo aumenta el porcentaje de límites detectados correctamente, PC, sino que también disminuye el número de puntos insertados erróneamente, PI. Por otro lado, la codificación basada en CME sólo aumenta el índice PC, mientras que PI sigue siendo alto.

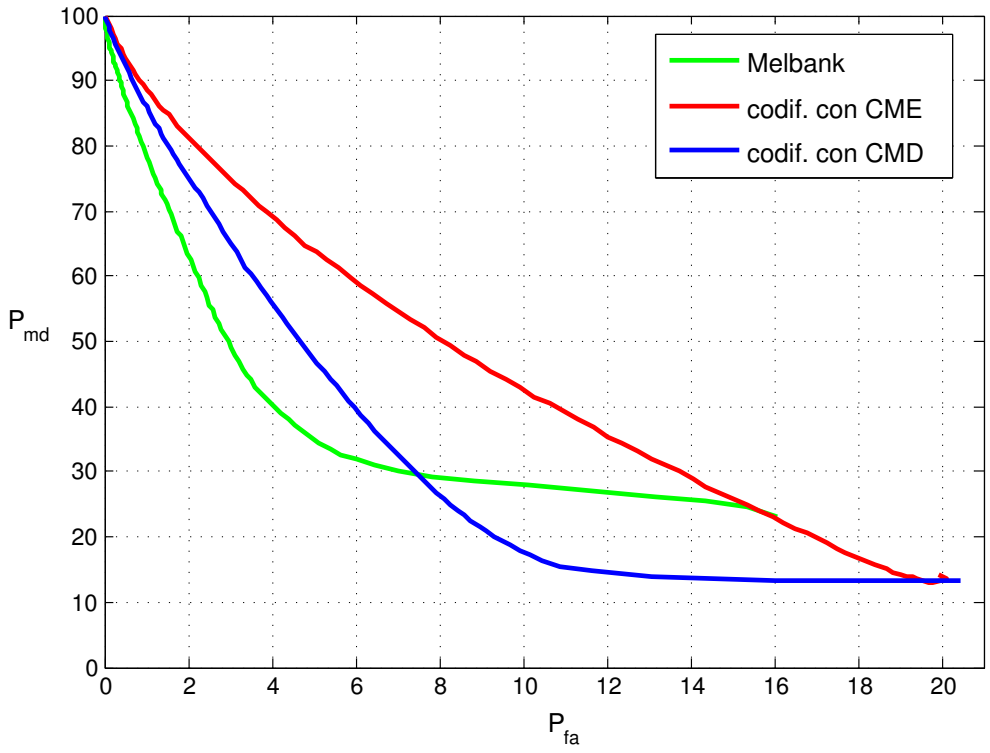

Figura 5.6: Curvas ROC para el algoritmo de segmentación de fonemas usando la parametrización Melbank clásica (línea verde), la codificación basada en CME (línea roja) y la codificación basada en CMD (línea azul). Los parámetros operacionales son $\alpha = 6$, for $\gamma = 3$ y β variando de 0 a 1 con incrementos de 0,01.

Los índices PC y PI por sí mismos no son indicadores fiables de buen rendimiento. Como resultado de esto, se obtienen y grafican las curvas ROC.

En la Figura 5.6 se muestran las curvas ROC, construidas con los índices P_{fa} y P_{md}, a fin de comparar el rendimiento del algoritmo de segmentación utilizando los tres esquemas de codificación considerados en este trabajo. En esta figura se ha utilizado $\alpha = 6$, $\gamma = 3$ y β variando entre 0 y 1 (con pasos de 0.01). Los valores altos de P_{fa} se producen debido a que se han obtenido muchas sobre-segmentaciones. Valores altos de P_{md} indican que el algoritmo ha segmentado en menos la señal. Como puede verse en la figura, cuando P_{fa} es bajo, P_{md} es alto y viceversa. Cuanto más cerca esta la curva al eje inferior y al eje de la izquierda, más precisa es la detección. Se puede observar que la parametrización basada en CMD da mejores resultados que Melbank, ya que permite una reducción de los P_{md}, por debajo del 30 %, con relativamente bajos incrementos de P_{fa}. Por otro lado, la codificación basada en CME no presenta buenos resultados. Esto confirma que la parametrización basada en CMD es la mejor codificación, en particular para requerimientos con baja tasa de sobre-segmentación.

Estos resultados indican que la parametrización basada en CMD proporciona información relacionada con cambios bruscos en la señal de voz, lo cual mejora el rendimiento de la tarea de segmentación de fonemas.

5.5. Conclusiones

En esta aplicación se han propuesto dos nuevas codificaciones de la señal de habla, basadas en la entropía de Shannon y la divergencia de Kullback-Leibler, obtenidas a través de una representación en el plano tiempo–escala de la señal. Estas parametrizaciones se utilizaron como entrada de un algoritmo de segmentación automática y se comparó su desempeño con el obtenido mediante la representación Melbank. Los resultados obtenidos indican que estas parametrizaciones incrementan significativamente la capacidad del algoritmo para realizar la tarea de segmentación. En particular, la parametrización basada en la divergencia de Kullback-Leibler muestra el mejor desempeño, ya que no solo incrementa el número de límites correctamente detectados, sino que, además, disminuye la cantidad de puntos erróneamente insertados.

Los resultados demuestran que los coeficientes basados en la CME y la CMD proveen información valiosa para realizar la segmentación, la cual está relacionada con características acústicas que tienen en cuenta las transiciones entre fonemas. Esto podría estar relacionado al hecho de que estas medidas brindan información acerca los cambios en la dinámica del tracto vocal, lo cual es importante para llevar a cabo la segmentación fonética.

5.6. Modificación del Algoritmo de Segmentación

A partir de las experiencias realizadas y las características de las codificaciones basadas en la CMD, se propuso modificar el algoritmo descrito en la sección 5.2 eliminando la primer etapa, correspondiente a la obtención de la función de salto. En la figura 5.7 se observan las etapas del método de segmentación modificado. Los parámetros operacionales que regulan el proceso son ahora dos: β y γ.

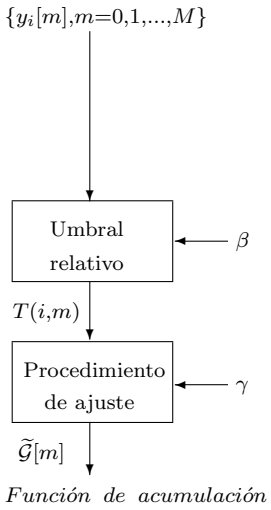

Figura 5.7: Esquema correspondiente a las etapas del algoritmo de segmentación automática de fonemas independiente del texto modificado, donde se ha eliminado la primera etapa de procesamiento, correspondiente a la función de salto. La versión original de este algoritmo fue descrita en la sección 5.2.

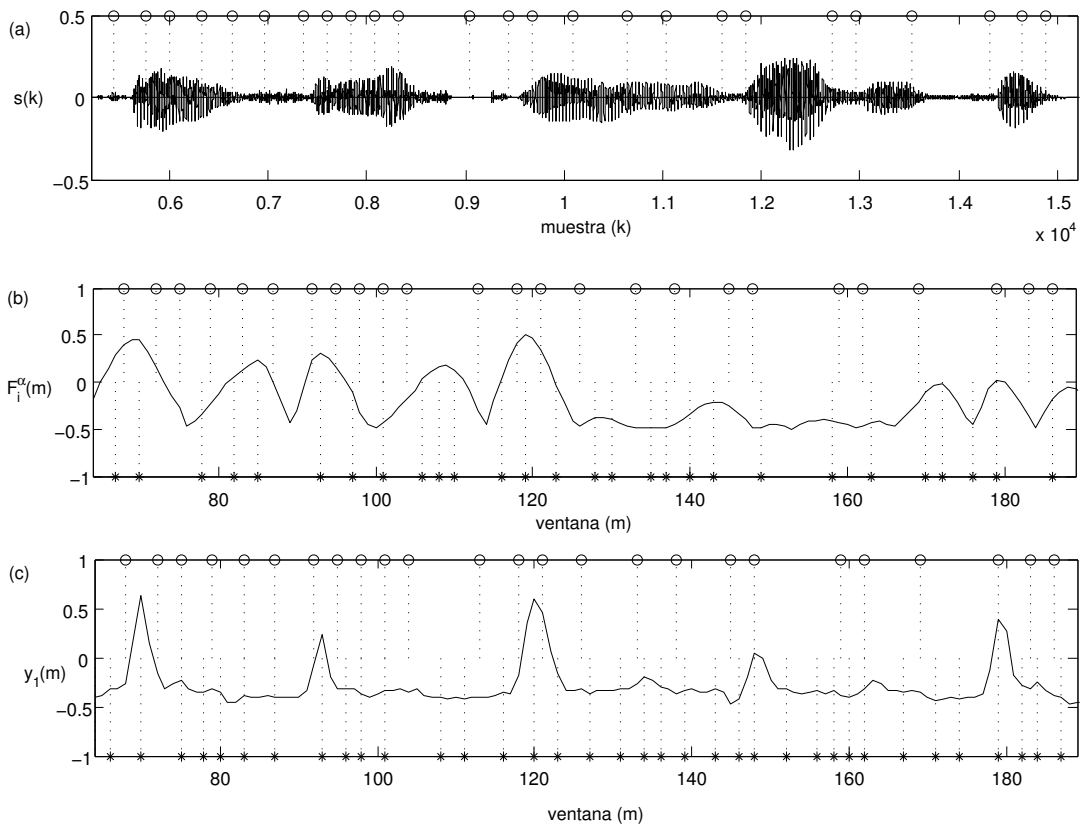

Figura 5.8: (a) Señal de habla etiquetada con Albayzin (líneas punteadas superiores con marcadores circulares). (b) Función de salto $\mathcal{F}_i^{\alpha}[m]$ correspondiente a la característica Melbank $y_1[m]$ de la señal mostrada en (a); las líneas punteadas superiores indican el etiquetado de referencia y las líneas punteadas inferiores con marcadores de asterisco muestran la segmentación obtenida usando la representación Melbank como entrada para el algoritmo original, descrito en la sección 5.2. (c) Característica $y_1[m]$ obtenida utilizando la codificación basada en la CMD de la señal mostrada en (a); las líneas punteadas superiores indican la segmentación de referencia y las inferiores, con marcadores de asterisco, la segmentación obtenida usando el algoritmo modificado. Los parámetros operacionales utilizados para el procedimiento de segmentación automática son: $\alpha = 6$, $\beta = 0{,}05$ y $\gamma = 3$.

En este caso, la parametrización basada en la CMD se utilizó no solo como entrada del algoritmo, sino que reemplaza la primera etapa del mismo. Esta propuesta fue motivada por el hecho de que la evolución temporal del vector de características ($\mathbf{y}_i$) obtenido mediante la CMD brinda una señal suficientemente abrupta, como se puede observar a partir de la figura 5.2 (a), por lo que la función de salto no mejora sustancialmente la misma.

A modo de ejemplo, en la figura 5.8 se muestra una parte de la señal de habla etiquetada, correspondiente a la frase: " ¿Cómo se llama el mar que baña Valencia?", junto con la evolución temporal de la función de salto ($\mathcal{F}_i^{\alpha}$, (5.1)) aplicada sobre una de las características de la codificación Melbank ($\mathbf{y}_1$) y la evolución de una de las características de la parametrización calculada utilizando la CMD (también $\mathbf{y}_1$). En

dicha figura se muestra, mediante las líneas punteadas inferiores con los marcadores asteriscos, los puntos de segmentación que se obtienen para ambas codificaciones. En el caso de la codificación basada en CMD, se utiliza el algoritmo modificado para obtener la segmentación. La líneas punteadas superiores, con marcadores circulares, indican la segmentación de referencia, provista por la base de datos. A partir de la figura 5.8 se puede observar que la parametrización basada en CMD, usada como entrada del algoritmo de segmentación modificado, detecta límites fonéticos que son ignorados en el algoritmo original, donde está presente la etapa de cálculo de la función de salto. Esto se observa, por ejemplo, en la tercer línea punteada inferior de la figura 5.8 (c). Por otro lado, el algoritmo modificado detecta adecuadamente algunos puntos erróneamente insertados en el algoritmo original, como por ejemplo, la sexta y séptima linea punteada inferior de la figura 5.8 (c). Esto sustenta la posibilidad de utilizar la codificación basada en CMD como entrada del algoritmo, reemplazando, además, la primera etapa del mismo.

5.6.1. Resultados del algoritmo modificado usando la parametrización basada en CMD

En la tabla 5.3 se muestran los índices PC y PI de la segmentación obtenida utilizando el algoritmo modificado. Estos valores se comparan con los obtenidos usando el algoritmo original con la codificación Melbank y la parametrización basada en CMD como entradas. Los parámetros operacionales utilizados para las pruebas fueron: $\gamma = 2$, 3 y 4; $\beta = 0{,}01$; 0,05; 0,1 y $\alpha = 6$ para el algoritmo original. Cabe destacar que no es necesario plantear un valor de α cuando se utiliza el algoritmo modificado. Los valores óptimos para los índices PC y PI son 100 % y 0 %, respectivamente.

Tabla 5.3: Porcentaje de límites fonéticos correctamente detectados (PC) y porcentaje de puntos erróneamente insertados (PI) obtenidos a partir del algoritmo modificado y el algoritmo original, usando como entradas, en este último caso, la codificación Melbank y la basada en CMD. Se evaluaron diferentes valores de parámetros operacionales γ y β. Para el algoritmo original se utilizó $\alpha = 6$.

Parámetros		Algoritmo modificado		Algoritmo original			
				Melbank		CMD	
γ	β	PC	PI	PC	PI	PC	PI
2	0.01	99.54	22.88	97.33	17.50	93.07	16.21
	0.05	97.81	14.10	95.44	16.10	91.27	9.26
	0.1	95.06	11.45	92.06	11.06	87.13	7.71
3	0.01	93.91	18.44	93.21	15.22	94.10	15.57
	0.05	94.62	11.22	90.44	12.01	91.47	9.83
	0.1	92.69	9.13	87.40	6.85	88.15	8.64
4	0.01	95.29	15.52	86.92	13.84	95.44	14.98
	0.05	94.17	9.25	83.55	9.21	93.05	9.98
	0.1	91.47	7.38	79.44	5.59	89.43	8.91

Se puede observar, a partir de la tabla 5.3, que para gran parte del conjunto de parámetros operacionales, el algoritmo modificado brinda mejores resultados, incrementando el número de límites correctamente detectados y disminuyendo la cantidad de puntos erróneamente insertados. Esto se evidencia, principalmente, para los valores de $\beta = 0{,}05$ y 0,1. Por otra parte, la mejora en el desempeño, provista por el algoritmo modificado, también se verifica si se comparan índices similares de PC y PI sin tener en cuenta los parámetros operacionales utilizados. Por ejemplo, con el algoritmo modificado se obtienen valores de PC=97.81 % y PI=14.10 % (segunda fila), mientras que el algoritmo original con la codificación Melbank muestra valores de PC=97.33 % y PI=17.50 % (primera fila). Asimismo, el algoritmo modificado brinda índices PC=94.62 % y PI=11.22 % (quinta fila) y el algoritmo original usando ahora la codificación basada en CMD da PC=94.10 % yPI=15.57 % (cuarta fila).

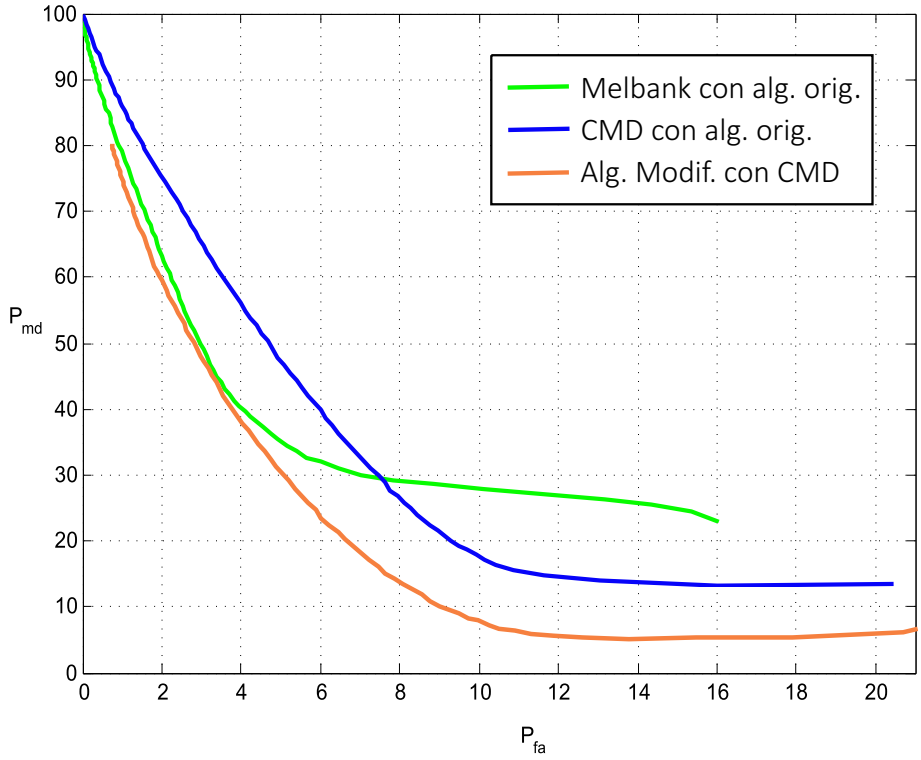

Figura 5.9: Curvas ROC para el algoritmo de segmentación usando el algoritmo modificado (línea naranja) y el algoritmo original con las codificaciones Melbank (línea verde) y la basada en la CMD (línea azul). Los parámetros operacionales son $\alpha = 6$, $\gamma = 3$ y β variando de 0 a 1 con incrementos de 0,01.

Por último, en la figura 5.9 se muestran las curvas ROC construidas con los índices P_{fa} y P_{md}, de manera tal de comparar el desempeño del algoritmo modificado y el algoritmo original con las dos codificaciones utilizadas para generar la tabla 5.3. En esta figura se utilizó el valor de $\gamma = 3$, con β variando entre 0 y 1 (en pasos de 0.01) y $\alpha = 6$ para el algoritmo original. Como puede observarse, a partir de la figura, dado que el algoritmo modificado permite reducir P_{fa}, reduciendo también P_{md}, se puede inferir que el mismo brinda mejores resultados que el algoritmo original utilizando la codificación Melbank o la basada en CMD.

Estos resultados indican que la parametrización basada en la CMD provee información relacionada con los cambios abruptos en la señal de habla, lo cual mejora el desempeño del algoritmo de segmentación. Asimismo, este tipo de codificación reduce la necesidad de procesamiento previo para realizar la segmentación.

5.7. Comentarios de cierre del capítulo

A partir de los resultados obtenidos utilizado las medidas de información multiresolución en un sistema de ASR, presentados en el capítulo 4, se propuso utilizar este tipo de procesamiento como parte del pre–procesamiento de la señal de habla para llevar a cabo la segmentación automática de fonemas, en un sistema independiente del texto. Para esta propuesta, la parametrización de la señal de habla se realizó usando sólo medidas basadas en la CME y la CMD.

Las medidas de información multiresolución fueron capaces de brindar información que permitió mejorar la detección de transiciones entre fonemas. Por otra parte, los resultados obtenidos sugieren que es posible eliminar la primera etapa del algoritmo de segmentación cuando se utiliza como codificación las medidas basadas en la entropía relativa Kullback-Leibler (CMD). Los resultados muestran que el rendimiento de esta propuesta mejora el desempeño de la segmentación alcanzada con algoritmo original. Asimismo, se reduce la cantidad de parámetros operacionales a configurar en el algoritmo.

A lo largo de este trabajo, las medidas de información multiresolución se utilizaron para realizar dos tipos de tareas diferentes, brindando información relevante acerca de los cambios en la dinámica de la señal de voz, aún en condiciones de ruido de la señal. En el capítulo siguiente, correspondiente a las conclusiones de esta tesis, se sintetizarán las actividades realizadas y los resultados obtenidos y se expondrán las principales conclusiones a las que ha permitido arribar este trabajo.

Conclusiones

En los capítulos anteriores se presentaron los diferentes aspectos de la emisión y percepción del habla humana, los cuales brindan muchas de las bases en las cuales se sustentan las técnicas de caracterización y modelado de la señal. En esta tesis se planteó el uso de procesamientos basados en medidas de información multiresolución para representar la señal de habla. Para obtener este tipo de representaciones se utilizaron medidas de información multiresolución basadas en entropías [38] y se propuso la utilización de medidas de divergencia [95] para evaluar la evolución temporal de la complejidad de los coeficientes de la CWT de la señal de voz. Para comprobar el desempeño de los mismos, se utilizaron en dos tipos de tareas: reconocimiento automático del habla y segmentación automática de fonemas.

Los resultados obtenidos para la tarea de reconocimiento automático del habla mostraron que estas medidas de información, calculadas en el plano tiempo-escala, proveen una mejora significativa en el desempeño del sistema. En particular, la CME mostró un desempeño similar a la codificación de referencia para la representación donde se tienen en cuenta los coeficientes correspondientes a las escalas bajas y altas (método PC_{SD}). Esto se evidenció, especialmente, para las señales mezcladas con ruido aditivo blanco. Este tipo de representación se basa en las características que presentan tanto la CME como la CMD de las señales con y sin ruido. En los escalogramas realizados se observa que las características propias de las señal de habla se manifiestan principalmente en las escalas más bajas, mientras que las estructuras pertenecientes al ruido se manifiestan en las escalas superiores.

Por otra parte, las medidas de información relativa mostraron los mejores resultados. La concatenación de dos coeficientes, como los correspondientes al primero y segundo componentes principales (método PC_{12}) y como los correspondientes al método PC_{SD} brindaron un desempeño que mejoró significativamente el reconocimiento, para ambos tipos de ruido, especialmente para las SNRs de 15, 10 y 5 dB.

Estos resultados hacen suponer que este tipo de procesamiento proporciona información valiosa al sistema de reconocimiento bajo condiciones de ruido. Esto puede estar vinculado con las características de la CME de brindar información acerca de los cambios en la dinámica subyacente de los sistemas no lineales. En el caso de la señal de voz, estos cambios están asociados a las variaciones que presenta el tracto vocal durante la emisión del habla. Sin duda, poder detectar estas variaciones es clave para realizar la tarea de reconocimiento. Por lo que, las medidas de información basadas en CME y CMD aportan mayor información al sistema, mejorando su desempeño [95, 96].

En la tarea de segmentación automática de fonemas se utilizaron codificaciones de la señal de habla basadas íntegramente en la CME y la CMD. Se encontró que la parametrización basada en la divergencia de Kullback-Leibler incrementa la habili-

dad del algoritmo para llevar a cabo la segmentación, ya que incrementa el número de límites correctamente detectados, disminuyendo la cantidad de puntos erróneamente insertados. Los resultados obtenidos demuestran que este tipo de medidas proveen información relacionada con características acústicas que tienen en cuenta las transiciones entre fonemas, lo cual permite mejorar la segmentación [97, 98].

Tanto para la aplicación de reconocimiento de habla como para la tarea de segmentación de fonemas, los resultados obtenidos muestran que es posible utilizar parametrizaciones de habla basadas en medidas de información multiresolución y que las mismas brindan información relevante acerca de los cambios en la dinámica de la señal de voz, que mejoran el desempeño de estos sistemas.

Bibliografía

[1] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(4):745–777, 2014.

[2] George Saon and Jen-Tzung Chien. Large-vocabulary continuous speech recognition systems: A look at some recent advances. *Signal Processing Magazine, IEEE*, 29(6):18–33, 2012.

[3] V Radha and C Vimala. A review on speech recognition challenges and approaches. *doaj. org*, 2(1):1–7, 2012.

[4] K.K. Paliwal, J.G. Lyons, S. So, A.P. Stark, and K.K. Wójcicki. Comparative evaluation of speech enhancement methods for robust automatic speech recognition. In *Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on*, pages 1–5, dec. 2010.

[5] D. O'Shaughnessy. Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, 41:2965–2979, 2008.

[6] Ryszard Makowski and Robert Hossa. Automatic speech signal segmentation based on the innovation adaptive filter. In *International Journal of Applied Mathematics and Computer Science*, volume 24, pages 259–270, jun. 2014.

[7] Jon Ander Gómez and Marcos Calvo. Improvements on automatic speech segmentation at the phonetic level. In César San Martin and Sang-Woon Kim, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 7042 of *Lecture Notes in Computer Science*, pages 557–564. Springer Berlin Heidelberg, 2011.

[8] Kris Demuynck and Tom Laureys. A comparison of different approaches to automatic speech segmentation. In Petr Sojka, Ivan Kopeček, and Karel Pala, editors, *Text, Speech and Dialogue*, volume 2448 of *Lecture Notes in Computer Science*, pages 277–284. Springer Berlin Heidelberg, 2006.

[9] Kun Han and DeLiang Wang. Towards generalizing classification based speech separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(1):168–177, 2013.

[10] C. Martínez, J. Goddard, D. Milone, and H. Rufiner. Bioinspired sparse spectro-temporal representation of speech for robust classification. *Computer Speech & Language*, 26(5):336–348, 2012.

[11] Heikki Kallasjoki, Jort F Gemmeke, and Kalle J Palomaki. Estimating uncertainty to improve exemplar-based feature enhancement for noise robust speech recognition. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(2):368–380, 2014.

[12] Ning Ma, Jon Barker, Heidi Christensen, and Phil Green. Combining speech fragment decoding and adaptive noise floor modeling. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(3):818–827, 2012.

[13] Jort F Gemmeke, Tuomas Virtanen, and Antti Hurmalainen. Exemplar-based sparse representations for noise robust automatic speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2067–2080, 2011.

[14] Leandro Di Persia, Diego Milone, Hugo Leonardo Rufiner, and Masuzo Yanagida. Perceptual evaluation of blind source separation for robust speech recognition. *Signal Processing*, 88(10):2578 – 2583, 2008.

[15] Roger K Moore. Spoken language processing: Where do we go from here? In *Your Virtual Butler*, pages 119–133. Springer, 2013.

[16] Ben Gold, Nelson Morgan, and Dan Ellis. *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons, 2011.

[17] Katharina von Kriegstein, David R. R. Smith, Roy D. Patterson, Stefan J. Kiebel, and Timothy D. Griffiths. How the human brain recognizes speech in the context of changing speakers. *The Journal of Neuroscience*, 30(2):629–638, 2010.

[18] Aitzol Ezeiza, Karmele López de Ipiña, Carmen Hernández, and Nora Barroso. Enhancing the feature extraction process for automatic speech recognition with fractal dimensions. *Cognitive Computation*, 5(4):545–550, 2013.

[19] Youssef Zouhir and Kaïs Ouni. Speech signals parameterization based on auditory filter modeling. In *Advances in Nonlinear Speech Processing*, pages 60–66. Springer, 2013.

[20] Carlos M Travieso, Jesús B Alonso, Juan Rafael Orozco-Arroyave, Jordi Solé-Casals, and Esteve Gallego-Jutglà. Automatic detection of laryngeal pathologies in running speech based on the hmm transformation of the nonlinear dynamics. In *Advances in Nonlinear Speech Processing*, pages 136–143. Springer, 2013.

[21] Daniel May, Tao Ma, Sundar Srinivasan, Georgios Lazarou, and Joseph Picone. Continuous speech recognition using nonlinear dynamic invariants. *submitted to the Association for Computational Linguistics: Human Language Technologies, Columbus, Ohio, USA*, 2008.

[22] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, New Jersey, 1993.

[23] Elizabeth D. Casserly and David B. Pisoni. Speech perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(5):629–647, 2010.

[24] H. Rufiner, C. Martínez, D. Milone, and J. Goddard. Auditory cortical representations of speech signals for phoneme classification. *Lecture Notes in Artificial Inteligence 4827: MICAI 2007*, 332:1004–1014, November 2007.

[25] J. Rafiee, M.A. Rafiee, N. Prause, and M.P. Schoen. Wavelet basis functions in biomedical signal processing. *Expert Systems with Applications*, 38(5):6190 – 6201, 2011.

[26] R. F. Leonarduzzi, G. Schlotthauer, and M. E. Torres. Short-time multifractal analysis: application to biological signals. *Journal of Physics*, 313(012012), 2011. ISSN: 1742-6596.

[27] Nicholas Stergiou and Leslie M Decker. Human movement variability, nonlinear dynamics, and pathology: is there a connection? *Human movement science*, 30(5):869–888, 2011.

[28] F. Lestussi, H. L. Rufiner, L. Di Persia, and D. H. Milone. Sparse coding for apnea-hipopnea detection. In *Anales de la XIV Reunión de Procesamiento de la Información y Control*, Oro Verde, Argentina, nov 2011.

[29] Kuang Chua Chua, Vinod Chandran, U. Rajendra Acharya, and Choo Min Lim. Application of higher order statistics/spectra in biomedical signals - a review. *Medical Engineering & Physics*, 32(7):679 – 689, 2010.

[30] D Puthankattil Subha, Paul K Joseph, Rajendra Acharya, and Choo Min Lim. Eeg signal analysis: a survey. *Journal of medical systems*, 34(2):195–212, 2010.

[31] Hasan Ocak. Automatic detection of epileptic seizures in eeg using discrete wavelet transform and approximate entropy. *Expert Systems with Applications*, 36(2):2027–2036, 2009.

[32] Anisoara Paraschiv-Ionescu and Kamiar Aminian. Nonlinear analysis of physiological time series. In *Advanced biosignal processing*, pages 307–333. Springer, 2009.

[33] Alain de Cheveigné and Jonathan Z. Simon. Denoising based on spatial filtering. *Journal of Neuroscience Methods*, 171(2):331 – 339, 2008.

[34] Arnaud Delorme, Terrence Sejnowski, and Scott Makeig. Enhanced detection of artifacts in eeg data using higher-order statistics and independent component analysis. *NeuroImage*, 34(4):1443 – 1449, 2007.

[35] M.B.I. Reaz, M.S. Hussain, and F. Mohd-Yasin. Techniques of emg signal analysis: detection, processing, classification and applications. *Biological Procedures Online*, 8(1):11–35, 2006.

[36] Christopher J James and Christian W Hesse. Independent component analysis for biomedical signals. *Physiological Measurement*, 26(1):R15, 2005.

[37] A. Cichocki. Blind signal processing methods for analyzing multichannel brain signals. *International Journal of Bioelectromagnetism*, 6(1):1–21, 2004.

[38] M. E. Torres, M. M. Añino, L. G. Gamero, and M. A. Gemignani. Automatic detection of slight changes in nonlinear dynamical systems using multiresolution entropy tools. *Int. J. of Bifurcations and Chaos*, 11(4):967–981, 2001.

[39] Soham Sarkar and Swagatam Das. Multilevel image thresholding based on 2d histogram and maximum tsallis entropy - a differential evolution approach. *Image Processing, IEEE Transactions on*, 22(12):4788–4797, 2013.

[40] G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner, and M. E. Torres. Evaluation of a new model for vowels synthesis with perturbations in acoustic parameters. *Latin American Applied Research*, 43(3), jul 2013.

[41] Gaurav Sharma, Sibt ul Hussain, and Frédéric Jurie. Local higher-order statistics (lhs) for texture categorization and facial analysis. In *Computer Vision–ECCV 2012*, pages 1–12. Springer, 2012.

[42] Michael Elad, Mario A.T. Figueiredo, and Yi Ma. On the role of sparse and redundant representations in image processing. *Proceedings of the IEEE*, 98(6):972–982, 2010.

[43] Stephen M Smith, Peter T Fox, Karla L Miller, David C Glahn, P Mickle Fox, Clare E Mackay, Nicola Filippini, Kate E Watkins, Roberto Toro, Angela R Laird, et al. Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, 106(31):13040–13045, 2009.

[44] Alfred M Bruckstein, David L Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review*, 51(1):34–81, 2009.

[45] Hara Stefanou, Thanasis Margaritis, Dimitris Kafetzopoulos, Konstantinos Marias, and Panagiotis Tsakalides. Microarray image denoising using a two-stage multiresolution technique. In *Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on*, pages 383–389. IEEE, 2007.

[46] J. Bobin, Y. Moudden, J. Starck, and M. Elad. Morphological diversity and source separation. *IEEE Signal Processing Letters*, 13(7):409–412, 2006.

[47] Yinpeng Jin, Elsa Angelini, and Andrew Laine. Wavelets in medical image processing: denoising, segmentation, and registration. In *Handbook of biomedical image analysis*, pages 305–358. Springer, 2005.

[48] J. Starck, M. Elad, and D. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Transaction On Image Processing*, 14(10):1570–1582, 2005.

[49] P. Gruber, F. Theis, A. Tomé, and E. Lang. Automatic denoising using local independent component analysis. In *Engineering of Intelligent Systems (EIS)*, pages 1–6, March 2004.

[50] S.V. Sunitha, T.M. Shashidhar, and S.R. Pavithra. Feature extraction of nonlinear chaotic characteristics for pathological voice recognition. 2014.

[51] Patricia Henríquez, Jesús B Alonso, Miguel A Ferrer, Carlos M Travieso, and Juan R Orozco-Arroyave. Nonlinear dynamics characterization of emotional speech. *Neurocomputing*, 132:126–135, 2014.

[52] Vahid Khanagha, Khalid Daoudi, Oriol Pont, Hussein Yahia, and Antonio Turiel. Non-linear speech representation based on local predictability exponents. *Neurocomputing*, 132:136–141, 2014.

[53] Chengli Sun, Qi Zhu, and Minghua Wan. A novel speech enhancement method based on constrained low-rank and sparse matrix decomposition. *Speech Communication*, 60:44–55, 2014.

[54] Leandro D. Vignolo, Diego H. Milone, and Hugo L. Rufiner. Genetic wavelet packets for speech recognition. *Expert Systems with Applications*, 40(6):2350–2359, 2013.

[55] Marius Crisan. New aspects of phoneme synthesis based on chaotic modeling. In *Instrumentation, Measurement, Circuits and Systems*, pages 605–614. Springer, 2012.

[56] Amparo Marti, Máximo Cobos, and José J. Lopez. Evaluating the influence of source separation methods in robust automatic speech recognition with a specific cocktail-party training. In *Audio Engineering Society Convention 132*, Apr 2012.

[57] VL Lajish, RK Sunil Kumar, and P Vivek. Speaker identification using a non-linear speech model and ann. *International Journal of Advanced Information Technology*, 2(5):15–24, 2012.

[58] Raghunath S Holambe and Mangesh S Deshpande. *Advances in Non-Linear Modeling for Speech Processing*. Springer Science & Business Media, 2012.

[59] C. E. Martínez, J. Goddard, L. Di Persia, D. H. Milone, and H. L. Rufiner. Denoising audio signals in the non-negative auditory cortical domain. In *12th Argentine Symposium on Technology (40th JAIIO)*, Córdoba, Argentina, aug 2011.

[60] Jong Won Shin, Joon-Hyuk Chang, and Nam Soo Kim. Voice activity detection based on statistical models and machine learning approaches. *Computer Speech & Language*, 24(3):515–530, 2010.

[61] Zhang Jun-chang and Zhao Li. A speech denoising method based on improved emd. In *Multimedia and Signal Processing (CMSP), 2011 International Conference on*, volume 2, pages 305–309. IEEE, 2011.

[62] Christian D Sigg, Tomas Dikk, and Joachim M Buhmann. Speech enhancement using generative dictionary learning. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(6):1698–1712, 2012.

[63] Zhou Yan. A new sparse representation algorithm for speech denoising. In *3rd International Conference on Computer Science and Service System*. Atlantis Press, 2014.

[64] Preety D Swami, Rupali Sharma, and Alok Jain. Speech enhancement by noise driven adaptation of perceptual scales and thresholds of continuous wavelet transform coefficients. *Speech Communication*, 2015.

[65] Yu Shao and Chip-Hong Chang. Bayesian separation with sparsity promotion in perceptual wavelet domain for speech enhancement and hybrid speech recognition. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 41(2):284–293, 2011.

[66] Khaled Daqrouq, Ibrahim N Abu-Isbeih, Omar Daoud, and Emad Khalaf. An investigation of speech enhancement using wavelet filtering method. *International Journal of Speech Technology*, 13(2):101–115, 2010.

[67] Hong-Kwang Jeff Kuo and Yuqing Gao. Maximum entropy direct models for speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(3):873–881, 2006.

[68] H. L. Rufiner, M. E. Torres, L. Gamero, and D. H. Milone. Introducing complexity measures in nonlinear physiological signal: application to robust speech recognition. *Physica A*, 332:496–508, 2004.

[69] S. Ghorshi, S. Vaseghi, and Q. Yan. Cross-entropic comparison of formants of british, australian and american english accents. *Speech Communication*, 50(7):564–579, July 2008.

[70] J. Chien, H. Hsieh, and S. Furui. A new mutual information measure for independent component alalysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, pages 1817–1820, 2008.

[71] B. Bigi, Y. Huang, and R. De Mori. Vocabulary and language model adaptation using information retrieval. In *International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, South Korea, 2004. International Speech Communication Association (ISCA).

[72] J. Weeds, D. Weir, and D. McCarthy. Characterizing measures of lexical distributional similarity. In *Proceedings of the 20th International Conference of Computational Linguistics, COLING-2004*, pages 1015–1021, Geneva, Switzerland, 2004. Morgan Kaufmann.

[73] L. Lee. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics 2001*, pages 65–72, Key West, Florida, 2001. Morgan Kaufmann.

[74] X. Li, H. Liu, Y. Zheng, and B. Xu. Robust speech endpoint detection based on improved adaptive band-partitioning spectral entropy. In *Bio-Inspired Computational Intelligence and Applications*, pages 36–45. Springer Berlin / Heidelberg, 2007.

[75] Aik Ming Toh, Roberto Togneri, and Sven Nordholm. Spectral entropy as speech features for speech recognition. *Proceedings of PEECS*, 1, 2005.

[76] Mohammed Ajmal, Azadeh Kushki, and Konstantinos N Plataniotis. Time-compression of speech in information talks using spectral entropy. In *Image Analysis for Multimedia Interactive Services, 2007. WIAMIS'07. Eighth International Workshop on*, pages 80–80. IEEE, 2007.

[77] M. Torres, L. Gamero, and E. D'Attellis. Pattern detection in EEG using multirresolution entropy. *Lat. Am. Appl. Res.*, 53:53–57, 1995.

[78] L. G. Gamero, A. Plastino, and M. E. Torres. Wavelet analysis and nonlinear dynamics in a non extensive setting. *Physica A*, 246:487–509, 1997.

[79] H. Torres, J. Gurlekian, H. Rufiner, and M. Torres. Self-organizing map clustering based on continuous multiresolution entropy. *Physica A: Statistical Mechanics and its Applications*, 361(1):337–354, February 2006.

[80] Brian N Pasley, Stephen V David, Nima Mesgarani, Adeen Flinker, Shihab A Shamma, Nathan E Crone, Robert T Knight, and Edward F Chang. Reconstructing speech from human auditory cortex. *PLoS biology*, 10(1):e1001251, 2012.

[81] Claus Christiansen, Michael Syskind Pedersen, and Torsten Dau. Prediction of speech intelligibility based on an auditory preprocessing model. *Speech Communication*, 52(7 - 8):678 – 692, 2010.

[82] Rungsun Munkong and Biing-Hwang Juang. Auditory perception and cognition. *Signal Processing Magazine, IEEE*, 25(3):98–117, 2008.

[83] H. Asari, B. Pearlmutter, and A. Zador. Sparse representations for the cocktail party problem. *Journal of Neuroscience*, 26(28):7477–7490, 2006.

[84] M. Lewicki. Effcient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, 2002.

[85] H. Rufiner, J. Goddard, L. Rocha, and M. Torres. Statistical method for sparse coding of speech including a linear predictive model. *Physica A*, 367:231–251, 2006.

[86] H. Rufiner, L. Rocha, and J. Goddard. Preserving acoustic cues in speech . In *Proc. of the 2nd Joint Meeting of the IEEE Engineering in Medicine and Biology Society and the Biomedical Engineering Society EMBS-BMES2002*, volume 1, pages 288–289, Houston, Texas, October 2002.

[87] R Stern and Nelson Morgan. Hearing is believing: Biologically-inspired feature extraction for robust automatic speech recognition. *IEEE signal processing magazine*, 29(34-43):170, 2012.

[88] Ramin Pichevar, Hossein Najaf-Zadeh, Louis Thibault, and Hassan Lahdili. Auditory-inspired sparse representation of audio signals. *Speech Communication*, 53(5):643–657, 2011.

[89] Yang Shao, Soundararajan Srinivasan, Zhaozhang Jin, and DeLiang Wang. A computational auditory scene analysis system for speech segregation and robust speech recognition. *Computer Speech & Language*, 24(1):77–93, 2010.

[90] Daniel M Rasetshwane, J Robert Boston, C-C Li, John D Durrant, and Gregory Genna. Enhancement of speech intelligibility using transients extracted by wavelet packets. In *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA'09. IEEE Workshop on*, pages 173–176. IEEE, 2009.

[91] U. Yapanel and J. Hansen. A new perceptually motivated mvdr-based acoustic front-end (pmvdr) for robust automatic speech recognition. *Speech Communication*, 50(2):142–152, February 2008.

[92] M. Holmberg, D. Gelbart, and W. Hemmert. Automatic speech recognition with an adaptation model motivated by auditory processing. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):43–49, 2006.

[93] N. Morgan, H. Bourlard, and H. Hermansky. *Automatic Speech Recognition: An Auditory Perspective*, volume 18, pages 309–338. Springer New York, 2004.

[94] M. M. Añino, M. E. Torres, and G. Schlotthauer. Slight parameter changes detection in biological models: A multiresolution approach. *Physica A*, 324(3–4):645–664, 2003.

[95] M. Torres, L. Rufiner, D. Milone, and A. Cherniz. Multiresolution information measures applied to speech recognition. *Physica A*, 385(1):319–332, 2007.

[96] M. E. Torres, H. L. Rufiner, D. H. Milone, and A. S. Cherniz. Comparison between temporal and time-scale information measures applied to speech recognition. *WSEAS Transactions on signal Processing*, 9(2):1153–1159, 2006.

[97] A. Cherniz, M. E. Torres, H. Rufiner, and A. Esposito. Time-scale information measures for text-independent phone segmentation. In *RPIC 2009, XIII Reunión de Trabajo en Procesamiento de la Información y Control*, Septiembre 2009.

[98] A. Cherniz, M. E. Torres, H. Rufiner, and A. Esposito. Multiresolution analysis applied to text-independent phone segmentation. *J. Phys.: Conf. Ser.*, 90, 2007.

[99] Real Academia Española.

[100] Enrique Obediente. *Fonética y fonología*. Universidad Los Andes, 1998.

[101] HL Rufiner. *Análisis y representación de la voz mediante técnicas no convencionales*. PhD thesis, Tesis de Doctorado en Ingeniería, Universidad de Buenos Aires, Argentina, 2005.

[102] J. Deller, J. Proakis, and J. Hansen. *Discrete Time Processing of Speech Signals*. Macmillan Publishing, New York, 1993.

[103] S. Greenberg. The ears have it: The auditory basis of speech perception. In *Proceedings of the International Congress of Phonetic Sciences*, volume 3, pages 34–41, 1995.

[104] E.R. Kandel, J.H. Schwartz, and T.M. Jessell. *Principios de Neurociencia*. McGraw-Hill, 2001.

[105] Antonio Quilis. *Tratado de Fonología y Fonética Españolas.* Biblioteca Románica Hispánica. Editorial Gredos, Madrid, 1993.

[106] Horacio E. Cingolani and Alberto B. Houssay. *Fisiología Humana*, volume 4. El Ateneo, Buenos Aires, 6 edition, 1988.

[107] Ana María Borzone Manrique. *Manual de Fonética Acústica.* Hachette, Buenos Aires, 1980.

[108] Emilio Alarcos Llorach. *Gramática de la Lengua Española.* Real Academia Española. Colección Nebrija y Bello. Editorial Espasa Calpe, Madrid, 1999.

[109] Oswald Ducrot and Tzvetan Todorov. *Diccionario enciclopédico de las ciencias del lenguaje.* Siglo Veintiuno, Mexico, 10 edition, 1984.

[110] Jacob Benesty, M Mohan Sondhi, and Yiteng Huang. *Springer handbook of speech processing.* Springer Science & Business Media, 2008.

[111] Kenneth N Stevens. *Acoustic Phonetics*, volume 30. MIT Press, 2000.

[112] Jack Mullen, David M Howard, and Damian T Murphy. Waveguide physical modeling of vocal tract acoustics: flexible formant bandwidth control from increased model dimensionality. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(3):964–971, 2006.

[113] William J Hardcastle, John Laver, and Fiona E Gibbon. *The handbook of phonetic sciences.* John Wiley & Sons, 2010.

[114] J. Gurlekian, L. Colantoni, H. Torres, A. Rincon, A. Moreno, and J. Mariño. Database for an automatic speech recognition system for argentine spanish. In Buneman & Liberman Bird, editor, *Proceedings of the IRCS Workshop on Linguistic Databases, Workshop on Linguistic Databases*, pages 92–98, Filadelfia, USA, December 2001.

[115] Rafael Monroy-Casas. *Aspectos fonéticos de las vocales españolas.* LibrosEnRed, 2004.

[116] David Pisoni and Robert Remez. *The handbook of speech perception.* John Wiley & Sons, 2008.

[117] NeurOreille. Viaje al mundo de la audición, 2014.

[118] P. Mason. *Medical Neurobiology.* OUP USA, 2011.

[119] G. Ehret and R. Romand. *The Central Auditory System.* Oxford University Press, 1997.

[120] G. Von Békésy. *Experiments in Hearing.* McGraw-Hill, New York, 1960.

[121] Anders Fridberger, Jacques Boutet de Monvel, Jiefu Zheng, Ning Hu, Yuan Zou, Tianying Ren, and Alfred Nuttall. Organ of corti potentials and the motion of the basilar membrane. *The Journal of neuroscience*, 24(45):10057–10063, 2004.

[122] Renato Nobili, Fabio Mammano, and Jonathan Ashmore. How well do we understand the cochlea? *Trends in neurosciences*, 21(4):159–167, 1998.

[123] D. Purves, G. Augustine, D. Fitzpatrick, L. Katz, A. LaMantia, and J. McNamara. *Invitación a la Neurociencia*. Editorial Médica Panamericana, 2001.

[124] Edward F Evans. Basic physiology of the hearing mechanism. In *Audio Engineering Society Conference: 12th International Conference: The Perception of Reproduced Sound*. Audio Engineering Society, 1993.

[125] M Charles Liberman. The cochlear frequency map for the cat: Labeling auditory-nerve fibers of known characteristic frequency. *The Journal of the Acoustical Society of America*, 72(5):1441–1449, 1982.

[126] Kiang, Watenabe, and Thomas. *Discharge Patterns of Single Fibers in the Catś Auditory Nerve*. MIT Press, Cambridge, MA, 1965.

[127] John F Brugge. Patterns of organization in auditory cortex. *The Journal of the Acoustical Society of America*, 78(1):353–359, 1985.

[128] Richard A Reale and Thomas J Imig. Tonotopic organization in auditory cortex of the cat. *Journal of Comparative Neurology*, 192(2):265–291, 1980.

[129] Don H Johnson. The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. *The Journal of the Acoustical Society of America*, 68(4):1115–1122, 1980.

[130] Jerzy E Rose, John F Brugge, David J Anderson, and Joseph E Hind. Phase-locked response to low-frequency tones in single auditory nerve fibers of the squirrel monkey. *Journal of neurophysiology*, 30(4):769–793, 1967.

[131] Robert Gilkey and Timothy R Anderson. *Binaural and spatial hearing in real and virtual environments*. Psychology Press, 2014.

[132] JL Goldstein and P Srulovicz. Auditory-nerve spike intervals as an adequate basis for aural frequency measurement. *Psychophysics and physiology of hearing*, pages 337–346, 1977.

[133] Bertrand Delgutte and Nelson YS Kiang. Speech coding in the auditory nerve: I. vowel-like sounds. *The Journal of the Acoustical Society of America*, 75(3):866–878, 1984.

[134] Richard A Reale and C Daniel Geisler. Auditory-nerve fiber encoding of two-tone approximations to steady-state vowels. *The Journal of the Acoustical Society of America*, 67(3):891–902, 1980.

[135] Eric D Young and Murray B Sachs. Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. *The Journal of the Acoustical Society of America*, 66(5):1381–1403, 1979.

[136] Bertrand Delgutte. Physiological models for basic auditory percepts. In *Auditory computation*, pages 157–220. Springer, 1996.

[137] M.I. Miller and M.B. Sachs. Representation of stop consonants in the discharge patterns of auditory-nerve fibers. *Journal of the Acoustic Society of America*, 74(2):502–517, August 1983.

[138] L. H. Carney and C. D. Geisler. A temporal analysis of auditory-nerve fiber responses to spoken stop consonant-vowel syllables. *Journal of the Acoustic Society of America*, 79(6):1896–1914, June 1986.

[139] Bertrand Delgutte. Auditory neural processing of speech. *The handbook of phonetic sciences*, pages 507–538, 1997.

[140] Secker-Walker and Searle. Time-domain analysis of auditory-nerve-fiber firing rates. *Journal of the Acoustic Society of America*, 88:637–642, 1990.

[141] Jan Schnupp, Israel Nelken, and Andrew King. *Auditory neuroscience: Making sense of sound*. MIT Press, 2011.

[142] A.C. Morris and J.M. Pardo. Phoneme transition detection and broad classification using a simple model based on the function of onset detector cells found in the cochlear nucleus. In *Proceedings of the Eurospeech'95*, pages 115–118, 1995.

[143] N. Suga. What does single-unit analysis in the auditory cortex tell us about information processing in the auditory system? In P. Rakic and W. Singer, editors, *Neurobiology of the neocortex*. John Wiley & Sons, 1988.

[144] P. Belin, R.J. Zatorre, R. Hoge, A.C. Evans, and B. Pike. Event-related fMRI of the auditory cortex. *NeuroImage*, 10:417–429, 1999.

[145] N. Castañeda, J.M. Cornejo, and P. Granados. Neuroanatomical representation of P1 component of the Long Latency Auditory Evoked Potential as a frequency function of a tone burst in normal hearing young children. In *XVIII IERASG Biennial Symposium*, Puerto de la Cruz, Tenerife, Canary Islands, Spain, June 2003.

[146] Gunnar Fant. *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations*, volume 2. Walter de Gruyter, 1971.

[147] J.M. Hillenbrand, M.J. Clark, and T.M. Nearey. Effects of consonant environment on vowel formant patterns. *Journal of the Acoustic Society of America*, 109(2):748–763, February 2001.

[148] J. E. Diaz Verdejo, A. M. Peinado, A. J. Rubio, E. Segarra, N. Prieto, and F. Casacuberta. Albayzin: a task-oriented spanish speech corpus. In *Proceedings of the First International Conference on Language Resources and Evaluation*, volume 1, pages 497–502, Granada, May 1998. European Language Resources Association.

[149] L.R. Rabiner and R.W. Schafer. *Introduction to Digital Speech Processing*. Foundations and Trends in Technology. Now Publishers, 2007.

[150] Alan V Oppenheim and Ronald W Schafer. From frequency to quefrency: A history of the cepstrum. *Signal Processing Magazine, IEEE*, 21(5):95–106, 2004.

[151] Robert B Randall. A history of cepstrum analysis and its application to mechanical problems. In *International Conference*, pages 29–30.

[152] Donald D Greenwood. Auditory masking and the critical band. *The journal of the acoustical society of America*, 33(4):484–502, 1961.

[153] Joseph W Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1215–1247, 1993.

[154] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[155] Hynek Hermansky and Nelson Morgan. Rasta processing of speech. *Speech and Audio Processing, IEEE Transactions on*, 2(4):578–589, 1994.

[156] Morten L. Jepsen, Stephan D. Ewert, and Torsten Dau. A computational model of human auditory signal processing and perception. *The Journal of the Acoustical Society of America*, 124(1):422–438, 2008.

[157] Hendrik Purwins, Benjamin Blankertz, and Klaus Obermayer. Computing auditory perception. In *Proceedings of the Cognition and Perception Issues in Computer Music*, 2000.

[158] Bertrand Delgutte. Physiological models for basic auditory percepts. In *Auditory computation*, pages 157–220. Springer, 1996.

[159] Oded Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 2(1):115–132, 1994.

[160] Akram Aldroubi and Michael Unser. *Wavelets in medicine and biology*. CRC press, 1996.

[161] O.A. Rosso, M.T. Martin, A. Figliola, K. Keller, and A. Plastino. Eeg analysis using wavelet-based information tools. *Journal of Neuroscience Methods*, 153(2):163–182, 2006.

[162] Mehmet Rahmi Canal. Comparison of wavelet and short time fourier transform methods in the analysis of emg signals. *Journal of medical systems*, 34(1):91–94, 2010.

[163] L. G. Gamero, J. Vila, and F. Palacios. Wavelet transform analysis of heart rate variability during myocardial ischemia. *Biological Engineering and Computing*, 40(1):72–78, 2002.

[164] Pietro Lio. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics*, 19(1):2–9, 2003.

[165] Mohamed N. Nounou, Hazem N. Nounou, Nader Meskin, Aniruddha Datta, and Edward R. Dougherty. Multiscale denoising of biological data: A comparative analysis. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 9(5):1539–1545, 2012.

[166] George Tzanetakis, Georg Essl, and Perry Cook. Audio analysis using the discrete wavelet transform. In *Proc. Conf. in Acoustics and Music Theory Applications*, 2001.

[167] A. Kandaswamy, C. Sathish Kumar, Rm.Pl. Ramanathan, S. Jayaraman, and N. Malmurugan. Neural classification of lung sounds using wavelet coefficients. *Computers in Biology and Medicine*, 34(6):523–537, 2004.

[168] Jean-Luc Starck, Fionn Murtagh, and Jalal M Fadili. *Sparse image and signal processing: wavelets, curvelets, morphological diversity*. Cambridge University Press, 2010.

[169] P. S. Addison, J. Walker, and R. C. Guido. Time–frequency analysis of biosignals. *IEEE Engineering in Medicine and Biology*, 28(5):14–29, 2009.

[170] S.G. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 2nd edition, September 1999.

[171] MATLAB®. *Wavelet Toolbox for Use with MATLAB*. The MathWorks, Inc., http://www.mathworks.com, 1 edition, March 1996. Users Guide.

[172] C.E. Shannon. A mathematical theory of comunication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[173] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[174] Entropy journal. Special issue: Entropy in genetics and computational biology, Julio 2010.

[175] G. Schlotthauer, M. E. Torres, and H. L. Rufiner. Determinación de la frecuencia fundamental de la voz basada en descomposición modal empírica por conjuntos y entropías. In *XIII Reunión de Trabajo en Procesamiento de la Información y Control (RPIC 2009)*, pages 387–392, 2009.

[176] R.A. Baldwin. Use of maximum entropy modeling in wildlife research. *Entropy*, 11:854–866, 2009.

[177] François Bavaud. Information theory, relative entropy and statistics. In *Formal Theories of Information*, pages 54–78. Springer, 2009.

[178] Inder Jeet Taneja. Statistical aspects of divergence measures. *Journal of statistical planning and inference*, 16:137–145, 1987.

[179] Aman Ullah. Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planning and Inference*, 49(1):137–162, 1996.

[180] Joshua S. Richman and J. Randall Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. of Physiol. Heart Circ. Physiol.*, 278:2039–2049, 2000.

[181] Xu-Sheng Zhang, Rob J. Roy, and Erik Weber Jensen. Eeg complexity as a measure of depth of anesthesia for patients. *IEEE Transactions on Biomedical Engineering*, 48(12):1424–1433, December 2001.

[182] Madalena Costa, Ary L. Goldberger, and C. K. Peng. Multiscale entropy analysis of biological signals. *Physical Review E*, 71:1–18, February 2005.

[183] Prasanta Kumar Ghosh, Andreas Tsiartas, and Shrikanth Narayanan. Robust voice activity detection using long-term signal variability. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(3):600–613, 2011.

[184] Bing-Fei Wu and Kun-Ching Wang. Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments. *IEEE Transactions on Speech and Audio Processing*, 13(5):762–775, September 2005.

[185] Wang Xu, Qi Ding, and Bing xi Wang. A speech endpoint detector based on space-energy-entropy. *Acoustical Science and Technology*, 25(1):54–57, 2004.

[186] Kim Weaver, Khurram Waheed, and Fathi M. Salem. An entropy based robust speech boundary detection algorithm for realistic noisy environments. In *Proceedings of the International Joint Conference on Neural Networks*, volume 1, pages 680–685, Portland, Oregon, USA, July 2003. IEEE.

[187] C. Tsallis. Some comments on boltzmann-gibbs statistical mechanics. *Chaos, Solitons and Fractals*, 6:539, 1995. and references therein.

[188] C. Tsallis, A. R. Plastino, and W. Zheng. Power-law sensitivity to initial conditions - new entropic representation. *Chaos, Solitons and Fractals*, 8(6):885–891, 1997.

[189] M. L. Lyra and C. Tsallis. Nonextensivity and multifractality in low-dimensional dissipative systems. *Physical Review Letters*, 80(1):53–56, 1998.

[190] Shachi Sharma. Power law and tsallis entropy: Network traffic and applications. *Chaos, Nonlinearity, Complexity*, pages 162–178, 2006.

[191] M. E. Torres and L. G. Gamero. Relative complexity changes in time series using information measures. *Physica A*, 286(3-4):457–473, 2000.

[192] M. E. Torres. *El procesamiento de señales ligadas a problemas no lineales.* PhD thesis, Universidad Nacional de Rosario - Argentine, 1999. (Math. D. Thesis).

[193] Ivo Grosse, Pedro Bernaola-Galván, Pedro Carpena, Ramón Román-Roldán, Jose Oliver, and H Eugene Stanley. Analysis of symbolic sequences using the jensen-shannon divergence. *Physical Review E*, 65(4):041905, 2002.

[194] Pedro W Lamberti and Ana P Majtey. Non-logarithmic jensen–shannon divergence. *Physica A: Statistical Mechanics and its Applications*, 329(1):81–90, 2003.

[195] M. Akay. *Detection and Estimation Methods for Biomedical Signals*. Academic Press, San Diego, CA, 1996.

[196] M. Cooke and D. Ellis. The auditory organization of speech and other sources in listeners and computational models. *Speech Comunication*, 35(3-4):141–177, 2001.

[197] J. C. Junqua and J. P. Haton. *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer Academic Publishers, Boston, 1996.

[198] O. Viikki. Editorial: Noise robust ASR. Speech Comunication, 2001.

[199] Gillian M Davis. *Noise reduction in speech applications*, volume 7. CRC Press, 2002.

[200] Richard P. Lippmann. Speech recognition by machines and humans. *Speech Comunication*, 22(19):1–15, 1997.

[201] Jürgen Tchorz. *Auditory-based signal processing for noise suppression and robust speech recognition*. BIS Verlag, 2000.

[202] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[203] Mark Gales and Steve Young. The application of hidden markov models in speech recognition. *Foundations and trends in signal processing*, 1(3):195–304, 2008.

[204] Steve Young, Dan Kershaw, Julian Odell, Dave Ollason, Valtcho Valtchev, and Phil Woodland. *HMM Toolkit*. Cambridge University, http://htk.eng.cam.ac.uk, htk v3.3 edition, 2000.

[205] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, pages 164–171, 1970.

[206] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. On the estimation of discount parameters for language model smoothing. In *Interspeech*, pages 1433–1436, 2011.

[207] Stanley F Chen and Ronald Rosenfeld. A survey of smoothing techniques for me models. *IEEE transactions on Speech and Audio Processing*, 8(1):37–50, 2000.

[208] Sven C Martin, Christoph Hamacher, Jörg Liermann, Frank Wessel, and Hermann Ney. Assessment of smoothing methods and complex stochastic language modeling. In *EUROSPEECH*, 1999.

[209] Steve Young. A review of large-vocabulary continuous-speech. *Signal Processing Magazine, IEEE*, 13(5):45, 1996.

[210] Stephen John Young, NH Russell, and JHS Thornton. *Token passing: a simple conceptual model for connected speech recognition systems*. Cambridge University Engineering Department Cambridge, UK, 1989.

[211] Stefan Ortmanns, Hermann Ney, and Xavier Aubert. A word graph algorithm for large vocabulary continuous speech recognition. *Computer Speech & Language*, 11(1):43–72, 1997.

[212] Xavier Aubert and Hermann Ney. Large vocabulary continuous speech recognition using word graphs. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 49–52. IEEE, 1995.

[213] J.J. Odell, V. Valtchev, Philip C. Woodland, and Steve J. Young. A one pass decoder design for large vocabulary recognition. In *Proceedings of the workshop on Human Language Technology*, pages 405–410. Association for Computational Linguistics, 1994.

[214] Frederick Jelinek. *Statistical methods for speech recognition*. MIT press, 1997.

[215] Mehryar Mohri, Fernando Pereira, and Michael Riley. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88, 2002.

[216] Douglas B Paul. Algorithms for an optimal a* search and linearizing the search in the stack decoder. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 693–696. IEEE, 1991.

[217] Andrew Varga and Herman JM Steeneken. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech communication*, 12(3):247–251, 1993.

[218] Steven B Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, 1980.

[219] Sadaoki Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(1):52–59, 1986.

[220] D. Michie, D. J. Spiegelhalter, and C. C. Taylor. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, University College, London, 1994.

[221] John-Paul Hosom. Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication*, 51(4):352–368, 2009.

[222] Diana Binnenpoorte, Simo Goddijn, and Catia Cucchiarini. How to improve human and machine transcriptions of spontaneous speech. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[223] Manish Sharma and Richard Mammone. Blind speech segmentation: Automatic segmentation of speech without linguistic knowledge. In *Proc. of Fourth International Conference on Spoken Language–ICSLP 96*, volume 2, pages 1237–1240, Philadelphia, USA, 1996. IEEE.

[224] Steven Greenberg. Strategies for automatic multi-tier annotation of spoken language corpora. In *Proceedings of the 8th European Conference on Speech Communication and Technology–Eurospeech '03*, pages 45–48, Geneva, Switzerland, 2003. ISCA Archive.

[225] Dac-Thang Hoang and Hsiao-Chuan Wang. Blind phone segmentation based on spectral change detection using legendre polynomial approximation. *The Journal of the Acoustical Society of America*, 137(2):797–805, 2015.

[226] Jiahong Yuan, Neville Ryant, Mark Liberman, Andreas Stolcke, Vikramjit Mitra, and Wen Wang. Automatic phonetic segmentation using boundary models. In *INTERSPEECH*, pages 2306–2310. Citeseer, 2013.

[227] V. Stouten, K. Demuynck, and H. Van hamme. Automatically learning the units of speech by non-negative matrix factorisation. In *8th Annual Conference of the International Speech Communication Association INTERSPEECH 2007*, pages 1937–1940, Antwerp, Bélgica, 2007. ISCA.

[228] D. H. Milone, J. J. Merelo, and H. L. Rufiner. Evolutionary algorithm for speech segmentation. In *Proc. of the 2002 IEEE World Congress on Evolutionary Computation*, volume 2, pages 1115–1120. IEEE Computer Society, 2002. Paper No. 7270.

[229] E. Vidal and A. Marzal. *A Review and New Approaches for Automatic Segmentation of Continuous Speech Signals*, pages 43–53. Elsevier, New-York, 1990.

[230] Kåre Sjölander. An hmm-based system for automatic segmentation and alignment of speech. In *Proceedings of Fonetik*, volume 2003, pages 93–96. Citeseer, 2003.

[231] Fabio Brugnara, Daniele Falavigna, and Maurizio Omologo. Automatic segmentation and labeling of speech based on hidden markov models. *Speech Communication*, 12(4):357–370, 1993.

[232] G. Almpanidis and C. Kotropoulos. Phoneme segmentation using the generalized gamma distribution and small sample bayesian information criterion. *Speech Communication*, 50(1):38–55, 2008.

[233] Sarah Hoffmann and Beat Pfister. Fully automatic segmentation for prosodic speech corpora. In *Interspeech*, pages 1389–1392, 2010.

[234] Doroteo Torre Toledano, Luis A Hernández Gómez, and Luis Villarrubia Grande. Automatic phonetic segmentation. *Speech and Audio Processing, IEEE Transactions on*, 11(6):617–625, 2003.

[235] A. Esposito and G. Aversano. Text independent methods for speech segmentation. In Gérard Chollet, Anna Esposito, Marcos Faundez-Zanuy, and Maria Marinaro, editors, *Nonlinear Speech Modeling And Applications: Advanced Lectures and Revised Selected Papers*, pages 261–290. Springer, Berlin, Germany, 2005.

[236] Andreas Tsiartas, Theodora Chaspari, Nassos Katsamanis, Prasanta Kumar Ghosh, Ming Li, Maarten Van Segbroeck, Alexandros Potamianos, and Shrikanth Narayanan. Multi-band long-term signal variability features for robust voice activity detection. In *INTERSPEECH*, pages 718–722, 2013.

[237] Sherry Y. Zhao and Nelson Morgan. Multi-stream spectro-temporal features for robust speech recognition. In *INTERSPEECH*, pages 898–901. Citeseer, 2008.

[238] M. Molla, K. Hirose, and N. Minematsu. Robust speaker identification system using multi-band dominant features with empirical mode decomposition. In *Computer and Information Technology, 2008. ICCIT 2008. 10th International Conference on*, pages 1–5, 2007.

[239] G. Evangelopoulos and P. Maragos. Multiband modulation energy tracking for noisy speech detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6), 2006.

[240] Khalid Daoudi, Dominique Fohr, and Christophe Antoine. Dynamic bayesian networks for multi-band automatic speech recognition. *Computer Speech & Language*, 17(2-3):263–285, 2003.

[241] N. Mirghafori. *A Multi-Band Approach to Automatic Speech Recognition*. PhD thesis, Berkeley, CA, 1998. Reprinted as ICSI Technical Report, TR-99-04.

[242] Carl D Mitchell, Mary P Harper, and Leah H Jamieson. Using explicit segmentation to improve hmm phone recognition. In *icassp*, pages 229–232. IEEE, 1995.

[243] Ta-Hsin Li and Jerry D Gibson. Speech analysis and segmentation by parametric filtering. *Speech and Audio Processing, IEEE Transactions on*, 4(3):203–213, 1996.

[244] Gunnar Ahlbom, Frédéric Bimbot, and Gérard Chollet. Modeling spectral speech transitions using temporal decomposition techniques. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87.*, volume 12, pages 13–16. IEEE, 1987.

[245] B. Petek, O. Andersen, and P. Dalsgaard. On the robust automatic segmentation of spontaneous speech. In *Proc. of the 4th International Conference on Spoken Language Processing - ICSLP 96*, pages 913–916, Philadelphia, USA, 1996. ISCA Archive.