UNIVERSIDAD DE GRANADA

TESIS DOCTORAL

Ecualización de histogramas en el procesado robusto de voz

Autora:
Dª. Luz García Martínez

Director:
Dr. D. Jose Carlos Segura Luna

DEPARTAMENTO DE TEORÍA DE LA SEÑAL, TELEMÁTICA Y COMUNICACIONES

Granada, Noviembre 2007

Editor: Editorial de la Universidad de Granada
Autor: Luz Garcia Martinez
D.L.: Gr. 2809 - 2007
ISBN: 978-84-338-4698-3

A mi tía Luz, que me enseñó a leer y a escribir...

... A Petra, por la otra tesis...

... A los valientes

Agradecimientos

Quiero dar las GRACIAS sinceras a mi director de tesis Jose Carlos Segura, por su ayuda, su paciencia y su generosidad. GRACIAS también a Carmen Benítez, Javier Ramírez, Angel de la Torre y Antonio Rubio por la ayuda científica y la cercanía personal.

GRACIAS a Luis. GRACIAS a Cristina y a Manu, porque los tres han puesto el umbral altísimo en el reconocimiento y la percepción del habla en humanos. Me han escuchado, han detectado emociones, han filtrado ruidos blancos, rosas y no tan rosas, y han revolucionado los niveles semántico y pragmático de la comprensión del lenguaje y los estados mentales en humanos. Para cuándo una máquina que alcance el 10 % de su precisión en reconocimiento.

GRACIAS a Anabel por los consejos y la oreja paciente.

GRACIAS a mis tíos Miguel y Carmen por estar cerca.

GRACIAS a *la* Liébana, mi *teleco del tipo dos* favorita.

GRACIAS *al despacho 21* por las *absurdeces* reconfortantes.

GRACIAS a las Isabelinas, guapas y finas, por el trote cochinero tan honrosamente mantenido.

Por último, GRACIAS a mi familia. GRACIAS a mis padres Queti y Juanfra porque sin ellos no estaría aquí en ninguno de los sentidos, y GRACIAS a mi hermano Nacho por estar. Dicen en cierta película que "*Familia es un lugar en el que nunca se deja atrás a nadie*". Palabras clave para búsquedas: nicho, rescoldo, llegada, estación, destino, momento, querencia, manantial, rito, estirpe, garbo, orgullo, pertenencia del ser.

Granada, 7 de Noviembre del 2007
Luz García Martínez

Resumen

Esta tesis se centra en una técnica de robustecimiento de las características cepstrales MFCC usadas en el reconocimiento automático del habla: **la Ecualización de Histogramas, HEQ**. La Ecualización de Histogramas es una transformación no lineal que se aplica al vector de características cepstrales en el *front-end* del reconocedor automático del habla. Su objetivo es transformar dichas características a un dominio (dominio ecualizado) invariante ante las distorsiones que el ruido provoca en las distribución de densidad de probabilidad. HEQ se puede situar dentro de un grupo de técnicas de robustecimiento del reconocimiento automático del habla, definidas como **técnicas de encuadre estadístico** cuya filosofía es normalizar parámetros estadísticos de las características (ya sea la media, la varianza, algunos momentos de orden superior o la función de densidad de probabilidad) para eliminar la distorsión provocada por el ruido. En sus orígenes HEQ era una técnica de procesado de imágenes, pero su bajo coste computacional, la simplicidad de su planteamiento que no necesita modelos del ruido que se combate, así como la versatilidad de sus aplicaciones (debida también al hecho de no presuponer ninguna característica sobre las distorsiones que elimina), hicieron atractiva su aplicación en el procesado de la señal acústica que se ha llevado a cabo en los últimos seis años.

El trabajo realizado en esta tesis analiza las prestaciones y peculiaridades de HEQ y sus limitaciones como técnica de robustecimiento. Estas limitaciones se deben fundamentalmente al hecho de que la calidad del encuadre estadístico depende en gran medida de la obtención de unas estadísticas fiables de la frase que se ecualiza. Cuando las frases son cortas se producen dos efectos no deseables: la fiabilidad de las estadísticas disminuye y el porcentaje de voz y silencio que tenga la frase pasa a ser un factor influyente en la transformación distorsionando con ello la información acústica. Existe además otra limitación de HEQ: por motivos de viabilidad

computacional y algorítmica, los coeficientes cepstrales son considerados independientes entre sí. Esta hipótesis implica una pérdida de información en el proceso de ecualización, al no capturar la información acústica dada por las correlaciones entre coeficientes.

Para afrontar estas limitaciones, el trabajo propone un algoritmo de **ecualización paramétrica** llamado **PEQ** en el que se definen dos clases que se ecualizan por separado: una clase para la tramas de voz y otra para las tramas de silencio, y una expresión paramétrica (Gausiana definida por su media y varianza) de las densidades de probabilidad de ambas clases. Los experimentos realizados muestran que PEQ mejora de manera eficaz la tasa de reconocimiento. Esta mejora es justificable ya que la expresión paramétrica de las densidades de probabilidad refleja las estadísticas de las frases de manera más fiable cuando hay pocos datos. Por otra parte los criterios para definir las clases de voz y datos capturan la correlación del coeficiente cepstral C0 (cuyo valor medio es usado como umbral de decisión al clasificar las tramas como pertenecientes a una u otra clase) con el resto de coeficientes.

Por último, este trabajo se hace eco de la conveniencia de capturar la información temporal en el vector de características acústicas como indican los métodos de parametrización basados en modelos perceptuales o el uso frecuente de las características dinámicas de los coeficientes MFCC. Para ello se propone un algoritmo llamado **TES (Suavizado temporal)** que captura para cada coeficiente cepstral la correlación entre tramas consecutivas. Esta correlación es portadora de información acústica, que se distorsiona a causa del ruido o del *mismatch* entre los datos de entrenamiento y *test*. TES normaliza dichas correlaciones inter-trama haciéndolas iguales a las de los datos del entorno limpio.

Índice general

Índice de tablas

Índice de figuras

Parte I

Introducción

Capítulo 1

Introducción

1.1. Contexto del trabajo

El reconocimiento automático de voz nace en los años 50 como respuesta científico-tecnológica al deseo del hombre de interaccionar oralmente con las máquinas. En un primer momento se fundamenta en los principios de la fonética acústica y se limita al reconocimiento de palabras aisladas de un vocabulario muy reducido, para un locutor exclusivo y utilizando para ello dispositivos electrónicos.

En la década de los 70 se empieza a tener en cuenta que el conocimiento sintáctico, semántico y contextual son fuentes de información muy útiles. Se utilizan microprocesadores y se aplican las técnicas de programación dinámica al reconocimiento de palabras conectadas. Empiezan a aparecer reconocedores independientes de locutor para tareas muy concretas.

Pero es en los años 80 cuando se da el giro metodológico fundamental con el modelado estadístico y el uso de los modelos ocultos de Markov o HMMs. A partir de ese momento, el reconocimiento de habla continua ha mejorado, aumentándose el tamaño de los vocabularios, diversificándose las aplicaciones y enfrentándose a situaciones cada vez más reales, en las que los locutores y las condiciones del entorno de reconocimiento difieren de los que se han utilizado para entrenar el reconocedor.

En la actualidad las tasas de reconocimiento más optimistas están en un orden de magnitud por debajo de las que serían atribuibles al ser humano y el reconocimiento automático de voz continua es un campo de trabajo al que la comunidad científica dedica un esfuerzo importante.

En lo que a su aplicación práctica se refiere, el reconocimiento automático del habla empezó como el particular reto científico de emular el comportamiento humano con máquinas, siendo objeto de interés para un público y unas aplicaciones bastante específicas y limitadas. La situación actual no podría ser más antagónica, habiendo sido denominada *tercera revolución industrial* [113]. Los avances tecnológicos previos y originados por el nacimiento de la *Sociedad de la Información* con las *Tecnologías de la Información y las comunicaciones* (TICs) asociadas a ella, han provocado una revolución en la demanda de interfaces usuario-máquina lo más amigables y transparentes posibles para el usuario. El diálogo hombre-máquina aparece en este escenario como mecanismo óptimo y natural de relación de los habitantes de esa Sociedad de la Información con *sus* TICs desde varios puntos de vista:

- Desde un punto de vista científico-tecnológico, la inteligencia artificial y en particular el reto de emular la comunicación oral humana sigue siendo un estímulo atractivo. Los avances en capacidad de computación y potencia del software y hardware, así como en el conocimiento del comportamiento humano reinventan continuamente el camino.

- Desde un punto de vista económico, hay que señalar tres factores :

  i) La tecnología se ha convertido en un bien de consumo.

  ii) El mercado ha cambiado sus protocolos y la compartición de información, y la globalización de los intercambios que conlleva el comercio electrónico exigen la presencia de las Tecnologías de la Información. La demanda de reconocimiento automático del habla crece de manera interesante en los sistemas de telecomu-

nicación, en los sistemas de control y en los sistemas de entrada de datos y de acceso a bases de datos.

iii) Las personas de edad avanzada en las poblaciones del primer mundo y los emigrantes de países con menor penetración de las TICs son dos sectores del mercado tecnológico con un alto potencial como consumidores. Estos dos sectores demandan interfaces universales, intuitivas y amigables.

- Desde un punto de vista social, la Sociedad de la Información ha creado oportunidades muy interesantes y al mismo tiempo ha generado la llamada *brecha digital* entre los usuarios de las TICs y los no usuarios de las TICs. El desarrollo de tecnologías amigables casi transparentes para el usuario, es una ayuda para combatir esta brecha[1].

El análisis anterior enmarca la motivación de este trabajo científico en reconocimiento automático de la voz. El reconocimiento automático debe ser intrínsecamente robusto. Es decir, debe dar las máximas prestaciones posibles en las condiciones más adversas imaginables. Las condiciones adversas para un reconocedor se definen como las diferencias o desajustes que puedan existir entre los datos con los que ha sido entrenado, y los datos que debe reconocer. El robustecimiento de un reconocedor de voz se puede definir como la aportación de mecanismos que lo hagan menos vulnerable a esos desajustes de las condiciones de entrenamiento y evaluación. Existe una importante línea científica para el estudio de estrategias de robustecimiento que atacan las debilidades del reconocedor desde puntos de vista diferentes. El objetivo general de esta tesis es crear una base sólida de conocimiento sobre el reconocimiento automático del habla y las es-

[1]Declaración de principios de la Cumbre Mundial de la Sociedad de la información, Ginebra, 2003. *Reconocemos que la educación, el conocimiento, la información y la comunicación son esenciales para el progreso, la iniciativa y el bienestar de los seres humanos. Es más, las tecnologías de la información y las comunicaciones (TIC) tienen inmensas repercusiones en prácticamente todos los aspectos de nuestras vidas. El rápido progreso de estas tecnologías brinda oportunidades sin precedentes para alcanzar niveles más elevados de desarrollo. La capacidad de las TIC para reducir muchos obstáculos tradicionales, especialmente el tiempo y la distancia, posibilitan, por primera vez en la historia, el uso del potencial de estas tecnologías en beneficio de millones de personas en todo el mundo.*

trategias de robustecimiento existentes. Con esta base se persigue analizar en profundidad la Ecualización de Histogramas como transformación no lineal de las características acústicas que robustece el reconocimiento. Una vez hecho esto, se proponen ciertas mejoras y derivaciones de la técnica original que serán analizadas en el entorno experimental del trabajo y que tienen como finalidad mejorar las tasas de reconocimiento automático del habla exitoso en entornos ruidosos.

1.2. Organización de la Tesis

Esta memoria se ha estructurado en tres partes que se detallan a continuación, con los capítulos que contiene cada una de ellas:

Parte I Introducción Introduce todo el documento y contiene el presente capítulo.

Capítulo 1. Introducción. Es el capítulo actual, en el que se recoge el contexto y la motivación de la investigación documentada en esta tesis, y se presenta la estructura del documento.

Parte II Estado de la cuestión Expone el estado de la cuestión de las materias en las que esta investigación está centrada o aquellas que son necesarias para entender su desarrollo. Se divide a su vez en varios capítulos.

Capítulo 2. Reconocimiento automático del habla. Este capítulo hace un encuadre científico-tecnológico del reconocimiento automático del habla, recogiendo los conceptos básicos, detallando las etapas que lo componen y los métodos de implementación existentes.

Capítulo 3. Robustecimiento del RAH. Después de describir el efecto del ruido en la señal de voz se hace un repaso exhaustivo de las estrategias de robustecimiento del reconocimiento existentes.

Capítulo 4. Objetivos de esta tesis. Se concretan cuáles han sido los objetivos abordados y las contribuciones que se espera obtener con este trabajo.

Parte III Propuesta Esta parte contiene la propuesta que se ha elaborado para satisfacer los objetivos planteados en la investigación. Los capítulos que contiene se describen a continuación.

Capítulo 5. Descripción del entorno de trabajo. Este capítulo describe el sistema de reconocimiento y las parametrizaciones utilizadas para evaluar las técnicas propuestas, así como las 3 bases de datos sobre las que se evalúan: AURORA2, AURORA4 y HIWIRE.

Capítulo 6. Ecualización de Histogramas. Se analiza la técnica de Ecualización de Histogramas *HEQ*, estudiando su fundamento teórico y detalles de implementación, para definir experimentos con las tres bases de datos del entorno de evaluación.

Capítulo 7. Ecualización paramétrica de Histogramas. Este capítulo propone el algoritmo de ecualización paramétrica *PEQ*, lo sitúa dentro del conjunto de técnicas de ecualización existentes y lo compara con HEQ basándose en los resultados de la experimentación con ambos algoritmos.

Capítulo 8. Normalización de las características estáticas y dinámicas. Propuesta de inclusión en el *front-end* del reconocedor de un filtro que introduce información temporal normalizada en el vector de características robustas para el reconocimiento. Análisis de los efectos de inclusión de dicho filtro *TES* mediante experimentación con las bases de datos del sistema.

Parte IV Evaluación y conclusiones Esta parte recoge la evaluación de los diferentes componentes de la propuesta, así como las conclusiones de esta investigación. Se divide a su vez en dos capítulos.

Capítulo 9. Evaluación. Se evalúa si la propuesta realizada cumple los objetivos fijados en este trabajo y se aportan los resultados del uso de los algoritmos propuestos en proyecto europeo HI-WIRE [1] en cuyo marco se ha realizado este trabajo.

Capítulo 10. Conclusiones. Las conclusiones a las que ha dado lugar esta tesis se recogen en este capítulo, así como un resumen de las aportaciones y de las líneas de trabajo futuras basadas en o relacionadas con esta investigación.

Parte V Bibliografía y anexos La última parte del documento contiene las referencias utilizadas en éste, una relación de las publicaciones e información adicional para la mejor comprensión de lo aquí expuesto.

Parte II

Estado de la cuestión

Capítulo 2

Reconocimiento Automático del Habla

Este capítulo hace un encuadre científico-tecnológico del reconocimiento automático del habla, definiendo los tipos de reconocedores, sus objetivos y las etapas de procesado de la señal de voz necesarias para implementarlos. Las diferentes posibilidades metodológicas son analizadas para cada una de dichas etapas. Se termina con una definición de los criterios de evaluación cuantitativa en los sistemas de RAH.

2.1. Planteamiento del problema: RAH

2.1.1. La comunicación oral

El proceso de comunicación oral es uno de los comportamientos humanos que más ampliamente ha sido estudiado por disciplinas como la biología, la física y la lingüística. La figura 2.1 muestra la visión esquemática de los elementos que intervienen en dicho proceso, y del flujo que lo origina: en la mente del emisor se crea un mensaje que, por medio de impulsos nerviosos a los nervios motores que activan los músculos vocales, se traduce en un discurso de palabras transmitido a través de una señal acústica. La señal acústica es recibida por el receptor que lleva a cabo el proceso inverso: el movimiento de la membrana basilar en el oído del receptor se convierte en impulso eléctrico que es transmitido al cerebro mediante los

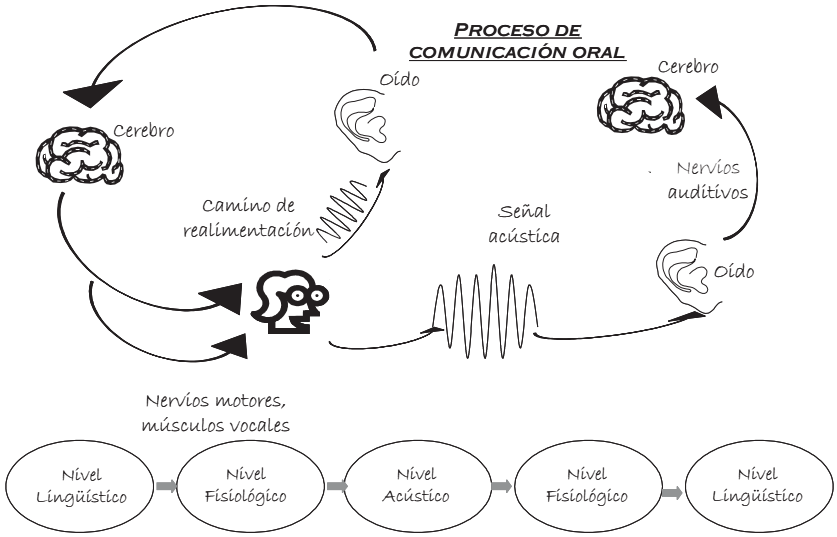


Figura 2.1. Esquema del proceso de comunicación oral

nervios auditivos. En el cerebro del receptor se produce el análisis y comprensión del mensaje.

2.1.2. El reconocimiento en la comunicación entre humanos

El objetivo del reconocimiento automático del habla (*RAH*) es imitar el proceso de reconocimiento que lleva a cabo el receptor en la comunicación oral. Hay varios niveles de reconocimiento en dicho proceso humano, y los diferentes sistemas de reconocimiento automático implementan todos, algunos o sólo los más básicos dependiendo de cuáles sean su aplicación y complejidad. Se pueden distinguir ocho niveles de reconocimiento en orden de complejidad ascendente:

- NIVEL ACÚSTICO: la señal acústica analógica que ha enviado el emisor es recibida y traducida a un conjunto de rasgos relevantes no redundantes. En la comunicación oral, este reconocimiento se hace en el oído. Hay cuatro operaciones incluidas en este nivel, que son total

o parcialmente implementadas para automatizarlo en el RAH:

Parametrización: la señal analógica se transforma en una señal numérica que pueda ser tratada por la máquina digital en la que se hace el reconocimiento. Hay varios métodos de parametrización en el dominio del tiempo y de la frecuencia, que dan lugar a diferentes parámetros de caracterización.

Segmentación: determina como separar la señal analógica continua en una cadena de sonidos cuya sucesión es la señal en el tiempo. Se lleva a cabo con métodos basados en las curvas de variación de la energía o de variabilidad de la señal.

Extracción de la información relevante: se busca retener solo aquellos datos que proporcionen información útil para el reconocimiento como pueden ser los espectros de los instantes de mayor estabilidad o de los instantes de transición

Información relativa a la prosodia: estudia la variación del armónico fundamental de la voz, variación de la intensidad, y el ritmo.

- NIVEL FONÉTICO: la secuencia de información relevante obtenida en el nivel acústico es traducida a una secuencia de fonemas
- NIVEL FONOLÓGICO: los fonemas de la lengua que hacen que el contenido fonético de las palabras se modifique en una articulación rápida o por una sucesión de términos léxicos son analizados. Las variedades dialectales son también tratadas.
- NIVEL LÉXICO: se identifican las palabras de la lengua en la que se produce la comunicación.
- NIVEL SINTÁCTICO: se detectan las reglas gramaticales que permiten describir y analizar el lenguaje, y que relacionan las palabras reconocidas a nivel léxico.
- NIVEL SEMÁNTICO: analiza el sentido de las palabras, buscando la comprensión del mensaje y eliminando las interpretaciones que no

tengan sentido. Es el nivel de conocimiento de las palabras que da un diccionario de la lengua.

- NIVEL PRAGMÁTICO: estudia el sentido del mensaje recibido teniendo en cuenta el contexto de su aplicación. Reconoce la información que viene determinada por la situación en la que se produce la comunicación.

- NIVEL PROSÓDICO: interviene de manera paralela al resto de niveles, sin formar parte de una estructura piramidal como los demás. Este nivel detecta la información que el mensaje comunica mediante los modos de pronunciación: palabras pronunciadas con cierto nivel de insistencia para ponerlas en relieve, fronteras entre grupos de palabras, naturaleza interrogativa o declarativa de una frase, etc.

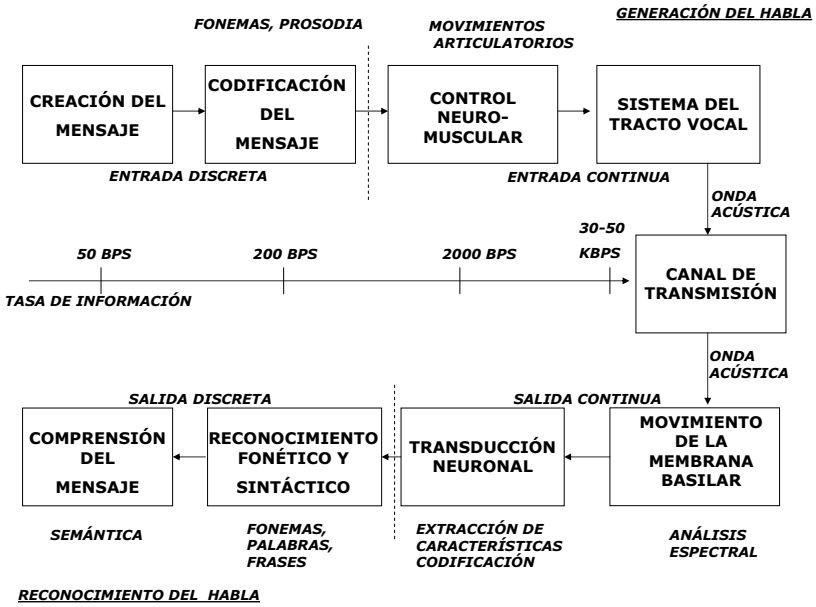


Figura 2.2. Otra esquematización del proceso de comunicación oral

La figura 2.2 muestra otra esquematización del modelo de producción y percepción de la voz propuesto por Rabiner y Levinson [103] en la que al igual que en la figura 2.1, se pueden identificar los niveles de producción y reconocimiento que se dan en la comunicación oral y las tareas necesarias para automatizar la comunicación, ya sea en la parte de producción de voz (síntesis automática de voz) o en la parte de reconocimiento (reconocimiento automático del habla).

En particular, la parte del reconocimiento del habla se automatiza implementando, en su totalidad o parcialmente, los ocho niveles de reconocimiento mencionados según la complejidad y necesidades del sistema de reconocimiento. Las posibilidades son muchas. En la tabla 2.1, Cole [18] recopila una visión global de las variables que definen un sistema de reconocimiento automático de habla y el rango de valores que pueden tomar:

- Hay reconocedores de palabras aisladas, de palabras conectadas y de habla continua, lo que supone un orden creciente de complejidad del reconocedor que tiene que delimitar palabras y frases.

- El habla puede ser no espontánea (leída o dirigida mediante un diálogo de opciones), o puede ser espontánea con el consiguiente incremento de la dificultad.

- El reconocedor de voz puede ser además dependiente de locutor teniendo que discernir la información acústica para un solo hablante, puede ser adaptado al locutor, multilocutor o independiente de locutor con lo que deberá filtrar las distorsiones acústicas debidas a las peculiaridades del hablante.

- Los fines específicos o generales del reconocedor y la perplejidad y tamaño del vocabulario que reconoce son características que aumentan la dificultad o simpleza de la tarea de reconocimiento.

- La distorsión acústica debida al ruido de canal y al ruido aditivo que acompañe a la voz incrementa la dificultad de la tarea de reconocimiento.

PARÁMETRO	RANGO
Forma de hablar	Palabras aisladas ↔ Habla continua
Estilo del habla	Texto leído ↔ Habla espontánea
Adaptación	Dependiente de locutor ↔ Independiente de locutor
Tamaño del vocabulario	Pequeño (<20 palabras) ↔ Grande (> 20.000 palabras)
Modelo de lenguaje	Estados finitos ↔ Dependiente de contexto
Perplejidad	Pequeña (<10) ↔ Grande(>100)
SNR	Alta (>30) ↔ Baja (<10)
Transductor	Micrófono de cancelación de eco ↔ Teléfono

Tabla 2.1. Parámetros que caracterizan el sistema de reconocimiento

Independientemente de las características que se acaban de describir, los bloques conceptuales necesarios para implementar un reconocedor son los que aparecen en la figura 2.3, que serán analizados en profundidad en el resto de este capítulo:

- **Datos de entrenamiento**. Son la información con la cual el sistema se entrena. Es un conjunto de datos que debe ser los suficientemente equilibrado como para que el reconocedor *aprenda* a reconocer ese vocabulario.

- **Bloque de parametrización**. Las entradas de señal de voz pasarán por una etapa en la que se extraerán sus características representativas antes de ser clasificadas.

- **Bloque de reconocimiento**. La clasificación de los parámetros se hará usando los datos de entrada y las referencias con las que cuenta el sistema: el aprendizaje de los datos de entrenamiento, y las referencias acústicas, léxicas y de lenguaje.

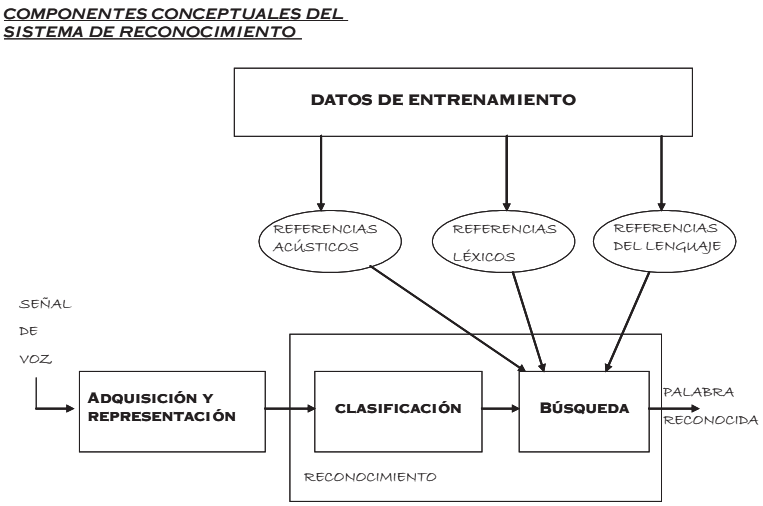


Figura 2.3. Componentes conceptuales de un sistema de reconocimiento

2.2. Parametrización de la señal de voz

El objetivo de la fase de parametrización en el proceso de reconocimiento, es la extracción de la información relevante de la señal acústica analógica, eliminando las redundancias y la información asociada a las fuentes de variabilidad que tiene la misma. La información relevante será aquella que permita:

- Diferenciar unos fonemas de otros. Los fonemas están caracterizados por:

  i) La envolvente espectral del fonema, determinada por los formantes que lo componen. Los formantes se definen como las frecuencias de resonancia del tracto vocal para cada fonema.

  ii) El tipo de excitación que los produce. Las vocales y consonantes sonoras estás generadas mediante una excitación periódica. La frecuencia fundamental de la excitación es también una carac-

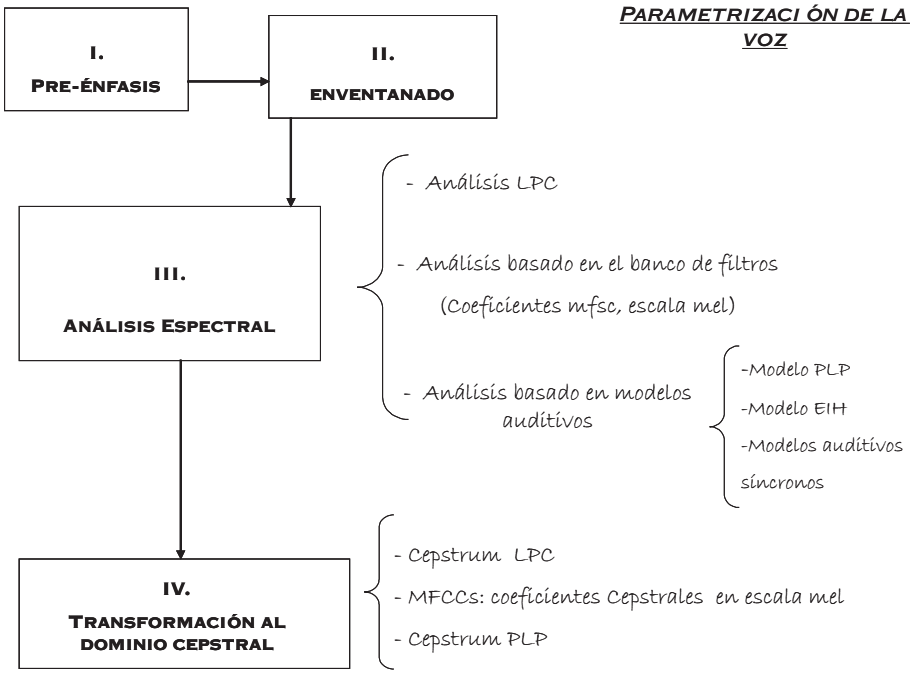


Figura 2.4. El proceso de parametrización

terística definitoria del fonema, aunque es variable para los diferentes hablantes y las diferentes entonaciones.

iii) La energía de la señal. Las vocales y consonantes sonoras tienen mayor energía que las sordas, siendo la energía un buen parámetro de caracterización ya que presenta poca variabilidad para un mismo fonema una vez que ha sido convenientemente normalizada.

- Aportar datos sobre la prosodia de la frase tales como el acento, los tonos y la entonación. Esta información se obtiene analizando:

  i) Las variaciones de la frecuencia fundamental.

  ii) Las variaciones de la duración de los fonemas.

iii) La variación en la intensidad de los fonemas diferenciados.

Teniendo en cuenta la información expuesta como necesaria para caracterizar los fonemas y su prosodia, es razonable que la mayor parte de los sistemas de parametrización se basen en el análisis de la potencia espectral en tiempo corto [103]. Al hacer este análisis la señal se divide en tramas los suficientemente cortas como para poder considerar la señal cuasi-estacionaria. Siendo cuasi-estacionaria, la trama se somete a un análisis espectral y queda caracterizada por un vector de características que suele tener de 10 a 20 parámetros. La figura 2.4 muestra de manera general el proceso de parametrización con las posibles variantes en cada una de las etapas [25], que a continuación serán descritas en detalle.

2.2.1. Filtro de Pre-énfasis

En primer lugar la señal de voz muestreada pasa un filtro de pre-énfasis (típicamente un filtro FIR de primer orden) que amplifica las altas frecuencias para compensar el efecto de los pulsos glotales y la impedancia de radiación. Por lo general este filtro sigue la expresión:

$$H(z) = 1 - \mu \cdot z^{-1}, \quad \text{siendo } 0{,}95 \leq \mu \leq 0,98 \tag{2.1}$$

2.2.2. Enventanado

A continuación la señal es segmentada en tramas de longitudes del orden de 25 ms. Para esta una trama temporal la señal es cuasi-estacionaria y se puede analizar como tal. Para segmentar se usan funciones *ventana* recibiendo este proceso también el nombre de *enventanado*. La ventana de Hamming es la más usada por su compromiso entre resolución de frecuencias y distorsión armónica, para un coste computacional medio. Su expresión es la mostrada en la ecuación 2.2 en la que L es la longitud de la ventana en número de muestras:

$$w(n) = 0{,}54 - 0{,}48 \cdot cos(2\pi \cdot \frac{n-1}{L}), \quad \text{siendo } 1 \leq n \leq L \tag{2.2}$$

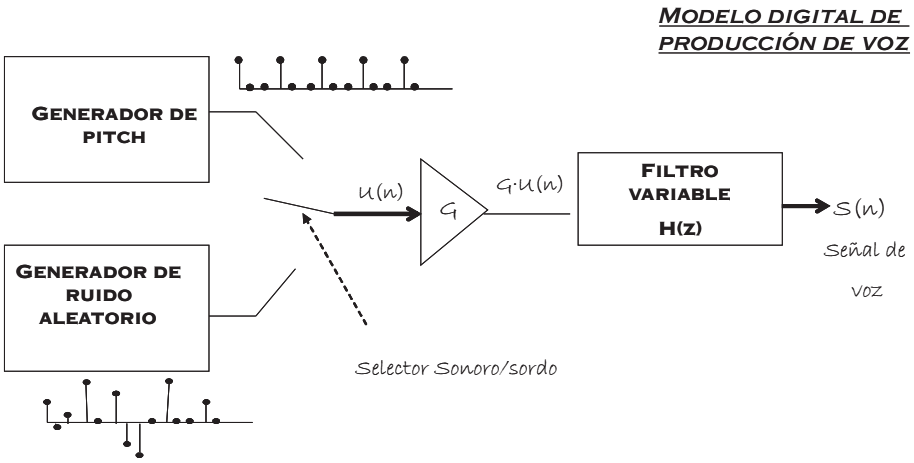


Figura 2.5. Modelo digital de producción de voz

Para una mejor resolución temporal, se usan ventanas solapadas en el tiempo, siendo 10 ms una longitud típica de solapamiento.

2.2.3. Análisis Espectral

El análisis de los segmentos de habla obtenidos se puede hacer tanto en el dominio del tiempo como en el dominio de la frecuencia. En el dominio del tiempo las magnitudes que se analizan son la energía local, la tasa de cruces por cero de la señal y su autocorrelación. Este dominio aporta un análisis de la señal rápido, sencillo y con una interpretación inmediata.

Sin embargo, el análisis espectral es el utilizado por su mayor potencia para caracterizar la información de la señal de voz. Las parametrizaciones usadas en RAH se derivan en su totalidad del análisis de la potencia espectral de las tramas de voz. El análisis de la fase del espectro de frecuencias se omite debido a que los oídos son insensibles a las variaciones de la fase y en consecuencia los equipos de comunicaciones de voz y de grabación no preservan la fase original, que también se ve alterada por factores no deseados como la acústica del entorno. El análisis de la potencia espectral

se hace además en escala logarítmica por motivos prácticos:

- La escala logarítmica hace que cuando la ganancia que tiene la señal cambia, la forma del espectro de potencias se mantenga, simplemente desplazándose hacia arriba o hacia abajo.

- El filtrado lineal debido a la acústica del entorno o a variaciones en el canal, tiene un efecto convolucional en el dominio del tiempo, un efecto multiplicativo para el espectro de potencias lineal, y un simple efecto de suma de una constante para los espectros logarítmicos de potencias.

- La forma de onda de la voz se puede modelar como la convolución en el dominio temporal de la excitación de una señal cuasi-periódica con un filtro variante en el dominio del tiempo que está determinado por la configuración del tracto vocal para la producción de dicha señal de voz (ver figura 2.5 que esquematiza el modelo digital de producción de voz). Esta configuración del tracto vocal como filtro variante en el tiempo va a ser la que nos dé la información sobre los fonemas articulados. Es deseable poder separar estos dos componentes de la forma de onda (excitación cuasi-periódica y filtro variante), y el dominio de la potencia espectral logarítmica es el óptimo para hacerlo ya que en dicho dominio ambos componentes son aditivos.

La escala logarítmica presentaría problemas para valores de frecuencia muy bajos cercanos a cero. El hecho de haber aplicado la ventana de preénfasis previamente elimina este peligro.

Como veíamos en el esquema de la figura 2.4, hay 3 técnicas de análisis espectral ([103], [61], [25], [18],[102], [24]) que serán descritas a continuación:

- **Representaciones basadas en el modelo LPC:** (*Linear Predictive Coding*) [75], [73]. El modelo digital de producción de voz de la figura 2.5 es usado para modelar el tracto vocal como un tubo acústico sin pérdidas ni bifurcaciones, a través del cual el sonido se propaga como

una onda plana. Los efectos del tracto vocal en la señal de excitación son la creación de una serie de resonancias, quedando así el tracto modelado como un filtro *todo polos* $H(z)$. El análisis LPC, también llamado modelo AR de estimación espectral, permite dar los coeficientes de dicho filtro $H(z)$ sin calcular el espectro explícitamente.

- **Análisis basado en el banco de filtros. Coeficientes MFSC:** (*Mel Frequency Spectral Coefficients*). Debido a que el espectro de potencias de la señal se obtiene aplicando la transformada de Fourier a las tramas de voz de ventanas que se solapan, aparecerán armónicos a frecuencias múltiplo de la frecuencia fundamental de las tramas. Este efecto se puede subsanar agrupando los conjuntos de componentes cercanos en unas 20 bandas de frecuencias antes de hacerles el logaritmo de la potencia. Cada filtro hará un promedio pesado de las componentes espectrales presentes en su banda, caracterizando el tracto vocal con la envolvente espectral suavizada. Es común usar la escala de resolución perceptual del oído humano, haciendo que las bandas que abarcan los filtros sean más anchas para frecuencias superiores a 1 Khz. Esta escala recibe el nombre de escala Mel o de frecuencias subjetivas. El logaritmo de la energía a la salida de los filtros en escala Mel da lugar a los coeficientes MFSC.

- **Análisis basado en modelos auditivos**. Este análisis tiene en cuenta aspectos fisiológicos y psicofísicos del proceso auditivo humano que incorpora a los criterios de parametrización. Las anteriores estrategias de parametrización estaban basadas en el modelo de producción de la voz. Este grupo de técnicas se basa sin embargo en el modelo de percepción de la voz humana para parametrizar el habla, intentando reproducir el comportamiento de la membrana basilar del oído humano en la percepción. Para ello se persigue implementar mecanismos que capturen la información fisiológica que caracteriza a la percepción humana:

  - Análisis de frecuencias en canales paralelos.

- Conservación de la estructura temporal fina del sonido.
- Rango dinámico limitado en los canales individuales.
- Realce de los contrastes temporales
- Realce de los contrastes espectrales en frecuencias adyacentes.

Hasta los años 80, los modelos auditivos de parametrización se caracterizaban por la siguiente estructura [18]:

i) Un banco de filtros paso banda que plasma la selectividad en frecuencias del modelo auditivo empleado. Para ello la anchura de los filtros crece de manera no lineal con la frecuencia central de los mismos. Ejemplos de escalas perceptuales para el banco de filtros son la escala *Mel*, la escala *Bark* o la escala *ERB*.

ii) Interacciones no lineales dentro del canal y/o entre canales. Estas interacciones plasman la transducción debida a las células ciliadas del oído, y la supresión lateral entre bandas de frecuencias adyacentes.

iii) A veces, algún mecanismo para aportar información temporal detallada como función de la frecuencia.

Los algoritmos más importantes que han sido propuestos dentro de esta filosofía son:

PLP *Perceptual Linear Prediction* [57]. Este modelo usa tres conceptos de la psicoacústica de la audición para calcular el espectro de la voz: la resolución espectral en la banda crítica, las curvas de igual potencia y la ley de intensidad-potencia. Una vez calculado el espectro de frecuencias, se aproxima usando un modelo AR *todo polos* igual al de *LPC*. Esta técnica será completada posteriormente con un filtrado RASTA que elimina la influencia no deseada de la respuesta en frecuencia del canal de comunicación, con el nombre de RASTA-PLP [55]

GSD *Generalized Synchrony Detector*, (Modelo de Seneff), [123]. Este modelo refleja la respuesta de la membrana basilar a los estímulos acústicos usando un banco de filtros cuya actividad promedia. A este promedio de las activaciones de cada filtro se suma un detector de sincronismo que detecta las activaciones síncronas de varios canales adyacentes.

EIH *Ensemble Interval Histogram*: [46] Da una representación de la voz con alta resolución espectral basada en el cómputo de los histogramas de las frecuencias de activación de los filtros con los que modela la membrana basilar.

Los resultados obtenidos con estas técnicas alrededor de los años 90 son bastante buenos, ligeramente mejores que los de los parámetros MFCC del dominio cepstral (que serán analizados a continuación y que son los más utilizados en la actualidad), ya que captan cierta información útil adicional:

- La estructura temporal detallada de la señal.
- La supresión lateral de los canales adyacentes.
- Los contrastes temporales.
- Otras características no lineales del proceso de audición.

Sin embargo, esta línea de investigación no siguió desarrollándose ya que aunque los resultados eran buenos, llevaban asociado un coste computacional y de almacenamiento que no compensaba ni era factible para reconocimientos en tiempo real con coste computacional razonable. En la actualidad, existe un resurgimiento de esta línea de trabajo motivado por las capacidades de computación y almacenamiento superiores a las existentes en los años 80, por la necesidad de encontrar parametrizaciones que mejoren MFCC y permitan enfrentarse a los actuales retos de reconocimiento que también han aumentado, y por el descubrimiento de que la información de la sincronía/asincronía temporal de la señal de voz que el oído humano

capta, es muy útil para caracterizar los formantes [63], [67] y no es capturada por los MFCCs. El resultado ha sido el interés en crear algoritmos que exploren los mecanismos de extracción de la información de sincronía de las salidas de los canales paralelos que modelan el canal auditivo, y capten mejor la relación entre frecuencias y tiempos. Ejemplos de esta línea de trabajo son:

- Uso de la transformada wavelet: el comportamiento del canal periférico auditivo es modelable mediante una transformada wavelet que captura la información en los dominios de la frecuencia y el tiempo mejorando algunas de las limitaciones de la Transformada de Fourier [115].
- Mejoras del modelado síncrono de Seneff como la aportada con el algoritmo ALSD (*Average Localized Synchrony Detection* [9]).
- Variaciones del ZPCA (*Zero-Crossing and Peak Amplitudes*) como las propuestas en [68] y [47], que son mejoras del antiguo EIH (*Enhanced Interval Histogram*) antes mencionado incorporando información de sincronismo.
- Modelo de la cóclea para extracción de parámetros, propuesto por R. Lyon en [72].
- Uso de redes neuronales para capturar las operaciones no lineales en el espectro logarítmico de frecuencias.
- Uso de paradigmas fractales para modelar los procesos no lineales de la percepción. [74]

2.2.4. El dominio cepstral

Las técnicas de análisis espectral que operan en el dominio de la potencia espectral logarítmica tienen la limitación de que debido a que los espectros de los filtros en bandas adyacentes están bastante correlados, originan coeficientes espectrales también bastante correlados. Es deseable eliminar esa correlación manteniendo solo la información que sea útil para el reconocimiento. Para ello, se utiliza un filtro de decorrelación homomórfica

o *Cepstrum* que, mediante la transformada inversa de Fourier del logaritmo del espectro de potencias, lleva los coeficientes espectrales al dominio de la *cuefrencia* convirtiéndolos en coeficientes *cepstrales*. Los coeficientes cepstrales representan la señal temporal que correspondiente al espectro logarítmico de potencias. El dominio de la cuefrencia es un dominio homomórfico del dominio temporal. Esto implica que las convoluciones en el dominio temporal se convierten en sumas en su dominio homomórfico de la cuefrencia. Esto será sumamente útil ya que permitirá separar las señales de voz de los ruidos convolucionales con los que estén mezcladas. Las componentes de excitación y envolvente espectral del tracto vocal aparecerán en zonas separadas del dominio transformado de la cuefrencia, que se podrán separar mediante ventanas. Haciendo un juego de paralelismos, los inventores de este operador homomórfico llamado Cepstrum (cuyo nombre crearon intercambiando la posición de las cuatro primeras letras del término *spectrum*), llamaron a ese enventanado en el dominio de la cuefrencia "*liftering*"(cambiando la posición de las primeras letras del termino correspondiente *filtering* del dominio spectrum) [92]. Los análisis en el dominio espectral tienen sus correspondientes homólogos en el dominio de la cuefrencia que reciben los nombres de coefientes cepstrales LPC, MFCCs (Mel Frequency Cepstral Coefficients), o Cepstrum PLP. Los coeficientes MFCC han demostrado ser los que mejores resultados dan como técnica de parametrización teniendo en cuenta el compromiso entre coste computacional y resultados obtenidos [24].

2.2.5. Post-procesado del vector de características

Existen herramientas para eliminar dependencias entre conjuntos de variables, que son frecuentemente usadas para depurar el vector básico de características obtenido con los métodos de parametrización descritos anteriormente. El análisis de componentes principales (*Principal Component Analysis*, **PCA** [51]) elimina completamente las dependencias lineales entre un conjunto de características y se usa para decorrelar los niveles de energía a lo largo del espectro o combinaciones de distintos parámetros

que puedan aportar información redundante. El análisis lineal discriminante (*Linear Discriminant Análisis*, **LDA** [100]) hace un análisis equivalente al de PCA al que se le añade una ponderación de las características basada en características discriminativas mediante aprendizaje supervisado. Si los parámetros de voz se pasan a través de un filtro paso banda que tenga un cero espectral a frecuencia cero, se eliminarán las características de variación muy constantes o muy lentas debidas a la respuesta en frecuencias del canal de comunicación. Este proceso recibe el nombre de filtrado **RASTA**([57]), y aporta robustez frente a las distorsiones espectrales lineales por lo que aparece combinado con algunas de las parametrizaciones descritas.

Es deseable además aumentar la información estática del espectro de potencias de tiempo corto con información acerca de sus cambios en el tiempo. Furui propone con éxito (ver [38]) calcular y añadir al vector de características la diferencia entre tramas correlativas. La **primera derivada** (velocidad) y **la segunda derivada** (aceleración) del vector de características son desde entonces la aproximación inmediata para tener en cuenta las variaciones temporales entre tramas.

2.2.6. Sistemas de Reconocimiento: aproximación estadística

En la actualidad el reconocimiento automático del habla se hace de manera unívoca mediante arquitecturas software que generan una secuencia de hipótesis de reconocimiento a partir de la señal acústica, basándose en métodos estadísticos. Las secuencias de parámetros acústicos que recibe el reconocedor se definen como observaciones de modelos de palabras O. Reconocerlos significa:

i) Calcular la probabilidad *a posteriori* de que el emisor haya pronunciado la palabra de transcripción W habiendo observado la secuencia O, para las distintas transcripciones W_i posibles.

ii) Identificar la observación con transcripción W que maximice esa probabilidad a posteriori.

El reconocedor elegirá como secuencia reconocida la transcripción $\hat{W}$ que maximice la probabilidad a posteriori. Usando la regla de Bayes, esto equivale a maximizar la expresión:

$$\begin{aligned} \hat{W} = \underset{W}{arg\,max} \quad \{p(W|O)\} &= \underset{W}{arg\,max} \quad \{\frac{p(O|W) \cdot p(W)}{p(O)}\} \\ \hat{W} &\sim \underset{W}{arg\,max} \quad \{p(O|W) \cdot p(W)\} \end{aligned} \tag{2.3}$$

El término $p(O)$ de la expresión 2.3 representa la probabilidad de que se dé una determinada observación. Su valor es constante e independiente de W por lo que se elimina del proceso de maximización. Los dos términos $p(W)$ y $p(O|W)$ estructuran el reconocimiento estadístico y serán analizados a continuación:

- *p(W)* es la probabilidad a priori de que aparezca la secuencia W, y se denomina modelo de lenguaje.

- $p(O|W)$ es la probabilidad de que la transcripción W tenga la representación acústica O. Esta probabilidad se denomina modelo acústico.

2.2.7. El modelo de lenguaje

El modelo de lenguaje también llamado gramática determina la probabilidad de una transcripción $p(W) = P(w_1, w_2, ..., w_N)$. El modelo de lenguaje más usado es el de las n-gramáticas en el que la probabilidad de cada transcripción depende de las $(n-1)$ transcripciones anteriores:

$$P(w_1, w_2, ..., w_N) = \prod_{n=1}^{n=N} p(w_n), \quad \text{para una uni-gramática} \tag{2.4}$$

$$P(w_1, w_2, ..., w_N) = p(w_1) \cdot \prod_{n=2}^{n=N} p(w_n|w_{n-1}), \quad \text{para una bi-gramática} \tag{2.5}$$

$$P(w_1, w_2, ..., w_N) = p(w_1) \cdot p(w_2|w_1) \prod_{n=3}^{n=N} p(w_n|w_{n-1}, w_{n-2}) \quad \text{para una tri-gramática} \tag{2.6}$$

También se puede usar un modelo más sencillo en el que todas las secuencias de transcripciones son equiprobables, y entonces $p(W)$ no influye en la maximización de la expresión 2.3.

En lo que se refiere al tamaño de la unidad de transcripción y de observación, hay varias posibilidades. En la comunicación oral entre humanos las unidades básicas son las palabras. Para los sistemas de reconocimiento automático con vocabularios grandes es necesario usar unidades menores que la palabras como pueden ser sílabas, fonemas, trifonemas, etc.

2.2.8. Aproximaciones al modelado acústico

El modelo acústico del sistema de reconocimiento se puede generar siguiendo diferentes estrategias. La bibliografía de clasificación de patrones recoge las tres aproximaciones siguientes, que conceptualmente se diferencian en la referencia que se toma para clasificar los patrones [103],[25],[48]:

Comparación de plantillas o patrones

Esta aproximación es la más antigua y toma como referencia plantillas de los objetos que se clasifican. Mediante técnicas de programación dinámica basadas en el algoritmo DTW (*Dinamyc Time Warping*, [13]), se entrena el sistema obteniendo plantillas o patrones de las unidades que habrá que reconocer. La plantilla es una secuencia de características acústicas ordenadas en el tiempo que son índices en un diccionario de centroides. La comparación con estas plantillas exige un alineamiento temporal no lineal de las

mismas con los datos de entrada, y una medida de distancia. Esta técnica tiene la limitación de que necesita mucha capacidad de almacenamiento para los centroides con los que se ha de comparar, y presenta inconvenientes en el habla continua por la dificultad de la alineación temporal.

Modelado estadístico: HMMs

Las aproximaciones estadísticas toman como referencia el modelo estocástico de los datos. Se hace también un alineamiento no lineal con el algoritmo de Viterbi [103], entre los datos de entrada y las palabras de un diccionario cuyos términos son patrones estocásticos. Los alineamientos se establecen como probabilidades de que la secuencia analizada sea generada por los distintos Modelos de Markov (*HMMs, Hidden Markov Models* [104]). Hasta la fecha este método es el que mejores resultados proporciona y el más utilizado.

Redes Neuronales

La aproximación basada en Redes Neuronales (*ANN, Artificial Neural Networks*, [6]) toma como referencia para el modelado los patrones de actividad. Se define como un modelo computacional paralelo compuesto de unidades procesadoras adaptativas con una alta interconexión entre ellas mediante pesos. Estas unidades se agrupan en diferentes capas, distinguiéndose la capa de entrada y la de salida. Se caracterizan por ser estructuras intrínsecamente no lineales con capacidad para aprender una determinada tarea a partir de pares de observación-objetivo sin hacer asunciones sobre el modelo subyacente. Según la topología de la red existen varios tipos de redes: Perceptrón multicapa, máquina de Boltzman, Adaline, etc. Se han conseguido buenos resultados utilizando redes neuronales para reconocimiento de patrones acústicos ([52], [134]), sin embargo su uso no ha superado al de los HMMs porque presentan limitaciones como un tiempo excesivamente elevado para su entrenamiento, o el desconocimiento a priori del número de nodos y capas necesarios para cada problema de clasificación. Existen sistemas híbridos de clasificación que combinan los

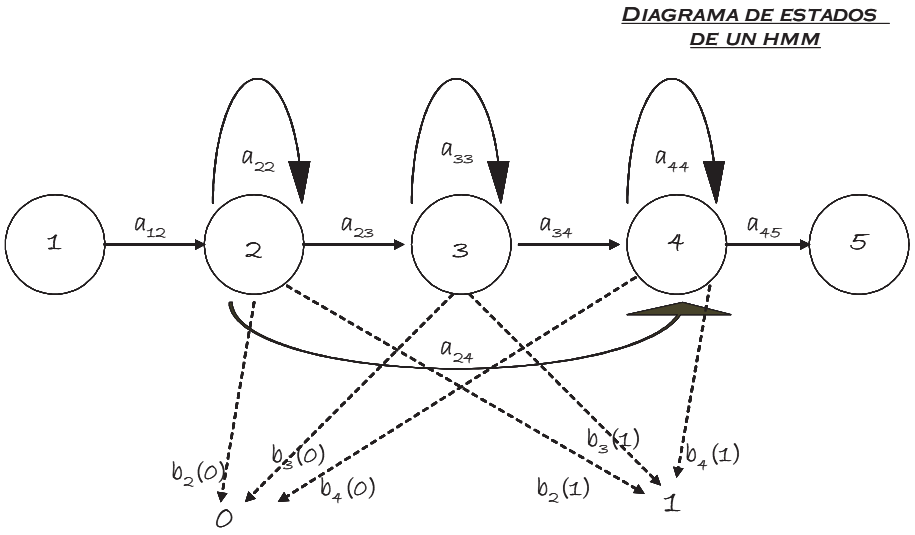


Figura 2.6. Diagrama de estados de un HMM

HMMs con las ANN consiguiendo resultados interesantes [84], [130] .

2.2.9. El modelo acústico: Modelos Ocultos de Markov

La probabilidad $P(X|W) = P(x_1, x_2, ..., x_T|W)$ es la probabilidad del modelo acústico de la transcripción W. Es decir, es la probabilidad de que la señal acústica que representa a W sea X. Es una práctica universal usar los modelos ocultos de Markov (HMM, *Hidden Markov Models*, [104]) para calcular las probabilidades acústicas debido a su capacidad de modelar estadísticamente de manera adecuada la generación de voz. Un modelo oculto de Markov es la composición de dos procesos estocásticos *(X,Y)* definidos como:

i) Una cadena oculta de Markov X que tiene en cuenta la variabilidad temporal, y que no es directamente observable.

ii) Un proceso observable Y que tiene en cuenta la variabilidad espectral y va tomando valores en el espacio de las características acústicas u

observaciones.

La combinación de ambos procesos modela las fuentes de variabilidad de la señal de voz y permite reflejar una secuencia de parámetros acústicos como concatenación de los procesos elementales del modelo con la flexibilidad suficiente para hacer sistemas de reconocimiento. Los modelos ocultos de Markov usados en el reconocimiento de voz tienen dos asunciones formales características:

a) La historia de la cadena no influye en la evolución futura de la misma si existe información actual (hipótesis de Markov de primer orden).

b) Ni la evolución de la cadena ni las observaciones pasadas determinan la observación actual si se ha especificado la ultima transición de la cadena (hipótesis de independencia de las salidas).

Una vez hechas esas asunciones, si llamamos $y \epsilon Y$ a una variable que representa las observaciones y llamamos $i, j \epsilon X$ a las variables que representan los estados del modelo, el modelo $\lambda = (A, B, \Pi)$ queda representado por las siguientes matrices de parámetros que se pueden ver en la figura 2.6:

$$\begin{aligned} A &\equiv a_{i,j}|i, j\epsilon X, \quad \text{probabilidades de transición} \\ B &\equiv b_{i,j}|i, j\epsilon X, \quad \text{distribuciones de las salidas} \\ \Pi &\equiv \pi_i|i\epsilon X, \qquad \text{probabilidades iniciales} \end{aligned} \tag{2.7}$$

donde los términos de las matrices se definen como:

$$\begin{aligned} a_{i,j} &\equiv p(X_t = j|X_{t-1} = i) \\ b_{i,j}(y) &\equiv p(Y_t = y|X_{t-1} = i|X_t = j) \\ \pi_i &\equiv p(X_o = i) \end{aligned} \tag{2.8}$$

Según la naturaleza de la matriz B de las distribuciones de probabilidad de las salidas, los HMMs se pueden clasificar en varios tipos, [104],[25]:

i) Modelos discretos, DHMMs

Las observaciones son vectores de símbolos de un alfabeto finito de N elementos diferentes. Para cada componente de dicho vector de símbolos se define una densidad discreta:

$$\{w(k)|k = 1, ..., N\}$$

y la probabilidad del vector se calcula multiplicando las probabilidades de cada componente siendo éstos independientes entre sí.

$$b_i(y) = p(y|x_i, \lambda) = \prod_k w_{y,x_i,k} \tag{2.9}$$

ii) Modelos continuos, CHMMs

Otra posibilidad es definir las distribuciones de probabilidad en espacios de observaciones continuos, lo cual puede ser conveniente ya que la señal de voz es continua. Las distribuciones de probabilidad necesitan ciertas restricciones en este caso para que el número de parámetros del sistema sea manejable y las re-estimaciones sean consistentes: las transiciones se definen con mezclas de distribuciones paramétricas básicas caracterizadas por pocos parámetros, que suelen ser Gausianas o Laplacianas. Cada estado x_i del modelo tendrá un conjunto específico $V(x_i, \lambda)$ de funciones densidad de probabilidad. Si llamamos v_k a cada una de esas *pdfs*, la expresión de las probabilidades de las salidas será:

$$b_i(y) = p(y|x_i, \lambda) = \sum_{v_k \epsilon V(x_i,\lambda)} p(y|v_k, x_i, \lambda) \cdot P(v_k|x_i, \lambda) \tag{2.10}$$

donde el término $P(v_k|x_i, \lambda)$ representa la probabilidad de aparición de la *pdf* v_k.

iii) Modelos semicontinuos, SCHMMs

Para modelar distribuciones complejas con la mezcla funciones paramétricas a veces es necesario un gran número de estas en cada mez-

cla, y un corpus de entrenamiento muy grande. Una solución efectiva es compartir las distribuciones entre diferentes transiciones del modelo. Esto es lo que hacen los modelos semicontinuos, en los que todos los estados comparten las mismas distribuciones de probabilidad con diferentes pesos:

$$\begin{aligned} V(x_i, \lambda) &= V, \quad \forall x_i, \lambda \\ b_i(y) = p(y|x_i, \lambda) &= \sum_{v_k \epsilon V} p(y|v_k) \cdot P(v_k|x_i, \lambda) \end{aligned} \tag{2.11}$$

iv) Modelos con cuantización vectorial múltiple: MVQHMMs y SCMVQHMMs

El cálculo de las probabilidades los modelos continuos es más lento que en el caso de los modelos discretos. Una alternativa que agiliza el aprendizaje es usar cuantización vectorial de las mezclas de Gausianas. Para ello se reduce la mezcla de distribuciones de probabilidad que definen la *pdf*, a la distribución más probable.

$$\begin{aligned} V(x_i, \lambda) &= V(\lambda), \quad \forall x_i \epsilon \lambda \\ b_i(y) = p(y|x_i, \lambda) &= p(y|o, \lambda) \cdot P(o|x_i, \lambda) \\ o = max^{-1}_{v_j \epsilon V(\lambda)}[P(v_j|y, \lambda)] &= max^{-1}_{v_j \epsilon V(\lambda)}[p(y|v_j, \lambda)] \cdot P(v_j|\lambda)] \end{aligned} \tag{2.12}$$

A la secuencia de observaciones $Y = y_1, y_2, ...y_T$, se le asocia una secuencia de símbolos discretos $O = o_1, o_2, ...o_T$, y la probabilidad acústica $p(Y|\lambda)$ se descompone en el producto de probabilidades:

$$p(Y|\lambda) = p(Y|O, \lambda) \cdot P(O|\lambda) \tag{2.13}$$

donde $p(Y|O, \lambda)$ es la probabilidad de cuantización y $P(O|\lambda)$ es la probabilidad de generación de símbolos discretos por parte del modelo. En el caso de la cuantización vectorial múltiple aplicada a los modelos de Markov semicontinuos (SCMVQHMMs), la filosofía es la misma y la probabilidad

de las salidas tendrá la expresión:

$$b_i(y) = p(y|x_i, \lambda) = \sum_{v_k \epsilon V(\lambda)} p(y|v_k, \lambda) \cdot P(v_k|x_i, \lambda) \tag{2.14}$$

2.2.10. El proceso de modelado

El proceso de modelado de unidades acústicas se divide comúnmente en 3 problemas [104], [25], [36]:

- Evaluación. Es el problema básico: dada una observación acústica y y un modelo oculto de Markov, determinar la probabilidad de que el modelo genere esa observación, es decir, la probabilidad acústica $p(Y|\lambda)$. Esta probabilidad se determina con el algoritmo *forward-backward* [104].

- Decodificación. Determinación de la secuencia óptima de estados $X = x_1, x_2, ...x_T$ dada la observación acústica y y el modelo oculto de Markov. Es decir, se busca la alineación de la observación con el modelo, asignando cada vector a un estado del modelo. Se lleva a cabo mediante el algoritmo de *Viterbi* [133].

- Estimación o entrenamiento de los HMMs. Consiste en el cálculo de los parámetros que caracterizan el modelo. Dados un conjunto de datos y una colección de secuencias observables, se determina el HMM que con mayor probabilidad ha generado la secuencia. Este problema se resuelve comúnmente con el algoritmo *Baum-Welch* [12].

2.3. Criterios de Evaluación del RAH

La evaluación cuantitativa del sistema de reconocimiento provee medidas estandarizadas del funcionamiento del mismo y da la posibilidad de compararlo con otros, posibilitando así la extracción de conclusiones al respecto. Para dar una evaluación cuantitativa del reconocedor hay que conocer:

i) La probabilidad de cometer errores de reconocimiento.

ii) La complejidad computacional y los requerimientos de memoria del proceso de reconocimiento.

iii) El tiempo de respuesta del sistema.

2.3.1. Tasa de Error y precisión del reconocimiento

La tasa de error se define como la capacidad del reconocedor automático del habla de cometer errores de reconocimiento. En el caso de reconocimiento de palabras aisladas, la tasa de error de palabra (*WER, Word Error Rate*) se define como:

$$WER = \frac{n_e}{n_p} \tag{2.15}$$

Donde n_p representa el número total de palabras reconocidas y n_e el número de palabras clasificadas erróneamente. Para el caso de reconocimiento de habla continua, el reconocimiento se hace frase a frase y se definen tres tipos de errores: errores de inserción n_i, errores de borrado n_b, y errores de sustitución n_s. En este caso la tasa de error de palabra se define como:

$$WER = \frac{n_i + n_s + n_b}{n_p} \tag{2.16}$$

Algunos autores trabajan con la tasa de aciertos de palabras (*WAcc, Word Accuracy*) cuya expresión es:

$$WAcc = 1 - WER = 1 - \frac{n_i + n_s + n_b}{n_p} \tag{2.17}$$

2.3.2. Intervalo de confianza de la medida del error

La tasa de error antes definida es una estimación de la probabilidad de error dentro de un determinado intervalo de confianza cuya amplitud dependerá del número total de pruebas con que se haya obtenido la tasa de error. Si se asume (ver [25]) una distribución binomial $B(n, p)$ del número de elementos reconocidos correctamente siendo p la probabilidad de aciertos y n el número total de ensayos, se puede definir un intervalo de con-

fianza centrado en el valor estimado de la probabilidad de acierto $\hat{p}$ que contendrá con probabilidad $(1-\alpha)$ la probabilidad de acierto. Mediante el teorema del límite central se demuestra que esta probabilidad $\hat{p}$ tiende a la distribución normal $N(0,1)$ y el intervalo de confianza para el valor $\hat{p}$ será:

$$[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}] \tag{2.18}$$

donde $z_{1-\frac{\alpha}{2}}$ es el cuantil $(1-\frac{\alpha}{2})$ de la distribución normal estándar. En la expresión 2.18 se puede apreciar que cuanto mayor es el número de elementos reconocidos, más estrecho será el intervalo de confianza.

2.3.3. Aspectos computacionales y tiempo de respuesta

El coste computacional del proceso de reconocimiento se puede dividir en dos partes, el coste computacional de la parametrización de señal de voz, y el coste computacional del proceso de decodificación:

- El coste computacional de la parametrización unido al coste de almacenamiento, aumentará con los algoritmos más elaborados como son las parametrizaciones basadas en los modelos auditivos, siendo mínimo para el caso de sustracción espectral de la media por ejemplo.
- En el proceso de reconocimiento, la decodificación es la que tiene un coste computacional que puede ser bastante elevado, especialmente para el reconocimiento de habla continua. Las búsquedas en las gramáticas con alta perplejidad hacen que el árbol de búsqueda se expanda. Para controlar esta expansión y con ella el coste y el tiempo de respuesta, se define un umbral de poda y se utilizan algoritmos heurísticos (ver [112]) que descartan las opciones menos probables.

Los costes mencionados se deben sintonizar según los requerimientos de tasa de error de la aplicación, y los medios con los que se cuenta para su implementación.

Capítulo 3

Robustecimiento del RAH

En este capítulo se describe el efecto del ruido en la señal de voz, y se hace un repaso exhaustivo de las estrategias de robustecimiento que existen en la actualidad, dividiéndolas para ello en técnicas de cancelación del ruido, técnicas de compensación de características y técnicas de adaptación de modelos. El ruido no estacionario y los algoritmos que lo combaten son también descritos. Por último se exponen los mecanismos de robustecimiento con *arrays* de micrófonos.

3.1. Reconocimiento Automático del Habla en entornos ruidosos

El proceso de reconocimiento se produce de manera óptima cuando las condiciones de los datos que se evalúan son idénticas a aquellas con las que entrenó el sistema de reconocimiento. Esto no ocurre casi nunca en el mundo real de las aplicaciones de reconocimiento del habla. Existen muchas fuentes de variabilidad que producen desajustes entre las condiciones de entrenamiento y las de evaluación.

La figura 3.1 muestra un esquema de estas fuentes de variabilidad y de las estrategias que existen para combatirlas. **Un mismo hablante** produce sonidos con variaciones indeseadas, que no transmiten información acústica relevante, según su estado físico o emocional, o según el contexto

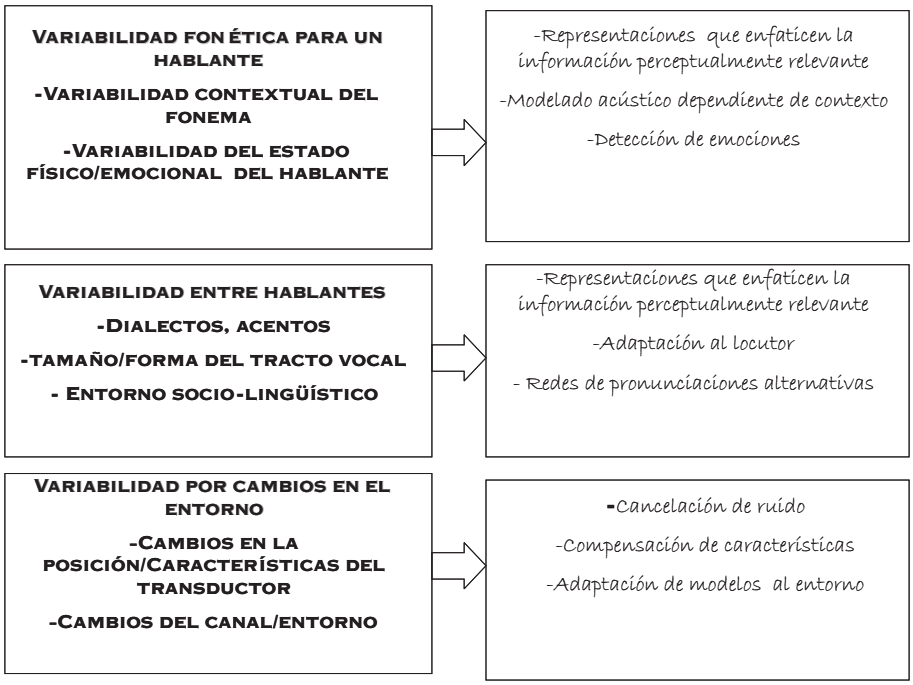


Figura 3.1. Estrategias de robustecimiento

fonético de los sonidos articulados. Las **variaciones entre hablantes** tienen que ver con las características del tracto vocal y el género de los mismos, y se suman a las antes mencionadas para un hablante concreto. Estas variabilidades se atenúan o eliminan mediante parametrizaciones que subrayan la información acústica con entrenamientos dependientes de contexto, redes de pronunciación paralelas, algoritmos de adaptación al locutor o de normalización del tracto vocal del locutor por ejemplo.

Cuando la variabilidad se produce por **cambios en el entorno del hablante**, y/o en la posición y características del **canal** que éste utiliza, las estrategias para combatirlas se denominan estrategias de robustecimiento del reconocedor de habla. El reconocimiento robusto es por tanto aquel

que está *inmunizado* o es lo menos vulnerable posible a los cambios de las condiciones del entorno en que se produce la evaluación.

Los algoritmos de robustecimiento constituyen un área de investigación y desarrollo fundamental en el procesado de la voz. Los retos actuales del reconocimiento automático del habla están enmarcados en las siguientes líneas de trabajo:

- Reconocimiento de voz codificada sobre líneas telefónicas: presenta la dificultad adicional de que cada canal telefónico tiene su propia SNR y respuesta en frecuencias. Las aplicaciones telefónicas de reconocimiento tienen que adaptarse a los canales usando muy pocos datos específicos de canal.

- Entornos con SNR baja. Si en los años 80 se hacía RAH en una habitación silenciosa con un micrófono de mesa, en la actualidad los escenarios en los que se demanda el reconocimiento son:

  - Teléfonos móviles.

  - Coches en marcha.

  - Voz espontánea.

  - Voz enmascarada por otra voz.

  - Voz enmascarada por música.

  - Ruidos no estacionarios.

- Interferencias de voz co-canal. Las interferencias causadas por otros hablantes son un reto para el reconocimiento robusto de mayor dificultad que los cambios del entorno de reconocimiento debidos a ruidos de banda ancha.

- Adaptación rápida para hablantes no nativos. Las aplicaciones actuales de voz demandan robustecimiento y adaptación a los acentos de hablantes no nativos.

- Bases de Datos con degradaciones realistas. La formulación, grabación y diseminación de bases de datos que contengan ejemplos

realistas de la degradación que existe en entornos prácticos son necesarias para enfrentarse a los retos de reconocimiento automático existentes.

En la siguiente sección se estudiarán en profundidad los efectos del ruido en la señal de voz. De un modo general se puede decir que la variabilidad introducida por el ruido degrada el reconocimiento por dos razones:

i) Debido a su naturaleza aleatoria, el ruido aumenta la incertidumbre de la señal. Esto provoca que la información mutua entre lo que el hablante ha dicho y lo que se ha grabado disminuya. La consecuencia inmediata es que la precisión que se puede obtener con un reconocedor optimo disminuye.

ii) La adición del ruido a la voz cambia la distribución de la señal de voz, haciéndola diferente a la usada para modelar el reconocedor de voz. La consecuencia es que los umbrales de clasificación se convierten en erróneos.

3.1.1. El ruido y sus efectos

El ruido se define como todo sonido no deseado, que distorsiona la información transmitida por la onda acústica dificultando su correcta percepción. Existen dos tipos fundamentales de distorsiones de la señal de voz: el ruido aditivo y la distorsión de canal. Sus efectos y las estrategias que existen para combatirlos que se describirán a continuación, se pueden ver esquematizados en la figura 3.2.

El ruido aditivo se define como el que se suma a la señal de voz en el dominio del tiempo, y será estacionario si además tiene una densidad de potencia espectral que no varía con el tiempo. Dentro de esta categoría existen ruidos aditivos blancos, que son aquellos que tienen un espectro de potencias plano, en contraste con los ruidos aditivos coloreados, cuyo espectro de potencias tienen peculiaridades para ciertas frecuencias (el ruido rosa por ejemplo, tiene mayor energía a bajas frecuencias). Ruidos aditivos

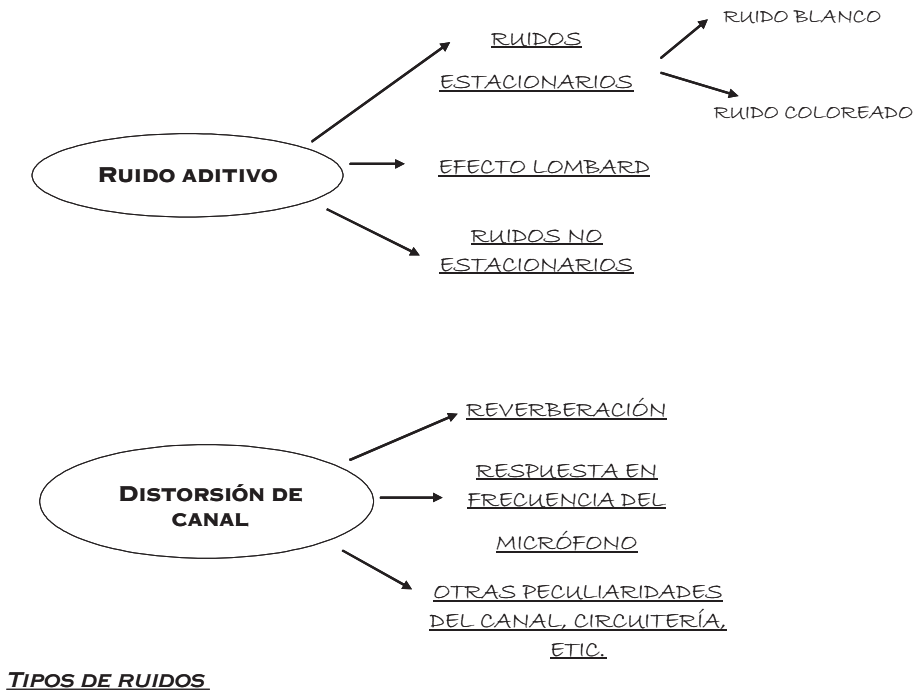


Figura 3.2. Tipos de ruido en la señal de voz

no estacionarios son aquellos cuyas propiedades estadísticas cambian con el tiempo. En esta categoría están los portazos, las voces espontáneas, efectos de los labios o la respiración, etc. Los efectos del ruido aditivo en la señal son los más difíciles de eliminar, ya que tienen la peculiaridad de transformarla no linealmente en ciertos dominios de análisis. El ruido aditivo se puede considerar el motor de la investigación que hay en marcha en el campo del reconocimiento automático robusto de habla en estos momentos.

La distorsión de canal es el ruido que se mezcla de manera convolucional con la señal de voz en el tiempo. Puede estar provocado por reverberaciones de la señal en el medio de transmisión, por la respuesta en frecuencias del micrófono que se utilice, o por peculiaridades del medio de

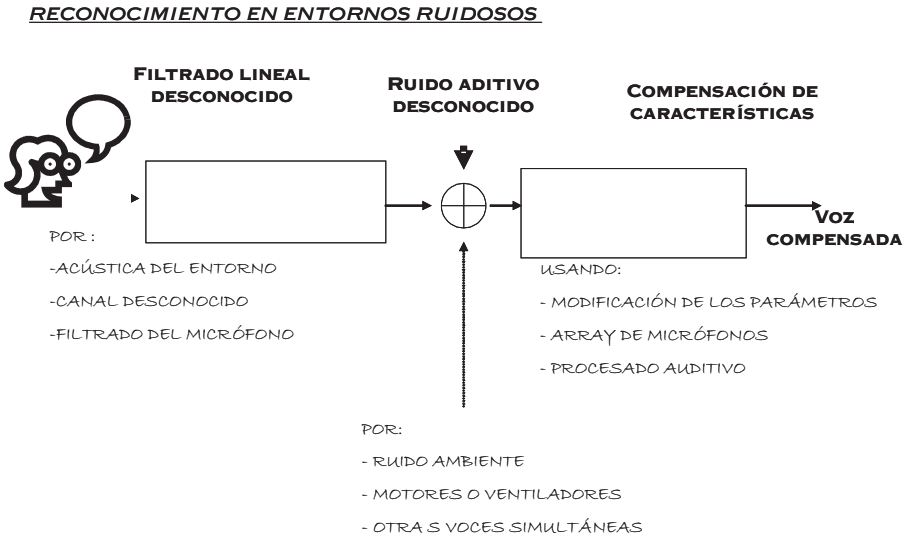


Figura 3.3. Ruido en el proceso de reconocimiento

transmisión tales como un filtro eléctrico en la circuitería A/D, etc. Sus efectos han sido combatidos con relativo éxito ya que son lineales, y se evitan procesando linealmente la señal como métodos como el filtrado RASTA, cancelación de ecos, o sustracción del valor medio de los coefientes MFCC por ejemplo.

El paradigma más utilizado para modelar el efecto del ruido en la comunicación oral (ver figuras 3.3 y 3.4), es el que lo define como una combinación de ruido aditivo y filtrado lineal o ruido convolucional de canal [61] que obedece a la expresión:

$$y[m] = x[m] * h[m] + n[m] \tag{3.1}$$

Teniendo en cuenta que el ruido n[m] y la seña x[m] son estadísticamente independientes, la señal contaminada y[m] tendrá la siguiente expresión en el dominio de la frecuencia espectral para el canal i del banco de

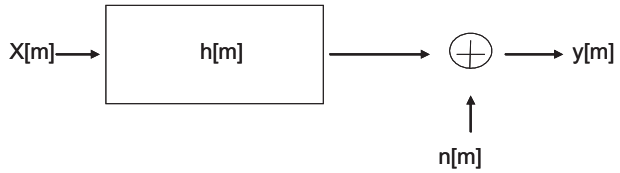


Figura 3.4. Modelo del ruido del entorno

filtros:

$$|Y(f_i)|^2 \cong |X(f_i)|^2 \cdot |H(f_i)|^2 + |N(f_i)|^2 \tag{3.2}$$

Tomando logaritmos en la expresión 3.2 y operando, se llega a la siguiente aproximación en el dominio de la frecuencia:

$$ln|Y(f_i)|^2 \cong ln|X(f_i)|^2 + ln|H(f_i)|^2 + ln(1 + exp(|N(f_i)|^2 - ln|X(f_i)|^2 - ln|H(f_i)|^2)) \tag{3.3}$$

Para pasar la ecuación 3.3 al dominio cepstral con $M + 1$ coeficientes cepstrales, se definen las siguientes 4 matrices con la transformada discreta del coseno (operador *C()* en la expresión 3.4) que representan vectores en el dominio cepstral:

$$\begin{aligned} x &= C(ln|X(f_0)|^2 \quad ln|X(f_1)|^2 \quad ... \quad ln|X(f_M)|^2) \\ h &= C(ln|H(f_0)|^2 \quad ln|H(f_1)|^2 \quad ... \quad ln|H(f_M)|^2) \\ n &= C(ln|N(f_0)|^2 \quad ln|N(f_1)|^2 \quad ... \quad ln|N(f_M)|^2) \\ y &= C(ln|Y(f_0)|^2 \quad ln|Y(f_1)|^2 \quad ... \quad ln|Y(f_M)|^2) \end{aligned} \tag{3.4}$$

Combinando las ecuaciones 3.3 y 3.4 podemos llegar a una expresión en el dominio cepstral de la señal y contaminada:

$$\hat{y} = \hat{x} + \hat{h} + g(\hat{n} - \hat{x} - \hat{h}) \tag{3.5}$$

siendo la función g de la expresión 3.5 definida como:

$$g(z) = C(ln(1 + e^{C^{-1}(z)}) \tag{3.6}$$

Consideraremos para simplificar el análisis que no hay ruido convolucional de canal o que este se elimina mediante algún proceso de filtrado lineal, es decir, consideraremos que $H(f) = 1$. La expresión en el dominio cepstral de la señal con ruido aditivo será:

$$y = x + ln(1 + exp(n - x)) \tag{3.7}$$

La relación entre la señal limpia x y la señal y contaminada con ruido aditivo n expresada por la ecuación 3.6 es lineal para valores de señal altos, y deja de serlo cuando la energía de la señal se aproxima o es menor que la del ruido. Este comportamiento queda reflejado en la gráfica 3.5, en la que se representa la energía logarítmica de una señal y contaminada con un ruido aditivo Gausiano de media $\mu_n = 3$ y desviación estándar $\sigma_n = 0,4$. La línea continua es el valor medio de la transformación no lineal que ha sufrido la energía logarítmica. Los puntos representan los datos transformados. La transformación media se puede invertir para conseguir el valor esperado de la señal limpia dada la señal ruidosa observada. Esta estimación tendrá en cualquier caso un grado de incertidumbre que dependerá de la SNR del punto transformado. Para valores de y con energía mucho mayor que el ruido, ese grado de incertidumbre será pequeño. Para los valores de y cercanos a la energía del ruido, la incertidumbre será alta.

Esta no linealidad de la distorsión es una característica del ruido aditivo en el dominio cepstral. Si se analizan los histogramas de la densidad de probabilidad de los coeficientes MFCC para una señal limpia y una señal contaminada con ruido aditivo se observa [27] que los efectos del ruido son:

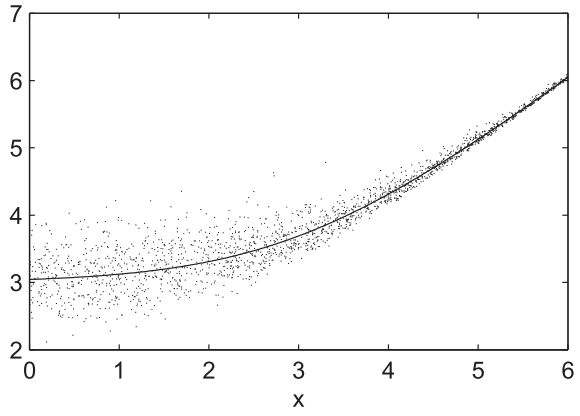


Figura 3.5. Transformación aleatoria debida al ruido aditivo.

- Un desplazamiento del valor medio del histograma del coeficiente de la señal contaminada.
- Una reducción de la varianza de dicho histograma.
- La modificación de la forma global del histograma (que equivale a una modificación de los momentos de orden superior del mismo). Esa modificación es especialmente acusada en la energía logarítmica y en los coeficientes cepstrales de menor orden C_0 y C_1.

Estrategias de robustecimiento frente al ruido

Existen varias clasificaciones de las técnicas clásicas de robustecimiento frente a la distorsión producida por el ruido. Una muy ampliamente aceptada es la se puede observar en la figura 3.6, que considera tres familias de técnicas basándose en la filosofía usada para afrontar los efectos del ruido:

i) Técnicas de pre-procesado de la señal para cancelar del ruido antes de parametrizar la señal de voz, con el objetivo de que al parametrizarla

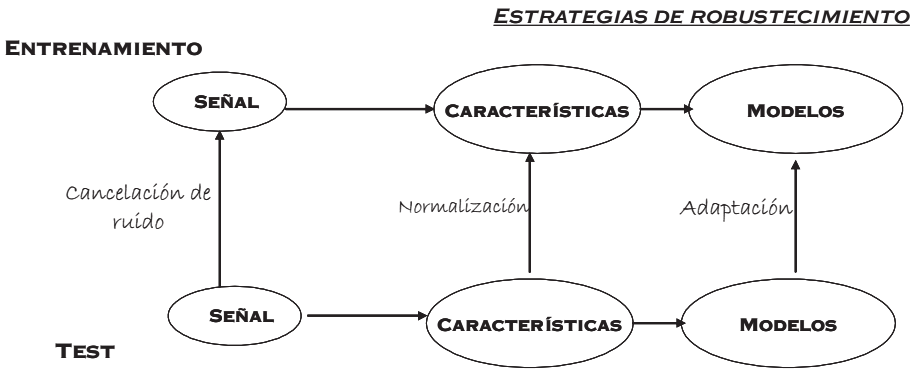


Figura 3.6. Estrategias clásicas de robustecimiento

sea lo más parecida posible a una señal limpia.

ii) Técnicas de compensación de características. La cancelación de la distorsión del ruido se hace una vez que la señal está parametrizada. Mediante diferentes operaciones como puede ser el filtrado cepstral paso alta, el uso de modelos del efecto del ruido, etc., se recuperan en la medida de lo posible los parámetros de voz limpia a partir de los de voz ruidosa.

iii) Técnicas de modificación del clasificador para que la clasificación sea óptima teniendo voz ruidosa. La primera intuición es reentrenar el modelo con datos contaminados. Esta opción podría ser útil aunque tiene un coste computacional, de capacidad de almacenamiento y tiempo elevado. Sería además necesario disponer de datos suficientes en las condiciones de evaluación, lo que no es posible en la mayor parte de las situaciones. En general es más deseable que el sistema se adapte a entornos cambiantes y no conocidos al entrenar y en un tiempo relativamente pequeño. Los mecanismos de adaptación de modelos permiten la modificación de éstos usando una pequeña porción de datos.

El avance del conocimiento de la percepción auditiva, de la capacidad computacional y de almacenamiento de las máquinas y las nuevas necesidades creadas al buscar sistemas de reconocimiento más complejos, han originado otras estrategias como son el reconocimiento multimicrófono, las técnicas de *missing features*, el reconocimiento multibanda, y por supuesto el reconocimiento multimodal, en el que la voz y la imagen se reconocen y sus informaciones se complementan.

3.1.2. Técnicas de cancelación del ruido

El objetivo de las técnicas de cancelación del ruido en la señal es eliminarlo antes de que la señal sea procesada y sometida al reconocimiento. Se basan en la asunción de que la voz y el ruido están incorrelados y son aditivos en el dominio del tiempo por lo cual el espectro de potencias de la señal ruidosa será la suma de los espectros de la voz el ruido. El ruido se considera estacionario, al menos en las tramas en las que se divide la señal para su tratamiento, y es posible estimar su densidad espectral de potencia. Estas técnicas siguen dos pasos que serán descritos:

i) Estimación del espectro del ruido.

ii) Atenuación de dicho espectro de ruido en el espectro de la señal contaminada.

Para una trama t, el espectro de la señal contaminada será la suma del espectro de la señal limpia $S(w)$ más el del ruido $N(w)$:

$$Y(w) = S(w) + N(w) \tag{3.8}$$

Mediante un detector de actividad de voz se separan las tramas de voz ruidosa de aquellas que sólo tienen ruido y se calcula así la estimación del espectro de ruido notada como $|M(t, f)|^2$:

$$|M(t, f)|^2 = \lambda |M(t-1, f)|^2 + (1-\lambda)|X(t, f)|^2 \tag{3.9}$$

siendo $|X(t,f)|^2$ el espectro de la señal ruidosa y λ el factor de actualización del ruido.

La estimación del espectro de voz limpia $Y(t,f)$ partiendo del espectro de voz ruidosa $X(t,f)$ se define con la función denominada *función de ganancia o de pesado*, $G(t,f)$:

$$\begin{aligned} Y(t,g) &= G(t,f) \cdot X(t,f) \\ G(t,f) &= f(X(t,f), N(t,f)) \end{aligned} \tag{3.10}$$

Los distintos algoritmos de atenuación espectral, se basan en diferentes métodos para calcular la función de ganancia de la ecuación 3.10. Una posible clasificación de los mismos es la que los divide en tres grupos: los algoritmos de sustracción espectral (de potencia o de amplitud), el filtrado de Wiener y los algoritmos basados en modelos estadísticos (como pueden ser máxima probabilidad o MMSE). A continuación se describe de un modo resumido su filosofía.

Sustracción Espectral lineal

La expresión general una señal $Y(t)$ sometida a sustracción espectral lineal [15], es la siguiente:

$$|Y(t,f)|^p = \begin{cases} \bullet & |X(t,f)|^p - \alpha|M(t,f)|^p \\ & \quad si \quad |X(t,f)|^p - \alpha|M(t,f)|^p > \gamma|M(t,f)|^p; \\ \bullet & \gamma \cdot |X(t,f)|^p \quad en \quad otro \quad caso \end{cases} \tag{3.11}$$

en la que se ha usado la notación:

- α es un factor de *sobresustracción* que suele tomarse como $1 \leq \alpha \leq 2$ y que compensa el hecho de que el ruido puede subestimarse.
- γ es el umbral mínimo espectral que evita que el espectro estimado sea cero o negativo. Los valores típicos de γ son 0.01 e inferiores

- p es el término exponencial, que puede tomar los valores $p = 1$ si el algoritmo hace sustracción espectral de la magnitud, o $p = 2$ si el algoritmo hace sustracción espectral de potencia.

Sustracción Espectral no lineal

Propuesta por Lockwood en [71]. La sustracción espectral lineal produce en ocasiones un ruido musical que se debe a la suposición de que el ruido es estacionario y de varianza cero. Esto no siempre es cierto del todo y se pueden producir estimaciones negativas del espectro de voz, sobre todo para SNRs bajas. La sustracción espectral no lineal propone un factor de *sobresustracción* que depende de la SNR. Para SNRs bajas el algoritmo resta el ruido más alto de las M últimas tramas, produciéndose sobresustracción. Para SNRs altas se resta una estimación suavizada del ruido (no hay sobresustracción). La ganancia $G(t, f)$ tiene la siguiente expresión:

$$G(t,f) = \frac{|\overline{X}(t,f)| - \alpha(t,f) \cdot (1 - sigmoide(\frac{|\overline{X}(t,f)|}{|\overline{N}(t,f)|}))(1 - \frac{|\overline{N}(t,f)|}{\alpha(t,f)})}{|\overline{X}(t,f)|} \quad (3.12)$$

donde

$$\begin{aligned} \alpha(t,f) &= \max_{Mtramas}\{N(t,f)\} \\ \overline{X}(t,f) &= \lambda_v \cdot X(t,f) + (1-\lambda_v) \cdot \overline{X}(t,f) \\ \overline{N}(t,f) &= \lambda_r \cdot N(t,f) + (1-\lambda_r) \cdot \overline{N}(t,f) \end{aligned} \quad (3.13)$$

siendo λ_v y λ_r los factores de actualización de la voz y el ruido respectivamente, y $\alpha(w)$ el factor de sobre-estimación del ruido que ahora es función de la SNR (es función del ruido de las M tramas anteriores).

Filtrado de Wiener

La estimación del espectro de señal sin ruido se calcula como una versión filtrada de la señal ruidosa, con un filtro de Wiener que obedece a la ecuación 3.14, en la que γ es de nuevo el umbral espectral mínimo para evitar espectros negativos o nulos:

$$H(t,f) = \frac{max\{E[|\overline{X}(t,f)|^2] - |[\overline{N}(t,f)|^2], \quad \gamma \cdot E[|\overline{X}(t,f)|^2]\}}{E[|\overline{X}(t,f)|^2]} \tag{3.14}$$

La diferencia entre el filtrado de Wiener y la sustracción espectral de potencias (ecuación 3.11) es que el filtrado de Wiener usa los valores esperados y la sustracción espectral usa los valores instantáneos. Son similares en la práctica aunque las filosofías subyacentes difieren. El filtrado de Wiener se puede hacer de manera iterativa.

Modelos estadísticos

Utilizan una aproximación estadística para calcular la función de ganancia. Las dos técnicas estadísticas más relevantes para la atenuación espectral son la estimación de Máxima Probabilidad y la regla de supresión del ruido de Ephraim-Malah. A continuación se esbozan.

- Estimación de Máxima probabilidad. Si se maximiza la probabilidad de la señal de voz ruidosa $X(t,f)$ dada la señal con cancelación de ruido $Y(t,f)$, se obtiene la siguiente expresión de la ganancia o función de pesado:

$$G(w) = \frac{1}{2}[1 + \sqrt{1 - \frac{\hat{\mu}^2(w)}{|Y(w)|^2}}] \tag{3.15}$$

- Regla de supresión del ruido de Ephraim-Malah [35]. Este algoritmo estima el espectro de ruido minimizando el error cuadrático medio los espectros logarítmicos de la señal de voz ruidosa $X(t,f)$ y la señal

de voz sometida a la cancelación de ruido $Y(t, f)$. La ecuación 3.16 minimizada:

$$E[(log_{10}|X(t,f)^2| - log_{10}|Y(t,f)|^2] \tag{3.16}$$

es la condición que define la función de pesado $G(t, f)$ que tiene la expresión:

$$G(t,f) = \frac{\pi}{2}\sqrt{(\frac{1}{1+SNR_{post}})(\frac{SNR_{prio}}{1+SNR_{prio}})}\cdot M[(1+SNR_{post})(\frac{SNR_{prio}}{1+SNR_{prio}})] \tag{3.17}$$

Siendo $M[\Theta]$ la función:

$$M[\Theta] = e^{-\Theta/2}\cdot[(1+\Theta)\cdot I_0(\frac{\Theta}{2}+\Theta\cdot I_1)(\frac{\Theta}{2})] \tag{3.18}$$

en la que I_0 y I_1 representan las funciones de Bessel modificadas de orden 0 y 1. Las expresiones de la SNR a priori y a posteriori se definen del siguiente modo:

$$SNR_{post} = \frac{|X(t,f)^2|}{|N(t,f)|^2} - 1 \tag{3.19}$$

$$SNR_{prior} = (1-\alpha)\cdot max\{SNR_{post}, 0\} + \alpha\frac{|H(t-1,f)\cdot X(t-1,f)|^2}{|N(t,f)|^2} \tag{3.20}$$

3.1.3. Compensación de características

Las técnicas de compensación de características operan sobre las características parametrizadas para recuperar en la medida de lo posible los vectores de características limpias. Esta familia de técnicas se puede dividir en tres grupos basándose en el método de compensación utilizado:

a) Técnicas de filtrado cepstral paso alta.

b) Técnicas de compensación con datos estéreo.

c) Técnicas de compensación con modelos del entorno.

d) Estrategias de encuadre estadístico.

Filtrado cepstral paso alta

Las técnicas de filtrado cepstral paso alta añaden un robustecimiento frente al ruido bastante elevado, con un coste computacional muy bajo. Por esta razón se incluyen en todos los *front-ends* de reconocimiento automático. El objetivo de estos algoritmos es forzar que los valores medios de los coeficientes cepstrales sean cero. De este modo compensan los efectos de filtrado lineal desconocidos que pueda tener el canal. En este conjunto de técnicas destacamos los clásicos CMN y RASTA que son descritos a continuación:

[RASTA] R*elative* S*pectral* A*mplitude, [54]*:

Esta técnica suprime mediante un filtro IIR paso banda los componentes de la señal que varían más despacio o más rápido que la señal de voz, reforzando con ello los más relevantes para la tarea de reconocimiento. El filtrado se hace bien en el dominio del logaritmo espectral de potencias, o en el dominio cepstral, y produce beneficios tanto en presencia de ruido aditivo como convolucional. La expresión de un filtro RASTA es:

$$H(z) = 0{,}1 \cdot z^4 \cdot \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0{,}98z^{-1}} \tag{3.21}$$

En la figura 3.7 se observa su respuesta en frecuencias.

[CMN] C*esptral* M*ean* N*ormalization, [37]*:

Esta técnica de normalización de la media cepstral consiste en calcular el valor medio de cada coeficiente cepstral en un segmento de tiempo de $2N + 1$ tramas, y normalizarlo sustrayendo ese valor medio. Con esta operación se elimina el filtrado lineal del canal de modo simple, con resultados muy parecidos a los de RASTA, siendo las prestaciones superiores

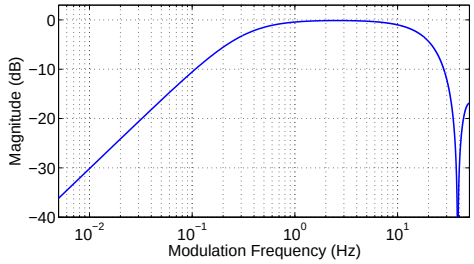


Figura 3.7. Respuesta en frecuencias del filtro RASTA

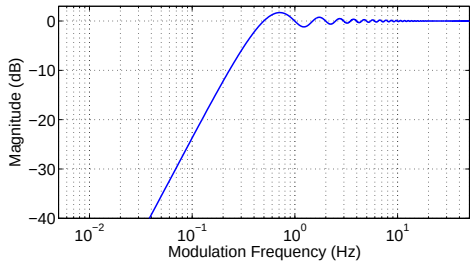


Figura 3.8. Respuesta en frecuencias del filtro que implementa CMN

para condiciones de entrenamiento y test no especialmente diferentes. La normalización de los coeficientes cepstrales se hace siguiendo la ecuación:

$$C_{\hat{x}}[m] = C_x[m] - \frac{1}{2N+1} \cdot \sum_{l=-N}^{N} C_x[m+l] \tag{3.22}$$

La figura 3.8 muestra la respuesta en frecuencia de un filtro FIR que implementa la sustracción cepstral de la media.

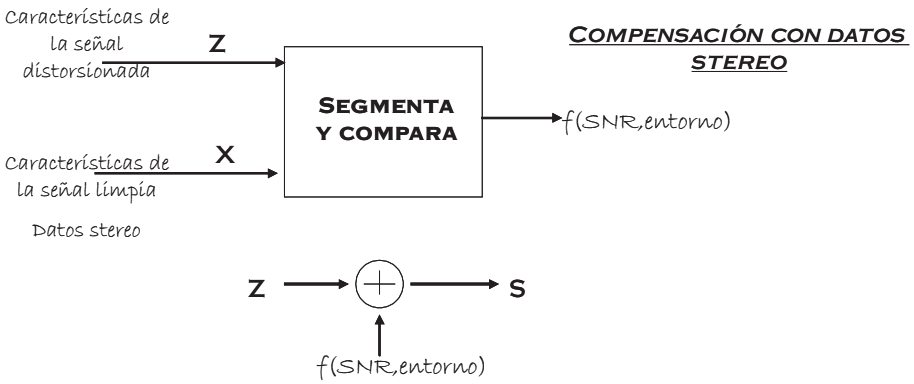


Figura 3.9. Compensación de características con datos estéreo

Compensación con datos estéreo

Las técnicas de compensación de características con datos estéreo siguen dos pasos:

i) En primer lugar, comparan las características de la voz ruidosa que se quiere compensar, con datos limpios estéreo.

ii) El resultado de esa comparación es una corrección del entorno que se suma al vector de características que son así compensadas antes de pasar al reconocedor (ver figura 3.9).

Los datos estéreo pueden obtenerse en una sesión previa al reconocimiento, o con un micrófono cercano, etc. A continuación se describen brevemente los algoritmos de compensación con datos estéreo más relevantes :

[RATZ] *multiva*R*iate g*A*ussian based cepsT*ral normali*Z*ation, [82]*:

Este algoritmo de compensación de características trabaja sobre dos hipótesis de partida:

- Las estadísticas de la voz limpia se pueden representar mediante una mezcla de M Gausianas.

- Los efectos del entorno en las estadísticas de la voz limpia se pueden modelar como compensaciones en los vectores media y varianza de la mezcla de Gausianas anteriormente referida.

El algoritmo sigue tres pasos. En primer lugar calcula las estadísticas de la voz limpia, modelando la *pdf* de las características de los datos estéreo como una mezcla de vectores Gausianos multivaluados. Para una señal limpia x tendremos la expresión:

$$\begin{gathered} x = [x_0 \quad ... \quad x_{p-1} \quad x_p] \\ p(x) = \sum_{k=0}^{M-1} P[k] \cdot N_x(\mu_{x,k}, \Sigma_{x,k}) \end{gathered} \tag{3.23}$$

en la que $P[k]$, $\mu_{x,k}$ y $\Sigma_{x,k}$ representan la probabilidad a priori, vector media y matriz covarianza de cada elemento K de la mezcla de Gausianas.

El segundo paso de RATZ es calcular las estadísticas de los datos de voz contaminada, z. El efecto del entorno en las estadísticas de la voz limpia se modela con los parámetros r_k y R_K:

$$\begin{aligned} \mu_{z,k} &= \mu_{x,k} + r_k \\ \Sigma_{z,k} &= \Sigma_{x,k} + R_k \end{aligned} \tag{3.24}$$

Los valores de r_k y R_k se calculan usando el criterio de máxima probabilidad. Se define la función de probabilidad $L(Z)$ sobre los vectores cepstrales ruidosos observados z_i, dado θ que representa a los términos r_k y R_k

que se quieren optimizar:

$$L(Z = z_0 \quad ... \quad z_i \quad ... \quad z_{N-1}|\Theta) = \sum_{i=0}^{N-1} log(p(z_i|\Theta)) \tag{3.25}$$

Esta maximización se resuelve utilizando los datos limpios estéreo, y origina las siguientes expresiones (ecuación 3.26) para los factores de corrección del entorno. Los factores de corrección para la media y la varianza de cada una de las M componentes del modelo de Gausianas de los datos limpios son:

$$\begin{aligned} r_k &= \frac{\sum_{i=o}^{N-1}(z_i - x_i) \cdot P(k|x_i)}{\sum_{i=0}^{N-1} P(k|x_i)} \\ R_k &= \frac{\sum_{i=0}^{N-1}(z_i - \mu_{z,k}) \cdot (z_i - \mu_{z,k})^T \cdot P(k|x_i)}{\sum_{i=0}^{N-1} P(k|x_i)} \end{aligned} \tag{3.26}$$

En tercer lugar, las características normalizadas usadas en el reconocimiento serán las $\hat{x}$ que tienen como media y varianza de sus mezclas de Gausianas la siguiente expresión:

$$\begin{aligned} z &= x + r(x) \\ \hat{x} = E[x|z] = \int_X x \cdot p(x|z)dx &= z - \int_X r(x) \cdot p(x|z)dx \\ \hat{x} &\simeq z - \sum_{k=0}^{M-1} P(k|z) \cdot r_k \end{aligned} \tag{3.27}$$

[SPLICE] *Stereo-based Piecewise LInear Compensation for Environments* [31]:

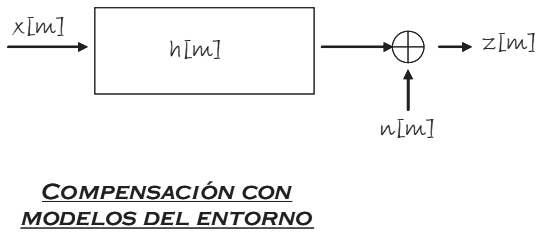


Figura 3.10. Compensación con modelos del entorno

Este algoritmo al igual que RATZ, utiliza los datos estéreo limpios para definir una función que origine características limpias. Hay autores que definen SPLICE como la imagen especular de RATZ en el dominio de la voz ruidosa. El primer paso de SPLICE es definir los vectores contaminados Z, con un modelo de mezclas de M Gausianas:

$$P(z) = \sum_{k=0}^{M-1} P(k) \cdot N(\mu_{z,k}, \Sigma_{z,k}) \tag{3.28}$$

El segundo paso es definir la probabilidad a posteriori de los vectores limpios X dados los vectores ruidosos Z y las M Gausianas del modelo de mezclas usado para su definición:

$$P(x|z,k) = N(x, z + r_k) \tag{3.29}$$

El término r_k se define como factor de corrección por Gausiana. Para determinarlo se maximiza la probabilidad a posteriori de los datos limpios dados los ruidosos, usando los datos estéreo limpios con los que se cuenta. Aplicando el algoritmo de máxima probabilidad se llega a la siguiente expresión de r_k:

$$r_k = \frac{\sum_{t=0}^{T-1} p(k|z_t) \cdot (x_t - z_t)}{\sum_{t=0}^{T-1} (p(k|z_t)} \tag{3.30}$$

Con este resultado, la estimación final de los coeficientes cepstrales limpios será:

$$\hat{x} = z - \sum_{k} P(k|z) \cdot r_k \tag{3.31}$$

Compensación con modelos del entorno

La filosofía de estos algoritmos es dar una función analítica de la naturaleza de la degradación del entorno, necesitando pocos datos empíricos para lograr la compensación de las características (término opuesto de los algoritmos de compensación con datos estéreo). Definen la degradación como un filtro y un ruido tales que cuando se aplican de manera inversa maximicen la probabilidad de las observaciones así compensadas (ver figura 3.10). El algoritmos más relevante de esta familia es *VTS*:

[VTS] V*ector* T*aylor* S*eries approach* [83]:

El objetivo de VTS es estimar la función de densidad de probabilidad de la voz ruidosa, dada la función de densidad de probabilidad de la voz limpia, un segmento de voz ruidosa, y el desarrollo en serie de Taylor que relaciona la voz limpia con la voz ruidosa. Una vez que la *pdf* de la voz ruidosa se ha calculado, se hace una estimación basada en el error cuadrático medio mínimo para predecir la secuencia de voz limpia no observada.

Utilizando el modelo del entorno acústico que hemos descrito en la sección 3.1.1 de este capítulo, la expresión de una señal contaminada en el dominio logarítmico del banco de filtros sería la de la ecuación 3.32, en la que se ignora el efecto del canal h para simplificar el análisis (ver [83] para más datos):

$$\hat{y} = \hat{x} + g(\hat{n} - \hat{x}) \tag{3.32}$$

siendo $g(\hat{n}, \hat{x})$ el término aditivo que representa el ruido en el dominio log-FBE, y $g(z)$ una función definida como $g(z) = ln(1 + e^{C^{-1}(z)})$.

Como función auxiliar para el desarrollo en serie de Taylor que utiliza

VTS, se define la función $f_i(x, n)$ para el canal i-ésimo del banco de filtros:

$$f_i(x, n) \equiv \frac{1}{1 + exp(x(i) - n(i))} \tag{3.33}$$

de tal modo que las derivadas parciales de la función $g_i(x, n)$, se pueden expresar como:

$$\begin{aligned} \frac{\partial g(i)}{\partial x(i)} = -f_i \qquad \frac{\partial g(i)}{\partial n(i)} = f_i \\ \frac{\partial^2 g(i)}{\partial^2 x(i)} = \frac{\partial^2 g(i)}{\partial^2 n(i)} = (1 - f_i) \cdot f_i = v_i \\ \frac{\partial^2 g(i)}{\partial x(i) \partial n(i)} = \frac{\partial^2 g(i)}{\partial n(i) \partial x(i)} = -(1 - f_i) \cdot f_i = -v_i \end{aligned} \tag{3.34}$$

La teoría del desarrollo en serie de Taylor es que una función infinitamente derivable como lo es $g(x(i), n(i))$ evaluada en el punto $(x_0(i), n_0(i))$se puede aproximar por su desarrollo en serie con la expresión:

$$\begin{aligned} g(x(i), n(i)) \approx g(x_0(i), n_0(i)) + \frac{\partial}{\partial x} g(x_0(i), n_0(i),) \cdot (x(i) - x_0(i)) + \\ + \frac{\partial}{\partial n} g(x_0(i), n_0(i)) \cdot (n(i) - n_0(i)) + \quad ... \end{aligned} \tag{3.35}$$

La ecuación 3.35 tiene infinitos términos de desarrollo. En el caso de distribuciones Gausianas como la que nos ocupa es suficiente con que la función se expanda con exactitud solamente en un espacio pequeño alrededor de la media del vector. Este factor se aprovecha para truncar el desarrollo después del primer o segundo orden.

El primer paso del algoritmo VTS es encontrar los parámetros estadís-

ticos de la distribución de voz ruidosa $y(i)$, dada la distribución Gausiana de los datos limpios $x(i)$ modelados como una mezcla de M Gausianas con medias $\mu_{x(i),m}$ y varianza $\Sigma_{x(i),m}$. Para un desarrollo de Taylor de orden uno, la media y varianza de $y(i)$ tendrán la expresión:

$$\mu_y(i) = E(x(i) + g(x_0(i), n_0(i)) = \mu_x(i) + E(g(x_0(i), n_0(i)) + \\ +E(\frac{\partial}{\partial x(i)} g(x_0(i), n_0(i)) \cdot (x(i) - x_0(i))) + E(\frac{\partial}{\partial n(i)} g(x_0(i), n_0(i)) \cdot (n(i) - n_0(i))) + \tag{3.36}$$

Del mismo modo, la expresión de la matriz de covarianzas será:

$$\Sigma_y(i,j) = (I + \frac{\partial}{\partial x(i)} g(x_0(i), n_0(i)))^T \cdot \Sigma_x(i,j) \cdot (I + \frac{\partial}{\partial x(i)} g(x_0(i), n_0(i))) + \\ +(\frac{\partial}{\partial x(i)} g(x_0(i), n_0(i)))^T \cdot \Sigma_n(i,j) \cdot \frac{\partial}{\partial x(i)} g(x_0(i), n_0(i)) \tag{3.37}$$

siendo $\Sigma_n(i,j)$ la varianza del ruido, que al igual que la estadística $\mu_n(i)$ se habrá calculado a partir de la secuencia de la señal ruidosa $y(i)$.

Si aplicamos las ecuaciones 3.34 a la expresión de $\mu_y(i)$ y $\Sigma_y(i,j)$, tendremos los valores de la media y varianza de la señal contaminada:

$$\mu_y(i) \approx \mu_x(i) + g(x_0(i), n_0(i)) + \frac{1}{2} v(x_0(i), n_0(i)) \cdot [\Sigma_x(i,j) + \Sigma_n(i,j)] \\ \Sigma_y(i,j) \approx (1 - f(x_0(i), n_0(i)) \cdot (1 - f(x_0(j), n_0(j)) \cdot \Sigma_x(i,j) \\ +f(x_0(i), n_0(i)) \cdot f(x_0(j), n_0(j)) \Sigma_n(i,j) \tag{3.38}$$

Las ecuaciones 3.36 y 3.37 se aplican de manera iterativa con EM, hasta obtener unos valores convergentes de la probabilidad de los datos ruidosos

observados:

i) Se obtiene un valor inicial de μ_n y Σ_n.

ii) Se expande la función $g(x, n)$ alrededor del vector de medias de cada Gausiana de la distribución de x, $\mu_{x,m}$ y la estimación de μ_n.

iii) Se estiman los parámetros de la distribución contaminada y: $\mu_{y,m}$, $\Sigma_{y,m}$.

iv) Se hace una reestimación con el algoritmo EM [30] para recalcular los valores de μ_n y Σ_n.

v) Si la probabilidad de los datos ruidosos observados no ha convergido, se repite el paso 2.

Una vez que los parámetros de la distribución contaminada se han calculado, se estiman los vectores de voz limpia usando el criterio de error cuadrático medio mínimo:

$$\begin{aligned} \hat{x}_{MMSE} = E(x|y) &= \int_{-\infty}^{+\infty} x \cdot p(x|y)dx \\ \hat{x}_{MMSE} = E(x|y) &= \int_{-\infty}^{+\infty} (y - g(x,n)) \cdot p(x|y)dx \end{aligned} \tag{3.39}$$

El modelo de voz limpia es una modelo de mezcla de M Gausianas con lo que la última ecuación de 3.39 puede escribirse como:

$$\begin{aligned} \hat{x}_{MMSE} = E(x|y) &= \int_{-\infty}^{+\infty} (y - g(x,n)) \cdot p(x|y)dx \\ &= y - \int_{-\infty}^{+\infty} \sum_{k=0}^{M-1} g(x,n) \cdot p(x,k|y)dx \\ &= y - \sum_{k=0}^{M-1} p(k|y) \cdot \int_{-\infty}^{+\infty} g(x,n) \cdot p(x|y,k)dx \end{aligned} \tag{3.40}$$

El término $g(n,x)$ se puede sustituir por su valor promedio si la desviación estándar de la Gausiana es pequeña:

$$\int_{-\infty}^{+\infty} g(x,n)\cdot p(x|y,k)dx \approx g(\mu_{x,k},\mu_n)\int_{-\infty}^{+\infty} p(x|y,k)dx \tag{3.41}$$

La integral de la expresión 3.41 vale la unidad, por ser la integral de una distribución de probabilidad. De este modo la expresión de la señal limpia x en función de la señal distorsionada y y del modelo de la señal limpia será:

$$\hat{x}(y) \approx y - \sum_{k=0}^{M-1} p(k|y)\cdot g(\mu_{x,k},\mu_n) \tag{3.42}$$

siendo la probabilidad de las distintas Gausianas dada una observación de voz distorsionada:

$$p(k|y) = \frac{p(k)\cdot N(y,\mu_{y,k}\Sigma_{y,k})}{\sum_{k'=1}^{K} p(k')\cdot N(y,\mu_{y,k'},\Sigma_{y,k'})} \tag{3.43}$$

VTS da resultados bastante buenos, a pesar de tener un coste computacional apreciable, cuando los ruidos son estacionarios. Sin embargo su limitación aparece al tratar ruidos no estacionarios o transitorios. Las alternativas más recientes para tratar estos ruidos transitorios son el reconocimiento sub-banda y la aproximación de las características ausentes (*Missing Features approach*) que serán analizadas posteriormente.

Estrategias de encuadre estadístico

Este conjunto de algoritmos de normalización de características, definen transformaciones lineales o no lineales para modificar las estadísticas de las características de voz ruidosa y hacerlas coincidir con las de una referencia limpia. La normalización cepstral de la media, que ha sido clasificada anteriormente como técnica de filtrado cepstral paso alta, también

puede incluirse en esta categoría. A continuación describimos las estrategias de encuadre estadístico más relevantes:

i) CMVN C*epstral* M*ean and* V*ariance* N*ormalization*, [131]:

La normalización de la media y la varianza en el dominio cepstral es una ampliación de la normalización de la media *CMN* definida en el grupo de técnicas de filtrado cepstral paso alta. El efecto aditivo del ruido supone un desplazamiento de la media de las densidades de probabilidad Gausiana de los MFCCs, y un escalado de la varianza de las mismas. Si llamamos y a la señal cepstral distorsionada con un ruido aditivo h, y x a la señal cepstral limpia de media μ_x y varianza σ_x, el efecto del desplazamiento de la media h y el escalado de la varianza α queda expresado por:

$$\begin{aligned} y &= \alpha \cdot x + h \\ \mu_y &= \alpha \cdot \mu_x + h \\ \sigma_y &= \alpha \cdot \sigma_x \end{aligned} \tag{3.44}$$

Las señales x e y normalizadas en media y varianza tendrán la expresión:

$$\begin{aligned} \hat{x} &= \frac{x - \mu_x}{\sigma_x} \\ \hat{y} &= \frac{(y - \mu_y)}{\sigma_y} \\ &= \frac{((\alpha \cdot x + h) - (\alpha \cdot \mu_x + h))}{\alpha \cdot \sigma_x} = \hat{x} \end{aligned} \tag{3.45}$$

ii) Normalización de un determinado número de momentos:

Una extensión natural de la técnica de *CMVN* es normalizar un número superior de momentos estadísticos como punto intermedio entre normalizar sólo dos (*CMVN*) y normalizarlos todos como es el caso de *HEQ* que será descrito a continuación. En el 2004, Khademul añade en [80] los cuatro primeros momentos de los MFCCs al conjunto de parámetros

usados en el reconocimiento automático obteniendo con ello beneficios en la tasa de reconocimiento y haciendo que el sistema converja más rápidamente. También en el 2004 Chang-Wen Hsu propone en [135] un método de normalización de los momentos cepstrales de orden más alto en el que además de la media, se puede normalizar un coeficiente de orden par y otro de orden impar consiguiendo resultados beneficiosos a partir del momento de orden 50 de la distribución original. Las exploraciones en esta dirección se han limitado a buscar aproximaciones paramétricas de la ecualización de ciertos momentos de la distribución que no han sido más de 3 simultáneamente, y que conllevan un coste computacional que no los hace atractivos frente a la Ecualización de Histogramas [97].

iii) HEQ *Histogram Equalization*, [26],[59]:

La transformación lineal en CMVN elimina únicamente los efectos lineales del ruido. La distorsión no lineal producida por el entorno, afecta no sólo a la media y la varianza de las distribuciones de probabilidad, sino también a los momentos de orden superior. La técnica de Ecualización de Histogramas propone generalizar la normalización de los dos primeros momentos de la *pdf* que hacía CMNV, trasformando la *función de distribución acumulada* de los coeficientes cepstrales para hacerla coincidir con la de los datos limpios. Ambas *CDFs* se transforman a una referencia común. Esta técnica tiene como atractivos el ser una técnica de bajo coste computacional y de almacenamiento, y el no usar datos estéreo ni suposición alguna sobre el tipo de ruido que se quiere eliminar, por lo que es útil además para combatir el ruido residual de otras normalizaciones basadas en modelos del efecto del entorno como es por ejemplo VTS, [117]. El capítulo 6 de esta tesis hace un estudio exhaustivo de la Ecualización de Histogramas, analizando sus ventajas y limitaciones y proponiendo estrategias para superar estas últimas.

3.1.4. Adaptación de modelos

Esta alternativa de robustecimiento consiste en adaptar los modelos acústicos obtenidos durante la fase de entrenamiento del sistema, a

las condiciones de evaluación. Los modelos son entrenados con señales limpias adquiridas en un entorno sin distorsiones y se adaptan a las nuevas condiciones de evaluación con datos de adaptación del entorno ruidoso. Estas técnicas se usan tanto para adaptación a entornos que difieren del de entrenamiento, como para adaptación a locutores. Las técnicas de adaptación de modelos más utilizadas son de dos tipos: **adaptación estadística de los modelos** mediante *MLLR*, *MAP* o combinaciones de ambos, y **descomposición de modelos en paralelo**, o *PMC*.

La adaptación mediante transformaciones estadísticas puede ser supervisada si se cuenta con las transcripciones de los datos de adaptación, o no supervisada si no se cuenta con las transcripciones de los datos de adaptación. Se puede hacer además de modo estático, usando los datos de adaptación antes de empezar el reconocimiento, o de modo incremental, adaptando a medida que avanza el reconocimiento de la señal de voz y se tienen resultados.

MLLR

Regresión Lineal de Máxima Probabilidad [40],[36]. Define un conjunto de transformaciones lineales de las medias y las varianzas de las Gausianas que forman los HMMs, con el objetivo de igualarlas a las de los datos de adaptación. El efecto de estas transformaciones es un desplazamiento en las medias y una alteración en las varianzas del sistema inicial de tal modo que cada estado del HMM tenga la máxima probabilidad de generar los datos de adaptación.

Existen dos versiones de MLLR: una llamada *MLLR con restricciones* y otra llamada *MLLR sin restricciones*. MLLR con restricciones busca una sola transformación común para la media y la varianza de las Gausianas de los modelos, que siendo única maximice la probabilidad de los datos de adaptación en los modelos adaptados. MLLR sin restricciones define dos transformaciones lineales: una primera transformación de las medias de los modelos M, que origina unos modelos $\hat{M}$ adaptados en media, y una segunda transformación de las varianzas, independiente de la primera

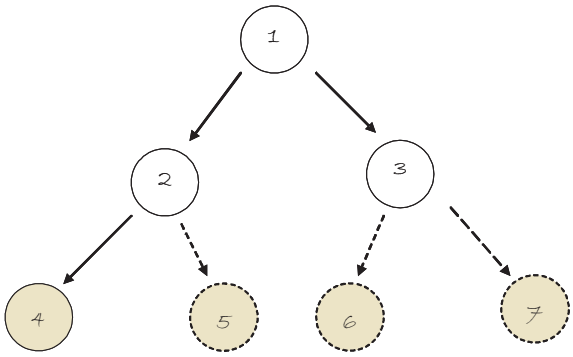


Figura 3.11. Ejemplo de árbol de regresión

y opcional, que origina unos modelos $\tilde{M}$ adaptados en media y varianza. Las máxima probabilidad para los datos de adaptación O_T es mayor en el caso de adaptar ambos parámetros de las Gausianas del modelo:

$$L(O_T|\tilde{M}) \geq L(O_T|\hat{M}) \geq L(O_T|M) \tag{3.46}$$

En el caso de MLLR sin restricciones, si llamamos ξ a la matriz extendida de las medias de las Gausianas del modelo, en la que el parámetro ω representa el desplazamiento del sesgo [36]:

$$\xi = [w \quad \mu_1 \quad \mu_2 \quad ... \quad \mu_n]^T \tag{3.47}$$

el objetivo del algoritmo es buscar una matriz de transformación W que origine la matriz de medias adaptadas $\hat{\mu}$:

$$\hat{\mu} = W \cdot \xi \tag{3.48}$$

De este modo se puede definir una transformación global para cada Gausiana. Sin embargo, si el número de datos de adaptación es suficiente,

se pueden agrupar las Gausianas en clases de regresión mediante un *árbol de clases de regresión*, definiéndose una transformación W_m para cada una de las clases en vez de la transformación general de la ecuación 3.48. Cada una de estas transformaciones será más específica, obteniéndose un grado mayor de adaptación. El árbol se construye con un algoritmo de división de centroides, que utiliza distancias euclídeas para agrupar Gausianas que representan componentes próximos entre sí en el espacio acústico. Mediante divisiones sucesivas, las Gausianas se agrupan hasta llegar a los nodos terminales que especifican las agrupaciones finales denominadas clases de regresión principales. Cada Gausiana presente en el modelo pertenece a una de estas clases. La figura 3.11 muestra un ejemplo de árbol de regresión con 4 clases terminales C4, C5, C6 y C7. Las clases cuyo círculo tiene un contorno de línea continua son las que tienen suficientes datos para formar una clase de regresión principal. Para las clases 2, 3 y 4 se definirá una transformación de adaptación.

Si tenemos R clases de regresión de Gausianas, $R = \{m_1, m_2...m_R\}$, la transformación definida por clase W_m cumple:

$$\hat{\mu}_m = W_m \cdot \xi_m \tag{3.49}$$

La matriz W_m se obtiene resolviendo la maximización de la probabilidad de la media transformada mediante la técnica *EM* (*Expected Maximization*) que maximiza la función auxiliar definida en 3.50 respecto a W_m:

$$Q(M, \bar{M}) = \sum_{Q} P(Q|X, M) \cdot log(P(X, Q|\bar{M})) \tag{3.50}$$

La expresión 3.50 será máxima cuando W_m cumpla:

$$\sum_{t=1}^{T} \sum_{r=1}^{R} L_{m_r}(t) \Sigma_{m_r}^{-1} \cdot O(t) \cdot \xi_{m_r}^{T} = \sum_{t=1}^{T} \sum_{r=1}^{R} L_{m_r}(t) \Sigma_{m_r}^{-1} \cdot W_m \cdot \xi_{m_r} \cdot \xi_{m_r}^{T} \tag{3.51}$$

En esta ecuación $O(t) = \{o(1), ..., o(T)\}$ representan los datos de adaptación y $L_{m_r}(t)$ será la probabilidad de ocupación de la clase de regresión m_r en el instante t. Si llamamos $q_{m_r}(t)$ a la componente Gausiana m_r en t y M al modelo HMMs original, $L_{m_r}(t)$ será:

$$L_{m_r}(t) = p(q_{m_r}(t)|M, O_T) \tag{3.52}$$

Se definen dos términos auxiliares para resolver 3.51 y dar una expresión de W_m:

En primer lugar, la parte izquierda de la ecuación 3.51 que será independiente de la matriz de transformación, la denominamos Z:

$$Z = \sum_{t=1}^{T} \sum_{r=1}^{R} L_{m_r}(t) \Sigma_{m_r}^{-1} \cdot O(t) \cdot \xi_{m_r}^{T} \tag{3.53}$$

En segundo lugar, se define una nueva variable G^i, que será una matriz de elementos:

$$g_{jq}^{i} = \sum_{r=1}^{R} v_{ii}^{r} \cdot d_{jq}^{r} \tag{3.54}$$

donde

$$V^r = \sum_{t=1}^{T} L_{m_r}(t) \cdot \Sigma_{m_r}^{-1} \tag{3.55}$$

y

$$D^r = \xi_{m_r} \cdot \xi_{m_r}^{T} \tag{3.56}$$

Usando las variables auxiliares de 3.53 y 3.54 con la ecuación 3.51, los términos de la matriz de transformación W_m se calculan en la expresión 3.57 en la que w_i y z_i son los vectores i-ésimos de Z:

$$w_i^T = G_i^{-1} \cdot z_i^T \tag{3.57}$$

En el caso de MLLR no restringido, las transformaciones de la media y

la varianza se calculan de manera independiente. La adaptación en varianzas se lleva a cabo una vez hecha la adaptación en medias, aumentando con ello la máxima probabilidad de los datos adaptados en el modelo adaptado. Una vez calculada W_m, la matriz de covarianzas de las Gausianas se transforma haciendo:

$$\hat{\Sigma}_m = B_m^T \cdot H_m \cdot B_m \tag{3.58}$$

donde H_m es la transformación lineal para las covarianzas que se busca y B_m es el factor Choleski de Σ_m^{-1}:

$$\Sigma_m^{-1} = C_m \cdot C_m^T \tag{3.59}$$

y

$$B_m = C_m^{-1} \tag{3.60}$$

La probabilidad de las Gausianas con la covarianza transformada Σ_m^{-1} será máxima cuando la matriz de transformación H_m tenga la expresión:

$$H_m = \frac{\sum_{r=1}^{R} C_{m_r}^T \cdot [L_{m_r}(t) \cdot (o(t) - \mu_{m_r})(o(t) - \mu_{m_r})^T]}{L_{m_r}(t)} \tag{3.61}$$

Para encontrar esta expresión de la transformación lineal de las covarianzas es necesario utilizar matrices de covarianzas diagonales. En caso de que estas no lo sean, hay métodos computacionalmente más complejos que dan soluciones iterativas.

Para el caso de MLLR con restricciones, el proceso es similar al descrito hasta este punto. Se busca A para definir una transformación única para media y varianza de las Gausianas de los modelos:

$$\begin{aligned} \hat{\mu} &= A^{-1} \cdot \xi - b^T \\ \hat{\Sigma}_m &= A^{-1} \cdot A^{T-1} \end{aligned} \tag{3.62}$$

y llegando a las siguientes expresiones de las media y varianza transformadas:

$$\hat{\mu}_m = \frac{\sum_{t=1}^{T}\sum_{r=1}^{R} L_{m_r}(t)\cdot o(t)}{\sum_{t=1}^{T}\sum_{r=1}^{R} L_{m_r}(t)} \tag{3.63}$$

$$\hat{\Sigma}_m = \sum_{t=1}^{T}\sum_{r=1}^{R} L_{m_r}(t)\cdot(o(t)-\hat{\mu}_m)\cdot(o(t)-\hat{\mu}_m)^T \tag{3.64}$$

MAP

Adaptación de Máximos a Posteriori, o adaptación Bayesiana [43]. Este método basa la adaptación de los modelos en información a priori sobre la distribución del modelo adaptado, que se obtiene de los datos de adaptación. MLLR buscaba maximizar la probabilidad de que los datos de adaptación $\{O_T\}$ se reconociesen con el modelo adaptado $\hat{M}$ como expresa 3.65:

$$\hat{M} = \underset{M}{max}\{L(O_1^T|M)\} \tag{3.65}$$

Lo que busca maximizar la adaptación bayesiana de modelos, es la probabilidad a posteriori del modelo $\hat{M}$ dados los datos $\{O_T\}$:

$$\hat{M} = \underset{M}{max}\{P(M|O_1^T)\} \tag{3.66}$$

La ecuación 3.66 se puede expresar en función de 3.65 usando la regla de Bayes:

$$\hat{M} = \underset{M}{max}\{P(M|O_1^T)\} = \underset{M}{max}\{L(O_1^T|M)\cdot p(M)\} \tag{3.67}$$

En 3.67, $p(M)$ es la probabilidad a priori de que el modelo M sea el que se utilice y la probabilidad maximizada con el planteamiento es $P(M|O_1^T)$, es decir, la probabilidad a posteriori de que M sea el modelo optimo para $\{O_T\}$. MAP se basa por tanto en transformar la probabilidad del modelo M, usando los datos de adaptación, para que sea la óptima dados esos datos. (La expresión 3.67 indica que si dicha probabilidad se desconociese, MAP

se convertiría en MLLR). Haciendo uso de la condición establecida se llega a la siguiente expresión para la media de las probabilidades a priori de las Gausianas del modelo adaptado:

$$\hat{\mu} = \frac{N}{N+\tau} \cdot \mu_o + \frac{\tau}{N+\tau} \cdot \mu_p \tag{3.68}$$

donde μ_o es la media de los datos de adaptación calculada usando máxima probabilidad y μ_p representa la media original del modelo sin adaptar. El parámetro τ representa un factor de pesado del conocimiento a priori en las medias de los datos de adaptación, y se regula para optimizar distintos entornos de evaluación.

La expresión de la varianza de las probabilidades a priori será:

$$\hat{\sigma}^2 = \frac{N}{N+\alpha+1} \cdot \sigma_o^2 + \frac{\tau \cdot (\mu_o - \hat{\mu})^2 + \sigma_p^2}{N+\alpha-1} \tag{3.69}$$

donde el parámetro α representa otro factor de pesado equivalente a τ.

Cuando la cantidad de datos con la que cuenta es grande, la adaptación MAP converge a un modelo dependiente del entorno, siendo un método de adaptación que da mejores resultados que MLLR. Las ventajas de MLLR son producir adaptaciones más rápidas, con menor número de datos y sin que sea condición necesaria que la adaptación sea supervisada.

Combinación de modelos paralelos

Parallel Model Combination, PMC, [39]. El método utilizado en esta técnica de combinación de modelos paralelos es crear dos modelos ocultos de Markov independientes: uno para la voz limpia y otro para el ruido, y combinarlos para obtener un modelo adaptado del entorno ruidoso con el que se vaya a evaluar.

En primer lugar los parámetros son movidos del dominio cepstral al dominio de las potencias espectrales logarítmicas mediante la inversa de la

transformada discreta del coseno (matriz C^{-1} en las expresiones):

$$\begin{aligned} \mu^l &= C^{-1} \cdot \mu^c \\ \Sigma^l &= C^{-1} \cdot \Sigma^c \cdot (C^{-1})^T \end{aligned} \tag{3.70}$$

Para una Gausiana limpia X, el elemento k-ésimo de sus vectores de media y varianza en el dominio espectral lineal, tendrá la siguiente relación con las ecuaciones 3.70 del dominio espectral logarítmico:

$$\begin{aligned} \mu_{X_k} &= e^{\mu^l_{X_k} + \frac{1}{2} \cdot \Sigma^l_{X_k}} \\ \Sigma_{X_k} &= \mu^2_{X_k} \cdot (e^{\Sigma^l_{X_k}} - 1) \end{aligned} \tag{3.71}$$

Las mismas expresiones se pueden aplican para obtener la media y varianza de un modelo del ruido N en el domino espectral lineal. Como en dicho dominio los efectos del espectro de voz y de ruido son aditivos, podemos sumar sus medias y varianzas:

$$\begin{aligned} \hat{\mu}_Y &= \mu_X + \tilde{\mu}_N \\ \hat{\Sigma}_Y &= \Sigma_X + \tilde{\Sigma}_N \end{aligned} \tag{3.72}$$

Una vez combinados los modelos en 3.72, la vuelta al dominio espectral logarítmico será la dada por las ecuaciones 3.73 supuesta una distribu-

ción log-normal de las densidades de probabilidad:

$$\hat{\mu}_Y^l = log(\hat{\mu}_Y) - \frac{1}{2} \cdot log(\frac{\hat{\Sigma}_Y}{\hat{\mu}_Y^2} + 1)$$
$$\hat{\Sigma}_Y^l = log(\frac{\hat{\Sigma}_Y}{\hat{\mu}_Y^2} + 1) \tag{3.73}$$

Finalmente, para trasladar de nuevo los parámetros al dominio cepstral se les hará la transformada del coseno discreta (matriz C en la siguiente expresión):

$$\hat{\mu}_Y^c = C \cdot \hat{\mu}_Y^l$$
$$\hat{\Sigma}_Y^c = C \cdot \hat{\Sigma}_Y^l \cdot C^T \tag{3.74}$$

Transformaciones no lineales

Las redes neuronales tienen la capacidad de aprender comportamientos no lineales de un conjunto de datos, y por lo tanto han sido utilizadas también para adaptar modelos limpios a entornos con unos determinados datos de adaptación [137],[138].

3.2. El ruido no estacionario

Los métodos de robustecimiento descritos hasta ahora, trabajan sobre la hipótesis de que el ruido aditivo es estacionario: su densidad espectral de potencia no cambia con el tiempo y son ruidos de banda estrecha. Para los ruidos transitorios cuyas propiedades estadísticas cambian con el tiempo, existen dos técnicas que se han desarrollado con la filosofía de emular el mecanismo de percepción de los humanos: se procesan las componentes

de señal cuya SNR es alta, y se suprimen o ignoran las componentes de SNR baja. Estas dos técnicas que serán analizadas a continuación son el reconocimiento multi-banda y la técnica de características ausentes (*Missing Features Approach*)

3.2.1. Missing features

La aproximación *Missing Features* (que se puede traducir como *características ausentes*) [107], [108] se basa en detectar los componentes espectrales ausentes en el espectrograma de la señal, o presentes pero no fiables debido a su baja SNR. Si el reconocimiento de una transcripción W de un sonido Y se basa en la maximización de la expresión:

$$P(W|Y) = \frac{P(Y|W) \cdot P(W)}{P(Y)} \tag{3.75}$$

y en concreto en la maximización de la probabilidad auditiva $P(Y|W)$, la aproximación que hace Missing Features es dividir los datos acústicos Y en particiones fiables Y_F y particiones no fiables Y_{NF}, para usar la probabilidad marginal $P(Y_F|W)$ en el reconocimiento.

Esta identificación de la fiabilidad de los datos acústicos, se denomina definición de la máscara de características ausentes y es la parte más delicada del proceso. Hay diferentes técnicas de identificación de la máscara: estimación basada en criterios de SNR, estimación bayesiana, o estimación mediante criterios perceptuales.

Una vez definida la máscara, existen dos estrategias de reconocimiento: reconocer directamente utilizando el espectrograma incompleto, o completar el espectrograma usando los datos fiables e información sobre los datos limpios.

i) **Métodos de compensación del clasificador de características** [20]. Una vez detectados y eliminados los datos no fiables, se modifica el modo de calcular las probabilidades de las clases o estados. Hay dos formas de hacerlo:

- Modificación mediante imputación condicional de clase ([64]): los componentes no fiables del vector de características son reemplazados por sus estimaciones MAP dada la distribución a priori de la clase o estado, que son las que se usan para calcular la probabilidad de la clase o estado.

- Marginalización([19]): Los datos no fiables del vector en el dominio del logaritmo espectral de potencias son sacados de la distribución de la clase y tabulados entre unos límites superior e inferior establecidos con datos fiables. Las distribuciones así resultantes que tienen un menor número de componentes son las que se usan para calcular la probabilidad del vector.

Estos métodos operan típicamente en el dominio de los coeficientes espectrales en escala Mel, y en algunos casos esto hace que los resultados sean peores que los obtenidos con los coeficientes MFCC sin la aproximación de características ausentes.

ii) **Métodos de compensación de características** ([105]). También llamados *de reconstrucción de espectrograma*. Estos métodos modifican las características de entrada en vez de los clasificadores del reconocedor. Los componentes del dominio espectral logarítmico que no son fiables se borran y se reconstruyen usando información estadística derivada de las características de voz fiables. Esto da lugar a un conjunto completo de características de las que se derivan los coeficientes cepstrales que son los óptimos para el reconocimiento. Hay varias alternativas para la reconstrucción de las características:

- Reconstrucción geométrica Las características no fiables se calculan por interpolación lineal o no lineal (polinómica, funciones racionales, *splines*), usando las características fiables adyacentes en el eje de tiempos o de frecuencias.

- Reconstrucción basada en *clusters*. Las características se distribuyen en clusters en el dominio espectral. Los componentes no fiables de un vector espectral no son tenidos en cuenta para determinar

el *cluster* al que pertenece el vector. La distribución de probabilidades del *cluster* es usada después para obtener las estimaciones MAP de los componentes no fiables.

- Reconstrucción basada en covarianzas [106]. Esta estrategia asume que las distribuciones de probabilidad de los vectores de características son estacionarias. Las correlaciones entre cuales quiera dos vectores se obtienen de los datos limpios, y usando estas correlaciones de los datos no fiables con fiables y suponiendo que las distribuciones de probabilidad son Gausianas, se calculan las estimaciones MAP de las características no fiables.

En todos los casos descritos, la identificación de cuáles son los elementos no fiables del vector de características es el factor clave de la eficacia del método. Para ello es necesario el conocimiento *a priori* del ruido real de los elementos del espectro. Esto es difícil, especialmente en el caso de los ruidos no estacionarios. El avance de esta técnica está por tanto condicionado al avance de los clasificadores para la estimación de la máscara siendo la marginalización el método más robusto a errores de máscara para algunos autores.

3.2.2. Reconocimiento multibanda

El origen de esta técnica [127],[89], está en el estudio de la percepción humana en la que el mensaje de voz es decodificado de manera independiente en las distintas sub-bandas de frecuencia y la decodificación final es el resultado de mezclar las distintas decisiones de las distintas sub-bandas. Si alguna de las sub-bandas tiene información no fiable el cerebro humano la *des-enfatiza* al tomar la decisión global. En la implementación tradicional del reconocimiento automático del habla, la parametrización se hace extrayendo el conjunto de características acústicas de toda la banda de frecuencias de la señal acústica. De este modo, incluso si sólo una parte de la banda de frecuencias está contaminada de ruido, todas las características acústicas se ven afectadas.

El reconocimiento multibanda divide el espectro de frecuencias en sub-bandas y las modela de manera independiente con lo que se consigue aislar las distorsiones de frecuencias específicas (emulando al proceso de percepción en humanos). Hay dos cuestiones de fondo a la hora de implementar el reconocimiento multibanda:

a) Cómo se definen las sub-bandas en el *front-end* del reconocedor: cuántas y de qué longitud. Estudios empíricos muestran un funcionamiento óptimo con cuatro sub-bandas. [16].

b) Cómo se recombinan las sub-bandas para que el reconocedor obtenga una decisión global. Para resolver esta pregunta hay varios enfoques con distintas estrategias de pesado de las sub-bandas. En general de manera intuitiva se acepta que no todas las sub-bandas contribuyen igual, aquellas que incluyan más formantes tendrán más información. Algunos autores recombinan los resultados del reconocimiento en cada banda, lo cual recibe el nombre de *recombinación de probabilidades*,[88]. Otros sin embargo yuxtaponen las bandas antes de estimar el modelo, recibiendo esta técnica el nombre de *recombinación de frecuencias*,[126].

3.3. *Arrays* de micrófonos

Para reconocimiento en entornos con SNRs bajas, se consiguen importantes mejoras en la tasa de reconocimiento usando *arrays* de micrófonos [62], [91]. Los *arrays* permiten tener ganancias direccionales que aumentan la sensibilidad respecto a la señal del hablante y la disminuyen respecto al entorno que lo rodea. Esto presenta ventajas en el caso de ruidos difusos especialmente provenientes de una posición espacial diferente a la del locutor. Los *arrays* de micrófonos permiten además focalizar la atención en el campo directo en un entorno con reverberación.

Hay tres enfoques clásicos para el diseño de *arrays* de micrófonos en reconocimiento robusto.

- Formación de haces combinados *de retardo y suma*. (*Delay and Sum Beamforming*). Las señales de los diferentes micrófono se retrasan y suman del tal modo que se realzan las que provienen de una determinada posición. A continuación, se aplica un algoritmo de postprocesado que compense la coloración espectral que introduce el *array*. Las técnicas más usadas en este grupo son las de *formación de haces superdirectivos* ([21], [22]).

- *Arrays* de micrófonos con filtrado clásico adaptativo basado en el error cuadrático medio, que dan buenos rendimientos para ruidos aditivos independientes. No son un buen método sin embargo para compensar la reverberación ([50]).

- Algoritmos basados en correlaciones cruzadas, que tienen la capacidad de reforzar las componentes de señal que vienen de un ángulo azimut particular. Estos algoritmos tienen el atractivo de que hacen un procesado similar al del sistema binaural humano, muy robusto ante el ruido aditivo y la reverberación ([66]).

Capítulo 4

Objetivos de la tesis

4.1. Objetivos de la tesis

Del encuadre científico-tecnológico de las técnicas de robustecimiento hecho en los dos capítulos anteriores, se pueden extraer como conclusiones los siguientes requisitos para los algoritmos que tengan dicho propósito:

R1 Cuanto menos específico del tipo de ruido sea el algoritmo, más capacidad para combatir ruidos no modelados o mezclas de ellos tendrá.

R2 Como argumentaron los creadores de los algoritmos de parametrización basados en modelos auditivos de la percepción, es deseable conservar la información temporal del sonido, y añadirla a la de la frecuencia en el proceso de parametrización.

R3 Son deseables algoritmos de robustecimiento con la menor carga computacional posible.

Dentro del conjunto de técnicas de compensación de características mediante encuadre estadístico, las llamadas transformaciones no lineales de características son bastante atractivas para cierto tipo de aplicaciones de reconocimiento que necesiten robustez con un tiempo de respuesta corto y carga computacional no muy elevada ya que:

i) Sus requerimientos de información sobre el ruido son mínimos, al no trabajar con ningún modelo ni hipótesis sobre el mismo, y sin embargo, son capaces de combatir la distorsión no lineal que éste produce.

ii) No usan modelos complejos del entorno, como puede ser el caso de VTS, con lo que su coste computacional no es excesivamente elevado.

iii) No necesitan datos estéreo ni de adaptación. Usan una referencia que de modo óptimo se calcula con datos limpios.

Por esta razón las transformaciones no lineales se han perfilado como una técnica potente de transformación de características que produce resultados satisfactorios con poca complejidad [27], [32], [58] [136] y van a ser objeto de estudio exhaustivo en el presente trabajo.

Basado en lo expuesto hasta ahora, **el objetivo general de esta tesis es crear una base sólida de conocimiento sobre las técnicas de compensación de características mediante encuadre estadístico, y en particular sobre la Ecualización de Histogramas, con el fin de permitir la construcción fundamentada de mecanismos que la mejoren atacando sus debilidades y potenciando sus fortalezas**. Para alcanzar este objetivo general se plantean los siguientes objetivos específicos:

O1 Hacer un **estudio escrupuloso de la Ecualización de Histogramas y de sus variantes**. Analizar en profundidad su comportamiento enfrentándolo con los requisitos de los algoritmos de robustecimiento expuestos anteriormente, para definir las posibles carencias y evoluciones de HEQ. Este objetivo se ha materializado en las siguientes aportaciones:

- **O1.1** Estudio de la **distribución de densidad acumulada óptima** que se elige como referencia en la ecualización.
- **O1.2** Propuesta de una **versión on-line de HEQ** que mejora el tiempo de respuesta del sistema de reconocimiento.
- **O1.3** Propuesta de la **ecualización parcial** de cierto número de coeficientes cepstrales.

El estudio exhaustivo de HEQ se ha materializado también en la detección de las siguientes carencias:

C1.1 La longitud de la frase que se ecualiza influye en la fiabilidad de su función de densidad acumulada, y con ello en la fiabilidad de la ecualización. Es deseable una ecualización robusta, independientemente de la longitud de la frase.

C1.2 El porcentaje de voz y ruido presente en la frase influye en la transformación de ecualización. Esto introduce distorsión en los parámetros ecualizados.

C1.3 La información temporal que se utiliza se limita al uso de los coeficientes $\triangle$ y $\triangle\triangle$ que son ecualizados como el resto de componentes del vector de características.

Los resultado del objetivo **O1** han tenido como consecuencia en el planteamiento de los siguientes dos objetivos:

O2 Propuesta de una **versión paramétrica del algoritmo de Ecualización** de Histogramas (*PEQ-Parametric Equalization*) que lo haga menos dependiente del número de datos que se ecualizan, y que ecualice por separado una clase de voz y otra de silencio.

O3 Propuesta de un **algoritmo** (*TES-Temporal Smothing Filter*) **que añada información temporal** al proceso de ecualización.

Parte III

Propuesta

Capítulo 5

Descripción del entorno de trabajo

Este capítulo describe el sistema de reconocimiento automático del habla empleado para evaluar las técnicas propuestas en el trabajo de tesis. Se describen las técnica de parametrización utilizadas, los modelos acústicos y de lenguaje empleados, y las tres bases de datos usadas para llevar a cabo la experimentación: AURORA2, AURORA4 y HIWIRE. Por último se describen los criterios de evaluación seguidos para analizar los resultados.

5.1. Extracción de características

Se han utilizado dos parametrizaciones en este trabajo. La primera de ellas es el sistema de parametrización básico de referencia de la herramienta HTK [36] para el *front-end*, que será denominado *Baseline* y que es muy similar al estandarizado y descrito por la ETSI en [4]. Como umbral de reconocimiento robusto estandarizado que es deseable alcanzar y superar, se ha implementado el estándar de parametrización de la ETSI *Advanced Front-End* [5]. A continuación se describen ambas parametrizaciones.

5.1.1. Sistema base de referencia: *Baseline*

La parametrización básica utilizada en este trabajo se crea utilizando el software de reconocimiento automático del habla HTK [36] para crear

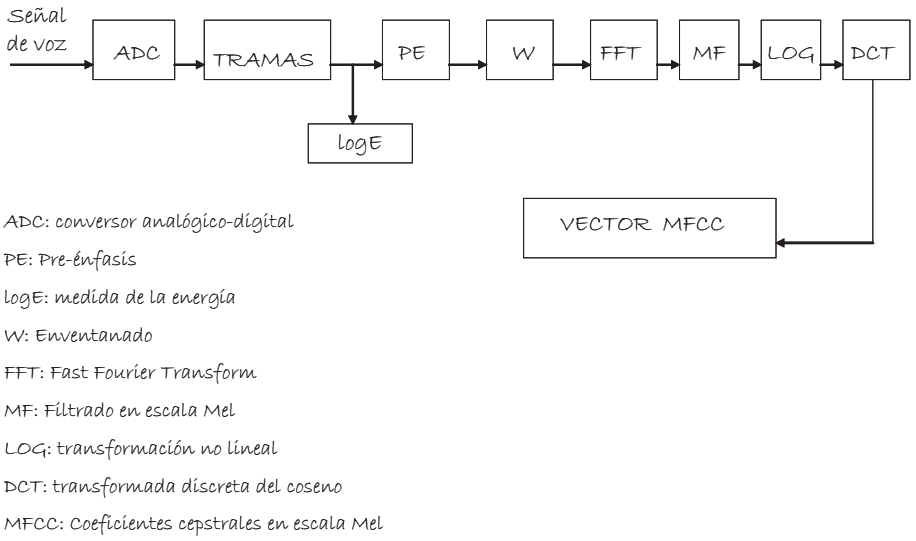


Figura 5.1. Estándar de parametrización

coeficientes cepstrales en escala Mel con sustracción cepstral de la media. Los bloques de los que consta el sistema de parametrización usado son los que se ven en la figura 5.1.

i) En primer lugar se hace la **conversión analógica-digital**, con la posibilidad de muestrear las señales a 8Khz, 11Khz y 16Khz. En el caso de este trabajo, las bases de datos usadas ya contienen la señal digital muestreada a 8 Khz.

ii) El siguiente paso es **dividir la señal en tramas** de longitud $25ms$ que contienen 200 muestras, con un intervalo de desplazamiento entre tramas consecutivas de 80 muestras.

iii) La **energía logarítmica** de la trama se mide después de haber dividido la señal en tramas de 200 muestras. La energía logarítmica de la

trama será:

$$logE = ln(\sum_{i=1}^{i=200} s(i)^2) \tag{5.1}$$

iv) El siguiente paso es filtrar la señal con un **filtro de pre-énfasis** de expresión:

$$s_{pe}(n) = s(n) - 0,97 \cdot s(n-1) \tag{5.2}$$

v) La salida del filtro de pre-énfasis se pasa a través de una **ventana de Hamming** con expresión:

$$s_h(n) = \{0,54 - 0,46 \cdot cos(\frac{2 \cdot \pi(n-1)}{200-1})\} \cdot s_{pe}, \quad 1 \leq \text{n} \leq 200 \tag{5.3}$$

vi) Las 200 muestras de la trama, se completan con ceros para formar una trama de 256 componentes a la que se le hace **FFT**. Los puntos de la FFT tendrán la expresión:

$$bin_k = |\sum_{n=0}^{255} s_h(n) \cdot e^{-j \cdot n \cdot k \cdot \frac{2 \cdot \pi}{256}}|, \quad \text{k=0,...,255} \tag{5.4}$$

vii) Habiendo pasado ya al dominio espectral, se ignoran los componentes por debajo de los 64 Hz, que se define como frecuencia inicial ($f_{inic} = 64Hz$). La banda de frecuencias considerada es por tanto 64Hz-4 Khz (mitad de la frecuencia de muestreo). Esta banda se divide en **23 canales equidistantes en escala Mel** superpuestos entre sí. Las frecuencias centrales de dichos canales en función de los puntos de la transformada FFT tienen la siguiente expresión:

$$\begin{aligned} Mel\{x\} &= 2595 \cdot log_{10}(1 + \frac{x}{700}) \\ f_{c_i} = Mel^{-1}\{Mel\{f_{inic}\} + \frac{Mel\{4000\} - \{Mel\{f_{inic}\}}{23+1} \cdot i\}, &\quad \text{i=1,...,23} \\ cbin_i &= entero\_sup\{\frac{f_{c_i}}{8000} \cdot 256\} \end{aligned} \tag{5.5}$$

La salida del filtro Mel será la suma ponderada de los valores espectrales de la FFT (bin_i) en cada banda. Para definir los filtros se usan ventanas triangulares solapadas del modo siguiente:

$$fbank_k = \sum_{i=cbin_{k-1}}^{cbin_k} \frac{i - cbin_{k-1} + 1}{cbin_k - cbin_{k-1} + 1} \cdot bin_i + \\ + \sum_{i=cbin_k}^{cbin_{k+1}} (1 - \frac{i - cbin_k}{cbin_{k+1} - cbin_k + 1}) \cdot bin_i, \quad \text{siendo k=1,...,23} \tag{5.6}$$

Los términos $cbin_0$ y $cbin_{24}$ de la expresión 5.6 se definirán como los puntos de la transformada FFT correspondientes a la frecuencia inicial y la mitad de la frecuencia de muestreo respectivamente:

$$cbin_0 = entero\_sup\{\frac{f_{inic}}{8000} \cdot 256\} \\ cbin_{24} = entero\_sup\{\frac{4000}{8000} \cdot 256\} \tag{5.7}$$

viii) A la salida de los filtros Mel se le aplica el **logaritmo natural**:

$$f_i = ln(fbank_i), \quad \text{i=1,...,23} \tag{5.8}$$

ix) Los coeficientes cepstrales se obtienen aplicando la **transformada discreta del coseno** a los f_i. Se utilizan solamente los trece primeros:

$$c_i = \sum_{j=1}^{23} f_j \cdot cos(\frac{\pi \cdot i}{23} \cdot (j - 0,5)), \quad 0 \leq i \leq 12 \tag{5.9}$$

De esta manera, la parametrización *Baseline* del *front-end* del reconocedor tiene 13 coeficientes cepstrales en escala Mel, más un término con la energía logarítmica de la trama. La información de la energía logarítmica de la trama y la del coeficiente *C0* de la misma son redundantes. En nuestro sistema de reconocimiento hemos utilizado *C0* en vez de la energía logarít-

mica por razones prácticas de compatibilidad con el software de reconocimiento. De este modo trabajamos con un vector inicial de 13 coeficientes cepstrales:

$$C = c_0, c_1, ..., c_{12} \tag{5.10}$$

A dichos 13 coeficientes se les calculan las derivadas primera y segunda como vemos en la expresión 5.11 en la que las longitudes de regresión tienen los valores $\Delta L = 3$ y $\Delta\Delta L = 2$:

$$\triangle c(t) = \frac{\sum_{\tau=-\Delta L}^{\Delta L} \tau \cdot c(t+\tau)}{\sum_{\tau=-\Delta L}^{\Delta L} \tau^2} \qquad \triangle\triangle c(t) = \frac{\sum_{\tau=-\Delta\Delta L}^{\Delta\Delta L} \tau \cdot \Delta c(t+\tau)}{\sum_{\tau=-\Delta\Delta L}^{\Delta\Delta L} \tau^2} \tag{5.11}$$

El vector de características usado tiene por tanto 39 componentes: los MFCCs *C0-C12*, sus 13 derivadas primeras, y sus 13 derivadas segundas:

$$X = \{C, \triangle C, \triangle\triangle C\} \tag{5.12}$$

La parametrización *Baseline* utilizada, lleva i**mplícita la sustracción cepstral de la media (CMN)**. Ésta se hace de manera automática durante el proceso de entrenamiento al crear los modelos HMMs con HTK.

5.1.2. Advanced Front-End

La parametrización del estándar *Advanced front-end* está descrita en las especificaciones [5] de la ETSI. La figura 5.2 muestra un diagrama de los bloques que atraviesa la señal de voz para obtener el vector de parámetros. En primer lugar la señal pasa por un bloque de reducción del ruido en el dominio temporal, al que le sigue otro bloque de reducción dinámica del ruido en función de la SNR de la señal, llamado bloque de procesado de la forma de onda. A continuación la señal se pasa al dominio cepstral, y los coeficientes cepstrales son sometidos a una ecualización ciega.

Bloque de reducción del ruido

El bloque de reducción del ruido es el que aparece en la figura 5.3, y está formado por dos etapas diferentes que comparten los elementos que a

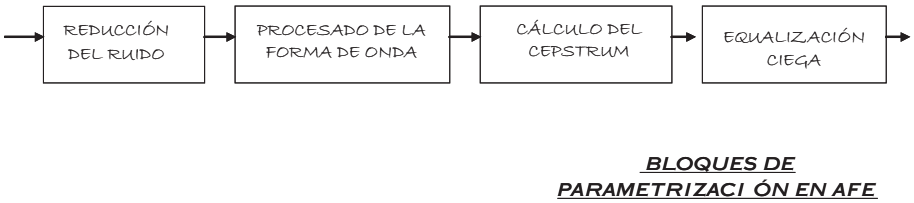


Figura 5.2. Estándar de parametrización AFE

continuación se describen, exceptuando el caso del bloque de factorización de la ganancia, que aparece sólo en la segunda etapa:

i) Estimación del espectro de la señal. La señal se divide en tramas que son pasadas por una ventana de Hamming. Se les aplica la transformada de Fourier y se obtiene la representación en frecuencias de cada trama:

$$X(bin) = FFT\{s(n)\}, \quad 0 \leq \text{bin} \leq \frac{256}{2} \tag{5.13}$$

ii) Cálculo de la **densidad de potencia espectral media**. Con los puntos de la transformada FFT de la señal de la ecuación 5.13, se calcula la potencia espectral de cada trama:

$$P(bin) = |X(bin)|^2, \quad 0 \leq \text{bin} \leq \frac{256}{2} \tag{5.14}$$

Y esta potencia espectral de la trama se suaviza con la expresión:

$$P(bin) = \frac{P(2 \cdot bin) + P(2 \cdot bin + 1)}{2}, \quad 0 \leq \text{bin} < \frac{256}{4}$$
$$P(\frac{256}{4}) = P(\frac{256}{2}) \tag{5.15}$$

Para calcular la potencia espectral media en las T_{PSD} tramas anteriores, estando en la trama t, se utiliza la siguiente expresión en la que

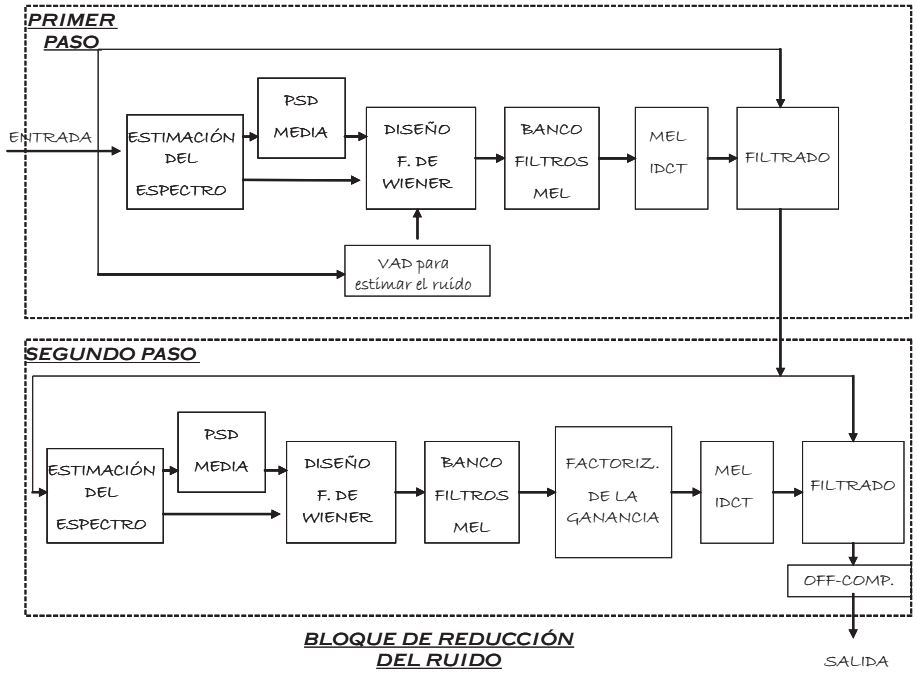


Figura 5.3. Bloque de reducción del ruido

$N_{SPEC} = \frac{256}{4} + 1$ es la longitud del espectro de potencias:

$$P_{PSD}(bin, t) = \frac{1}{T_{PSD}} \sum_{i=0}^{T_{PSD}-1} P(bin, t-i), \quad 0 \leq \text{bin} \leq N_{SPEC} - 1 \tag{5.16}$$

iii) Diseño del filtro de Wiener. El filtro de Wiener es la principal técnica de reducción de ruido usada en los estándares de AURORA. Se encarga de hacer una estimación de la densidad de la potencia espectral limpia a partir de la potencia espectral de la señal ruidosa con la ayuda de un detector de actividad de voz. El filtro aparece en las dos etapas del bloque de reducción de ruido, siendo la salida del filtro de Wiener de la primera etapa la que se usa como entrada del filtro de

Wiener de la segunda etapa. Su funcionamiento es idéntico en ambas etapas, exceptuando el modo en el que se calcula la potencia de ruido ($P_{ruido}(bin, t)$) que es utilizada para calcular las SNRs *a priori* $\eta(bin, t)$ y $\eta_2(bin, t)$.

iii.a) En el filtro de Wiener de la **primera etapa** del bloque de reducción de ruido, la **potencia espectral del ruido** se calcula del modo siguiente:

$$\begin{aligned} &P_{noise}^{1/2}(bin, t_n) = \\ &\quad = max(\lambda_{NSE} \cdot P_{noise}^{1/2}(bin, t_n - 1) + (1 - \lambda_{NSE}) \cdot P_{PSD}^{1/2}(bin, t_n), e^{-10,0}) \\ &P_{noise}^{1/2}(bin, t) = P_{noise}^{1/2}(bin, t_n) \end{aligned} \tag{5.17}$$

La trama t_n de la expresión 5.17 representa la última trama de silencio detectada por el VAD, y P_{PSD} representa la potencia espectral media de la trama, calculada en el paso $ii)$ de esta sección. El término λ_{NSE} es el factor de olvido para la estimación del espectro de ruido.

iii.b) En el caso del filtro de Wiener de la **segunda etapa**, $P_{ruido}(bin, t)$ se estima con la siguiente función:

iii.b.1 Para las tramas del principio de la frase ($t < 11$), la potencia de ruido será:

$$P_{ruido}(bin, t) = \lambda_{NSE} \cdot P_{ruido}(bin, t-1) + (1-\lambda_{NSE}) \cdot P_{PSD}(bin, t) \tag{5.18}$$

iii.b.2 Para las tramas $t \geq 11$, se define la variable $update$ como:

$$update = 0,9 + 0,1 \cdot \frac{P_{PSD}(bin,t)}{P_{PSD}(bin,t) + P_{ruido}(bin,t-1)} \cdot (...) \\ (...) \cdot (1 + \frac{1}{1 + 0,1 \cdot \frac{P_{PSD}(bin,t)}{P_{ruido}(bin,t-1)}}) \tag{5.19}$$

y la **potencia espectral de ruido** se calcula usando la variable $update$:

$$P_{noise}(bin,t) = P_{noise}(bin,t-1) \cdot update \tag{5.20}$$

Una vez conseguida la expresión de la potencia de ruido $P_{ruido}(bin,t)$ el resto del proceso de filtrado de Wiener es igual para las dos etapas del bloque de reducción del ruido.

iii.c) Se calcula la potencia de la señal limpia una vez eliminada la potencia de ruido:

$$P^{1/2}_{limpia}(bin,t) = 0,98 \cdot P^{1/2}_{limpia3}(bin,t-1) + ... \\ (1 - 0,98) \cdot T(P^{1/2}_{PSD}(bin,t) - P^{1/2}_{ruido}(bin,t)) \tag{5.21}$$

La función $T(x)$ usada en la ecuación 5.21 es la función umbral definida como:

$$T(x) = \begin{cases} x, & \text{si x>0;} \\ 0, & \text{en otro caso.} \end{cases} \tag{5.22}$$

iii.d) Con estos datos la **SNR *a priori*** se define como:

$$\eta(bin,t) = \frac{P_{limpia}(bin,t)}{P_{ruido}(bin,t)} \tag{5.23}$$

y la función de transferencia del filtro tiene entonces la expre-

sión:

$$H(bin,t) = \frac{\sqrt{\eta(bin,t)}}{1+\sqrt{\eta(bin,t)}} \tag{5.24}$$

que será utilizada para mejorar la estimación del espectro de la señal limpia:

$$P^{1/2}_{limpia2}(bin,t) = H(bin,t) \cdot P^{1/2}_{PSD}(bin,t) \tag{5.25}$$

y con ello obtener una versión mejorada de la SNR *a priori* $\eta_2(bin,t)$:

$$\eta_2(bin,t) = max(\frac{P_{limpia2}(bin,t)}{P_{ruido}(bin,t)}, \eta^2_{TH}) \tag{5.26}$$

siendo η_{TH} un umbral mínimo correspondiente a una SNR de -22dB.

iii.e) Con el nuevo valor de SNR a priori η_2, **se recalcula la función de transferencia del filtro de Wiener**:

$$H_2(bin,t) = \frac{\sqrt{\eta_2(bin,t)}}{1+\sqrt{\eta_2(bin,t)}}, \tag{5.27}$$

iii.f) Por último la versión re-calculada de la **densidad espectral de potencia de la señal limpia** a la salida del filtro de Wiener será:

$$P^{1/2}_{limpia3}(bin,t) = H_2(bin,t) \cdot P^{1/2}_{PSD}(bin,t) \tag{5.28}$$

iv) El **detector de actividad de voz** (VAD). Para la estimación de las tramas de voz y de silencio usadas en el filtro de Wiener definido en el punto anterior de esta sección, se usa un detector de actividad de voz que calcula dos parámetros: la energía logarítmica de las 80 últimas muestras de la trama que se analiza, y el valor medio de la energía de trama. Con estos dos parámetros clasifica las tramas como tramas de voz o de silencio.

v) Los **coeficientes lineales de frecuencia del filtro de Wiener** $H_2(bin)$ calculados en el punto iii) de esta sección, son suavizados y transfor-

mados a una **escala Mel**, originando así los coeficientes del filtro de Wiener en escala Mel $H_{2\_Mel}(k)$. Estos coeficientes se calculan igual que los coeficientes del banco de filtros en escala Mel referidos en las especificaciones del estándar ETSI para el *front-end* que se ha descrito en la sección 5.1.1 de este capítulo, siguiendo las ecuaciones 5.5, 5.6 y 5.7.

vi) Factorización de la ganancia. El objetivo de este proceso, que se da solamente en la segunda etapa del bloque de reducción del ruido, es suavizar el efecto de la reducción de ruido aplicada a los coeficientes del filtro de Wiener en escala Mel mediante una ganancia α_{GF}. La intención es aplicar una reducción de ruido más agresiva a las tramas que tengan exclusivamente ruido, y aplicar una reducción de ruido un poco menos agresiva a las tramas que tengan voz y ruido. Para ello se siguen los siguientes pasos:

vi.a) Se calcula la **energía de la señal sin ruido** para una trama utilizando el espectro de potencia de señal limpia $P_{limpia3}$ calculado para definir el filtro de Wiener:

$$E_{limpia}(t) = \sum_{bin=0}^{N_{SPEC}-1} P_{limpia3}(bin, t) \tag{5.29}$$

vi.b) Se calcula **la energía del ruido de la trama** t utilizando la densidad espectral de potencia de ruido $P_{ruido}(bin, t)$:

$$E_{ruido}(t) = \sum_{bin=0}^{N_{SPEC}-1} P_{ruido}(bin, t) \tag{5.30}$$

vi.c) Se define una **SNR suavizada media** utilizando las energías medias de tres tramas limpias:

$$SNR_{media}(t) = (20 \cdot log_{10}(\frac{E_{limpia}(t-2) \cdot E_{limpia}(t-1) \cdot E_{limpia}(t)}{E^3_{ruido}(t)}))/3 \tag{5.31}$$

vi.d) Se define una variable SNR_{baja} que hace un seguimiento del

nivel de SNR que se define como umbral para determinar que una trama es de ruido con el siguiente criterio:

if $\{SNR_{media}(t) - SNR_{baja}(t-1) < 0 \quad o \quad t < 10\}$

$SNR_{baja}(t) = \lambda_{SNR}(t) \cdot SNR_{baja}(t-1) + (1 - \lambda_{SNR}(t)) \cdot SNR_{media}(t)$

else

$SNR_{baja}(t) = SNR_{baja}(t-1)$

vi.e) La factorización de la ganancia del filtro de Wiener se hace usando **un factor de ganancia** α_{gf} definido con la siguiente lógica y haciendo uso de la variable SNR_{baja}:

if $\{SNR_{media}(t) < SNR_{baja}(t) + 3,5\}$

$\alpha_{gf}(t) = \alpha_{gf}(t-1) + 0,15$

if $\{\alpha_{gf}(t) > 0,8\}$

$\alpha_{gf}(t) = 0,8$

else

$\alpha_{gf}(t) = \alpha_{GF}(t-1) - 0,3$

if $\{\alpha_{gf}(t) < 0,1\}$

$\alpha_{gf}(t) = 0,1$

Como se ve en este pseudo-código, el factor α_{gf} toma valores entre $0,1$ y $0,8$ lo que conlleva un ajuste fino de la *erosión* que hace el filtro de Wiener desde un 10 % para tramas de voz, hasta un 80 % para tramas de ruido.

vi.f) Por último, el filtro de Wiener de la segunda etapa se modifica con la factorización de la ganancia α_{GF} del siguiente modo:

$$H_{2\_Mel\_gf}(k,t) = 1 + \alpha_{gf}(t) \cdot H_{2\_Mel}(k,t), \quad 0 \leq k \leq K_{FB} + 1 \tag{5.32}$$

vii) El paso $vi)$ ha originado unos coeficientes del Filtro de Wiener en escala Mel y con una ganancia modificada. Para pasarlos al dominio temporal, se utiliza la **transformada del coseno inversa en escala**

Mel ($IDCT_{Mel}$) . La respuesta al impulso del filtro de Wiener en el dominio temporal será:

$$h_{WF}(n) = \sum_{k=0}^{K_{FB}+1} H_{2\_Mel}(k) \cdot IDCT_{Mel}(k,n), \quad 0 \leq n \leq K_{FB}+1 \tag{5.33}$$

y su extensión responderá a la expresión:

$$h_{WF\_mirr}(n) = \begin{cases} h_{WF}(n), & 0 \leq n \leq K_{FB}+1; \\ h_{WF}(2 \cdot (K_{FB}+1)+1-n), & . \end{cases} \tag{5.34}$$

viii) Una vez calculada la expresión en el tiempo del filtro de Wiener, se aplica a la señal de voz sucia original en el caso de la primera etapa, o a la señal ya filtrada una vez en el caso de la segunda etapa. Los pasos que se siguen son:

viii.a) La respuesta al impulso **causal** $h_{WF\_causal}(n,t)$ se obtiene de la expresión 5.34 con la siguiente relación:

$$\begin{aligned} h_{WF\_causal}(n,t) &= h_{WF\_mirr}(n+K_{FB}+1,t), && n = 0,...,K_{FB} \\ h_{WF\_causal}(n,t) &= h_{WF\_mirr}(n-K_{FB}-1,t), && n = K_{FB}+1,...,2\cdot(K_{FB}+1) \end{aligned} \tag{5.35}$$

viii.b) Esta respuesta causal **se trunca** para una longitud del filtro $FL = 17$:

$$h_{WF\_trunc}(n,t) = h_{WF_caus}(n+K_{FB}+1-\frac{(FL-1)}{2},t), \quad n = 0,...,FL-1 \tag{5.36}$$

viii.c) La respuesta al impulso causal truncada, se pasa por una **ven-**

tana de Hamming:

$$h_{WF\_w}(n,t) = \{0,5 - 0,5 \cdot cos(\frac{2 \cdot \pi \cdot (n + 0,5)}{FL})\} \cdot h_{WF\_trunc}(n,t), \quad \text{para } 0 \leq n \leq FL - 1 \tag{5.37}$$

viii.d) Finalmente, la señal inicial con ruido s_{in}, se filtra con el filtro de Wiener construido para obtener **la señal con cancelación de ruido** s_{filt}:

$$s_{filt}(n) = \sum_{i=-\frac{(FL-1)}{2}}^{\frac{(FL-1)}{2}} h_{WF\_w}(i + \frac{FL-1}{2}) \cdot s_{in}(n-i), \quad 0 \leq n \leq 80 - 1 \tag{5.38}$$

ix) Compensación del *offset*. Como último paso del bloque de reducción del ruido, la señal filtrada es pasada a través de un filtro *ranura* con el que se elimina la componente DC de la señal:

$$s_{filt\_off}(n) = s_{filtrada}(n) - s_{filt}(n-1) + (1 - \frac{1}{1024} \cdot s_{filt\_off}(n-1)) \quad 0 \leq n \leq 80 - 1 \tag{5.39}$$

Bloque de procesado de la forma de onda

El bloque de procesado de forma de onda, es el que se representa en la figura 5.4 y recibe el nombre de SWP (*SNR-dependent Waveform Processing*). Recibe la señal filtrada desde el bloque de reducción del ruido, y tiene las siguientes etapas:

i) Estimación de la energía instantánea y suavizado. La energía instantánea de la señal para cada trama que entra al bloque se calcula usando el operador *Teager* como vemos en las expresiones 5.40 en las que N_{in} representan en número de tramas almacenadas en el *buffer* del

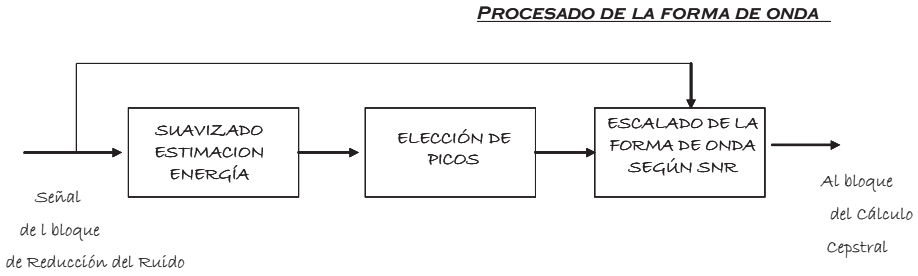


Figura 5.4. Procesado de la forma de onda

bloque, y que en el estándar es tomado como 200:

$$
\begin{aligned}
E_{Teag}(n) &= |s^2_{filt\_off}(n) - s_{filt\_off}(n-1) \cdot s_{filt\_off}(n+1)|, \\
&\qquad para \quad 1 \leq n < N_{in} \\
E_{Teag}(0) &= |s^2_{filt\_off}(0) - s^2_{filt\_off}(0) \cdot s_{filt\_off}(1)|, \\
&\qquad para \quad n = 0 \\
E_{Teag}(N_{in}-1) &= |s^2_{filt\_off}(N_{in}-1) - s_{filt\_off}(N_{in}-2) \cdot s_{filt\_off}(N_{in}-1)|, \\
&\qquad para \quad n = N_{in}-1
\end{aligned}
\tag{5.40}
$$

La energía instantánea calculada en 5.40 se suaviza con un filtro FIR de longitud 9:

$$E_{Teag\_suav}(n) = \frac{1}{9} \sum_{i=-4}^{4} E_{Teag}(n+i) \tag{5.41}$$

ii) En el **bloque de elección de los picos** se detectan los máximos de la estimación de la energía suavizada que están relacionados con la frecuencia fundamental:

- Se busca el máximo global en todas las tramas de $E_{Teag\_suav}(n), \quad 0 \leq n \leq N_{in} - 1$.

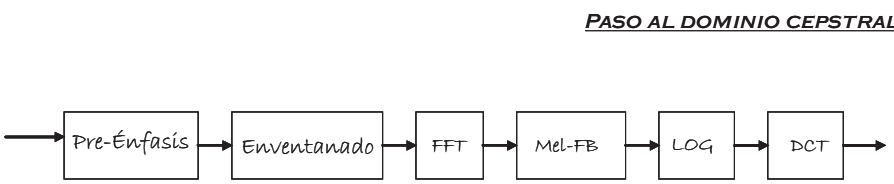


Figura 5.5. Transformación al dominio cepstral

- Se identifican los máximos a la derecha e izquierda del máximo global.

iii) Una vez obtenidos el número de máximos de energía suavizada N_{max} y sus posiciones $pos_{max}(n_{max})$, **el bloque de escalado de la forma de onda según la SNR**, aplica un factor de escala $w_{swp}(n)$ a las tramas con el siguiente criterio:

- En el intervalo $\langle [pos_{max}(n_{max}) - 4], [pos_{max}(n_{max}) - 4] + 0,8 \cdot [pos_{max}(n_{max} + 1) - pos_{max}(n_{max})] \rangle$, el factor de escalado valdrá $w_{swp}(n) = 1,0$.
- $w_{swp}(n) = 0,0$ para el resto de muestras.
- $w_{swp}(n) = 0,5$ para las tramas que son transición entre intervalos con factor de escala 0 y 1.

Finalmente, la señal filtrada y con compensación de *offset* del bloque de reducción de ruido es escalada con el factor dependiente de la forma de onda que se acaba de describir:

$$s_{swp}(n) = 1,2 \cdot w_{swp}(n) \cdot s_{filt\_off}(n) + 0,8 \cdot (1 - w_{swp}(n)) \cdot s_{filt\_off}(n), \quad 0 \leq n \leq N_{in} - 1 \tag{5.42}$$

Bloque de cálculo del cepstrum

Como muestra la figura 5.2, el bloque siguiente en el proceso de extracción de características es el bloque de cálculo de los parámetros cepstrales.

El cálculo de los parámetros cepstrales se divide en seis etapas que aparecen desglosadas en la figura 5.5: la señal pasa por un filtro de preénfasis, posteriormente se le aplica una ventana de Hamming y se le hace la transformada discreta de Fourier. De este modo, se pasa al dominio de la frecuencia en el que se aplica la escala Mel a la salida del banco de filtros. La escala logarítmica y la posterior transformada discreta del coseno, originan los doce coeficientes cepstrales, a los que se añade el logaritmo de la energía lnE (o en el caso de nuestro entorno de trabajo el coeficiente *C0*), dando un vector de características con trece componentes. Este proceso que se menciona ahora de manera resumida, es idéntico al que se hace en el estándar ETSI para el *front-end* básico (la descripción detallada se puede ver en la sección 5.1.1 de este capítulo).

Bloque de ecualización ciega

Los trece coeficientes cepstrales $c(0), c(1)..., c(12)$ se ecualizan siguiendo el algoritmo *LMS* (*Least Mean Square*) del modo que se describe a continuación:

i) En primer lugar se definen las variables $peso$ y $salto$ usando el logaritmo de la energía de la trama como:

$$\begin{aligned} peso &= Min(1, Max(0, lnE - \frac{211}{64})) \\ salto &= 0,0087890625 \cdot peso \end{aligned} \tag{5.43}$$

ii) Los coeficientes cepstrales ecualizados se calculan con las siguientes

ecuaciones:

$$\begin{aligned} c_{eq}(i) &= c(i) - bias(i), \quad 1 \leq i \leq 12 \\ bias(i) &+= salto \cdot (c_{eq}(i) - RefCep(i)), \quad 1 \leq i \leq 12 \end{aligned} \tag{5.44}$$

Los valores iniciales usados en 5.44 son los que corresponderían al Cepstrum de un espectro plano:

bias(i)=0,0 , $1 \leq i \leq 12$,

RefCep(1)=-6,618909,	RefCep(2)=0,198269
RefCep(3)=-0,740308,	RefCep(4)=0,055132
RefCep(5)=-0,227086,	RefCep(6)=0,144280
RefCep(7)=-0,112451,	RefCep(8)=-0,146940
RefCep(9)=-0,327466,	RefCep(10)=0,134571
RefCep(11)=0,027884,	RefCep(12)=-0,114905

5.2. El Reconocedor de habla

El trabajo propuesto se ha llevado a cabo usando un reconocedor de habla continua construido con la herramienta de software para reconocimiento automático de Voz HTK, en su versión 3.3 de Entropic [36] (*HTK-Hidden Markov Models ToolKit*). HTK permite crear y manipular modelos ocultos de Markov. Podemos dividir el trabajo del Reconocedor de Habla en tres etapas: parametrización, entrenamiento de los modelos y evaluación del reconocedor.

5.2.1. Parametrización

El software de HTK tiene como entrada ficheros de audio con frases muestreadas a frecuencia de 8 Khz.

- En el caso de la parametrización *Baseline*, HTK transforma directamente los ficheros audio en vectores de 39 coeficientes MFCC automatizando el proceso descrito en la sección 5.1.1 de este capítulo.

- Para las transformaciones no lineales que se estudian en este trabajo, HTK da la parametrización *Baseline* de los MFCCs, que posteriormente son transformados usando Matlab.

- En el caso del AFE, todo el proceso de parametrización necesario se realiza con la implementación de referencia en lenguaje C del estándar ETSI-AFE [5].

5.2.2. Entrenamiento

Cada palabra de las bases de datos se modela usando un HMM continuo con los siguientes parámetros:

- Cada palabra tiene 18 estados: un estado inicial y otro final que no consumen observaciones, más 16 estados intermedios.

- La topología definida para los HMMs es de izquierda a derecha con un salto máximo permitido de un estado.

- Cada estado se define como una mezcla de tres Gausianas multivaluadas con 39 componentes.

- Las Gausianas multivaluadas tienen las matrices de covarianzas diagonales.

- Se definen dos modelos adicionales: el modelo de pausa para el comienzo y fin de la frase, y el modelo de los silencios entre palabras.

El entrenamiento se hace aplicando el algoritmo de Baum-Welch en varias iteraciones, mediante los programas de entrenamiento *HRest* y *HERest* de la herramienta HTK. La decodificación se efectúa con el programa de decodificación *HVite*.

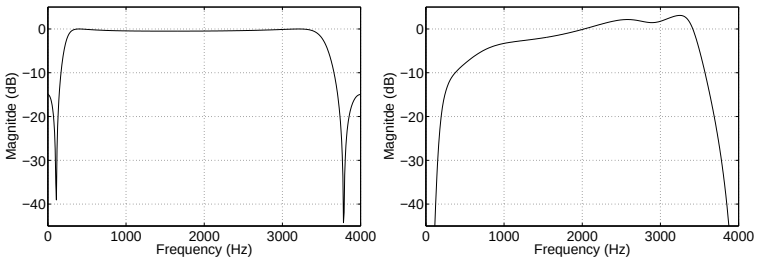


Figura 5.6. Respuesta en frecuencias de los filtros G.712 y MIRS

5.3. Bases de Datos utilizadas

El análisis de los beneficios de las transformaciones que se proponen en este trabajo se ha hecho usando tres bases de datos. Dos de ellas, AURORA2 y AURORA4, son subconjuntos de la base de datos del proyecto AURORA (ETSI STQ-AURORA Proyect Database 2.0, [10]). La tercera, denominada *HIWIRE Database* ha sido construida en el desarrollo del proyecto europeo HIWIRE [1]. Es un corpus específico para la comunicación entre los pilotos de aviones y las torres de control. Tiene además la peculiaridad de haber sido grabada por locutores no nativos.

5.3.1. AURORA2

La base de datos AURORA2 [96] se crea añadiendo artificialmente ruidos a la base de datos limpios TIDigits [69]. Contiene grabaciones de adultos de EE.UU. pronunciando dígitos aislados y secuencias de hasta 7 dígitos en inglés.

Tareas de Entrenamiento y Evaluación

La tarea de entrenamiento limpio se lleva a cabo con 8440 frases limpias grabadas en condiciones de alta SNR y pronunciadas por 55 locutores adultos masculinos y 55 locutoras adultas femeninas. Estas grabaciones son filtradas con un filtro G.712 o MIRS cuya respuesta en frecuencias se observa en la figura 5.6 (ver la recomendación G.712 de la ITU [3]

para más información). El objetivo de este filtrado es considerar las características realistas de frecuencias y terminales que son usadas en el área de las telecomunicaciones.

Para el entrenamiento multicondición, la partición de datos de entrenamiento se divide en 20 subconjuntos de 422 frases, en los que se añaden los ruidos pertinentes para representar los cuatro posibles escenarios de ruido de metro, murmullos, coche y sala de exposiciones, a cada una de las posibles SNRs de 20dB, 15 dB, 10dB, 5dB o señal limpia.

En el caso de la tarea de evaluación, los ruidos se añaden artificialmente a los datos limpios, habiendo sido grabados en entornos probables para las potenciales aplicaciones de reconocimiento automático del habla: ruido del metro, ruido de una multitud de gente, de un coche, de una sala de exposiciones, de un restaurante, de un aeropuerto y de una estación de trenes. Estos ruidos son añadidos a la base de datos TIDigits a SNRs de 20dB, 15dB, 10dB, 5dB, 0dB y -5dB para el subconjunto de datos de evaluación. Con estos ruidos , las 4004 frases de test pronunciadas por 52 locutores hombres y 52 locutores mujeres, se dividen en 4 grupos con 1001 frases, y definen 3 conjuntos de pruebas diferentes:

- *Set A*: Los ruidos del *metro*, *murmullos* de gente, *coche*, y *sala de exposiciones*, son sumados artificialmente a la señal limpia, a las diferentes SNRs.

- *Set B*: Los ruidos del *restaurante*, de la *calle*, del *aeropuerto* y la *estación de tren*, se suman a la señal limpia las diferentes SNRs.

- *Set C*: En este conjunto de pruebas, los ruidos de *metro* y de *calle* son alterados previamente mediante un *filtrado de característica MIRS* antes de añadirlos a la señal limpia, para añadirles los efectos del canal de comunicación. Este subconjunto de test tendrá pues los efectos del ruido aditivo y del ruido convolucional juntos.

5.3.2. AURORA4

AURORA4 [60] es una base de datos de voz continua estandarizada por el grupo STQ de la ETSI, construida basándose en una tarea de dictado sobre textos del periódico Wall Street Journal con un vocabulario de 5000 palabras.

Tareas Entrenamiento y Evaluación

El conjunto de datos de entrenamiento está formado por 7318 frases grabadas con micrófono de proximidad y sin ruido añadido.

Existe también un conjunto de **datos de entrenamiento** multicondición en AURORA4, que se crea dividiendo las 7318 frases de entrenamiento en dos subconjuntos. Un subconjunto grabado con la característica de micrófono de proximidad, y otro grabado con un micrófono de peor calidad. Después, los 6 ruidos diferentes descritos a continuación son añadidos a las frases de entrenamiento de manera artificial aleatoriamente (una vez que éstas han sido filtradas usando un filtro G.712 o MIRS) eligiendo SNRs entre 10dB y 20dB de modo que los tipos de ruido, incluidos los datos limpios, queden distribuidos de modo uniforme y la SNR media sea de 15dB.

Los 14 *tests* que forman la **tarea de evaluación** se construyen añadiendo ruido grabado a las muestras de voz limpia, con unos niveles de SNR que van desde los 5dB a los 15 dB. Se añaden 7 tipos de ruido diferentes que dan lugar a los 7 primeros tests formados por 166 frases cada uno:

- *set 01*: *test* limpio.
- *set 02*: ruido de coche. Éste es un ruido relativamente estacionario.
- *set 03*: ruido *babble*, de murmullos de conversaciones.
- *set 04*: ruido de un restaurante.
- *set 05*: ruido de la calle, que contiene ruidos no estacionarios.
- *set 06*: ruido de un aeropuerto. Al igual que el *set 05*, contiene ruidos no estacionarios.

- *set 07*: ruido de un tren.

Los 7 *tests* restantes desde el 08 hasta el 14 se forman sumando los mismos 7 ruidos descritos, a grabaciones hechas con 18 micrófonos diferentes que no son de proximidad, por lo que añaden al ruido aditivo de los *tests*, un ruido de canal convolucional. Cada *test* contiene también 166 frases de evaluación.

5.3.3. HIWIRE

La base de datos HIWIRE ([122],[1]) está compuesta por órdenes orales pertenecientes al sistema de comunicación CPDLC [7] (*Controller Pilot Data Link Communications*) entre la tripulación de la cabina de un avión y los controladores de tráfico aéreo. Tradicionalmente estas órdenes son escritas y transmitidas como texto a través de conexiones de datos para evitar usar el canal radio HF entre el avión y la torre de control debido a su poca calidad. El reconocimiento automático hace en este caso que las órdenes orales CPDLC sean escritas automáticamente en el enlace de datos.

Las características de la base de datos son:

- Longitud de las frases: variable desde 1 a 12 palabras, que pueden ser números, caracteres alfabéticos (del alfabeto fonético de la NATO), nombres comunes de instrumentos y órdenes.
- Se usa una gramática determinista.
- Hay 133 palabras y 331 frases diferentes en la base de datos.
- La perplejidad de la gramática es de 14,9.

Tareas de Entrenamiento y Evaluación

La base de datos tiene dos tipos de material:

(i) Una partición limpia: conjunto original de frases limpias grabadas con un micrófono de proximidad.

País	♯ **Hablantes**	♯ **Frases**
Francia	31	3100
Grecia	20	2000
Italia	20	2000
España	10	999
Total	81	8099

Tabla 5.1. Distribución de hablantes por país y número de frases

(ii) Una partición contaminada, que se ha obtenido añadiendo ruidos reales grabados en la cabina de un avión a la partición limpia. La partición contaminada cuenta con tres subconjuntos que tienen SNRs medias de 10dB (*Low Noise)*, 5dB (*Mid Noise*) y -5dB (*High Noise*).

Cada frase de la partición limpia, tiene una homóloga en cada uno de los tres subconjuntos de ruido alto, medio y bajo. El ruido es real y se ha grabado en la cabina de un Boeing 737, con un micrófono de superficie AKG Q300 situado en el cuadro de mandos de dicha cabina. Las grabaciones se han hecho durante vuelos normales y el ruido obtenido es bastante estacionario.

Se han grabado un total de 8099 frases en inglés pronunciadas por 81 hablantes no nativos de Francia, Grecia, Italia y España. Para cada hablante, se han grabado 100 frases por sesión. La tabla 5.1 muestra la distribución de frases por países: las frases limpias han sido grabadas en un PC y muestreadas a 16 Khz, para grabarlas en formato WAV PCM de 16 bits, siendo su SNR media estimada de 30dB. Para generar la partición ruidosa, se le ha añadido artificialmente el ruido grabado a los niveles establecidos, manteniendo el nivel de la señal original.

La base de datos HIWIRE no cuenta con una partición de datos de entrenamiento. Para crear los modelos entrenados, se ha utilizado la base de datos TIMIT [2] con 630 locutores ingleses nativos.

5.4. Resultados de referencia

Los resultados de las tareas de reconocimiento para las dos parametrizaciones de referencia utilizadas en este trabajo, *Baseline* y *Advanced Front-End*, se enumeran a continuación para las 3 bases de datos que son usadas en los trabajos descritos en los siguientes capítulos. Se muestran además los resultados del reconocimiento con modelos entrenados con multicondición, excepto para el caso de la base de datos HIWIRE que no dispone de un conjunto de datos de entrenamiento.

	Set A	**Set B**	**Set C**	**Valor Medio**	**Mejora Relativa**
BASELINE	46,8	51,1	41,3	47,42	0
MULTITRAIN	12,33	13,67	17,73	13,95	70,6 %
AFE	14,14	15,2	18,25	15,39	67,5 %

Tabla 5.2. WER para las tareas de evaluación de AURORA2

	Test 01-07	**Test 08-14**	**Valor Medio**	**Mejora Relativa**
BASELINE	39,63	49,87	44,75	0
MULTITRAIN	25,36	31,30	28,33	36,7 %
AFE	27,65	36,07	31,86	28,8 %

Tabla 5.3. WER para las tareas de evaluación de AURORA4

	Clean	**Low N.**	**Mid N.**	**High N.**	**Valor Medio**	**Mejora Relat.**
BASELINE	8,52	52,05	74,43	97,49	58,12	0
AFE	12,35	28,78	42,47	85,15	42,19	27,4 %

Tabla 5.4. WER para las tareas de evaluación de HIWIRE

Capítulo 6

Ecualización de Histogramas

En este capítulo se hace un análisis exhaustivo de la técnica de ecualización de histogramas como transformación no lineal de los coeficientes cepstrales para robustecer el reconocimiento automático de voz. Se describe el fundamento teórico de la técnica y se analizan los detalles del mismo y de su implementación. Como consecuencia de ese análisis, se plantean experimentos con las tres bases de datos del entorno de trabajo y se extraen conclusiones de los mismos.

6.1. Filosofía de la Ecualización de Histogramas

El efecto del ruido sobre los parámetros MFCC de la señal de voz ha sido caracterizado en el capítulo 3 como:

i) una transformación lineal del vector de características debida al ruido convolucional de canal más,

ii) una transformación no lineal de las los coeficientes cepstrales unida a una transformación aleatoria de los mismos, debida a la naturaleza aleatoria del ruido.

La transformación aleatoria mencionada en $ii)$ es irreversible dada su naturaleza. Para contrarrestar las transformaciones lineales se aplican las técnicas mencionadas en el capítulo 3. Es en el contexto de robustecimiento

del reconocimiento automático de voz frente a las transformaciones no lineales debidas al ruido aditivo, en el que la ecualización de histogramas se presenta como una técnica prometedora.

La ecualización de histogramas es una técnica frecuentemente usada en el procesado digital de imágenes [49],[114], para mejorar el brillo y el contraste, y optimizar el rango dinámico de la escala de grises. De manera simple corrige automáticamente las imágenes demasiado brillantes, demasiado oscuras o con poco contraste. Los valores de los niveles de gris se ajustan dentro de un determinado rango y la entropía de la imagen se maximiza.

A partir de 1.998 con el trabajo de Balchandran [11] la Ecualización de Histogramas (*HEQ-Histogram Equalization*) se empieza a utilizar también en el procesado robusto de voz. Se puede situar dentro de la familia de técnicas de normalización de características basadas en el encuadre estadístico *(statistical matching techniques)*. La filosofía aplicada al usarla en reconocimiento robusto es transformar no linealmente los parámetros de la voz, tanto los de entrenamiento como los de *test*, de modo que se ajusten dentro de un determinado rango común. Esa *igualación* de los rangos o márgenes (*equalize* en inglés, de ahí la expresión) de los parámetros de la emisión original con los que se ha entrenado el reconocedor estadístico, y de los parámetros de con los que se evalúa, tiene el efecto siguiente: el sistema de reconocimiento basado en el clasificador de Bayes se hace idealmente invulnerable a las transformaciones lineales y no lineales que el ruido aditivo Gausiano pueda provocar en los parámetros de *test* una vez ecualizados, siempre que dichas transformaciones sean invertibles.

En otras palabras, el reconocimiento se traslada a un dominio en el que cualquier transformación invertible no perturba ni varía el error de clasificación del clasificador de Bayes. Si *CMN* y *CMVN* normalizan la media y la varianza de las distribuciones de densidad de probabilidad de los coeficientes cepstrales, lo que hace *HEQ* es normalizar las funciones de densidad de probabilidad de los datos de entrenamiento y *test*, transformándolas en una tercera *pdf* común que se convierte en referencia.

El fundamento teórico [26] en el que se apoya la técnica es la propiedad de las variables aleatorias según la cual una variable aleatoria *x* con densidad de probabilidad $p_x(x)$ y función de densidad de probabilidad acumulada $C_x(x)$ puede ser transformada en una variable aleatoria $\hat{x} = T_x(x)$ con una función densidad de probabilidad de referencia $\phi_{\hat{x}}(x)$, conservando idéntica función densidad acumulada, ($C_x(x)$ = $\Phi(\hat{x})$), siempre que la transformación aplicada $T_x(x)$ sea invertible [99].

El hecho de que las *CDFs* se conserven, proporciona una expresión unívoca de la transformación invertible $T_x(x)$ que hay aplicar para que la variable transformada $\hat{x} = T_x(x)$ tenga la *pdf* $\phi_x(\hat{x})$ deseada:

$$\Phi(\hat{x}) = C_x(x) = \Phi(T_x(x)) \tag{6.1}$$

$$\hat{x} = T_x(x) = \Phi_{\hat{x}}^{-1}(C_x(x)) \tag{6.2}$$

La transformación $T_x(x)$ definida en 6.2 es una función monótona no decreciente y no lineal en el caso general, y como se puede ver en 6.2 su expresión está definida en función de la *CDF* de la variable que se transforma.

Una vez que las variables aleatorias han sido ecualizadas, se hacen invulnerables a cualquier transformación lineal o no lineal que se les aplique siempre que sea invertible. Sea *x* una variable aleatoria que sufre una transformación genérica no lineal invertible *G* para convertirse en la variable aleatoria distorsionada *y=G(x)*. Si ambas variables aleatorias distorsionada y no distorsionada se ecualizan a una pdf de referencia ϕ_{ref}, tendremos las variables ecualizadas:

$$\hat{x} = T_x(x) = \Phi_{ref}^{-1}(C_x(x)) \tag{6.3}$$

$$\hat{y} = T_y(y) = \Phi_{ref}^{-1}(C_y(G(x)) \tag{6.4}$$

Si la función *G* es invertible, entonces las *CDFs* de *x* e *y=G(x)* serán iguales:

$$C_x(x) = C_y(G(x)) \tag{6.5}$$

y del mismo modo lo serán las variables transformadas:

$$\hat{x} = T_x(x) = \Phi_{ref}^{-1}(C_x(x) = \Phi_{ref}^{-1}(C_y(G(x))) = \hat{y} \tag{6.6}$$

Las conclusión que se pueden sacar de la expresión 6.6 es que si las variables se ecualizan, el hecho de que hayan sido sometidas a una distorsión invertible, es irrelevante para el entrenamiento y reconocimiento. En el dominio ecualizado tienen idéntico valor.

La efectividad de este método de normalización para reconocimiento robusto en entornos ruidosos se basa en la suposición de que el ruido, que hemos denominado *G* en desarrollo anterior, es una transformación invertible en el espacio de características. En el capítulo 3 vimos que eso no es exacto. El ruido es una variable aleatoria cuyo efecto promedio se supondrá invertible, siendo ese efecto promedio del ruido el que *HEQ* persigue eliminar.

HEQ fue utilizada en procesado de voz por primera vez en 1.998 por Balchandran y Mammone [11]. En esa primera incursión de la ecualización en el campo de la voz, se usó para eliminar las distorsiones no lineales en el Cepstrum LPC de un sistema de identificación de hablantes, usando como distribución de referencia los datos de entrenamiento limpios. En el 2000 Dharanipragada [32] emplea *HEQ* para eliminar el *mismatch* entre el entorno de los auriculares y el del micrófono en un sistema de reconocimiento de voz, y le añade un paso de adaptación *MLLR* concluyendo que trabaja mejor que *MLLR* no supervisado, sumando beneficios al usar ambas técnicas de manera conjunta. A partir de ese momento, la ecualización de histogramas ha sido ampliamente probada e incorporada a los *front-ends* de los reconocedores de voz en entornos ruidosos. Molau, Hilger y Herman Ney la aplican a partir del 2001([79],[59]) en el dominio del banco de

filtros en escala Mel del *front-end* del reconocedor. Implementan *HEQ* conjuntamente con otras técnicas de robustecimiento como LDA *(LDA-Linear Discriminant analysis)* posterior, o VTLN *Vocal Track Length Normalization* consiguiendo resultados de reconocimiento satisfactorios. De la Torre y Segura ([27], [118], [26]) implementan *HEQ* en el dominio de los coeficientes cepstrales y estudian sus beneficios al usarlo conjuntamente con la normalización *VTS*.

6.1.1. Elección del dominio de Ecualización

La parametrización usada por la práctica totalidad de la comunidad científica para el reconocimiento de voz son los coeficientes *MFCC* (Mel frequency cepstral coefficients). Esta parametrización se obtiene al llevar el análisis espectral pasado por un banco de filtros con escala Mel al dominio de la cuefrencia, que se define como la transformada inversa de Fourier del logaritmo del espectro [28]. En dicho dominio temporal, las muestras se llaman coeficientes cepstrales. Los coeficientes cepstrales en escala Mel proporcionan resultados sensiblemente mejores que los del cepstrum LPC, siendo comparables a los de los modelos auditivos, sin la elevada carga computacional de los últimos [24]. En la figura 6.1 se ve el proceso de generación de los MFCCs para una señal de voz.

Hilger y Molau ([59], [79],[78]) ecualizan la salida logarítmica de los bancos de filtros, razonando que la compresión logarítmica disminuye el error de discretización de los histogramas (de ahí que se haga después del logaritmo). Las razones aducidas para ecualizar antes de volver al dominio del tiempo son que de este modo se pueden compensar distorsiones específicas de determinadas frecuencias que tienen efectos independientes en determinados componentes del banco de filtros. Una vez en el dominio del tiempo, esas distorsiones en bandas específicas se han redistribuido y forman parte de todos los coeficientes cepstrales ya que la *DCT* hace una combinación lineal de las salidas de todos los filtros. Sin embargo, en el dominio del banco de filtros las características están fuertemente correladas y una transformación independiente no parece adecuada.

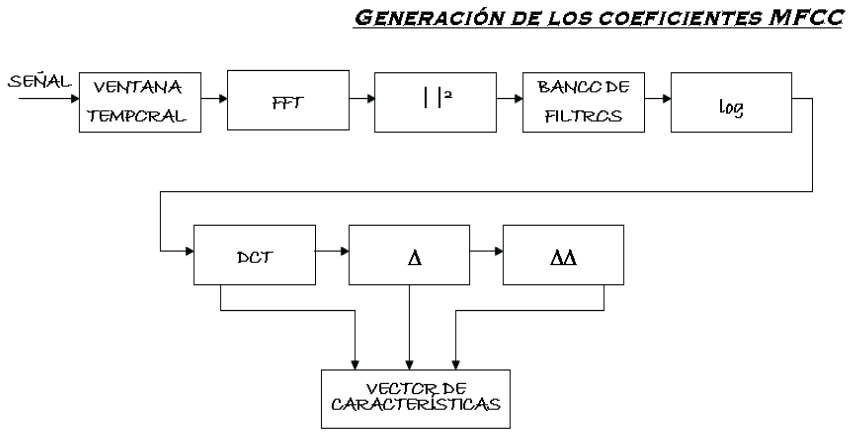


Figura 6.1. Proceso de generación de los coeficientes MFCC

El resto de autores que han usado *HEQ* en procesado de voz, han obtenido mejores resultados normalizando los parámetros MFCC en el dominio del Cepstrum. Tanto los pioneros Balchrandran ([11]) y Darhanipragada ([32]) como De la Torre y Segura ([26], [118]) hacen la ecualización en el dominio de la cuefrencia, actuando sobre los coeficientes cepstrales en escala Mel y sobre sus derivadas.

Hay que mencionar al hablar de dominios la técnica de realimentación de la ecualización que utiliza Obuchi en [87]. Argumenta que los coeficientes de regresión temporal Δ, y $\Delta\Delta$ no son independientes de las características estáticas y propone el cálculo de los coeficientes dinámicos usando los coeficientes estáticos ya ecualizados y reajustando después los estáticos de manera óptima.

6.1.2. Estudio de la distribución de referencia

La elección de la distribución de referencia Φ_{ref} que se utiliza como *CDF* común con la que se ecualizan las variables aleatorias, es una decisión relevante en el proceso de reconocimiento ya que la función de densidad de

probabilidad representa las estadísticas globales de la voz. El desarrollo de la expresión 6.1 da la relación entre las densidades de probabilidad original y de referencia en el dominio ecualizado:

$$p_x(x) = \frac{dC_x(x)}{dx} = \frac{d\Phi(T_x(x))}{dx} = \phi(T_x(x))\frac{dT_x(x)}{dx} = \phi(\hat{x})\frac{dT_x(x)}{dx} \qquad (6.7)$$

Dharanipragada argumenta en [32] la relación que debe haber entre las *pdfs* original y de referencia en términos de información. Utilizando la distancia de Kullback-Liebler como medida de la información mutua existente entre la *pdf* de la variable aleatoria en el dominio original y la *pdf* de referencia para el dominio ecualizado:

$$D(\phi|p_x) = \int_{\hat{x}} \phi(\hat{x}) * log(\phi(\hat{x})) * d\hat{x} - \int_{\hat{x}} \phi(\hat{x}) * log(p_x(\hat{x})) * d\hat{x} \qquad (6.8)$$

se concluye que esta distancia será cero en el caso de que se cumpla 6.9:

$$\phi(\hat{x}) = p_x(\hat{x}) \qquad (6.9)$$

Es difícil encontrar transformaciones $T_x(x)$ que cumplan la condición expresada en 6.9 considerando x y $\hat{x}$ variables aleatorias de dimensión N. Si se acepta la simplificación de independencia entre las dimensiones del vector de características, la expresión 6.9 se puede buscar unidimensionalmente.

Dos distribuciones de referencia han sido utilizadas al implementar *HEQ* para las características de voz. La primera de ellas es la distribución Gausiana, que ha sido ampliamente usada, recibiendo la ecualización con *pdf* de referencia Gausiana el nombre de *Gausianización*. La razón de su uso es principalmente que la función de densidad de probabilidad de la señal de voz tiene una forma cercana a la de una Gausiana bimodal.

Una segunda distribución de referencia que ha reportado resultados mejores que la Gausiana, ha sido la *pdf* de los datos de entrenamiento construida de manera empírica mediante el uso de histogramas acumulativos de los mismos. A continuación se analiza la Ecualización de Histogramas usando ambas *pdfs* de referencia, y sus resultados con el entorno de evaluación manejado.

Gausianización

Chen y Gopinath proponen [17] una transformación de Gausianización para modelado de datos con muchas dimensiones, que alterna pasadas de transformaciones lineales para conseguir independencia entre las dimensiones, con pasadas de Gausianización marginal individual de esas dimensiones con técnicas de una variable. Ese es el origen de la Gausianización como técnica de escalado de la distribución de probabilidad. Ha sido aplicada con éxito por muchos autores [136], [116], [95], [98], [27]. Saon y Darhanipragada señalan una ventaja fundamental al usarla, que tiene que ver con el hecho de que en la mayoría de los sistemas las distribuciones de salida de los *HMMs* se modelan como mezcla de Gausianas con covarianzas diagonales. Es razonable esperar que el hecho de *Gausianizar* las características refuerce esta asunción.

Referencia limpia

La elección de la densidad de probabilidad de referencia como la de los datos del entorno de entrenamiento, se puede analizar como la versión no paramétrica de la Gausianización que asumía que la densidad de probabilidad de los datos de entrenamiento tiene forma Gausiana. La realidad es que la *pdf* de estos datos tiene una forma muy parecida a una mezcla de dos Gausianas, cuyos valores de medias y varianzas tienen que ver con peculiaridades de la base de datos y el proceso de transmisión. El calcular de manera empírica esa *pdf* exige la condición de que los datos con los que se cuenta sean suficientes para representar de manera no sesgada la

estadística global de la voz.

6.1.3. Ecualización combinada con otros métodos

La ecualización de histogramas es una técnica de robustecimiento que no hace ninguna suposición a priori sobre el modelo de la distorsión que sufre el sistema. Esto tiene sus ventajas ya que la hace enfrentarse a distorsiones de comportamiento no predecible o no modelado. Por esa razón *HEQ* es un candidato interesante para eliminar distorsiones residuales después de la aplicación de técnicas de robustecimiento basadas en normalización a priori del ruido, ya sea en el dominio del tiempo (ver [120]), o en el de la frecuencia (ver [121]). La aplicación de *HEQ* en el dominio cepstral después de aplicar VTS [81] en el dominio de la energía logarítmica del banco de filtros, merece una mención especial por los buenos resultados que da al combatir el ruido residual de *VTS* [117].

6.2. Aspectos de la implementación de HEQ

6.2.1. Estudio de la transformación

El proceso que se sigue para ecualizar cada componente del vector de características es el siguiente [97], [27], [26]:

i) Se calcula el histograma acumulativo de referencia del entorno ecualizado $\Phi(\hat{x})$ usando los datos de entrenamiento limpio. La CDF de referencia de los datos de entrenamiento se aproxima con este histograma.

ii) Para cada frase se calcula el histograma acumulativo de todos los valores del componente que se quiere ecualizar, usándose para aproximar su CDF $C_x(x)$. Tanto en el caso del histograma de los datos de entrenamiento como en éste, para construir el histograma se utilizan 100 *bins* igualmente espaciados en el rango de $[\mu - 4\sigma, \mu + 4\sigma]$ siendo

μ y σ la media y varianza del coeficiente que se ecualiza. Si tenemos un conjunto de N observaciones correspondientes a los valores de un coeficiente cepstral en una frase dada, la *pdf* se puede aproximar por su histograma como:

$$p_x(x \epsilon B_i) = \frac{n_i}{N} \tag{6.10}$$

siendo n_i el número de observaciones del bin B_i.

La función de densidad acumulada CDF se aproximará por:

$$C_x(x_i) = C_x(x \epsilon B_i) = \Sigma_{j=1}^{i} \frac{n_j}{N} \tag{6.11}$$

iii) Se eligen una serie de puntos del histograma acumulativo del dominio original (por ejemplo los centros de los *bins* del histograma), a los que se les aplica la expresión de la transformación de ecualización definida en 6.3:

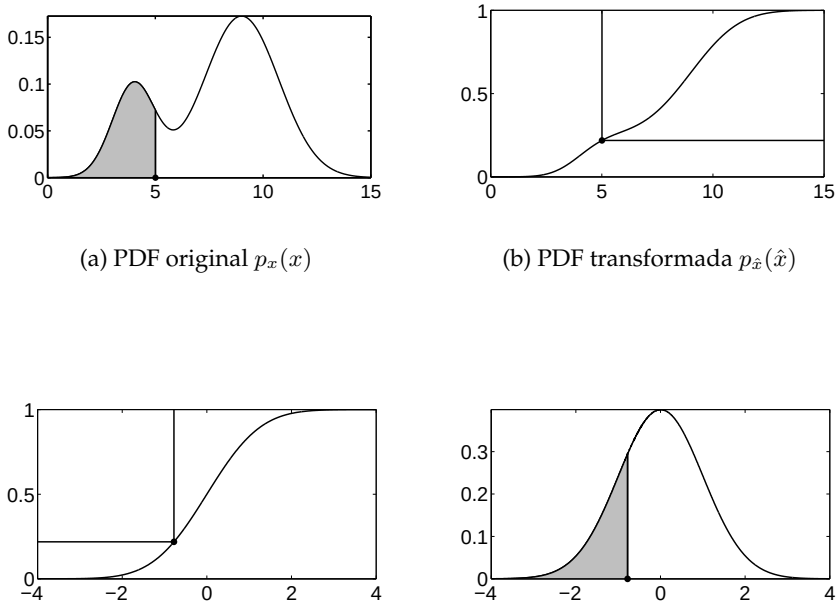
$$\hat{x} = T_x(x) = \Phi^{-1}(C_x(x)) \tag{6.12}$$

iv) Para el resto de puntos del dominio original, se aproxima la transformación utilizando la tabla de pares de puntos así obtenida.

La manera más efectiva y común de definir la transformación es mediante aproximación lineal a trazos que conecta los puntos de la tabla. Hilger propone en [58] el uso de una función cuadrática en vez de lineal, con el objetivo de eliminar los puntos de discontinuidad de la transformación lineal a trozos. Esta transformación más compleja le reporta mejores resultados para datos con un rango dinámico grande. La figura 6.3 muestra un ejemplo de la transformación definida siendo la PDF de referencia elegida una Gausiana Normal. Las sub-figuras (a) y (b) muestran las PDFs original y transformada. Para encontrar el valor transformado $\hat{x_0}$ dado el valor x_0, se siguen dos pasos:

1. Se busca el valor de la CDF original $C_x(x_0)$.
2. Se encuentra el valor $\hat{x_0}$ como aquel cuya CDF transformada tiene el mismo valor que $C_x(x_0)$. (ver sub-figuras (c) y (d) de la figura 6.2).

La transformación así definida cuando el proceso se repite para todos los puntos del dominio original es la que aparece en la figura 6.3.



(a) PDF original $p_x(x)$

(b) PDF transformada $p_{\hat{x}}(\hat{x})$

(c) CDF original $C_x(x)$

(d) CDF transformada $\Phi(\hat{x})$

Figura 6.2. Estadísticas en los dominios original y transformado

6.2.2. *QBEQ*

El primer paso de optimización computacional del algoritmo es la utilización de cuantiles muestrales en lo que se llama Ecualización Basada en Cuantiles, *QBEQ* [58], [118]. La implementación de QBEQ es la siguiente:

i) En primer lugar se genera la estadística ordenada de la frase que se ecualiza. Si el número total de tramas de la frase son $2T + 1$, esos $2T + 1$ valores ordenados estadísticamente serían los de la expresión

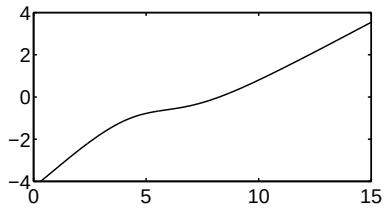


Figura 6.3. Transformación $T_x(x)$ entre las pdfs (a) y (b)

6.13, en la que $x_{(r)}$ es el la trama cuyo valor ocupa la posición r-ésima en la secuencia ordenada:

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(r)} \leq ... \leq x_{(2T+1)} \tag{6.13}$$

ii) Se calculan un conjunto de cuantiles muestrales de la *CDF* de referencia. Se elige el número de cuantiles N_Q de la muestra y se registran los valores de la *CDF* para los valores de probabilidad p_r de cada cuantil:

$$Q_{\hat{x}}(p_r) = \Phi^{-1}(p_r) \tag{6.14}$$

$$p_r = (\frac{r - 0,5}{N_Q}), \forall r = 1, ..., N_Q \tag{6.15}$$

iii) Los cuantiles muestrales de los datos originales obedecen a la expresión 6.16 en la que *k* y *f* son la parte entera y fraccionaria de $(1+2Tp_r)$:

$$Q_x(p_r) = \begin{cases} (1-f)x_k + fx_{k+1}, & 1 \leq k \leq 2T; \\ x_{(2T+1)}, & k = 2T+1. \end{cases} \tag{6.16}$$

iv) Cada una de las parejas de cuantiles $(Q_x(p_r), Q_{\hat{x}(p_r)})$ representan un punto de la transformación de ecualización que se aproxima lineal-

mente a trazos.

El coste computacional de esta versión optimizada de *HEQ* es en promedio de $(2T+1)log_2(2T+1)$ comparaciones para obtener la estadística ordenada, más $2N_Q$ productos y N_Q sumas para el cálculo de los N_Q cuantiles, a los que se les suma dos productos y dos sumas en el proceso de interpolación. Si se utiliza un número reducido de cuantiles el coste computacional del algoritmo es significativamente inferior al de *HEQ*.

6.2.3. *OSEQ*

Una versión aún más eficaz computacionalmente es la propuesta por Segura en [118] con el nombre de Ecualización basada en la Estadística Ordenada. Es adecuada cuando las *CDFs* se estiman sobre una ventana deslizante de longitud fija. Esta versión propone calcular un estimador puntual no sesgado del valor de la $C_x(x)$ con la expresión 6.17:

$$\hat{C_x}(x_{(r)}) = \frac{r-0,5}{2T+1} \forall r = 1, ..., 2T+1 \tag{6.17}$$

a partir de la cual se puede obtener el valor transformado de x_t del modo:

$$\hat{x_t} = \Phi^{-1}(\frac{r(x_t-0,5)}{2T+1}) \tag{6.18}$$

donde $1 \leq r(x_t) \leq 2T+1$ es el número de orden de x_t , es decir el índice *r* de la estadística ordenada con igual valor que x_t. OSEQ es especialmente eficiente cuando se implementa con una ventana deslizante. El aumento de la eficacia computacional del algoritmo viene dado por el hecho que ya que tanto la *CDF* Φ como la longitud de la frase o segmento *T* son fijos y por lo tanto se puede tabular una función $G(r)$:

$$G[r] = \Phi^{-1}(\frac{r-0,5}{2T+1}) \forall r = 1, ..., 2T+1 \tag{6.19}$$

De esta manera, el valor ecualizado $\hat{x_t}$ se obtiene por simple indexación, para lo que sólo hacen falta $2T$ comparaciones para ecualizar

cada componente.

6.2.4. Ecualización *on-line*

Existen dos escenarios en los que la ecualización de histogramas puede presentar ciertas limitaciones:

- Aplicaciones con condicionamientos de tiempo real en las que es deseable eliminar el retardo que implica tener todas las tramas de la frase para calcular su *CDF* original antes de hacer la ecualización.
- Tareas de evaluación con frases de una sola palabra, o demasiado cortas como para obtener de ellas una estadística fiable de los datos acústicos.

Una posible solución para estos escenarios es construir una *CDF global de los datos de test*, de manera que cada frase se ecualiza en tiempo real usando esta *CDF global* acumulada de las frases de *test* anteriores. Una vez ecualizada la frase sus datos estadísticos se incorporan a la CDF global de los datos de *test*, con un determinado peso α. Como valor inicial esta *CDF* se iguala a la *CDF* de referencia. La figura 6.4 muestra el proceso:

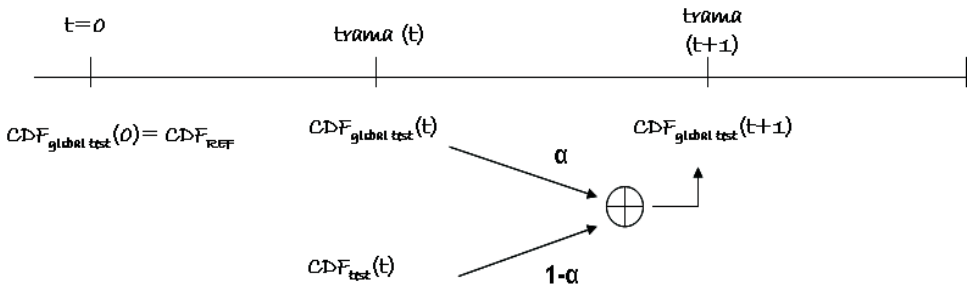


Figura 6.4. Proceso de Ecualización *on-line*

- CDF_{test} representa las función de densidad acumulada de la frase, utilizando sólo las t primeras tramas de la misma para calcularla.

- CDF_{global_test} representa las estadísticas globales acumuladas de las frases de *test* en las frases y tramas anteriores. Los n cuantiles muestrales ($Q(i), i = 1, ..., n$) de esta estadística global acumulada se actualizan para cada trama con la contribución de los cuantiles muestrales de CDF_{test} de acuerdo con la siguiente expresión:

$$Q(i)_{global\_test}(t+1) = \alpha \cdot Q(i)_{global\_test}(t) + (1-\alpha) \cdot Q(i)_{test}(t) \quad (6.20)$$

6.3. Experimentación y resultados

6.3.1. Análisis de la distribución de referencia

La mayor parte de los autores que usan Ecualización de Histogramas utilizan una CDF de referencia construida con los datos de entrenamiento [79], [59], [32], obteniendo con ello mejores resultados que al usar una referencia Gausiana. Esta tendencia se encuentra también en la ecualización implementada en nuestro entorno de desarrollo, como demuestran los resultados de los experimentos enumerados en las tablas 6.1, 6.2 y 6.3. Este comportamiento es explicable con la expresión vista en la ecuación 6.9. La distancia de Kullback-Liebler entre la *pdf* original y la de referencia es menor, (conservándose por tanto en la transformación mayor información sobre la estadística de voz original) cuanto más se parecen ambas *pdfs*. La *pdf* de referencia construida empíricamente es más parecida a la *pdf* original que una Gausiana.

Las tablas 6.1, 6.2 y 6.3 muestran los resultados de la implementación del algoritmo QBEQ con 31 cuantiles muestrales para ecualizar los coeficientes MFCC de las bases de datos AURORA2, AURORA4 y HIWIRE. La evaluación de dichas bases de datos muestra que los resultados de los ex-

perimentos con la *pdf* de los datos de entrenamiento limpios (llamado en las tablas experimento *REFCLEAN*) son mejores que los resultados de los experimentos con la *pdf* Gausiana (llamados en las tablas *ECDFG*). El caso más evidente es el de la base de datos HIWIRE para hablantes no nativos, en el cual, el usar una pdf Gausiana da resultados un 20,9 % peores que no hacer ningún tipo de ecualización (experimento *BASELINE* en las tablas). Esto debe a que la distribución de probabilidad de los datos en el dominio original dista mucho de ser Gausiana y la información mutua conservada al ecualizar a una Gausiana es muy deficiente.

	Set A	**Set B**	**Set C**	**Valor Medio**	**Mejora Relativa**
BASELINE	46,8	51,1	41,3	47,42	0
ECDFG	20,72	18,14	20,44	19,63	58,6 %
REFCLEAN	19,33	17,3	18,97	**18,41**	**61,1** %
MULTITRAIN	12,33	13,67	17,73	13,95	70,6 %
AFE	14,14	15,2	18,25	15,39	67,5 %

Tabla 6.1. WER en AURORA2. Estudio de las CDFs de referencia

	Test 01-07	**Test 08-14**	**Valor Medio**	**Mejora Relativa**
BASELINE	39,63	49,87	44,75	0
ECDFG	35,92	46,13	41,03	8,3 %
REFCLEAN	32,35	42,00	**37,18**	**16,9** %
MULTITRAIN	25,36	31,30	28,33	36,7 %
AFE	27,65	36,07	31,86	28,8 %

Tabla 6.2. WER en AURORA4. Estudio de las CDFs de referencia

	Clean	Low N.	Mid N.	High N.	Valor Medio	Mejora Relat.
BASELINE	8,52	52,05	74,43	97,49	58,12	0
ECDFG	56,12	63,30	72,04	89,57	70,26	-20,9 %
REFCLEAN	14,02	47,16	61,84	88,70	**52,93**	**9,0** %
AFE	12,35	28,78	42,47	85,15	42,19	27,4 %

Tabla 6.3. WER en HIWIRE. Estudio de las CDFs de referencia

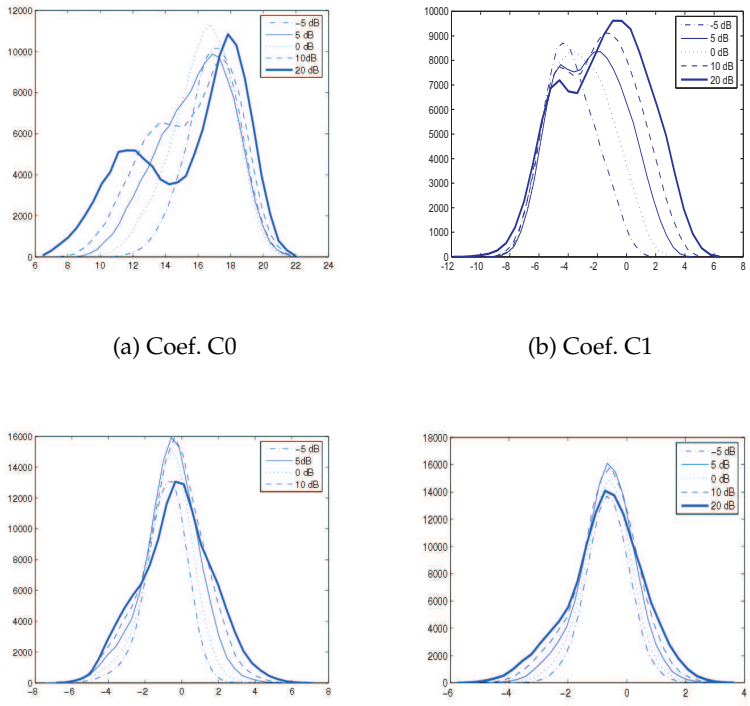
6.3.2. Ecualización progresiva de coeficientes cepstrales

La ecualización de los MFCCs de la frases consigue robustecerlos ya que su *pdf* se ajusta a la de referencia, siendo ésta una representación equilibrada de las estadísticas globales de la señal de voz. Sin embargo este robustecimiento lleva unida una distorsión.

La razón de que aparezca esta distorsión es que la transformación se define usando la *pdf* de la frase que se ha de ecualizar como podemos ver en la ecuación 6.2, y dicha *pdf* se calcula con pocos datos. Debido a esto, la *pdf* de la frase que se ecualiza es sólo una aproximación más o menos fiable (según el tamaño de la frase) de la estadística global de la señal de voz en el entorno. Este grado de fiabilidad se traslada a la transformación con ella definida.

Del estudio individual de los coeficientes cepstrales [65],[94] se concluye que tienen propiedades estadísticas diferentes y capacidad discriminativa decreciente al aumentar el orden del coeficiente, siendo afectados de manera desigual por el ruido.

En lo que a las propiedades estadísticas se refiere, los coeficientes cepstrales de menor orden tienen una varianza mayor que los de orden alto [128] y su capacidad discriminativa es mucho mayor. En el proceso de cálculo de los MFCCs, la FFT inversa equivale a un escalado de orden bajo mediante coseno de los logaritmos de las energías espectrales a la salida del banco de filtros en escala Mel. El coeficiente *C0* representa la energía y el *C1* el balance global de las energías entre las altas y bajas frecuen-



(a) Coef. C0

(b) Coef. C1

(c) Coef. C2

(d) Coef. C3

Figura 6.5. Histograma de los MFCCs *C0*,*C1*,*C2* y *C3*, a diferentes SNRs

cias, también llamado decaimiento promedio del espectro o *tilt* espectral. El resto de *Ci*s son difícilmente relacionables con aspectos fundamentales de la producción o percepción de voz. Se usan sabiendo que contienen detalles espectrales cuya fineza aumenta al incrementarse el orden, y que su conjunto permite discriminar entre sonidos similares. Esta falta de identificabilidad con aspectos concretos los hace vulnerables a las condiciones acústicas. En particular, el hecho de que cada *Ci* esté afectado por todo el rango de frecuencias dada su construcción como combinación lineal de las salidas del banco de filtros, limita la capacidad de la parametrización para resistir ruidos específicos de una determinada frecuencia.

	Set A	**Set B**	**Set C**	**Valor medio**	**Mejora Relativa**
BASELINE	46,8	51,1	41,3	47,42	0
ECDF, C0	26,93	22,65	28,05	25,44	40,08 %
ECDF, C0-C1	23,18	20,13	24,05	22,14	53 %
ECDF, C0-C2	20,8	18,33	21,74	20	57,8 %
ECDF, C0-C3	19,76	18,85	20,12	19,47	54,4 %
ECDF, C0-C4	19,13	16,83	19,53	18,29	62,1 %
ECDF, C0-C5	18,55	16,82	19,03	17,95	62,1 %
ECDF, C0-C6	16,39	15,66	18,64	**16,55**	**65,1 %**
ECDF, C0-C7	18,22	16,32	18,45	17,50	63,1 %
ECDF, C0-C8	18,25	16,29	18,37	17,49	63,1 %
ECDF, C0-C9	18,17	16,35	18,29	17,47	63,2 %
ECDF, C0-C10	18,31	16,43	18,95	17,69	62,7 %
ECDF, C0-C11	18,64	16,64	18,41	17,79	62,5 %
ECDF, C0-C12	19,33	17,3	18,97	18,44	61,1 %
MULTITRAIN	12,33	13,67	17,73	13,95	70,6 %
AFE	14,14	15,2	18,25	15,39	67,5 %

Tabla 6.4. WER para AURORA2. Ecualización progresiva de los MFCCs

La figuras 6.5 muestran la *pdf* de cada uno de los 4 primeros coeficientes cepstrales para un determinado grupo de frases con diferentes SNRs. Se puede apreciar que el coeficiente *C0* es el más afectado por el nivel de ruido, siendo este efecto mucho menos apreciable en los coeficientes de orden alto. En los coeficientes de orden 0 hasta 2 se ve la distorsión lineal debida al ruido que se traduce en un desplazamiento de la media y un escalado de la varianza. Las figuras muestran también los efectos de la distorsión no lineal: el cambio en la forma de la *pdf* que lleva incluso a transformar una *pdf* bimodal (con los picos en los valores medios del ruido y la voz) en una unimodal para los 3 primeros MFCCs.

Esta diferencia de comportamiento frente al ruido y de capacidad discriminativa de los distintos coeficientes ha sido atacada de modos diversos. En [128] se propone definir una distancia de cepstral entre los MFCCs de

	Test 01-07	**Test 08-14**	**Valor Medio**	**Mejora Relativa**
BASELINE	39,63	49,87	44,75	0
ECDF, C0	34,34	43,06	38,70	13,5 %
ECDF, C0-C1	33,81	44,03	38,92	13 %
ECDF, C0-C2	32,47	42,20	37,33	16,5 %
ECDF, C0-C3	32,85	42,80	37,82	15,4 %
ECDF, C0-C4	31,85	41,85	**36,85**	**17,6 %**
ECDF, C0-C5	32,62	42,60	37,61	15,9 %
ECDF, C0-C6	32,03	42,03	37,03	17,02 %
ECDF, C0-C7	32,54	42,33	37,43	16,3 %
ECDF, C0-C8	32,47	41,75	37,11	17,0 %
ECDF, C0-C9	32,95	42,47	37,71	15,7 %
ECDF, C0-C10	32,94	42,55	37,75	15,6 %
ECDF, C0-C11	32,54	42,62	37,58	16,0 %
ECDF, C0-C12	32,35	42,00	37,18	16,9 %
MULTITRAIN	25,36	31,30	28,33	36,7 %
AFE	27,65	36,07	31,86	28,8 %

Tabla 6.5. WER para AURORA4. Ecualización progresiva de los MFCCs.

entrenamiento y *test*, y pesar esa distancia con unos factores de correción definidos como la inversa de la varianza estadística del *Ci*. Dado que la varianza del *Ci* es mayor en los primeros coeficientes, este factor de escalado implica una influencia mucho mayor de éstos en la caracterización de la información de voz. Otros autores como [65] usando coeficientes LPC proponen usar una ventana de *liftering* para reducir los efectos no deseados y aplican esa ventana de *liftering* a los coeficientes LPC de menor orden solamente.

En el contexto de la Ecualización de Histogramas, Wet señala en [29] los beneficios de incluir en el conjunto de coeficientes que se ecualizan el logaritmo de la energía, siendo ésta la responsable del mayor incremento de la tasa de reconocimiento exitoso. En el trabajo [119] de Segura se estudia el efecto de la ecualización progresiva de un determinado número de

	Clean	**Low N.**	**Mid N.**	**High N.**	**Valor Medio**	**Mejora Relat.**
BASELINE	8,52	52,05	74,43	97,49	58,12	0
ECDF, C0	11,07	41,73	62,12	93,17	52,02	10,5 %
ECDF, C0-C1	10,83	44,58	67,49	95,53	54,61	6,0 %
ECDF, C0-C2	12,59	38,39	53,67	88,66	47,33	18,6 %
ECDF, C0-C3	11,73	38,06	52,44	83,89	46,53	19,9 %
ECDF, C0-C4	12,31	38,70	52,25	81,43	**46,17**	**20,6 %**
ECDF, C0-C5	13,35	39,92	53,91	82,35	47,38	18,5 %
ECDF, C0-C6	13,63	42,84	55,31	82,35	48,53	16,5 %
ECDF, C0-C7	13,23	42,32	54,99	82,99	48,38	16,8 %
ECDF, C0-C8	13,22	43,13	56,02	84,36	49,18	15,4 %
ECDF, C0-C9	14,21	44,58	57,83	84,93	50,38	13,3 %
ECDF, C0-C10	13,46	44,98	59,10	86,56	51,03	12,3 %
ECDF, C0-C11	12,87	45,77	59,22	86,57	51,11	12,1 %
ECDF, C0-C12	14,02	47,16	61,84	93,17	52,02	10,5 %
AFE	12,35	28,78	42,47	85,15	42,19	27,4 %

Tabla 6.6. WER para HIWIRE. Ecualización progresiva de los MFCCs.

coeficientes. Dado que la ecualización robustece frente al ruido pero al mismo tiempo introduce una cierta distorsión debida al cálculo aproximado de la *pdf* de la frase que se ecualiza, y dado que la mayor capacidad discriminativa se encuentra en los coeficientes de menor orden, tiene interés considerar la ecualización de un determinado número de *Cis*. Las conclusiones de [119] son que la tasa de reconocimiento se incrementa a medida que se ecualizan más características, hasta llegar a una tasa de reconocimiento asintótica que se obtiene con un determinado índice de *Ci*. Dicho índice será menor cuanto mayor sea la SNR y una vez superado la tasa de reconocimiento se puede deteriorar levemente.

Las tablas 6.4, 6.5 y 6.6 muestran los resultados de los *test* de AURORA2, AURORA4 y HIWIRE respectivamente para una ecualización de un número progresivo de coeficientes MFCC. Podemos observar la siguiente pauta de comportamiento. Para las tres bases de datos, los valores ópti-

mos de reconocimiento no se obtienen ecualizando todos los coeficientes cepstrales. La ecualización de *C0* y *C1* es responsable del mayor incremento en la tasa de reconocimiento, y el número óptimo de coeficientes que es deseable ecualizar aumenta cuando hay ruido de canal añadido al ruido aditivo. (Es el caso del *Set C* de AURORA2, y los *Tests* desde el *08* al *14* para el caso de AURORA4). En general, AURORA2 funciona de manera óptima con mayor número de coeficientes ecualizados que AURORA4.

La mejora obtenida al ecualizar la cantidad óptima de coeficientes cepstrales para los *tests* que tienen solamente ruido aditivo, y ecualizar la cantidad óptima diferente para los *tests* que tienen el ruido aditivo y ruido de canal en otro caso se puede observar con los mejores resultados en las tablas 6.7, 6.8 y 6.9. En estas tablas se han elegido las ecualizaciones con un número óptimo de coeficiente ecualizados para cada tipo de ruido. En el caso de la base de datos HIWIRE, la mejora al ecualizar un número distinto de coeficientes para cada condición de ruido se incrementa en un 13,5 %.

	Set A	**Set B**	**Set C**	**Valor medio**	**Mejora Relativa**
ECDF, C óptimo	16,39 *(C0-C6)*	15,66 *(C0-C6)*	17,47 *(C0-C9)*	16,32	65 %
ECDF, C0-C12	19,33	17,3	18,97	18,44	61,1 %

Tabla 6.7. Mejores *WER*s en AURORA2. Ecualización progresiva.

	Test 01-07	**Test 08-14**	**Valor Medio**	**Mejora Relativa**
ECDF, C óptimo	31,85 *(C0-C4)*	41,75 *(C0-C8)*	36,08	19,4 %
ECDF, C0-C12	32,35	42,00	37,18	16,9 %

Tabla 6.8. Mejores *WER*s en AURORA4. Ecualización progresiva.

	Clean	**Low N.**	**Mid N.**	**High N.**	**Valor medio**	**Mejora Relativa**
ECDF, C óptimo	8,52 *(Baseline)*	38,06 *(C0-C3)*	52,25 *(C0-C4)*	81,43 *(C0-C4)*	45,2	22,4 %
ECDF, C0-C12	14,02	47,16	61,84	88,70	52,93	*8,9 %*

Tabla 6.9. Mejores *WER*s en HIWIRE. Ecualización progresiva.

6.3.3. Ecualización *on-line*

Las tablas 6.10, 6.11 y 6.12 muestran los resultados de aplicar esta variante de la ecualización descrita en la sección 6.2.4, que tiene los mejores resultados para AURORA2, en cuyo caso y para un $\alpha = 0,4$ se disminuye la mejora relativa solo en un 0,8 % respecto a la versión original de la ecualización de histogramas que aparece en las tablas como *REFCLEAN*. Para el caso de AURORA4, el α óptimo es también 0,4 y el empeoramiento relativo es del 6,1 % respecto a REFCLEAN. En el caso de la base de datos HIWIRE los resultados son malos en cualquier caso, siendo mejor no ecualizar si las restricciones temporales impiden aplicar la Ecualización en su versión "*lenta*".

	Set A	**Set B**	**Set C**	**Valor medio**	**Mejora Relativa**
BASELINE	46,8	51,1	41,3	47,42	0
ECDF,i2r, α =0,8	20,88	19,78	20,29	20,32	57 %
ECDF,i2r, α =0,6	19,67	19,01	19,68	19,41	59 %
ECDF,i2r, α =0,4	19,08	18,55	19,43	18,94	60 %
ECDF, i2r, α =0,2	19,41	19,08	20,11	19,42	59 %
REFCLEAN	19,33	17,3	18,97	18,41	61,1 %
MULTITRAIN	12,33	13,67	17,73	13,95	70,6 %
AFE	14,14	15,2	18,25	15,39	67,5 %

Tabla 6.10. WER en AURORA2. Ecualización *on-line* con diferentes α .

6.4. Resultados y conclusiones

Una vez vista su eficacia en los resultados numéricos presentados en este capítulo, las ventajas de *HEQ* para ser usado como algoritmo de robustecimiento del *front-end* de un sistema ASR se pueden resumir como:

	Test 01-07	**Test 08-14**	**Valor Medio**	**Mejora Relativa**
BASELINE	39,63	49,87	44,75	0
ECDF, i2r, α =0,8	34,64	46,20	40,42	9,6 %
ECDF, i2r, α =0,6	34,68	45,15	39,92	10,7 %
ECDF, i2r, α =0,4	34,83	44,98	39,91	10,8 %
ECDF, i2r, α =0,2	34,73	45,22	39,98	10,6 %
REFCLEAN	32,35	42,00	37,18	16,9 %
MULTITRAIN	25,36	31,30	28,33	36,7 %
AFE	27,65	36,07	31,86	28,8 %

Tabla 6.11. WER en AURORA4. Ecualización *on-line* con diferentes α.

	Clean	**Low N.**	**Mid N.**	**High N.**	**Valor medio**	**Mejora Relat.**
BASELINE	8,52	52,05	74,43	97,49	58,12	0
ECDF, i2r, α =0,8	21,17	56,03	70,93	92,40	60,13	-3,5 %
ECDF, i2r, α =0,6	22,36	58,14	72,11	92,76	61,34	-5,5 %
ECDF, i2r, α =0,4	24,62	60,41	74,16	93,27	63,11	-9 %
ECDF, i2r, α =0,2	29,28	64,66	76,25	93,86	66,01	-13,6 %
REFCLEAN	14,02	47,16	61,84	88,70	52,93	8,9 %
AFE	12,35	28,78	42,47	85,15	42,19	27,4 %

Tabla 6.12. WER en HIWIRE. Ecualización *on-line* con diferentes α.

- Se aplica en el dominio de las características, siendo independiente del *back-end* del reconocedor.
- No requiere información a priori sobre el tipo de ruido o SNRs que se esperan durante el reconocimiento. Esto hace que sea útil para ruidos cuyo modelo se desconoce o para combinaciones de varios tipos de ruido.
- Su éxito no depende de una detección fiable de los silencios ni de la estimación del ruido durante la detección de características.

- Es computacionalmente barato.
- Se puede aplicar a sistemas en tiempo real como aplicaciones de comandos o control de sistemas de diálogo.

Existen una serie de limitaciones del algoritmo descrito que justifican el desarrollo de nuevas versiones del mismo para combatirlas:

- Su efectividad depende del cálculo adecuado de las *CDFs* original y de referencia de los datos que se ecualizan. No en todos los escenarios posibles las frases tienen la longitud suficiente para dar una estadística global de la voz y definir así la *CDF* original de la frase y con ella la transformación que hay que aplicar.
- Esta técnica considera independencia estadística de los MFCCs entre sí. Esto no es del todo correcto. En el caso real la matriz de covarianzas de los *Cis* no es diagonal. Sería deseable capturar de algún modo esta dependencia entre componentes.

Capítulo 7

Ecualización Paramétrica de Histogramas

Este capítulo propone un algoritmo de ecualización paramétrica en dos clases llamado *PEQ* cuyo objetivo es aproximarse mediante una expresión paramétrica a la Ecualización de Histogramas tratada en el capítulo 6. La aproximación propuesta utiliza un modelo de dos Gausianas para definir dos clases en las que separan las tramas voz y silencio, que son ecualizadas por separado. El capítulo hace un encuadre del algoritmo propuesto *PEQ* dentro del conjunto de técnicas de ecualización paramétrica existentes, y una comparación exhaustiva del mismo con *HEQ* mediante experimentos con las bases de datos AURORA2, AURORA4 y HIWIRE.

7.1. Filosofía de la Ecualización Paramétrica en clases

Introducción

El capítulo anterior presentaba la Ecualización de Histogramas como una técnica de normalización no lineal de características, atractiva para aplicaciones de reconocimiento automático de voz en entornos ruidosos. Las razones de su conveniencia son su bajo coste computacional y la ausencia de modelos predeterminados de ruido que la hace útil para entornos

cuyas características de ruido se desconocen y/o tienen mezcla de varios tipos. La capacidad para combatir los efectos no lineales del ruido es su mayor aliciente. Las limitaciones de HEQ son fundamentalmente las que ya esbozamos en el capítulo anterior:

i) En primer lugar, es necesaria una cantidad mínima de datos en las frases que se ecualizan para definir de manera fiable la estadística global de voz de la frase del entorno ruidoso, y con ella la transformación de ecualización. Esta **falta de datos para generar estadísticas fiables** por frase se refleja también en el hecho de que el porcentaje de tramas de voz y silencio que contiene una frase, influye de manera no deseable en la *CDF* calculada y con ello en la transformación definida para ecualizar la frase.

La **influencia del porcentaje de tramas de silencio** se puede observar en la figura 7.1, en la que la sub-figura (a) muestra el valor en el tiempo del coeficiente cepstral *C1* para una frase típica, y la sub-figura (b) muestra el valor en el tiempo del coeficiente *C1* para la misma frase quitando parte del silencio inicial. Las funciones de densidad acumulada de ambas frases se muestran en la figura (c) en la que es apreciable que aunque ambas frases tienen el mismo valor para las tramas de voz, la cantidad diferente de tramas de silencio en ambos casos altera la *CDF* global. Esta diferencia en las *CDF* estimadas, induce una variabilidad no deseada en la transformación estimada como se puede ver en la sub-figura (d).

La razón de esta variabilidad no deseada es clara si expresamos la *CDF* como mezcla de dos *CDFs* correspondientes a las tramas de voz y de silencio, siendo α la fracción de tramas de silencio, y $C_{nx}(x)$ y $C_{sx}(x)$ las *CDFs* respectivas de las clases de silencio:

$$C_x(x) = \alpha \cdot C_{nx}(x) + (1 - \alpha) \cdot C_{sx}(x) \tag{7.1}$$

Incluso en el caso de que las distribuciones de probabilidad no se alterasen, diferentes valores de α originan diferentes $C_x(x)$.

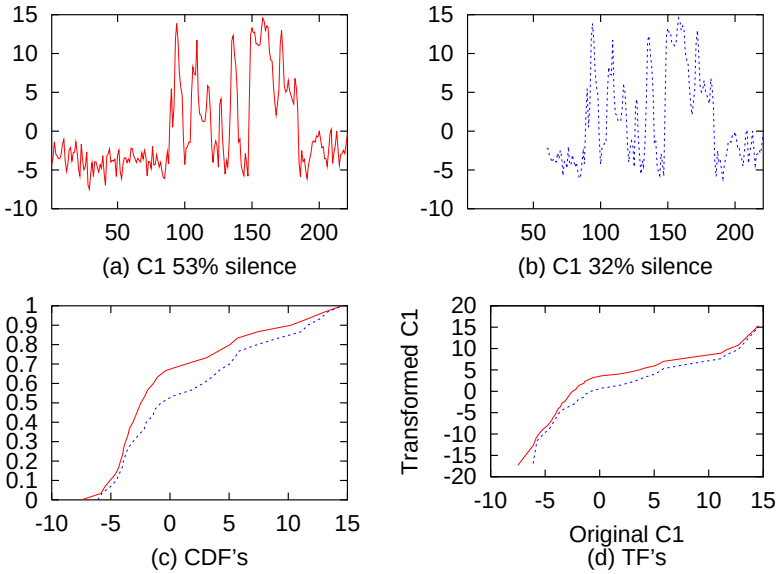


Figura 7.1. Influencia del porcentaje de silencio

Las estrategias existentes para abordar la exigencia de frases con una longitud mínima son principalmente el utilizar una expresión paramétrica de la CDF como proponen varios autores (ver [77], [53], [70]). Se parametrizan los histogramas de las *CDFs* de entrenamiento y *test* con Gausianas identificadas por sus medias y varianzas. El uso del método de la estadística ordenada [118] también mejora la estimación de la CDF cuando las frases del entorno ruidoso son cortas aunque en menor medida que una expresión paramétrica de las mismas.

La influencia del porcentaje de ruido de la frase de *test* al definir la transformación es, como hemos dicho, otra consecuencia de la falta de datos para definir la estadística global de la frase. Esta limitación se aborda con distintos enfoques, siendo lo más común el uso de clases o modelos de mezclas de Gausianas para agrupar las características siguiendo criterios fonéticos como hace Youngjoo Suh en [125] o de

SNR (ver [23]).

ii) La segunda limitación de HEQ que es necesario mencionar es el hecho de que al ecualizar componente a componente **se está desaprovechando la información de la relación entre características** para el proceso de reconocimiento, siendo deseable capturarla. Si el ruido ha producido una rotación del espacio de características sería conveniente deshacerla. Esta limitación ha originado toda una familia de técnicas que capturan relaciones entre coeficientes, ya sea definiendo cuantizaciones vectoriales con diferentes criterios, definiendo clases mediante modelos de mezclas de Gausianas (*GMMs*). Como algoritmos interesantes de cuantizaciones vectoriales hay que mencionar [76], que hace una ecualización seguida de una cuantización vectorial de los coeficientes cepstrales en un espacio 4D añadiendo información temporal, [23], o [125]. Hay muchos algoritmos que definen clases de características para capturar relaciones entre coeficientes usando *GMMs*: [90], [132], [124]. Utilizando el mismo algoritmo no lineal, los modelos de mezclas de Gausianas mejoran los resultados en comparación con la cuantización vectorial debido a la decisión *suave* entre clases (*soft decision*) dada por los GMMs.

PEQ: Ecualización Paramétrica en dos clases

Como alternativa efectiva que ataca las limitaciones de HEQ mencionadas, en este capítulo se propone (ver [42] y [41]) el uso de una forma paramétrica de la transformación de ecualización, que se basa en modelar la densidad de probabilidad de las características cepstrales con una mezcla de dos Gausianas. En condiciones ideales no ruidosas, la voz sigue una distribución que se asemeja mucho a una Gausiana bimodal. Por esta razón Sirko Molau propone en [77] el uso de dos histogramas acumulativos independientes para voz y silencio, para lo que separa las tramas como tales con un detector de actividad de voz. El resultado no es todo lo óptimo que se desearía ya que la discrimación entre voz y silencio es muy agresiva.

Bo Liu propone en [70] el uso de dos histogramas acumulativos Gausianos para definir la *pdf* de cada uno de los coeficientes cepstrales, y resuelve la distinción entre las clases de voz y ruido mediante un factor de pesado de las probabilidades de pertenecer a una u otra clase.

La alternativa propuesta en este trabajo recibe el nombre de *PEQ* y define una transformación de ecualización paramétrica basada en un modelo de mezcla de dos Gausianas. La primera Gausiana se usa para representar las tramas de ruido, y la segunda para modelar las tramas de voz. Para cada una de estas dos clases se define una transformación lineal paramétrica que mapea los espacios de representación limpio y ruidoso del siguiente modo:

$$\hat{x} = \mu_{n,x} + (y - \mu_{n,y}) \left(\frac{\Sigma_{n,x}}{\Sigma_{n,y}} \right)^{\frac{1}{2}}, \quad \text{si } y \text{ es una trama de silencio} \tag{7.2}$$

$$\hat{x} = \mu_{s,x} + (y - \mu_{s,y}) \left(\frac{\Sigma_{s,x}}{\Sigma_{s,y}} \right)^{\frac{1}{2}}, \quad \text{si } y \text{ es una trama de voz} \tag{7.3}$$

Los términos de las ecuaciones 7.2 y7.3, se definen del siguiente modo:

- $\mu_{n,x}$ y $\Sigma_{n,x}$ son la media y varianza de las distribuciones Gausianas de referencia limpia para la clase de ruido.

- $\mu_{s,x}$ y $\Sigma_{s,x}$ son la media y varianza de las distribuciones Gausianas de referencia limpia para la clases de voz.

- $\mu_{n,y}$, $\Sigma_{n,y}$ corresponden a la media y varianzas del modelo Gausiano de ruido para las frases del entorno ruidoso.

- $\mu_{s,y}$ y $\Sigma_{s,y}$ corresponden a la media y varianzas del modelo Gausiano

de voz para las frases del entorno ruidoso.

Con estas definiciones, las medias ruidosas $\mu_{n,y}$ y $\mu_{s,y}$ se transforman en medias limpias $\mu_{n,x}$ y $\mu_{s,x}$, así como las matrices de covarianzas del dominio ruidoso $\Sigma_{n,y}$ y $\Sigma_{s,y}$ se transforman en las matrices de covarianza del entorno limpio $\Sigma_{n,x}$ y $\Sigma_{s,x}$.

Los parámetros de referencia de las Gausianas del modelo limpio se calculan usando los datos de entrenamiento limpio. Sin embargo los parámetros de las dos Gausianas del entorno ruidoso se estiman individualmente para cada frase ecualizada.

Para cada trama que se ecualiza, hay que elegir si pertenece a la clase de voz o de silencio. Una forma de hacerlo es usar un detector de actividad de voz. Eso implicaría una decisión binaria entre ambas transformaciones lineales (transformación de acuerdo a los parámetros de la clase de voz o transformación de acuerdo a los parámetros de la clase de ruido). En el límite entre clases, la decisión binaria implicaría una discontinuidad. Para suavizar esta discontinuidad en vez de un *VAD* hemos utilizado una decisión *suave* basada en incluir las probabilidades condicionales de que cada trama sea de voz o de silencio siguiendo la expresión de la ecuación 7.4. La figura 7.2 muestra un esquema del proceso.

$$\hat{x} = P(n|y)(\mu_{n,x} + (y - \mu_{n,y}) \left(\frac{\Sigma_{n,x}}{\Sigma_{n,y}}\right)^{\frac{1}{2}}) + P(s|y)(\mu_{s,x} + (y - \mu_{s,y}) \left(\frac{\Sigma_{s,x}}{\Sigma_{s,y}}\right)^{\frac{1}{2}}) \tag{7.4}$$

En esta ecuación 7.4, los términos $P(n|y)$ y $P(s|y)$ son las probabilidades a posteriori de que la trama pertenezca a la clase de silencio y a la clase de voz respectivamente. Para obtenerlas se ha utilizado un clasificador Gausiano de dos clases, usando el logaritmo del término de la energía (el coeficiente cepstral *Co*) como criterio de clasificación. Inicialmente las tramas de la frase cuyo valor de *Co* es menor que la media de la frase se clasifican como tramas de ruido, y aquellas cuyo valor de *Co* es mayor que la media se clasifican como tramas de voz. Usando esta clasificación inicial

se estiman los valores iniciales de las medias, varianzas y probabilidades a priori de las clases y con el algoritmo *EM* (*Expected Maximization*) se itera hasta la convergencia. Esta clasificación origina los valores de $P(n|y)$ y $P(s|y)$, así como las matrices de media y covarianza de las clases de silencio y voz de la frase en proceso de ecualización: $\mu_{n,y}$, $\mu_{s,y}$, $\Sigma_{n,y}$ y $\Sigma_{s,y}$. Si denominamos n a las tramas de ruido presentes en la frase x , y s a las tramas de voz presentes en la frase x que se ecualiza, los parámetros se definen de manera iterativa mediante *EM* como vemos en las expresiones 7.5:

$$\begin{aligned} n_n &= \sum_x p(n|x) \cdot x \\ n_s &= \sum_x p(s|x) \cdot x \\ \mu_n &= \frac{1}{n_n} \cdot \sum_x p(n|x) \\ \mu_s &= \frac{1}{n_s} \cdot \sum_x p(s|x) \\ \bar{\Sigma}_n &= \frac{1}{n_n} \cdot \sum_x p(n|x) \cdot (x-\mu_n) \cdot (x-\mu_n)^T \\ \bar{\Sigma}_s &= \frac{1}{n_s} \cdot \sum_x p(s|x) \cdot (x-\mu_s) \cdot (x-\mu_s)^T \end{aligned} \tag{7.5}$$

siendo las probabilidades a posteriori utilizadas calculadas con la regla de Bayes:

$$\begin{aligned} p(n|x) &= \frac{p(n) \cdot (N(x, \mu_n, \bar{\Sigma}_n))}{p(n) \cdot (N(x, \mu_n, \bar{\Sigma}_n)) + p(s) \cdot (N(x, \mu_s, \bar{\Sigma}_s))} \\ p(s|x) &= \frac{p(s) \cdot (N(x, \mu_s, \bar{\Sigma}_s))}{p(n) \cdot (N(x, \mu_n, \bar{\Sigma}_n)) + p(s) \cdot (N(x, \mu_s, \bar{\Sigma}_s))} \end{aligned} \tag{7.6}$$

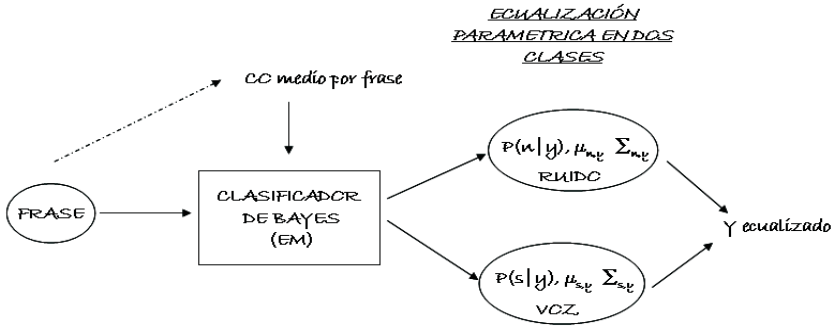


Figura 7.2. Proceso de parametrización PEQ

En la figura 7.3, las subfiguras (a) y (b) muestran el modelo paramétrico de dos Gaussianas para las funciones densidad de probabilidad de los coeficientes cepstrales *C0* y *C1*, superpuesto a los histogramas de las tramas de voz y de silencio para un conjunto de frases limpias. Las subfiguras (c) y (d) muestran los mismos modelos e histogramas para un conjunto de de frases ruidosas. Se puede apreciar la conveniencia de las Gausianas bimodales para aproximar los histogramas de las dos clases especialmente en el caso del coeficiente *C0*, y cómo la distancia entre ambos modos se reduce en condiciones ruidosas.

La figura 7.4 representa la transformación definida para una frase ruidosa según la ecualización paramétrica en dos clases *PEQ*. En las gráficas se representa también la transformación HEQ. Debido a que la ecualización paramétrica se basa en las probabilidades de clase $P(n|y)$ y $P(s|y)$ que dependen del nivel de coeficiente cepstral C0, la ecuación 7.4 define un $\hat{x}$ en función de y que será una transformación no lineal que tenderá al mapeo lineal dado por:

i) la ecuación 7.3 cuando se cumpla que $P(s|y) \gg P(n|y)$

ii) la ecuación 7.2 cuando se cumpla que $P(n|y) \gg P(s|y)$

Para el caso del coeficiente *C1*, debido a que las probabilidades $P(n|y)$

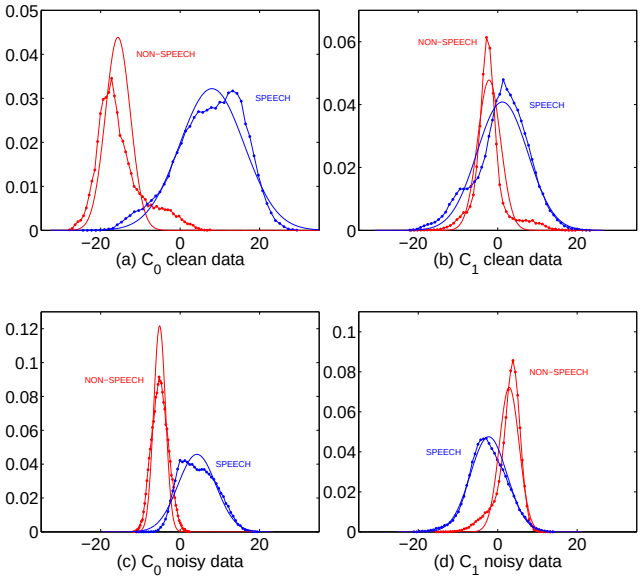


Figura 7.3. Histograma versus modelo paramétrico de dos Gausianas

y $P(s|y)$ dependen del valor de *C0*, la relación entre los datos limpios y sucios no es una función monótona. Un valor de *C1* ruidoso puede producir diferentes valores de *C1* ecualizado dependiendo del valor de *CO* para dicha trama. Este comportamiento concuerda con la distribución de probabilidad del coeficiente *C1*. Se observa también una tendencia no lineal como en el caso de *C0*, siendo la transformación HEQ que se ve superpuesta la función monótona no lineal que más se aproxima a la ecualización paramétrica propuesta.

7.2. Experimentos y Resultados

7.2.1. HEQ versus Ecualización Paramétrica en dos clases, PEQ

Las tablas 7.1, 7.2 y 7.3 muestran los resultados comparativos de la implementación de *PEQ* y *HEQ*. Para las tres bases de datos se produce una

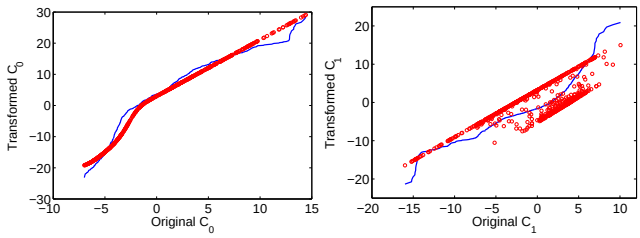


Figura 7.4. Transformación dada por PEQ versus HEQ

	Set A	**Set B**	**Set C**	**Valor medio**	**Mejora Relativa**
BASELINE	46,8	51,1	41,3	47,42	0
REFCLEAN	19,33	17,3	18,97	18,41	61,1 %
PEQ	18,44	16,52	20,32	**18,05**	**62 %**
AFE	14,14	15,2	18,25	15,39	67,5 %

Tabla 7.1. WER en AURORA2 para PEQ

mejora en los resultados al usar *PEQ*. La mejora relativa a *HEQ* más alta se obtiene para la base de datos HIWIRE, siendo de un 20 % la disminución de la tasa de error de palabra. En el caso de AURORA4 dicha mejora relativa es del 14,5 %. Para AURORA2, la mejora respecto al uso de *HEQ* es menor: solamente de un 0,9 %. Analizando en detalle el comportamiento de esa base de datos se puede ver que para el Set C en el que el ruido es una mezcla de aditivo y convolucional (frente a los *Sets A* y *B* que sólo tienen ruido aditivo), *HEQ* funciona mejor que *PEQ*.

7.2.2. Ecualización progresiva de coeficientes

Las tablas 7.4, 7.5 y 7.6 muestran el resultado de ecualizar un número progresivamente mayor de coeficientes cepstrales *Ci* para las diferentes bases de datos. El objetivo de este experimento es estudiar cuáles son los coeficientes cepstrales que más conviene ecualizar y hasta qué punto es

	Test 01-07	**Test 08-14**	**Valor Medio**	**Mejora Relativa**
BASELINE	39,63	49,87	44,75	0
REFCLEAN	32,35	42,00	37,18	16,9 %
PEQ	27,93	33,49	**30,71**	**31,4 %**
AFE	27,65	36,07	31,86	28,8 %

Tabla 7.2. WER en AURORA4 para PEQ

	Clean	**Low N.**	**Mid N.**	**High N.**	**Valor Medio**	**Mejora Relat.**
BASELINE	8,52	52,05	74,43	97,49	58,12	0
REFCLEAN	14,02	47,16	61,84	88,70	52,93	9,0 %
PEQ	10,55	27,37	43,10	84,14	**41,29**	**29,0 %**
AFE	12,35	28,78	42,47	85,15	42,19	27,4

Tabla 7.3. WER en HIWIRE para PEQ

conveniente ecualizarlos todos, teniendo en cuenta que la ecualización introduce una cierta distorsión al ser una transformación basada en aproximaciones de densidad de probabilidad, y bajo una serie de hipótesis que hacen el proceso manejable desde el punto de vista práctico. Al igual que se hizo en el capítulo 6 con *HEQ*, se han seleccionado las ecualizaciones progresivas que mejores resultados dan y se han comparado estos resultados con los obtenidos al ecualizar todos los coeficientes. Dichas comparaciones aparecen en las tablas 7.7 para AURORA2, 7.8 para AURORA4 y 7.9 para HIWIRE.

Las conclusiones que se pueden obtener del análisis de las tablas expuestas son las siguientes. La ecualización progresiva de los coeficientes cepstrales sigue siendo beneficiosa obviamente en términos de coste computacional (son menos los coeficientes que se transforman), y en resultados de reconocimiento. Al igual que ocurría con *HEQ*, en el caso de *PEQ*, la principal mejora del sistema ocurre al ecualizar los primeros coeficientes.

	Set A	**Set B**	**Set C**	**Valor medio**	**Mejora Relativa**
BASELINE	46,80	51,10	41,30	47,42	0
PEQ, C0	27,55	23,15	28,55	26,00	45,2 %
PEQ, C0-C1	22,53	18,97	24,27	21,46	54,7 %
PEQ, C0-C2	19,76	16,99	21,54	19,01	59,9 %
PEQ, C0-C3	19,40	17,58	20,98	18,99	60,0 %
PEQ, C0-C4	18,29	16,17	20,15	17,81	62,4 %
PEQ, C0-C5	18,14	16,23	19,76	17,70	62,7 %
PEQ, C0-C6	18,10	16,21	19,78	**17,68**	**62,71 %**
PEQ, C0-C7	18,35	16,60	20,02	17,99	62,1 %
PEQ, C0-C8	18,30	16,57	20,01	17,97	62,1 %
PEQ, C0-C9	18,20	16,35	19,85	17,79	62,5 %
PEQ, C0-C10	18,49	16,50	20,15	18,03	62 %
PEQ, C0-C11	18,33	16,35	20,33	17,94	62,3 %
PEQ, C0-C12	18,44	16,52	20,32	18,05	62 %
MULTITRAIN	12,33	13,67	17,73	13,95	70,6 %
AFE	14,14	15,20	18,25	15,39	67,50 %

Tabla 7.4. WER para AURORA2. Ecualización progresiva con PEQ

	Test 01-07	**Test 08-14**	**Valor Medio**	**Mejora Relativa**
BASELINE	39,63	49,87	44,75	0
PEQ, C0	34,10	43,23	38,66	13,6 %
PEQ, C0-C1	30,97	37,04	34,01	24 %
PEQ, C0-C2	28,97	34,83	31,90	28,7 %
PEQ, C0-C3	28,91	34,27	31,59	29,4 %
PEQ, C0-C4	28,55	34,36	31,46	29,7 %
PEQ, C0-C5	28,53	33,69	31,11	30,5 %
PEQ, C0-C6	28,25	33,50	30,87	31 %
PEQ, C0-C7	28,14	34,28	31,21	30,3 %
PEQ, C0-C8	28,20	33,82	31,01	30,7 %
PEQ, C0-C9	28,05	33,98	31,02	30,7 %
PEQ, C0-C10	27,95	34,06	31	30,7 %
PEQ, C0-C11	27,79	33,81	30,80	31,2 %
PEQ, C0-C12	27,93	33,49	30,71	31,4 %
MULTITRAIN	25,36	31,30	28,33	36,7 %
AFE	27,65	36,07	31,86	28,8 %

Tabla 7.5. WER para AURORA4. Ecualización progresiva con PEQ

Hay que destacar dos novedades. Para el caso de AURORA4, la mejora relativa que proporciona PEQ respecto a un *front-end* sin algoritmo de normalización es sólo una décima menor que la que proporciona PEQ con ecualización progresiva de los *Cis* como se ve en la tabla 7.8. Este es el límite de una tendencia que apuntan las tablas 7.7 y 7.9: la ecualización progresiva mejora la no progresiva, pero la mejora al comparar ambas es menor para *PEQ* que en el caso de *HEQ*. De nuevo esto se debe a que la ecualización paramétrica en clases elimina en gran parte las limitaciones de HEQ introduciendo por tanto menos distorsión en el proceso de reconocimiento.

7.2.3. Ecualización Paramétrica de dos clases *on-line*

Es deseable estudiar el potencial de una versión de *PEQ* en tiempo real que permita su uso en aplicaciones de voz que no permitan el retardo

	Clean	**Low N.**	**Mid N.**	**High N.**	**Valor Medio**	**Mejora Relat.**
BASELINE	8,52	52,05	74,43	97,49	58,12	0
PEQ, C0	9,20	34,76	54,98	92,07	47,75	17,8
PEQ, C0-C1	9,90	29,30	45,59	85,01	42,45	27
PEQ, C0-C2	9,85	26,59	42,01	80,93	39,85	31,4
PEQ, C0-C3	9,67	24,83	38,57	78,07	37,79	35,0
PEQ, C0-C4	9,71	23,17	36,27	76,70	36,46	37,3
PEQ, C0-C5	9,92	23,08	35,93	77,15	36,52	37,2
PEQ, C0-C6	10,25	23,56	36,80	78,31	37,23	36,0
PEQ, C0-C7	10,21	24,46	37,52	79,87	38,01	34,6
PEQ, C0-C8	10,13	25,03	38,96	80,47	38,65	33,5
PEQ, C0-C9	10,53	25,51	40,45	81,75	39,56	31,9
PEQ, C0-C10	10,48	26,02	41,24	82,94	40,17	30,9
PEQ, C0-C11	10,54	26,74	42,04	83,18	40,62	30,1
PEQ, C0-C12	10,55	27,37	43,10	84,14	41,29	29
AFE	12,35	28,78	42,47	85,15	42,19	27,4

Tabla 7.6. WER para HIWIRE. Ecualización progresiva con PEQ

de la longitud de una frase para definir la transformación de ecualización, o aplicaciones en las que la longitud de la frase sea demasiado pequeña para obtener estadísticas fiables. Para *HEQ* se definió (ver sección 6.2.4 del capítulo 6) una implementación *on-line* de la ecualización consistente en utilizar una CDF global de los datos de *test*, que se va actualizando con las estadísticas de cada nueva frase de *test* que son calculadas después de que ésta haya sido ya ecualizada. Cada frase modifica la CDF global con un cierto peso α. En el caso de PEQ, la versión *on-line* de la ecualización consta de los siguientes pasos:

i) Se definen una media y varianza globales de los datos de *test*, para las clases de voz y silencio. Para inicializarlas se les asignan los valores de la media y varianza de las distribución de referencia para las clases

	Set A	**Set B**	**Set C**	**Valor medio**	**Mejora Relativa**
PEQ, C óptimo	18,14 *(C0-C5)*	16,17 *(C0-C4)*	19,76 *(C0-C5)*	17,7	62,7 %
PEQ, C0-C12	18,44	16,52	20,32	18,05	62 %

Tabla 7.7. Mejores *WER*s en AURORA2. Ecualización progresiva con PEQ

	Test 01-07	**Test 08-14**	**Valor Medio**	**Mejora Relativa**
PEQ, C óptimo	27,79 *(C0-C11)*	33,50 *(C0-C6)*	30,65	31,5 %
PEQ, C0-C12	27,93	33,49	30,71	31,4 %

Tabla 7.8. Mejores *WER*s en AURORA4. Ecualización progresiva con PEQ

de voz y silencio:

$$\begin{aligned} \mu_{n,test\_global} &= \mu_{n,ref} \\ \mu_{s,test\_global} &= \mu_{s,ref} \\ \Sigma_{n,test\_global} &= \Sigma_{n,ref} \\ \Sigma_{s,test\_global} &= \Sigma_{s,ref} \end{aligned} \tag{7.7}$$

ii) Para ecualizar paramétricamente una trama y, se utilizan la media y varianza globales de los datos de *test* acumuladas de frases anteriores de modo que cada trama de la frase ecualizada $\hat{x}$ tendrá la expresión:

	Clean	**Low N.**	**Mid. Noise**	**High N.**	**Valor medio**	**Mejora Relat.**
PEQ, C óptimo	8,52 *(Baseline)*	23,08 *(C0-C5)*	35,93 *(C0-C5)*	76,70 *(C0-C4)*	36,06	38,0 %
PEQ, C0-C12	10,55	27,37	43,10	84,14	41,29	29 %

Tabla 7.9. Mejores *WER*s en HIWIRE. Ecualización progresiva con PEQ

$$\begin{aligned}\hat{x} &= \mu_{s,ref} + (y - \mu_{s,test\_global}) \left(\frac{\Sigma_{s,ref}}{\Sigma_{s,test\_global}} \right)^{\frac{1}{2}}, \text{para } y \text{ trama de voz} \\ \hat{x} &= \mu_{n,ref} + (y - \mu_{n,test\_global}) \left(\frac{\Sigma_{n,ref}}{\Sigma_{n,test\_global}} \right)^{\frac{1}{2}}, \text{para } y \text{ trama de silencio}\end{aligned} \tag{7.8}$$

iii) Una vez ecualizada la frase y, los parámetros globales de las frases de *test* se actualizan con los parámetros de la frase y ($\mu_{n,y}$, $\mu_{s,y}$, $\Sigma_{n,y}$ y $\Sigma_{s,y}$), convenientemente pesados con el parámetro α:

$$\begin{aligned}\mu_{n,test\_global} &= \alpha \cdot \mu_{n,test\_global} + (1-\alpha) \cdot \mu_{n,y} \\ \mu_{s,test\_global} &= \alpha \cdot \mu_{s,test\_global} + (1-\alpha) \cdot \mu_{s,y} \\ \Sigma_{n,test\_global} &= \alpha \cdot \Sigma_{n,test\_global} + (1-\alpha) \cdot \Sigma_{n,y} \\ \Sigma_{s,test\_global} &= \alpha \cdot \Sigma_{s,test\_global} + (1-\alpha) \cdot \Sigma_{s,y}\end{aligned} \tag{7.9}$$

Las tablas 7.10, 7.11 y 7.12 muestran los resultados de los experimentos con distintos valores de α para las 3 bases de datos de referencia. La versión *on-line* da resultados bastante aceptables produciendo mejoras relativas de hasta el 61,2 % en el caso de AURORA2, hasta el 21 % en el caso

de *AURORA4* y del 25,5 % en el caso de *HIWIRE*. Este último resultado es especialmente interesante, ya que la versión on-line de *HEQ* para *HIWIRE* que vimos en el capítulo 6 daba muy malos resultados, siendo preferible utilizar un *front-end* sin algoritmo de robustecimiento ninguno en el caso de tener requerimientos de tiempo real, o frases demasiado cortas para estimar su *pdf* de modo fiable. PEQ on-line sin embargo, funciona de manera muy aceptable con esta base de datos, debido a que las CDFs paramétricas que calcula para los datos de *test* son más fiables para representar las estadísticas de voz de la frase que se ecualiza. Respecto al valor del coeficiente de memoria α para definir la CDF, los resultados muestran que es un parámetro que conviene ajustar para cada base de datos. En el caso de *AURORA2* α=0,6 es el valor que mejor funciona, siendo α=0,4 el óptimo para AURORA4 y α=0,2 el conveniente para HIWIRE.

	Set A	**Set B**	**Set C**	**Valor medio**	**Mejora Relativa**
BASELINE	46,8	51,1	41,3	47,42	0
PEQ,i2r, α =0,8	19,01	17,82	20,55	18,84	60,3 %
PEQ,i2r, α =0,6	18,69	17,4	19,92	18,42	**61,2 %**
PEQ,i2r, α =0,4	18,93	17,77	20	18,68	61 %
PEQ, i2r, α =0,2	19,38	18,24	20,23	19,1	59,7 %
PEQ	18,44	16,52	20,32	18,05	62 %
MULTITRAIN	12,33	13,67	17,73	13,95	70,6 %
AFE	14,14	15,2	18,25	15,39	67,5 %

Tabla 7.10. WER en AURORA2. PEQ *on-line* con diferentes α .

7.2.4. PEQ frente a diferentes tipos y niveles de ruido

Análisis del comportamiento para diferentes niveles de ruido

Las figuras 7.5 y 7.6 hacen un análisis comparativo del comportamiento específico de *HEQ* y *PEQ* para los diferentes niveles de ruido en AURO-

	Test 01-07	**Test 08-14**	**Valor Medio**	**Mejora Relativa**
BASELINE	39,63	49,87	44,75	0
PEQ, i2r, α =0,8	33,87	40,66	37,24	16,8 %
PEQ, i2r, α =0,6	32,60	39,75	36,18	19,2 %
PEQ, i2r, α =0,4	32,26	38,43	35,35	**21 %**
PEQ, i2r, α =0,2	32,21	39,44	35,82	20 %
PEQ, No seg	27,93	33,49	30,71	31,4 %
MULTITRAIN	25,36	31,30	28,33	36,7 %
AFE	27,65	36,07	31,86	28,8 %

Tabla 7.11. WER en AURORA4. PEQ *on-line* con diferentes α .

	Clean	**Low N.**	**Mid N.**	**High Noise**	**Valor medio**	**Mejora Relativa**
BASELINE	8,52	52,05	74,43	97,49	58,12	0
PEQ, i2r, α =0,8	11,52	38,79	54,91	86,72	47,98	17,4 %
PEQ, i2r, α =0,6	10,41	35,99	52,60	85,81	46,20	20,5 %
PEQ, i2r, α =0,4	10	33,84	49,99	84,89	44,68	23,1 %
PEQ, i2r, α =0,2	9,77	31,53	47,79	84,02	43,28	**25,5 %**
PEQ	10,55	27,37	43,10	84,14	41,29	29 %
AFE	12,35	28,78	42,47	85,15	42,19	27,4 %

Tabla 7.12. WER en HIWIRE. PEQ *on-line* con diferentes α .

RA2 y HIWIRE. Para el caso de señales limpias ambos métodos aumentan la tasa de error de reconocimiento al aplicarlos. Introducen una distorsión innecesaria, siendo *HEQ* el que más distorsiona dada su menor exactitud en el cálculo de las *CDFs* y con ello en el cálculo de la transformación de ecualización. Cuando la tasa de ruido va aumentando, para *SNRs* de 15, 10 y 5 dB, es mucha la mejora introducida por *HEQ* y *PEQ* dando este último las tasas de error de reconocimiento más baja. Se podría decir que 10 dB es el punto óptimo de funcionamiento de estos métodos de normalización. Para señales de muy poca *SNR* como pueden ser 5dB, *HEQ* y *PEQ* mejoran

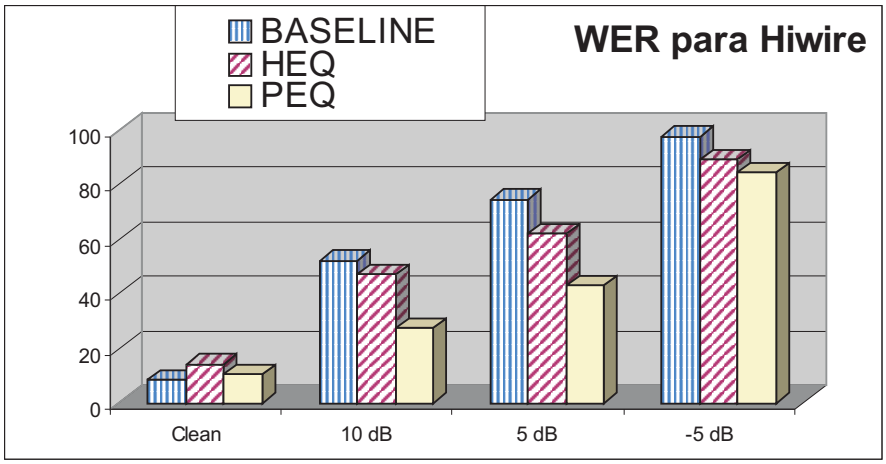


Figura 7.5. WER para HIWIRE,versus SNRs y Ecualizaciones

el BASELINE pero mucho más levemente. Es interesante el hecho de que para AURORORA2 en los casos de 20dB , 15dB y 10 dB, PEQ produce una tasa de error levemente superior a HEQ. Esta tendencia se invierte para 5, 10 y -5 dB, en los que *PEQ* es bastante más eficaz que *HEQ*. Esto se puede deber a que la filosofía de *PEQ* es clasificar las tramas como pertenecientes a la clase de voz o a la clase de silencio basándose en el coeficiente *C0*. Esta clasificación es beneficiosa para frases con una SNR no demasiado alta. Si la SNR es alta se distorsionan más los resultados.

Análisis del comportamiento para ruidos aditivos y convolucionales

Las figuras 7.7 y 7.8 muestran el comportamiento de *PEQ* y *HEQ* al actuar sobre ruidos aditivos y al actuar sobre ruidos aditivos mezclados con ruidos convolucionales. No se pueden extraer conclusiones respecto a la influencia del tipo de ruido en el funcionamiento del método. Para el caso de AURORA2, *PEQ* funciona ligeramente peor que *HEQ* en presencia de ruido convolucional. Para AURORA4 sin embargo, *PEQ* da mejores

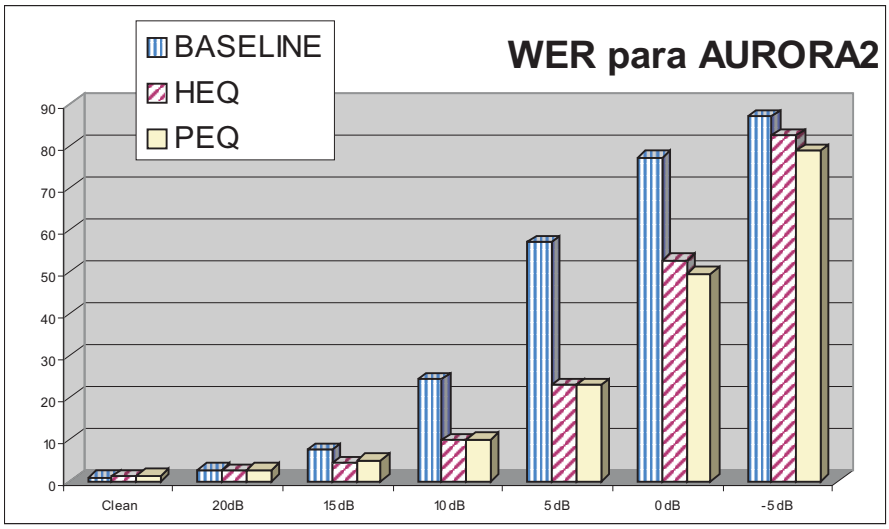


Figura 7.6. WER para AURORA2,versus SNRs y Ecualizaciones

resultados en cualquier caso.

Análisis del comportamiento para distintos tipos de ruido

La figura 7.9 muestra la comparativa de los dos métodos de normalización estudiados por tipos de ruido para AURORA2. En el caso de los datos limpios, se corrobora una vez más que la normalización aumenta el error de reconocimiento como era de esperar por la distorsión introducida. En el resto de casos *PEQ* da mejores resultados manteniéndose la mejora relativa a HEQ prácticamente constante. En el ruido *Babble*, la distancia relativa entre la tasa de error del método paramétrico y el no paramétrico aumenta siendo éste un ruido difícil de combatir por su similitud con la voz.

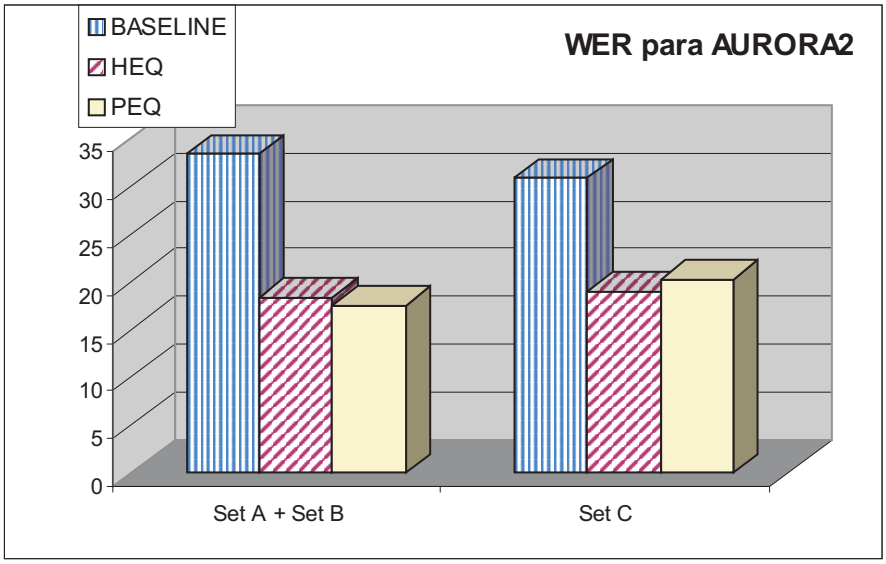


Figura 7.7. WER para AURORA2,ruidos aditivos y convolucionales

7.2.5. PEQ sobre MLLR

Se ha evaluado la conveniencia de *PEQ* como algoritmo de robustecimiento frente al ruido en sistemas que emplean además adaptación al locutor. La tabla 7.13 muestra los resultados de aplicar el método de adaptación supervisada de modelos *MLLR* [40] descrito en detalle en el capítulo 3, directamente sobre los coeficientes cepstrales, comparados con los resultados de aplicar *MLLR* sobre los coeficientes ecualizados con *PEQ*. Se hacen pruebas para 10, 20 y 50 frases de adaptación. Para los *tests* limpios *PEQ+MLLR* da errores de reconocimiento más altos que *MLLR*. Esto es razonable debido a la distorsión que *PEQ* introduce, sin contrapartida de beneficio alguno cuando no hay ruido que compensar como es en el caso de los *tests* limpios. En cuanto la SNR empeora, PEQ+MLLR produce mejoras relativas que superan a las de PEQ en un 20 % de media. Esto se debe a que PEQ mejora los alineamientos iniciales sobre los que MLLR trabaja.

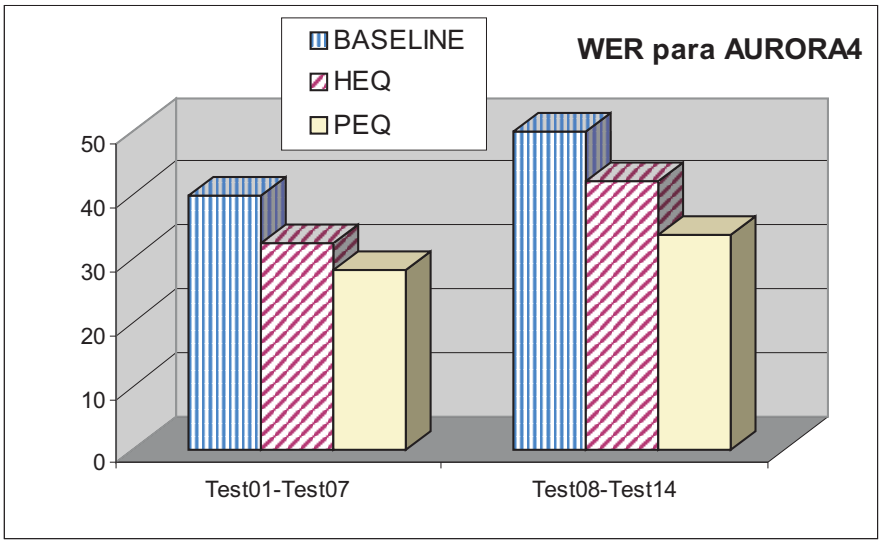


Figura 7.8. WER para AURORA4, ruidos aditivos y convolucionales

7.3. Conclusiones

Este capítulo presenta *PEQ*, una versión paramétrica de la Ecualización de Histogramas en la que se diferencian dos clases, modeladas mediante sendas Gausianas, de tramas : tramas de voz y de silencio. Las conclusiones que se pueden extraer del trabajo realizado son las que siguen:

- La ventaja de usar una expresión paramétrica de las funciones de densidad de probabilidad es que se consiguen estadísticas globales de la voz más fieles a los datos que son difíciles de obtener cuando las frases no son demasiado largas.

- El hecho de definir las clases de voz y silencio con el criterio del coeficiente *C0*, captura la dependencia entre dicho coeficiente y los demás, caracterizando así esa correlación entre el *C0* y el resto de *Cis*. Estas ventajas se materializan en los buenos resultados del método

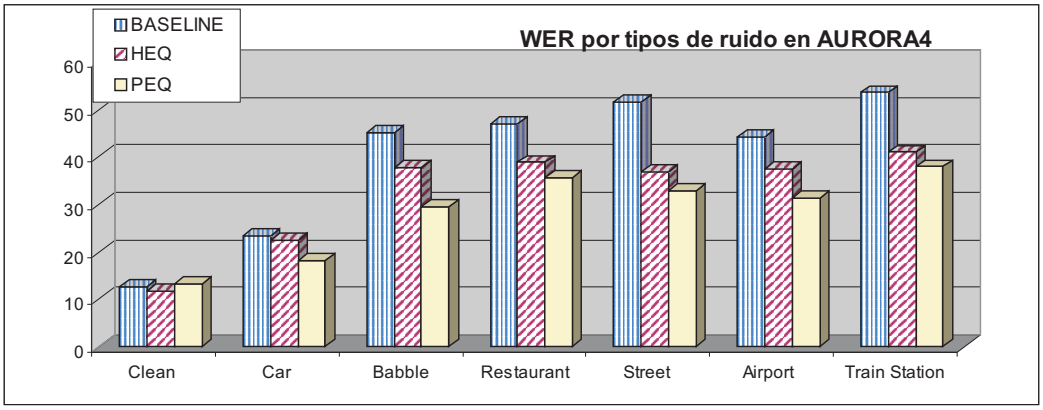


Figura 7.9. WER para AURORA4, diferentes tipos de ruidos

en las pruebas realizadas. Su rendimiento frente a HEQ es bastante mejor para AURORA4 y para HIWIRE, siendo la mejora en AURORA2 menor, aunque considerable.

- Hay que mencionar que la versión on-line de PEQ produce unos resultados bastante interesantes, mucho mejores que en el caso de HEQ on-line (en el caso de HIWIRE era preferible no hacer ningún tipo de ecualización antes que hacer *HEQ on-line*). De nuevo esto se debe a la mejor caracterización de los parámetros y sus estadísticas, que hace que funcione el utilizar la estadística global acumulada de las frases anteriores de *test* para ecualizar cada frase de *test*, evitando así el retardo que supone calcularle su estadística a cada frase en el momento de ecualizarla.

PEQ tiene una serie de caminos abiertos cuyo estudio es deseable, entre los cuales podemos mencionar los siguiente:

- El método sigue utilizando matrices de covarianza diagonales al igual que HEQ. La clasificación de las tramas en base a *C0* captura su corre-

	Clean	Low N.	Mid N.	High N.	Valor medio	Mejora Relativa
BASELINE	8,52	52,05	74,43	97,49	58,12	0
MLLR,10 frases	4,56	34,80	60,15	94,92	48,61	16,4 %
PEQ+MLLR,10 frases	7,60	21,38	35,11	80,21	36,07	37,9 %
MLLR,20 frases	3,78	31,21	56,71	94,34	46,51	20 %
PEQ+MLLR,20 frases	6,32	18,62	31,36	78,86	33,79	41,9 %
MLLR,50 frases	2,72	26,85	51,94	93,19	43,67	24,9 %
PEQ+MLLR,50 frases	4,83	14,17	26,59	76,94	30,63	47,3 %
PEQ	10,55	27,37	43,10	84,14	41,29	29 %

Tabla 7.13. WER para HIWIRE. Adaptación: MLLR versus PEQ+MLLR

lación con el resto de *Cis*. Sería deseable ampliar el estudio de las correlaciones entre características *Ci* que pertenezcan a la misma clase.

- La correlación temporal entre los coeficientes de tramas consecutivas aporta una información valiosa para el robustecimiento que sería interesante capturar.
- Cuando los datos de *test* son limpios, al igual que ocurría en el caso de HEQ aunque en menor medida, la tasa de reconocimiento se empeora. Es deseable eliminar este comportamiento.

Capítulo 8

Normalización de las características estáticas y dinámicas

Este capítulo propone el uso de un filtro que introduce información temporal en el conjunto de características de voz utilizadas en el reconocimiento automático. El filtro propuesto añade las autocorrelaciones de los coeficientes cepstrales al conjunto de características empleadas en entrenamiento y *test*. La autocorrelación de los coeficientes cepstrales no depende de las condiciones del entorno, sino exclusivamente de la información que dichos coeficientes representan. Por esta razón normalizar esas autocorrelaciones temporales tiene interés y puede aportar robustez al reconocimiento. En el desarrollo del capítulo la normalización de las autocorrelaciones será analizada en profundidad estudiando el mejor dominio para hacerla y sus efectos en las tres bases de datos con que se trabaja.

8.1. Introducción

La información temporal inter-trama, definida como las variaciones de los coeficiente entre tramas consecutivas, puede contribuir valiosamente al proceso de reconocimiento ya que estas variaciones son características unívocas de los datos de voz. En el capítulo 3 señalábamos que la información temporal añadía información valiosa a la dada por la frecuencia, y su

captura era uno de los objetivos de las parametrizaciones basadas en modelos perceptuales de la audición. Sin embargo, son pocos los algoritmos de parametrización no basados en modelos perceptuales que tienen en cuenta esta información. En el caso de los coeficientes cepstrales en escala Mel , el procesado temporal más básico se hace con las derivadas primera y segunda de los mismos. Otras técnicas para calcular información inter-trama son las que se derivan del filtrado *RASTA* ([56], [129], [93]) ya sea en su versión básica, u optimizado con un análisis *LDA* o *PCA* ([14]).

Al igual que el resto de componentes del vector de características, la información inter-trama es sensible a los desajustes entre el entorno de entrenamiento y el entorno de *test* como pueden ser el ruido de fondo o la distorsión de canal. Estos desajustes, hacen que el reconocimiento automático de voz se degrade cuando las condiciones son adversas. Por esta razón es deseable normalizar también la información temporal al igual que se hace con el resto de parámetros. La Ecualización de Histogramas aplicada a las derivadas temporales de los coeficientes MFCC ha sido estudiada en [86] y [85], analizando la conveniencia de ecualizar las deltas temporales de manera independiente o dependiente del resto de coeficientes cepstrales. La conclusión obtenida es que el proceso óptimo de normalización es usar realimentación: los coeficientes dinámicos se utilizan para ecualizar los coeficientes estáticos, y una vez obtenida la versión ecualizada de los estáticos, se recalculan los coeficientes dinámicos a partir de ellos.

La motivación de este capítulo es proponer una técnica muy simple de captura y normalización de la información temporal mediante un filtro de suavizado de las características de voz que se añade al proceso de Ecualización de Histogramas.

8.2. Filtro de suavizado temporal

Se propone un filtro *ARMA* para restablecer la estructura de las autocorrelaciones temporales de cada uno de los coeficientes cepstrales. El filtro se puede obtener como cascada de dos filtros: un primer filtro blanqueador

que elimina la estructura de correlaciones temporales existente en los datos de entrada, al que le sigue un segundo filtro que restaura las correlaciones temporales deseadas. Dada una secuencia temporal de observaciones de un determinado coeficiente cepstral $x(n)$, los pasos son los siguientes:

i) Se diseña un filtro de blanqueo con el que se obtiene la secuencia incorrelada $u(n)$, usando por ejemplo una aproximación lineal. Si llamamos $A(z)$ al predictor lineal, la secuencia decorrelada en el dominio *Z* será:

$$U(Z) = A(Z) \cdot X(Z) \tag{8.1}$$

ii) La secuencia blanqueada $U(Z)$ se pasa por un segundo filtro, también obtenido como filtro LPC, que es el encargado de restaurar la estructura de correlación deseada:

$$Y(Z) = \frac{U(Z)}{B(Z)} \tag{8.2}$$

El filtro *ARMA* que engloba el proceso de normalización se define como la cascada de los dos anteriores. Los datos normalizados en el dominio *Z* se definen como:

$$Y(Z) = \frac{A(Z)}{B(Z)} \cdot X(Z) \tag{8.3}$$

Los coeficientes del filtro blanqueador $A(Z)$ se derivan de las autocorrelaciones de la frase cuya información temporal se quiere normalizar. Los coeficientes del segundo filtro $\frac{1}{B(Z)}$ se obtienen como estimación de las autocorrelaciones de los datos de referencia es decir, de los datos de entrenamiento o de adaptación si fuese el caso.

8.2.1. Localización del filtro

El hecho de usar Y_s, la versión suavizada temporalmente de Y, implica aplicar una transformación lineal de los coeficientes cepstrales con-

sistente en aplicarles cierto escalado que hace que las correlaciones temporales mantengan las mismas proporciones que en el dominio de referencia. El efecto de este escalado será eliminar la distorsión inter-trama provocada por un entorno de *test* diferente al entorno de entrenamiento. El dominio en el cual se lleva a cabo dicho escalado tiene efecto en las mejoras conseguidas. Es deseable que el filtro se defina en el dominio en el cual la información inter-trama relevante sea más claramente separable de la distorsión que se quiere eliminar. Al aplicar la Ecualización de Histogramas ya sea en su versión no paramétrica *HEQ* o en su versión paramétrica con dos clases *PEQ*, lo que estamos haciendo es mover las características de la voz a un dominio más robusto frente a las peculiaridades del entorno. En el capítulo 6 se vio que hay dos funciones de densidad de probabilidad acumulada *CDF* que han demostrado ser útiles como *CDFs* de referencia para ecualizar: la *CDF* de los datos de entrenamiento y la *CDF* Gausiana. De acuerdo con esto, existen tres escenarios posibles en los que implementar el filtro *ARMA* para normalizar las autocorrelaciones inter-trama de los MFCCs. Las tres posibilidades se muestran en las figuras 8.1, 8.2 y 8.3, en las que se nota la autocorrelación en el dominio de referencia como R_x y la autocorrelación de la frase que se quiere filtrar como R_y:

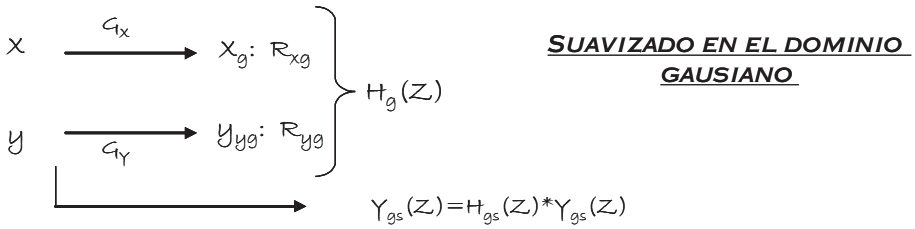


Figura 8.1. Suavizado en el dominio gausiano

- En la figura 8.1 las autocorrelaciones de los datos limpios R_{xg} y las de la frase de *test* R_{yg} se calculan una vez que las características se han ecualizado a una *CDF* de referencia Gausiana. El filtro se define en

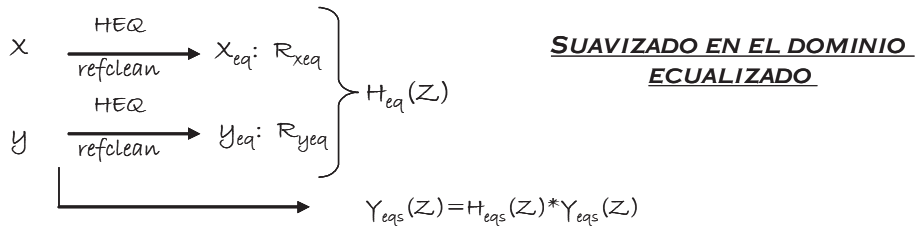


Figura 8.2. Suavizado en el dominio ecualizado

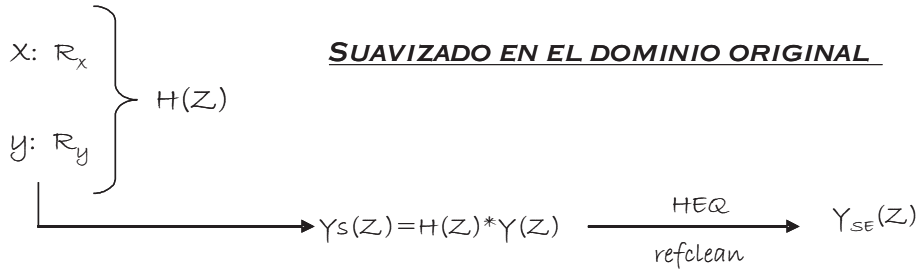


Figura 8.3. Suavizado en el dominio original

el **dominio Gausiano**, y una vez que la información inter-trama ha sido normalizada a través del filtro las características *suavizadas* son ecualizadas a una CDF de referencia limpia, que era la que mejores resultados de reconocimiento daba como se vio en el capítulo 3.

- En la figura 8.2 los coeficientes MFCC son ecualizados utilizando la *CDF* de referencia de los **datos limpios**, y es en ese **dominio ecualizado** en el que se calculan las autocorrelaciones R_{xeq} y R_{yeq} que definen el filtro *ARMA* que se aplica posteriormente.

- El último escenario analizado es aquel en el que el suavizado temporal se hace antes que la ecualización. Las autocorrelaciones R_x y R_y se calculan en el **dominio original**, los MFCCs se suavizan pasándolos

por el filtro temporal definido y una vez hecho esto se ecualizan con la *CDF* de los datos de entrenamiento.

Se han hecho pruebas para analizar los tres escenarios posibles de aplicación del filtro dando los resultados que vemos en la tabla 8.1, en la que hemos llamado *TES* al proceso de suavizado temporal (*temporal smoothing*), *GAUS* a la ecualización usando una *CDF* de referencia Gausiana, y *HEQ* a ecualización usando como *CDF* de referencia la de los datos de entrenamiento.

GAUS+TES+HEQ	**HEQ+TES**	**TES+HEQ**
19,0 %	19,4 %	17,2 %

Tabla 8.1. Comparacion: 3 escenarios de suavizado TES

La tabla 8.1 presenta la mejora relativa respecto al BASELINE que se obtiene al aplicar el suavizado temporal *TES* a los *tests* de AURORA4 sobre los 3 escenarios que se analizan. La posición óptima para el filtro de suavizado temporal es la que se esquematiza en la figura 8.2. Las autocorrelaciones de los coeficientes ya ecualizados han eliminado una parte muy importante de la distorsión, y son más representativas de la información de voz. El filtro definido con dichas autocorrelaciones funciona mejor que el definido en el dominio de ecualización Gausiana y el definido en el dominio sin ecualizar y será el utilizado en los experimentos estudiados en este capítulo.

En la figura 8.4 se muestra la amplitud del coeficiente *C0* para una frase de AURORA4 comparando cuatro procesados diferentes. La curva denominada *test 01* es la amplitud del coeficiente *C0* en el *test* limpio. La llamada *test 07* es la amplitud del *C0* de esa misma frase cuando se ha añadido el ruido aditivo del *test* 07 de las especificaciones de AURORA4 [60]. Las amplitudes etiquetadas como *HEQ* y *HEQ+TES* muestran la normalización mediante ecualización de histogramas, y la suma de ésta más el posterior suavizado temporal respectivamente. Se aprecia que *HEQ* amplía

el rango de amplitudes del coeficiente, que en la curva del *test* 07 había disminuido mucho en relación al *test* limpio original. *TES* suaviza picos de la amplitud que había introducido *HEQ* y acerca más la curva a la de la señal limpia original.

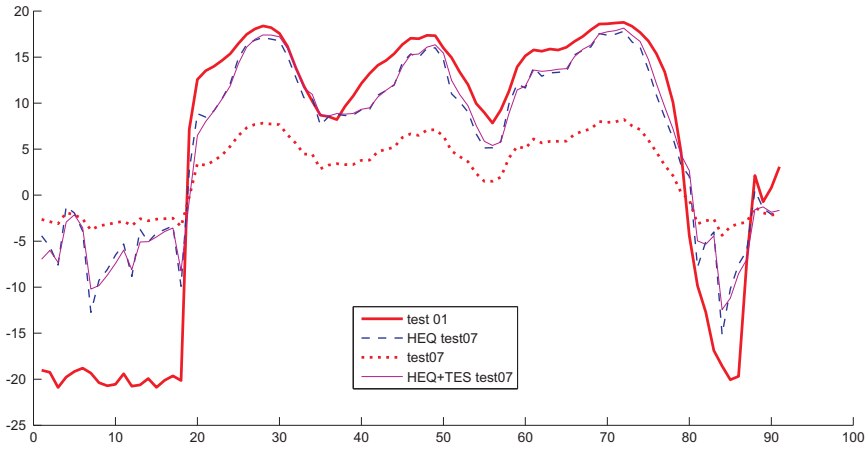


Figura 8.4. Efecto de del filtro TES en la señal

8.3. Experimentos y Resultados

El algoritmo de suavizado temporal propuesto ha sido estudiado para las 3 bases de datos con las que contamos, y su comportamiento ha sido analizado al ser aplicado directamente sobre la parametrización BASELINE, sobre la parametrización HEQ descrita en el capítulo 6 y sobre la parametrización PEQ descrita en el capítulo 7. Las tablas 8.2, 8.3 y 8.4 muestran los resultados. En el caso de AURORA2, la tabla 8.2 muestra que el suavizado temporal (llamado *Rx*) mejora el reconocimiento sea cual sea la parametrización sobre la que se produzca el filtrado temporal. Es especial-

	Set A	**Set B**	**Set C**	**Valor medio**	**Mejora Relativa**
BASELINE	46,8	51,1	41,3	47,42	0
BASELINE + Rx	34,97	29,68	34,37	32,73	31,0 %
HEQ	19,33	17,3	18,97	18,41	61,1 %
HEQ+ Rx	15,01	14,37	16,1	**14,7**	**69,0 %**
PEQ	18,44	16,52	20,32	18,05	62,0 %
PEQ + Rx	16,57	14,67	18,72	16,24	65,6 %
AFE	14,14	15,2	18,25	15,39	67,5 %

Tabla 8.2. WER en AURORA2 para Rx

mente remarcable el resultado obtenido para *HEQ+Rx*, que consigue una mejora de reconocimiento respecto al BASELINE mejor que AFE, mejor que PEQ e incluso mejor que PEQ+Rx.

Para AURORA4 los resultados de añadir Rx son de nuevo bastante positivos, para cualquier parametrización como podemos ver en la tabla 8.3:

	Test 01-07	**Test 08-14**	**Valor Medio**	**Mejora Relativa**
BASELINE	39,63	49,87	44,75	0
BASELINE + Rx	38,54	47,16	42,85	4 %
HEQ	32,35	42	37,18	16,9 %
HEQ + Rx	31,24	40,89	36,06	19,4 %
PEQ	27,93	33,49	30,71	31,4 %
PEQ + Rx	27,79	32,91	**30,35**	**32,2 %**
AFE	27,65	36,07	31,86	28,8 %

Tabla 8.3. WER en AURORA4 para Rx

En el caso de HIWIRE que podemos ver en la tabla 8.4, el suavizado temporal es sin embargo contraproducente para la parametrización *baseline*, empeorando la tasa de reconocimiento en un 2,53 %. Esto se debe a que

	Clean	**Low N.**	**Mid N.**	**High Noise**	**Valor Medio**	**Mejora Relativa**
BASELINE	8,52	52,05	74,43	97,49	58,12	0
BASELINE + Rx	9,28	55,66	76,06	97,36	59,59	-2,53 %
HEQ	14,02	47,16	61,84	88,70	52,93	8,9 %
HEQ + Rx	20,1	47,97	58,47	81,15	51,91	10,7 %
PEQ	10,55	27,37	43,10	84,14	41,29	29,0 %
PEQ + Rx	11,42	27,14	40,44	78,74	**39,43**	**32,2 %**
AFE	12,35	28,78	42,47	85,15	42,19	27,4 %

Tabla 8.4. WER en HIWIRE para Rx

HIWIRE es una base de datos muy ruidosa y difícil de evaluar al mezclarse los efectos del entorno con los de los hablantes no nativos. Las correlaciones de los datos limpios se calculan con TIMIT, que es la base de datos usada para entrenar los experimentos de HIWIRE como describimos en el capítulo 5. Esto hace que realmente exista un importante desajuste en las correlaciones de los datos con parametrización BASELINE, y que el filtro sea más agresivo de la cuenta. Una vez que los datos han sido ecualizados, los beneficios de Rx son de nuevo importantes.

La figura 8.5 da una visión conjunta para las 3 bases de datos de la mejora que el filtro de suavizado temporal hace sobre cada una de las tres parametrizaciones estudiadas: BASELINE, HEQ y PEQ:

- En las dos bases de datos de AURORA, se puede ver la tendencia a aportar mejoras cada vez menores, cuanto mejor es el robustecimiento de los parámetros. Por ejemplo en el caso de AURORA4 la mejora es del 4 % para BASELINE, del 2,5 % para HEQ y del 0,9 % para PEQ. Esto indica que hay una parte del robustecimiento que hacen en común las parametrizaciones y el filtro.

- El caso de HIWIRE es el opuesto. La mejora de Rx se incrementa al incrementarse la robustez de la paremetrización (en BASELINE empeora, en HEQ mejora un 1,8 % y en PEQ mejora un 3 %). Esto se debe

a que la distorsión en HIWIRE es mucha como ya hemos dicho. Los algoritmos de parametrización van dejando la información temporal más definida cuanto más eliminan dicha distorsión

8.4. Conclusiones

Este capítulo propone un método novedoso para añadir información temporal de la señal de voz al vector de parámetros que la caracterizan. Esto es deseable, y es buscado por muchos métodos de parametrización con resultados favorables que lo avalan. El método propuesto hace además una normalización de la información temporal, eliminando así las posibles variaciones de la misma debidas a ruidos o distorsiones de otros tipos entre el entorno de entrenamiento y el de evaluación. Los resultados son bastante interesantes, ya que aumenta la tasa de reconocimiento. Hay que señalar que el aumento es menos espectacular cuanto mejor es la parametrización.

Las siguientes **vías de trabajo** están **abiertas** con respecto a este algoritmo de normalización temporal:

- Sería deseable comparar su rendimiento con el de otras técnicas existentes como pueden ser los análisis LDA o PCA aplicados a tramas consecutivas en el tiempo.

- Es posible que las clases de voz y de ruido con las que trabaja la ecualización paramétrica PEQ, tengan comportamientos diferentes respecto a las correlaciones de componentes. Sería interesante explorar la definición de un filtro de normalización temporal para cada clase

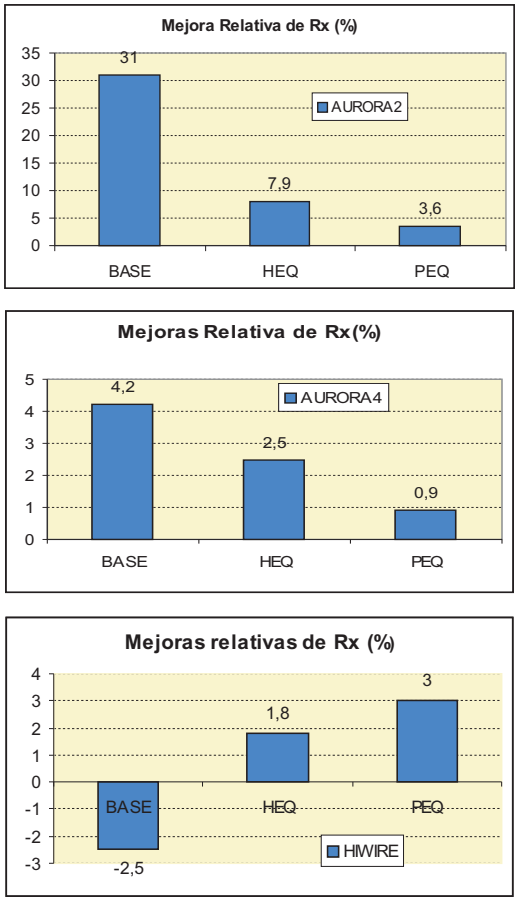


Figura 8.5. Mejoras de Rx

Parte IV

Evaluación

Capítulo 9

Evaluación

El objetivo de este capítulo es evaluar si los algoritmos de parametrización propuestos en esta tesis solventan las carencias señaladas en el capítulo 4, sección 4.1, y cumplen por tanto los objetivos expuestos. El uso de los algoritmos propuestos en esta tesis en el *Advanced front-end* creado para la base de datos HIWIRE dentro del proyecto HIWIRE (VI Programa Marco de la EU) será descrito a continuación, como ejemplo de implementación de los algoritmos de robustecimiento en una plataforma fija. Por último se describen los resultados de su implementación en el *front-end* del reconocedor de Loquendo usado en plataformas móviles para PDAs.

9.1. Análisis de los resultados

El primer objetivo planteado en la tesis, objetivo **O1** del capítulo 4, era el estudio exhaustivo de la Ecualización de Histogramas y sus variantes, para enfrentarla con los requisitos de los algoritmos de robustecimiento expuestos. El estudio exhaustivo se desmenuzaba en un estudio de la distribución de probabilidad acumulada óptima, el estudio de la ecualización parcial de ciertos coeficientes cepstrales y la propuesta de una versión *online* de la ecualización para aplicaciones con requisitos temporales que no permitan la ecualización con retardos del orden de la duración de una frase.

El segundo objetivo planteado en esta tesis, **O2** en el capítulo 4, es la propuesta de una versión paramétrica de la ecualización que elimine las

carencias de la ecualización de histogramas en términos de dependencia de la longitud de la frase (**C1.1** en el capítulo 4), y en términos de dependencia de los porcentajes de voz y ruido (**C1.2** en el capítulo 4). A continuación se analizan los resultados.

9.1.1. Ecualización paramétrica versus Ecualización de Histogramas

La figura 9.1 muestra la comparación entre la mejora en la *WER* respecto a la parametrización BASELINE, conseguida por HEQ y por PEQ para las tres bases de datos. La base de datos menos sensible a las diferencias HEQ-PEQ es AURORA2, que sólo ve incrementada la mejora en un 1 %. En el caso de HIWIRE y AURORA4 esa mejora es muy apreciable. Estos resultados avalan la teoría con la que se ha definido PEQ para hacer frente a las carencias de HEQ:

- Una expresión paramétrica de la distribución de probabilidad necesita menos datos que una distribución construida con histogramas. Esto debilita la dependencia de la longitud de la frase para asegurar una distribución fiel, en el orden de cantidades de datos por frase con el que trabajan AURORA2, AURORA4 y HIWIRE.

- El hecho de definir dos clases, una para voz y otra para silencio, elimina la dependencia del porcentaje de datos de voz y silencio y en la expresión de la transformación.

- Los resultados de PEQ mejoran los de AFE para AURORA4 y HIWIRE.

9.1.2. Ecualización progresiva

Dentro del objetivo **O1** de estudio exhaustivo de la ecualización, se ha estudiado el efecto de distorsión producido al ecualizar cada uno de los coeficientes cepstrales tanto en la Ecualización de Histogramas como en la

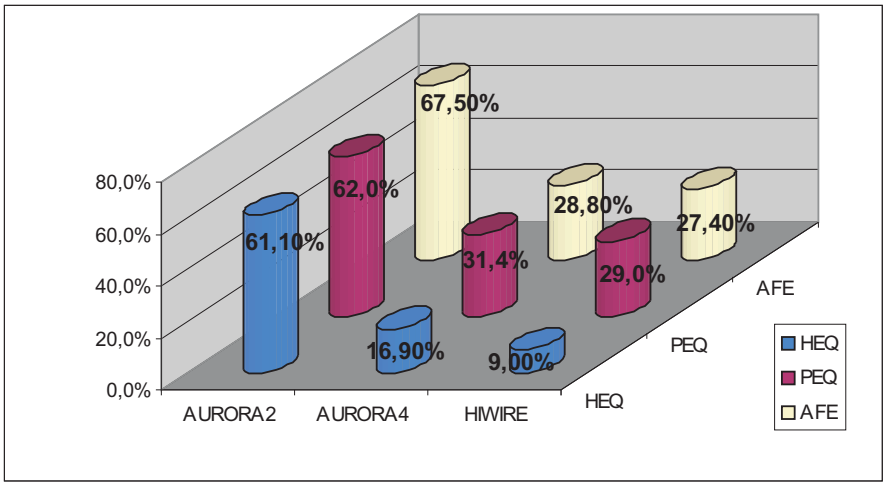


Figura 9.1. WER: PEQ versus HEQ

Ecualización Paramétrica de Histogramas. Los resultados de dicho análisis se pueden ver en las figuras 9.2 y 9.3:

- La mayor parte de la información acústica se concentra en los coeficientes cepstrales de primer orden. La ecualización de los coeficientes de orden alto, introduce distorsiones que empeoran la tasa de reconocimiento.

- Los beneficios de la ecualización progresiva son mayores al usar HEQ que al usar PEQ. Esto se debe a que la parametrización PEQ es más robusta e introduce menos distorsiones en general y en particular en la ecualización de los coeficientes de orden alto. La base de datos HIWIRE es la que más beneficiada se ve en cualquier caso al usar esta ecualización progresiva.

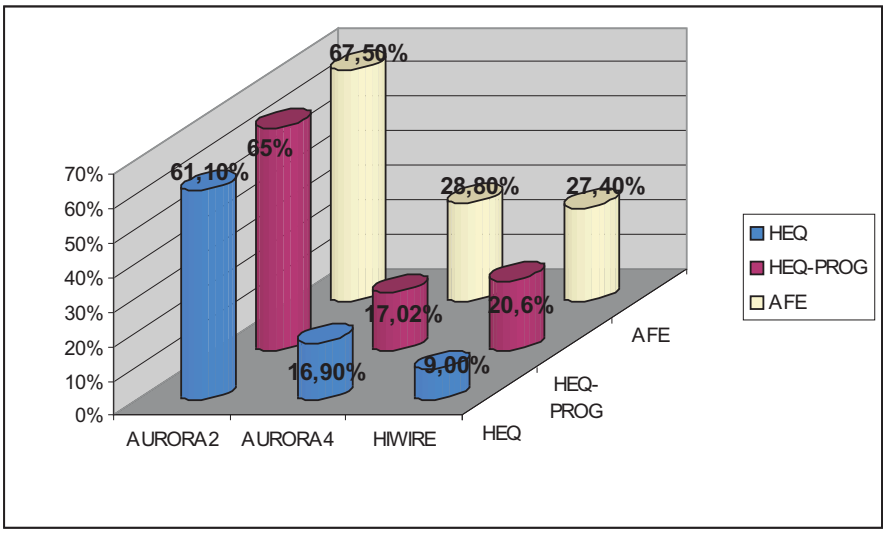


Figura 9.2. Mejora Relativa: Ecualización progresiva en HEQ

9.1.3. Parametrizaciones *on-line*

Las aplicaciones con requisitos de tiempo real, pueden tener exigencias que no permitan usar HEQ o PEQ, dado que es necesario haber la frase entera, antes de poder ecualizarla y entonces reconocerla. Las versiones on-line de HEQ y PEQ propuestas, ecualizan cada trama sin esperar a la recepción de la frase completa. Esto supone una disminución de la tasa de reconocimiento exitoso, que es bastante menor en el caso de PEQ y que puede ser aceptable según el compromiso que sea necesario entre velocidad de reconocimiento y precisión. Las figuras 9.4 y 9.5 la muestran los resultados comparativos:

- Al usar HEQ *on-line* el detrimento de la tasa de reconocimiento en AURORA2 es mínimo. Es algo mayor en AURORA4 y es bastante malo en el caso de HIWIRE, en el que es mejor usar directamente la parametrización BASELINE.

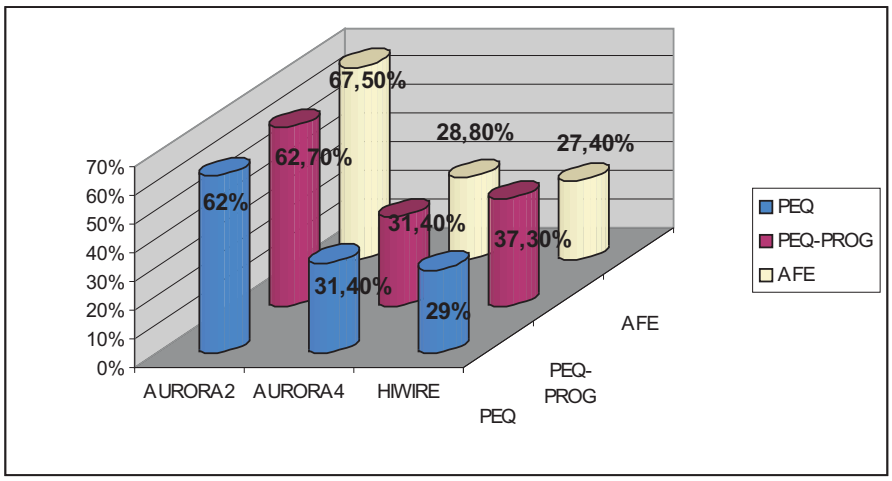


Figura 9.3. Mejora Relativa: Ecualización progresiva en PEQ

- La versión on-line de PEQ tiene mayor robustez, siendo útil para todas las bases de datos. Hay que destacar su robustez para HIWIRE.

9.1.4. Suavizado Temporal: Rx

El objetivo **O3** planteado en esta tesis, era la construcción de un algoritmo que capturase información temporal y la añadiese al conjunto de características de la frase. Era deseable que dicha información temporal estuviese normalizada puesto que también es sensible a la distorsión del entorno. Los resultados obtenidos con el algoritmo propuesto, llamado *TES-Temporal Smoothing Filter*, avalan la consecución de dicho objetivo. La gráfica 9.6 muestra los resultados comparativos de añadirlo a HEQ y a PEQ

Por último, en la figura 9.7 se ofrece una comparativa de los algoritmos estudiados y propuestos en esta tesis para cada una de las tres bases de datos usadas. En la gráfica se sitúan los porcentajes de mejora de reco-

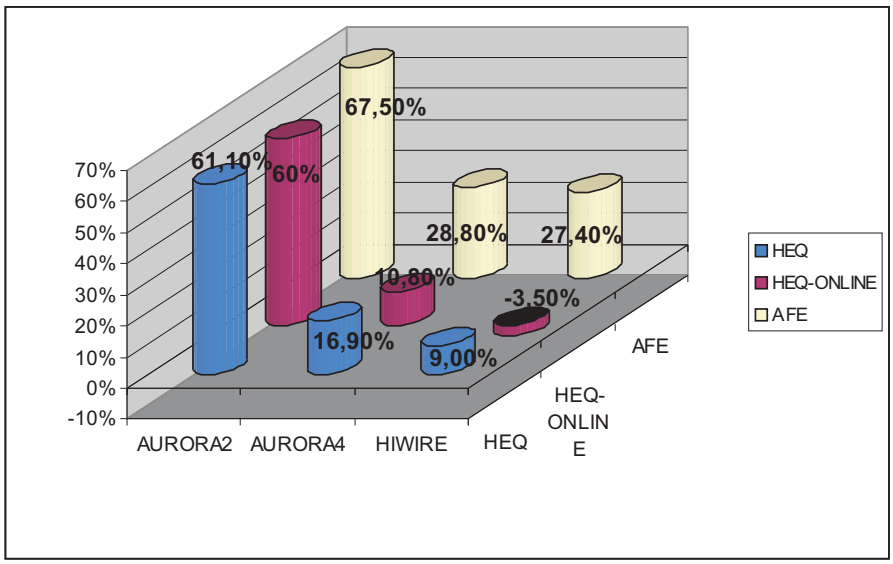


Figura 9.4. Mejora relativa para HEQ versión *online*

nocimiento para cada parametrización y para cada base de datos:

- Podemos ver que AURORA2 es la base de datos más susceptible de robustecerse con las parametrizaciones propuestas, le sigue AURORA4 y en último lugar HIWIRE. Esto está directamente relacionado con la dificultad de las tareas de cada una de ellas.

- Para AURORA2 los mejores resultados los obtiene HEQ con la referencia limpia, y el filtro de suavizado temporal (*HEQ-Rx*).

- La parametrización óptima para AURORA4 es PEQ-Rx (PEQ más filtrado temporal).

- En el caso de HIWIRE, la parametrización que más mejoras produce es *PEQ-PROG*, (PEQ con ecualización progresiva de los coeficientes, que en el caso de HIWIRE, es optima ecualizando los coeficientes *C0-C4*.

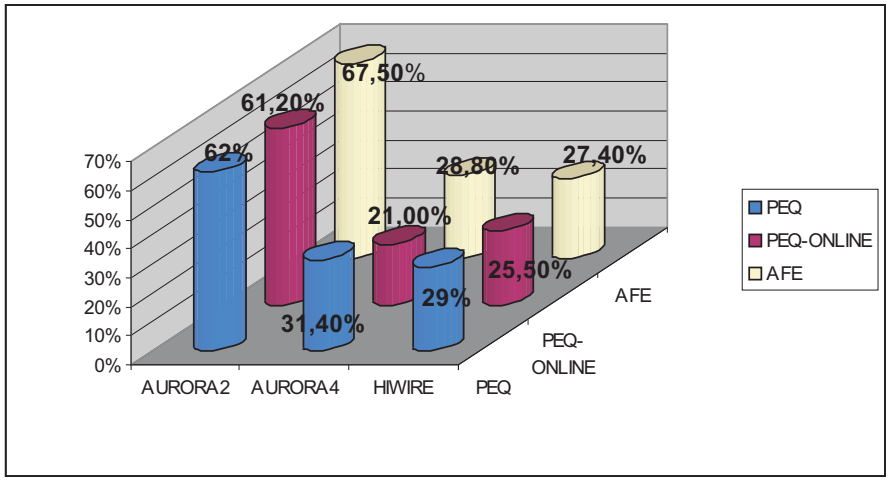


Figura 9.5. Mejora relativa para PEQ versión *online*

9.2. Resultados de Aplicación en el Proyecto HIWIRE

9.2.1. Advanced *front-end* para el proyecto HIWIRE

La utilidad de la parametrización PEQ queda también avalada por su inclusión en el *front-end* avanzado obtenido como resultado del proyecto HIWIRE y denominado HAFE (*Hiwire Advanced Front-End)* [33]. Los bloques incluidos en dicho *front-end* son los siguientes:

- Un subsistema de supresión del ruido formado por:

  i) Un detector de actividad de voz, propuesto y descrito por Ramírez en [111], [110], [109]. Este VAD basado en información contextual temporal es usado para estimar las características del ruido del entorno en el Filtro de Wiener y para el algoritmo de *frame dropping*.

  ii) Un filtro de supresión del ruido basado en el filtrado de Wiener.

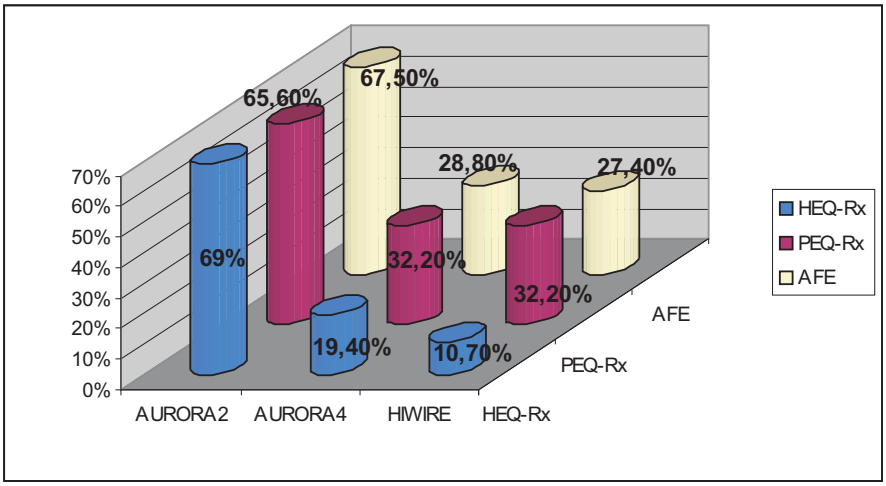


Figura 9.6. WER: suavizado temporal

iii) Un algoritmo de *frame dropping*, que descarta las tramas catalogadas como silencios por el VAD, con el objetivo de eliminar los periodos largos de silencio en las frases,

- Un subsistema de extracción robusta de características. Este subsistema se divide en dos que paralelamente extraen dos tipos de información de los segmentos de voz:

  i) Coeficientes MFCC o una versión mejorada de los mismos, los coeficientes TECC (*TECC-Teager-Kaiser Energy Coeficcients)*que añaden la energía instantánea de Teager-Kaiser a la información que dan los MFCCs. Esta parametrización está enfocada a la captura de la estructura de los formantes.

  ii) Características de modulación de la voz, cuyo objetivo es capturar la estructura fina de la naturaleza variante de la voz en el tiempo. Hay dos conjuntos de características dentro de este grupo:

ii.a) Coeficientes *IFM* (*Weighted Mean Instantaneous Frequency Coefficients*):basados en el modelo de modulación en amplitud y en frecuencia de la señal, que sostiene que las frecuencias de los formantes no son constantes durante la duración de un pitch, sino que fluctúan alrededor de frecuencias centrales. Estos coeficientes capturan dicha fluctuación [34].

ii.b) Características *FMP* (*Frecuency Modulation Percentages*) [101]: dan una estimación normalizada y más robusta frente al ruido de las de las frecuencias de resonancia.

- Un subsistema de normalización de características, en el que la parametrización utilizada es la propuesta en este trabajo de tesis con el nombre de PEQ, Ecualización Paramétrica.

Los resultados de reconocimiento para la base de datos HIWIRE con este *front-end*, se presentan a continuación en la tabla 9.1, en la que las distintas parametrizaciones que se analizan han sido obtenidas con el HIWIRE Advanced *front-end*, tienen filtrado de Wiener, Ecualización paramétrica, CMS o CMVN y *Frame Dropping*:

	Clean	**Low N.**	**Mid N.**	**High N.**	**Valor Medio**	**Mejora Relat.**
MFCC (Base.)	7,49	54,04	76,69	97,85	59,20	0
MFCC	14,2	30,39	46,18	86,74	44,38	25,03 %
TECC	7,2	23,44	46,16	88,19	41,25	30,32 %
TECC+FMP	6,14	18,89	38,23	84,39	36,91	37,65 %
TECC+IFM	7,87	25,25	41,32	86,55	40,25	32,01 %

Tabla 9.1. WER en HIWIRE. Resultados del HAFE

9.2.2. PEQ sobre otras técnicas de reducción del ruido

Se ha estudiado en [41] el efecto de aplicar *PEQ* en el *front-end* del reconocedor automático de voz de Loquendo *Inc*.

Este reconocedor usa modelos basados en una mezcla híbrida de modelos ocultos de Markov *(HMMs)* y Perceptrón Multicapa (*MLP*) [8]. Cada unidad fonética se caracteriza como un autómata de izquierda a derecha de uno o dos estados y auto-bucles, con probabilidades de transición de los *HMMs* uniformes y fijas. El Perceptrón Multicapa está caracterizado mediante una ventana de entrada que modela un contexto temporal y que recibe parametrizaciones Rasta-PLP con los 12 coeficientes cepstrales y el logaritmo de la energía, más las primeras y segundas derivadas de los mismos. La primera capa oculta se divide en tres bloques, uno para la trama central y otros dos para los contextos izquierdo y derecho. Cada bloque a su vez se divide en seis sub-bloques dedicados a contabilizar los seis tipos diferentes de parámetros de entrada.

A este sistema de reconocimiento se le ha añadido un método filtrado espectral y reducción del ruido desarrollado en el proyecto HIWIRE [44] con el nombre de *SNR dependent Ephraim-Malah Sepctral Attenuation (EM)* descrito en detalle en [45].

Las tablas 9.2, 9.3 y 9.4 dan los resultados de estas pruebas en los que podemos ver que para AURORA2 los mejores resultados se obtienen usando solamente *PEQ*. Cuando se aplica previamente la técnica de reducción de ruido (etiquetada como + *EM SNR* en las tablas) hay un empeoramiento relativo del 3,6 % de la tasa de reconocimiento. En el caso de AURORA4 y HIWIRE los mejores resultados se obtienen al sumar *EM* y *PEQ*.

Clean Tr	**Set A**	**Set B**	**Set C**	**Valor Medio**	**Mejora Relativa**
RPLP	24,4	22,5	24,7	23,7	0
+ PEQ	12,9	12,7	13,1	12,9	45,6 %
+ EM SNR	14,7	15,8	15,2	15,2	36,0 %
+ EM SNR + PEQ	13,6	14,2	13,4	13,8	42,0 %
Multicond. Tr.	**Set A**	**Set B**	**Set C**	**Valor Medio**	**Mejora Relativa**
RPLP	6,5	8,9	9,8	8,1	0
+ PEQ	7,2	8,6	7,8	7,9	0,02 %
+ EM SNR	6	8	8,9	7,4	0,09 %
+ EM SNR + PEQ	7,3	8,9	8,3	8,1	0 %

Tabla 9.2. WER para AURORA2. PEQ + EM SNR

	Test 01-07	**Test 08-14**	**Valor Medio**	**Mejora Relativa**
RPLP	66,3	76,9	71,6	0
+ PEQ	44,3	53,3	48,8	31,8 %
+ EM SNR	55	67,1	61,05	14,7 %
+ EM SNR + PEQ	43,4	53,6	48,5	32,3 %

Tabla 9.3. WER para AURORA4. PEQ + EM SNR

	Clean	**Low N.**	**Mid. N.**	**High N.**	**Valor Medio**	**Mejora Relativa**
RPLP	10,8	55,9	79,3	98,1	61,0	0
+ PEQ	14,3	32,4	48,9	85	45,2	25,9
+ EM SNR	10,7	30,6	46,1	84,3	42,9	29,7
+ EM SNR + PEQ	14,8	26,7	40,5	80,2	40,5	34 %

Tabla 9.4. WER para HIWIRE. PEQ + EM SNR

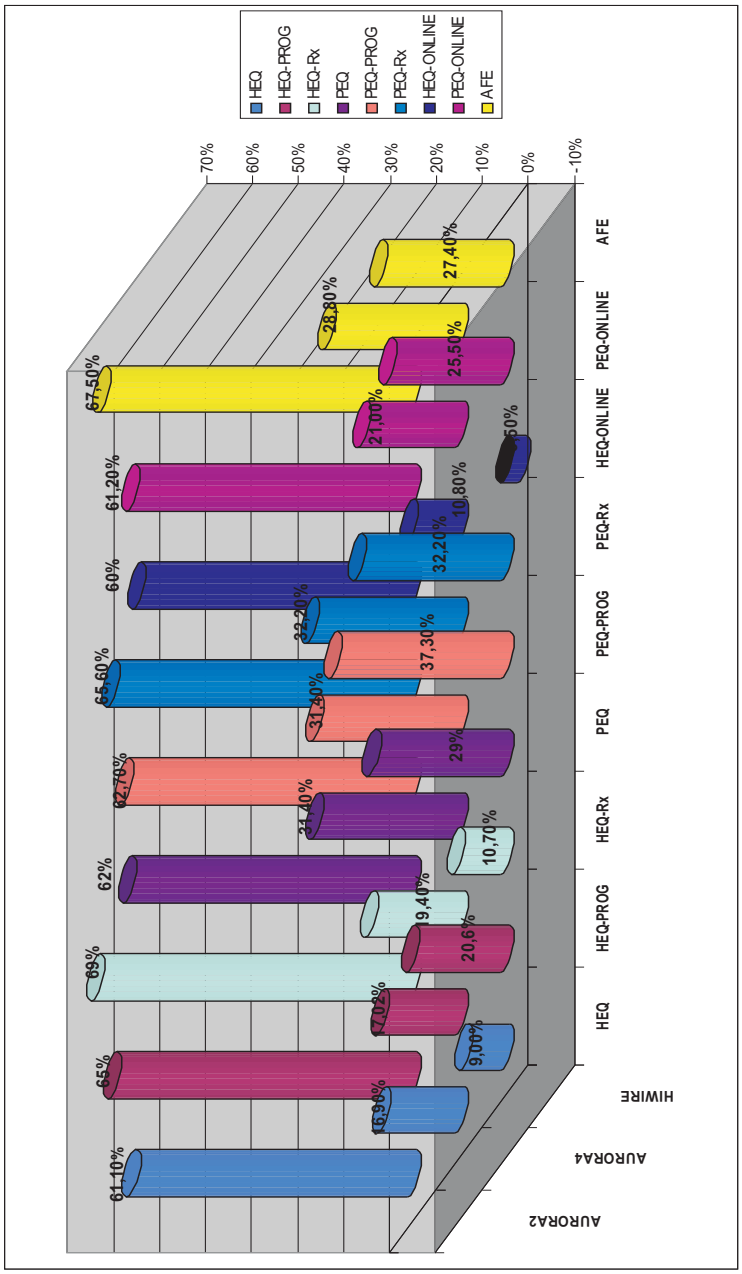


Figura 9.7. Mejora relativa de reconocimiento por parametrización

Capítulo 10

Conclusiones

En este capítulo se incluye un resumen de las principales conclusiones y aportaciones de esta tesis, resumiendo las vistas en cada uno de los capítulos anteriores.

10.1. Conclusiones y análisis de las aportaciones

10.1.1. Sobre la Ecualización de Histogramas

1. La Ecualización de Histogramas es un mecanismo de robustecimiento útil, de bajo coste computacional y que no necesita de la utilización de modelos a priori del tipo de distorsión.

2. Su peculiaridad frente a otras aproximaciones basadas en transformaciones lineales es que implementa una transformación no lineal de los datos.

3. La selección de la distribución de referencia tiene un impacto directo sobre el rendimiento de las técnicas de ecualización. Si bien la utilización de una referencia Gausiana reporta beneficios, éstos son mayores cuando se utiliza una referencia estimada en base a los datos de entrenamiento no contaminados.

4. El proceso de ecualización de los coeficientes cepstrales introduce una cierta distorsión. Debido a que no todos los coeficientes cepstrales

son igualmente discriminativos, ni están igualmente afectados por la distorsión debida al entorno, la ecualización de un subconjunto de coeficientes cepstrales de orden bajo ofrece el mejor compromiso. El número óptimo de coeficientes depende de la base de datos y el nivel de degradación.

5. Se ha propuesto un esquema de ecualización on-line que permite la implementación en tiempo real, a costa de un ligero deterioro en el rendimiento.

10.1.2. Sobre la Ecualización Paramétrica de Histogramas

1. Se ha propuesto un algoritmo de ecualización basado en un modelo paramétrico de dos clases (voz y silencio) que mejora la aproximación de ecualización previamente propuesta de forma que:

   - Al utilizar un modelo paramétrico se obtiene una estimación más suave de las funciones de ecualización cuando el número de datos disponibles es reducido (las observaciones de una única clase).
   - Al utilizar un modelo separado para las tramas de voz y silencio, el proceso de ecualización resultante es independiente de la proporción en que éstas aparecen en la frase.

2. Los resultados del esquema de ecualización paramétrica son superiores a los de la aproximación no paramétrica, especialmente en las tareas complejas de las bases de datos AURORA4 y HIWIRE.

3. Se ha propuesto una versión on-line del algoritmo que permite su utilización en aplicaciones de tiempo real.

10.1.3. Sobre la inclusión y normalización de la información temporal

1. Se ha propuesto una aproximación viable para la incorporación de información sobre correlación temporal en el proceso de ecualización, basada en la normalización de las correlaciones temporales de los coeficientes ecualizados, mostrándose un significativo aumento en el rendimiento del sistema combinado.

10.2. Análisis crítico del trabajo realizado y futuras líneas de investigación

1. En el presente trabajo se ha considerado, por simplicidad, que los coeficientes MFCC son estadísticamente independientes. Sin embargo, aunque pequeña, existe cierta correlación entre dichos coeficientes, y la incorporación de la información sobre estas correlaciones podría incrementar el rendimiento de las técnicas de ecualización.

2. Tanto HEQ como PEQ introducen una cierta degradación cuando no existe distorsión en los datos de evaluación. La reducción de esta degradación es una línea de trabajo abierta.

3. El concepto de ecualización de histogramas ha sido aplicado en este trabajo únicamente al diseño de técnicas de normalización de características. Su aplicación al desarrollo de técnicas de adaptación de modelos es una línea de investigación en curso.

Parte V

Bibliografía y anexos

Capítulo 11

Acrónimos y terminología

11.1. Lista de acrónimos

AFE Advanced Front End

ALSD Average Localized Synchrony Detection

ANN Artificial Neural Networks

CDF Cumulative Density Function

CHMMs Continuous Hidden Markov Models

CMN Cepstral Mean Normalization

CMVN Cepstral Mean and Variance Normalization

CPDLC Controller Pilot Data Link Communications

DHMMs Discrete Hidden Markov Models

DTW Dynamic Time Warping

EIH Ensemble Interval Histogram

EM Expected Maximization

EM SNR Ephraim Malah SNR denoising technique

FFT Fast Fourrier Transform

GMM Gaussian Mixture Models

GSD Generalized Synchrony Detector

HAFE HIWIRE Advanced Front End

HEQ Histogram Equalization

HMM Hidden Markov Models

HIWIRE Human Input that Work in Real Environments

LDC Linear Discriminant Analysis

LMS Least Mean Square

LPC Linear Predictive Coding

MAP Maximum a Posteriori

MFSC Mel Frequency Spectral Coefficients

MFCC Mel Frequency Cepstral Coefficients

MLLR Maximum Likelihood Linear Regression

MLP Multi Layer Perceptron

MMSE Minimum Mean Square Error

MVQHMMs Multiple Vector Quantization Hidden Markov Models

OSEQ Order Statistic-Based Transformations

PCA Principal Component Analysis

pdf probability density function

PEQ Parametric Histogram Equalization

PMC Parallel Model Combination

PLP Perceptual Linear Prediction

QBEQ Quantile Based Equalization

RAH Reconocimiento Automático del Habla

RASTA RelAtive SpecTral Amplitude

RATZ MultivaRiate gAussian based cepsTral normaliZation

SCHMMs Semi-Continous Hidden Markov Models

SCMVQHMMs Semi-Continous Multiple Vector Quantization Hidden Markov Models

SNR Signal to Noise Ratio

SPLICE Stereo-based Piecewise LInear Compensation for Environments

TES TEmporal Smoothing

TICs Tecnologías de la Información y las Comunicaciones

VAD Voice Activity Detector

VTLN Vocal Track Length Normalization

VTS Vector Taylor Series

WACC Word ACCuracy

WER Word Error Rate

ZPCA Zero-Crossing and Peak Amplitudes

11.2. Terminología científica en inglés utilizada

Los siguientes términos en lengua inglesa han sido utilizados por ser ampliamente reconocidos en la comunidad científica y con el objeto de evitar expresiones correctas equivalentes demasiado largas:

Arrays de micrófonos: Estructura ordenada de transductores dispuestos

según una distribución conocida, con lo que es posible relacionar las señales que obtiene cada uno de ellos para obtener un único valor en una dirección determinada.

back-end: Segundo módulo del reconocedor automático del habla, que toma como entradas las salidas del *front-end* del reconocedor y las procesa, entrenando con ellas los modelos HMMs (si la tarea que hace es de entrenamiento) o reconociendo si se trata de una tarea de evaluación.

baseline: Esquema básico de parametrización de partida, que contiene el procesado mínimo necesario para el reconocimiento y a partir del cual se mejoran las tasas de reconocimiento obtenidas.

bins: Son las clases muestrales en las que se divide el histograma acumulativo.

buffer: Ubicación de memoria en una computadora reservada para el almacenamiento temporal de información.

cluster: Agrupación de elementos (parámetros cepstrales en el contexto de esta tesis) con similitudes al aplicarles ciertos criterios.

front-end: Es el primer módulo del reconocedor automático del habla. Representa la interfaz de entrada de datos, y su función es convertir la señal acústica en un conjunto de vectores de parámetros apropiados.

mismatch: Desajustes, o diferencias. Habitualmente se habla de mistmatch entre las condiciones de entrenamiento y *test* del sistema de reconocimineto

set: Conjunto. Habitualmente se usa el término *set* para conjunto de datos.

splines: En el subcampo matemático del análisis numérico, se define

como curva definida a trozos mediante polinomios.

test: Prueba de evaluación

Capítulo 12

Publicaciones

A continuación se presentan las referencias de las publicaciones generadas a partir de los trabajos realizados durante la elaboración de esta tesis doctoral.

[1] **L. García**, J. C. Segura, J.Ramírez, A, de la Torre, C. Benítez. Parametric Non-Linear Feature Equalization for Robust Speech Recognition. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, Francia, 2006.

[2] **L.García**, J.C. Segura, C. Benítez, J. Ramírez, A. de la Torre, Normalization of the interframe information using smoothing filtering. *International Conference on Spoken Language Processing*,Pitsburg, USA, Septiembre 2006.

[3] D, Dimitriadis, J. C. Segura, **L. García**, A. Potamianos, P. Maragos, V. Pitsikalis. Advanced Front-end for Robust Speech Recognition in Extremely Adverse Environments. *International Conference on Spoken Language Processing* Belgium, 2007.

[4] Pedro M. Martínez, J.C. Segura, **L. García**. Robust Distributed Speech Recognition Using Histogram Equalization and Correlation Information. *International Conference on Spoken Language Processing*, Belgium 2007.

[5] **L. García**, Roberto Gemello, Franco Mana, Jose Carlos Segura. Recent Evolutions of Parametric non Linear Feature Equalization. *Enviado y pendiente de aceptación en el ICASSP 2008*.

[6] J. Ramirez, J. C. Segura, C. Benítez, **L. García** and A. Rubio. Statistical Voice Activity Detection Using a Multiple Observatiuon Likelihood Ratio Test. *IEEE Signal Processing Letters, volume 12, nº 10, pages 689-692*, 2005.

[7] J. Ramirez, J. C. Segura, J.M. Górriz, **L. García**. Improved Voice Activity Detection Using Contextual Multiple Hypothesis Testing for Robust Speech Recognition. *IEEE Transactions on Audio Speech and Language Processing, volume 1, nº 11, pages 1-13*, 2002.

Bibliografía

[1] Human inputs that work in real environments. *http://www.hiwire.org*.

[2] Timit acoustic-phonetic continuous speech corpus. *http://www.hiwire.org*.

[3] Itu recommendation g.712, transmission performance characteristics of pulse code modulation channels. 1996.

[4] Speech processing, transmission and quality aspects; distributed speech recognition; front-end feature extraction algorithm; compression algorithms. *ESTSI ES 201 108 v1.1.2 Recommendation*, 2000-04.

[5] Speech processing, transmission and quality aspects; distributed seech recognition; advanced front-end feature extraction algorithm; compression algorithms. *ESTSI ES 202 050 v1.1.1 Recommendation*, 2002-10.

[6] Hervé Abdi. A neural network primer. *Journal of Biological Systems*, 2:247–283, 1994.

[7] Federal Aviation Administration. Controller-pilot datalinl communications. *http://hf.tc.faa.gov/capabilities/cpdlc.htm*.

[8] D. Albesano, R. Gemello, and F. Mana. Hibrid hmm-nn modelling of stationary-transitional units for continuous speech recognition. *Int. Conf. on Neural Information Processing*, pages 1112–1115, 1997.

[9] A. M. Ali, J. Van Der Spiegel, and P. Muller. Robust auditory-based speech processing using the average localized synchrony detection. *IEEE Transactions on Acoustic, Speech, Signal Processing*, 10:279–292, 2002.

[10] European Language Resource Association. Etsi stq aurora proyect database. *http://www.elra.info*.

[11] R. Balchandran and R. Mammone. Non-parametric estimation and correction of non-linear distortion in speech systems. *Proceedings of ICSLP'98*, 1998.

[12] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Statist.*, 41:164–171, 1970.

[13] R.E. Belman. *Dinamic Programming*. Princeton Univ. Press, 1957.

[14] C. Benitez, L. Burget, B. Chen, S. Dupont, H. Garudadri, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas. Robust asr front-end using spectral-based and discriminant features: experiments on the aurora tasks. *Proc. of EUROSPEECH*, 2001.

[15] S. F. Boll. Suppression of acoustic noise in speech using spectral sustraction. *IEEE Transactions on Acoustic, Speech, Signal Processing*, ASSP-27, nº2:113–120, 1979.

[16] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. *Proc. of ICSLP '96*, 1:426–429, 1996.

[17] Scott Shaobing Chen and Ramesh A. Gopinath. Gaussianization. *Proceedings of NIPS 2000, Denver Colorado*, 2000.

[18] A. Cole, J. Mariani, H. Uzkoreit, A. Zaenen, and V. Zue. *Survey of the State of the Art in Human Language Technology*. National Science Foundation, Directorate XIII-E of the Commision of the European Commnunities Center for Spoken Language Understanding, Oregon Graduate Institute, 1995.

[19] M.P. Cooke, P. D Green, L.Jovsifovski, and A. Vizino. Robust asr with unreliable data and minimal assumptions. *Proc. of Robust'99*, pages 195–198, 1999.

[20] M.P. Cooke, A. Morris, and P. D Green. Recognizing occluded speech. *ESCA Tutorial and Workshop on Auditory Basis of Speech Perception, Keele University*, 1996.

[21] H. Cox, R. Zeskind, and I. Kooij. Practical supergain. *IEEE Transactions on Acoustics, Speech and Signal Processing*.

[22] H. Cox, R. Zeskind, and M. Owen. Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech and Signal Processing*.

[23] Tran Huy Dat, Kazuya Takeda, and Fumitada Itakura. A speech enhancement system based on data clustering and cumulative histogram equalization. *Proceedings of ICDE'05*, 2005.

[24] Steven B. Davis and Paul Merlmenstein. Comparison of parametric representations for monosyllabic word recognition in continously spoken sentences. *IEEE Transactions on Acoustic, Speech, Signal Processing*, ASSP-28, 4:357–365, 1980.

[25] A. de la Torre, A. Peinado, and A. Rubio. *Reconocimiento Automático de Voz en condiciones de ruido*. Monografías del Departamento de Electrónica, nº 47. Departamento de Electrónica y Tecnología de Computadores, Universidad de Granada, 2001.

[26] A. de la Torre, A. Peinado, J. C. Segura, J.L. Pérez Córdoba, C. Benítez, and A. Rubio. Histogram equalization of speech representation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13,n3:355–366, 2005.

[27] A. de la Torre, J. Segura, C. Benitez, A. Peinado, and A. Rubio. Non linear transformation of the feature space for robust speech recognition. *Proc. of ICASSP*, 2002.

[28] Angel de la Torre, Antonio Peinado, and Antonio Rubio. *Reconocimiento Automatico de voz en condiciones de ruido*. Monografias del Departamento de Electronica, n 47, Universidad de Granada, Granada, Espa$\tilde{n}$a, 2001.

[29] F. de Wet, J. de Veth, B. Cranen, and L. Boves. The impact of spectral and energy mismatch on the aurora2 digit recognition task. *Proceedings of ICASP'03*, II:105–108, 2003.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39,nº1:1–38, 1977.

[31] L. Deng, A. Acero, M. Plumpe, and X. Huang. Large-vocabulary speech recognition under adverse acoustic environments. *Proc. of ICSLP'00*, 2000.

[32] S. Dharanipragada and M. Padmanabhan. A nonlinear unsupervised adaptation technique for speech recognition. *Proceedings of ICSLP'00*, pages 556–559, 2000.

[33] D. Dimitriadis, L. García, A. Potamianos, P. Maragos, and V. Pitsikalis. Advanced front-end for robust speech recognition in extremly adverse environments. *Proceedings of ICSLP'07*, 2007.

[34] D. Dimitris, P. Maragos, and A. Potamianos. Robust am-fm features for speech recognition. *IEEE Signal Processing Letters*, 2005.

[35] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean square error log-spectral amplitude estimator. *IEEE Transactions on speech and audio processing*, 20, nº33:443–445, 1985.

[36] S. Young et alt. *The HTK Book*. Microsoft Corporation & Cambridge University Engineering Department, 1995.

[37] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Speech and Audio Processing*, 29,2:254–272, 1981.

[38] S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. ASSP*, 34:52–59, 1986.

[39] M. J. Gales and S. J. Young. Cepstral parameter compensation for the update of the parameters of a single mixture density hmm recognition in noise. *Speech Communications*, 12:231–239, 1993.

[40] M.J.F. Gales and P.C. Woodland. Mean and variance adaptation within the mllr framework. *Computer Speech & Language*, 10:249–264, 1996.

[41] Luz Garcia, Roberto Gemello, Franco Mana, and J.Carlos Segura. Recent evolutions of parametric non-linear feature equalization. *Pendiente de aceptación en ICASSP'8*, 2008.

[42] Luz Garcia, J.Carlos Segura, Javier Ramirez, Angel de la Torre, and Carmen Benitez. Parametric nonlinear feature equalization for robust speech recognition. *Proceedings of ICASSP'06*, pages 529–532, 2006.

[43] J.L. Gauvain and C.H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on speech and audio processing*, 2, nº2:291–298, 1994.

[44] R. Gemello, Franco Mana, and Renato De Mori. Hiwire, human inputs that work in real environments. *contract number IST-2002-507943*, 2002.

[45] R. Gemello, Franco Mana, and Renato De Mori. Automatic speech recognition with a modified ephraim-malah rule. *IEEE Signal Processing Letters*, 13,1:56–59, 2006.

[46] O. Ghitza. Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Transactions on Acoustic, Speech, Signal Processing*, 2:52–59, 1986.

[47] M. Ghulam, J. Horikawa, and T.Ñitta. A pitch-synchronous peak-amplitude based feature extraction method for robust asr. *Proc. of ICSLP'05*, 1:517–520, 2005.

[48] A. Gómez. *Tratamiento de la degradación debida al Canal en Sistemas de Reconocimiento Robusto*. Tésis Doctoral, Departamento de Teoría de la Señal, Telemática y Comunicaciones, Universidad de Granada, 2006.

[49] R.C González and P. Wintz. *Digital Image Processing*. Addison-Wesley, 1987.

[50] L. Griffiths and C. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*.

[51] Joseph Hair, Rolph Anderson, and William Black. *Análisis Multivariante*. Prentice Hall, 1999.

[52] John B. Hampshire and A. Waibel. A novel objective function for improved phoneme recognition using time-delay neural networks. *IEEE Trans. on Neural Networks*, 1, nº 2:216–228, 1990.

[53] Hemmo Haverinen and Irme Kiss. On-line parametric histogram equalization techniques for noise robust embedded speech recognition. *Proceedings of EUROSPEECH'03*, pages 3061–3064, 2003.

[54] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Transactions on acoustics, speech, and signal processing*, 2, nº 4:578–589, 1994.

[55] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. Compensation for the effect of the communication channel in auditory-like analysis of speech (rasta-plp). *Proceedings of Second European Conference on Speech Comm. and Tech.*, pages 121–124, 1991.

[56] H. Hermansky, N. Morgan, A. Bayyman, and P. Kohn. Rasta-plp speech analysis. *ICSI Technical Report, no TR-91-069*, 1991.

[57] H. Hermansky, K. Tsuga, S. Makino, and H. Wakita. Perceptually based processing in automatic speech recognition. *Proceedings of ICASSP'86*, pages 1971–1974, 1986.

[58] F. Hilger and H.Ney. Quantile based histogram equalization for noise robust speech recognition. *Proceedings of EUROSPEECH'01*, 2001.

[59] F. Hilger and H.Ney. Quantile based histogram equalization for noise robust large vocabulary speech recognition. *IEEE Transactions on speech and audio processing*, 2006.

[60] H.G. Hirsch. Experimental framework for the performance evaluation of speech recognition front-ends of large vocabulary task. *STQ AURORA DSR Working Group*, 2002.

[61] X. Huang, A. Acero, and H-W Hon. *Spoken Language Processing, A guide to theory, algorithms and system development*. Prentice Hall, 2001.

[62] T. B. Hughes, S. S. Kim, J. H. DiBiase, and H. F. Silverman. Performance of an hmm speech recognizer using a real-time tracking microphone array as input. *IEEE Trans. Speech Audio Processing*.

[63] Kentaro Ishizuka and Noboru Miyazaki. Speech feature extraction method representing periodicity and aperiodicity in sub bands for robust speech recognition. *Proceedings of ICASSP'04*, 2:1483–1486, 2006.

[64] L. Josifovski, M. Cooke, P. Green, and A. Vizinho. State based imputation of missing data for robust speech recognition and speech enhancement. *Proc. Eurospeech 1999*, 1999.

[65] Biing-Hwang Juang, L.R. Rabiner, and J.G. Wilpon. On the use of bandpass liftering in speech recognition. *IEEE Transactions on acoustics, speech, and signal processing*, 35(7):947–954, 1987.

[66] Yang-Won JUNG, Hong-Goo KANG, Chungyong LEE, Dae-Hee YOUN, Changkyu CHOI, and Jaywoo KIM. Adaptive microphone array system with two-stage adaptation mode controller. *IEICE TRANS. FUNDAMENTALS*.

[67] Chanwoo Kim, Yu-Hsiang Chiu, and Richard M. Stern. Physiologically motivated synchrony-based processing for robust automatic speech recognition. *Proceedings of ICSLP'06*, 2:1483–1486, 2006.

[68] D. S. Kim, S. Y. Lee, and R. M. Kil. Auditory processing of speech signals for robust speech recognition in real world noisy environments. *IEEE Transactions on Acoustic, Speech, Signal Processing*, 7:55–69, 1999.

[69] R.G. Leonard. A database for speaker independent digit recognition. *Proc. of ICASSP'84*, 3, 1984.

[70] Bo Liu, Li-Rong Dai, Jin-Lu Li, and Ren-Hua Wang. Double gaussian based feature normalization for robust speech recognition. *Proceedings of ISCLP'04*, pages 253–246, 2004.

[71] P. Lockwood and J. Boudy. 11,:572–578, 1992.

[72] R. Lyon. A computational model of filtering, detection and compresion in the cochlea. *Proc. of ICASSP'82*, 7:1282–1285, 1982.

[73] J. Makhoul. Spectral analysis of speech by linear prediction. *IEEE Transactions on Acoustic, Speech, Signal Processing*, 21, 3:140–148, 1973.

[74] P. Maragos and A. Potamianos. Fractal dimensions of speech sounds: computation and application to automatic speech recognition. *J. Acoustic Soc. Am.*, 105,3:1925–1932, 1999.

[75] J.D. Markel and A.H. Gray. *Linear Predictionm of Speech*. Springer-Verlag, 1976.

[76] Pedro M. Martinez, Jose C. Segura, and Luz Garcia. Robust distributed speech recognition using histogram equalization and correlation information. *Proceedings of Interspeech'07*, 2007.

[77] S. Molau, F. Hilger, D. Keysers, and H.Ñey. Enhanced histogram equalization in the acoustic feature space. *Proceedings of ICSLP'02*, pages 1421–1424, 2002.

[78] S. Molau, F. Hilger, and H.Ñey. Feature space normalization in adverse acoustic conditions. *Proceedings of ICASSP'03*, pages 1421–1424, 2003.

[79] S. Molau, M. Pitz, and H.Ñey. Histogram based normalization in the acoustic feature space. *Proc of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001.

[80] Md. Khademul Islam Molla and Keikichi Hirose. On the effectiveness of mfccs and their statistical distribution properties in speaker identification. *IEEE International Conference on Virtual Environments, Human-Computer Interface and Measurements Systems*, pages 136–141, 2004.

[81] P.J. Moreno. Speech recognition in noisy environments: A survey. *Ph.D. Disertation, Carnegie Mellon University, Pittsburgh, PA,1996*.

[82] P.J. Moreno, B. Raj, E. Gouvea, and R. Stern. Multivariate-gaussian-based cepstral normalization for robust speech recognition. *Proc. of ICASSP'95*, pages 137–140, 1995.

[83] P.J. Moreno, B. Raj, and R. Stern.

[84] N. Morgan and H.Boulard. Continuous speech recognition using multilayer perceptrons with hidden markov models. *Proc. of ICASSP'90*, pages 413–416, 1990.

[85] Y. Obuchi. Improved histogram-based feature compensation for robust speech recognition and unsupervised adaptation. *Proc. of ICSLP*, 2004.

[86] Y. Obuchi and R. Stern. Normalization of time-derivative parameters using histogram equalization. *Proc. of EUROSPEECH*, 2003.

[87] Y. Obuchi and R.M. Stern. Normalization of time-derivative parameters using histogram equalization. *Proceedings of EUROSPEECH'03*, pages 665–668, 2003.

[88] S. Okawa, E. Bocchieri, and A. Potamianos. Multiband speech recognition in noisy environments. *Proc. of ICASSP'98*.

[89] S. Okawa, T.Ñakajima, and K. Shirai. A recombination strategy for multi-band speech recognition based on mutual information criterion. *Proc. of Europ. Conf. Speech Comm. Tech., Budapest*, pages 603–606, 1999.

[90] Peder A. Olsen, Scott Axelrod, Karthik Visweswariah, and Ramesh Gopinath. Gaussian mixture modeling with volume preserving nonlinear feature space transforms.

[91] M. Omologo. Hands-free speech recognition: current activities and future trends. *Proc. of Int. Workshop Hands-Free Speech Communication*.

[92] A. Openheim and W. Schafer. From frequency to quefrency: a history of the cepstrum. *IEEE Signal Processing Magazine*, pages 95–106, 2004.

[93] J.P. Openshaw, Z.P. Sun, and J.S. Mason. A comparison of composite features under degraded speech in speaker recognition. *Acoustics, Speech, and Signal Processing,IEEE International Conference on*, 2:371–374, 1993.

[94] Douglas O'Shaughnessy and Marcel Gabrea. Efficient recognition of continuosly-spoken numbers. *IEEE Canadian Conference on Electrical and Computer Engineering*, pages 505–509, 2001.

[95] Pierre Ouellet, Gilles Boulianne, and Patrick Kenny. Flavours of gaussian warping. *Proceedings of INTERSPEECH'05*, pages 2957–2960, 2005.

[96] D. Pearce and H.G. Hirsch. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. *Proc. of ICSLP'00*, 2000.

[97] Antonio Peinado and Jose Carlos Segura. *Speech Recognition Over Digital Channels*. John Wiley, 2006.

[98] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. *Proceedings of Speaker Odyssey 2001 conference*, 2001.

[99] Z. Peyton and JR. Peebles. *Probability, Random Variables and Random signal principles*. Mac-Graw Hill, 1993.

[100] P.N.Belhumeur, J.P.Hespanha, and D.J.Kriegman. Eigenfaces vs fisherfaces: Recognition using classspecific linear projection. *IEEE Trans. Pattern Anal. Machine Intell*, 19:711–720, 1997.

[101] A. Potamianos and P. Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *Journal on Ac. Soc. Am.*, 1996.

[102] John G. Proakis and Dimitris G. Manolakis. *Digital Signal Processing*. Prentice Hall PTR, 1996.

[103] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice Hall PTR, 1993.

[104] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of IEEE,*, 77, 2:257–286, 1989.

[105] B. Raj, M.Seltzer, and R. M. Stern. Reconstruction of damaged spectrographic features for robust speech recognition. *Proc. ICSLP 2000*, 2000.

[106] B. Raj, M.Seltzer, and R. M. Stern. Reconstruction of missing features for robust speech recognition. *Speech Comm. Journal*, 43,nº4:275–296, 2004.

[107] B. Raj, M. Seltser, and R. Stern. Robust speech recognition: the case for restoring missing features. *Proc. of CRAC'01*, pages 301–304, 2001.

[108] B. Raj and R. M. Stern. Missing-feature approaches in speech recognition. *IEEE Signal Processing Magazine*, pages 101–116, 2005.

[109] J. Ramirez, J. C. Segura, C. Benítez, L. García, and A. Rubio. Statistical voice activity detection using a multiple observatiuon likelihood ratio test. *IEEE Signal Processing Letters,*, 12, nº 10:689–692, 2005.

[110] J. Ramirez, J. C. Segura, J.M. Górriz, and L. García. Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition. *IEEE Transactions on Audio Speech and Language Processing*, 1, nº 11:1–13, 2002.

[111] J. Ramirez, J.C. Segura, A. de la Torre, and A. Rubio. Efficient voice activity detection algorithms using long-term speech information. *Speech Commnunications*, 2004.

[112] Mosur K. Ravishankar. Efficient algorithms for speech recognition. *Ph.D. Thesis, School of Computer Science, Computer Science Division, Carnegie Mellon University*, CMU-CS-96-143, 1996.

[113] Virgilio Roel. *La Tercera Revolución Industrial y la Era del Conocimiento*. Fondo editorial UNMSM, 1998.

[114] J.C. Russ. *The Image Processing Handbook*. Boca Ratón, 1995.

[115] Y. Salimpour and M. D. Albohassani. Auditory wavelet transform based on auditory wavelet families. *Proc. of the 28th EMBS Annual International Conference*, 87:1731–1734, 1990.

[116] Georges Saon, Satya Dharanipragada, and Dan Povey. Feature space gaussianization. *Proceedings of ICASSP'04*, pages 329–332, 2004.

[117] J. C. Segura, M. C. Benitez, A. de la Torre, S. Dupont, and A. Rubio. Vts residual noise compensation. *Proceedings of ICASP'02*, pages 409–412, 2002.

[118] J. C. Segura, C. Benítez, A. de la Torre, and A. Rubio. Cepstral domain segmental nonlinear feature transformations for robust speech recognition. *IEEE Signal Processing Letteres*, 11, nº5:517–520, 2004.

[119] J. C. Segura, C. Benítez, J. Ramírez, A. de la Torre, and A. Rubio. Efectos no lineales del entorno acústico en parametrizaciones para reconocimiento automático de voz basadas en mfcc. *Actas de las terceras Jornadas de la Red Tecnológica del habla*, pages 27–32, 2004.

[120] J. C. Segura, M.C. Benítez, A. de la Torre, S. Dupont, and A. Rubio. Improved feature extraction based on spectral noise reduction and non linear feature normalization. *Proceedings of ICASSP'02*, pages 409–412, 2002.

[121] J. C. Segura, J. Ramírez, C. Benítez, A. de la Torre, and A. Rubio. Improved feature extraction based on spectral noise reduction and non linear feature normalization. *Proceedings of EUROSPEECH'03*, pages 353–356, 2003.

[122] J.C. Segura, T. Ehrette, A. Potaminaos, and D. Fhor et alter. The hiwire database, a noisy and non-native english speech corpues for cockpit communications. *http://www.hiwire.org*.

[123] S. Seneff. A computational model for the peripheral auditory system: application to speech recognition research. *Proceedings of ICASSP'86*, pages 1983–1986, 1986.

[124] Youngjoo Suh, Mikyong Ji, and Hoiring Kim. Probabilistic class histogram equalization for robust speech recognition. *IEEE Signal Processing Letters*.

[125] Youngjoo Suh and Hoirin Kim. Class-based histogram equalization for robust speech recognition. *ETRI Journal*, 28, nº4:502–505, 2006.

[126] Yik-Cheung Tam and Brian Mak. Optimization of sub-band weights using simulated noisy speech in multi-band speech recognition. *Proc. of ICSLP 2000*.

[127] S. Tibrewala and H. Hermansky. Multi-band and adaptation approach for robust speech recognition. *Proc. of EUROSPEECH'97*, 1997.

[128] Yohichi Tohkura. A weighted cepstral distance measure for speech recognition. *IEEE Transactions on acoustics, speech, and signal processing*, ASSP-35,10:1414–1422, 1987.

[129] S. van Vuuren and H. Hermansky. Data driven design of rasta-like filters. *Proc. of EUROSPEECH*, 1:409–412, 1997.

[130] J. Díaz Verdejo, J.C. Segura, A. Rubio, A. Peinado, and J.L. Pérez-Córdoba. A new neuron model for an alpahanet-semicontinuous hmm. *Proc. of ICASSP'93*, pages 529–532, 1993.

[131] O. Viiki, B. Bye, and K. Laurila. A recursive feature vector normalization approach for robust speech recognition in noise. *Proc. of ICASSP'98*, 1998.

[132] Karthik Visweswariah and Ramesh Gopinath. Feature adatation using proyection of gaussian posteriors. *Proc. of ICASSP 2002*, 2002.

[133] Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13,2:260–269, 1967.

[134] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme recognition using time-delay neural networks. *IEEE Trans. Acoust. Speech Signal Processing ASSP*, 37:328–339, 1989.

[135] Chang wen Hsu and Lin shan Lee. Higher order cepstral moment normalization (hocmn) for robust speech recognition. *Proceedings of ICASP'04*, pages 197–200, 2004.

[136] Bing Xiang, Upendra V. Chaudhari, Jiri Navratil, GaneshÑ. Ramaswamhy, and Ramesh A. Gopinath. Short-time gaussianization for robust speaker verification. *Proceedings of ICASSP'02*, pages 681–684, 2002.

[137] D. Yuk, C. Che, L. Jin, and Q. Lin. Environment independent continuous speech recognition using neural networks and hidden markov models. *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

[138] DongSuk Yukyz and James Flanagany. Telephone speech recognition using neural networks and hidden markov models. *Proc. of ICASSP'99*.