# Técnicas de reconocimiento robusto de la voz basadas en el pitch


Juan Andrés Morales Cordovilla

Dpto. de Teoría de la Señal Telemática y Comunicaciones

Universidad de Granada

Editor: Editorial de la Universidad de Granada
Autor: Juan Andrés Morales Cordovilla
D.L.: GR 967-2012
ISBN: 978-84-694-9344-1

**D. Antonio M. Peinado Herreros** y **Dña. Victoria Sánchez Calle**, Catedrático y Profesora Titular de Universidad del Departamento de Teoría de la Señal, Telemática y Comunicaciones

**CERTIFICAN:**

Que la memoria titulada: **"Técnicas de reconocimiento robusto de la voz basadas en el pitch"** ha sido realizada por **Juan Andrés Morales Cordovilla** bajo nuestra dirección en el Departamento de Teoría de la Señal, Telemática y Comunicaciones de la Universidad de Granada para optar al grado de Doctor en Ingeniería Electrónica.

Granada, a de de 2011

Fdo. Antonio M. Peinado Herreros
Director de la Tesis

Fdo. Victoria Sánchez Calle
Directora de la Tesis

A mis padres, por tanto como han hecho por mi.

*Hasta el pensamiento, hasta la invención, son hechos colectivos, producto del pasado y del presente.*
Piotr Kropotkin

*Comprensión es compresión*
Gregorio Chaitin

# Agradecimientos

Quisiera expresar mi mas sincero agradecimiento a todos los que han hecho que esto sea posible, especialmente a Antonio Peinado y Victoria Sánchez, directores de esta Tesis, por su excelente dirección e instrucción y por haber dedicado tanto esfuerzo en este trabajo. A Ning Ma por su calurosa acogida y haberme enseñado tanto en mis estancias. A José Andrés y José Luis por su compañerismo y su ayuda con los "ordenadores". Al departamento de Teoría de la Señal, Telemática y Comunicaciones y a la Universidad de Granada por sus becas, las cuales me han permitido viajar al extranjero y mejorar mi formación enormemente. A mis familiares y amigos por llenar mi vida de alegría y buenos momentos. Y como no, a mi novia Yaba por todo su apoyo y amor.

# Resumen

Esta Tesis propone y hace un estudio de técnicas que emplean de una forma u otra el pitch, el cual será entendido como la frecuencia fundamental en cada instante de tiempo de la voz, para transcribirla o reconocerla de forma robusta en condiciones de ruido. No pretende buscar un modo robusto de extraer el pitch, sino y sobre todo, una vez conocido este, ver como emplearlo de manera adecuada para robustecer el reconocimiento.

Se hará un estudio bibliográfico de las técnicas que han empleado el pitch intentando una primera clasificación de las mismas. Después, se propondrán tres técnicas de reconocimiento robusto basadas en el pitch comparándolas con otras similares. Estas técnicas son: *ventanas asimétricas* que se aplican sobre la autocorrelación de una señal para extraer un espectro menos afectado por el ruido, *autocorrelación cribada y (promediada)* que es capaz de estimar completamente la autocorrelación limpia de una señal periódica empleando el pitch bajo ciertas suposiciones de ruido, y *estima del ruido basada en el pitch* que es capaz de estimar ruidos no estacionarios a partir del pitch mediante lo que se denomina estima túnel y que se empleará sobre un reconocedor de MD (Missing Data) basado en marginalización.

Aparte de esto, se intentarán mostrar los resultados límite en el reconocimiento de las técnicas basadas en el pitch y que emplean la mínima información posible sobre el ruido. Para ello se identificarán los mecanismos básicos de reconocimiento robusto de los sonidos sonoros empleados por estas técnicas, se verá cuales son los óptimos (mostrando equivalencias) y se mostrarán experimentalmente esos resultados límite a partir del uso de máscaras oráculo de MD y de valores de pitch ideales. Concluiremos que la técnica *estima del ruido basada en el pitch* se acerca idealmente a los límites del reconocimiento basado en el pitch (suponiendo pitch ideal) pero que queda (aunque no por una excesiva diferencia) lejos de los límites de las máscaras oráculo. Finalmente, se

dará un pequeño bosquejo de como podría abordarse el reconocimiento cuando no hay pitch (habla susurrante) reciclando ciertas ideas presentadas en la Tesis.

# Abstract

This Thesis proposes and carries out a study of different techniques which, in some way, use the pitch (which will be understood as the fundamental frequency of speech) in order to carry out robust ASR (Automatic Speech Recognition) under noise conditions. The Thesis is not concerned with pitch extraction itself, but with the best way of using pitch for robust speech recognition.

We will also carry out a study of the related bibliography and the state of art regarding these pitch-based techniques for robust ASR. Then, we will propose three pitch-based techniques which will be compared to other similar ones. Our three proposals are: application of *asymmetric windows* to the noisy signal autocorrelation which tries to provide a spectrum less sensitive to noise, two estimators, named as *averaging and sifting estimators*, of the autocorrelation of the clean quasi-periodic signal, and a *noise estimation technique* which can deal with non stationary noise by employing pitch information and which is used to estimate the reliability masks required by a marginalization MD (Missing Data) recognizer.

Additionally, we will discuss the performance limits of the pitch-based techniques for robust ASR which employ minimal assumptions about the noise. In order to do so, we will identify the basic robust mechanisms employed by these techniques for recognizing voiced frames, the optimum mechanisms will be identified (by means of some equivalences), and the corresponding limit results will be experimentally obtained by applying MD oracle masks and ideal pitch. One of our conclusions is that our noise estimation technique for MD recognition is close to the limits of the pitch-based robust ASR techniques, although it would require additional information in order to achieve the performance with MD oracle masks. Finally, we will comment some possibilities (some of them related to speech without pitch) for future research from the ideas developed in this Thesis.

# Índice general

# Índice de figuras

## ÍNDICE DE FIGURAS

# Índice de tablas

# Siglas y términos en inglés

**AMFCC** Autocorrelation Mel-Frequency-Cepstral-Coeficients (Coeficientes Mel-Frecuenciales-Cepstrales derivados de la Autocorrelación)

**ASA** Auditory Scene Analysis (Análisis de Escenas Auditivas)

**ASR** Automatic Speech Recognition (Reconocimiento Automático de la Voz)

**DDR** Double Dynamic Range (ventana con Rango Dinámico Doble)

**HASE** High-lag Autocorrelation Spectrum Estimation (Estimación Espectral con coeficientes Altos de la Autocorrelación)

**HMM** Hidden Markov Models (Modelos Ocultos de Markov)

**HT** Harmonic Tunnelling (técnica [38] de Tunelaje Armónico)

**MD** Missing Data (Datos Perdidos)

**MSD** Magnitude Spectral Density (Magnitud de la Densidad Espectral)

**OSA** One Side Autocorrelation (Una de las Partes de la Autocorrelación)

**Pitch** Tono, periodo o frecuencia fundamental de la voz

**Píxel** Elemento espectro-temporal o cepstro-temporal de una representación acústica

**SFD** Speech Fragment Decoding (Decodificación de Fragmentos de Voz)

**SS** Spectral Subtraction (Sustracción Espectral)

**WAcc** Word Accuracy (tasa de Acierto de Palabra)

# Capítulo 1

# Introducción

## 1.1. Introducción

### 1.1.1. Motivación y planteamiento del problema

#### Reconocimiento automático de la voz en condiciones de ruido

Los sistemas de ASR (Automatic Speech Recognition, Reconocimiento Automático de la Voz) encargados de transcribir la información lingüística de la voz o el habla en texto, más desarrollados y comercializados hoy día, aún están muy lejos de reconocer con la misma exactitud y robustez con la que reconoce el ser humano. Para ello basta con probar cualquiera de estos sistemas que traen incorporados muchos de nuestros móviles, bien hablándoles con rapidez o en ambientes ruidosos.

Reconocer voz de forma automática no es más que comparar una representación de la señal de voz con una serie de patrones previamente establecidos. La implementación de los sistemas de ASR requiere el desarrollo de dos etapas diferenciadas: una de entrenamiento, en la que se establecen los patrones, y otra de test para validar el sistema. El que en la etapa de test no se obtengan buenos resultados se debe principalmente a que la voz se presenta de una forma distinta e incontrolable a la prevista por la etapa de entrenamiento.

Son muchas las formas en las que se puede presentar la voz y los investigadores que trabajan en el problema del ASR suelen poner restricciones respecto a la cantidad de formas en las que esta se puede presentar para así limitar el problema del reconocimiento. Una posible clasificación de estas formas consiste en hacerlas depender de los siguientes parámetros [109]: conjunto de locutores, modo de pronunciación, complejidad de la gramática, tamaño del vocabulario y tipos de ruidos posibles. En la Fig. 1.1 podemos

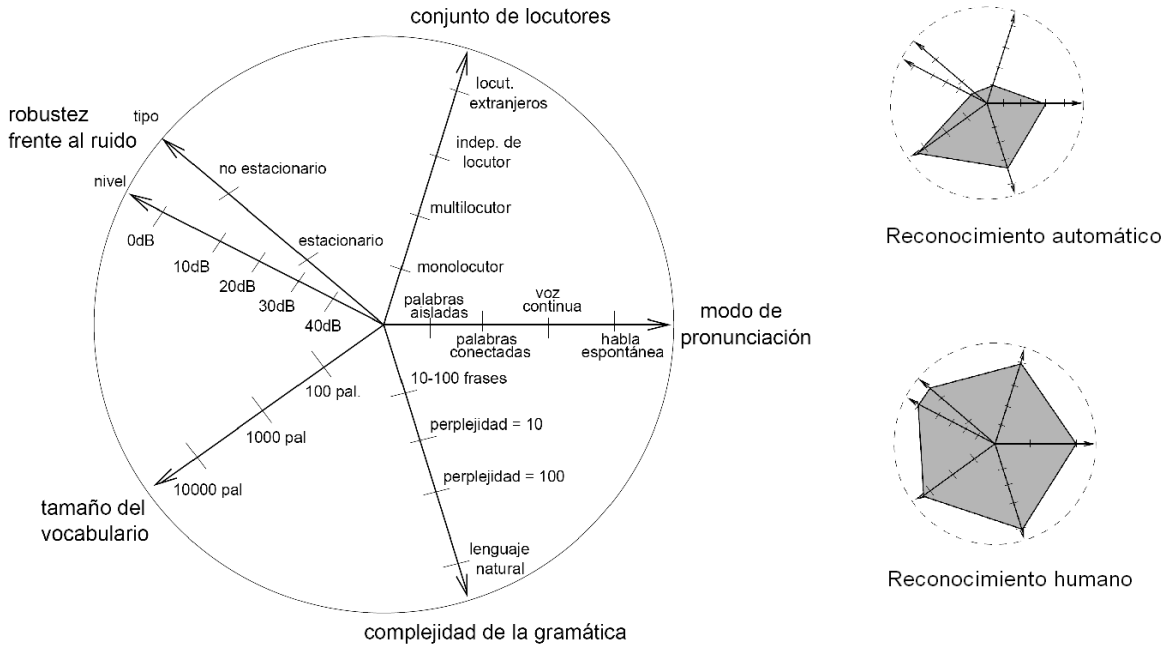



Figura 1.1: [109] Diagrama de Kiviat que nos muestra la variedad de formas en las que se puede presentar la voz y como el reconocedor automático (en este caso un dictáfono) solo puede abarcar de forma totalmente fiable un conjunto restringido de las mismas frente al hombre que puede abarcarlas todas.

ver un diagrama de Kiviat de esta variedad de formas. Lo interesante de este diagrama es que nos muestra que debido a la capacidad limitada de computación, los sistemas de reconocimiento solo pueden abordar regiones restringidas de este diagrama frente al humano que puede llegar a abordarlo todo completamente.

Según la región del diagrama de Kiviat a abordar podemos distinguir diferentes áreas de investigación en el campo del ASR. Entre ellas podemos mencionar las de los sistemas de diálogo, las del reconocimiento de habla continua con grandes vocabularios, y las del reconocimiento robusto en condiciones de ruido entre otras.

Los investigadores que trabajan en los sistemas de dialogo emplean gramáticas muy restrictivas y dirigidas que hacen que sus sistemas solo puedan reconocer ciertas palabras o frases aisladas en determinados instantes del proceso de reconocimiento. Sus aplicaciones suelen ser sistemas expendedores o de consulta telefónica de forma que no suelen imponer restricciones respecto al conjunto de locutores (edad, genero, acento, etc.) pero si respecto al tamaño de vocabulario dirigido en cada instante por la gramática.

Los que trabajan en el reconocimiento de habla continua intentan crear sistemas capaces de reconocer voz sin preocuparse por el tamaño del vocabulario, complejidad gra-

matical y modo de pronunciación, aunque suelen ser bastantes restrictivos respecto a la robustez frente al ruido y al conjunto de locutores (pues en cuanto se les hace reconocer voz con acento extraño suelen fallar). El dictáfono Dragon Dictation desarrollado por Nuance es un ejemplo de aplicación comercial de estos sistemas.

Los que trabajan en reconocimiento robusto intentan crear sistemas que no se vean afectados por la variabilidad del ruido (por esto el nombre de robusto) pero suelen restringirse a reconocer pronunciaciones de un número finito de secuencias de palabras conectadas que siguen una gramática muy simple. Por palabra conectada se entiende que no tiene porque haber pausa entre las distintas palabras. Estos investigadores no suelen crear aplicaciones directas pero proponen técnicas y procedimientos que esperan ser añadidos a posteriori en sistemas tales como los de habla continua gracias al empleo de interfaces comunes. Estas interfaces son las gramáticas, los modelos de reconocimiento y las características de la voz. La mejor forma de fusionar los sistemas de habla continua con los de reconocimiento robusto es un asunto que aún no está claro pero que ya se está empezando a investigar [127, 54]. Es más, estos investigadores también ayudan a mejorar la comprensión de como el ser humano realiza la audición y separación de fuentes sonoras por lo que su trabajo está muy relacionado con la psicoacústica.

En esta Tesis nos centraremos en el reconocimiento robusto de la voz en condiciones de ruido. Para evitar emplear reglas complejas de alto nivel lingüístico nuestras palabras serán secuencias aleatorias de dígitos conectados por lo que esto reducirá el problema a prácticamente la localización de la voz y el limpiado de la misma respecto del ruido. Este ruido podrá ser de muchos tipos (estacionario, no estacionario, armónico, inarmónico, etc.) y podrá provenir de muchas fuentes distintas (otras voces o sonidos, reverberaciones, filtrados, etc.) pero en cualquier caso provocará una distorsión de la señal de voz limpia. Denominaremos señal ruidosa o contaminada a la señal de voz limpia distorsionada por el ruido. Para simplificar nuestro problema, evitaremos el empleo de varias tomas de la señal ruidosa en varios puntos espaciales, es decir haremos reconocimiento robusto solo a partir de señal monofónica.

### Importancia del pitch

En reconocimiento monofónico con ruido, se presenta el problema de que la representación de la información de la voz limpia llega al reconocedor entremezclada con la del ruido y separar la información de esta respecto de la del ruido puede ser complicado. Como veremos, la mayoría de los procedimientos o técnicas que han abordado esta

separación lo han hecho empleando cierta información previa sobre la forma del ruido, sin embargo, cuando uno quiere abordar todos los ruidos posibles llega a la conclusión de que la información que más hay que tener en cuenta es la que ayuda a distinguir la voz respecto del ruido. Hay muchos tipos de pistas e informaciones que nos ayudan a distinguir ambas señales, pero al final la elección adecuada de estas pistas dependerá en gran medida de qué es lo que sea definido como voz.

La voz puede ser emitida de muchas formas dependiendo principalmente del tipo de «fuente principal» empleada. Estas formas pueden ser susurrantemente, con segundas voces musicales, etc. Aquí consideraremos que la voz es emitida de la forma habitual, es decir, con vibración de las cuerdas vocales. Se suele usar el término inglés «pitch» para hacer referencia a la correspondiente frecuencia de vibración (frecuencia fundamental).

Continuando con la búsqueda de las pistas más adecuadas de la voz que nos ayuden a distinguirla del ruido, en esta Tesis consideraremos específicamente al pitch por los tres motivos siguientes. El primero es que multitud de experimentos psicoacústicos como los de Darwin [33] muestran que el humano emplea el pitch no solo para distinguir y reconocer mejor una vocal respecto a un ruido inarmónico, también respecto a un ruido tipo armónico como puede ser otra vocal. Otro experimento psicoacústico que muestra la importancia del pitch es la capacidad que tenemos de reconocer el valor del pitch de la voz en altas condiciones de ruido y sin haber entendido nada de lo que se está hablando. Esto muestra que la localización del pitch es lo primero que hacemos antes de empezar el reconocimiento, por lo que consideramos que es la pista más primitiva de todas. El segundo motivo es que la mayor parte del tiempo, la voz emitida es periódica por lo que conocido el pitch, este puede ser empleado, aparte de para separar la voz sonora (con pitch) del ruido, también para localizar el resto de sonidos de la voz (los sonidos sordos y los silencios). El tercer motivo es que la mayor parte de las técnicas de reconocimiento automático robusto inspiradas en el ser humano, tales como las basadas en MD (Missing Data) [155], emplean el pitch como la pista principal para separar la voz del ruido, reforzando esto su importancia.

### Técnicas de robustecimiento basadas en el pitch

Cuando se intentan comparar las distintas técnicas de ASR robusto basadas en el pitch de forma justa, se evidencia la dificultad de esta tarea. Los motivos principales de esto son, entre otros, el que cada autor emplea un extractor de pitch diferente para evaluar su técnica, el no saber de donde proviene la fuente de la mejora (debido a otras técnicas

extras añadidas, o a al empleo de diferentes mecanismos de robustecimiento sobre los sonidos sordos y los silencios, siendo el de los sonoros el mismo), y el que muchas veces el autor no deja claro si se está proponiendo una nueva técnica para reconocimiento robusto, un nuevo extractor de pitch robusto o ambas cosas.

Debido a estos motivos vemos necesario hacer una comparación justa de las diferentes técnicas basadas en el pitch, tratando de ver las equivalencia entre las mismas y hasta donde podemos llegar a robustecer el reconocimiento de la voz conocido el pitch. Esta Tesis intentará resolver estas cuestiones.

Aparte de esto propondremos tres nuevas técnicas basadas en el pitch pero sin ocuparnos de la extracción del mismo, ya que consideramos que este es un aspecto importante para nosotros pero que queda fuera del alcance de esta Tesis.

Por último añadir dos cosas más. La primera es que dado que no estamos interesados en reconocimiento de alto nivel no emplearemos el pitch para reconocer la prosodia, ni tampoco para reconocer lenguas tonales como el Chino, pero en un futuro muchas de las técnicas e ideas presentadas en esta Tesis podrían ser empleadas para tal fin incluso para reconocimiento musical. Y la segunda es que, aunque las técnicas presentadas aquí no sirvan para voz sin pitch (p. ej. voz susurrante), tal y como veremos al final muchas de las ideas presentadas en esta Tesis pueden ser igualmente empleadas para el reconocimiento de este tipo de voz (Sec. 7.3.2).

### 1.1.2. Objetivos

Teniendo en cuenta las motivaciones anteriores, los objetivos principales de esta Tesis los podemos resumir de la siguiente forma:

1. Reconocer voz (o transcribir a texto) secuencias aleatorias de palabras conectadas y pronunciadas de la forma habitual (es decir con pitch) contaminadas por ruido a partir de señal monofónica.

2. Hacer un estudio comparativo de las diferentes técnicas de la bibliografía, tanto clásicas como basadas en el pitch, que robustecen el reconocimiento de la voz frente al ruido. Siempre intentándolas comparar con lo que se conoce sobre el reconocimiento humano.

3. Desarrollar y mejorar técnicas de robustecimiento de la voz basadas en el pitch que hagan las mínimas suposiciones posibles sobre el ruido. Para ello emplearemos

otras técnicas y esquemas de reconocimiento tales como sustracción espectral o MD (Missing Data, Datos Perdidos) que contribuyan a mejorar el rendimiento.

4. Mostrar la equivalencia entre algunas de estas técnicas basadas en el pitch, hacer una comparación justa de las mismas e intentar responder a la pregunta de hasta donde podemos mejorar el reconocimiento conocido el pitch.

### 1.1.3. Estructura de la Tesis

La Tesis aparte de esta introduccíon y los apéndices (entre los que se encuentran el resumen y las conclusiones en inglés), presenta siete capítulos más que se estructuran de la siguiente forma:

- Los capítulos segundo, tercero y cuarto de la Tesis son de fundamentos. En el segundo se estudia la voz y la audición humana. En el tercero se presentan los diferentes tipos de representaciones de la señal de voz, las máscaras de reconocimiento y los tipos de extractores de pitch que existen. En el cuarto los tipos de reconocedores de voz existentes, pero centrándonos y justificando el empleo de los reconocedores de MD basados en HMMs (Hidden Markov Models, Modelos Ocultos de Markov).

- En el quinto se estudian y comparan tanto algunas técnicas de robustecimiento convencionales como algunas de las técnicas basadas en el pitch encontradas en la bibliografía.

- En el sexto se proponen tres técnicas basadas en el pitch: ventanas asimétricas, autocorrelación cribada y estimación del ruido basada en el pitch.

- En el séptimo se muestran equivalencias entre las distintas técnicas y se intenta responder a la pregunta de cuales son los límites en el reconocimiento basado en el pitch.

- Por último en el octavo se resumen las conclusiones y aportaciones más importantes de esta Tesis y se bosquejan los trabajos futuros.

# Capítulo 2

# Fundamentos I: Voz y Audición

## 2.1. La voz

La voz es una secuencia de sonidos, generada por el aparato fonador humano (o por una imitación de este mediante una máquina), que codifica cierta información lingüística o un mensaje. En este trabajo únicamente revisaremos diversos aspectos de la señal de voz relevantes para el mismo. Para mas detalles consultar [134, 32, 109].

### 2.1.1. Elementos de la voz

Podemos decir que la voz está compuesta por tres tipos de elementos: los silencios (que aunque no son voz propiamente, si llevan información del mensaje), los sonidos sonoros y los sonidos sordos.

#### Sonidos sonoros

Los sonidos sonoros se caracterizan por tener una estructura temporal periódica (o cuasi periódica) y por lo tanto por tener un pitch (periodo o frecuencia fundamental). Se corresponden con las vocales, aunque también con ciertas consonantes como la «l» o la «m». Su espectro contiene una serie de armónicos separados aproximadamente la frecuencia del pitch. Este espectro puede considerarse como el producto de dos espectros, un tren de pulsos y una envolvente suave (envolvente espectral).

Consideraremos que la información lingüística portada por sonido sonoro reside en la envolvente espectral suave y no así en el tren de pulsos a la frecuencia fundamental (no consideramos lenguas tonales ni se presta atención a los aspectos prosódicos del mensaje). En particular serán de especial importancia la magnitud y posición de los picos

(formantes) de dicha envolvente. Menos importancia tendrán los valles de la envolvente, pues su profundidad puede variar considerablemente de unos locutores a otros [112].

La representación 2D de la posición de los dos primeros formantes en los sonidos vocálicos de un mismo locutor forman una curiosa forma denominada «triángulo vocálico» [134]. Este triángulo puede estar más arriba o abajo, o ser más pequeño o grande en función del locutor (si el pitch es más agudo suele estar más alto) pero siempre tenderá a tener un aspecto triangular.

### Sonidos sordos

Los sonidos sordos incluyen todo tipo de ruidos producidos por la voz y por lo general varían su energía espectral de forma rápida, estando esta concentrada a más de 2000 Hz.

Los sonidos sordos llevan la información del mensaje de texto en estas rápidas variaciones de las altas energías espectrales. Son más difíciles de clasificar pero se suele distinguir entre fricativos (de más larga duración, asociados a los sonidos de la «s» o la «f») y plosivos (de muy corta duración, asociados a los sonidos de la «p» o la «k»).

### Relación sonoro-sordo

Los sonidos sonoros y los sordos aparecen íntimamente relacionados. Los tres puntos siguientes lo muestran:

1. Aunque es cierto que existen sonidos que son mezcla de sonoros y sordos (tales como la «z» de la palabra inglesa«zip» o la «r» de «roble»), esta mezcla ocurre con tan poca frecuencia y rapidez que por simplicidad podemos considerar que nunca se da al mismo tiempo.

2. Podemos considerar que (en las lenguas de interés) los sonidos sordos nunca ocurren aisladamente [134] y que estos se encuentran como mucho a unos 0.2 segundos alrededor de los sonidos sonoros (antes o después).

3. Podemos considerar a los sonidos plosivos como inapreciables cuando el ruido es lo suficientemente fuerte (0 dB) y que lo que realmente nos da información de su existencia es la forma en la que se ataca o apaga (en las zonas de transito) el sonido sonoro. Esto es debido a la baja energía de los sonidos sordos en comparación con la de los sonoros.

Estas consideraciones son muy importantes pues permiten entre otras cosas localizar voz a partir del pitch (sonidos sonoros) y delatar la presencia de ciertos sonidos de la voz en función de otros cuando hay ruido.

### Unidades lingüísticas

Siguiendo la idea de búsqueda de unos elementos básicos o unidades lingüísticas de la voz que codifiquen el mensaje escrito combinando un conjunto finito de sonidos, las teorías clásicas (tal como la teoría de rasgos binarios de Jakobson [68]) han propuesto los «fonemas» como unidades básicas de la voz. Según estas teorías, los fonemas se diferencian claramente entre ellos por alguna característica acústica (como posición de los formantes o velocidad de cambio energético) o por alguna característica del modo y lugar en el que han sido articuladas en aparato fonador (si son plosivas alveolares o fricativas labiodentales, etc.).

Fenómenos como el de la coarticulación (que dan lugar a que se modifique la forma de pronunciar un fonema en función de los fonemas de alrededor) hacen que estás teorías no consigan diferenciar completamente las unidades de la voz debido a la gran variabilidad de formas en las que se pueden presentar los distintos fonemas (sobre todo los relacionados con las sonidos sordos), y debido a la dificultad de realizar una adecuada segmentación de las unidades en el tiempo [65].

Este tipo de dificultades han llevado a desechar el ideal de que los fonemas son las unidades básicas de la voz ([134, 156]) y a que los ASR de hoy en día tengan en cuenta estas tres consideraciones:

1. Usar las características dinámicas de velocidad y aceleración porque parte de la información lingüística se debe a como cambian las energías espectrales de la voz.

2. Emplear estructuras mayores como trifonemas e incluso palabras para definir las unidades de la voz.

3. Hacer el reconocimiento del texto y la segmentación temporal de las unidades lingüísticas al mismo tiempo.

Mencionar que las dos primeras consideraciones también las usan los sintetizadores de voz actuales.

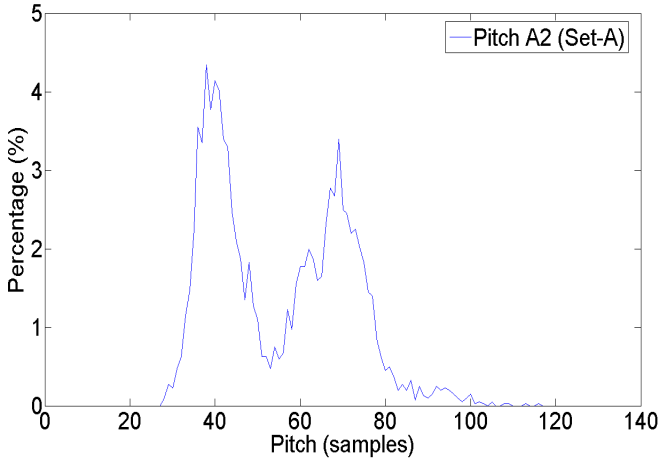



Figura 2.1: Histograma del pitch promedio de las frases limpias de conjunto Set-A de Aurora-2. Se observan dos modos, correspondientes a los distintos géneros.

### 2.1.2. El pitch

El pitch es el tono de los sonidos sonoros de la voz. Usamos la palabra inglesa pitch por su extenso uso en la jerga empleada en el campo de las tecnologías de la voz y porque esta significará para nosotros la función que nos indica en cada instante de tiempo el periodo o frecuencia fundamental de los sonidos sonoros, empleando el valor 0 o indefinido para señalar que en esos instantes la voz está en silencio o no es periódica.

El pitch de la voz humana suele ir variando a lo largo de una frase por varias razones, entre ellas la de poder expresar aun más información lingüística de la que se expresaría sin modular el pitch (información prosódica) y la de dar más robustez de entendimiento frente al ruido (p. ej. ayudándonos en la segmentación de las unidades lingüísticas como fonemas, palabras e incluso frases). A pesar de esta variación podemos decir que cada hablante suele hablar en torno a un pitch promedio (determinado por la longitud de sus cuerdas vocales) y que será más agudo para las voces de las mujeres y niños, y más grave para las voces de los hombres. En la Fig. 2.1 podemos observar el histograma del periodo de pitch promedio (en número de muestras, para una frecuencia de muestreo de 8000 Hz) de las diferentes frases limpias del conjunto Set-A de Aurora-2 (Set-A posee 4004 frases limpias, Sec. A.2). Podemos ver que el pitch humano se suele encontrar en el intervalo [30,100] muestras ([80,270] Hz). También podemos apreciar que hay dos grandes grupos de locutores, los que tienen un pitch agudo o femenino con media de 40 muestras (200 Hz), y los que tienen un pitch grave o masculino con media 69 muestras (116 Hz). El pitch promedio de Aurora-2 está alrededor de las 55 muestras (145 Hz).

### 2.1.3. Modelos de la voz

#### Modelo de fuente principal

Teniendo en cuenta la forma en la que se combinan los distintos elementos de la voz (silencios, sonoros y sordos) consideraremos el siguiente «modelo de fuente principal» de la voz:

***La voz es una señal de excitación o fuente principal que puede ser modulada espectralmente y en intensidad, y que a veces, y solo cuando esta fuente principal es apagada, puede ser sustituida por cortas señales correspondientes a ruidos.***

En el caso de voz emitida de la forma habitual la fuente principal puede ser considerada como una señal periódica (posee un pitch que es producto de la vibración de las cuerdas vocales) sin embargo, en voz de tipo susurrante (en el que las cuerdas vocales no intervienen [159]) la fuente principal puede ser considerada como un ruido. Las cortas señales de ruido se corresponden con los sonidos sordos. Este modelo es una definición simplificada de la voz que será empleada para proponer un VAD (Voice Activity Detector) basado en el pitch.

#### Modelo de producción de voz

El «modelo simplificado de producción de voz» se inspira en el modelo de fuente principal para generar casi cualquier secuencia de sonidos que produzca el aparato fonador [43, 111]. En la Fig. 2.2 mostramos su esquema. Para producir voz basta indicar en ciertos instantes de tiempo el valor de cada uno de los parámetros del modelo: periodo de pitch, decisión sonoro/sordo, ganancia y tipo de filtro (normalmente todo polos con 10-12 polos). Esto nos da un total de unos aproximadamente 15 parámetros que varían en el tiempo para transportar prácticamente toda la información referente a la voz. Mencionar que muchos codificadores y sintetizadores de voz se basan en este modelo o en modificaciones del mismo [65] para sus respectivos propósitos.

## 2.2. Audición

### 2.2.1. Sistema auditivo

El sistema auditivo se puede dividir en dos partes. El oído o sistema periférico, que se encarga de transducir al nervio auditivo características acústicas de la onda sonora

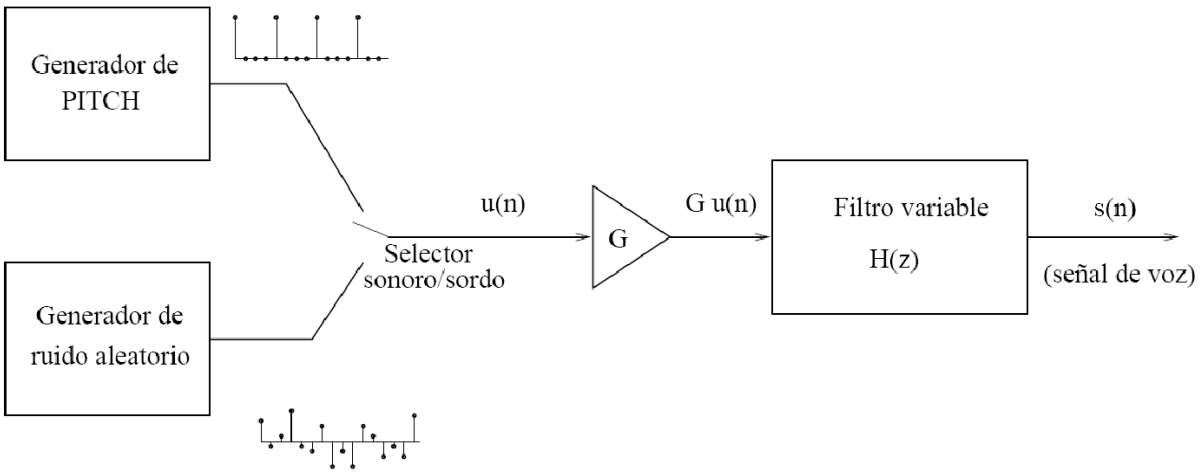



Figura 2.2: [109] Modelo de producción de voz. La fuente principal es el generador de pitch que produce los sonidos sonoros. En determinados y cortos instantes de tiempo esta fuente es sustituida por el generador de ruido para producir los sonidos sordos.

de entrada (principalmente la intensidad de cada frecuencia) en impulsos nerviosos. Y el sistema central, que se encarga de recoger y reconocer estos impulsos nerviosos.

A continuación explicaremos de forma breve las partes mas interesantes del sistema auditivo en relación al reconocimiento. Para más información consultar [122, 102, 55].

### Oído externo y medio

El oído se puede dividir en externo, medio e interno. El oído externo y medio se encargan de convertir, a través del pabellón aricular, el tímpano y los huesecillos (ver Fig. 2.3), las variaciones de presión sonora en variaciones de movimiento mecánico del liquido que llena la cóclea (la perilinfa). La señal de movimiento resultante al pasar a través de estos es amplificada en las altas frecuencias.

### Oído interno: membrana basilar

El oído interno contiene la cóclea, que desenrollada, no es más que un tubo dividido en tres regiones (vestibular, media y timpánica) por medio de dos membranas (basilar y de Reissner, Fig. 2.4). Las escala vestibular y la timpánica están conectadas y rellenas con la perilinfa, la cual al moverse produce un movimiento en la membrana basilar. La membrana basilar posee la característica de que va incrementando su tensión gradualmente. Esto provoca que cuando la señal de entrada sea un seno, se produzca una onda viajera a lo largo de la membrana basilar, haciendo que todos los puntos de la membrana basilar

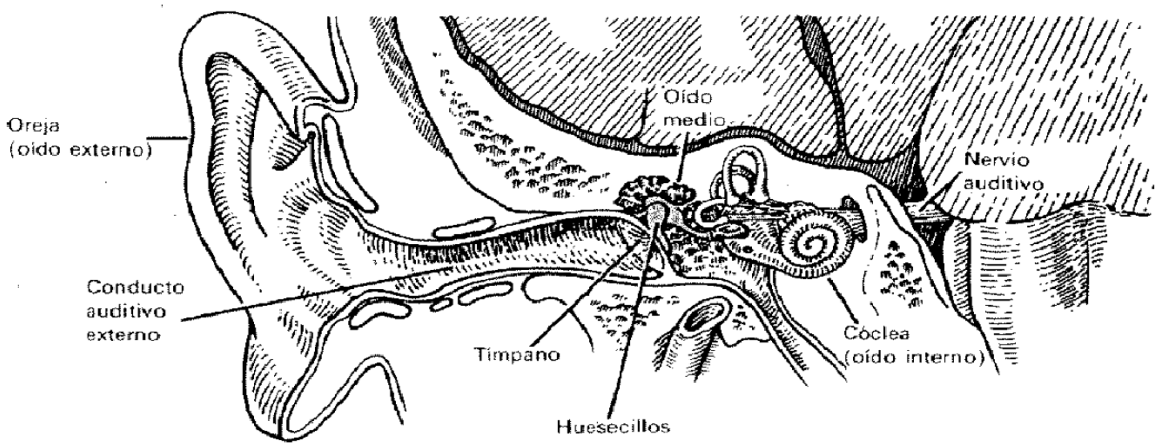



Figura 2.3: [109] El oído, compuesto por la oreja o pabellón auricular, tímpano, huesecillos, cóclea y nervio auditivo.

vibren a la frecuencia de entrada, aunque alcanzándose un máximo de amplitud en un único punto. Este punto máximo es único para esa frecuencia (organización tonotópica) por lo que, teniendo en cuenta que el principio de superposición también se da en la membrana basilar, se puede considerar a esta como si fuera un analizador de Fourier aunque con ciertas limitaciones. La limitación más importante es que no resuelve las frecuencias por igual, disminuyendo la resolución con el logaritmo de la frecuencia. Esto implica que si la señal de entrada son dos senos muy cercanos (no resolubles) la membrana basilar vibrará con un solo máximo, llegándose a oír solo el más «fuerte» de los dos (Sec. 2.2.4).

### Oído interno: pulsos nerviosos

La membrana basilar, al moverse de arriba a abajo provoca el movimiento de los estereocilios que están unidos a las células ciliadas internas y en consecuencia, el disparo de pulsos en la fibra nerviosa correspondiente. Debido a que la acción potencial del disparo solo se inicia en una dirección, solo se tiene en cuenta media onda del movimiento. También, debido al enganche de fase (phase locking) de las células ciliadas internas, solo se emite un pulso cada vez que el movimiento pasa por un punto. Esta emisión no tiene porque producirse cada ciclo, si no más bien cada múltiplo entero del ciclo. Si la amplitud de la señal es muy grande es más probable que se emita un pulso cada ciclo. La colección de los diferentes disparos, de las diferentes células ciliadas internas, es recogida en la fibra

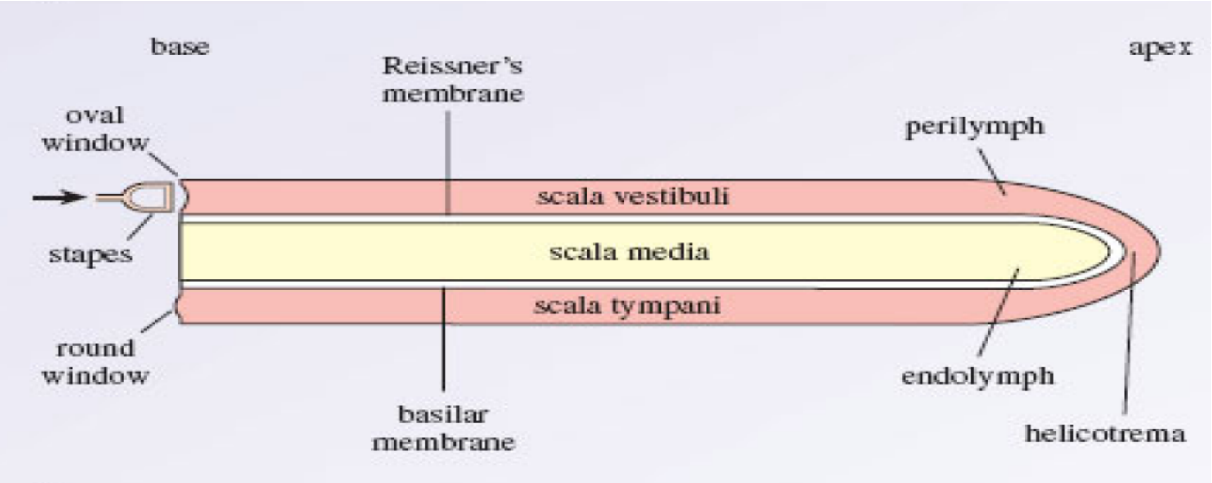



Figura 2.4: [78] Cóclea desenrollada dividida en tres regiones: vestibular, media y timpánica.

nerviosa correspondiente a esa frecuencia. La intensidad que se percibirá a esa frecuencia estará relacionada con el número de disparos por segundo (velocidad de disparo).

El nervio auditivo esta formado por las diferentes fibras nerviosas estando las fibras de las altas frecuencias en la periferia del nervio auditivo y las de las bajas frecuencias hacia el centro del nervio auditivo (se sigue manteniendo la organización tonotópica de la membrana basilar [71]).

Otros detalles interesantes sobre el funcionamiento de la cóclea, a tener en cuenta son: el papel de las células ciliadas externas en el control del movimiento basilar (ordenado por el sistema central [133, 130] y que hacen que la transducción del sonido a impulsos eléctricos no solo dependa de las propiedades físicas de la señal recibida), las saturaciones y no linealidades que aparecen en los diferentes niveles (saturación del movimiento de la membrana basilar o del ritmo de disparo, etc.) y el aumento repentino en la velocidad de disparo al producirse zonas de tránsito (zonas de cambios bruscos de amplitud).

### Sistema central

A medida que nos adentramos más en el sistema auditivo, más desconocido es este, siendo la forma en que opera el sistema central lo más desconocido. A pesar de esto, se pueden diferenciar las siguientes estructuras neuronales interconectadas entre sí: nervio auditivo, núcleo coclear, oliva superior, colículo inferior, núcleo geniculado medial y cortex auditivo. La mayoría de las reglas que se conocen sobre como reconoce el sistema central

provienen de experimentos en percepción auditiva. Un ejemplo de tales reglas son las propuestas por el esquema ASA (Auditory Scene Analysis, ver Sec. 2.2.6).

### 2.2.2. Filtros auditivos

#### Fundamentación

Diferentes resultados experimentales han llevado a la conclusión de que el oído computa (teniendo en cuenta el principio de equivalencia computacional [161]) la señal de entrada como si de un banco de filtros se tratase, donde a cada filtro se le denomina filtro auditivo y su anchura es función del logaritmo de la frecuencia. La salida de este banco de filtros viene codificada en el nervio auditivo (Sec. 2.2.1).

Entre los experimentos de percepción más destacados que justifican esto podemos mencionar: Los de Fletcher [45] y Zwicker [166] que supusieron la existencia de bandas críticas para explicar el enmascaramiento de un tono sobre ruido pasa-banda (Sec. 2.2.4). Los de Patterson que dieron con la forma exacta de los filtros auditivos ([119]) y los de Moore ([102]), que mediante el empleo de bancos de filtros, ha conseguido fusionar las dos teorías sobre percepción del pitch (temporal y del lugar, Sec. 2.2.5).

Entre los experimentos fisiológicos más destacados podemos mencionar los de Beckesy ([13]), que midieron el movimiento de la membrana basilar y los de Liberman ([81]) que obtuvieron las curvas de disparo, en función de la frecuencia, para una sola neurona, concluyendo que esta se dispara principalmente para frecuencias que estén dentro de su filtro auditivo correspondiente.

#### Forma del filtro y escalas auditivas

Patterson [119] dedujo la forma del filtro auditivo humano, mediante medidas del enmascaramiento entre un tono y un ruido rechazabanda situado alrededor de ese tono y del cual se fue variando su anchura de rechazo. En la Fig. 2.5 de la izquierda podemos observar la forma de un filtro auditivo a la frecuencia de 1000Hz. Dado que la forma exacta de este puede variar de unos oyentes a otros y de unas condiciones físicas a otras, se prefiere indicar su forma mediante el ERB (Equivalent Rectangular Bandwidth, Ancho de banda Rectangular Equivalente). El ERB de un filtro es la anchura que debe de tener un filtro rectangular (con la misma altura que el original) para que su area sea equivalente

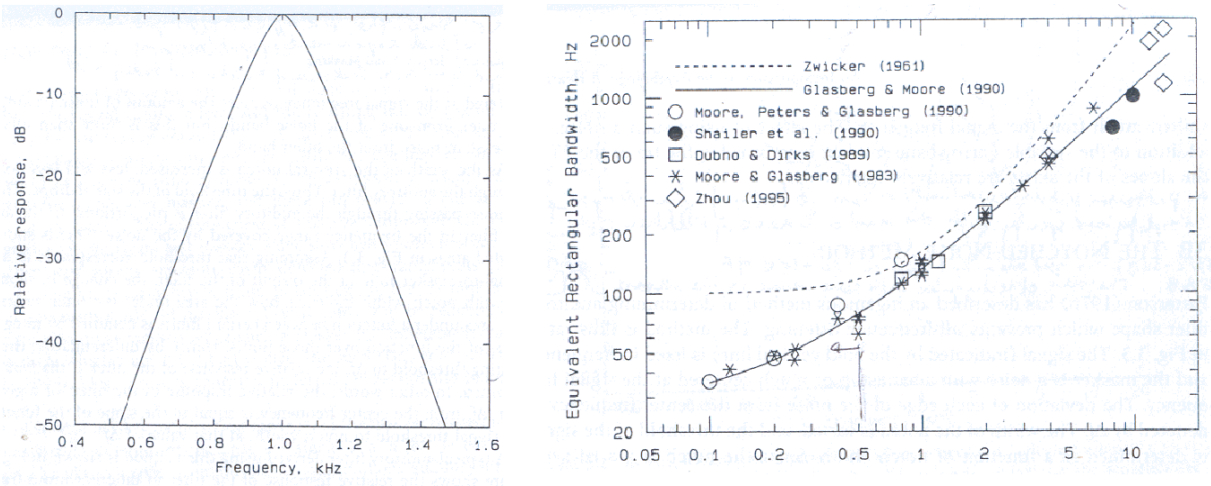



Figura 2.5: [102] Izquierda, forma de un filtro auditivo obtenida mediante el experimento de Patterson con frecuencia central de 1000 Hz. Derecha, función ERB de Glasberg y Moore y otras funciones y estimas relacionadas.

a la del original. Glasberg y Moore [49] han resumido en la siguiente ecuación el ERB promedio de muchos oyentes 'normales' en función de la frecuencia:

$$ERB(f) = 24{,}7(4{,}37 \cdot 10^{-3} f + 1) \tag{2.1}$$

donde $ERB$ y $f$ están expresadas en Hz. A esta ecuación se le conoce con el nombre de «función ERB». En la Fig. 2.5 de la derecha se aprecia esta función junto a los diferentes experimentos realizados para estimarla.

Una escala derivada del ERB y la cual resulta muy útil, para tener una idea del patrón de excitación que produce una señal en la membrana basilar [101], es la «escala ERB»:

$$ERB_{number}(f) = 21{,}4 log_{10}(4{,}37 \cdot 10^{-3} f + 1) \tag{2.2}$$

Esta nos indica el número ERB ($ERB_{number}$) en función de la frecuencia $f$ en Hz. Un incremento de un $ERB_{number}$ se corresponde con un incremento de 0.9 mm en la membrana basilar. Esta escala es similar a otras escalas auditivas como la Bark de Zwicker [167] y la Mel de Steven [144].

### 2.2.3. Filtro gammatone

Un filtro gammatone ([69]) es un filtro pasabanda simétrico que se define mediante su repuesta impulsiva de la siguiente manera:

$$g(t) = at^{n-1}cos(2\pi ft + \phi)e^{-2\pi bt} \quad (t > 0) \tag{2.3}$$

donde $a$ es la amplitud; $n$ es orden del filtro el cual determina la pendiente de caída de la falda del filtro; $f$ es la frecuencia central del filtro; $\phi$ es la fase y $b$ el ancho de banda del filtro (a -3dB) el cual determina la duración de la respuesta impulsiva. La importancia de estos filtros para la audición reside en que, como han mostrado Patterson y Moore [118], pueden generar una respuesta en frecuencia muy parecida a la de los filtros auditivos humanos obtenidos de forma perceptual por Patterson (Sec. 2.2.2). Es más, son capaces de indicarnos en cierta manera como se mueve la membrana basilar frente a un estimulo dado (experimentos similares a los de von Békésy [13] de observación del movimiento de la membrana basilar lo confirman [102]). Para que esto ocurra se suele tomar $n = 4$ y un $b = 1{,}019ERB(f)$. Con estos valores se consigue que cada filtro gammatone tenga, al menos, el mismo ERB que el del oído humano y un movimiento basilar parecido. Existe una implementación rápida del filtro gammatone (ver [63, 29]).

### 2.2.4. Enmascaramiento auditivo

#### Definición

Existen dos formas básicas de enmascaramiento: no simultaneo (que se produce cuando un sonido impide que otro se oiga, estando los dos separados temporalmente aunque muy cercanos en el tiempo) y frecuencial. El enmascaramiento frecuencial, se da cuando habiendo un sonido enmascarador con una determinada frecuencia (normalmente un tono puro o un ruido paso-banda), no se oye otro tono o banda de ruido objetivo cercano a la máscara. Existen multitud de experimentos que muestran el enmascaramiento frecuencial [102]. Muchos de estos experimentos se han usado para determinar la forma de los filtros auditivos tal y como hemos visto.

#### Causas fisiológicas

Respecto a las causas fisiológicas que lo producen se puede decir que son varias las que contribuyen a este fenómeno. Las más importantes de todas son la limitada resolución

frecuencial de la membrana basilar y su no linealidad de respuesta frente a la amplitud de la señal de entrada (Sec. 2.2.1). Otra causa que además muestra que la cóclea es activa, es la que se deriva del experimento de «supresión de dos tonos» [132]. De este experimento se deduce que la misma cóclea es capaz de disminuir e incluso suprimir el ritmo de disparo de la neurona correspondiente a la frecuencia que esta siendo enmascarada, cuando en principio el movimiento de la membrana basilar permitiría su escucha sin problemas.

#### Principios del enmascaramiento y aproximación log-max

El enmascaramiento puede quedar resumido bajo estos dos principios:

1. que el oído actúa como si de un banco de filtros auditivos se tratase
2. que la intensidad percibida en un filtro auditivo (representada mediante la velocidad de disparos neuronales, Sec. 2.2.1) es el logaritmo (o una función similar como la raíz cúbica) de la suma de las distintas amplitudes que llegan al filtro.

Veamos con un ejemplo como estos dos principios producen enmascaramiento. Sea una senoidal (o ruido) de amplitud (o desviación típica) $A_1$ que entra junto con una senoidal (o ruido), cercano en frecuencia, de amplitud $A_2$ en un mismo filtro (principio uno). La intensidad total $I_{1+2}$ que se percibirá en ese banco será la siguiente según el principio dos:

$$I_{1+2} = log(A_1 + A_2) \approx max(log(A_1), log(A_2)) = max(I_1, I_2) \tag{2.4}$$

donde vemos que debido a la aproximación log-max ([91, 150]) lo que se percibirá será la intensidad de la señal más fuerte quedando la débil enmascarada. Esta importante aproximación será la que justifique el empleo de las técnicas de missing data para el reconocimiento robusto de la voz (Sec. 4.2.2).

### 2.2.5. Percepción del pitch

#### Primeras teorías

Las dos clases de teorías que durante mucho tiempo han intentado explicar la percepción del pitch tal y como se explica en [101, 156] han sido: las espectrales [153, 50] (que resaltan el papel de los armónicos resolubles por la membrana basilar), y las temporales [136, 66] (que resaltan el papel de los armónicos sin resolver). Según las teorías espectrales el cerebro obtiene el pitch a partir del patrón que se produce en la membrana basilar de

los armónicos más energéticos y resueltos por la misma. Según las teorías temporales el cerebro obtiene el pitch a partir de la forma de onda creada en la membrana basilar por los armónicos más energéticos y no resueltos por la misma. Esta forma de onda se corresponderá con una modulación AM debido a la interacción entre los distintos armónicos en la membrana basilar (Sec. 3.3).

Cada teoría explicaba unos cuantos experimentos perceptivos pero dejaba otros tantos sin explicar. Las espectrales no podían explicar los siguientes experimentos: la percepción de pitch cuando solo se presentan armónicos agudos e irresolubles y la percepción de pitch cuando se multiplica una senoidal por un ruido blanco (en este caso el módulo del espectro largo (long-term spectrum) es plano y la membrana basilar no presenta máximos). Las temporales no podían explicar los siguientes experimentos: la dominancia por los armónicos bajos y resolubles en la percepción del pitch y la percepción binaural de pitch cuando se presenta un ruido blanco a cada oído variándose aleatoriamente la fase de una banda de frecuencias de uno de los ruidos (en este caso no aparece ninguna información temporal de máximos en la vibración de la membrana basilar).

Estos experimentos han llevado a la siguiente conclusión: el pitch no se procesa en el oído si no en zonas del sistema central por lo que la información temporal de la fase debe ser mantenida por los impulsos del nervio auditivo hasta llegar al sistema central.

#### Teoría espectro-temporal

Todo esto ha provocado el nacimiento de las teorías espectro-temporales [82] basándose en los conocimientos que se tienen sobre como el oído separa la señal en un banco de filtros y la transduce en impulsos nerviosos (Sec. 2.2.1). Un ejemplo de estas es la propuesta por Moore que consta de dos etapas [101]. En la etapa temporal de la teoría de Moore, el sistema central hace una especie de histograma para encontrar el intervalo más frecuente entre pulsos nerviosos dentro de un mismo canal frecuencial (o nervio auditivo). En la etapa espectral se escoge el intervalo más frecuente a lo largo de los diferentes canales, siendo este intervalo el periodo de pitch percibido.

### 2.2.6. Análisis de Escenas Auditivas

#### Definición

El ASA (Auditory Scene Analysis, Análisis de Escenas Auditivas) es un campo de la psicoacústica que se basa en analizar la percepción auditiva, siguiendo una serie de reglas,

de forma similar a como se analiza una escena visual. El padre fundador de ASA ha sido Bregman [16] y sus reglas están muy relacionadas con las de la visión propuestas por los psicólogos de la Gestalt [114]. Su éxito se ha debido a que las reglas de ASA se han podido implementar computacionalmente mejorando diversas aplicaciones tecnológicas como el ASR en condiciones de ruido, la transcripción musical o las prótesis auditivas. Esta implementación computacional se conoce con el nombre de CASA (Computational ASA, ASA Computacional) [155] y está ayudando al mismo tiempo a mejorar la comprensión del ASA.

### Esquema

Veamos con un ejemplo, como trabaja ASA para producir el reconocimiento de una frase contaminada por ruido:

1) Se proporciona al sistema central, mediante las transducciones del oído, una representación de la escena auditiva denominada cocleograma (similar al espectrograma, Sec. 3.1.2) que se compone de «píxeles» frecuencio-temporales. En esta escena habrá píxeles dominados por la voz y otros por el ruido.

2) Se aplican «reglas primitivas» (botton-up, abajo-arriba) de agrupación (o segmentación) de píxeles creando segmentos, grupos, fragmentos, etc. (según el tamaño o la regla empleada se le suele dar un nombre diferente a la agrupación) que provienen de una misma fuente. Las reglas primitivas son reglas innatas. Algunos ejemplos son: «agrupar píxeles con pitch común», «agrupar píxeles con comienzo/final común», etc. (ver [155] para ver implementaciones computacionales de estas reglas).

3) Se aplican «reglas basadas en modelos» (top-down, arriba-abajo) para agrupar los fragmentos que sean de la voz. Las reglas basadas en modelos son aprendidas. Un ejemplo de tales reglas para el reconocimiento musical es: «agrupar los fragmentos que encajen dentro del patrón rítmico esperado y desechar el resto». Para el caso del reconocimiento de la voz, agrupación e identificación de palabras se hacen al mismo tiempo (se prueban patrones de palabras que ayudan a agrupar fragmentos y al mismo tiempo, se elige la palabra que mejor encaje con los fragmentos existentes). Esto significa que al reconocer se aplica un «método de pizarra» (blackboard, ver SFD en la Sec. 5.1.6 para ver la implementación computacional de este método).

# Capítulo 3

# Fundamentos II: Representaciones, Máscaras y Extractores de Pitch

## 3.1. Representaciones acústicas

### 3.1.1. Definición y notación

#### Representación acústica

Los sistemas de reconocimiento intentan reducir la cantidad de información de la señal de voz antes de enviarla al reconocedor. Para ello se elimina información redundante y se intentan extraer las características más posiblemente relacionadas con el mensaje texto (Sec. 2.1). Una representación acústica de una señal es una matriz 2D, que nos informa sobre diferentes tipos de características acústicas en cada instante de tiempo. La representación acústica la obtienen los FE (Feature Extractor, Extractores de Características) y la usan los reconocedores para decodificar la señal de entrada. El más claro ejemplo de representación acústica es el espectrograma, pero hay otras muchas representaciones tales como el cepstrograma, el formantograma o el de los parámetros del modelo de producción de voz.

#### Notación

Usaremos la siguiente notación para referirnos a los distintos elementos de la representación acústica: La *matriz de características* $X(c,t)$ es la representación acústica en su conjunto. Un *canal de información* $c$ es cualquiera de las filas de la representación acústica. Se encarga de informar sobre una determinada característica acústica, como por

ejemplo puede ser la energía de una banda de frecuencias (si es un canal espectral) o la cantidad de sonoridad de la señal (si es un canal de sonoridad). Un *vector de características* es una de las columnas de la representación acústica y nos informa sobre las diferentes características de un segmento (o trozo) de señal en un instante de tiempo dado $t$. Un *coeficiente*, o simplemente *píxel*, es cualquier elemento de la representación acústica en un canal $c$ y en un instante $t$ (p. ej. coeficiente espectro-temporal). Dicho ésto, pasemos a estudiar las tres representaciones acústicas que emplearemos en esta Tesis: cocleograma, espectrograma y cepstrograma.

### 3.1.2. Cocleograma

El objetivo de un cocleograma es representar, en cada instante de tiempo y de la manera más fielmente posible el ritmo de disparo de cada nervio auditivo que sale de la cóclea. Este tipo de representación acústica, teniendo en cuenta el funcionamiento del oído (Sec. 2.2.1), nos informa sobre la energía de las diferentes frecuencias de la señal de entrada.

Se han propuesto diferentes modelos computacionales de cocleograma, con diferentes niveles de detalle del oído en su conjunto. En general más que ser modelos detallados del oído son más bien modelos funcionales que tienden a imitarlo solo en algunas partes (p. ej. una de las funciones más difíciles de imitar del oído es la de las células ciliadas externas en la cóclea). El modelo propuesto por Meddis [96, 97, 99] y en el cual se inspira el cocleograma que explicaremos, es un ejemplo de modelo detallado del oído. Otros modelos propuestos se pueden consultar en [87, 141]. El cocleograma que explicaremos aquí es un **Log-Gamm-Cocleograma** (Cocleograma gammatone con compresión Logarítmica) [91].

Para la obtención del Log-Gamm-Cocleograma tendremos en cuenta las siguientes características del oído: que las altas frecuencias son aumentadas por el oído externo y medio, que la membrana basilar actúa como un banco de filtros no linealmente distribuidos, y que el movimiento de cada filtro se traduce en el nervio auditivo en un ritmo de disparo dependiente de la amplitud del movimiento. La obtención del Log-Gamm-Cocleograma la podemos resumir en las dos etapas siguientes: banco de filtros y suavizado-muestreo.

#### Banco de filtros gammatone

La señal muestreada es pasada a través de un banco finito de filtros o canales gammatone, distribuidos equitativamente en la escala ERB y cuya anchura de banda crece con

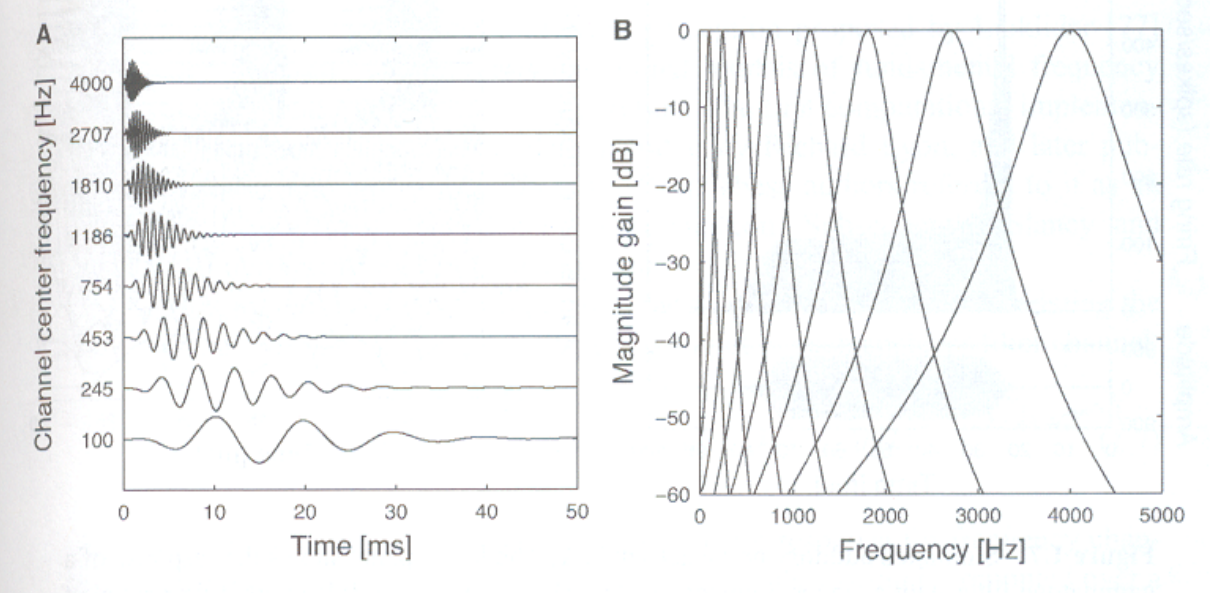



Figura 3.1: [155] Banco de filtros gammatone. Izquierda, respuestas impulsivas de los filtros. Derecha, respuestas en frecuencia de los filtros.

la frecuencia según la función ERB (ver Sec. 2.2.2). Esta elección se hace así, porque es una forma de simular el movimiento de la membrana basilar (ver Sec. 2.2.3).

En la Fig. 3.1 derecha, podemos apreciar un banco con 8 filtros gammatone. Se aprecia como se solapan entre sí. Aunque en la figura no se muestre, lo normal es que la ganancia de los filtros crezca según la frecuencia para imitar el comportamiento de realce de las altas frecuencias producido por el oído externo y medio. Si colocamos en filas las distintas salidas de los filtros gammatone obtenemos lo que denominaremos matriz de movimiento basilar. En la Fig. 3.1 izquierda, podemos observar una representación de esta matriz para un impulso unitario (respuestas impulsivas de los distintos filtros).

Al tomar un numero finito de filtros lo que obtenemos es, en realidad, un muestreo del movimiento de la membrana basilar en distintos puntos. Esto puede dar la sensación de que la matriz de movimiento basilar no transporta toda la información que usa el ser humano para reconocer. Sin embargo debido a que se solapan los filtros entre si, y a que en verdad el oído sufre de enmascaramiento frecuencial (Sec. 2.2.4) la matriz de movimiento basilar transporta prácticamente toda la información que usa el ser humano para reconocer.

La cantidad de filtros usados dependerá de la frecuencia de muestreo de la señal y del hecho empírico de que para reconocimiento lo aconsejable es tomar unos 3 filtros por octava. Esto nos da 32 filtros para una frecuencia de muestreo de 8 kHz que es la que

suelen emplear muchos cocleogramas. Usar más conlleva más coste computacional sin prácticamente ganancia en el reconocimiento.

### Suavizado-muestreo

Para obtener el ritmo de disparo de cada nervio auditivo aplicaremos un suavizado a cada canal de la matriz de movimiento basilar y luego un muestreo temporal. Este suavizado-muestreo tiene las siguientes cuatro etapas que se justifican viendo como opera el oído (Sec. 2.2.1).

1) Rectificado de media onda para simular el hecho de que las células ciliadas internas de los nervios auditivos solo se disparan en una dirección del movimiento de la membrana basilar.

2) Extracción de la envolvente mediante la transformada de Hilbert (Modulación AM/FM [155]) y suavizado mediante filtrado lineal paso-baja de primer orden con una constante de tiempo de 8 ms para obtener el ritmo de disparo del nervio auditivo de forma proporcional a la amplitud de vibración.

3) Muestreo temporal cada 10 ms para reducir la cantidad de información con la que trabajar. Mencionar que la matriz resultante en este punto es similar (salvo por una constante y quizás número de canales) a la matriz Mel-Espectrograma del espectrograma por lo que a esta matriz le denominaremos Gamm-Espectrograma.

4) Comprensión mediante la función logaritmo neperiano para imitar la comprensión en el ritmo de disparo con la amplitud, Mencionar que para imitar la saturación en el ritmo de disparo (Sec. 2.2.1) se suele limitar el valor mínimo que puede dar la función logaritmo.

El resultado final de este suavizado es la matriz Log-Gamm-Cocleograma también conocida como ratemap (mapa de disparos). A pesar de que en este tipo de cocleograma faltan muchos detalles para imitar con exactitud el ritmo de disparo de los nervios auditivos (como la saturación o el aumento en el ritmo de disparo en las zonas de transito [99]) se puede decir que el Log-Gamm-Cocleograma es una buena aproximación al ritmo de disparo. Por razones de mejora en las tasas de reconocimiento, es habitual complementar esta representación con las velocidades de los vectores cocleares (obtenidas por medio de derivadas discretas entre vectores de características cercanos en el tiempo). En la Fig. 3.2 podemos ver un ejemplo de Log-Gamm-Cocleograma para una señal de voz limpia.

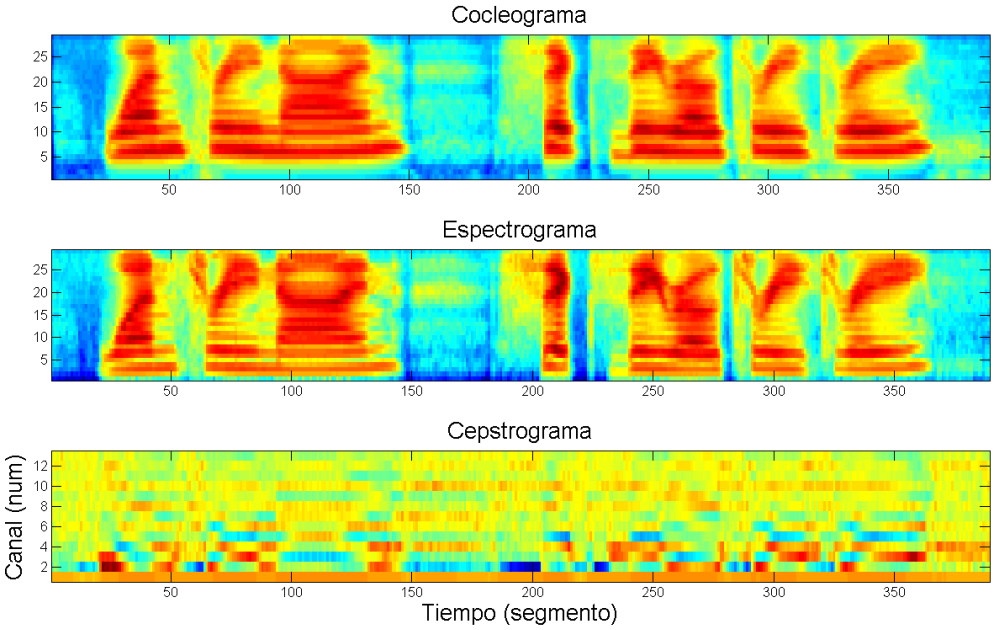



Figura 3.2: Comparación de las tres representaciones acústicas para una señal de voz limpia: Cocleograma (Sec. 3.1.2), Espectrograma (Sec. 3.1.3) y Cepstrograma (Sec. 3.1.4).

### 3.1.3. Espectrograma

El objetivo de un espectrograma es representar, en cada instante de tiempo, la energía de las diferentes frecuencias de la señal de entrada. Existen diferentes variantes del espectrograma (transformada de Fourier de tiempo corto, espectrograma dB, etc.). El espectrograma que explicaremos aquí es un *Log-Mel-Espectrograma* (Espectrograma en la escala Mel con compresión logarítmica) y el cual se obtiene a partir de las indicaciones del extractor de características FE de la ETSI [149].

La obtención del Log-Mel-Espectrograma imita en algunas partes al oído humano, pero en otras, procesa la señal de manera que su justificación no es más que la de dar buenos resultados de reconocimiento. Podemos resumir su obtención en las dos etapas siguientes: preprocesamiento-segmentación y Log-Mel-espectro.

#### Preprocesamiento-segmentación

En primer lugar la señal muestreada ($s(n)$) es preprocesada usando dos filtros: un eliminador de offset (que elimina la componente continua):

$$s_{of}(n) = s(n) - s(n-1) + 0{,}999 s_{of}(n-1) \tag{3.1}$$

y un filtro de preénfasis (que imita el efecto del oído externo y medio de aumentar las altas frecuencias):

$$s_{pe}(n) = s_{of}(n) - 0{,}97 s_{of}(n-1); \tag{3.2}$$

después de esto, la señal $s_{pe}(n)$ es segmentada en trozos (o segmentos) que se solapan entre ellos. Valores típicos de esta segmentación son: $N = 32ms$ (longitud del segmento) y $FS = 10ms$ (desplazamiento entre segmentos). El resultado de esto es la matriz de segmentación (para entendernos, cada columna representará un segmento de señal).

### Log-Mel-espectro

Para cada segmento se estima la magnitud de la densidad espectral discreta (con un número de puntos entre 0 y $2\pi$ por lo general igual a $N$), obteniéndose la matriz de densidad espectral. La densidad espectral puede ser estimada mediante la transformada de Fourier $\mathcal{F}$ de tres formas diferentes [123, 2]:

$$M_x(\omega) = \frac{\sum_{n=0}^{N} x(n) w_x(n) e^{-i\omega n}}{\sqrt{N}} = \frac{\mathcal{F}[x(n)w_x(n)]}{\sqrt{N}} \tag{3.3}$$

$$M_{r_x}(\omega) = \sqrt{\mathcal{F}[r_x(k) w_{r_x}(k)]} \tag{3.4}$$

$$M_{ARMA}(\omega) = \sigma_e^2 \frac{\mathcal{F}[\vec{b}]}{\mathcal{F}[\vec{a}]} \tag{3.5}$$

desde el segmento (espectro directo a través de un enventanamiento, normalmente Hamming, Ec. 3.3), desde la autocorrelación (periodograma a través de un enventanamiento, normalmente Káiser, Ec. 3.4) o desde los parámetros ARMA (espectro que por lo general no necesita enventanamiento, Ec. 3.5). Cabe mencionar que la matriz de densidad espectral no es más que una transformada de Fourier de tiempo corto normalizada y muestreada cada $FS$ (ver [152, 65]).

Cada vector de densidad espectral es pasado a través de un «banco de filtros» con un número finito de canales distribuidos equitativamente en la escala Mel obteniéndose la matriz Mel-Espectrograma y que presenta bastante similitud con la matriz Gamm-Espectrograma del cocleograma. En verdad, no se trata de un banco de filtros como tal. Esto es debido a que lo que en verdad se hace es multiplicar las diferentes componentes en frecuencia por un conjunto de ventanas triangulares distribuidas logarítmicamente, por lo que es más bien un suavizado del espectro. Se elige esta distribución porque relaciona la

distinción de tonos con la frecuencia y por lo tanto la distribución de los filtros auditivos humanos con la frecuencia (Sec. 2.2.2).

Finalmente cada elemento de la matriz Mel-Espectrograma es comprimido con la función logaritmo neperiano, para simular la forma en que el ser humano percibe la intensidad a las diferentes frecuencias (Sec. 2.2.2), obteniéndose la matriz Log-Mel-Espectrograma. Hay que mencionar que, por razones de mejora en las tasas de reconocimiento, se suele limitar el valor mínimo que puede dar la función logaritmo y que es habitual complementar esta representación con las velocidades de los vectores espectrales. En la Fig. 3.2 podemos ver un ejemplo de Log-Mel-Espectrograma para una señal de voz limpia.

### 3.1.4. Cepstrograma

El objetivo de un cepstrograma es representar, en cada instante de tiempo, los valores de las diferentes componentes cepstrales de la señal de entrada. Existen diferentes variantes del cepstrograma (cepstrograma-LPC, cepstrograma-IFFT, etc.). El cepstrograma que explicaremos aquí es un *Log-Mel-Cepstrograma* (Cepstrograma en la escala Mel con compresión Logarítmica) que se obtiene a partir de las indicaciones del extractor de características de la ETSI [149].

El Log-Mel-Cepstrograma se obtiene aplicando una simple DCT (Discrete Cosine Transform, Transformada Discreta del Coseno) con $NDCT$ puntos a cada vector de la matriz Log-Mel-Espectrograma explicada anteriormente.

A la parametrización obtenida se la conoce como MFCC (Mel-Frequency-Cepstral-Coeficients, Coeficientes Cepstrales Mel-Frecuenciales). Si la matriz de densidad espectral es obtenida mediante la autocorrelación (y no directamente de la señal), hablamos de AMFCC (Autocorrelation Mel-Frequency-Cepstral-Coeficients, Coeficientes Cepstrales Mel-Frecuenciales). Una de las técnicas propuesta en esta Tesis emplea AMFCCs. Por razones de mejora en las tasas de reconocimiento, es habitual complementar esta representación con las velocidades y aceleraciones de los vectores cepstrales. En la Fig. 3.2 podemos ver un ejemplo de Log-Mel-Cepstrograma para una señal de voz limpia. Observamos que se trata de una representación muy distinta de las anteriores en la que las correlaciones verticales (en el dominio cepstral) se han reducido considerablemente.

### 3.1.5. Comparación de las representaciones

Fijándonos en la Fig. 3.2 se puede observar que no hay prácticamente diferencia entre cocleograma y espectrograma, por lo tanto se puede decir que ambas representaciones son

igualmente buenas para hacer reconocimiento. Esta igualdad se debe a que las matrices Gamm-Espectrograma y Log-Espectrograma son muy similares. La similitud se puede demostrar atendiendo a las siguientes cuatro razones:

1. La relación que se puede establecer entre banco de filtros gammatone (cocleograma) y transformada de Fourier de tiempo corto (espectrograma) [65].

2. La similitud que existe entre las escalas ERB y Mel.

3. Ambas representaciones emplean el logaritmo para comprimir las representaciones.

4. Ambas representaciones obtienen un nuevo vector de características cada 10ms.

La diferencia entre cocleograma y espectrograma proviene de los «subproductos» que generan sus pasos intermedios. Por ejemplo, a veces puede resultar más ventajoso emplear el cocleograma porque se quiera aplicar una técnica de extracción de pitch espectro-temporal o porque se quiera aplicar una técnica de extracción de zonas de tránsito (estas suelen depender de la matriz de movimiento basilar [155]). Sin embargo, otras veces puede resultar más ventajoso emplear el espectrograma porque se quiera aplicar alguna técnica de robustecimiento que requiera del uso de la matriz densidad espectral o de la matriz de autocorrelación (varias de las técnicas propuestas en esta Tesis usan estas dos matrices).

El cepstrograma sin embargo es totalmente diferente a los otros dos debido a la DCT. El cepstrograma ofrece las tres ventajas siguientes frente a las otras dos representaciones: reducir el número de componentes en la representación acústica (haciendo más ligera la carga computacional con la que debe trabajar el reconocedor y por lo tanto permitiendo trabajar con grandes vocabularios), obtener una representación acústica en la que los distintos canales estén decorrelados (haciendo que cada canal se pueda modelar independientemente del resto aligerando aun más la carga computacional en el reconocedor) y hacer más robusta la representación acústica (disminuyendo la diferencia test-entrenamiento frente a variabilidad entre hablantes y frente a ruidos).

El cepstrograma tiene el inconveniente de hacer muy difícil la localización y recuperación de los elementos que han sido contaminados por ruido aditivo, por lo que se prefiere usar como representación final de reconocimiento una vez que la señal limpia ha sido previamente estimada. Las otras dos representaciones no sufren de este problema (ver Sec. 3.2.1) por lo que son fácilmente aplicables en técnicas de reconocimiento robusto con información incompleta (Sec. 5.1.6).

## 3.2. Máscaras

### 3.2.1. Enmascaramiento de las representaciones

#### Fenómeno de la dominancia

Las representaciones acústicas anteriores (cocleograma, espectrograma y cepstrograma) sufren de enmascaramiento al igual que la audición humana (Sec. 2.2.4). Veamos en qué sentido se produce este efecto. Sea $y(t)$ una señal contaminada que es suma de una limpia $x(t)$ y un ruido $n(t)$. Si se compara la representación limpia correspondiente $X(c,t)$ ($c$ indica canal y $t$ tiempo, Sec. 3.1.1) con la sucia $Y(c,t)$, tendremos que muchos de los píxeles o elementos de la representación limpia aparecerán ahora, en la representación sucia, ocluidos o enmascarados por el ruido. Es más, se puede decir que cada píxel, o bien está dominado casi completamente por la señal limpia (es decir, su valor es casi el mismo que el que tiene $X(c,t)$) o bien que está dominado casi completamente por el ruido (su valor es casi el mismo que el que tiene $N(c,t)$). Denominaremos a este fenómeno «fenómeno de la dominancia» y se puede resumir en la siguiente ecuación:

$$Y(c,t) \approx Y^{dom}(c,t) = \begin{cases} X(c,t), & \text{si } |Y(c,t) - X(c,t)| < Thr \\ N(c,t), & \text{en caso contrario} \end{cases} \tag{3.6}$$

Donde $Y^{dom}(c,t)$ es lo que denominaremos «representación dominante».

#### Comprobación de la dominancia mediante imágenes

El «fenómeno de la dominancia» se puede comprobar si comparamos $Y(c,t)$ con $Y^{dom}(c,t)$ en imágenes. En la Fig. 3.3 podemos ver el cocleograma contaminado $Y(c,t)$ y más abajo la representación dominante $Y^{dom}(c,t)$. Se puede comprobar como ambas representaciones son muy similares. En las Fig. 3.4 y Fig. 3.5 tenemos lo mismo pero para un espectrograma y un cepstrograma.

La razón de este fenómeno, en cocleograma y espectrograma, es debida a que las representaciones son comprimidas logarítmicamente en algún momento de su obtención perdiéndose la linealidad en la suma de señales y produciéndose la aproximación log-max ($log(Y) = log(X+N) \approx max(log(X), log(N))$, Sec. 2.2.4). En el cepstrograma (que es la DCT del espectrograma) se sigue manteniendo la dominancia también.

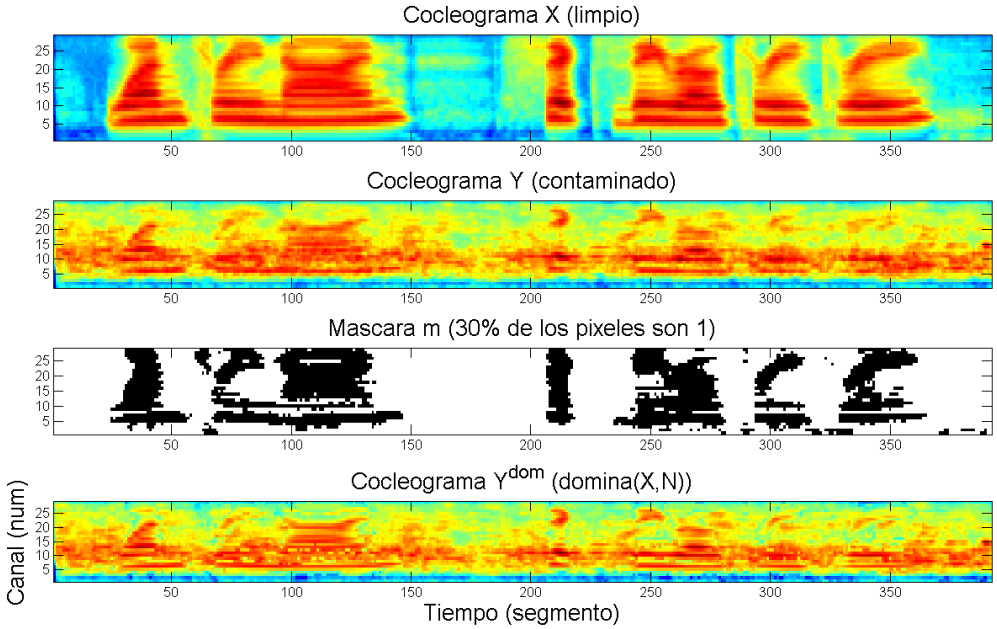



Figura 3.3: Enmascaramiento en el Cocleograma.

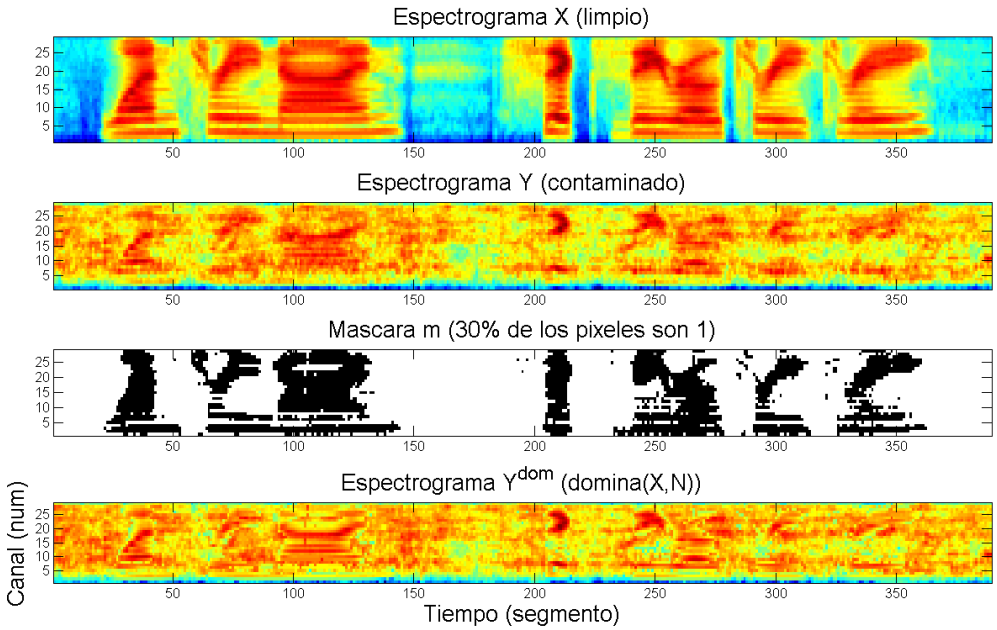



Figura 3.4: Enmascaramiento en el Espectrograma.

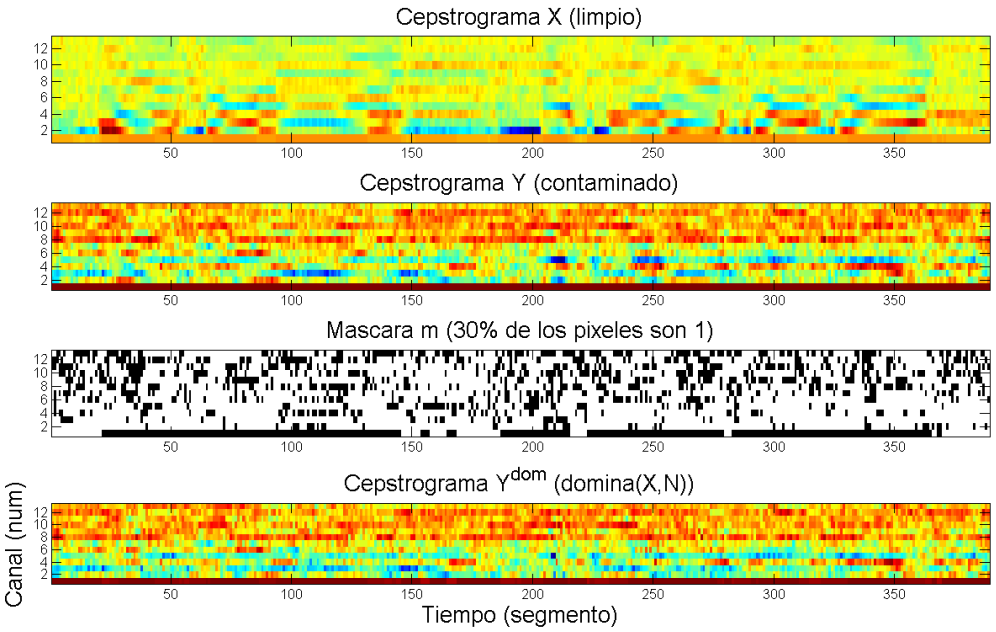



Figura 3.5: Enmascaramiento en el Cepstrograma.

### 3.2.2. Máscara discreta y analógica

#### Máscara discreta

Teniendo en cuenta lo anterior, si representamos con un 1 los píxeles en los que domina la voz y con un 0 en los que domina el ruido acabamos obteniendo lo que se denomina «máscara discreta de reconocimiento de la voz» o simplemente «máscara discreta». En las Fig. 3.3, 3.4 y 3.5 podemos observar las respectivas máscaras discretas de las representaciones acústicas.

Las técnicas de reconocimiento que emplean máscaras como MD (Missing Data) o SFD (Speech Fragment Decoding) (Sec. 5.1.6), denominan «máscara oráculo» a la máscara que indica sin equivocación cuando la voz domina sobre el ruido. Está máscara ideal es la que da mayor porcentaje de reconocimiento y es a la que debe de aproximarse cualquier estimación de máscara realizada. Las máscaras de las figuras anteriores son mascaras oráculo.

#### SNR de cada píxel

Para el caso del cocleograma y del espectrograma es posible obtener la máscara discreta $m(c,t)$ a partir de una umbralización de lo que se denomina «SNR de cada píxel». La

formula siguiente indica como hacerlo:

$$m(c,t) = \begin{cases} 1, \text{si } SNR(c,t) > Thr \\ 0, \text{en otro caso} \end{cases} \tag{3.7}$$

$$\text{donde} \quad SNR(c,t) = 20log_{10}\frac{M_X(c,t)}{M_N(c,t)} \tag{3.8}$$

donde $Thr$ es el valor umbral y donde $M_X(c,t)$ y $M_N(c,t)$ son las representaciones limpias y del ruido ($X(c,t)$ y $N(c,t)$) respectivamente llevadas al domino de la magnitud espectral (Sec. 3.1.3) mediante una transformación inversa. Para el caso del Log-Gamm-Cocleograma y el Mel-Log-Espectrograma estudiados en la Sec. 3.1, esta transformación inversa es la exponenciación (debido a la linealidad de los bancos de filtros y demás operaciones que se aplican) por lo que $M_X(c,t) = exp(X(c,t))$ y $M_N(c,t) = exp(N(c,t))$. Es más, despreciando el efecto de la fase en el domino de la magnitud espectral, es también posible obtener la $SNR$ a partir de la representación sucia $Y(c,t)$ de las siguientes dos maneras:

$$SNR(c,t) = 20log_{10}\frac{M_Y(c,t) - M_N(c,t)}{M_N(c,t)} \tag{3.9}$$

$$SNR(c,t) = 20log_{10}\frac{M_X(c,t)}{M_Y(c,t) - M_X(c,t)} \tag{3.10}$$

donde por lo general se limita el valor mínimo de las restas para evitar valores menores que cero.

### Máscara analógica

Los elementos de una «máscara analógica» están comprendidos entre 0 y 1, indicándonos de esta manera la probabilidad de que un píxel esté dominado por la voz. Este tipo de máscaras se emplean cuando el mecanismo de medida de la dominancia de la voz da lugar a valores continuos que además pueden estar afectados por error. Esta medida de la dominancia puede ser p. ej. una estima de la SNR de cada píxel (vista anteriormente) o la armonicidad de cada píxel (ver Sec. 3.3). La forma más habitual de adaptar estas medidas de la dominancia $md(c,t)$ (comprendidas en un intervalo cualquiera) al intervalo [0, 1] y obtener la máscara analógica $m_a(c,t)$, suele ser mediante la función sigmoide definida de la siguiente manera:

$$m_a(c,t) = \frac{1}{1 + e^{-\alpha(md(c,t)-\beta)}} \tag{3.11}$$

donde $\alpha$ se conoce como pendiente y $\beta$ como umbral.

#### Elección de los umbrales y pendientes

En el caso de la máscara discreta, el valor de umbralización $Thr$ (threshold) por lo general suele estar en torno a los 3 dB [27]. Este valor es tal que nos permite asegurar que si el píxel es fiable la contribución del ruido a la señal observada es prácticamente nula y el valor observado lo domina prácticamente la señal limpia.

En el caso de la máscara analógica, los valores de pendiente y umbral ($\alpha$ y $\beta$) se suelen escoger experimentalmente eligiendo aquellos que maximizan la tasa de reconocimiento. Lo normal es que el umbral óptimo continuo sea parecido al umbral óptimo discreto. Si tenemos en cuenta que por lo general es peor tomar un píxel de ruido como voz, que uno de voz como ruido, el valor umbral debe ser elegido de «manera conservadora» procurando que no se tomen muchos píxeles de ruido como fiables. La pendiente debe ser tal que, dentro del intervalo donde está el 65 % de los valores de SNR, la sigmoide cambie de 0.2 a 0.8 apróximadamente [91].

### 3.2.3. Técnicas de estimación de máscaras

Existen infinidad de técnicas para estimar máscaras [155]. La mayoría de las estimas de las máscaras son empleadas en reconocimiento MD, pero otras estimas pueden ser empleadas para hacer realce de voz directamente [117]. Existen técnicas de estima de máscaras que están especialmente pensadas para tratar ciertos ambientes o situaciones. Por ejemplo, en [115] se emplean técnicas específicas para ambientes reverberantes y en [53] se hace estimación de máscara a partir de señales estéreo y de la localización espacial de la voz.

Las técnicas para el cocleograma por lo general se basan en agrupar píxeles o conjuntos de píxeles a partir del empleo de reglas ASA (reglas primitivas o de alto nivel, Sec. 2.2.6). En la Sec. 5.2.3 se explican técnicas de este tipo.

Las técnicas para el espectrograma por lo general se basan en la estima o bien de la representación del ruido o bien de la señal limpia para a partir de las Ec. 3.8 y 3.10, estimar la SNR local de cada píxel y por lo tanto la máscara. Este tipo de técnicas se describen en la Sec. 5.2.3 ya que en esta Tesis se propone un método relacionado de obtención de la máscara. En el caso del espectrograma (o cocleograma) de la señal limpia, esta no suele ser la forma habitual de obtener máscaras ya que se suelen obtener mejores resultados de reconocimiento enviándolo directamente en forma de cepstrograma al reconocedor que

empleándolo como estimador de máscaras. Esto se debe a que el cepstrograma es una representación más robusta que el espectrograma (Sec. 3.1.5) y a que el cepstrograma se puede combinar con técnicas sencillas de robustecimiento tales como CMN (Sec. 5.1.3) que mejoran aún más los resultados.

Las técnicas para el cepstrograma podrían estar basadas también en la estima del espectrograma del ruido (o de la señal limpia) y en la aplicación de la DCT. Sin embargo, debido a que estas estimas del espectrograma nunca son perfectas y debido a que la DCT expande los errores a lo largo de la representación espectral, la estimación de máscaras del cepstrograma es una tarea abandonada [91] debido a que se obtienen mejores resultados reconociendo directamente con la estima limpia del cepstrograma que aplicando MD sobre el mismo.

## 3.3. Correlograma

### Correlograma

La importancia del correlograma fue primeramente señalada por Lickleder [82] como modelo auditivo de percepción del pitch. Posteriormente ha sido desarrollada por diferentes autores, entre ellos: Lyon y Weintraub [88, 157] (que crearon las primeras implementaciones computacionales), Slaney [143] (que le puso el nombre) y otros [98, 74, 91] (que lo han usado para obtener el pitch y separar señales simultáneamente).

El correlograma de un segmento de señal es la autocorrelación de cada una de las salidas de un banco de filtros (p. ej. un banco gammatone, Sec. 2.2.3) para ese trozo de señal y por lo tanto es una función 2D. El correlograma completo de una señal $x$ es una función 3D y se obtiene de la siguiente manera:

$$A_x(f,k,t) = \frac{1}{N}\sum_{n=k}^{N-1} x(f,t-n)x(f,t-n-k)w(n) \quad (0 \leq k < N) \tag{3.12}$$

donde $x(f,t)$ es la salida del banco de filtros con frecuencia central $f$, $k$ es el retardo de autocorrelación, $t$ es el instante de tiempo del trozo de señal de tamaño $N$ y $w$ es una ventana aplicada sobre el correspondiente segmento de señal. Como vemos aquí se está empleando la parte positiva de la autocorrelación biased (sesgada). Existe un algoritmo rápido para la obtención del correlograma considerando la FFT y el teorema de Wiener-Khinchin [154].

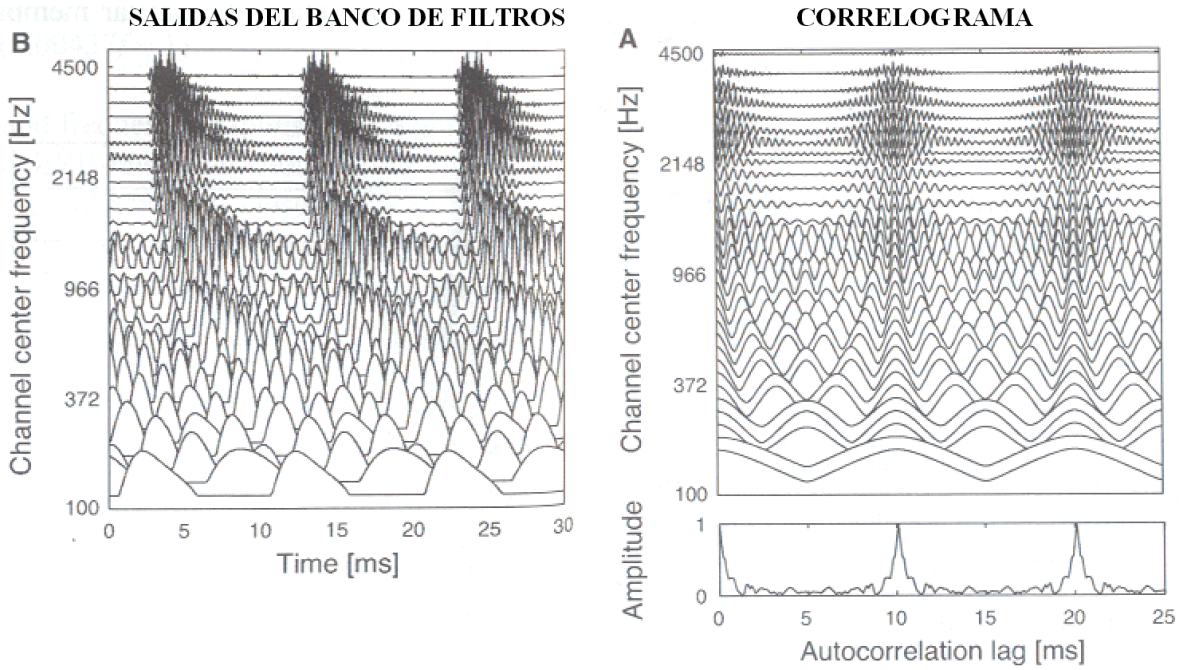



Figura 3.6: [155] Izquierda, salidas del banco de filtros para la señal de una vocal de 500 Hz. Derecha arriba, correlograma del segmento de una vocal de 100 Hz. Derecha abajo, autocorrelación sumada (suma de las autocorrelaciones de los distintos canales).

### Altas y bajas frecuencias en el correlograma

Teniendo en cuenta que para imitar al oído, el banco de filtros aumenta el ancho de banda de sus filtros conforme crece la frecuencia central, las salidas del banco de filtros para una señal armónica tendrán la siguiente forma: en los filtros graves donde los armónicos son resolubles serán senoidales. En los filtros agudos, donde entran a la vez más de dos armónicos, tendremos una señal modulada AM cuya frecuencia de modulación es la fundamental de la señal armónica de entrada. Por lo tanto las autocorrelaciones del correlograma compartirán un máximo común en el retardo correspondiente a la frecuencia fundamental y esto puede ser empleado para extraer el pitch (Sec. 3.4). En la Fig. 3.6 podemos observar a la izquierda las salida del banco de filtros (rectificadas en media onda) para una señal armónica de 500 Hz de pitch. A la derecha el correlograma de un trozo de señal 100 Hz de pitch.

Este tipo de representación (que trabaja de forma diferente las altas y bajas frecuencias tal y como las evidencias psicoácusticas indican [21]) es la que ha llevado al correlograma a ser empleado como método de obtención del pitch y de separación de fuentes.

#### Armonicidad

Mediante el correlograma podemos obtener lo que se denomina la armonicidad de un píxel $(f, t)$ para un determinado pitch $p$ (medido en número de muestras). Esta se obtiene dividiendo el valor del correlograma para retardo $p$ entre el valor del correlograma para retardo 0 de la siguiente manera:

$$H_x^p(f,t) = A_x(f,t,p)/A_x(f,t,0) \tag{3.13}$$

Dado que el valor máximo de la autocorrelación reside en el retardo 0, está armonicidad se acercará a 1 si el píxel está dominado por una fuente armónica con pitch $p$ y se acercará a 0 en caso contrario. De esta forma, la armonicidad puede ser empleada para asociar píxeles a determinadas fuentes de las cuales se conoce su pitch y por lo tanto para estimar máscaras y separar la voz del ruido.

#### Autocorrelación sumada

En Fig. 3.6 de la derecha abajo, podemos observar lo que se denomina autocorrelación sumada. Una autocorrelación sumada se obtiene sumando las autocorrelaciones de un cierto conjunto de canales ($f \in F$) de la siguiente manera:

$$SA_x^F(k,t) = \sum_{f \in F} A_x(f,k,t) \quad (0 \leq k < N) \tag{3.14}$$

Si la suma se realiza sobre todos los canales obtenemos la autocorrelación total del segmento suma de las fuentes presentes (p. ej. voz+ruido). Si se hace sobre ciertos canales dominados por una misma fuente (los de la voz) la autocorrelación sumada se acerca bastante a la total de esa fuente sola (la de la voz sola). La autocorrelación sumada puede ser empleada para obtener el pitch de una fuente (o de un conjunto de píxeles) mediante el máximo de la autocorrelación sumada tal y como se estudia en la Sec. 3.4.

## 3.4. Extractores del Pitch

### 3.4.1. Tipos de técnicas

De manera similar a las teorías sobre percepción del pitch (Sec. 2.2.5), podemos clasificar las técnicas computacionales de extracción de pitch en espectrales, temporales y

espectro-temporales. Veamos algunas de las técnicas más significativas pensadas para extraer el valor o los valores de pitch que hay en un segmento de señal [155].

### Espectrales

Las técnicas espectrales usan el modulo del espectro para obtener el pitch.

Para el caso de un solo pitch la técnica del histograma de Shcroeder [137] proporciona muy buenos resultados porque obtiene el pitch para todas las formas de señales periódicas que se pueden dar (espectros sin el armónico fundamental, espectros que les falta parte de sus armónicos, etc.). Una técnica parecida a esta es la [23, 95] que se basa en obtener el producto escalar entre el espectro y un tren de pulsos espectrales (espectro peine o comb spectrum) de una determinada frecuencia, y en tomar como pitch la frecuencia que proporcione mayor producto. Para evitar que unos armónicos pesen mucho más que otros y que estos lleguen a dirigir la obtención del pitch, se suele trabajar con una compresión del espectro (como el espectro en dB). El extractor de pitch [106] empleado en esta Tesis, y el cual es una modificación del xFE de la ETSII [148], usa esta técnica.

Para el caso de varios valores de pitch (pensamos en dos voces sonando a la vez aunque se puede extender a más de dos voces) podemos mencionar la técnica supresiva-iterativa de Parson [116] que también es válida para separación de voces es decir, para obtener la forma espectral de una voz y la otra. Esta técnica en el paso 0, extrae un pitch $F0$ mediante alguna técnica de un solo pitch. En el paso 1, suprime los armónicos correspondientes a $F0$ (mediante un filtrado peine supresivo) y obtiene el pitch $F1$. En el paso 2, suprime los armónicos correspondientes a $F1$ y obtiene de nuevo el pitch $F0$. De esta forma se van repitiendo los pasos 1 y 2 hasta que se tienen los dos valores de pitch.

### Temporales

Las técnicas temporales usan o bien la representación temporal de la señal o bien una función de autosimilitud como la autocorrelación para obtener el pitch.

Para el caso de un solo pitch la técnica de Rabiner [125], basada en tomar como pitch el máximo de la autocorrelación de la señal (más bien de un preprocesado de esta mediante *clipping* de picos máximos), proporciona muy buenos resultados ya que obtiene el pitch para todas las formas de señales periódicas que se pueden dar (señales periódicas con dos picos máximos, etc). Una técnica similar es la de Cheveigne [26] que en lugar de emplear la autocorrelación como función de autosimilitud emplea la SFD (Squared Difference Function, Función de Diferencia Cuadrática). El extractor de pitch YIN [26]

usa esta técnica también. Otro extractor de pitch basado en filtrado peine temporal (filtro que resta dos muestras separadas un periodo determinado y que su respuesta en frecuencia es como el de un filtro peine espectral) es el propuesto por Droppo en [37] que más que dar un pitch para todo el segmento de señal, da un pitch para cada muestra temporal resultando útil cuando la señal, siendo periódica, sufre pequeñas modulaciones en amplitud y frecuencia.

Para el caso de varios pitches podemos mencionar la técnica supresiva-iterativa de Frazier [46] o Cheveigne [25] basada en filtro peine temporal. Otra técnica no supresiva-iterativa es la de Weintraub [158] que usando la autocorrelación encuentra pistas que indican los dos pitch existentes.

#### Espectro-temporales

Las técnicas espectro-temporales suelen emplear el correlograma (Sec. 3.3) o alguna función de autosimilitud aplicada sobre los diferentes canales de un banco de filtros para obtener el pitch.

Para el caso de un solo pitch podemos mencionar las técnicas que emplean la correlación sumada (suma de las correlaciones en los diferentes canales, Sec. 3.3) para tomar como pitch el máximo de esta [24, 140]. Estas técnicas pueden llevar el añadido de que a la suma solo contribuyan los canales que se sepa que son sonoros (autocorrelación con forma periódica) descartando aquellos que se sepa que son de ruido (autocorrelación con forma parecida a la del ruido blanco).

Para el caso de varios pitches podemos mencionar la técnica supreso-iterativa de Meddis [98], la de Wu [162] (que emplea un criterio heurístico para descartar los canales de ruido en la autocorrelación sumada) y la de Ma [90] (que usa un SFD para reconocer y extraer el pitch al mismo tiempo). Como vemos estas técnicas también son válidas para separación de voces.

### 3.4.2. Comparación

Empleando el teorema de Wiener-Khinchin (que la autocorrelación es la IFT o transformada inversa de Fourier de la densidad espectral de potencia) se puede mostrar la similitud entre las técnicas espectrales y temporales tal como ha mostrado Ellis [39]. Es más, si tomamos el logaritmo al espectro antes de aplicar la IFT (tal y como hemos dicho que hacen algunas técnicas espectrales para evitar la dominancia de ciertos armónicos)

acabamos obteniendo el cepstrum, el cual también se puede emplear para estimar el pitch [110].

En general, respecto al tiempo de cómputo, los tres tipos de técnicas tienden a ser igualmente rápidas debido a que poseen algoritmos rápidos derivados de la FFT (Fast Fourier Transform) para su obtención.

Tal y como ha señalado Klapuri [75] la principal ventaja de las técnicas espectro-temporales sobre las otras dos es que permiten trabajar mejor con señales periódicas ligeramente inarmónicas gracias a la modulación AM de las altas frecuencias (en este caso los canales agudos tendrían una envolvente de periodo igual al pitch de la señal, ver Sec. 3.3). Esto conlleva que no sea necesario una ventana temporal demasiado ancha para resolver las altas frecuencias (como sí necesitarían las espectrales) o de un pitch perfectamente establecido para separar canales (como sí necesitarían las temporales en sus filtros peine temporales para separar señales).

### 3.4.3. Detalles de implementación

Lo que hace ser más efectivos a unos extractores de pitch respecto a otros, no es tanto la técnica empleada a nivel de segmento, si no los detalles en la implementación global. Estos detalles suelen ser restricciones que dependen del objeto que emite el pitch. Por ejemplo, si vamos a extraer el pitch de voces humanas podemos decir que este debe estar en torno al intervalo $80 - 270Hz$ (Sec. 2.1.2). Si vamos a extraer el pitch de un instrumento musical como el piano, podemos emplear un modelo de evolución temporal de la envolvente espectral que nos indique como se va apagando el sonido y que nos ayude a buscar sus diferentes armónicos. O si sabemos que el pitch debe variar suavemente (como en el caso del habla) podemos aplicar un suavizado a los pitches de los diferentes segmentos que evite así los saltos bruscos. Este suavizado puede ser tan complejo como se quiera (p. ej. en la técnica de Ma [90] que se estudiará en la Sec. 5.2.3 se aplica un suavizado basado en HMMs a una serie de candidatos a pitch). El extractor de pitch [106] que emplearemos en esta Tesis aplica este tipo de restricciones para hacerlo más robusto frente al ruido.

# 3. FUNDAMENTOS II: REPRESENTACIONES, MÁSCARAS Y EXTRACTORES DE PITCH

# Capítulo 4

# Fundamentos III: Reconocedores

## 4.1. Reconocedor basado en HMMs

### 4.1.1. Justificación de los HMMs

#### Aproximaciones al ASR

En [85, 124] se da una clasificación (no muy rigurosa pero útil) de las tres aproximaciones principales que se han hecho para abordar el problema del ASR (Automatic Speech Recognition, Reconocimiento Automático de la Voz): La aproximación acústico-fonética, que se basa en la teoría de rasgos binarios de Jakobson [68], separa los fonemas que componen la señal y con estos se reconoce el mensaje usando arboles de decisión. La aproximación desde la inteligencia artificial, que se basa en tener un conjunto de reglas lógicas de clasificación en un Sistema Experto para cada nivel de lenguaje (acústico, léxico, sintáctico,..), usa métodos inductivos (botton-up), deductivos (top-down) o de pizarra (botton-up más top-down) para aunar los diferentes niveles y reconocer el mensaje. Y la aproximación de reconocimiento estadístico de patrones, que se puede abordar mediante DTW (Dynamic Time Warping, Alineamiento Temporal basado en programación Dinámica), NN (Neural Network, Redes Neuronales) o HMMs (Hidden Markov Models, Modelos Ocultos de Markov), la cual trocea la señal en segmentos de un tamaño que no tienen porqué corresponderse con los fonemas y a partir de estos se reconoce usando el modelo estadístico.

La aproximación acústico-fonética no ha resultado ser una buena solución debido a la dificultad que hay en separar y distinguir unos fonemas de otros (fenómenos de coarticulación y variabilidad entre hablantes respectivamente, Sec. 2.1). La aproximación desde la

inteligencia artificial es una buena solución pero está más orientada a hacer reconocimiento de alto nivel tal como resolver ambigüedades léxicas usando la sintaxis. [80, 86]. La aproximación de patrones es una buena solución porque no necesita separar con exactitud los fonemas permitiendo reconocer cualquier unidad lingüística (desde fonemas, palabras, hasta frases completas, Sec. 2.1.1) y por que permiten capturar la variabilidad entre hablantes.

Dentro de la aproximación de patrones, todos los reconocedores tienen en común que constan de una primera etapa de entrenamiento, donde se entrenan sus patrones o modelos, y otra de test, donde se da la solución de reconocimiento en términos de probabilidad. Los DTWs miden el coste del alineamiento mínimo entre dos secuencias, la de test y la de referencia. Las NNs primero entrenan los pesos de la red (normalmente de un perceptrón multicapa) y en la etapa de test, la red da en su salida la solución de una forma codificada. Los HMMs son entrenados para modelar las distintas unidades lingüísticas a reconocer. En la etapa de test se selecciona el modelo que más probablemente represente a la señal que se esté testeando.

#### Éxito de los HMMs

El motivo principal por el que los HMMs, a diferencia de los DTWs y las NNs, se han erigido como la opción más usada hoy día para resolver el problema del ASR es de que estos han podido incorporar en un modelo común tanto el modelado acústico de bajo nivel (unidades lingüísticas y silencios) como el del lenguaje de alto nivel (gramática).

Esto ha permitido que se pueda realizar al mismo tiempo la segmentación y el reconocimiento de las unidades lingüísticas mediante un método tipo pizarra, sin necesidad de emplear un detector de silencios como sí lo necesitan los DTWs y las NNs. A su vez, esto ha dado origen al éxito de los HMMs en el reconocimiento de habla continua con grandes vocabularios empleando como unidades lingüísticas fonemas o trifonemas.

### 4.1.2. Reconocimiento mediante HMMs

#### Modelado de la voz

La forma que tienen los HMMs de modelar el habla continua consiste en crear un macromodelo HMM que une pequeños modelos HMMs representantes de las diferentes unidades lingüísticas consideradas.

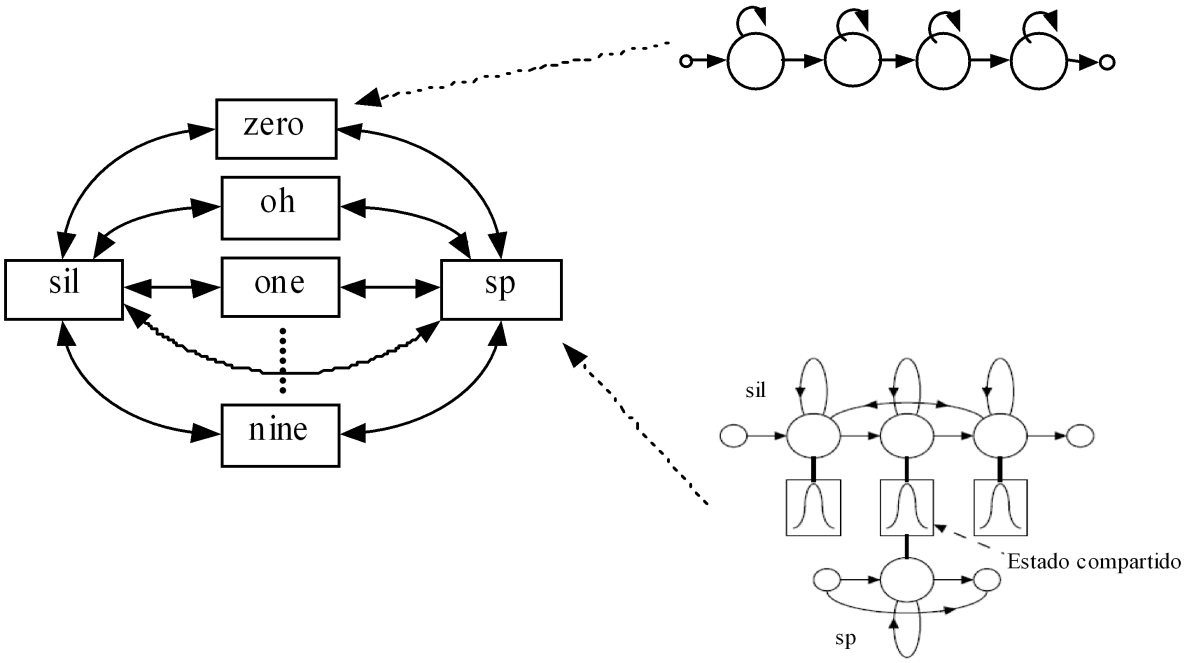



Figura 4.1: Macromodelo HMM para reconocimiento de dígitos conectados. Se observa como el silencio *sil* comparte un estado con la pausa corta *sp*.

Las unidades lingüísticas (y los silencios) son HMMs definidos por sus estados $q$, sus probabilidades de transición entre estados $a_{i,j}$ y sus probabilidades de emisión $p(\mathbf{x}|q)$ de la observación $\mathbf{x}$ dado el estado $q$ a las que nos referiremos como probabilidades de observación. Las probabilidades de transición entre las palabras (en un macromodelo) vienen dadas por el modelo o gramática del lenguaje. Cada estado suele representar un segmento de señal cuasiestacionario (casi un fonema). La topología de los HMMs de cada palabra es normalmente «hacia delante».

La Fig. 4.1 muestra de forma simplificada, el macro-modelo que se emplea para modelar las frases de dígitos conectados de Aurora (Aurora-2 y Aurora-3). Aquí las unidades lingüísticas consideradas no son fonemas o trifonemas sino palabras (representantes de los dígitos) y estas se interconectan por medio de la pausa corta (*sp*) o el silencio largo (*sil*). La forma de modelar las interconexiones directas entre palabras de forma que sea un modelo de dígitos conectados, es mediante la transición directa que tiene la pausa corta *sp*. El silencio y la pausa corta comparten una distribución de emisión de estado. En la Sec. A.1 se dan más detalles sobre el macromodelo de Aurora.

Las probabilidades de emisión de estado se suelen modelar mediante funciones de densidad de probabilidad separables tales como GMMs (Gaussian Mixture Models, Modelos

de Mezcla de Gaussianas) con matriz de covarianza diagonal. La separabilidad implica suponer que las componentes $x_c$ del vector de características (o canales de la representación acústica) son independientes entre si. Esto hace que esta probabilidad se estime de la siguiente manera:

$$p(\mathbf{x}|q) = \sum_{M}^{k=1} P(k,q)p(\mathbf{x}|q,k) = \sum_{M}^{k=1} P(k,q) \prod_{i} p(x_i|q,k) \tag{4.1}$$

donde $M$ es el número de gaussianas empleado y suele depender del tipo de representación acústica empleada para que se cumpla la hipótesis de separabilidad. Para el cepstrograma suele ser menor que para el espectrograma (y cocleograma) debido a que la independencia entre componentes es mayor. Este aumento del nº de Gaussianas en el espectrograma hoy día ya no supone un coste computacional elevado posibilitando esto el desarrollo del reconocimiento espectral tal y como hacen los sistemas de MD.

Por ultimo mencionar que una vez establecidos correctamente los parámetros del macromodelo, si este se emplease como «generador» de señal, el macromodelo empezaría a pasar de unos estados a otros emitiendo vectores de características de forma que la secuencia producida nos «recordaría» a una persona diciendo dígitos conectados. Esto es el fundamento de los modernos sintetizadores de voz basados en HMMs [41].

### Entrenamiento

El entrenamiento de un HMM, previamente fijada la topología (nº de estados, enlaces, etc), consiste en disponer de múltiples representaciones acústicas (conjunto de entranamiento) del sistema a modelar y a partir de ellas estimar los valores $a_{i,j}$ y $p(x|q)$ que mejor representen al conjunto de entrenamiento y por lo tanto del sistema a modelar. El algoritmo más comúnmente empleado para estimar estos valores es el de Baum-Welch [126] el cual es un algoritmo tipo EM (Expectation-Maximization, Expectación-Maximización).

### Reconocimiento

El reconocimiento empleando HMMs consiste en averiguar la secuencia de palabras $W = w_1, w_2, ...., w_T$ más probable dada la representación acústica o secuencia de observación $X = \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T$. Teniendo en cuenta que en el macromodelo cada secuencia de estados $Q = q_1, q_2, ...., q_T$ se corresponde con una secuencia de palabras $W$ el problema del reconocimiento se reduce a averiguar la secuencia de estados óptima dada la secuencia

de observación:

$$\hat{Q} = \arg\max_{Q} P(Q|X) \tag{4.2}$$

Empleando la regla de Bayes y teniendo en cuenta que la probabilidad total de observar la secuencia $X$ es el producto de las probabilidades de emisión de la secuencia $Q$ considerada, nos queda:

$$\hat{Q} = \arg\max_{Q} \frac{P(X|Q)P(Q)}{P(X)} = \arg\max_{Q} \prod_{t=1}^{T} p(\mathbf{x}_t|q_t)P(Q) \tag{4.3}$$

donde $P(X)$ se toma constante, $p(\mathbf{x}_t|q_t)$ se obtiene mediante la Ec. 4.1 y $P(Q)$ depende de las probabilidades de transición. Este problema de averiguar la secuencia oculta (hidden) de estados más probable (o de decodificar la secuencia de observación), se podría resolver de forma «exhaustiva» probando todas las posibles secuencias de estados existentes y eligiendo aquella que de mayor probabilidad. Sin embargo, gracias a que los HMMs de la voz tienen topología «hacia delante» existe un algoritmo rápido para encontrar o decodificar la secuencia de estados más probable. Este es el conocido algoritmo de Viterbi [126].

## 4.2. Reconocedor de MD basado en HMMs

### 4.2.1. Introducción

#### Orígenes

En el mundo del reconocimiento de las señales suele ocurrir que la información disponible para reconocer la señal deseada esté incompleta (posea partes no fiables). Las primeras técnicas desarrolladas para reconocer señal a partir de información incompleta no fueron desarrolladas en el campo del ASR robusto, sino en el del reconocimiento de objetos en visión [1] o en el de reconocimiento de voz con perdidas de paquetes por transmisión (Weighted Viterbi o Soft Decoding [121]).

Este retraso de aplicación en el campo del ASR fue debido a que por ejemplo en el campo de la visión era más patente el fenómeno de la oclusión (los objetos se tapan los unos a los otros) que en el del sonido (donde normalmente tenemos la impresión de poder percibir varios sonidos a la vez). Sin embargo, tanto en visión como en sonido ocurren con la misma frecuencia la oclusión y percepción simultanea de objetos (todo depende de que es lo que a la mente se le haga consciente).

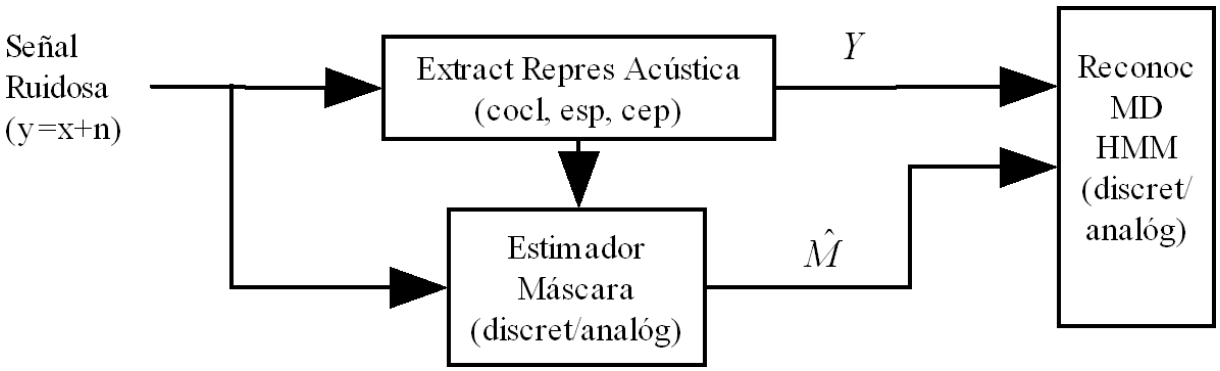



Figura 4.2: Sistema de reconocimiento compuesto por el extractor de la representación acústica (cocleograma, espectrograma o cepstrograma), el estimador de máscaras (discreta o analógica) y el reconocedor de MD basado en HMMs que puede trabajar con máscaras discretas o analógicas.

Desde que se ha sido tenido en cuenta este hecho, diversos autores han intentado crear técnicas de ASR que trabajen con información incompleta. Técnicas como Multistream [15, 59] (Sec. 5.1.6) son un ejemplo de esto. Sin embargo, no ha sido hasta comienzos del 2000 cuando, gracias a los trabajos de los investigadores de la Universidad de Sheffield (especialmente Cooke [27]), se han empezado a obtener buenos resultados de reconocimiento considerando que el espectrograma de la voz tiene partes o datos perdidos (MD, Missing Data). El avance de estos investigadores ha consistido en desarrollar un aparato matemático que ha permitido incorporar, sin apenas cambios, técnicas estadísticas de MD a los ya bien establecidos reconocedores-HMM (ver Sec. 4.2.3). Esto ha supuesto emplear las ventajas que ofrecen los HMMs frente a otro tipo de aproximaciones como NN o DTW (Sec. 4.1.1). En esta Tesis trabajaremos con este tipo de reconocedor de MD basado en HMMs.

### Sistema de reconocimiento de MD

En la Fig. 4.2 se puede ver el esquema general de un sistema de reconocimiento de MD basado basado en HMMs. Podemos ver que posee tres subsistemas: el extractor de la representación acústica, el estimador de máscaras y el reconocedor de MD. Los dos primeros se han estudiado en las Sec. 3.1 y 3.2.1. El reconocedor de MD puede ser para máscaras discreta o analógicas y es el que estudiaremos aquí.

### 4.2.2. Justificación del empleo

A continuación estudiaremos los motivos que nos han llevado a emplear un sistema de reconocimiento de MD en esta Tesis. Los motivos son tanto de naturaleza psicoacústica como técnica.

#### Motivos psicoacústicos

- Psicoacústicamente se ha demostrado que la información que llega al sistema central por parte del nervio auditivo sufre de enmascaramiento (Sec. 2.2.4), siendo habitual para el humano reconocer voz a partir de información incompleta ([44, 58]). Esto ha motivado la búsqueda de sistemas automáticos que trabajen con medidas de incertidumbre (Sec. 5.1.6).

- Los sistemas de MD han permitido que se puedan implementar computacionalmente, y de una forma sencilla y elegante, muchas de las ideas sobre percepción que durante mucho tiempo venía proponiendo la psicoacústica tales como el efecto de enmascaramiento o las reglas de agrupación ASA (Auditory Scene Analysis, Sec. 2.2.6). De todo esto ha surgido un nuevo campo de investigación denominado CASA (Computational ASA) que intenta, a diferencia de la separación ciega de fuentes, separar sonidos siguiendo los mecanismos de audición humana. La novedosa técnica de reconocimiento SFD (Speech Fragment Decoding, Sec. 5.1.6) es un resultado claro del desarrollo de CASA, y muestra como se pueden aunar con gran éxito principios de percepción psicoacústica con técnicas de MD.

#### Motivos técnicos

- Se han observado una serie de características en las representaciones espectro-temporales de la voz (cocleograma y espectrograma) que han permitido a los reconocedores de MD poder obtener buenos resultados de reconocimiento. Estas características son las dos siguientes [8]: 1) La voz concentra su energía en ciertas regiones espectro-temporales (formantes y armónicos) que, incluso en condiciones de ruido muy altas (0dB), sobresalen sobre el ruido. Esto permite que la identificación de estas regiones sobre el ruido sea relativamente sencilla aplicando técnicas de estimación de máscaras. 2) Estas regiones están distribuidas de forma redundante por todo el espectrograma de forma que si el ruido enmascara gran parte de estas

regiones es posible reconocer con alta fiabilidad (Cooke demostró en [28] que bastan el 10 % de los pixeles totales para reconocer un mensaje).

- Los sistemas de MD han reducido el problema del ASR robusto a prácticamente la estima de máscaras evitando así los problemas de tener que averiguar con mucha exactitud (Sec. 5.1.6) las partes de la voz enmascaradas por el ruido. Con una buena estimación de la máscara se pueden llegar a obtener porcentajes de reconocimiento del orden o incluso superior al del humano (del 90 % a 0 dB). Por todo esto autores como Wang [155] han propuesto que el problema del reconcimiento robusto es el problema de la estima de la máscara oráculo.

En la Sec. 5.1.6 se añaden otro tipo de motivos que justifican el empleo de técnicas de reconocimiento con incertidumbre en la información frente a técnicas con información completa o sin incertidumbres.

### 4.2.3. Técnicas de estimación de probabilidades

#### Incorporación de las técnicas de MD a los HMMs

Veamos como se incorporan las técnicas de MD a los reconocedores basados en HMM tal y como han propuesto los investigadores de Sheffield [27]. Supongamos que tenemos una secuencia de observación o representación acústica $X$ que intentamos reconocer y de la cual poseemos su correspondiente máscara $M$. Como hemos visto en la Sec. 4.1.2 la forma de hacerlo es resolviendo la Ec. 4.3 mediante el algoritmo de Viterbi (que nos permite averiguar la secuencia de estados $Q$ más probable dada la observación $X$). Tal y como vemos en esta ecuación, este algoritmo requiere del cómputo de las probabilidades de emisión de estado $p(\mathbf{x}_t|q_t)$ y que en lo que sigue denominaremos $p(x|q)$. Cuando parte de los elementos de $x$ no se conocen el cálculo de estas probabilidades se debe de hacer de una forma distinta a la normal (Ec. 4.1) y es aquí donde se incorporan las técnicas de MD de estimación de probabilidad.

Las técnicas de estima de las probabilidades se pueden dividir en dos grupos: técnicas de imputación y técnicas de marginalización.

#### Imputación

Las técnicas de imputación se emplean, más que para estimar probabilidades, para estimar el vector de características limpio ($\hat{x}$) a partir de las componentes fiables del

mismo y de un modelo estadístico que nos indica como se distribuyen y relacionan las diferentes componentes del vector. Estas técnicas se describen en más detalle en [27, 127]. Mencionar que en [27] se puede observar que las fórmulas del aparato probabilístico que nos permite hacer estimas de los elementos no fiables mediante imputación, comparten muchos términos en común con las formulas de marginalización que estudiaremos a continuación.

### Marginalización

La técnica de marginalización que estudiaremos a continuación sí que nos permite estimar directamente las probabilidades del algoritmo de Viterbi a partir de los datos fiables (sin tener que estimar los no fiables). Es la técnica que llevan incorporados la mayoría de los reconocedores de MD hoy en día y el reconocedor que usaremos para evaluar nuestras técnicas. Veamos como opera.

Marginalizar consiste en estimar la probabilidad "apartando" o "marginalizando" a los elementos que no son fiables. Este método propone usar la probabilidad marginal como una buena estima de la probabilidad total de observación:

$$p(x|q) \approx p(x_r|q) = \int p(x_r, x_u|q)dx_u \tag{4.4}$$

Donde hemos separado el vector $x$ en el conjunto de sus elementos fiables $x_r$ ($r$ de reliable) y el de los no fiables $x_u$ ($u$ de unreliable). La contribución, al valor de la probabilidad total, de los elementos fiables se deja como está y la de los no fiables se promedia integrando sobre el conjunto de los posibles valores que pueden llegar a tomar. Aunque no aparezca en la formula, esta integración debe ser normalizada por el intervalo de integración para que tenga el efecto de un promediado.

Hasta aquí este mecanismo sirve para cualquier tipo de función de probabilidad. Sin embargo, tal y como dijimos en la ec. 4.1.2 por razones de coste computacional, lo normal es trabajar con funciones de probabilidad separables tales como GMMs con matriz diagonal. Teniendo en cuenta esto nuestra función de probabilidad se puede separar en productos, quedándonos de la siguiente manera:

$$p(x|q) = \sum_{k=1}^{M} P(k|q) \prod_{i \in r} p(x_i|q,k) \prod_{i \in u} \int p(\hat{x}|q,k)d\hat{x} \tag{4.5}$$

En la Fig. 4.3 podemos ver un ejemplo del cálculo de esta probabilidad del vector $x$. El dibujo podría pensarse como un espectrograma y las zonas marcadas con *speech* representa las zonas fiables (donde la voz domina al ruido).

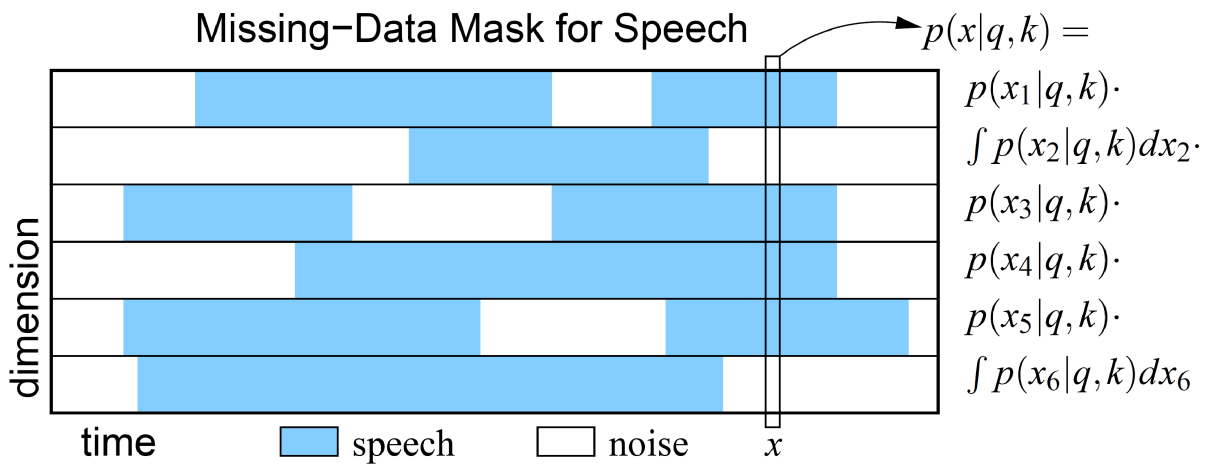



Figura 4.3: [91] Estimación de la probabilidad marginal en un instante de tiempo teniendo en cuenta la máscara de reconocimiento de la voz.

Como se ha dicho anteriormente la integral es en realidad un promediado de la contribución a la probabilidad total de los elementos no fiables. Esta integral se hace sobre el rango de posibles valores que pueden llegar a tomar los elementos no fiables. Este rango en el domino espectral suele ser entre el valor mínimo posible $x_i^{lb}$ ($lb$ de low bound) y el valor observado $x_i$ (con ruido aditivo el valor real estará entre estos dos valores). Teniendo en cuenta esto la probabilidad queda de las siguiente forma:

$$p(x|q) = \sum_{k=1}^{M} P(k|q) \prod_{i \in r} p(x_i|q,k) \prod_{i \in u} \frac{1}{x_i - x_i^{lb}} \int_{x_i^{lb}}^{x_i} p(\hat{x}|q,k) d\hat{x} \tag{4.6}$$

Si no se conocieran los límites de los elementos no fiables (como suele ocurrir en transmisión donde se pierden completamente algunos elementos) se integraría entre $-\infty$ y $+\infty$. En [28] se muestra que siempre que se pueda es mejor emplear conocimiento sobre los límites. Las integrales, al ser sobre gaussianas, se pueden evaluar de forma muy rápida empleando diferencias de la función error [27].

La Ec. 4.6 puede derivarse también en el marco de la aproximación soft-data (en la que los datos dejan de ser deterministas para convertirse en pdfs de evidencia) suponiendo que los datos se ajustan a una pdf uniforme en el rango $[x_i^{lb}, x_i]$ [121].

### Marginalización Soft

Los errores en una máscara discreta (valores 0 o 1) son irreversibles y pueden tener un gran impacto en el rendimiento del reconocimiento. Sin embargo, en una máscara analógica (con valores entre 0 y 1, Sec. 3.2.2) al no rechazarse o aceptarse completamente los píxeles, se permite recuperarlos o desecharlos en función de lo bien que encajen en el modelo HMM de reconocimiento. En [128] tenemos los primeros pasos del empleo de máscaras analógicas en MD, aunque más bien aplicadas sobre imputación. En [7, 6] es donde se demuestra que el empleo de máscaras analógicas en marginalización, consigue incrementar las tasas de reconocimiento respecto a las máscaras discretas de una forma notable (de unos 15 puntos más sobre 100 a 0 dB).

Si llamamos $w_i$ a la probabilidad (entre 0 y 1) de que el elemento observado $x_i$ sea fiable, el cálculo de la probabilidad de observación se convierte en:

$$p(x|q) = \sum_{k=1}^{M} P(k|q) \prod_{i=1}^{N} \left( w_i p(x_i|q,k) + (1-w_i) \frac{1}{x_i - x_i^{lb}} \int_{x_i^{lb}}^{x_i} p(\hat{x}|q,k) d\hat{x} \right) \tag{4.7}$$

Se puede observar que cuando las probabilidades de fiabilidad $w_i$ de la máscara analógica se hacen discretas, esta ecuación se convierte en la de la máscara discreta (Ec. 4.6).

# Capítulo 5

# Técnicas de Robustecimiento Convencionales y Basadas en el Pitch

## 5.1. Técnicas de robustecimiento convencionales

### 5.1.1. Clasificación

Son muchas las técnicas que se han propuesto para hacer robustos a los sistemas de ASR frente al ruido (ya sea aditivo, convolutivo o ambos, Sec. 1.1). Muchas de las técnicas existentes han sido ideadas propiamente para ASR robusto, sin embargo otras muchas provienen de otros campos que están más orientados a que el ser humano perciba la señal con inteligibilidad y/o calidad (p. ej., realce de la voz [83, 111] o transmisión robusta de la voz [121, 104, 22]). También se han empleado técnicas estéreo (basadas en arrays de micrófonos) para separar la voz del ruido y así robustecer el reconocimiento. Técnicas como separación ciega de fuentes (BBS) [70], basadas en análisis de componentes independientes (ICA) [145] y en que las señales se mezclan linealmente, pueden ser empleadas para esto. Sin embargo, teniendo en cuenta que en esta Tesis estamos interesados en técnicas monofónicas, podemos clasificar las técnicas de robustecimiento de la siguiente manera [121].

**Preprocesamiento en el domino temporal**: cuando se modifica la señal de test contaminada para conseguir un mayor rendimiento del sistema ASR.

**Parametrización robusta**: cuando se selecciona una representación acústica adecuada que no se vea afectada por las variabilidades del ruido y la señal de voz.

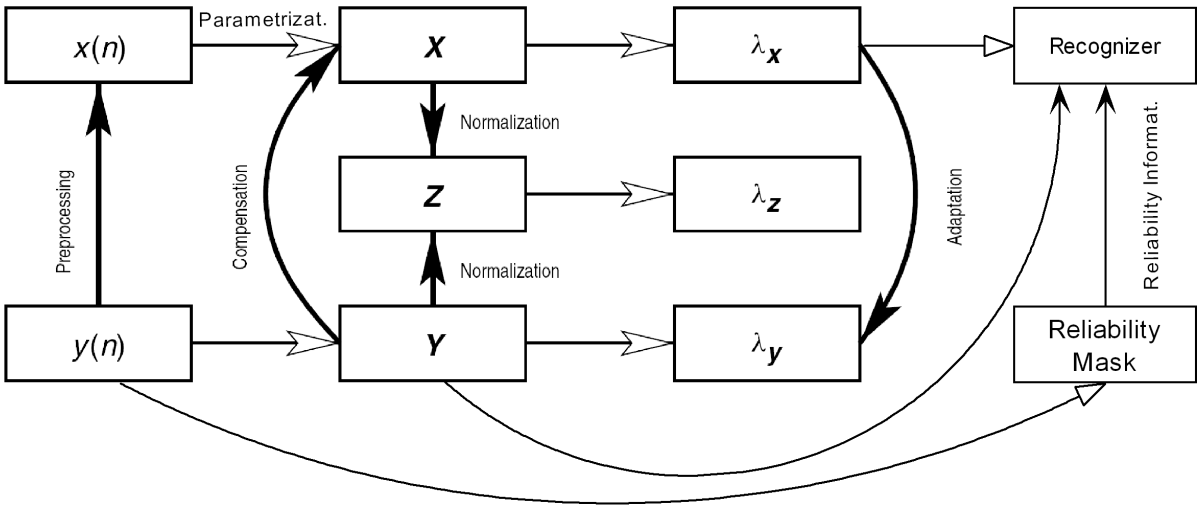



Figura 5.1: ([121] adaptada) Posible clasificación de las diferentes técnicas clásicas de robustecimiento.

**Compensación**: cuando se modifica la representación contaminada para hacerla lo más parecida posible a la limpia.

**Normalización**: cuando se transforman tanto la representación limpia como la distorsionada para llevarlas a un nuevo dominio menos afectado por el ruido (se aplica tanto en la etapa de entrenamiento como en la de test).

**Adaptación**: cuando se modifican los modelos limpios para hacerlos parecidos al entorno de test sucio.

**Procesamiento de incertidumbre**: cuando se tiene en cuenta la fiabilidad de cada uno de los segmentos de los parámetros de la representación acústica en el propio motor de reconocimiento.

En la Fig. 5.1 podemos ver un resumen de esta clasificación. Esta clasificación no es del todo completa ya que muchas de las técnicas existentes pueden encajar en varias clases a la vez y otras en ninguna. Cabrían otro tipo de clasificaciones como aquella basada en la cantidad de conocimiento del ruido requerida pero a pesar de todo, seguiremos clasificación anterior por su utilidad.

### 5.1.2. Técnicas de preprocesamiento y de parametrización robusta

**SWP** (SNR dependent Waveform Processing, SNR dependiente del Procesamiento de la Forma de Onda) [92] primero hace una búsqueda temporal (por medio de la extracción de la envolvente) de picos importantes en la señal respetando una separación mínima entre picos (en el caso de una señal sonora estos picos se corresponderán con los pulsos glotales, separados un periodo de pitch). Después se multiplica cada muestra por un peso amplificador o atenuador dependiendo de si la muestra es cercana al pico máximo del pitch o no. El efecto global es que aumenta la SNR de la señal. Esta técnica se suele aplicar normalmente sobre una señal que ya ha sido limpiada previamente mediante algún otro tipo de técnica de preprocesamiento que elimina ruido y devuelva el resultado en el dominio temporal. Técnicas de preprocesamiento de este tipo son las ventanas temporales (Hamming), filtrado offset y de preénfasis [149] (esta última mejora los resultados de reconocimiento realzando las altas frecuencias). Otras son el doble filtrado temporal de Wiener que lleva el AFE (Advance Front-End) [147], la técnica presentada en [151] y en general las técnicas de realce de la voz [83, 111] con aplicación al reconocimiento y que pueden llegar a ser consideradas como técnicas de compensación (Sec. 5.1.4).

**PLP** (Perceptual Linear Predictive, Predicción Lineal Perceptual) [56] de cada segmento de señal deriva un espectro tipo MEL (que intenta imitar el patrón de excitación de la membrana basilar, Sec. 3.1.3). De este se obtiene la autocorrelación y los parámetros LPC. De estos se deriva o un cepstrum o un espectro LPC según se desee. Minimiza la diferencia entre hablantes preservando la información relevante al habla. Se puede combinar con otras técnicas como RASTA [57]. Otras parametrizaciones robustas relacionadas son MFCC (Sec. 3.1.4) e incluso la técnica HASE (Sec. 5.2.1). En [135, 109] se puede ver una comparación de diferentes parametrizaciones robustas.

### 5.1.3. Técnicas de normalización

**HEQ** (Histogram Equalization, Ecualización del Histograma) [34] aplica una transformación a cada canal cepstral. Cada coeficiente cepstral del canal es cambiado por otro mediante esta función. Hace que el histograma de distribución de los coeficientes cepstrales contaminados se asemeje a uno de referencia (normalmente gaussiano). La transformación se obtiene a partir de la estimación del histograma contaminado. Se aplica en la etapa de

entrenamiento y de test haciendo más similares los vectores de características de ambas etapas.

**CMN**(Cepstral Mean Normalization, Normalización de Media Cepstral) [108, 84] obtiene la media de cada canal cepstral y esta media se resta a cada coeficiente cepstral del canal. Combate bastante bien los ruidos convolutivo y los aditivos muy estacionarios. Otra técnica de normalización relacionada es RASTA [57], CTN [146] e incluso VTLN aplicado a ASR robusto [48].

### 5.1.4. Técnicas de compensación

**SS** (Spectral Subtraction, Sustracción Espectral) [65, 121] da una estimación de la magnitud del vector espectral limpio restando al sucio una estimación del ruido de fondo. El ruido de fondo se puede estimar de muchas maneras [61, 129, 38, 42] pero clásicamente se estima a partir de las partes de silencio (empleando un VAD). En caso de mala estimación del ruido, dado que la magnitud no puede ser nunca negativa, se limita el valor mínimo del espectro limpio estimado. Esta limitación produce una distorsión conocida como ruido musical. Una interpretación muy usada es la de ver a la SS como un filtrado (multiplicación en el domino espectral [14]) dependiente de la SNR de cada píxel frecuencio-temporal. Según esta interpretación el espectrograma limpio se estima como:

$$\hat{X}(f,t) = Y(f,t)H_{ss}(f,t) \tag{5.1}$$

$$donde \quad H_{ss}(f,t) = \sqrt{max(1-\frac{1}{SNR(f,t)},a)} \tag{5.2}$$

$$donde \quad SNR(f,t) = \frac{Y(f,t)^2}{\hat{N}(f,t)^2} \tag{5.3}$$

donde $a$ es el factor de atenuación y suele estar en torno a 0.005. Para reducir el ruido musical, $SNR(f,t)$ y el filtro $H_{ss}(f,t)$ suelen ser suavizados en el tiempo y la frecuencia respectivamente [65]. SS da muy buenos resultados si el ruido está bien estimado (esto suele ocurrir en ruidos aditivos y bastante estacionarios). Existen muchas variantes no lineales para hacer frente al ruido musical como las de [40, 10, 73]. Otras técnicas muy relacionadas son VTS [72, 109], el filtrado de Wiener [12], e incluso los filtros de Kalman (o los de partículas que son una extensión de los de Kalman). Estos últimos intentan realizar a la misma vez la estimación del ruido y de la señal limpia [160, 163]. Un ejemplo típico de sistema de reconocimiento que incorpora la SS y que emplearemos en esta Tesis

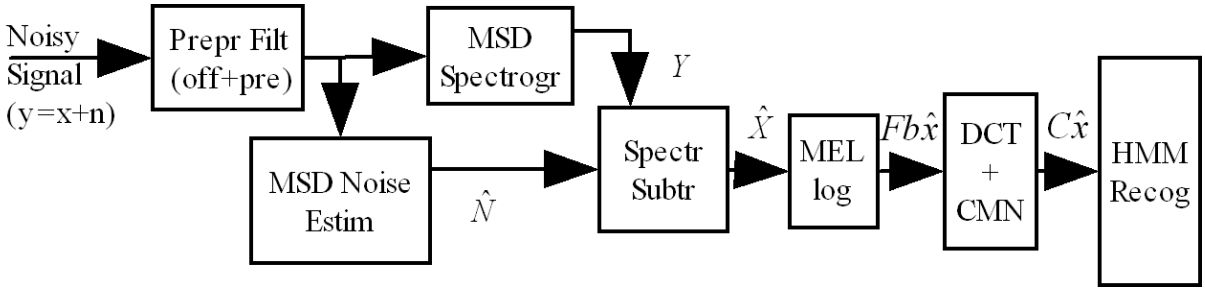



Figura 5.2: Sistema de reconocimiento que incorpora sustracción espectral.

lo podemos observar en la Fig. 5.2 donde podemos ver el estimador de la MSD (magnitud de la densidad espectral) del ruido ($\hat{N}$), la conversión al domino cepstral de la señal limpia estimada $C_x$ y el reconocedor basado en HMMs.

**Compensación MMSE** (Minimun Mean Square Error, Error Cuadrático Medio Mínimo): reemplaza cada vector cepstral sucio por una estimación MMSE limpia del mismo. La estimación MMSE se obtiene integrando sobre todos los valores posibles limpios ponderando cada uno de ellos por su correspondiente probabilidad de observación. Por ejemplo, en *VQ-MMSE Compensation* (Vector Quantization Minimun Mean Square Error Compensation, Compensación MMSE basada en Cuantización Vectorial) [51] la estimación se obtiene a partir de la media ponderada (o combinación lineal) de los diferentes vectores limpios estimados en cada una de las clases limpias cuantizadas. Los pesos de la ponderación son las probabilidades de que el vector limpio derive en el sucio observado. Estas probabilidades se obtienen de una base de datos estéreo (limpio-sucio) cuantizada. Si el ruido a atacar está registrado en la base de datos obtiene buenos resultados. Otras técnicas relacionadas son RATZ [108], SPLICE [36] y MEMLIN [20].

**Imputation Techniques** (Técnicas de Imputación) [27, 127] estiman las partes no fiables de la representación acústica contaminada (normalmente el espectrograma) empleando modelos de la representación limpia y reconocen con el cepstrograma. Están muy relacionadas con las técnicas de procesamiento de incertidumbre (Sec. 5.1.6).

### 5.1.5. Técnicas de adaptación de modelos

**PMC** (Parallel Model Combination, Combinación de Modelos Paralelos) [47] transforma las medias y covarianzas cepstrales de los HMM limpios en función del ruido ambiental. El nuevo modelo HMM resultante es muy parecido al que se obtendría entrenando con los vectores contaminados de ese ambiente. La transformación es una suma en el dominio espectral de la media limpia con la media del ruido que después hay que pasar al dominio

cepstral. Si se acierta con el modelo de ruido da muy buenos resultados de reconocimiento incluso con ruidos no estacionarios. Otras técnicas relacionadas son MLLR [79] (técnica similar pero más eficiente), descomposición HMM [150] y entrenamiento multicondición [120].

### 5.1.6. Técnicas de procesamiento de incertidumbre

Aunque podríamos considerar como técnicas de procesamiento de incertidumbre a un amplio grupo en el que se podrían incluir las técnicas de imputación de MD o de estimación Bayesiana de características, nos restringiremos exclusivamente a las que se aplican exclusivamente en el motor de reconocimiento. Estas técnicas no pretenden minimizar el desajuste entre entrenamiento y test, ni estimar las partes no fiables de la representación acústica. En lugar de ello, modifican el motor de reconocimiento HMM para que este tenga en cuenta la fiabilidad de la representación.

Esto evita los problemas, con respecto a los sistemas de información completa (compensación y demás), de tener que averiguar con mucha exactitud las partes de la voz enmascaradas por el ruido en las que puede ocurrir que la estima de las probabilidades del vector de características (por parte del decodificador) se vuelva muy inexacta si solo una o unas cuantas componentes del mismo no están estimadas con exactitud.

**Multistream Recognition** (Reconocimiento de Multi-Canales) [59, 15] toma una sección de señal, donde se espera que haya una unidad lingüística (p. ej. un fonema), y reconoce por separado cada uno de los canales espectrales (normalmente se toman 7 canales) obteniéndose una matriz de probabilidad para cada canal y cada unidad reconocida. Esta matriz es analizada en una etapa de mezcla de probabilidades para decidir finalmente la unidad lingüística presente. Si se sabe (mediante el conocimiento del ruido) qué canales deben ser desechados (por estar dominados por el ruido) la etapa de mezcla se simplifica mucho y se pueden obtener muy buenos resultados de reconocimiento. Si no se sabe qué canales deben ser desechados, la mezcla se complica y puede ser realizada de diversas formas heurísticas como: la lineal (en la que se ponderan las distintas probabilidades [15]) o la no lineal (en la que se emplean perceptrones multicapa [59] o modelos de unión probabilísticos [100]). Esta técnica es muy útil para ruidos estacionarios que dominan siempre los mismos canales espectrales.

**WVA** (Weighted-Viterbi Algorithm, Algoritmo de Viterbi con Pesos) [11] se basa en el uso de una estima muy simple de las características no fiables (p. ej. mediante una simple repetición del vecino más próximo) que luego es empleada en el decodificador de Viterbi

en la manera usual aunque pesando exponencialmente las probabilidades de observación con un peso relacionado con la fiabilidad de la observación acústica (0 no fiable, 1 fiable). Soft-Data [121], otra técnica relacionada, considera la fiabilidad de cada dato mediante una pdf (normalmente gaussiana) cuya anchura se traduce también en una modificación de las probabilidades de observación. Se puede demostrar que si la pdf es de evidencia uniforme este método degenera en la marginalización MD que veremos a continuación.

**Marginalización MD** (Missing Data, Datos Perdidos) [27] toma el espectrograma de la señal contaminada y mediante el empleo de una máscara, que indica qué coeficientes espectro-temporales son dominados por la voz frente al ruido, reconoce la señal de voz (Sec. 4.2). Tiene la virtud de no requerir ninguna suposición sobre el tipo de ruido a combatir por lo que en principio teniendo una buena estimación de la máscara da muy buenos resultados para cualquier tipo de ruido (estacionario o no estacionario). Existen infinidad de técnicas para estimar la máscara (Sec. 3.2.3 y 5.2.3). Una de las propuestas de esta Tesis se centra precisamente en la obtención de una máscara para aplicar la marginalización MD.

**SFD** (Speech Fragment Decoding, Decodificación de Fragmentos de Voz) [5], a partir de una fragmentación del espectrograma, genera todas las posibles máscaras (que nacen de combinar los diferentes fragmentos suponiendo que son de voz o de ruido), las reconoce todas y elige aquella cuya secuencia de palabras reconocida es la más probable en el modelo HMM. Los fragmentos se obtienen mediante reglas primitivas de CASA [155] (como agrupar píxeles que compartan un pitch común [90]). Es un claro ejemplo de técnica de pizarra donde se combinan reglas primitivas con reglas de alto nivel basadas en modelos. Da tan buenos resultados como MD pues en verdad su núcleo es un reconocedor de MD. La única diferencia es que, para ruidos que son difíciles de distinguir de la voz mediante reglas primitivas (como otras voces), hace más fácil la obtención de las máscaras. En [89] se muestra como pueden ayudarse mutuamente MD y SFD.

### 5.1.7. Debilidades de las técnicas convencionales

En general no existe la técnica perfecta para resolver el problema de la robustez. Todas pueden tener, aparte de las virtudes antes mencionadas, alguno de los siguientes cuatro defectos los cuales pueden ser usados para compararlas:

### Combatir solo determinados tipos de ruido

Este es el defecto de que la técnica es demasiado *ad hoc*. Por ej. CMN combate muy bien ruidos convolutivos lentos pero falla para ruidos aditivos no estacionarios. Multistream va bien en ruidos estacionarios de canal pero no para ruidos no estacionarios o que siendo estacionarios ocupen solo una fracción de un canal. La compensación MMSE puede fallar si el ruido presente no está considerado en la base de datos estéreo. En general solo MD y SFD se libran de este defecto.

### Depender de otras técnicas

Este es el defecto de que la técnica *pasa el problema a otra*. Por ej. MD y SFD necesitan de un buen extractor de máscaras o de un buen segmentador para su buen funcionamiento. Algo parecido pasa con SS que normalmente depende de un buen VAD, o de PCM que requiere de un buen reconocedor de ruido de ambiente. En general solo las técnicas muy básicas como las de normalización se libran de este defecto.

### Tener un elevado coste computacional

Este defecto debe de evaluarse en función de la cantidad de tipos de ruidos que combata y de las otras técnicas asociadas que dependan de esta. Por ello, quizás es más apropiado hablar del defecto de que la técnica *no combate eficientemente los ruidos* para los que está diseñada. Por ej. técnicas como MEMLIM o PCM sufren de este defecto ya que pueden ser sustituidas por técnicas similares más eficientes como VQ-MMSE Compensation o MLLR, respectivamente. Técnicas como CMN (que combate ruido convolutivo de forma muy eficiente) o SFD (que posee un algoritmo inteligente de Viterbi para probar a la vez muchas máscaras) no sufren de este defecto.

### No imitar el reconocimiento humano

Este es el defecto de que la técnica *no es biomimética*. Aunque no es un defecto crítico, tiene su importancia en el hecho de que la experiencia ha mostrado que las técnicas que se enfrentan al ruido de forma parecida a como lo hace el humano, son más eficientes en el sentido de emplear solo las pistas que de verdad son importantes en el reconocimiento, soliendo combatir más cantidad de tipos de ruido. De lo que se sabe sobre la forma en que el ser humano combate el ruido podemos decir que las técnicas que no trabajan con modelos limpios (como multicondition) no imitan la forma humana mientras que las

técnicas de procesamiento de incertidumbre y en especial SFD, por lo que dicen las reglas ASA, tienden a imitar en mayor medida el reconocimiento humano.

## 5.2. Técnicas de robustecimiento basadas en el pitch

Las técnicas de reconocimiento robusto basadas en el pitch pueden ser divididas en tres grandes grupos dependiendo del uso que hagan del pitch. Estos grupos son: Las que se basan en aprovechar la **estructura armónica** (que no emplean la estimación del pitch de cada segmento de señal, pero si ciertas propiedades derivadas de la periodicidad o de la estructura armónica de la señal), las que se basan en **estimar la voz limpia** (que sí que emplean directamente el pitch para estimar la señal limpia) y las que se basan en **estimar máscaras** (que también emplean el pitch de cada segmento para indicar qué píxeles del cocleograma están dominados por la voz frente al ruido). A continuación estudiaremos las técnicas más importantes del estado del arte para reconocimiento robusto basado en el pitch.

### 5.2.1. Técnicas de apovechamiento de la estructura armónica

**HASE** (High-lag Autocorrelation Spectrum Estimation, Estimación Espectral con coeficientes Altos de la Autocorrelación) [142] obtiene la OSA (One Sided Autocorrelation o Parte positiva o negativa de la Autocorrelación) de un segmento de señal, elimina los primeros $L$ coeficientes (presumiblemente degradados por ruido), les aplica una ventana como la DDR (Double Dynamic Range, Rango Dinámico Doble) y obtiene una estimación del espectro limpio. Una ventana DDR de tamaño $L$ se obtiene convolucionando con ella misma una Hamming de tamaños $L/2$. De estos espectros se obtiene el cepstrograma AMFCC (Sec. 3.1.4) que es finalmente enviado al reconocedor. HASE funciona bien en ruidos poco autocorrelados tipo blanco (ruidos cuya autocorrelación se hace pequeña a partir del coeficiente $L$). En los segmentos sonoros es posible probar que el espectro HASE es muy similar al espectro limpio habitual (con toda la autocorrelación o con toda la OSA). Para ello hay que tener en cuenta que su OSA posee una estructura periódica (se repite cada periodo del pitch), y que esto provoca que la información referente a la envolvente espectral no solo se encuentre en los primeros coeficientes de autocorrelación eliminados, sino también en sus respectivas repeticiones (efecto de modulación en el dominio de la autocorrelación). En los segmentos sordos esto no ocurre habiendo un mismatch o desajuste entre

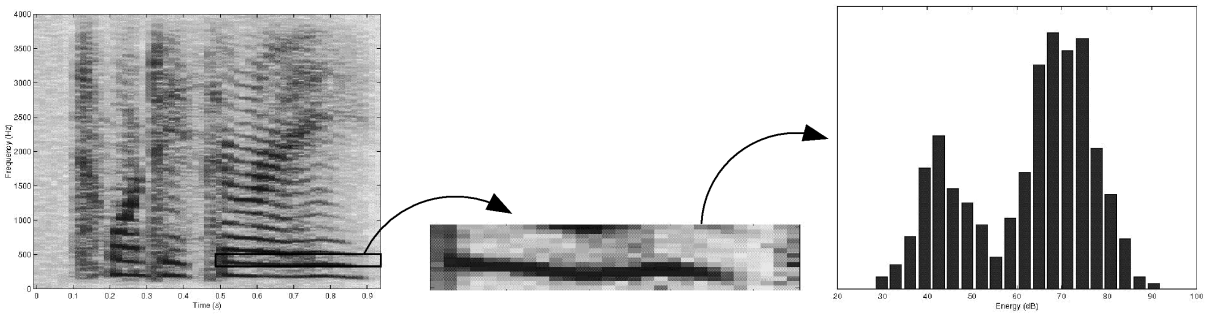


Figura 5.3: Filtrado armónico u obtención del nivel de ruido de un segmento (con varios armónicos de la voz) del espectrograma estrecho a partir del histograma de energías propuesto en [129].

espectro HASE y espectro limpio. Sin embargo, se puede evitar este mismatch aplicando HASE tanto al entrenamiento como al test.

Según todo esto HASE puede ser considerada como una técnica de normalización e incluso de parametrización robusta por lo que está relacionada con técnicas como PLP (Sec. 5.1.3). Otras técnicas similares son Cyclic-Spectrum [113], OSALPC [60], SMC [93] y LSMYWE [94] que se basan en el empleo de los coeficientes altos de la autocorrelación teniendo en cuenta que estos también contienen información sobre la envolvente espectral. Algunas de las técnicas propuestas en esta Tesis se inspiran en HASE.

**HF** (Harmonic Filtering, Filtrado Armónico) [129] mejora la estima del espectrograma del ruido realizada por cualquier técnica clásica de estimación de ruido tal como la basada en el histograma de Hirsch [61] o las basadas en un VAD. Para ello, obtiene un espectrograma estrecho (que permita distinguir los armónicos de la voz sonora), toma segmentos de este espectrograma de longitud 0.5 segundos y de ancho 200 Hz, obtiene el histograma de energías de cada segmento y teniendo en cuenta que los armónicos de la voz sonora tendrán energía más alta que el ruido, obtiene el valor de energía del ruido para ese segmento. La Fig. 5.3 muestra un ejemplo de este proceso. Cuando el ruido es armónico esta técnica puede dar un valor erróneo, por lo que el valor final del ruido para ese segmento lo da un algoritmo que mezcla la estima clásica con la estima HF. Esta técnica mejora los resultados de las técnicas clásicas cuando el ruido es poco estacionario. El ruido final estimado se puede aplicar sobre técnicas como SS o en MD. Una técnica relacionada, en el sentido de que mejora la estimación del ruido empleando la estructura armónica de la voz, es la basada en la envolvente LPC [42].

SWP (estudiada en la Sec. 5.1.2) podría ser incluida como otra técnica que emplea la estructura armónica de la señal debido a que saca partido de los pulsos glotales para

hacer robustecimiento. Otras técnicas que han empleado la estructura armónica, ya no con aplicaciones al reconocimiento robusto si no al realce de la voz, han sido [164, 76] (basadas en estimación MMSE espectral de la voz limpia) y [163] (basada en filtros de Kalman para seguir el pitch y los formantes). Estas tres técnicas mejoran el problema del ruido musical que provoca la SS tomando en consideración la estructura armónica de la voz.

### 5.2.2. Técnicas para estimación de la señal limpia

**WHNM** (Weighted Harmonic+Noise Model, Pesado basado en Modelo Hamonico+Ruido) [138] obtiene de cada segmento de señal ruidoso $y$ la señal armónica $y_h$ (se puede demostrar que es como la IDFT del espectro resultante de muestrear el espectro ruidoso cada múltiplo del pitch) y su correspondiente señal aleatoria o de ruido $y_r = y - y_h$. El espectro Mel limpio estimado se obtiene mediante la siguiente ecuación:

$$\hat{X} = \alpha_h Y_h + \alpha_r Y_r, \quad 0 \leq \alpha_h, \alpha_r \leq 1 \tag{5.4}$$

donde $Y_h$ es el espectro Mel de la señal armónica e $Y_r$ el de la señal aleatoria. El valor de $\alpha_h$ es una medida de la SNR del segmento y se obtiene como:

$$\alpha_h = \frac{\sum_i y_h(i)^2}{\sum_i y(i)^2} \tag{5.5}$$

$\alpha_r$ es constante y aproximadamente igual a 0.10 (estimado de forma experimental). Si el segmento de señal no tiene pitch (es sordo o de silencio) se pone uno ficticio de 150 Hz (valor no importante en el resultado final) y se aplica la misma técnica. Una vez obtenido el espectrograma Mel se obtiene el cepstrograma y se reconoce. Esta técnica va bien cuando el ruido no es armónico y la SNR no es muy baja.

Otras técnicas relacionadas son PHCC [52] (vectores de características robustos basados en darle más peso a los armónicos del pitch), la técnica de Kuroiwa [77] (basada en obtener una señal periódica promedio a partir de muestras de diferentes periodos), y la técnica de Parson [116] (que separa señales armónicas con diferentes pitch). Estas tres técnicas se pueden reducir a variantes temporales o espectrales de filtros peine (comb filter) que muestrean la señal en los armónicos espectrales del pitch [83, 111].

**FPM-SE** (Fine Pitch Model Signal-Estimation, Estimacion de Señal basada en Modelo Fino del Pitch) [19] es una técnica muy relacionada con WHNM aunque algo más

complicada. Esta estima la señal limpia en cada instante de tiempo de la siguiente manera:

$$\hat{x}(n) = \gamma(n)y(n) + (1 - \gamma(n))a(n)\hat{x}(n - \tau(n)) \tag{5.6}$$

donde $\tau(n)$ es el pitch en ese instante de tiempo (obtenido mediante un extractor fino o preciso como el de [37]), y donde $\gamma(n)$ y $a(n)$ son obtenidos mediante complejas estimaciones MMSE (Sec. 5.1.4). Las probabilidades MMSE de los diferentes valores de $\gamma(n)$ y $a(n)$ son dependientes de unos datos de entrenamiento estéreo con ruido y de la señal sucia observada. Esta técnica va bien cuando el ruido está incluido dentro de los datos de entrenamiento. La técnica CASA de Weintraub [158] está relacionada con esta en el sentido de que emplea datos de entrenamiento para comenzar la estima de la señal limpia.

**HT** (Harmonic Tunnelling, Tunelaje Armónico) [38] trata de obtener una estima del espectrograma del ruido a partir de los picos armónicos (dependientes del pitch) de cada segmento. Los picos armónicos y el pitch los obtiene de la siguiente manera: obtiene el espectrograma estrecho de la señal ruidosa, localiza los picos espectrales más significativos mediante derivadas, obtiene una primera estima del pitch basándose en la autocorrelación del espectro de cada segmento y obtiene la estima final del pitch basándose en esta primera estima del pitch y en tres medidas (local, global y temporal) que nos indican la probabilidad de que cada pico anterior forme parte del pitch final. Los picos armónicos (por lo general relacionados con los armónicos de la voz sonora) se obtienen cribando o eliminando aquellos picos que tengan baja probabilidad. En las Fig. 5.4 podemos observar el espectrograma estrecho, los picos iniciales detectados y los picos armónicos finales tras la criba.

El ruido lo obtiene buscando los túneles o las regiones entre-picos espectrales supuestamente dominados por el ruido. Para ello aplica un algoritmo que va tomando parejas de picos adyacentes y decide donde residen los límites de comienzo de los túneles. Una vez obtenidos los túneles aplica una interpolación y un suavizado que tiene en cuenta estos túneles para obtener una estima final del espectrograma del ruido. Con este ruido se obtiene el espectrograma limpio (a partir de una SS dependiente de la SNR) y su correspondiente cepstrograma el cual se envia al reconocedor. Esta técnica va bien cuando el ruido es poco estacionario y se puede conjugar con otras técnicas de estimación de ruido tal y como se ha hecho en [72, 165] para realce de la voz. Otras técnicas relacionadas con esta son FPM-NE [19] o la de Frazier [46] basadas en variantes temporales de filtros peine (con respuesta impulsiva tipo $h_T(t) = \delta(t) - \delta(t - T)$) que obtienen el ruido que hay entre los armónicos del pitch. HT tiene el defecto de no considerar los sonidos sordos y de ser

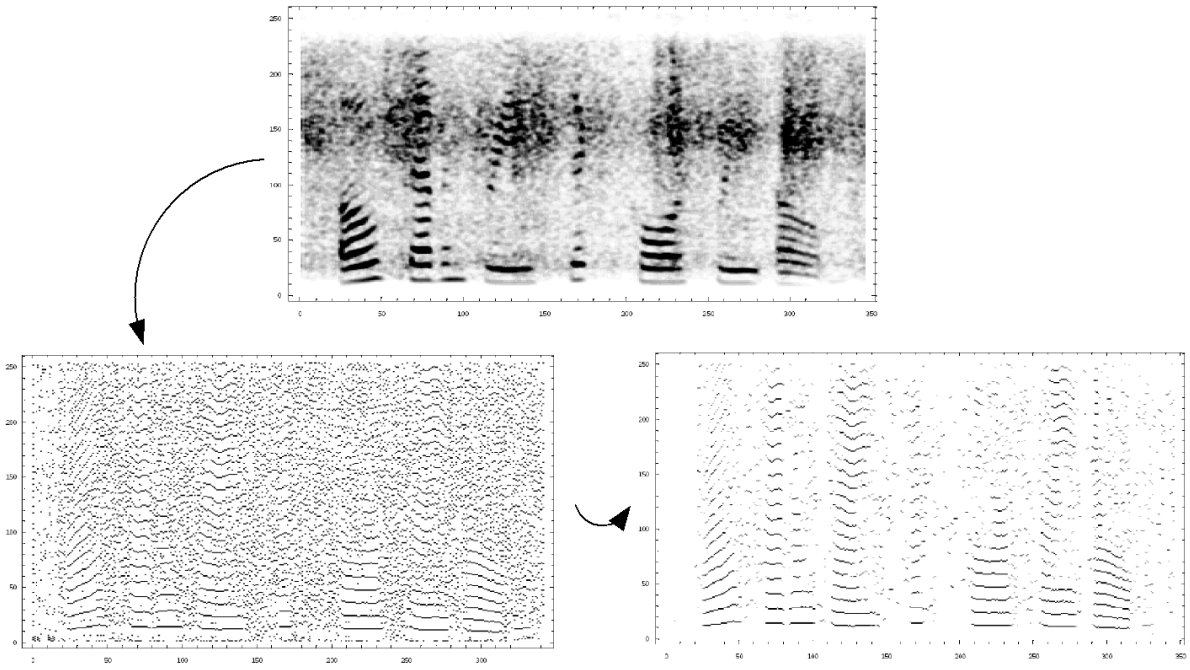


Figura 5.4: Espectrograma estrecho, picos iniciales detectados y picos armónicos finales tras la selección. Estos picos finales son empleados en el tunelaje armónico de [38].

sensible a la no precisión en la estima del ruido. Una de las técnicas propuestas en esta Tesis propone una variante de HT que evita este tipo de defectos.

### 5.2.3. Basadas en estimar máscaras

**La técnica de Barker** [9, 6] supone que la voz es la única fuente armónica de la señal (el ruido es inarmónico). La resumimos en los siguientes cuatro pasos: 1) Se extrae el pitch $p(t)$ de cada segmento de señal mediante el máximo de la autocorrelación sumada $SA_y(t,k)$ del correlograma contaminado $A_y(f,t,k)$ (ver Sec. 3.3) y se obtiene una medida de la sonoridad de cada segmento de señal como $V(t) = SA_y(t,p(t))/SA_y(t,0)$. 2) Se estima la armonicidad de cada píxel como $H(f,t) = A_y(f,t,p(t))/A_y(f,t,0)$ y se pasa esta armonicidad a través de una sigmoide para obtener la «máscara armónica» $M_h$ analógica. 3) Por otro lado se obtiene la «máscara ruido» $M_n(f,t)$ analógica basada en la SNR local de cada píxel (Sec. 3.2.2) mediante una estimación del cocleograma del ruido ($\hat{N}_{gam}$) basada en los 10 primeros segmentos del cocleograma contaminado ($Y_{gam}$). 4) La máscara final de la voz es una combinación lineal de ambas máscaras (donde domina la máscara $M_h$ si la sonoridad es alta y donde domina la $M_n(f,t)$ si la sonoridad es baja).

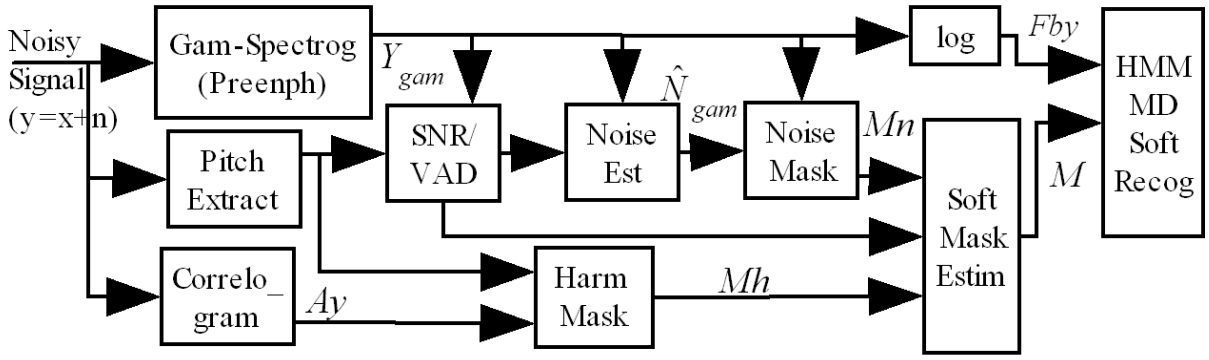



Figura 5.5: Sistema de reconocimiento basado en la técnica de Barker [6] para los propósitos de esta Tesis. Se estiman dos máscaras, una ($M_n$) basada en la estimación mediante un VAD del ruido y otra ($M_h$) basada en la armonicidad mediante el correlograma. La máscara final es una combinación lineal de ambas máscaras.

Esta técnica da buenos resultados pero siempre y cuando el ruido no tenga armonicidad (ruido tipo voz).

Una técnicas relacionada con ésta es la propuesta en [128], la cual hace una estimación de la SNR local de cada píxel usando un modelado Gaussiano del ruido para producir una máscara analógica. Otra relacionada es la propuesta en [139] la cual se basa en usar unas características especiales (entre ellas la armonicidad basada en filtros peine) que hacen que la distinción voz-ruido (mediante un clasificador bayesiano) sea robusta y casi independiente del ruido.

En la Fig. 5.5 podemos ver un sistema de reconocimiento basado en la técnica de Barker, el cual emplearemos en esta Tesis. Podemos observar la estima de la máscara-armónica analógica $M_h$ (basada en el correlograma $A_y$) y la estima de la máscara-ruido analógica $M_n$ (basada en una estima del cocleograma del ruido $\hat{N}_{gam}$). Las diferencias con la técnica de Barker son: primero, que la estima del ruido es más completa por estar basada en un VAD (bloque *SNR/VAD*) y, segundo, que en los segmentos de señal con pitch se pone directamente $M_h$ y en los otros $M_n$ (bloque *Soft Mask Estimation*) es decir, no se hace una combinación lineal de las máscaras.

**La técnica de Brown** [18, 155] se basa en agrupar segmentos siguiendo las reglas computaciones de ASA [155] y su idea principal consiste en agrupar píxeles que tengan un contorno de pitch similar. La resumimos en los siguientes cuatro pasos: 1) Se obtienen segmentos de píxeles (pequeñas agrupaciones de píxeles) que compartan similar modulación FM y/o que tengan alto correlograma-cruzado . 2) Se extrae el contorno de pitch de cada segmento mediante la autocorrelación sumada y un suavizado. 3) Se van

comparando los distintos segmentos (empezando por el mayor) y se van agrupando si su medida de similitud es parecida. Esta medida de similitud tiene en cuenta que tengan un comienzo/final común y un contorno de pitch parecido. El proceso termina cuando ya no se pueden agrupar más segmentos teniendo al final, al menos, una gran agrupación de píxeles que se corresponderá con la máscara de los sonidos sonoros. 4) La máscara final de la voz se puede obtener combinando la máscara sonora junto con alguna otra técnica que obtenga la máscara de los sonidos sordos tal como el "algoritmo watershed" [31]. El origen de esta técnica lo podemos encontrar en la propuesta de Cooke en [29].

Otra técnica relacionada con esta, en el sentido de que intenta seguir reglas de agrupamiento ASA, es la de Hu y Wang [64, 155], la cual tiene en cuenta la evidencia psicoacústica de que el ser humano trata las bajas frecuencias de forma diferente a las altas [74]. Para ello obtiene el pitch mediante una red neuronal de osciladores [17]. La armonicidad de las bajas frecuencias la obtiene como en la técnica de Barker y la de las altas frecuencias comparando la envolvente AM de las salidas del banco de filtros (Sec. 3.1.2) con un seno de frecuencia la del pitch.

**La técnica de Ma** [90] se basa en obtener fragmentos de voz (sonoros y sordos, dominados por una única fuente) y obtener la máscara final de la voz mediante un reconocedor SFD (Sec. 5.1.6). Está pensada para trabajar con ruido tipo voz (voz+voz). La resumimos en los siguientes seis pasos: 1) Se obtienen pequeños grupos de píxeles que tengan alto correlograma-cruzado [155]. 2) Se hace agrupamiento espectral por cada segmento de señal. Para ello, mediante un filtrado de Gabor, se realza el correlograma para obtener 0, 1 o 2 dendritas (cada dendrita es consecuencia de un pitch, y dado que es para voz+voz, como máximo habrá 2 dendritas) y en consecuencia 0, 1 o 2 grupos sonoros de píxeles asociados a cada dendrita. De esto se derivan de 1 a 4 candidatos a pitch por cada segmento de señal. 3) Se obtienen segmentos de pitch (sin identificar a que voz pertenecen) mediante un suavizado temporal (basado en HMMs [30]) que deja solo en dos los cuatro candidatos a pitch anteriores. 4) Se hace agrupamiento temporal uniendo los grupos espectrales de píxeles que forman un mismo segmento de pitch (en los cruces de segmentos de pitch se comienzan a obtener nuevos agrupamientos). 5) Se obtienen grupos inarmónicos (producidos por sonidos fricativos) mediante el "algoritmo watershed" [31]. 6) Finalmente cuando se tienen todos los grupos o fragmentos de voz se emplea un reconocedor SFD para para reconocer y agrupar los fragmentos mediante el empleo de los modelos de las palabras a reconocer. El resultado final es la frase reconocida junto con su máscara de reconocimiento.

Mejoras de esta técnica se han propuesto recientemente en [8] (donde se hace reconocimiento dependiente del hablante) y en [89] (donde se mezcla MD con SFD obteniendo mejoras y mostrando que ambas técnicas sirven para ruidos complementarios tales como los estacionarios y los impulsivos).

### 5.2.4. Debilidades de las técnicas basadas en el pitch

Hacer una comparación justa de las diferentes técnicas basadas en el pitch es difícil. Entre los motivos más importantes de esta dificultad destacan:

1. El hecho de emplear cada una un extractor de pitch y una base de test diferente: este es el caso si comparamos WHNM, HT y la técnica de Barker (que cada uno usa un extractor de pitch diferente); o de la técnica de Ma (que es evaluada sobre una base de ruidos compuesta por voces y no sobre ruido no-vocal como es más usual).

2. No dejar claro si se está hablando de una nueva técnica para reconocimiento robusto, de un nuevo extractor de pitch robusto o ambas cosas a la vez: este es el caso de HT, de la técnica de Brown y de la técnica de Ma que incluyen su propio extractor de pitch, sin embargo, técnicas como WHNM son propuestas como nuevas formas de robustecer los segmentos sonoros conocido el pitch.

3. No saber al compararlas, de donde proviene la fuente de la mejora: si por el empleo de diferentes mecanismos de robustecimiento sobre los silencios y los sonidos sordos, siendo el de los sonoros el mismo (p. ej. la técnica de Ma obtiene máscara en los sonidos sonoros de forma muy similar a la de Barker, sin embargo, difieren en la forma de extraerla en los sonidos sordos y los silencios); si por los conocimientos previos sobre el ruido empleados (FPM-SE entrena la técnica para ruidos similares a los que va a combatir); si por las técnicas extra añadidas (tales como CMN, SWP, frame-dropping [35], etc.) o si por el esquema de reconocimiento empleado (SS, MD, SFD, etc.).

De estas dificultades se deriva la necesidad de buscar equivalencias entre las diferentes técnicas para poder compararlas de una manera adecuada y el Cap. 7 se dedica a ello. A pesar de estas dificultades, y de forma similar a como hicimos con las técnicas convencionales de reconocimiento robusto, podemos encontrar los siguientes defectos en las distintas técnicas de pitch, los cuales pueden ser usados para compararlas.

### No abordar todo tipo de ruidos

La técnica HF sufre este problema porque su estimación de ruido falla cuando la SNR es muy baja o cuando el ruido aumenta repentinamente (en estos casos no se observa distinción en el histograma entre el ruido y la voz). Similar problema tiene HASE (que no es capaz de abordar ruidos armónicos), FPM-SE (que puede fallar si el ruido no se ha empleado en el entrenamiento) o la técnica de Ma (que está orientada a ruido tipo voz).

### El Problema de los sonidos sordos

Muchas de las técnicas propuestas no indican qué hacer con los sonidos sordos llegando a eliminar su información y a hacer solo reconocimiento con los sonoros. Tal es el caso de HASE que elimina prácticamente la información de los sonidos sordos, aunque el problema es aliviado al usar HASE en ambas fases, test y entrenamiento. Problemas similares lo tienen WHNM (que suponiendo un pitch ficticio para los sordos elimina parte de su información), HT (que llega a tomar como ruido los sonidos sordos) y la técnica de Brown (que no indica qué hacer con los sonidos sordos).

### Necesitar de un pitch preciso

FPM-SE sufre de este defecto ya que una pequeña desviación en el valor de pitch podría provocar que la diferencia entre periodos no sea correcta. WHNM también sufre este defecto pues es en el fondo es un muestreo espectral en cada armónico del pitch. Técnicas como HT no sufren tanto de este problema debido a que estimar el ruido entre los huecos de los armónicos del pitch requiere menos precisión que estimar el armónico con precisión. Las técnicas de estimación de máscaras, al trabajar con el correlograma, no sufren tanto este defecto.

### Detectar el pitch de forma imprecisa

Esto se refiere a las técnicas que proponen al mismo tiempo un nuevo mecanismo de robustecimiento junto con un nuevo extractor de pitch el cual puede tener el defecto de no ser robusto. Tal es el caso de la técnica de Barker basada solo en tomar como pitch el máximo de la autocorrelación. HT también sufre de este defecto (ya que el extractor de pitch que propone no da muy buenos resultados de reconocimiento).

### Ser compleja y no biomimética

Teniendo en cuenta lo que se conoce sobre la forma humana de reconocer, se puede decir que ninguna técnica se asemeja completamente al ser humano (no es biomimética) salvo quizás, la técnica de Ma que tiene en cuenta conceptos de ASA. A pesar de esto, esta técnica, que en el fondo no es más que un separador de valores de pitch, tiene el defecto de abusar de las reglas de alto nivel (o basadas en modelo) para separar y asociar los valores de pitch de los dos hablantes en situaciones donde el ser humano lo hace de manera más sencilla (p. ej. teniendo en cuenta la diferencia de altura entre valores de pitch) por lo que podemos decir que esta técnica es compleja computacionalmente hablando respecto a la forma en que lo hace el hombre. Algo parecido podemos decir respecto a FPM-SE (que requiere de un cómputo elevado para obtener un pitch preciso, entrenar los datos estéreo y estimar las probabilidades MMSE).

# Capítulo 6

# Técnicas Propuestas

## 6.1. Ventanas asimétricas

### 6.1.1. Introducción

La técnica que presentamos a continuación [107] es una técnica que intenta, con poca cantidad de cálculo y sin hacer estimación del ruido, obtener vectores de características más robustos. Debido a que intenta disminuir la diferencia test-entrenamiento puede ser considerada como una técnica de parametrización robusta (Sec. 5.1). Debido a que para su justificación emplea la estructura armónica de la señal de voz (más que el pitch de cada segmento) puede ser considerada como una técnica de pitch basada en la estructura armónica (Sec. 5.2.1).

Esta técnica está inspirada en la técnica *HASE* (High-lag Autocorrelation Spectrum Estimation) [142] estudiada en detalle en la Sec. 5.2.1. Esta se basa en anular los primeros coeficientes de la OSA (One Side Autocorrelation) los cuales están más contaminados por el ruido, para obtener una estimación limpia del espectro. Este procesado puede ser interpretado alternativamente como una ventana asimétrica aplicada a la OSA. De aquí surge la idea de buscar una ventana asimétrica adecuada y que pondere adecuadamente las repeticiones debidas al pitch que se observan en la función OSA.

Las secciones subsiguientes explicarán esta técnica y la compararán solo con HASE debido a que esta última supera en resultados a otras muchas relacionadas tales como SMC [93] y OSALPC [60].

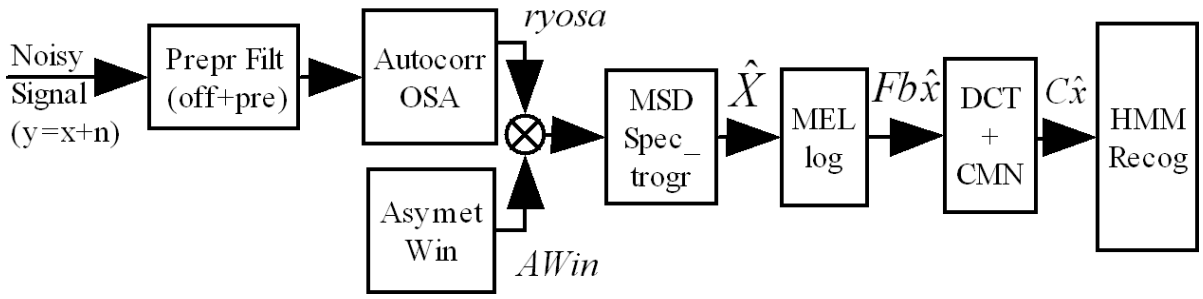



Figura 6.1: Sistema de reconocimiento donde se ve como se aplica la técnica de las ventanas asimétricas sobre la OSA.

### 6.1.2. Sistema de reconocimiento

En la Fig. 6.1 podemos observar el sistema de reconocimiento propuesto para estimar y evaluar las ventanas asimétricas propuestas de estima del espectro limpio. Este toma como entrada la señal ruidosa de una frase, la cual es suma de la voz limpia y el ruido ($y = x+n$). El bloque *Preprocesing Filter* filtra la señal contaminada mediante un filtrado de offset y de preénfasis (Sec. 3.1.3). Este último realza las altas frecuencias. El bloque *Autocorrelation OSA* obtiene la autocorrelación ruidosa OSA ($r_y\hat{o}sa$) de cada segmento de señal y el bloque *Asymetric Window* proporciona una ventana asimétrica que se aplica (mediante multiplicación) sobre esta OSA. Los tres bloques siguientes se encargan de obtener el cepstrograma (Sec. 3.1.4). *MSD Spectrogram, Mel-log y DCT* obtienen el espectrograma de la densidad de la magnitud espectral ($\hat{X}$), la representación en el banco de filtros ($F\hat{b}_x$) y el cepstrum AMFCC ($\hat{C}_x$), respectivamente, a partir de la OSA enventanada. Para obtener una densidad de magnitud espectral con energía similar a la que obtendríamos empleando toda la autocorrelación, habría que multiplicar por dos la MSD de la Ec. 3.4 (Sec. 3.1.3) empleando la OSA enventanada en lugar de $r_x$. Finalmente, la estima cepstral es pasada al *HMM Recognizer* para obtener una transcripción de la frase.

### 6.1.3. Conjunto de ventanas asimétricas

Suponiendo que entrenamos y testeamos con la misma ventana, la búsqueda de una ventana adecuada de reconocimiento se puede ver como un problema de optimización en los resultados de reconocimiento en función de los pesos que se aplican sobre la OSA. Hacer una búsqueda exhaustiva de esta manera es inabarcable computacionalmente, pues supondría hacer una cantidad ingente de pruebas de reconocimiento. Teniendo en cuenta

esto, limitaremos la búsqueda a un conjunto de posibles ventanas seleccionado, que reducirá la búsqueda a un problema bidimensional. La elección de este conjunto de ventanas se basa en los tres criterios heurísticos siguientes:

1. Los coeficientes bajos de la OSA deben de tener menos peso debido a que suelen ser los más contaminados por el ruido. Este criterio es bastante conocido y en varias técnicas como HASE se demuestra su efectividad.

2. Debe de haber un conjunto de coeficientes que deben de tener más peso que el resto debido a que suelen estar menos afectados por el ruido y debido a que transportan más información lingüística (como mostraremos más adelante estos se corresponderán con los múltiplos del pitch).

3. El conjunto va a incluir ventanas típicas que se hayan ya empleado sobre la OSA tales como la DDR (Double Dynamic Range) o la ventana HASE de Shannon (Sec. 5.2.1).

Teniendo en cuenta estos tres criterios proponemos el siguiente conjunto de ventanas asimétricas $DDR_{c,w}$ dependientes de dos parámetros:

$$DDR_{c,w}(k) = \begin{cases} DDR_w(\frac{w}{2} - (c+1) + k) & c - \frac{w}{2} < k \leq c + \frac{w}{2} \\ 0 & otherwise \end{cases}$$
$$k = \{0, ..., L-1\} \tag{6.1}$$

donde $L$ es el tamaño total que coincide con el de la OSA, $c$ el centro y $DDR_w$ es una ventana DDR de anchura $w$ que es obtenida autocorrelacionando un ventana de Hamming de tamaño $w/2$. En la Fig. 6.2 podemos ver una de las ventanas asimétricas de este conjunto, la $DDR_{50,250}$ superpuesta a la OSA de una vocal.

Este conjunto cumple con los tres criterios anteriores de manera que variando $c$ podemos darle mayor peso a ciertos coeficientes (criterio 2), variando $w$ podemos aumentar o disminuir el peso dado a los primeros coeficientes de autocorrelación (criterio 1) y fijando los parámetros a por ejemplo $L = 256$, $c = 135$ y $w = 240$ ($DDR_{135,240}$) podemos obtener una de las típicas ventanas empleadas sobre la OSA (criterio 3), que con estos valores coincide la HASE de Shannon.

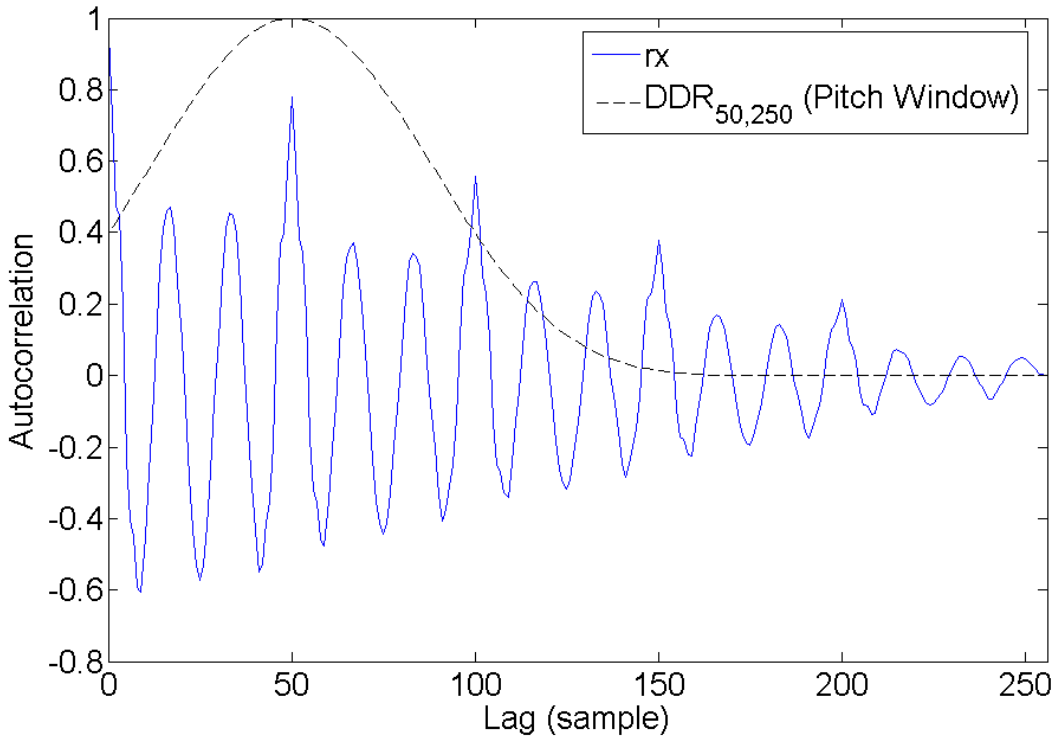



Figura 6.2: Ejemplo de una ventana asimétrica $DDR_{50,250}$ aplicada sobre la OSA de un segmento sonoro de una vocal con pitch 50 muestras.

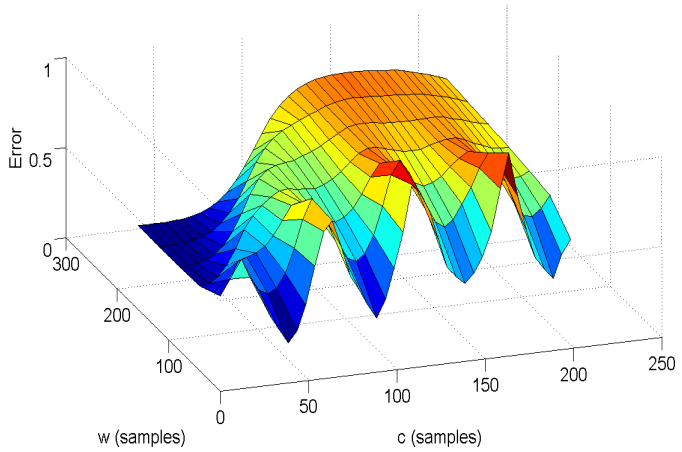



Figura 6.3: Superficie de error cepstral $Err(c, w)$ para un segmento sonoro (pitch=50 muestras) contaminado con ruido blanco en función del centro $c$ y ancho $w$ de la ventana de análisis $DDR_{c,w}$. Se observan mínimos de error cepstral cuando la ventana está centrada sobre los coeficientes del pitch ($c = 50, 100, 150, ..., etc.$).

.

### 6.1.4. Ventana para segmentos sonoros

#### Anchura y centro de las ventanas

A continuación estudiaremos cual de nuestras ventanas $DDR_{c,w}$ es la mejor para segmentos de señal sonoros. Para hacerlo, un segmento de voz sonora limpia ha sido contaminado con diferentes realizaciones de un ruido blanco a una SNR de 0dB. Más concretamente, este segmento de voz limpia ha sido extraído de una vocal «e» con un pitch de 50 muestras. Mostramos su OSA en la Fig. 6.2.

Lo que buscamos es qué ventana es la que genera un menor desajuste entre las representaciones cepstrales limpia y ruidosa. La superficie de error obtenida variando los parámetros $c$ y $w$ es dibujada en la Fig. 6.3 donde el error es la distancia promedio entre el cepstrum AMFCC limpio $C^{c,w}$ y los diferentes cepstrums AMFCCs ruidosos $C^{c,w}_{yn}$ cuando una ventana $DDR_{c,w}$ es aplicada sobre ambos. La siguiente ecuación muestra como se obtiene este error:

$$\text{Err(c,w)} = \frac{1}{N}\sum_{n=1}^{N} dist(C^{c,w}_{x}, C^{c,w}_{y_n}) \tag{6.2}$$

donde $dist$ es la distancia euclídea y $N$ el número de diferentes segmentos o realizaciones de ruido blanco empleadas para contaminar la señal de voz (100 segmentos en nuestro experimento).

Puede observarse que aparecen diferentes valles profundos localizados en $c = 50, 100, 150, ...$ muestras cuando el ancho de la ventana $w$ no es muy grande. De esto podemos conjeturar las dos hipótesis siguientes:

- « Se alcanzará menos error cepstral (y por lo tanto de reconocimiento) cuando la ventana tenga su centro o peso máximo sobre el pitch de la señal limpia o sus múltiplos enteros (H1)».

Esto es debido a que en estos puntos, por lo general, la SNR es la máxima debido a que se corresponden con los picos máximos de energía de la autocorrelación de la señal limpia. Es más, estos puntos son los que más información lingüística (de la envolvente espectral) transportan. Efectivamente, al ser la autocorrelación aproximadamente periódica, en los sucesivos múltiplos del pitch encontramos repetidas las mismas correlaciones cortas responsables de la envolvente espectral.

- «En general se alcanzará menos error cepstral (y por lo tanto de reconocimiento) cuando el ancho de la ventana $w$ no sea muy grande aunque tampoco muy pequeño pues llegamos a perder demasiada información de la señal (H2)».

Es decir, debemos de encontrar un compromiso entre darle poco peso a los primeros coeficientes de autocorrelación más contaminados y hacer que esté incluida la máxima información posible de reconocimiento dentro de la ventana. Estas hipótesis han sido extraídas para ruido blanco pero los resultados de reconocimiento obtenidos en las secciones siguientes las validarán para otro tipo de ruidos.

### Analisis espectral de las ventanas

Analicemos ahora qué ocurre en el dominio espectral. La Fig. 6.4 muestra el espectro limpio y el espectro ruidoso promedio de las misma señal anterior para cuatro ventanas diferentes: $DDR_{127,256}$(Standard), $DDR_{135,240}$(Shannon), $DDR_{50,40}$(Thin) y $DDR_{50,250}$(Broad) (esta última es muy parecida a la óptima para Aurora-2 tal y como veremos).

El rango dinámico de una ventana es la distancia en dB entre el lóbulo principal y el secundario y las ventanas que se aplican sobre la autocorrelación deben de tener un rango de unos 80 dB. Observando los espectros limpios, podemos ver que las ventanas centradas sobre el pitch ($DDR_{50,40}$ y $DDR_{50,250}$) tienen un corto rango dinámico elevando así los valles espectrales. Esto, más que ser un problema, llega a ser una ventaja en condiciones

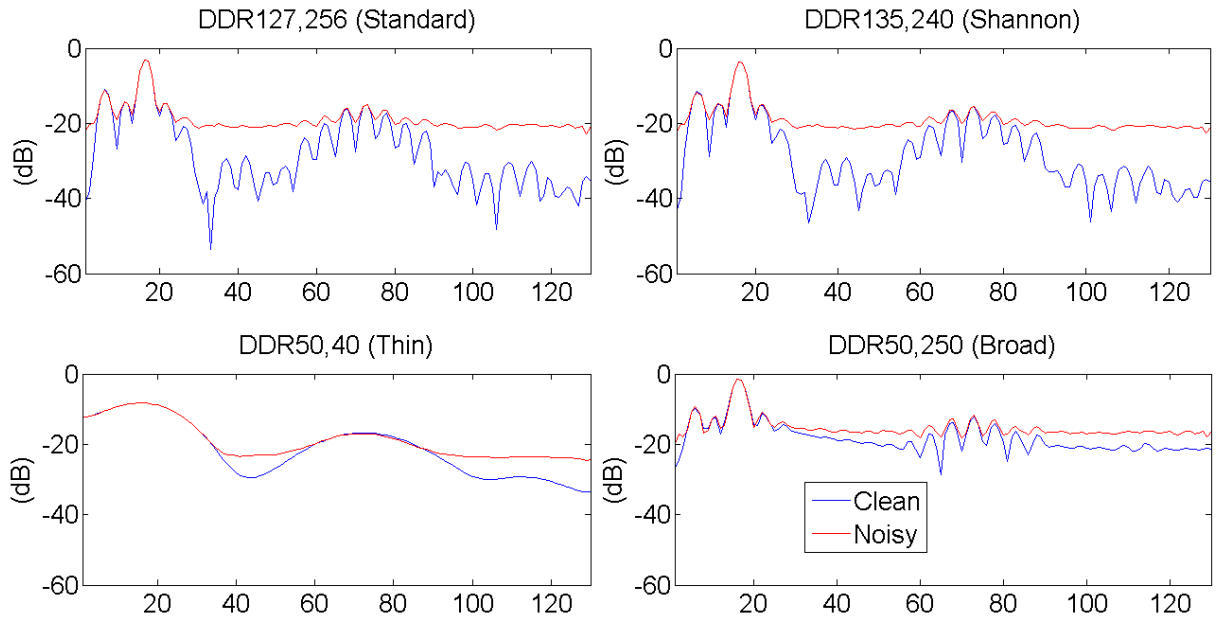



Figura 6.4: Espectro promedio de cuatro ventanas diferentes aplicadas a una vocal con pitch=50 muestras contaminada con ruido blanco. Observar el agotamiento del rango dinámico sobre los espectros limpios de las dos ventanas de abajo, $DDR_{50,40}$ y $DDR_{50,250}$.

de ruido debido a que, como se observa al comparar con los espectros sucios, se disminuye la discrepancia limpio-sucio y por lo tanto la discrepancia entrenamiento-test. Es más, en condiciones limpias podemos conjeturar que tampoco llegará a ser una desventaja debido a que, tal y como mencionamos en la Sec. 2.1, lo importante en el reconocimiento no son tanto los valles (que tienen una alta variabilidad entre locutores) como los formantes, y estos siguen quedando bien caracterizados por las ventanas centradas sobre el pitch como podemos observar en la 6.4. Los resultados en condiciones limpias confirmarán esta hipótesis de que «el corto rango dinámico no tiene grandes efectos negativos sobre el reconocimiento (H3)».

### 6.1.5. Ventanas para segmentos sordos y de silencio

Hasta ahora las ventanas $DDR_{c,w}$ han sido justificadas para los segmentos sonoros. Veamos su justificación sobre los sordos y los silencios teniendo en cuenta que aplicaremos la misma ventana a todo tipo de segmento. Para los segmentos sordos, dar poco peso a los primeros coeficientes de autocorrelación podría suponer una pérdida de información y una reducción del porcentaje de reconocimiento en condiciones limpias. Sin embargo, si el

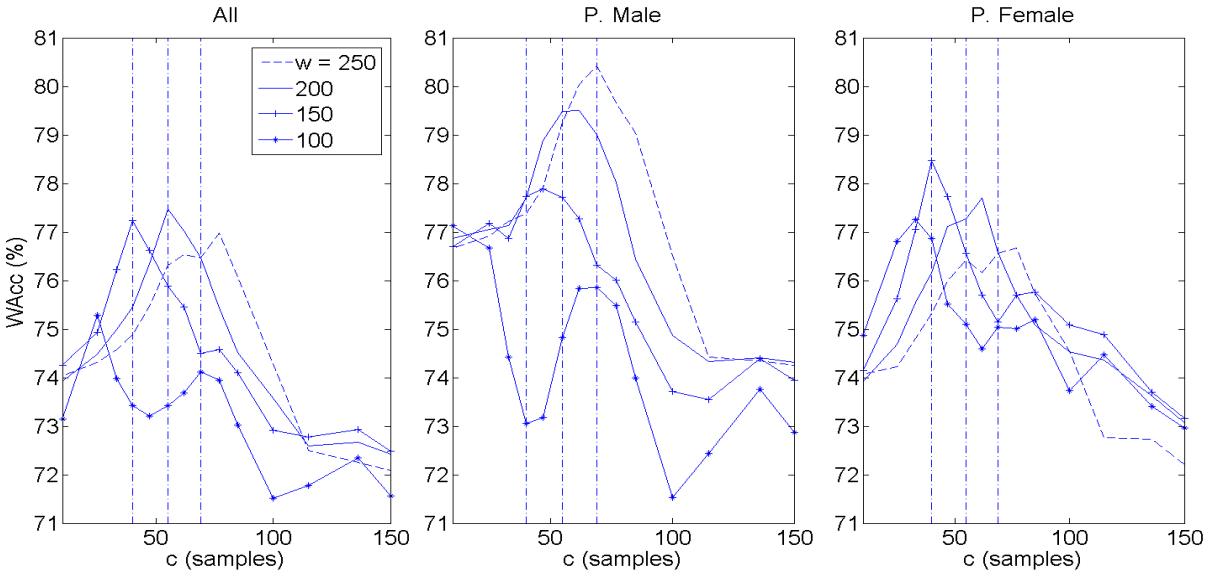



Figura 6.5: WAcc ( %) para toda Aurora-2 (0-20 dB) empleando en entrenamiento y test todas las frases, sólo las que tienen pitch masculino y sólo las que tienen pitch femenino, en función de $c$ (centro) para diversos valores de ancho de ventana $w$ (100, 150, etc.). Las tres lineas verticales se corresponden con el pitch femenino, promedio y masculino (40, 55 y 69 muestras respectivamente).

entrenamiento y el test son hechos con la misma ventana (que es lo que se hará), podemos hacer la hipótesis de que «esta pérdida de información en los sonidos sordos no afectará al porcentaje de reconocimiento (H4)» tal y como mostrarán los resultados experimentales. Para los segmentos de silencio, no hay información que perder, por lo que dar poco peso a los primeros coeficientes será siempre beneficioso tanto en condiciones limpias como, especialmente, en las sucias.

### 6.1.6. Resultados experimentales

Los parámetros de nuestro sistema de reconocimiento de la Fig. 6.1 están descritos en la Sec. A.1 de forma conjunta con otros sistemas para poder hacer una comparación justa entre ellos. Solamente añadir que, para disminuir cualquier tipo de discrepancias, la misma técnica y parámetros que son empleados en el test también son empleados para el entrenamiento.

### Análisis de los resultados

La Fig. 6.5 muestra los resultados de reconocimiento WAcc (Word Accuracy, tasa de Acierto de Palabra en tanto por ciento, Sec. A.3) promediados sobre toda Aurora-2 (Set A, C y B, Sec. A.2) y para las SNRs de 0-20 dB, en función de la ventana $DDR_{c,w}$ empleada y de si se ha empleado en el entrenamiento y test todas las frases, sólo las que tienen pitch tipo masculino ($>$ 55 muestras, Sec. 2.1.2) o sólo las que tienen pitch tipo femenino ($<$ 55 muestras).

Lo interesante de estas figuras es que muestran claramente que los mejores resultados de reconocimiento son obtenidos cuando las ventanas quedan centradas alrededor de los respectivos pitches promedio de los distintos conjuntos entrenamiento-test empleados. Para el conjunto que emplea todas las frases el mejor resultado es 77.47 % con $DDR_{55,200}$ (precisamente su centro está donde está el pitch promedio de la voz, en 55 muestras), para el que emplea solo las masculinas es 80.43 % con $DDR_{69,250}$ (su centro está donde está el pitch promedio de la voz masculina, en 69 muestras) y para el que emplea solo las femeninas es 78.47 % con $DDR_{40,150}$ (su centro está donde está el pitch promedio de la voz femenina, en 40 muestras). Es más, centrándonos en el conjunto que emplea todas las frases, podemos ver que la ventana centrada sobre el pitch promedio ($DDR_{55,200}$) supera notablemente los resultado de HASE ($DDR_{135,240}$) que proporcionan un 72.43 %. Todo ésto viene a fortalecer nuestra hipótesis (H1) de que la mayor robustez en contra del ruido es alcanzada cuando las ventanas $DDR_{c,w}$ están centradas alrededor del valor del pitch debido a que aquí la SNR local es más alta y debido a que estos son los coeficientes que más información lingüística transportan.

Otra cosa interesante que deducimos de esta figura es que el ancho de ventana $w$ debe ser lo suficientemente grande como para cubrir los diferentes valores de pitch y que capture suficiente información lingüística, pero no demasiado porque esto podría sobrepesar los primeros coeficientes de autocorrelación y entonces reducir los resultados de reconocimiento al introducir coeficientes más afectados por el ruido. Esto lo muestran los tres resultados máximos anteriores ($DDR_{40,150}$, $DDR_{55,200}$ y $DDR_{69,250}$) en los que a medida que crece el centro óptimo de la ventana también crece el ancho óptimo, confirmándose así la hipótesis (H2). En este sentido hay que mencionar que nuestra propuesta se ve favorecida por valores de periodo del pitch alto tal y como reflejan los resultados.

La Tab. 6.1 muestra los resultados tomando todas las frases de Aurora-2 en función de la SNR. La fila señalada como *Hamming* es cuando el espectro es obtenido directamente desde la señal (sin pasar por la OSA) enventanando cada segmento con una ventana de

| Ventana | Clean | 20 dB | 15 dB | 10 dB | 5 dB | 0 dB | -5 dB | Media (20-0 dB) |
|---|---|---|---|---|---|---|---|---|
| Hamming (FE) | 99.14 | 97.21 | 92.57 | 76.72 | 44.28 | 22.99 | 13.00 | $66{,}76 \pm 0{,}80$ |
| $DDR_{135,240}$ (HASE) | **99.15** | **97.47** | **94.37** | 84.26 | 58.35 | 27.69 | 14.72 | $72{,}43 \pm 0{,}76$ |
| $DDR_{55,200}$ (Pitch medio) | 98.85 | 96.12 | 93.21 | **85.91** | **70.00** | **42.09** | **18.07** | $\mathbf{77{,}47 \pm 0{,}71}$ |

Tabla 6.1: Resultados de reconocimiento WAcc (Word Accuracy %) de diferentes tipos de ventanas para toda Aurora-2 (Set A, B y C) en función de la SNR. Los intervalos de confianza de las medias han sido obtenidos tal y como se explica en la Sec. A.3.

Hamming. Los resultados que se obtienen son muy similares a los que daría el FE de la ETSI [149] con CMN añadido. Las otras dos filas muestran los resultados de las ventanas $DDR_{135,240}$ (equivalente a HASE) y $DDR_{55,200}$ (centrada alrededor del pitch promedio). Los intervalos de confianza de los resultados promedio (20-0dB) han sido obtenidos tal y como se explica en la Sec. A.3.

Podemos ver cómo los resultados de la ventana propuesta $DDR_{55,200}$ son superiores a los de las ventanas convencionales mejorando en más de 5 puntos los resultados promedio de HASE. Otra cosa interesante que podemos ver es que a pesar del agotamiento del rango dinámico en los sonidos sonoros y de la pérdida de información en los sonidos sordos que produce la ventana centrada alrededor del pitch, los resultados en limpio son casi tan buenos como los que dan las ventanas convencionales que no sufren de alguno de estos defectos. Esto verifica las otras dos hipótesis (H3 y H4) que hemos mencionado anteriormente.

La Tab. 6.2 muestra los resultados para Aurora-3 Spanish (base de datos de ruido real, Sec. A.2) en función de las discrepancias test-entrenamiento. Los intervalos de confianza se han obtenido siguiendo la Sec. A.3 y en las tablas que siguen de la Tesis serán omitidos para evitar sobrecargarlas más. Puede observarse que la ventana centrada en el pitch supera de nuevo los resultados de HASE principalmente para la peor condición (High Mismatch). Para este caso $DDR_{55,200}$ mejora 3.76 puntos.

Teniendo todo esto en cuenta, podemos considerar la ventana $DDR_{55,200}$ como una buena ventana de reconocimiento. Adicionalmente podemos concluir que las ventanas asimétricas centradas en el pitch pueden proporcionar incluso mejores resultados si el sistema discrimina las locuciones por su pitch promedio. En la Sec. 8.3 se tratan los trabajos futuros relacionados con las ventanas asimétricas.

| Ventana | WM | MM | HM | Media |
|---|---|---|---|---|
| Hamming (FE) | 89.08 | 82.15 | 64.51 | $78{,}58 \pm 0{,}64$ |
| $DDR_{135,240}$ (HASE) | 89.76 | **83.16** | 76.39 | $83{,}10 \pm 0{,}58$ |
| $DDR_{55,200}$ (Pitch medio) | **89.85** | 82.87 | **80.15** | $\mathbf{84{,}29 \pm 0{,}57}$ |

Tabla 6.2: Resultados de reconocimiento WAcc ( %) de diferentes ventanas para Aurora-3 Spanish (ruido real) en función del tipo de discrepancia test-entrenamiento: Well, Medium y High Mismatch (WM, MM, y HM).

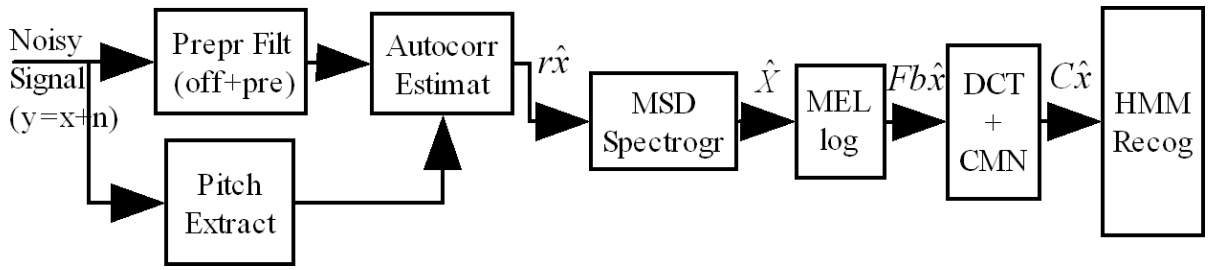



Figura 6.6: Sistema de reconocimiento donde se muestra como son aplicadas las técnicas de estimación de la autocorrelación limpia basadas en el pitch.

## 6.2. Autocorrelación promediada y cribada

### 6.2.1. Introducción

Las técnicas que presentamos a continuación [106] intentan, empleando el pitch de cada segmento y sin hacer estimación del ruido, obtener una estimación de la autocorrelación limpia y por lo tanto del espectro. Pueden ser consideradas como unas técnicas de preprocesamiento (Sec. 5.1.2) porque modifican la señal en un dominio muy cercano al temporal (el de la autocorrelación y sus productos). Dentro de las técnicas que emplean el pitch las consideramos como unas técnicas de estimación de la señal limpia (Sec. 5.2.2).

Presentamos dos técnicas. La primera, estimación mediante promediado o simplemente estimación promediada, se puede interpretar como en un sencillo promediado de la señal ruidosa para incrementar la SNR en los segmentos sonoros. Veremos que esta técnica es un tipo de filtrado peine (o de muestreo de los armónicos del pitch) por lo que puede tratar ruidos armónicos que no estén relacionados con el pitch de la voz.

La segunda es una modificación de la anterior que se inspira nuevamente en la técnica HASE [142] debido a que emplea su idea de que el ruido suele estar contenido en los coeficientes de autocorrelación más bajos. Como se verá más adelante se puede demostrar que esta técnica reúne las ventajas de HASE junto con las de las técnicas basadas en muestreo de los armónicos del pitch (WHNM, técnica de Kuroiwa, filtros peine, etc., Sec. 5.2.2) por lo que puede tratar ruidos que son mezcla de señales poco autocorreladas y señales armónicas no relacionadas con el periodo del pitch.

### 6.2.2. Sistema de reconocimiento

En la Fig. 6.6 podemos observar el sistema de reconocimiento propuesto donde se muestra como son aplicadas las técnicas de estimación de la autocorrelación limpia basadas en

el pitch. Este toma como entrada la señal ruidosa de una frase, la cual es suma de la voz limpia y el ruido ($y = x + n$). El bloque *Pitch extractor* (extractor de pitch) toma esta señal y obtiene el pitch en cada segmento de señal. El resto de los bloques toman la señal sucia pasada a través de un un filtro de preprocesado. El bloque *Autocorrelation Estimator* obtiene una estima de la autocorrelación limpia ($\hat{r}_x$) de cada segmento empleando el pitch. Los tres bloques siguientes se encargan de obtener el cepstrograma (ver Sec. 3.1.4). *MSD Spectrogram, Mel-log y DCT* obtienen el espectrograma de la densidad de la magnitud espectral ($\hat{X}$), la representación en el banco de filtros ($\hat{Fb}_x$) y el cepstrum AMFCC ($\hat{C}_x$) respectivamente a partir de la estima de la autocorrelación limpia multiplicada normalmente por una ventana DDR o de Kaiser (típicas para la autocorrelación). Finalmente, la estima cepstral es pasada al *HMM Recognizer* (reconocedor basado en HMMs).

### 6.2.3. Estimas de la autocorrelación para segmentos sonoros

En esta sección presentamos los dos métodos propuestos de estimación de la autocorrelación limpia, promediado y cribado, para segmentos sonoros (publicados en [106]). En la Sec. 6.2.4 veremos como extender su uso a segmentos sordos y silencios.

#### Notación y consideraciones

Sea $x(n)$ $(n = 0, \ldots, N-1)$ un segmento ruidoso suma de la señal de voz sonora cuasiperiódica (con periodo $T$ dado en número de muestras) y un ruido. Por simplicidad asumiremos que $x(n)$ es la superposición de una señal periódica pura $p(n)$ y una señal de distorsión $d(n)$:

$$x(n) = p(n) + d(n) \quad (n = 0, \ldots, N-1) \tag{6.3}$$

Consideraremos que la señal $d(n)$ recoge todo tipo de distorsiones, entre ellas: las posibles no periodicidades de la señal sonora (debido a la cuasiperiodicidad de la misma) y el ruido aditivo. Por simplicidad también supondremos que el segmento ruidoso posee un número entero de periodos $N_p$ por lo que $N = TN_p$. Esta suposición puede ser eliminada sin problema tal y como se explica en [106], sin más que adaptar los promedios que aparecerán en las fórmulas de las estimas propuestas, al número de muestras disponibles. Siguiendo esta notación y estas consideraciones, el objetivo de las estimaciones que estudiaremos es intentar obtener una estima de la autocorrelación biased (sesgada) de la señal periódica pura ($\hat{r}_p(k)$) (la cual será muy parecida a la autocorrelación de la señal de voz sonora cuasiperiódica limpia).

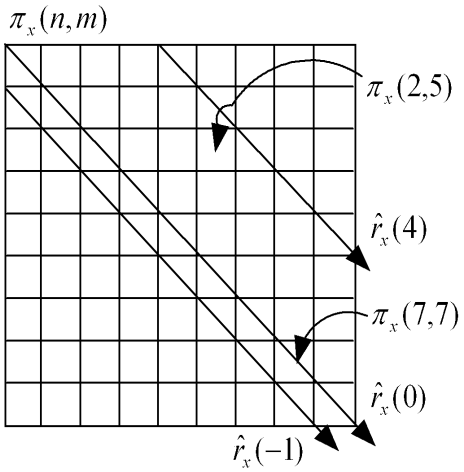



Figura 6.7: Tabla de productos para una señal de nueve elementos. Se ilustran ciertos productos y las flechas diagonales indican los elementos a sumar para obtener los distintos coeficientes de autocorrelación.

### Tabla de productos y autocorrelación biased

Las estimas de la autocorrelación que explicaremos a continuación se pueden formular mediante una tabla que recoge todas las combinaciones de productos entre muestras de señal que aparecen en las autocorrelaciones. La tabla simétrica de productos $\pi_x(n,m)$ de la señal $x$ la definimos como:

$$\pi_x(n,m) = x(n)x(m) \quad (n,m = 0,\ldots,N-1) \tag{6.4}$$

Por lo tanto, el elemento $k-esimo$ de la autocorrelación biased $\hat{r}_x(k)$ (y de forma similar el de la unbiased) puede ser obtenido sin más que sumar los diferentes elementos de la correspondiente diagonal $k-esima$ de la tabla:

$$\hat{r}_x(k) = \frac{1}{N}\sum_{n=k}^{N-1}\pi_x(n,n-k) \quad (k = 0,\ldots,N-1) \tag{6.5}$$

En la Fig. 6.7 podemos ver un ejemplo de estima de la autocorrelación mediante estas tablas para una señal con 9 muestras.

Para ver la precisión con que la autocorrelación biased nos acerca al valor teórico de la autocorrelación de la señal periódica pura $r_p(k)$ desarrollamos su valor esperado. Teniendo en cuenta que la distorsión y la señal periódica no están correlacionadas y teniendo en cuenta la definición de autocorrelación teórica de una señal estacionaria

$r_x(k) = E[x(n)x(n-k)]$, se puede ver fácilmente que el valor esperado de la estima biased es el siguiente:

$$E[\hat{r}_x(k)] = w_B^N(k)\,(r_p(k) + r_d(k)) \tag{6.6}$$

donde $w_B^N(k)$ es una ventana de Barlett de tamaño $N$ y $r_d(k)$ la autocorrelación de la distorsión. Se ve que el valor esperado de esta estima (a parte de sufrir de sesgo debido a la ventana $w_B^N(k)$) no se acerca mucho al valor teórico de la periódica debido a que cada coeficiente de autocorrelación está afectado por un error igual $r_d(k)$. En definitiva podemos decir que esta estima no aporta robustez ninguna. En la Fig. 6.8a podemos ver cuán lejos está la autocorrelación sucia biased de la autocorrelación limpia biased para una señal sonora de voz contaminada con ruido AR. Debajo (Fig. 6.8b) podemos ver que su espectro también dista mucho del espectro limpio.

### Autocorrelación promediada

Siguiendo con la tabla de productos, se puede llegar a ver que en el caso de la señal periódica $p(n)$ cada producto $\pi_p(n,m)$ debe de aparecer repetido $N^2$ veces en la tabla. La siguiente ecuación nos indica, de forma general, los diferentes productos $\pi_p(n,m)$ que son los mismos:

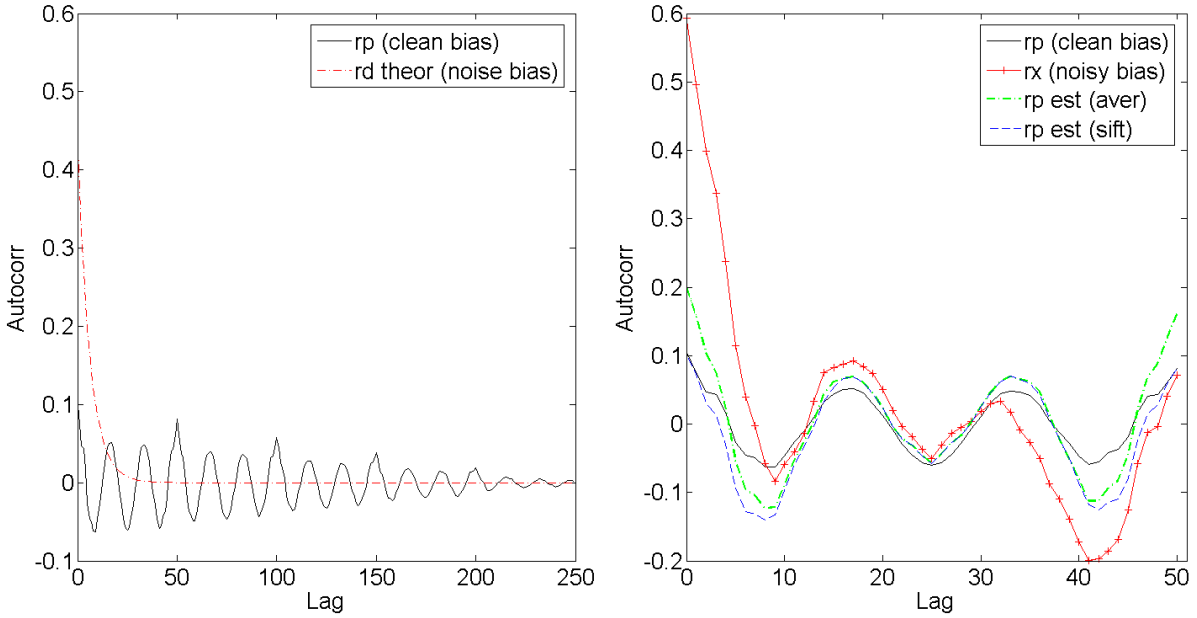
$$\begin{aligned} \pi_p(n,m) = \pi_p(iT + \underline{n}, jT + \underline{m}), \quad \forall \quad (i,j = 0,1,\ldots,N_p - 1) \\ (n,m = 0,1,\ldots,N-1) \end{aligned} \tag{6.7}$$

donde $\underline{n} = 0,\ldots,N-1$ y $\underline{m} = 0,\ldots,N-1$ son los módulos en base $T$ o restos de la división $n/T$. La Fig. 6.9 muestra (señalados con X) los productos que debieran de ser los mismo si $x$ fuera una señal periódica pura de longitud $N = 9$ y periodo $T = 3$ muestras.

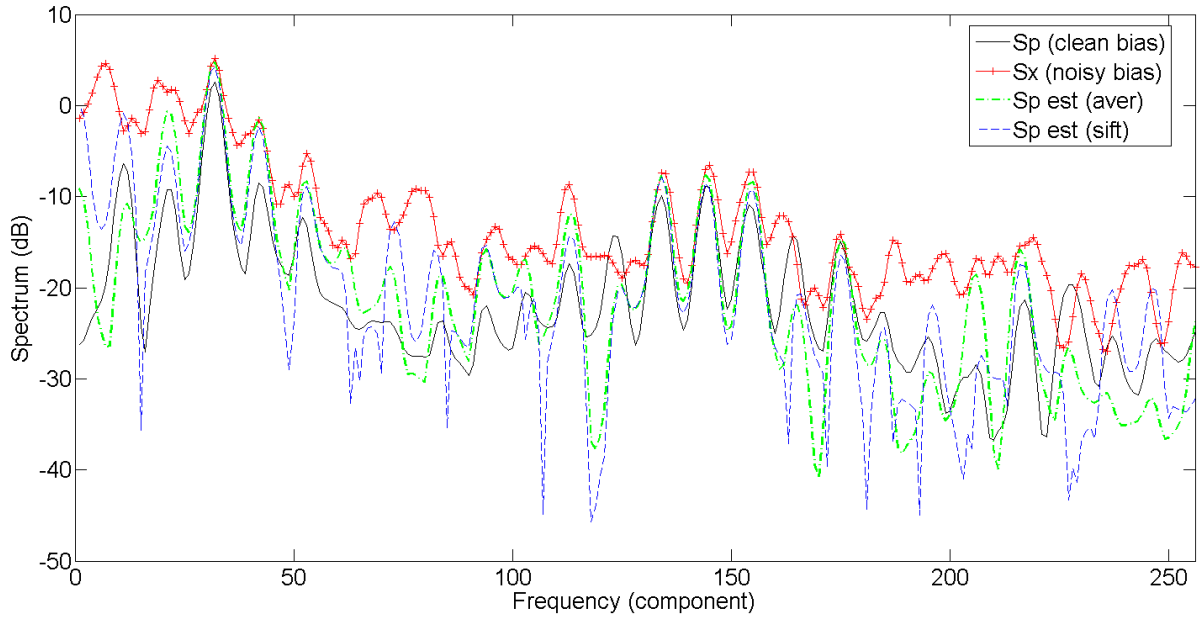
Si la señal periódica ahora es contaminada por la distorsión $d(n)$ la nueva tabla ruidosa ($\pi_x(n,m)$) ya no será periódica en el sentido anterior debido a que cada producto $\pi_p(n,m)$ estará afectado por un error $\epsilon(n,m)$ tal y como muestra la siguiente deducción:

$$\pi_x(n,m) = x(n)x(m) = p(n)p(m) + p(n)d(m) + p(m)d(n) + d(n)d(m) = \pi_p(n,m) + \epsilon(n,m) \tag{6.8}$$

Suponiendo que este error sea de media 0 ($\epsilon(n,m) \to 0$), y teniendo en cuenta las repeticiones antes mencionadas, podemos obtener una buena estima de la tabla de productos de la señal periódica limpia promediando los diferentes productos ruidosos de la




(a) Izquierda, autocorrelación biased de la señal limpia y teórica del ruido AR empleado para contaminarla. Derecha, autoc. limpia biased (clean), sucia biased (noisy), estima promediada (aver) y estima cribada (sift) ($\delta = 16$).




(b) Espectro derivado de la autocorrelación limpia (clean), sucia biased (noisy), estima promediada (aver) y estima cribada (sift).

Figura 6.8: Arriba, comparación de las autoc. propuestas para una vocal con pitch 50 muestras contaminada por ruido AR. Abajo los correspondientes espectros.

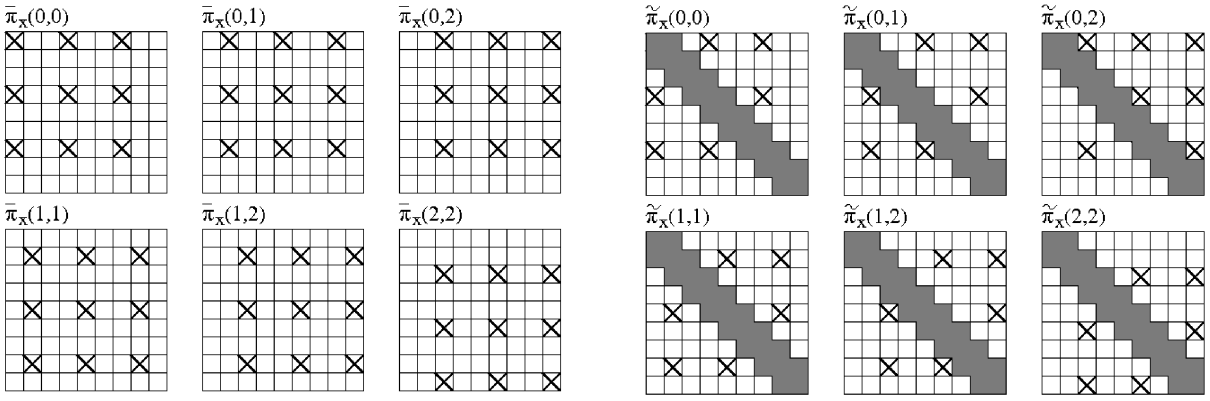



Figura 6.9: Tabla de productos $\pi_x(n,m)$ (repetida 12 veces) para una señal $x$ de longitud $N=9$ y periodo $T=3$ muestras. Izquierda, obtención de los diferentes productos promedio $\bar{\pi}_x(n,m)$ para la autoc. promediada. Derecha, obtención de los diferentes productos cribados $\tilde{\pi}_x(n,m)$ para la autoc. cribada con $\delta=2$.

siguiente manera:

$$\pi_p(n,m) \approx \bar{\pi}_x(n,m) = \frac{1}{N_p^2} \sum_{i=0}^{N_p-1} \sum_{j=0}^{N_p-1} \pi_x(iT+\underline{n}, jT+\underline{m}) \tag{6.9}$$

La Fig. 6.9 muestra la obtención de los diferentes productos promedio $\bar{\pi}_x(n,m)$ a partir de los productos $\pi_x(n,m)$. Se muestra sólo la obtención de los productos base ($\bar{\pi}_x(0,0)$, $\bar{\pi}_x(0,1)$, etc.) debido a que (por las simetrías) el resto de productos $\bar{\pi}_x(n,m)$ son lo mismo que estos.

Teniendo en cuenta todo esto, nuestra «autocorrelación promediada» (estima de la autocorrelación periódica limpia $r_p(k)$) nos queda como:

$$r_p(k) \approx \bar{r}_x(k) = \frac{1}{N} \sum_{n=k}^{N-1} \bar{\pi}_x(n, n-k) \tag{6.10}$$

Se puede demostrar rigurosamente (Sec. 6.2.7) que el valor esperado de esta estima es el siguiente:

$$E[\bar{r}_x(k)] = w_B^N(k) \left( r_p(k) + \frac{N_1(k)\bar{s}_d(\underline{k}) + N_2(k)\bar{s}_d(\underline{k}-T)}{N-k} \right) \tag{6.11}$$

donde vemos que posee un error que depende de la función $\bar{s}_d(\underline{k})$ (Ec. 6.18) que nos indica cuan lejos está dicha estima del valor teórico de la señal periódica $r_p(k)$. Para entender

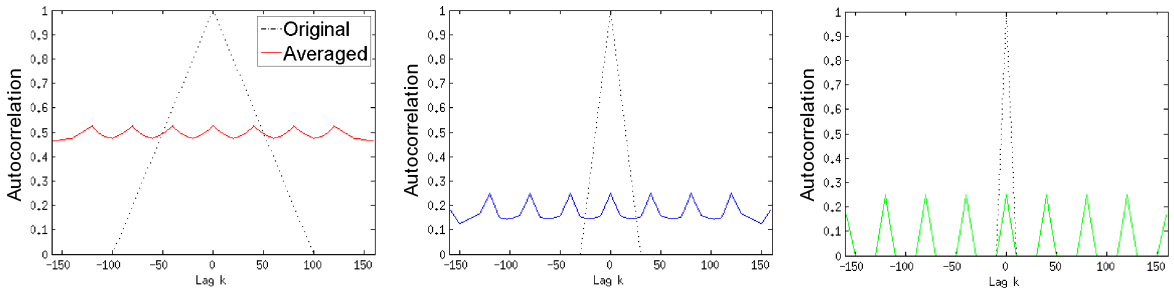



Figura 6.10: Ejemplos de autocorrelaciones promediadas considerando un periodo de $T = 40$ muestras y número de periodos $N_p = 4$) para diferentes tipos de distorsiones coloreadas cuya autocorrelación esta contenida en un intervalo $\delta_d = 100 > T$ (izquierda), $\delta_d = 30 > T/2$ (centro) y $\delta_d = 10 < T/2$ (derecha).

cuanto vale este error y la forma que tiene podemos fijarnos en la Fig. 6.10 la cual muestra cómo actúa la autocorrelación promediada sobre una distorsión. Vemos que lo que hace es convertir su autocorrelación original $r_d(k)$ en periódica con una energía (coeficiente $k = 0$) menor y proporcional al número de periodos $N_p$. De esto se deduce que la estima promedio aumentará la SNR de la estimación un número proporcional a $N_p$. Por otro lado, en la Fig. 6.8a podemos ver como la estima promedio está mucho más cerca de la autocorrelación limpia de lo que lo está la biased contaminada. Debajo podemos ver que con los espectros ocurre lo mismo.

Esta estima de la autocorrelación tiene muchas equivalencias. Por ejemplo, se puede demostrar [106] que es equivalente en promedio, a la correlación cruzada de dos señales permutadas periódico-aleatoriamente, donde cada señal permutada $x_p(n)$ se obtiene intercambiando aleatoriamente las posiciones de las correspondientes muestras periódicas de la siguiente manera:

$$x_p(n) = x(aT + \underline{n}) \tag{6.12}$$

donde $a$ es un número aleatorio entre $[0, N_p - 1]$. Esto se explica en más detalle en la «autocorrelación por entremezclado» del DEA (Diploma de Estudios Avanzados) que precedió a esta Tesis [103].

También se puede demostrar (Sec. 6.2.8) que la autocorrelación promediada es un tipo de filtrado peine (su espectro es equivalente a un muestreo en los armónicos del pitch de la señal contaminada $x(n)$). Esto le da la capacidad, respecto a la biased, de eliminar todo el ruido que hay entre los túneles o armónicos del pitch siendo una estima muy efectiva frente a ruidos armónicos (eso sí, si el ruido posee componentes justamente en los armónicos del pitch esta técnica no consigue combatirlos).

### Autocorrelación cribada

En la tabla de productos $\pi_x(n,m)$ podemos considerar que la distorsión no afecta a todos los elementos por igual, de donde surge la idea de «cribar» dicha tabla, no empleando los productos menos fiables. En particular, podemos mejorar la autocorrelación promediada teniendo en cuenta que muchas veces la distorsión puede considerarse contenida en los primeros coeficientes de la autocorrelación [142] (en un intervalo $\delta$ alrededor de la diagonal 0 de la tabla de productos). La Fig. 6.9 derecha muestra un ejemplo de criba de distorsión contenida en un $\delta = 2$ muestras. Podemos ver que aunque se eliminan productos de la zona de la diagonal, la periodicidad de la tabla aún permite estimar los diferentes productos y por lo tanto todos los coeficientes de autocorrelación de la señal periódica. Para esta mejora basta modificar la Ec. 6.9 de estima promedio de la tabla limpia, no considerando o cribando en el promediado los productos de la diagonal contaminada, tal y como muestra la siguiente ecuación:

$$\pi_p(n,m) \approx \tilde{\pi}_x(n,m) = \frac{1}{N_\delta(\underline{n},\underline{m})} \sum_{(i,j)\in S_\delta(\underline{n},\underline{m})} \pi_x(iT+\underline{n}, jT+\underline{m}) \tag{6.13}$$

donde $\delta$ es el intervalo de criba y $N_\delta(\underline{n},\underline{m})$ es el número de parejas $i,j = 0,\ldots,N_p-1$ que se conservan, contenidas en un conjunto $S_\delta(\underline{n},\underline{m})$ definido como:

$$S_\delta(\underline{n},\underline{m}) = \{(i,j) : |(i-j)T + \underline{n} - \underline{m}| \geq \delta\} \tag{6.14}$$

La Fig. 6.9 muestra un ejemplo de obtención de los productos cribados $\tilde{\pi}_x(n,m)$.

Teniendo en cuenta todo esto, nuestra «autocorrelación cribada» (estima de la autocorrelación periódica limpia $r_p(k)$) nos queda como:

$$r_p(k) \approx \tilde{r}_x(k) = \frac{1}{N}\sum_{n=k}^{N-1} \tilde{\pi}_x(n,n-k) \quad (k=0,\ldots,N-1) \tag{6.15}$$

Se puede demostrar (Sec. 6.2.7) que el valor esperado de esta estima viene dado por la Ec. 6.11 pero sustituyendo los valores de $\bar{s}_d(\underline{k})$ por los de su versión cribada $\tilde{s}_d(\underline{k})$ (Ec. 6.28). Esta autocorrelación tiene las siguientes propiedades interesantes:

1. Se puede demostrar (Sec. 6.2.7) que si la autocorrelación de la distorsión $r_d(k)$ está contenida en el intervalo de criba, esta estima en promedio nos da exactamente el valor teórico de la autocorrelación biased de la señal periódica limpia $r_p(k)$. En la

Fig. 6.8a podemos ver cómo la estima cribada está un poco más cerca (en término medio) de la autocorrelación limpia de lo que lo está la promediada. Debajo podemos ver que con los espectros ocurre lo mismo.

2. Se puede demostrar (Sec. 6.2.7) que la autocorrelación cribada da exactamente los mismos valores que la promediada en el intervalo $\delta \leq k \leq T-\delta$ y que en los intervalos $0 \leq k < \delta$ y $T - \delta \leq k < T$ la cribada tiende a acercarse más a la limpia. De esto se deduce que la cribada es una extensión de la promediada haciendo $\delta = 0$. Los intervalos en los que la cribada se acerca más a la limpia son precisamente los más significativos para el reconocimiento ya que son los que transportan la información relativa a la envolvente espectral. En la Fig. 6.8a podemos observar este efecto para el caso de $\delta = 16$.

En definitiva, podemos hacer la hipótesis de que la autocorrelación cribada dará mejores resultados de reconocimiento que la promediada (H1), debido a que reúne las ventajas de la promediada (como eliminar el ruido entre los armónicos del pitch o ruidos no armónicamente relacionados con el pitch) más las ventajas que ofrece la criba (como eliminar ruidos con autocorrelación contenida dentro del intervalo de criba).

### 6.2.4. Estimas de la autocorrelación para segmentos sordos y de silencio

La autocorrelación biased puede aplicarse a todo tipo de segmentos: sonoros, sordos y silencios. Sin embargo, las autocorrelaciones promediadas y cribadas requieren que el segmento tenga un pitch, por lo que en principio no producirán estimaciones limpias de los segmentos sordos y de silencio. Para evitar tener que emplear un VAD, así como nuevas técnicas de robustecimiento para este tipo de segmentos extenderemos las autocorrelaciones promediadas y cribadas a este tipo de segmentos suponiéndoles un pitch ficticio de 145 Hz (pitch promedio de la voz humana en el que experimentos preliminares han mostrado que el valor de este no afecta en gran medida al resultado final). También en lo que sigue elegiremos el mismo valor de criba $\delta$ para segmentos sonoros, sordos y de silencio.

Esta idea de extender la misma técnica tanto a los sonidos sonoros como al resto por motivos de simplicidad, es común en las técnicas de robustecimiento basadas en el pitch tal y como hemos visto en las ventanas asimétricas, HASE, SWP y WHNM (Sec. 6.1 y 5.2.2). Los motivos que permiten esta extensión en estas técnicas de estimación de la

autocorrelación son los mismos que para las otras técnicas, y se basan en las dos hipótesis siguientes:

1) En los segmentos de silencio siempre es mejor aplicar estas técnicas (promediado y cribado) que no hacer nada (biased) (H2). Por ejemplo, la Fig. 6.10 muestra como la aplicación de la promediada siempre tiende a disminuir la energía del ruido.

2) Si se entrena y testea aplicando siempre la misma técnica de robustecimiento se disminuyen las discrepancias test-entrenamiento, entre ellas las debidas a las perdidas de información de los sonidos sordos al cribar (H3). Esto último se debe de verificar especialmente al comparar los resultados en limpio.

### 6.2.5. Extractor de pitch

Con el fin de emplear el mismo extractor de pitch para todas las técnicas presentadas en esta Tesis, elegiremos en lo que sigue, el extractor espectral de un solo pitch descrito en [106] (ver Sec. 3.4 para entender a qué nos referimos con espectral y de un solo pitch). La elección de este extractor de pitch se justifica porque es el que mejores resultados da en las diferentes técnicas de la Tesis frente a otro tipo de extractores probados tales como el extractor temporal YIN [26] o ciertos extractores espectro-temporales que han sido diseñados por nosotros y que son similares a los que se emplean en [90].

Este extractor toma el pitch proporcionado por el extractor xFE de la ETSI [148] y le aplica un proceso de suavizado típico de los extractores de pitch (Sec. 3.4.3). Este suavizado se basa en aplicar restricciones que consiguen eliminar ciertos fallos tales como saltos de octava y demás errores producidos por el extractor xFE principalmente a bajas SNRs.

### 6.2.6. Resultados experimentales

Los parámetros de nuestro sistema de reconocimiento de la Fig. 6.6 están descritos en la Sec. A.1 de forma conjunta con otros sistemas para poder hacer una comparación justa entre ellos. Solamente añadir que, para disminuir cualquier tipo de discrepancias, la misma técnica y parámetros que son empleados en el test también lo son empleados para el entrenamiento.

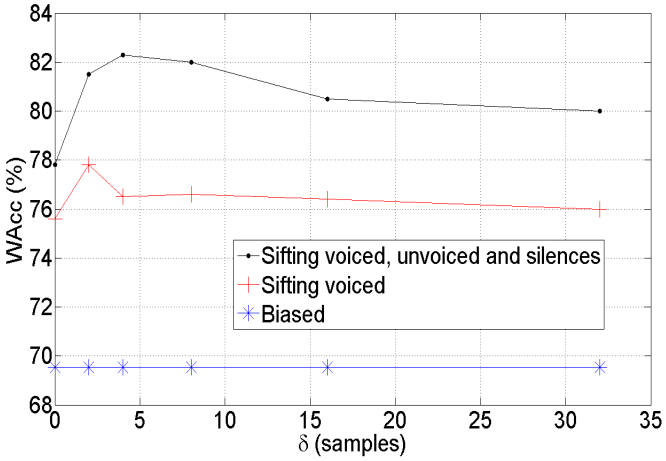



Figura 6.11: Resultados de reconocimiento del Set-A de Aurora-2 en función del intervalo de criba, aplicando siempre autocorrelación biased $*$, aplicando cribada solo a los segmentos sonoros $+$ (resto con biased) y aplicando cribada a todo tipo de segmentos $\bullet$ (sonoros, sordos y de silencio). Para $\delta = 0$ los resultados son los de la autocorrelación promediada.

.

### Valor óptimo de cribado

La Fig. 6.11 muestra los porcentajes de reconocimiento promedios (de 0-20 dB) sobre el conjunto Set-A de Aurora-2 en función de $\delta$ para tres tipos de situaciones: cuando no se aplica técnica de robustecimiento (autocorrelación biased siempre), cuando la autocorrelación cribada es aplicada sobre los segmentos sonoros únicamente (para sordos y silencio se aplica biased) y cuando la cribada es aplicada sobre todos. Podemos sacar las siguientes conclusiones:

1) Cribar siempre es más beneficioso que promediar en término medio, teniendo en cuenta que la autocorrelación promediada es equivalente a la cribada con $\delta = 0$, confirmándose la hipótesis H1 anterior.

2) Se ve que es más beneficioso aplicar la autocorrelación cribada sobre todo tipo de segmentos que sólo sobre los sonoros, confirmándose así las hipótesis H2 y H3 mencionadas anteriormente.

3) El valor de criba óptimo es $\delta = 8$, valor ni muy grande ni muy pequeño. Esto es debido al compromiso que produce el cribado entre eliminación de ruido y pérdida de información de la voz. Este compromiso consiste en que si $\delta$ es muy grande se elimina mucho ruido (aumentando los resultados de reconocimiento) pero a su vez se produce un borrado excesivo de productos y en consecuencia de pérdida de información de la voz (disminuyendo el reconocimiento).

| Técnica | Clean | 20 dB | 15 dB | 10 dB | 5 dB | 0 dB | -5 dB | Media (20-0 dB) |
|---|---|---|---|---|---|---|---|---|
| A. Bias (FE) | 99.06 | 97.65 | 94.74 | 84.06 | 55.30 | 26.53 | 13.63 | 71.65 |
| HASE ($\delta = 15$) | 99.15 | 97.47 | 94.37 | 84.26 | 58.35 | 27.69 | 14.72 | 72.43 |
| A. Aver ($\delta = 0$) | 99.36 | 97.99 | 95.85 | 89.98 | 72.36 | 36.55 | 12.94 | 78.55 |
| A. Sift ($\delta = 8$) | 98.63 | 96.69 | 94.50 | 89.39 | 76.30 | 44.60 | 14.75 | 80.30 |
| A. Sift Ideal ($\delta = 8$) | 98.63 | 97.06 | 95.48 | 91.84 | 82.52 | 61.00 | 29.93 | 85.58 |
| AFE | 99.11 | 97.72 | 96.05 | 91.84 | 82.19 | 59.91 | 28.87 | 85.54 |

Tabla 6.3: Resultados de reconocimiento WAcc ( %) sobre toda Aurora-2 (Set A, B and C), en función de la SNR, obtenidos por diferentes técnicas de robustecimiento.

Teniendo en cuenta estas tres conclusiones, en lo que sigue aplicaremos la técnica de cribado con un $\delta = 8$ sobre todos los segmentos.

### Comparación de técnicas

La Tab. 6.3 muestra los resultados de reconocimiento de diferentes técnicas de robustecimiento sobre toda Aurora-2 en función de la SNR. *A. Bias*, es el resultado obtenido cuando aplicamos las autocorrelación biased que es nuestro resultado base. *HASE* es la técnica de Shannon (Sec. 5.2.2) la cual posee un valor equivalente de criba de $\delta = 15$. *A. Aver* y *A. Sift* son los resultados obtenidos cuando aplicamos las autocorrelaciones promediada (o lo que equivale a la cribada con $\delta = 0$) y cribada (con $\delta = 8$) respectivamente. *A. Sift Ideal* es la autocorrelación cribada pero con pitch ideal (extraído de los correspondientes ficheros limpios que se están evaluando). Por último *AFE* es el extractor de características de la ETSI y que se pone aquí como punto de referencia superior.

De estos resultados podemos sacar las siguientes conclusiones:

1) *A. Sift*, al combatir ruidos armónicos debido al promediado que lleva incorporado, mejora los resultados de HASE que solo lleva criba, pero también los de *A. Aver* que no lleva criba. Esto vuelve a confirmar nuestra hipótesis H1 de que la autocorrelación cribada reúne las ventajas de la promediada más la criba de *HASE*.

2) Los resultados en limpio de *A. Sift* respecto a *A. Bias* o *HASE* son similares mostrando esto que la perdida de información de los sonidos sordos debido al cribado no es un grave problema y por lo tanto verificándose la hipótesis H3 antes mencionada.

3) Los resultados *A. Sift Ideal* muestran las posibilidades de esta técnica si se dispusiese de un extractor de pitch robusto frente a ruido acústico.

| Técnica | WM | MM | HM | Media |
|---|---|---|---|---|
| A. Bias (FE) | 84.03 | 62.15 | 37.85 | 61.34 |
| HASE ($\delta = 15$) | 85.91 | 64.69 | 43.34 | 64.65 |
| A. Sift ($\delta = 8$) | 76.80 | 50.14 | 39.11 | 55.35 |
| A. Sift Ideal ($\delta = 8$) | 84.52 | 71.47 | 61.44 | 72.48 |

Tabla 6.4: Resultados de reconocimiento WAcc ( %) obtenidos por diferentes técnicas para Aurora-3 Danish (ruido real).

4) Los resultados *AFE* son superiores a los de *A. Sift* debido a que este incluye técnicas de estimación de ruido que siempre serán más potentes que las simples suposiciones sobre el ruido que hace *A. Sift* (ruido contenido en el intervalo de criba o ruido armónicamente no relacionado con el pitch), aunque cabe destacar que *A. Sift Ideal* proporciona resultados muy similares a *AFE*.

La Tab. 6.4 muestra los resultados obtenidos sobre la base de datos de ruido real Aurora-3 Danish (Sec. A.2). Podemos ver que *A. Sift* requiere un mejor estimador de pitch para mejorar los resultados de HASE. Esto se ve observando los resultados de *A. Sift Ideal* con pitch ideal en los que se mejora en 18 puntos los resultados de HASE para la peor condición (la de high mismatch).

### Cribado dinámico

La Tab. 6.5 muestra los resultados de Aurora-2 en función del tipo de ruido. Podemos ver como en general *A. Sift* supera a *A. Aver* excepto para ruido tipo *Restaurant* y *Airport*. Las causas de estas deficiencias pueden ser varias (errores en el pitch, un valor de criba no adecuado, etc.). Por ejemplo con pitch ideal en ambas, si se toma un valor de $\delta = 4$ en *Aiport*, la cribada puede superar a la promediada en 0.77 puntos. Si se toma un $\delta = 2$ en *Restaurant*, se puede reducir la distancia en 0.56 puntos. Todo esto sugiere la necesidad de, aparte de mejorar el extractor de pitch, de un $\delta$ dinámico variable en función del ruido. Experimentos oráculos tomando el mejor $\delta$ de reconocimiento para cada frase han mostrado una notable mejora respecto a $\delta$ estático. En la fila denotada como *A. Sift* ($\delta = Ideal$) podemos observar los resultados de esta mejora. En la Sec. 8.3 de trabajos futuros se discute más esta idea del cribado dinámico.

| Técnica | Set A | | | | Set B | | | | Set C | | Media (20-0 dB) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Subw | Babb | Car | Exhi | Rest | Stre | Airp | Trai | Subw MIRS | Stre MIRS | |
| HASE ($\delta = 15$) | 71.02 | 73.22 | 69.67 | 68.11 | 75.67 | 73.34 | 76.38 | 73.79 | 70.74 | 72.31 | 72.43 |
| A. Aver ($\delta = 0$) | 79.19 | 80.14 | 77.36 | 76.54 | 81.03 | 79.08 | 80.73 | 78.73 | 75.63 | 77.01 | 78.55 |
| A. Sift ($\delta = 8$) | 83.62 | 81.96 | 80.56 | 80.80 | 78.45 | 82.15 | 80.16 | 80.63 | 76.16 | 78.47 | 80.30 |
| A. Sift ($\delta = Ideal$) | 89.07 | 87.49 | 86.68 | 86.88 | 85.03 | 88.07 | 85.92 | 86.03 | 85.17 | 85.96 | 86.63 |
| A. Sift Ideal ($\delta = Ideal$) | 93.40 | 92.10 | 91.44 | 90.49 | 91.06 | 92.28 | 91.11 | 92.49 | 91.43 | 91.40 | 91.72 |

Tabla 6.5: Resultados de reconocimiento WAcc ( %) obtenidos por diferentes técnicas para Aurora-2 en función del tipo de ruido.

### 6.2.7. Demostración I: Estadística de las autocorrelaciones

#### Valor esperado de la autocorrelación promediada y cribada

Primero vamos a obtener el valor esperado de la autocorrelación promediada (Ec. 6.10) y después, a partir de este, el de la cribada. La mayoría de los símbolos que aquí se emplean ($T$ periodo, $N_p$ número de periodos, etc.) se encuentran descritos en la sección correspondiente a la Ec. 6.10, los que no se describen a continuación. El valor esperado de la autocorrelación promediada vale lo siguiente,

$$E[\bar{r}_x(k)] = \frac{w_B^N(k)}{N-k} \sum_{n=k}^{N-1} E[\bar{\pi}_x(n, n-k)] \tag{6.16}$$

El valor esperado de la tabla promedio $\bar{\pi}_x(n,m)$ puede ser estimado considerando que $x(n)$ es un proceso aleatorio estacionario tal y como se muestra,

$$\begin{aligned} E[\bar{\pi}_x(n,m)] &= \frac{1}{N_p^2}\sum_{i,j=0}^{N_p-1} E[\pi_x(iT+\underline{n}, jT+\underline{m})] \\ &= \frac{1}{N_p^2}\sum_{i,j=0}^{N_p-1} r_x\left((i-j)T+(\underline{n}-\underline{m})\right) \\ &= \frac{1}{N_p^2}\sum_{l=-(N_p-1)}^{N_p-1} (N_p-|l|) r_x\left(lT+(\underline{n}-\underline{m})\right) \end{aligned} \tag{6.17}$$

Si definimos la siguiente función par,

$$\begin{aligned} \bar{s}_x(j) = \frac{1}{N_p^2}\sum_{l=-(N_p-1)}^{N_p-1} (N_p-|l|) r_x(lT+j) \\ (j=-(T-1),\ldots,T-1) \end{aligned} \tag{6.18}$$

entonces nos queda que,

$$E[\bar{\pi}_x(n,m)] = \bar{s}_x(\underline{n}-\underline{m}) \tag{6.19}$$

por lo que el valor esperado de la Ec. 6.16 se convierte en,

$$E[\bar{r}_x(k)] = \frac{w_B^N(k)}{N-k}\sum_{n=k}^{N-1} \bar{s}_x(\underline{n}-\underline{n}k) \tag{6.20}$$

Podemos considerar dos posibilidades:

1. Caso $\underline{n} \geq \underline{n-k}$. Entonces, $\underline{n-k} = \underline{n}-\underline{k}$ y los elementos de la diagonal $k$-esima de la tabla $E[\bar{\pi}_x(n,m)]$ pueden expresarse como,

   $$E[\bar{\pi}_x(n,n-k)] = \bar{s}_x(\underline{k}) \tag{6.21}$$

   El número de elementos contenidos en esta diagonal es,

   $$N_1(k) = (N_p-\overline{k})(T-\underline{k}) \tag{6.22}$$

   donde $\overline{k}$ es el cociente o valor entero de la división $k/T$.

2. Caso $\underline{n} < \underline{n-k}$. Entonces, $\underline{n-k} = \underline{n} - \underline{k} + T$ y los elementos de la diagonal $k$-esima de la tabla $E[\bar{\pi}_x(n,m)]$ pueden expresarse como,

$$E[\bar{\pi}_x(n, n-k)] = \bar{s}_x(\underline{k} - T) \tag{6.23}$$

El número de elementos contenidos en esta diagonal es,

$$N_2(k) = (N_p - \overline{k} - 1)\underline{k} \tag{6.24}$$

donde puede ser mostrado fácilmente que,

$$N_1(k) + N_2(k) = N - k \tag{6.25}$$

Finalmente, podemos expresar,

$$E[\bar{r}_x(k)] = w_B^N(k) \frac{N_1(k)\bar{s}_x(\underline{k}) + N_2(k)\bar{s}_x(\underline{k} - T)}{N - k} \tag{6.26}$$

Cuando $x(n)$ es una señal periódica limpia de periodo $T$, podemos ver fácilmente que $\bar{s}_x(j) = r_x(j)$ $(j = -(T-1), \ldots, T-1)$ y también que $\bar{s}_x(\underline{k} - T) = r_x(\underline{k})$ dado la periodicidad de $r_x(k)$. Por lo tanto, $E[\bar{r}_x(k)] = w_B^N(k) r_x(k)$. De hecho, no hay aleatoriedad en este caso, así que $\bar{r}_x(k) = w_B^N(k) r_x(k)$.

Cuando $x(n)$ es la suma de una señal periódica $p(n)$ y un proceso estacionario $d(n)$ (no correlado con $p(n)$), entonces el valor esperado de la autocorrelación promediada finalmente vale,

$$E[\bar{r}_x(k)] = w_B^N(k) \left( r_p(k) + \frac{N_1(k)\bar{s}_d(\underline{k}) + N_2(k)\bar{s}_d(\underline{k} - T)}{N - k} \right) \tag{6.27}$$

El valor esperado de la autocorrelación cribada puede ser obtenido de la misma manera. Todas las expresiones anteriores pueden ser igualmente empleadas aunque la función $\bar{s}_x(j)$ debe ser sustituida por su versión cribada la cual es,

$$\tilde{s}_x(j) = \frac{1}{N_\delta(j)} \sum_{l \in L_\delta(j)} (N_p - |l|) r_x(lT + j)$$
$$(j = -(T-1), \ldots, T-1) \tag{6.28}$$

donde,

$$L_\delta(j) = \{l \in [-(N_p - 1), N_p - 1] : |lT + j| \geq \delta\} \tag{6.29}$$

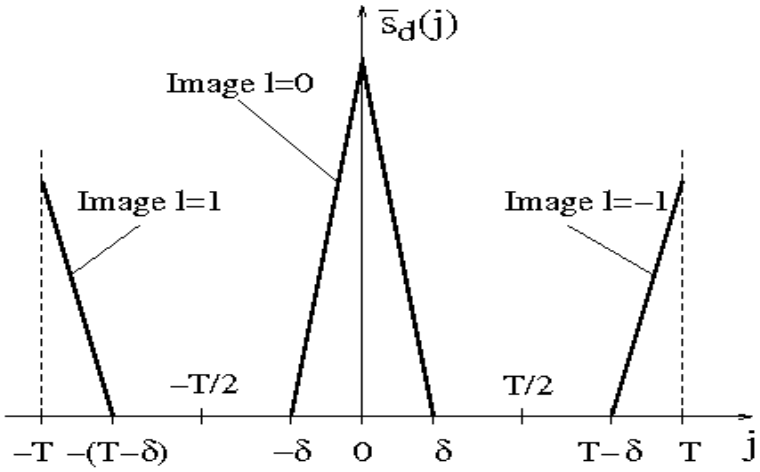



Figura 6.12: Ejemplo de la función $\bar{s}_d(j)$ en el intervalo $[-T, T]$ cuando la distorsión está contenida en el intervalo de criba ($r_d(k) = 0$ si $|k| < \delta$) y el intervalo no es muy grande ($\delta < T/2$)

y $N_\delta(j)$ es el número de elementos en el conjunto $L_\delta(j)$.

### Interpretación estadística

Veamos algunas propiedades interesantes de la autocorrelación cribada partiendo del análisis del error de la promediada (Ec. 6.27).

Aunque la función $\bar{s}_d(j)$ haya sido definida (Ec. 6.18) solamente en el intervalo $[-(T-1), T-1]$, esta es en realidad una serie de $2N_p - 1$ imágenes (separadas un retardo $T$ y escaladas un factor $N_p^2$) de la autocorrelación de la distorsión original $r_d(k)$. Por simplicidad, asumamos que la autocorrelación de la distorsión $r_d(k)$ está contenida en el intervalo de criba $\delta$ y que $\delta < T/2$ (este es el caso de la Fig. 6.10 de la derecha).

Teniendo en cuenta esto, la función $\bar{s}_d(j)$ solo posee las contribuciones de las imágenes $l = -1, 0, +1$ (ya que esta está solamente definida en el intervalo $[-(T-1), T-1]$) y puede ser simplificada como:

$$\bar{s}_d(j) = \frac{N_p - 1}{N_p^2} r_d(j - T) + \frac{1}{N_p} r_d(j) + \frac{N_p - 1}{N_p^2} r_d(j + T) \tag{6.30}$$

La Fig. 6.12 muestra las tres imágenes de $r_d(k)$ (las cuales corresponden a los tres términos de esta ecuación).

Para la estimación cribada, debemos de considerar $\tilde{s}_d(j)$ (Ec. 6.28) en lugar de $\bar{s}_d(j)$. La función $\tilde{s}_d(j)$ solo incluye aquellos términos de 6.30 ($l = -1, 0, +1$) pertenecientes al

conjunto $L_\delta(j)$ (Ec. 6.29). Para computar esta nueva función $\tilde{s}_d(j)$, distinguiremos tres casos diferentes según $j$, teniendo en cuenta la definición de $L_\delta(j)$ y la forma de la función $\bar{s}_d(j)$ original representada en la Fig. 6.12. Consideraremos solamente $0 \leq j < T$, aunque el resultado puede ser directamente extendido a $|j|$ ($j \in [-(T-1), T-1]$) dado que es una función par. Los tres casos son:

1. Caso $0 \leq j < \delta$. La imagen de $r_d(j)$ correspondiente a $l = 0$ no está incluida ($l = 0 \notin L_\delta(j)$) por lo que $\tilde{s}_d(j) = 0$.

2. Caso $\delta \leq j \leq T - \delta$. Las tres imágenes $l = -1, 0, +1$ son empleadas por lo que $\tilde{s}_d(j) = \bar{s}_d(j) = 0$.

3. Caso $T - \delta \leq j < T$. La imagen de $r_d(j)$ correspondiente a $l = -1$ no está incluida ($l = -1 \notin L_\delta(j)$) por lo que $\tilde{s}_d(j) = 0$.

De esto tenemos que $\tilde{s}_d(j) = 0$ para todo $j \in [-(T-1), T-1]$ y considerando la Ec. 6.27 podemos concluir que:

$$E[\tilde{r}_x(k)] = w_B^N(k) r_p(k) \tag{6.31}$$

Esto muestra que la influencia de la distorsión es eliminada completamente en un sentido estadístico. En otras palabras, si obviamos la ventana de Barlett, podemos decir que la autocorrelación cribada es un estimador unbiased de la autocorrelación de la señal periódica limpia $r_p(k)$.

De los tres casos anteriores también se puede deducir que en el intervalo $\delta \leq j \leq T - \delta$ las autocorrelaciones promediadas y cribadas coincidirán debido a que en ese intervalo $\tilde{s}_d(j) = \bar{s}_d(j)$ (independientemente de si la distorsión esta contenida en el intervalo de criba).

### 6.2.8. Demostración II: Filtrado peine mediante autocorrelación promediada

Sea $x(n) = p(n) + d(n)$ un segmento de señal contaminado de tamaño $N$ suma de una señal periódica limpia $p(n)$ de periodo $T$ muestras (frecuencia en radianes $\omega_0 = 2\pi/T$) y una distorsión $d(n)$. Por simplicidad en las demostraciones que haremos supondremos que dentro del segmento hay un número entero de periodos $N_p$, es decir $N = TN_p$. Kuroiwa en

[77] propone emplear la señal promediada periódica $z(n)$ definida de la siguiente forma,

$$z(n) = \frac{1}{N_p} \sum_{i=0}^{N_p-1} x(iT + \underline{n}) \tag{6.32}$$

donde $\underline{n}$ es el resto de dividir $n$ entre $T$, como estima de la señal limpia $p(n)$. Nosotros pretendemos demostrar que este promediado es un tipo de filtrado peine o que es equivalente a un muestreo espectral en los armónicos del periodo de la señal contaminada. Esto se puede expresar mediante la trasformada de Fourier de la señal promediada $Z(\omega_k)$ de la siguiente manera,

$$Z(\omega_k) = \begin{cases} X(\omega_k), & \text{si } \omega_k = m\omega_0 \\ 0, & \text{en otro caso} \end{cases} \tag{6.33}$$

donde $m$ es un entero. Demostrar los casos en que vale 0 es trivial si se tiene en cuenta que tenemos un número entero de periodos y que la transformada de Fourier de una señal periódica pura, como lo es $z(n)$, vale 0 salvo en los armónicos del periodo. La demostración del valor en los armónicos del periodo se reduce a demostrar que estas dos ecuaciones son iguales:

$$X(\omega_0 m) = \sum_{n=0}^{N-1} x(n) e^{-im\omega_0 n} \tag{6.34}$$

$$Z(\omega_0 m) = \sum_{n=0}^{N-1} z(n) e^{-im\omega_0 n} \tag{6.35}$$

Pasemos a desarrollar el espectro $Z(\omega_0 m)$, el cual se puede expresar como:

$$Z(\omega_0 m) = \sum_{n=0}^{N-1} \left( \frac{\sum_{l=0}^{N_p-1} x(lT + \underline{n})}{N_p} \right) e^{-im\omega_0 n} \tag{6.36}$$

Haciendo las siguientes definiciones:

$$e(n) \equiv e^{-im\omega_0 n} \tag{6.37}$$

$$S(\underline{n}) \equiv \sum_{l=0}^{N_p-1} x(lT + \underline{n}) \tag{6.38}$$

y teniendo en cuenta que $e^{-im\omega_0 n} = e^{-im2\pi n/T}$ es una señal de periodo $T$ independiente-

mente del valor $m$ (debido a las propiedades de los números complejos) podemos decir que $e(n) = e(\underline{n})$ y reescribir el espectro de la siguiente manera:

$$Z(\omega_0 m) = \frac{1}{N_p} \sum_{n=0}^{N-1} S(\underline{n}) e(\underline{n}) = \sum_{n=0}^{T-1} S(\underline{n}) e(\underline{n}) \tag{6.39}$$

donde se ha tenido en cuenta que $N = N_p T$ para llegar al último miembro. Sustituyendo por las definiciones e igualdades anteriores podemos hacer el siguiente desarrollo de la ecuación anterior,

$$\sum_{n=0}^{T-1} \left( \sum_{l=0}^{N_p-1} x(lT + \underline{n}) \right) e(\underline{n}) = \sum_{n=0}^{T-1} \sum_{l=0}^{N_p-1} x(lT + \underline{n}) e^{-im\omega_0 n} \tag{6.40}$$

Teniendo en cuenta que $n = lT + \underline{n}$ finalmente tenemos que,

$$Z(\omega_0 m) = \sum_{n=0}^{N-1} x(n) e^{-im\omega_0 n} \tag{6.41}$$

por lo queda demostrado que las Ec. 6.34 y Ec. 6.35 son iguales y por lo tanto que la señal promediada es un tipo de filtrado peine.

Teniendo en cuenta que la autocorrelación de la señal promediada $r_z(k)$ es equivalente a la autocorrelación promediada propuesta $\bar{r}_x(k)$ tal y como muestra el siguiente desarrollo:

$$\begin{aligned} r_z(k) &= \frac{1}{N} \sum_{n=k}^{N-1} z(n) z(n-k) = \frac{1}{N} \sum_{n=k}^{N-1} \frac{1}{N_p^2} \sum_{i=0}^{N_p-1} \sum_{j=0}^{N_p-1} x(iT + \underline{n}) x(jT + \underline{n-k}) \\ &= \frac{1}{N} \sum_{n=k}^{N-1} \bar{\pi}_x(n, n-k) = \bar{r}_x(k) \end{aligned} \tag{6.42}$$

y teniendo en cuenta que la densidad espectral puede ser estimada igualmente desde la señal o desde su autocorrelación, podemos decir que la autocorrelación promediada propuesta también es un tipo de filtrado peine.

## 6.3. Estima del ruido basada en el pitch para reconocimiento con MD

### 6.3.1. Introducción

La técnica que presentamos a continuación (publicada en [105]) es una técnica que, mediante el pitch de cada segmento, intenta estimar el ruido presente. Teniendo en cuenta el efecto de enmascaramiento (Sec. 3.2.1), la única manera de estimar el ruido, sin conocer la señal de voz, es interpolandolo a partir de zonas donde se supone que este es conocido. Las estimas del ruido basadas en un VAD siguen esta idea. Sin embargo, cuando el ruido es poco estacionario este tipo de estimas pueden fallar. Técnicas como **HF** (Harmonic Filtering [129]) o **HT** (Harmonic Tunnelling [38]) estudiadas en la Sec. 5.2.2, mejoran este problema obteniendo más cantidad de muestras del ruido a partir de la separación de los armónicos espectrales del pitch del resto del ruido. La «estima del ruido basada en el pitch» propuesta, primero realiza una estimación VAD del ruido (estando el VAD basado en el pitch) y después mejora esta estima empleando una modificación de la técnica HT basada en filtrado peine del ruido.

Aparte de las modificaciones que se le hacen a la técnica HT para mejorarla, como no incluir como ruido a los sonidos sordos, evitar la sobre-estimación del ruido a altas SNRs y emplear MD (Missing Data) en lugar de SS (Spectral Subtraction), lo interesante de la propuesta es que esta explota de forma óptima la información del pitch para hacer ASR robusto tal y como estudiaremos en el Cap. 7.

Las secciones subsiguientes explicarán esta técnica y la compararán con otras técnicas similares, entre ellas con una estimación del ruido basada solo en VAD y con la técnica de Barker estudiada en la Sec. 5.2.3 que emplea, al igual que esta, MD y el pitch.

### 6.3.2. Sistema de reconocimiento

En la Fig. 6.13 podemos observar el sistema de reconocimiento propuesto para estimar y evaluar la estima del ruido basada en el pitch. Este toma como entrada la señal ruidosa de una cierta locución, la cual es suma de la voz limpia y el ruido ($y = x + n$). El bloque *Pitch extractor* (extractor de pitch) toma esta señal y obtiene el pitch en cada segmento de señal. El resto de los bloques toman la señal sucia pasada a través de un un filtro de preprocesado.

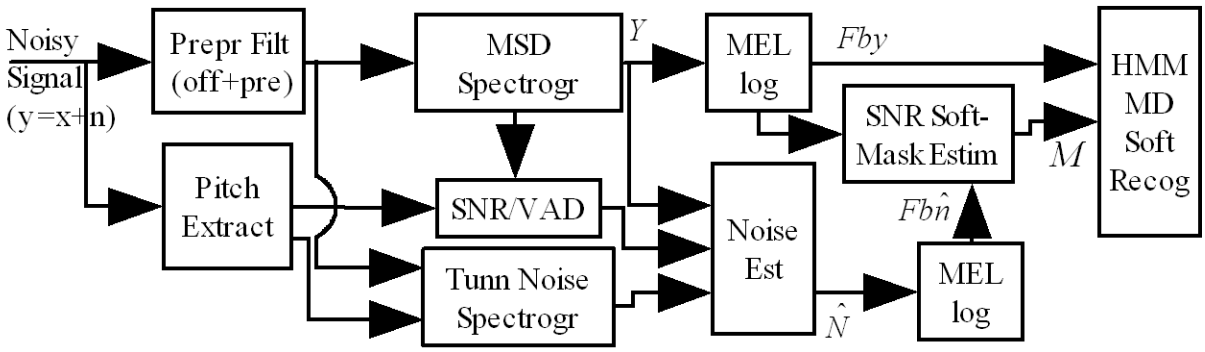



Figura 6.13: Sistema de reconocimiento propuesto para evaluar la estima del ruido basada en el pitch.

.

Los bloques *SNR* (estimador de la SNR de la frase) y VAD (detector de actividad de voz) toman como entrada el espectrograma de la densidad de la magnitud espectral de la señal ruidosa ($Y$ obtenida por el bloque *MSD Spectrogram*) y el pitch. *Tunnel noise spectrogram* (espectrograma túnel del ruido) estima el ruido en los segmentos sonoros usando el pitch y la señal ruidosa. Para ello emplea una variante de la técnica HT. Nuestro bloque central *Noise estimator* (estimador de ruido) toma $Y$, $SNR$, $VAD$ y *Tunnel noise* para dar una estima del espectrograma del ruido ($\hat{N}$). $Y$ y $\hat{N}$ son pasados a través de un banco de filtros Mel (Sec. 3.1.3) y una compresión logarítmica (obteniéndose $Fby$ y $Fb\hat{n}$). Estas dos ultimas salidas son usadas para estimar la SNR de cada píxel espectro-temporal y consecuentemente la correspondiente máscara analógica. Finalmente, la máscara analógica y $Fby$ son pasados al *MD Soft Recognizer* (reconocedor MD analógico) para obtener una transcripción de la frase.

### 6.3.3. Estima del ruido basada en el pitch

A continuación pasamos a describir con más detalle las funciones y bloques del sistema de reconocimiento. Mencionar que los parámetros de los diferentes bloques han sido determinados a través de experimentos preliminares sobre un conjunto de frases de entrenamiento (no de test) de Aurora-2 contaminadas con ruido aditivo. Concrétamente hemos empleado las 50 frases más largas del conjunto de entrenamiento «clean» y se han contaminado a 20 y 0 dB con los ruidos «subway» y «babble».

### Función de estimación temporal del ruido

Una importante función, muy usada por los estimadores de ruido, es la función estimación de ruido basada en las partes conocidas:

$$\hat{N}(\omega_j, t_k) = NEstimaTF((\mathbf{t}_{kn}, \boldsymbol{\omega}_{\boldsymbol{kn}}), Y(\boldsymbol{\omega}_{\boldsymbol{kn}}, \mathbf{t_{kn}}), (\omega_j, t_k)) \tag{6.43}$$

Esta función tiene como entradas las posiciones $(\boldsymbol{\omega}_{\boldsymbol{kn}}, \mathbf{t_{kn}})$ y los valores $Y(\boldsymbol{\omega}_{\boldsymbol{kn}}, \mathbf{t_{kn}})$ espectro-temporales de los píxeles donde el ruido es conocido, además de la posición del píxel donde se desea conocer la estima del ruido $(\omega_j, t_k)$. Como salida nos da el valor de la estima del ruido en este último píxel $(\hat{N}(\omega_j, t_k))$. Mencionar que $kn$ es de known o conocido.

Caben muchas posibilidades para esta función y una de ellas es la que sólo tiene en cuenta píxeles dentro de un mismo canal frecuencial. A esta función, que solo tiene en cuenta píxeles dentro de un mismo canal frecuencial, la llamaremos función de estimación temporal del ruido:

$$\hat{N}_{temp}(\omega_j, t_k) = NEstimaT(\mathbf{t_{kn}}, Y(\omega_j, \mathbf{t_{kn}}), t_k) \tag{6.44}$$

también caben muchas posibilidades para esta función pero, por simplicidad y porque los experimentos preliminares muestran que da buenos resultados de reconocimiento, usaremos la siguiente función temporal: Un píxel de ruido conocido, mantiene el mismo valor de ruido que el original. Un píxel de ruido desconocido es sustituido por el ruido promedio de los 10 píxeles de ruido conocidos más cercanos en tiempo dentro de un mismo canal frecuencial.

### Extractor de pitch

Nuestro extractor de pitch es exactamente el mismo que el que se emplea en la Sec. 6.2.5, por lo que no entraremos en detalles.

### Función de estimación espectral del ruido

Siguiendo la filosofía de la estimación del ruido basada en las partes conocidas del mismo, presentamos otra variante de la Ec. 6.43 que solo tiene en cuenta píxeles dentro de un mismo segmento temporal y que notaremos como:

$$\hat{N}_{freq}(\omega_j, t_k) = NEstimF(\boldsymbol{\omega}_{\boldsymbol{kn}}, Y(\boldsymbol{\omega}_{\boldsymbol{kn}}, t_k), \omega_j) \tag{6.45}$$

donde ahora tenemos una función de estimación frecuencial del ruido. Caben varias posibilidades para esta función. Una de ellas es la propuesta por la técnica HT [38] y que se basa en buscar, sobre el espectro discreto, las componentes espectrales que pertenezcan al ruido a partir de las componentes armónicas del pitch. Aquí proponemos una variante de esta técnica basada en el espectro continuo y que es un tipo de filtrado peine del ruido. Esta propuesta se basa en obtener una estima de la MSD (Magnitude Spectral Density) discreta del ruido interpolando muestras espectrales de la MSD continua de la señal ruidosa. Estas muestras son tomadas en los valles entre los armónicos del pitch (muestras túnel). Veamos como hacerlo. La MSD continua de un segmento ruidoso $y$ con $N$ muestras se obtiene, tal y como se explica en la Sec. 3.3, de la siguiente manera:

$$Y(\omega) = \left| \frac{\sum_{n=0}^{N-1} y(n) win(n) e^{-i\omega n}}{\sqrt{N}} \right| \tag{6.46}$$

donde $\omega$ indica la frecuencia en radianes y $win(n)$ es la ventana usada para la estimación espectral (en nuestro caso será una de Hamming). Las muestras túnel $Y(\omega_l)$ son obtenidas evaluando la Ec. 3.3 en las frecuencias correspondientes a los huecos. La estima frecuencial de la MSD discreta del ruido o estima túnel del ruido, de un segmento $t_k$ con $NFT$ puntos espectrales entre 0 y $2\pi$ es obtenida interpolando entre estas muestras túnel:

$$\begin{aligned} \hat{N}_{tun}(\omega_j, t_k) &= Interp(\boldsymbol{\omega_l}, Y(\boldsymbol{\omega_l}, t), \omega_k) \\ \omega_l = \omega_0(l + \frac{1}{2}),\ l &= \{-1/2, 0, 1, 2, .., ceil(\pi/\omega_0)\} \\ \omega_j = \frac{2\pi j}{NFT},\ j &= \{0, .., NFT/2 - 1\} \end{aligned} \tag{6.47}$$

donde $\omega_0$ es la frecuencia de pitch del correspondiente segmento sonoro e $Interp$ es la función de interpolación para la cual caben muchas posibilidades pero que en nuestro sistema será lineal. En la Sec. 7.2.1 se discuten los efectos de elegir diferentes tipos de interpolaciones.

La Fig. 6.14 muestra un ejemplo de estima túnel. Las muestras túnel son mostradas con cuadrados y el espectro túnel del ruido con lineas entre puntos. Puede observarse que la estima túnel se acerca al ruido real (linea con puntos). Un problema de esta estimación es que cuando la energía del ruido es muy baja comparada con la de la señal de voz,

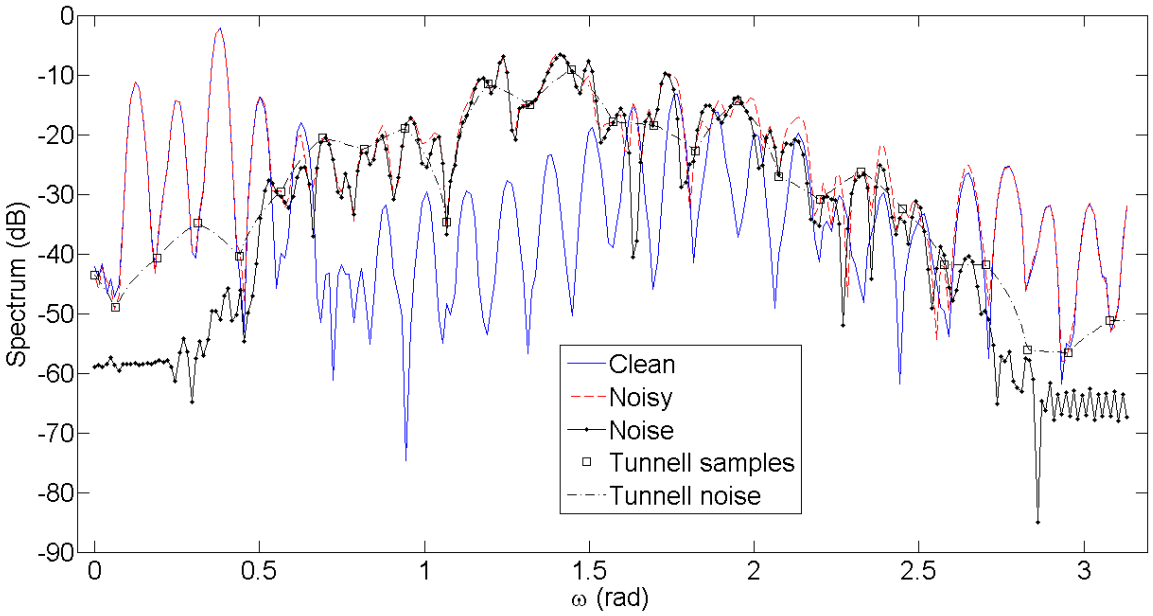



Figura 6.14: Ejemplo de la estima túnel del ruido sobre un segmento de voz sonoro con pitch $\omega = 0{,}126$ rad.

el ruido tiende a ser sobrestimado (p. ej. en los dos extremos de la Fig. 6.14). Esto es debido a que en estas regiones las muestras túnel toman valores que son consecuencia de la ventana usada en la MSD continua más que del propio ruido, y es imposible recuperar completamente el valor real del ruido. Este efecto no es importante a bajas SNRs pero a altas SNRs puede llegar a ser más problemático. Por lo tanto, a altas SNRs, la estima túnel será usada como límite superior del ruido más que como una adecuada estimación del mismo.

### Estimador de la SNR global

Si se tiene una estima espectral del ruido ($\hat{N}(\omega_j, t_k)$) y de la señal limpia ($\hat{X}(\omega_j, t_k)$), empleando el teorema de Parseval, es posible obtener las correspondientes energías ($E_{\hat{N}}(t_k)$ y $E_{\hat{X}}(t_k)$) en cada segmento temporal. La Ec. 6.48 muestra cómo estimar la SNR global

de la locución completa usando estas energías.

$$S\hat{N}R = 10 * log_{10}(\sum_{t_k \in voiced}^{nf} E_{\hat{X}}(t_k) / \sum_{t_k=1}^{nf} E_{\hat{N}}(t_k)) \tag{6.48}$$

$$\text{where } E_S(t_k) = \sum_{j=0}^{NFT/2-1} |S(\omega_j, t_k)|^2 \tag{6.49}$$

donde $nf$ es el número de frames o de segmentos de señal. Solamente los segmentos sonoros (voiced) son empleados para estimar la energía total de la señal limpia ya que los sordos y los silencios prácticamente no contribuyen a la energía total (Sec. 2.1). Esta energía total será similar a la empleada por Aurora-2 para obtener las SNRs de mezcla (Aurora-2 sigue la recomendado de la ITU *P.56* que dice que no hay que considerar las partes de silencio en el computo de la energía total).

Para obtener $\hat{N}(\omega_j, t_k)$ asumiremos que la voz está ausente en los diez primeros y diez últimos segmentos del espectrograma ruidoso ($Y(\omega_j, t_k)$). Estas dos regiones de ruido conocido son pasadas a la función de estimación temporal del ruido (Ec.6.44) para obtener una estima completa del espectrograma del ruido. El espectrograma limpio es estimado mediante una simple SS (Spectral Subtraction): $\hat{X}(\omega_j, t_k) = Y(\omega_j, t_k) - \hat{N}(\omega_j, t_k)$ (0.06 es tomado como valor umbral mínimo).

### Detector de Actividad de Voz

Una característica importante del esquema propuesto es que no tratamos a los sonidos sordos como parte del ruido (cosa que sí hace la técnica HT original [38]) si no que estos son localizados con un VAD para evitar su inclusión. El VAD que proponemos se basa en el modelo de fuente principal de la Sec. 2.1.3, de forma que una vez localizada la fuente principal (en nuestro caso el pitch) es posible localizar el resto de la voz.

Este detecta tres clases diferentes de segmentos: silencio, sordos y sonoros (Sec. 2.1.1). Los segmentos etiquetados como sonoros se corresponden con los segmentos donde el extractor de pitch da un pitch válido (distinto de 0). Para los segmentos sordos asumimos que cumplen las dos propiedades siguientes [134]: Primera, sus energías están localizadas principalmente entre 1800 y 4000 Hz. Segunda, sólo pueden ser localizados antes o después de una secuencia de segmentos sonoros y nunca ocurren aisladamente. Siguiendo la primera propiedad, y de forma similar a como se hace en la detección de «zonas de comienzo/final común» [155], podemos estimar una SNR instantánea de las altas frecuencias HF (High

Frequency) como:

$$\hat{SNR}^{HF}(t_k) = 10 * log_{10}(E^{HF}_{\hat{X}}(t_k)/E^{HF}_{\hat{N}}(t_k)) \tag{6.50}$$

donde el espectrograma limpio $\hat{X}$ y el del ruido $\hat{N}$ son estimados de la misma manera que en el apartado anterior, por medio de una simple sustracción espectral. Las energías de los segmentos son estimadas empleando la Ec. 6.49 pero en lugar de sumar sobre todo el rango de frecuencias, se emplean solamente las frecuencias entre 1800 y 4000 Hz. Teniendo en cuenta la segunda propiedad antes mencionada y esta medida instantánea de la SNR, consideraremos que los segmentos con $\hat{SNR}^{HF}(t_k) > 3dB$ y que ocurren hasta 20 segmentos antes o después de una secuencia sonora, son sordos. Experimentos preliminares han mostrado también que a bajas SNRs, esta estimación de los sordos toma muchos segmentos de ruido como sordos. Por lo tanto, cuando $\hat{SNR} < 10dB$ será asumido que las señales sordas están demasiado mezcladas con el ruido y no se llevará a cabo detección de los segmentos sordos. Finalmente, los segmentos de silencio son aquellos que no han sido clasificados ni como sordos ni como sonoros.

### Estimador de ruido

Nuestra estimación del ruido es llevada a cabo en dos etapas.

En la primera, es supuesto que en las regiones de silencio (detectadas con nuestro VAD) el espectrograma ruidoso ($Y(\omega_j, t_k)$) está dominado por el ruido, de forma que estas regiones de ruido conocidas son pasadas a la función de estimación temporal del ruido (Ec. 6.44) para obtener una primera estima del ruido denominada «ruido VAD».

En la segunda etapa, los correspondientes segmentos sonoros de la primera estimación son revisados usando el «ruido túnel» con el objetivo de mejorar esta estima y por lo tanto los resultados de reconocimiento (esta es la hipótesis H1). Tal y como mencionamos en la Sec. 6.3.3, el ruido túnel proporciona una buena estimación del ruido cuando la SNR es baja pero a altas SNRs, es mejor usar el ruido túnel como un límite superior del ruido real. Siguiendo esta idea, cuando $\hat{SNR} < 10dB$ los segmentos sonoros de la primera estima del ruido son reemplazados por el ruido túnel, en caso contrario, el ruido túnel es usado como límite superior para estos segmentos. Esto podría suponer un mal seguimiento del ruido a altas SNRs si este fuera poco estacionario. Sin embargo, tal y como se explica en [89] al analizar la base de datos de ruido real CHiME, esto no suele ocurrir en situaciones reales debido a que a más SNR, el ruido tiende a ser más estacionario.

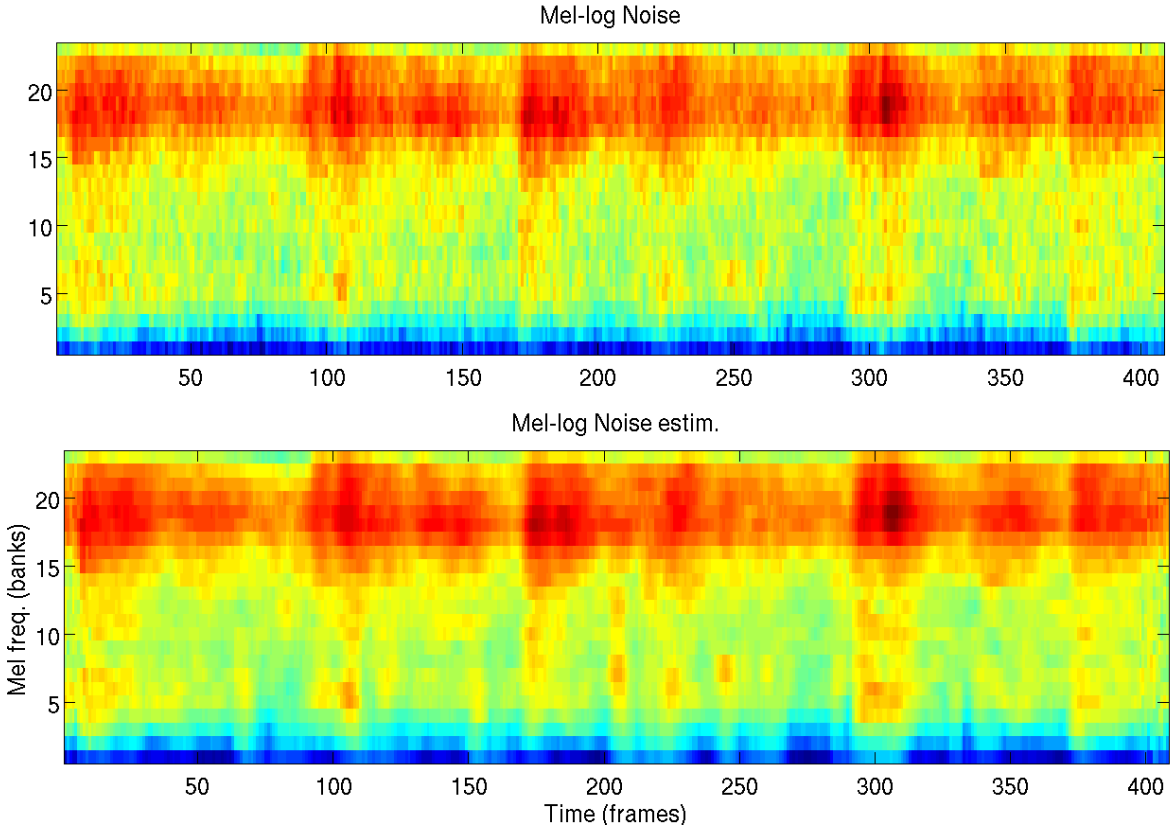



Figura 6.15: Abajo, estima del ruido basada en el pitch. Arriba, el ruido que se intenta estimar que es el de la frase 4460806 de Aurora-2 con ruido subway a 0dB.

Finalmente, el ruido revisado en la segunda etapa es pasado a través de un filtro temporal de media de tamaño 5 segmentos para suavizar posibles errores y el producto final es nuestra estima propuesta del espectrograma del ruido. La Fig. 6.15 muestra un ejemplo comparativo de esta estima (una vez pasada por el banco de filtros Mel y comprimida mediante la función logaritmo neperiano).

### Estimador de máscara analógica basado en la SNR

Una vez estimado el ruido, este puede ser empleado en diferentes tipos de técnicas de robustecimiento (p. ej. SS que es lo que hace HT [38]), sin embargo lo emplearemos para estimar máscaras de MD porque suponemos que MD da mejores resultados de reconocimiento que SS (hipótesis H2). Veamos cómo estimar la máscara analógica. Mediante la estima de la SNR local de cada píxel podemos obtener la máscara analógica tal y como se explica en la Sec. 3.2.2. La SNR de cada píxel Mel-Log ruidoso ($Fby(ch_j, t_k)$) la

obtenemos como:

$$S\hat{N}R(ch_j, t_k) = 20 * log_{10}(e^{Fb\hat{x}(ch_j,t_k)}/e^{Fb\hat{n}(ch_j,t_k)}) \tag{6.51}$$

donde $Fb\hat{n}(ch_j, t_k)$ es la estima Mel-log del ruido en el dominio del banco de filtros (canal Mel $ch_j$ y segmento $t_k$ ) y donde el espectrograma limpio $Fb\hat{x}(ch_j, t_k)$ es estimado mediante una sustracción espectral simple después de deshacer la comprensión logarítmica: $e^{Fb\hat{x}(ch_j,t_k)} = e^{Fby(ch_j,t_k)} - e^{Fb\hat{n}(ch_j,t_k)}$ (donde 0.06 es tomado como valor de suelo mínimo). La máscara analógica es generada comprimiendo $S\hat{N}R(ch_j, t_k)$ entre [0,1] con una función sigmoide (Sec. 3.2.2). Los valores de umbral y pendiente de esta función son $\beta = -3$ (i.e. SNR -3 dB) y $\alpha = 0{,}2$, respectivamente, y han sido determinados empíricamente sobre el conjunto de entrenamiento mencionado al comienzo de esta sección.

### 6.3.4. Resultados experimentales

Los parámetros de nuestro sistema de reconocimiento de la Fig. 6.13 no se explican porque están descritos en la Sec. A.1 de forma conjunta con otros sistemas para poder hacer una comparación justa entre ellos.

#### Resultados con Aurora-2

La Tab. 6.6 muestra las tasas de reconocimiento (WAcc) de diferentes técnicas de robustecimiento para Aurora-2 en función de la SNR. Los cuatro primeros sistemas, etiquetados con *Ceps*, emplean como entrada al reconocedor una estimación del cepstrograma limpio de la voz y todos aplican CMN. Los cuatro últimos, etiquetados con $MD$, emplean un reconocedor de MD. *FE* corresponde a reconocer diréctamente con los MFCCs derivados los vectores espectrales Mel-Log ruidosos. Es nuestro resultado base y es muy similar al que daría el FE [149] de la ETSI con CMN. *AFE* es el extractor de la ETSI [147].

*N. VAD+Tun, SS* corresponde a la estima propuesta del ruido (basada en ruido VAD más ruido túnel) es usada en un sistema con sustracción espectral para estimar la señal limpia. La Fig. 5.2 y la Sec. 5.1.4 explican este sistema. La SS empleada tiene los dos siguientes parámetros: Factor de atenuación $A = 10$ dB, y suavizado mediante filtros de mediana temporales de tamaño 9 segmentos para suavizar la estima $SNR$ y el filtro $H_{ss}$. Esto último reduce el ruido musical. *A. Sift* es la autocorrelación cribada [106] de la Sec. 6.2 (con $\delta = 8$ y mismo extractor de pitch que el resto de técnicas) y es presentada aquí como una técnica que emplea el pitch para hacer reconocimiento robusto. *N. VAD*

| Sistema | Clean | 20 dB | 15 dB | 10 dB | 5 dB | 0 dB | -5 dB | Media (20-0 dB) |
|---|---|---|---|---|---|---|---|---|
| FE (Ceps) | 99.14 | 97.21 | 92.57 | 76.72 | 44.28 | 22.99 | 13.00 | 66.76 |
| N. VAD+Tun, SS (Ceps) | 99.36 | 96.66 | 92.09 | 81.84 | 64.09 | 37.06 | 9.72 | 74.35 |
| A. Sift (Ceps) | 98.63 | 96.69 | 94.50 | 89.39 | 76.30 | 44.60 | 14.75 | 80.30 |
| AFE (Ceps) | 99.11 | 97.72 | 96.05 | 91.84 | 82.19 | 59.91 | 28.87 | 85.54 |
| N. VAD+Harm (MD, Cocl) | 98.67 | 96.18 | 92.67 | 84.17 | 74.21 | 50.41 | 17.65 | 79.53 |
| N. VAD (MD) | 98.76 | 96.19 | 93.38 | 88.42 | 77.92 | 49.52 | 15.56 | 81.09 |
| N. VAD+Tun (MD) | 98.78 | 95.79 | 92.04 | 86.66 | 78.03 | 54.43 | 18.40 | 81.39 |
| N. VAD+Tun Ideal (MD) | 98.78 | 95.97 | 92.81 | 88.57 | 84.24 | 74.43 | 55.83 | 87.21 |

Tabla 6.6: Resultados de reconocimiento WAcc ( %) obtenidos por diferentes técnicas para toda Aurora-2 (Set A, B and C) en función de la SNR.

*+ Harm* es un sistema MD basado en la técnica de Barker [6] que emplea el cocleograma (no el espectrograma) como representación acústica y que se basa en estimar dos tipos de máscaras: máscara-ruido $M_n$ y máscara-armónica $M_h$. La Fig. 5.5 y la Sec. 5.2.3 explican este sistema. Para hacer una comparación justa con nuestro sistema, este sistema emplea el mismo VAD y la misma Ec. 6.44 que nuestro sistema para estimar su ruido $\hat{N}_{gam}$ pero adaptado al cocleograma. Los parámetros de este sistema son los siguientes: Umbrales y pendientes de las sigmoides -6 dB y 0.8 para $M_n$, y 0.8 y 200 para $M_h$. Los parámetros del cocleograma se describen en la Sec. A.1 y son tales que hacen que este sea muy similar al espectrograma. Finalmente, *N. VAD* es la estima propuesta del ruido VAD (sin el añadido del ruido túnel) sobre el sistema de MD explicado anteriormente (Fig. 6.13).*N. VAD + Tun* es la estima propuesta del ruido completa y *N. VAD+Tun Ideal* es la estima completa cuando se emplea pitch ideal (pitch obtenido del fichero limpio que se esté testeando).

De esta tabla podemos extraer las siguientes conclusiones:

1) En condiciones limpias las técnicas basadas en el cepstrum obtienen resultados ligeramente mejores que las basadas en MD. Esto es debido a que el cepstrograma es una representación más robusta que el espectrograma (ver Sec. 3.1.5).

2) En general, las técnicas basadas en MD obtienen mejores resultados que las basadas en SS verificándose la hipótesis H2. Esto se ve comparando los resultados de *N. VAD+Tun, SS* con los de *N. VAD+Tun* en las que empleando el mismo ruido, MD obtiene mejores resultados que SS. La SS empleada aquí es demasiado simple ya que para su buen funcionamiento requiere que el ruido esté por lo general más bajo que la voz. Empleando una *SS* más compleja como la de [10] se podría disminuir esta diferencia, pero por lo general

MD siempre tendrá la ventaja de no tener que conocer el ruido con exactitud, bastando con saber simplemente si domina la voz o el ruido (Sec. 5.1.6).

3) Comparando *N. VAD+Tun* con *A. Sift* y *N. VAD+Harm Cocl* vemos que la estima propuesta del ruido hace un mejor aprovechamiento del pitch que estas dos. Sin embargo, no podemos concluir esto definitivamente, ya que varias causas pueden estar influyendo en estas diferencias, entre otras, el que las dos técnicas sean más sensibles a los errores de pitch y el que sus parámetros no se hayan tuneado perfectamente. Esto nos lleva a la pregunta de qué técnica es la que mejor aprovecha la información del pitch para combatir el ruido. La respuesta a esta pregunta la damos en el Cap. 7.

4) Si *N. VAD* es comparado con *N. VAD+Tun* deducimos que la adición del ruido túnel al ruido VAD supone un beneficio (principalmente a bajas SNRs) aunque pequeño. Esto confirma nuestra hipótesis H1 aunque no fuertemente debido principalmente a que los ruidos de Aurora-2 (como es bien conocido [62]) son en general bastante estacionarios. Sin embargo, en ruidos más esporádicos esta adición podría potencialmente dar mayores beneficios. Esta diferencia se hace mucho más patente cuando se emplea pitch ideal indicándonos que otro culpable de esta modesta mejora es la mala estimación del pitch. Todo esto se comprueba en la Tab. 6.7 al comparar los resultados con pitch ideal de *N. VAD Ideal* y *N. VAD+Tun Ideal* prestando atención a los ruidos *bable* (ruido menos estacionario) y *car* (ruido más estacionario). Se observa que la mejora de añadir ruido túnel en *bable* es de 14 puntos mientras que en *car* empeora 2 puntos.

5) Otra cosa interesante que muestra la Tab. 6.7 es que la autocorrelación cribada (*A. Sift*) al ser una técnica cepstral lleva incorporado CMN y obtiene mejores resultados que las técnicas de MD para los ruidos convolutivos del conjunto Set-C. Esto muestra la debilidad que tienen las técnicas de MD frente a ruidos tipo convolutivos. Está debilidad está en fase de investigación [115].

6) Comparando *N. VAD+Tun* con *N. VAD+Tun Ideal* vemos que un mejor extractor de pitch mejoraríamos enormemente los resultados (más de 20 puntos a 0 dB), superándose los de *AFE* (que de todas es la técnica más potente sin emplear información oráculo).

Las pruebas con Aurora-3 no se muestran debido a los dos motivos siguientes: El primero es que el extractor de pitch empleado no es lo suficientemente robusto como para hacer frente a esta base de datos tal y como vimos en la Sec. 6.2.6 al probar la autocorrelación cribada sobre Aurora-3. El segundo es que esta base de datos está pensada para entrenar los modelos con frases contaminadas y esto no va con la filosofía de MD que necesita entrenar en limpio. Como trabajo futuro podríamos mejorar el extractor de

| Técnica | Media (20-0 dB) [0 dB] | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Set A** | | | | **Set B** | | | | **Set C** | | Media |
| | Subw | Babb | Car | Exhi | Rest | Stre | Airp | Trai | Subw MIRS | Stre MIRS | |
| A. Sift (Ceps) ($\delta = 8$) | 84 [53] | 82 [48] | 81 [40] | 81 [45] | 78 [46] | 82 [48] | 80 [48] | 81 [43] | 76 [33] | 78 [40] | 80 [45] |
| N. VAD (MD) | 82 [53] | 83 [52] | 83 [47] | 84 [58] | 82 [51] | 83 [55] | 84 [56] | 81 [47] | 74 [36] | 76 [40] | 81 [50] |
| N. VAD+Tun (MD) | 85 [64] | 83 [58] | 83 [52] | 84 [63] | 80 [53] | 83 [59] | 82 [56] | 80 [49] | 76 [44] | 77 [46] | 81 [54] |
| N. VAD Ideal (MD) | 85 [66] | 85 [63] | 90 [76] | 87 [71] | 86 [66] | 87 [70] | 89 [73] | 88 [72] | 81 [62] | 83 [65] | 86 [68] |
| N. VAD+Tun Ideal (MD) | 89 [80] | 88 [77] | 88 [74] | 87 [75] | 89 [76] | 88 [75] | 89 [77] | 87 [74] | 83 [69] | 83 [67] | 87 [74] |

Tabla 6.7: Resultados de reconocimiento WAcc ( %) (20-0 dB) obtenidos por diferentes técnicas para Aurora-2 en función del tipo de ruido. El resultado a 0 dB se muestra entre corchetes.

pitch e intentar hacer una adaptación de Aurora-3 para conseguir unos modelos limpios. En la Sec. 8.3 se detalla todo esto.

# Capítulo 7

# Equivalencias y Límites de las Técnicas Basadas en el Pitch

## 7.1. Mecanismos básicos y equivalencias

### 7.1.1. Mecanismos básicos sonoros

#### Equivalencias entre técnicas

En la Sec. 5.2 y Cap. 6 nos hemos dedicado a estudiar y proponer diferentes técnicas de robustecimiento basadas en el pitch. Ahora vamos a intentar compararlas de forma justa atendiendo a algunas equivalencias encontradas. En principio podemos suponer que estas técnicas son diferente si atendemos a los detalles de implementación (extractor de pitch resultante y empleado, mecanismo de actuación sobre los segmentos sonoros, sordos y de silencio, forma de reutilizar los productos generados por los diferentes módulos, etc.). Sin embargo, olvidándonos de estos detalles y atendiendo solamente a cómo actúan las técnicas sobre los segmentos sonoros podemos decir que muchas de las técnicas son equivalentes y que estas obedecen a uno de los cuatro mecanismos básicos que explicamos a continuación.

#### Mecanismos básicos de los segmentos sonoros

Consideramos que los mecanismos básicos para robustecer un segmento sonoro basados en el pitch son:

1) **Aprovechamiento de la estructura armónica**: estos mecanismos no requieren de un extractor de pitch, tal y como se explica en la Sec. 5.2.1, sino de los efectos que

este produce sobre la señal. Podemos destacar HASE y las ventanas asimétricas [107] (mecanismos de realce espectral que emplean el cepstrograma), la técnica HF [129] (que estima el ruido y el cual puede ser empleado en SS para reconocer con el cepstrograma o en MD para reconocer con el espectrograma) y otras técnicas relacionadas tales como SWP [92], etc..

2) **Estima peine de la señal limpia**: basado en aplicar algún tipo de filtrado peine o algoritmo relacionado (bien sea en el dominio temporal, de la autocorrelación o del espectro) al segmento sonoro contaminado por ruido, de forma que el espectro resultante tienda a disminuir la energía de las componentes espectrales entre los armónicos del pitch (que son de ruido) y deje intactas las componentes del pitch (que son de voz más ruido). El espectro resultante es una estima espectral limpia que puede ser empleada para reconocer mediante su cepstrograma. Técnicas que emplean este mecanismo son WHNM [138], sus técnicas relacionadas (PHCC [52], etc.), y la Autocorrelación Promediada (y Cribada) propuesta [106] tal y como se demuestra en la Sec. 6.2.8.

3) **Estima túnel del ruido**: mecanismo opuesto al anterior y basado en aplicar algún tipo de filtrado peine o algoritmo relacionado (bien sea en el dominio temporal, de la autocorrelación o del espectro) al segmento sonoro contaminado por ruido de forma que el espectro resultante tienda a disminuir la energía de las componentes armónicas del pitch y deje por igual las muestras túnel (las componentes espectrales entre los armónicos del pitch). Estas muestras túnel son empleadas para estimar el ruido total (ruido túnel) mediante alguna interpolación o ajuste a un modelo de ruido. El ruido túnel puede ser empleado para hacer SS (sustracción espectral o similares) y reconocer con el cepstrograma, o para estimar máscaras y reconocer mediante MD. Técnicas que emplean esto son HT [38], sus técnicas relacionadas (FPM-NE [19], etc.) y el Ruido Basado en el Pitch [105].

4) **Estima de máscaras mediante armonicidad**: basado en estimar la armonicidad de cada píxel frecuencio-temporal a través del correlograma tal y como se explica en la Sec. 3.3. Esta armonicidad es empleada para estimar una máscara discreta o analógica (ver técnica de Barker, Sec. 5.2.3) y reconocer con MD. Técnicas que emplean esto son casi todas las basadas en el cocleograma tales como la técnica de Barker [6], la de Brown [18] y la de Ma [90].

Mencionar que la técnica FP-MSE [19] (y similares) no ha sido clasificada debido a que emplea información previa sobre el ruido y limita su aplicabilidad a cualquier tipo de ruido (cosa en la que no estamos interesados). A pesar de esto, podríamos incluir esta técnica en los mecanismos 2 o 3 debido a que limpia y estima el ruido al mismo tiempo.

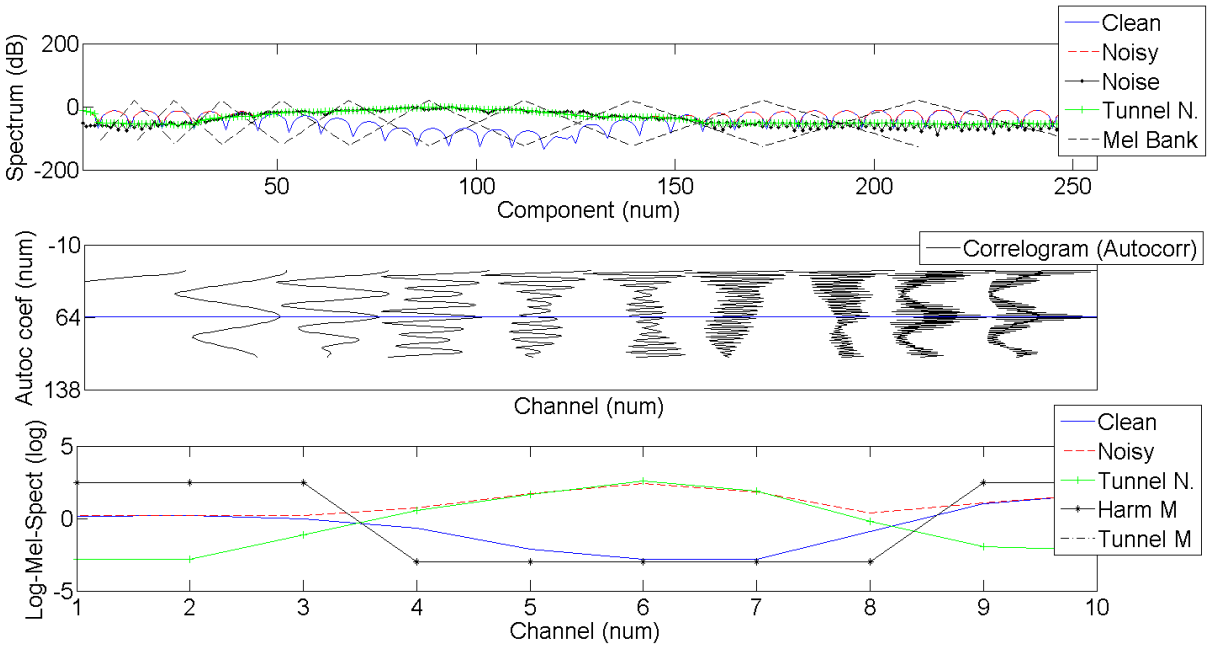



Figura 7.1: Equiparación entre el mecanismo de estima de la máscara túnel y de la máscara armónica.

Este estudio nos lleva a la pregunta de si existe un mecanismo óptimo de robustecimiento de los segmentos sonoros basado en el pitch y de si este ya está siendo empleando por alguno de los mecanismos básicos encontrados. La Sec. 7.2 intentará responder a estas preguntas.

### 7.1.2. Equiparación máscara túnel y armónica

Vamos a mostrar que el mecanismo de estima de la máscara a partir de una estimación túnel del ruido es equiparable al mecanismo de estimación a partir de la armonicidad. Primero vamos a mostrarlo con un ejemplo y luego vamos a razonar esta equiparabilidad.

#### Ejemplo que muestra la equiparabilidad

Supongamos que tenemos un segmento ruidoso ($y = x + n$) suma de una señal limpia sonora ($x$) (con un determinado pitch) y un ruido ($n$). En la Fig. 7.1 de arriba, se muestran los correspondientes espectros limpio, ruidoso, del ruido real, de la estima túnel del ruido (que se acerca mucho al real) y el banco de filtros Mel empleado compuesto por diez canales (este ha sido implementado tanto en su versión de pesado espectral como en su versión de

convolución temporal para poder obtener el correlograma). En el centro podemos ver el correlograma compuesto por las diez autocorrelaciones correspondientes a cada filtro Mel. Estas autocorrelaciones están cruzadas por una línea que nos indica cual es el coeficiente del pitch y de la cual se puede estimar la armonicidad de cada autocorrelación y por lo tanto la máscara armónica. Abajo podemos ver los espectros Log-Mel limpios, ruidosos y de la estima túnel del ruido. Comparando el espectro Log-Mel ruidoso con el del ruido túnel se puede estimar la máscara túnel. También abajo tenemos las estimas de la máscara túnel y de la máscara armónica (si la voz domina sobre el ruido se representa con valor alto y bajo en caso contrario). Podemos observar que ambas máscaras se superponen mostrándose la equiparabilidad entre ambos mecanismos.

### Razonamiento de la equiparabilidad

La razón de que ambas estimas obtengan la misma máscara se explica considerando los dos casos siguientes:

1) Supongamos un canal del banco de filtros en el que la máscara debe de valer 1 (p. ej. el canal 9 de la Fig. 7.1). Esto implica que debido a la dominancia y a la periodicidad de la señal limpia tendremos que la forma espectral ruidosa para ese canal será de tipo peine (un conjunto de armónicos separados la frecuencia del pitch). La forma peine implica que el espectro Log-Mel de la estima túnel del ruido (obtenido mediante el pesado Mel del ruido túnel) siempre estará por debajo del espectro ruidoso por lo tanto la máscara túnel valdrá 1. La forma peine también implica que la autocorrelación del correlograma de ese canal tendrá un pico alto en el coeficiente del pitch (debido a la modulación AM que producen el conjunto de armónicos que entran en el canal, Sec. 3.3), produciéndose una alta armonicidad y por lo tanto la máscara armónica valdrá 1 también.

2) Supongamos un canal del banco de filtros en el que la máscara debe de valer 0 (p. ej. el canal 6 de la Fig. 7.1). Esto implica que, debido a la dominancia del ruido, el espectro o bien tiene forma aleatoria o bien tiene forma peine (si es otra fuente armónica pero con un pitch diferente). Tenga la forma que sea, si se piensa, el ruido Log-Mel túnel estimado siempre estará al mismo nivel o por encima del espectro ruidoso y por lo tanto la máscara túnel valdrá 0. También tenga la forma que sea, la armonicidad (guiada por la línea del pitch) será baja y por lo tanto la máscara túnel valdrá 0.

Estos dos casos nos hacen ver que ambos mecanismos (uno inspirado en cómo se produce la voz y otro en cómo se percibe esta) van a tender a dar siempre las mismas máscaras y, por lo tanto, resultados de reconocimiento similares (hipótesis H1). Mencionar

que obviamente, esto ocurrirá siempre que se haga una correcta elección del umbral de la SNR en la máscara túnel (Sec. 3.2.2) y del umbral de la armonicidad en la máscara armónica (Sec. 5.2.3).

## 7.2. Mecanismos óptimos sonoros

### 7.2.1. Estimación óptima del ruido basada en el pitch

#### Límites en la recuperación de información a partir del pitch

Para responder a la pregunta de cuál es el mejor mecanismo de robustecimiento de un segmento sonoro, antes debemos de conocer la máxima información que podemos recuperar a partir del pitch de un segmento periódico puro contaminado por ruido.

Para ello debemos de plantear estas cuestiones de manera formal, aunque con cierta pérdida de generalidad, siguiendo las idealizaciones de la sección Sec. 6.2.3. Supongamos que tenemos un segmento ruidoso $x(n)$ de longitud $N$ muestras el cual es suma de una señal limpia periódica pura $p(n)$ de periodo o pitch $T$ muestras (u $\omega_0$ radianes) y un ruido o distorsión $d(n)$ que recoge, además del ruido, las posibles no periodicidades del segmento sonoro. Por simplicidad también suponemos que tenemos un número de periodos $N_p$ entero ($N = N_p * T$). Según esto, nuestra pregunta se transforma ahora en saber qué porcentaje de la información contenida en las $N$ muestras de la señal periódica pura $p(n)$ podemos llegar a recuperar empleando $x(n)$, $T$ y cualquier tipo de procesado o transformación que no añada información extra sobre la señal periódica o el ruido.

La respuesta a esta pregunta es que el porcentaje máximo de información contenida en las $N$ muestras de la señal periódica que podemos llegar a recuperar empleando solamente la señal ruidosa y el pitch es nulo, y que a lo máximo que podemos aspirar idealmente es a recuperar solamente un $100(N_p - 1)/N_p$ por ciento de la información del ruido.

Esto se demuestra fácilmente trasladando la información de las señales al dominio espectral complejo (no de la magnitud que produce pérdidas de información) tal y como mostramos a continuación. En el dominio espectral complejo tenemos que la señal ruidosa vale (aplicando simplemente una DFT de $N = TN_p$puntos):

$$X(\omega_k) = P(\omega_k) + D(\omega_k) \qquad (k = 0, ..., N-1) \tag{7.1}$$

Teniendo en cuenta que la transformada de Fourier de una señal periódica pura (al tener

un número entero de periodos) vale 0 salvo en los armónicos del pitch tenemos que:

$$X(\omega_k) = \begin{cases} P(\omega_k) + D(\omega_k) & \text{si} \quad \omega_k = \omega_0 m \\ D(\omega_k) & \text{en otro caso (muestras túnel)} \end{cases} \tag{7.2}$$

donde $m = 0, 1, .., T-1$. Esta ecuación muestra que la información de la señal periódica siempre queda modificada por el ruido sea cual sea el procesado que apliquemos, y que del ruido podemos llegar a recuperar solamente las $N(Np-1)/Np$ muestras espectrales contenidas entre los armónicos del pitch (muestras túnel del ruido).

### Ruido túnel como estima óptima

Olvidándonos de la fase del espectro de la Ec. 7.1 (que no da información de reconocimiento), los límites de recuperación nos señalan que a partir de las muestras túnel de la magnitud espectral podemos estimar el ruido de manera óptima siempre y cuando poseamos un modelo para el ruido. La estima túnel del ruido, tal y como se ha propuesto en la Sec. 6.3.3, parece desperdiciar información de la que se podría llegar a emplear para estimar el ruido (pues sólo se toma una muestra túnel entre dos armónicos cuando en verdad se podrían llegar a tomar hasta $N_p - 1$ muestras). Es más, podría parecer que el modelo o interpolación lineal de ruido empleado (basado en suponer que el ruido continúa linealmente entre dos muestras túnel) es demasiado simple y que se podría haber empleado un modelo más complejo y adecuado.

Sin embargo, experimentos preliminares en los cuales se ha controlado tanto la anchura del segmento de muestras túnel tomadas como el tipo de modelo de ruido o interpolación supuesta, nos han mostrado que aumentando la anchura túnel no se gana en los resultados y que usando un modelo polinómico o ARMA en lugar de uno lineal tampoco se mejoran mucho los resultados. La razón que explica el primer fenómeno reside en que, al no ser perfectamente periódica la señal sonora así como por tener aplicada una ventana de longitud finita, su espectro llega a ocupar muestras de la región túnel, por lo que tomar como ruido toda la región puede ser más perjudicial que beneficioso. Esto se ha comprobado experimentalmente incluso tomando tamaños de ventana que ocupen un número entero de periodos.

La razón que explica el segundo fenómeno es que, en principio, el ruido puede obedecer a cualquier modelo y que las ventajas que parece ofrecer el suavizado de una interpolación ARMA o polinómica, también las lleva el modelo lineal gracias a que al final el banco de filtros MEL siempre se encarga de suavizar el espectro estimado.

Todos estos razonamientos nos llevan a la conclusión de que la estima túnel del ruido de la Sec. 6.3.3 y estimas similares basadas en filtros peine del ruido pueden considerarse óptimas (en condiciones ideales) pues aprovechan al máximo toda la información posible que se puede obtener con el pitch suponiendo muy poca información sobre el ruido (tal como el modelo de interpolación).

### 7.2.2. Mecanismos óptimos sonoros

Teniendo en cuenta los tres puntos siguientes:

1. Que la estimación túnel del ruido es óptima en el sentido de aprovechar al máximo la información del pitch (Sec. 7.2.1).

2. La equivalencia entre la máscara túnel y máscara armónica (Sec. 7.1.2).

3. Las ventajas que ofrece el reconocimiento de MD empleando máscaras frente a otro tipo de técnicas como SS (Sec. 5.1.6 y 4.2.2).

Podemos decir que los mecanismos de estima de máscaras basados en el ruido túnel o en armonicidad para un reconocedor de MD constituyen una excelente aproximación al problema del reconocimiento robusto basado en el pitch de los sonidos sonoros, y que en condiciones ideales los podemos considerar como mecanismos óptimos (hipótesis H2).

### 7.2.3. Resultados experimentales

Para mostrar experimentalmente las diferentes hipótesis hechas en esta sección sobre los cuatro mecanismos básicos sonoros, vamos a comparar en el dominio espectral (o coclear) y con MD, los resultados de reconocimiento de diferentes técnicas, representantes de cada uno de los mecanismos básicos. Para sacar a la luz el resultado exclusivo del mecanismo sonoro, emplearemos pitch ideal y máscara oráculo sobre los segmentos sordos y de silencio.

Todo esto posibilitará una comparación justa, evitando que los resultados mostrados estén influenciados entre otras cosas: por el empleo de un dominio diferente (tales como el cepstral), por el empleo de técnicas de compensación extras añadidas (tales como CMN) y por la mala estima del pitch. En la Tab. 7.1 podemos ver estos resultados.

En la primera columna (*Técnica «per se»*) se muestran los resultados de las diferentes técnicas sin emplear información oráculo (solamente el pitch ideal). En la columna central se muestran los resultados de cada uno de lo mecanismos básicos de los segmentos

| | Media (20-0 dB) [0 dB] | | |
|---|---|---|---|
| Técnica | Técnica «per se» (sin oráculos) | Máscara oráculo en sordos y sil. | Máscara oráculo en todos |
| FE (Espectr.) | 33.30 [7.66] | 64.25 [25.04] | 95.01 [90.18] |
| $DDR_{55,200}$ (Espectr.) | 35.84 [5.84] | 73.16 [37.98] | 90.35 [82.75] |
| A. Sift ($\delta = 8$) (Espectr.) | 36.61 [8.09] | 77.92 [47.72] | 93.36 [88.94] |
| N. VAD+Harm (Cocl.) | 85.95 [72.21] | 89.15 [73.13] | 95.11 [89.40] |
| N. VAD+Tun (Espectr.) | 87.21 [74.43] | 90.87 [79.46] | 95.01 [90.18] |

Tabla 7.1: Resultados de reconocimiento WAcc % sobre toda Aurora-2 (20-0 dB), obtenidos por las diferentes técnicas representantes de los cuatro mecanismos básicos sonoros. Entre corchertes se muestra el resultado a 0 dB.

sonoros (en segmentos sordos y de silencio empleamos máscara oráculo). En la columna de la derecha mostramos los resultados empleando la máscara oráculo sobre todos los segmentos. A continuación explicamos las diferentes técnicas empleadas en relación a la primera columna:

- *FE* reconoce directamente empleando el espectrograma contaminado (con máscara todo 1s). Es representante de no aplicar ningún mecanismo de robustecimiento sobre los segmentos sonoros .

- $DDR_{55,200}$ reconoce empleando la estima del espectrograma limpio dado por la ventana asimétrica (Sec. 6.1) con máscara todo 1s. Es representante de los mecanismos basados en la estructura armónica.

- *A. Sift* reconoce empleando la estima del espectrograma limpio dado por la autocorrelación cribada (Sec. 6.2) con máscara todo 1s. Es representante de los mecanismos de estima peine de la señal limpia.

- *N. VAD+Harm* reconoce empleando el espectrograma contaminado y la máscara estimada mediante la adaptación de la Técnica de Barker (Sec. 5.2.3). Es representante de los mecanismos de estima de la máscara mediante la armonicidad.

- *N. VAD+Tun* reconoce empleando el espectrograma contaminado y la máscara estimada mediante la estima del ruido basada en el pitch propuesta en la Sec. 6.3. Es representante de los mecanismos de estima túnel del ruido.

Los parámetros de umbrales y pendientes de las sigmoides (Sec. 3.2.2) de las técnicas *N. VAD+Harm* y *N. VAD+Tun* han sido re-optimizados para obtener los mejores resultados posibles en la segunda columna valiendo ahora: $\beta = -6$ dB y $\alpha = 1{,}6$ para $M_n$, $\beta = 0{,}75$ y $\alpha = 200$ para $M_h$ (máscara armónica), y $\beta = -3$ dB y $\alpha = 0{,}2$ para la máscara túnel.

Fijándonos en la columna central, podemos sacar las siguientes conclusiones respecto a los mecanismos sonoros:

1. Los mecanismos basados en la estructura armónica, al emplear poco conocimiento sobre el ruido y no emplear el pitch de la señal, obtienen los peores resultados de reconocimiento, aunque producen mejoras en comparación a no hacer nada (*FE*).

2. Los mecanismos basados en estimar la señal limpia mediante filtros peine obtienen mejores resultados que los de la estructura armónica debido a que emplean el valor de pitch para eliminar el ruido en las regiones túnel. A pesar de esto, no pueden alcanzar resultados óptimos por dos motivos: El primero es por no limpiar el ruido de los armónicos del pitch. Sin embargo, si se elimina parte de este ruido haciendo ciertas suposiciones sobre el mismo, se pueden mejorar los resultados como es el caso de la técnica *A. Sift* (respecto a un promediado). El segundo es por la «no perfecta periodicidad de la señal sonora» haciendo que el muestreo de los armónicos del pitch no sea perfecto. Este es el motivo de que este tipo técnicas, para incrementar sus resultados, tengan que emplear un pitch muy fino (tal y como hace la técnica FPM-SE [19]) o tengan que aplicarse también en el entrenamiento para equilibrar las no periodicidades (tal y como hace *A. Aver* y *A. Sift*).

3. Los mecanismos básicos que mejores resultados de reconocimiento dan son los basados en la estima de las máscaras mediante el ruido túnel y armonicidad para reconocimiento con MD. Es más, se ve que ambos resultados son muy parecidos aunque siendo un poco mejor la estima túnel. Este incremento puede ser debido a la diferencia entre la escala Mel del espectrograma y la ERB del cocleograma. Salvando esta diferencia, podemos decir que ambos mecanismos son equiparables y óptimos en el sentido de ser los que mejor aprovechamiento hacen del pitch (emplean la máxima información que se puede obtener sobre el ruido a partir del pitch). Todo esto confirma las hipótesis H1 y H2 de las secciones precedentes.

## 7.3. Limitaciones del reconocimiento basado en el pitch

### 7.3.1. Límites en el rendimiento

Si comparamos las columnas primera y segunda de la Tab. 7.1 para la técnica propuesta *N. VAD+Tun* y tenemos en cuenta que la segunda columna contiene los límites de las técnicas basadas en el pitch (pues los sordos y los silencios llevan máscara oráculo y los sonoros son robustecidos mediante uno de los mecanismos sonoros óptimos), podemos concluir que la técnica propuesta de estima del ruido basada en el pitch (primera columna) es casi óptima (empleando pitch ideal) pues se acerca a los límites del reconocimiento basado en el pitch (segunda columna) empleando la mínima información posible sobre el ruido. Sus resultados no están excesivamente lejos de los de las máscaras oráculo (columna tercera). Sin embargo, si se quiere alcanzar estos resultados se debe de añadir más información (referente al ruido o la voz) en la estima de máscaras para alcanzar los límites oráculo.

### 7.3.2. Reconocimiento de voz sin valores de pitch

Toda esta Tesis está pensada para reconocer voz suponiendo que esta posee un solo pitch, es decir bajo la hipótesis de que la fuente principal de excitación es periódica (o cuasi-periódica), que es tal y como normalmente se presenta (2.1.3). Sin embargo, la voz a veces se puede presentar sin pitch (voz susurrante, [159]) o incluso con múltiples valores de pitch (segundas voces musicales) y el ser humano puede reconocerla sin problemas incluso en condiciones de ruido.

Todo esto podría llegar a dar la sensación de que que el pitch no es importante en el reconocimiento robusto. Sin embargo, tal y como se ha mencionado en la introducción de la Tesis (Sec. 1.1), debemos considerar que el pitch es una pista muy importante para separar la voz del ruido, aunque no la única. Pistas como las propuestas por el marco CASA (comienzo/final común de fragmentos frecuencio-temporales, modelos de alto nivel, etc., Sec. 2.2.6) podrían ser empleadas para abordar este tipo de voz. El estudio e implementación de tales pistas es un campo aún no muy explorado [67, 159], y es una de las lineas futuras de investigación que nos gustaría desarrollar aplicando ciertas ideas presentadas en esta Tesis. Entre estas ideas podemos mencionar:

- Tener en cuenta el modelo de fuente principal de la voz (Sec. 2.1.3) para localizarla y separarla del ruido de forma similar a como hace el VAD propuesto en la Sec. 6.3.3.

La fuente principal, en la Tesis, está donde hay vibración de las cuerdas vocales o pitch, pero ahora podría estar donde se detecten fragmentos frecuencio-temporales largos y con alta SNR local (en el caso de voz susurrante) o donde se detecten apariciones simultaneas de múltiples valores de pitch (en segundas voces).

- Reconocer empleando MD, aunque quizás para la voz susurrante lo ideal sea emplear SFD [5] ya que esta técnica permite emplear reglas de alto nivel para separar la voz del ruido.

Otras ideas y técnicas que se deberían desarrollar para abordar este tipo de voz (sobre todo la susurrante) podrían ser:

- Adaptar y mejorar los modelos a este nuevo tipo de voces considerando su nueva forma espectral respecto a la voz normal (tendencia al aplanamiento espectral de los formantes, disminución de la energía de los sonidos sonoros, etc..) [159, 67].

# 7. EQUIVALENCIAS Y LÍMITES DE LAS TÉCNICAS BASADAS EN EL PITCH

# Capítulo 8

# Conclusiones, Contribuciones y Trabajo Futuro

## 8.1. Conclusiones

Esta Tesis tiene como motivación principal la de proponer y hacer un estudio comparativo de las técnicas de ASR (Automatic Speech Recognition) robusto basadas en el pitch, entendiendo por técnicas basadas en el pitch aquellas que aprovechan la presencia del pitch en la voz para robustecer el reconocimiento en condiciones de ruido. A continuación resumimos las conclusiones más importantes obtenidas en esta Tesis:

- Teniendo en cuenta que el mensaje de la señal de voz se codifica mediante tres tipos de elementos (los sonidos sonoros, los sonidos sordos y los silencios) y la forma en la que estos se combinan, podemos decir que la señal de voz consiste «principalmente» de sonidos sonoros rodeados por sonidos sordos. Esto se ha denominado «modelo de fuente principal» el cual es una definición simplificada de voz que ha sido usada para desarrollar un VAD (Sec. 6.3.3). Este modelo también es valido en el caso de voz susurrante si se tiene en cuenta que en este caso la fuente principal es más bien un ruido.

- El estado actual de las técnicas convencionales de ASR robusto nos lleva a concluir que las técnicas de MD pueden obtener resultados de reconocimiento muy elevados (similares a los del ser humano) sin necesidad de estimar perfectamente el ruido o la señal limpia. Sin embargo, estas trasladan el problema a la estimación de la máscara de reconocimiento.

- El estudio comparativo de las diferentes técnicas de ASR robusto basadas en el pitch (técnicas de aprovechamiento de la estructura armónica, de estimación de la señal limpia y de estimación de máscaras) no es sencillo debido a que cada autor emplea un extractor de pitch diferente, al empleo de técnicas extras añadidas y a que puede llegar a confundirse la técnica de robustecimiento basada en el pitch con la técnica de extracción del pitch. Por estas razones, se han establecido ciertas equivalencia entre las diferentes técnicas, así como los límites del reconocimiento basado en el pitch.

- Se ha propuesto un conjunto de ventanas asimétricas denominado $DDR_{c,w}$ que extiende la técnica HASE [142] empleada para robustecer la extracción de características cepstrales a partir de la OSA en ruidos poco autocorrelados (contenidos en los primeros coeficientes de autocorrelación). Se ha concluido que los coeficientes de autocorrelación que más peso deben de tener son los correspondientes al pitch pues son los más energéticos (con mayor SNR) y los que más información de reconocimiento transportan. Los de menos peso deben ser los primeros por ser los más contaminados por el ruido.

- Se ha propuesto una estimación de la autocorrelación denominada *cribada* (basada a su vez en otra estimación propuesta denominada *promediada*). Esta usa el pitch y depende de un parámetro de criba $\delta$ que indica la cantidad de productos de autocorrelación rechazados, los cuales se supone que están más contaminados por el ruido. Se ha demostrado que eligiendo un valor de $\delta$ de forma que incluya los principales coeficientes de autocorrelación de un ruido poco autocorrelado, la estima puede ser igual a la de la señal limpia bajo ciertas suposiciones.

- Teniendo en cuenta que para $\delta = 0$ la autocorrelación cribada se convierte en un filtrado peine (o muestreo de los armónicos de la señal contaminada) y que muchas de las técnicas basadas en el pitch se pueden reducir a un filtrado peine, podemos concluir que la autocorrelación cribada es una representante de las técnicas peine, que reúne las ventajas de estas (de eliminar el ruido entre los armónicos del pitch) y de las técnicas tipo HASE (de eliminar ruidos poco autocorrelados).

- La extensión a los segmentos sordos, tanto de las ventanas $DDR_{c,w}$ como de la técnica de cribado podría rebajar el rendimiento, principalmente en condiciones limpias, debido a que la información de estos segmentos está contenida principalmente en los primeros coeficientes de autocorrelación, los cuales tienden a ser eliminados. Sin

embargo este problema puede ser paliado aplicando la técnica tanto en la etapa de entrenamiento como en la de test.

- Técnicas como HT [38] o la de Frazier [46] basadas en estimar el espectro del ruido de los segmentos sonoros contaminados empleando las muestras túnel (muestras espectrales entre los armónicos del pitch), sufren de incluir como ruido a los segmentos sordos (no usan VAD) y de sobrestimarlo rebajando el rendimiento debido a que también emplean SS, la cual es muy sensible a estas sobrestimaciones. Para evitar estos problemas se ha propuesto un sistema de reconocimiento que hace una estimación de ruido VAD+Túnel y que emplea MD en lugar de SS.

- El VAD propuesto parte del pitch para localizar el resto de los elementos de la voz considerando el modelo de fuente principal. La estima túnel también emplea el pitch. Por ello, podemos concluir que la estimación propuesta VAD+Túnel se trata de una *estima del ruido completamente basada en el pitch*.

- Si no atendemos a los detalles de como procesan los segmentos sordos y los silencios, el extractor de pitch empleado, etc., podemos considerar que las técnicas basadas en el pitch emplean uno estos cuatro mecanismos básicos de robustecimiento de los segmentos sonoros: aprovechamiento de la estructura armónica, filtrado peine para estimar la señal limpia, estima túnel del ruido (o anti-filtrado peine para estimar el ruido) que puede ser empleada en SS (HT) o para estimar máscaras (p. ej. nuestra propuesta de estima de ruido) y estimación de la máscara mediante la armonicidad.

- La cantidad máxima de muestras espectrales del ruido que se pueden recuperar de un segmento sonoro contaminado empleando solamente el pitch son (en condiciones ideales) las $N(N_p - 1)/N_p$ muestras túnel, donde $N$ es el tamaño de segmento y $N_p$ el número de periodos de la señal sonora. De esto se deduce que para estimar el ruido es necesario añadir más información sobre el mismo y, precisamente, esto es lo que hacen las estimas túnel (empleadas en técnicas como HT, FPM-NE o nuestra propuesta) al interpolar el ruido a partir de estas muestras túnel. De esto podemos concluir que (idealmente) este tipo de técnicas hacen estimaciones óptimas del ruido basándose en el pitch y en muy poca información sobre el ruido (modelo de interpolación).

- Se puede mostrar que las máscaras de los segmentos sonoros obtenidas mediante el ruido túnel y la armonicidad son muy similares. Teniendo en cuenta que el ruido

túnel es óptimo y las ventajas que ofrece MD frente a SS podemos considerar que los mecanismos de reconocimiento basados en este tipo de máscaras pueden pueden ser considerados como mecanismos óptimos (al menos, bajo ciertas condiciones) de aprovechamiento de la información del pitch para reconocer los segmentos sonoros. Los resultados experimentales ayudándose de máscaras oráculo así lo han demostrado.

- Teniendo en cuenta los mecanismos óptimos de los segmentos sonoros y los resultados empleando máscaras oráculo (sobre los segmentos sordos y de silencio), podemos concluir que la técnica propuesta de estima del ruido basada en el pitch se aproxima al rendimiento óptimo (empleando pitch ideal) pues se acerca a los límites del reconocimiento basado en el pitch (empleando la mínima información posible sobre el ruido). Adicionalmente, sus resultados no están muy lejos de los de las máscaras oráculo. Si se quiere alcanzar estos resultados será necesario añadir más información (referente al ruido o la voz) en la estima de máscaras para alcanzar los límites oráculo.

- Algunas de las ideas presentadas en la Tesis tales como el empleo de MD o el modelo de fuente principal para obtener un VAD, pueden ser recicladas para reconocer voz susurrante en la que no hay pitch.

## 8.2. Contribuciones

Las principales contribuciones de esta Tesis se pueden resumir en:

- Proponer un conjunto de ventanas asimétricas, que se aplican sobre la OSA para hacer estimación espectral robusta las cuales, con poca cantidad de cálculo, ayudan a mejorar el reconocimiento en condiciones de ruido [107].

- Proponer dos estimadores de la autocorrelación limpia que usan el pitch y que pueden hacer frente tanto a ruidos tipo armónicos (autocorrelación promediada y cribada) como ruidos poco autocorrelados (cribada). Se ha mostrado que la cribada puede llegar a estimar de forma exacta la autocorrelación limpia bajo ciertas condiciones [106].

- Proponer un VAD y un estimador del ruido basado en el pitch a partir de un modelo simplificado de la voz (modelo de fuente principal) el cual soluciona muchos de los

problemas asociados a técnicas de estima de ruidos similares, tales como la inclusión como parte del ruido de los sonidos sordos y las sobrestimaciones del mismo [105].

- Estudiar las diferentes técnicas basadas en el pitch, clasificarlas, mostrar ciertas equivalencias y señalar los límites del reconocimiento basado en el pitch, mostrando que la técnica propuesta de estimación del ruido basada en el pitch se acerca a estos límites [Tesis].

## 8.3. Trabajo Futuro

Muchos de los experimentos realizados en la Tesis (tales como los basados en pitch ideal) nos indican qué trabajos futuros son de mayor interés a partir de las ideas y técnicas desarrolladas en la Tesis. A continuación hacemos una síntesis de los mismos:

- Respecto a las **ventanas asimétricas** podríamos realizar reconocimiento en función del pitch promedio del hablante (relacionado con el género) empleando ventanas centradas sobre dicho pitch ya que esto mejoraría en gran medida los resultados tal y como señalan los experimentos de la Sec. 6.1.6.

- Respecto a la **autocorrelación cribada**, tal y como hemos visto en los resultados con $\delta$ oráculo, se podría emplear un $\delta$ dinámico dependiente del ruido para mejorar los resultados. Es más, podríamos extender la idea de criba eliminando no sólo los productos que están alrededor de la diagonal principal sino alrededor de otras diagonales en función del ruido presente.

- Respecto a la **estima del ruido basada en el pitch** podemos decir que un punto clave será el de mejorar el extractor de pitch, pues tal y como muestran los resultados de la Tab. 7.1, haciendo esto estaríamos prácticamente alcanzando los límites del reconocimiento basado en el pitch (incluso sin necesidad de mejorar el VAD). Esta extracción se podría realizar al mismo tiempo junto con la estima del ruido y el reconocimiento de la voz mediante el empleo de un reconocedor SFD (Speech Fragment Decoding) de forma similar a como lo hace la técnica de Ma (Sec. 5.2.3). Para ello, el extractor de pitch podría considerar diferentes candidatos de pitch (segmentos de pitch superpuestos) y cada candidato podría resultar en una estima del ruido diferente. Estas hipótesis paralelas podrían ser evaluadas separadamente con un reconocedor de MD y elegir aquella que resulte en mayor probabilidad de reconocimiento.

- Otro trabajo muy interesante que nos señala la tabla 7.1 es el de intentar alcanzar los límites de las máscaras oráculo, sobre todo a bajas SNRs. Como hemos visto, únicamente mediante el pitch no podemos alcanzar esos límites y la forma de hacerlo es añadir información del ruido o de la señal de voz. Esta información podría actualizarse dinámicamente en función del ruido de las partes de silencio y ser empleada en la estimación de la máscara.

- Por último mencionar que el reconocimiento de voz sin o con múltiples valores de pitch (voz susurrante y con segundas voces) es una línea de gran interés. Ha sido comentada en detalle en la Sec. 7.3.2.

# Apéndice A

# Anexos de la Tesis

## A.1. Parámetros de reconocimiento

Vamos a detallar los parámetros de los distintos sistemas de reconocimiento empleados en esta Tesis y que se corresponden con los sistemas de las Fig. 5.2, 5.5, etc. En general podemos decir que todos los sistemas están compuesto por un front-end (que lleva incorporado las técnicas de robustecimiento) y el reconocedor. Con el objetivo de hacer una comparación justa de las distintas técnicas que estudiamos a lo largo de este capítulo (ventanas asimétricas, autocorrelación cribada y ruido basado en el pitch), hemos procurado que las distintas representaciones acústicas (cocleograma, espectrograma y cepstrograma) que usan los reconocedores sean lo más parecidas posibles. Teniendo en cuenta esto, tomaremos los siguientes parámetros.

**Respecto al extractor de características** usaremos los parámetros puestos como ejemplos a la hora explicar las distintas representaciones acústicas (Sec. 3.1) debido a que las hacen muy parecidas entre ellas (Sec. 3.1.5) y porque son muy similares a los que lleva el FE estándar de la ETSII [149, 120]. Los parámetros son los siguientes: Frecuencia de muestreo 8000 Hz, realce de las altas frecuencias (con preénfasis para espectrograma y con ganancia en los filtros gammatone para cocleograma), longitud y desplazamiento entre segmentos 10 y 32 ms (80 y 256 muestras), longitud-ventana 256-Hamming para señal, $256 - DDR_{c,w}$ para la OSA y 511-DDR para autocorrelaciones completas, componentes de la MSD 512 (rango $[0, 2\pi]$), canales del banco de filtros 23 (ya sea mel o gammatone), valor mínimo -2.80 para espectrograma y -6.20 para cocleograma, y coeficientes cepstrales 13 (C0,..,C12, no empleamos logE o logaritmo de la energía ya que este no lleva ningún mecanismo de compensación) todos con CMN.

**Respecto al reconocedor** (Sec. 4.1.2) usaremos los parámetros más comúnmente empleados para evaluar las bases de datos Aurora-2 y Aurora-3 [120]. Los parámetros son los siguientes: Tamaño de los vectores de características 46 componentes para espectrales y cocleares (23-estáticos + 23-velocidades), y 39 para los cepstrales (13-estáticos + 13-velocidades + 13-aceleraciones). Número de estados: 1 para la pausa, 3 para el silencio y 16 para las palabras. Número de gaussianas por estado para espectrograma y cocleograma: 11 para silencio y pausa y 9 para las palabras, para cepstrograma es: 6 y 3 respectivamente (el cepstrograma requiere menos gaussianas debido a la decorrelación entre canales).

Por último mencionar que, salvo las técnicas que emplean reconocedor de MD (que no lo requieren), el entrenamiento y el test se harán con los mismos parámetros de la técnica que se esté evaluando.

## A.2. Bases de datos

Evaluamos nuestros sistemas sobre dos bases de datos clásicas Aurora-2 y Aurora-3.

**Aurora-2** [120, 62] posee frases contaminadas artificialmente con 10 tipos de ruidos diferentes: subway, babble, car y exhibition para Set-A, restaurant, street, airport y train para Set-B, y subway-mirs y street-mirs para Set-C. Cada uno de estos ruidos es mezclado a 7 niveles de SNR diferentes: clean, 20, 15, 10, 5, 0 y -5 dB. Todo esto nos da un total de 70 conjuntos de test de 1001 frases cada uno. Los ruidos de Set-C son convolutivos (no aditivos) y tratan de imitar situaciones más realistas. El entrenamiento se puede hacer en limpio (que es el que emplearemos nosotros siempre) o ruido para robustecer los modelos (entrenamiento Multicondición [120] con los mismos ruidos del Set-A).

**Aurora-3** [4, 3] posee frases contaminadas realmente con ruido de coche. Atendiendo a si el micrófono está cerca (ch0) o lejos (ch1) de la boca y atendiendo a si el ruido del motor es silencioso (q), medio (m) o fuerte (l) existen 6 conjuntos de frases. Según los conjuntos empleados para entrenar y testear se distinguen 3 condiciones de prueba o de discrepancias entrenamiento-test: well-matched (WM), medium mismatch (MM) y high mismatch (HM). Podemos decir que la peor condición o la que dará peores resultados de reconocimiento será HM ya que es casi el equivalente a entrenar con limpio y testear con ruido. En las otras se entrena con ruido y se testea con ruido también en mayor o menor grado por lo que son un tipo de entrenamiento Multicondición. Existen varias clases de Aurora-3 según el idioma de las frases. En esta Tesis emplearemos Español (Spanish) [4] y Danés (Danish) [3].

| | **Intervalos de confianza (%)** | |
|---|---|---|
| WAcc (%) | Aurora-2 | Aurora-3 |
| 70,00 | $70{,}00 \pm 0{,}78$ | $70{,}00 \pm 0{,}71$ |
| 80,00 | $80{,}00 \pm 0{,}68$ | $80{,}00 \pm 0{,}62$ |
| 90,00 | $90{,}00 \pm 0{,}51$ | $90{,}00 \pm 0{,}47$ |

Tabla A.1: Intervalos de confianza con un 95 % de probabilidad, en función del WAcc, para los conjuntos de test completos de Aurora-2 y Aurora-3.

## A.3. Tasas de acierto e intervalos de confianza

El WAcc (Word Accuracy, tasa de Acierto de Palabra) es una medida usualmente empleada para medir el rendimiento de un sistema de reconocimiento. Su valor es opuesto al WER (Word Error Rate, WAcc $= 1 -$ WER) y se obtiene según la formula WAcc $= (H - I)/N$, donde $H$ es número de palabras acertadas, $I$ el número de palabras insertadas y $N$ el número total de palabras testadas.

El intervalo de confianza del WAcc (o intervalo en el que podemos asegurar que siempre estarán nuestros resultados con un $(1 - \alpha)$ de probabilidad y por lo tanto que nos indicará como de seguras son nuestras conclusiones) dependerá en última instancia de $N$, siendo más estrecho a más palabras testadas. El WAcc puede ser visto como una distribución binomial (ya que se trata de una medida de clasificación acierto/error) o, si $N$ es lo suficientemente grande (mediante aplicación del teorema central del límite), como una distribución normal $N(0, 1)$ de forma que el intervalo de confianza puede obtenerse como,

$$\text{WAcc} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\text{WAcc}(1 - \text{WAcc})}{N}} \tag{A.1}$$

donde para un $(1 - \alpha) = 0{,}95$ (probabilidad del 95 %), $z_{1-\frac{\alpha}{2}} \approx 1{,}96$. En Aurora-2 (Set A, B y C) el número total de palabras distintas testadas (sin considerar que se repiten con distintos tipos de ruidos) es de $N = 13159$, por lo que para unos resultados típicos de reconocimiento (WAcc $= 70, 80, 90$ %) podemos establecer los intervalos de confianza mostrados en la Tab. A.1. En el caso de Aurora-3, en el que $N = 15834$, los intervalos de confianza son un poco más pequeños. Estos intervalos de confianza tan estrechos justifican las conlusiones extraidas de los resultados mostrados en la Tesis con estas bases de datos.

# A. ANEXOS DE LA TESIS

# Apéndice B

# Summary of the Thesis: Pitch-based Robust Speech Recognition Techniques

## B.1. Introduction

### B.1.1. Motivations

#### Importance of pitch in robust speech recognition

Acoustic noise represents one of the major challenges for ASR (Automatic Speech Recognition) systems. Many different approaches have been proposed to deal with this problem in monaural signal [121, 65, 155] and many of them try to employ some kind of noise information to do robust ASR. However, when one wants to deal with all kind of noises it is clear that the most important information to separate noise from speech is just speech information. There exits many cues and informations which help to distinguish speech from noise but at the end the correct choice will depend on what is defined as speech. Speech can be emitted in many different ways which mainly depend on the considered type of the «main source». These ways can be whispering, vocal harmony speech (in music), etc.. In this Thesis it will be considered that speech is emitted in its normal way, with vibration of the vocal folds and with only one pitch at each time instant.

Continuing with the search for the most important cues, this Thesis will particularly consider the signal pitch due to the three following reasons:

1. Many psychoacoustics experiments, such as those shown in [33, 155], reach the conclusion that very often humans use pitch to separate speech from noise.

2. Pitch is a useful information to distinguish different types of speech segments (voiced, unvoiced and silence) and to separate speech and noise signals.

3. Many robust ASR techniques inspired in human recognition, as shown in [155], use pitch.

**Robust techniques based on pitch**

The comparison of the different ASR techniques based on pitch is not an easy matter because of several reasons:

1. Each author uses a different pitch extractor to evaluate his technique.

2. It is not clear which is the real cause for obtaining different results: different methods applied to voiced and unvoiced sounds, application of additional techniques (such as cepstral normalization, missing data approaches,...), etc.

3. Sometimes it is not clear whether an author is proposing either a new technique for robust ASR based on pitch or a new robust pitch extractor (or both at the same time).

Because of these reasons, we consider it necessary to do a fair comparison of these pitch-based techniques, trying to show the equivalences between some of them and trying to see the limits of pitch-based recognition. Apart from this, we will propose three new pitch-based techniques but without paying attention to the pitch extractor because this is beyond the scope of this Thesis.

## B.1.2. Objectives

Taking into account the previous motivations, the main objectives of the Thesis can be summarized as follows:

1. Recognition of monaural speech which is emitted in its normal way (i.e. with pitch) and contaminated with acoustic noise.

2. Development of a comparative study of both classical and pitch-based robust speech recognition techniques considered as the state of the art.

3. Development and improvement of robust ASR techniques based on pitch, trying to do minimal assumptions about the noise. In order to do so, we will employ other techniques and recognition schemes such as SS (Spectral Subtraction) or MD (Missing Data).

4. We will show the equivalences between some of the different techniques, doing a fair comparison and trying to answer the question of to what extent recognition can be made more robust by means of the pitch.

## B.2. Principles of Automatic Speech Recognition

The first chapters are devoted to explaining some important concepts which will be used throughout the Thesis. These concepts refer to: speech, hearing, signal processing, acoustic representations (cochleagram, spectrogram and cepstrogram) and their masks, pitch extractors, and MD (Missing Data) recognizer based on HMM (Hidden Markov Models).

The most important issues described in these chapters are:

- The «main source model» of speech which considers that speech is a main source which is intensity and spectrally modulated and sometimes replaced by short duration noises (unvoiced sounds). The main source can be a noise in the case of whispered speech, but in a normal situation speech will be identified with a voiced sound and, if pitch is known, the rest of the elements of the speech can be also located (unvoiced sounds and silences) as well. This model is a simplified definition of speech which will be considered to develop a VAD.

- The soft mask of a given time-frequency signal representation (i.e. spectrogram or cochleagram) can be estimated through local SNR estimates or through harmonicity (in the case of voiced frame with pitch $p(t)$) by means of a sigmoid function. The local SNR and the harmonicity can be estimated by means of a noise estimate $M_{\hat{N}}(f,t)$ and a correlogram $A_y(f,t,p(t))$ as follows:

$$SNR(f,t) = 20log_{10}\frac{M_Y(f,t) - M_{\hat{N}}(f,t)}{M_{\hat{N}}(f,t)} \quad \text{(B.1)}$$

$$H(f,t) = A_y(f,t,p(t))/A_y(f,t,0) \quad \text{(B.2)}$$

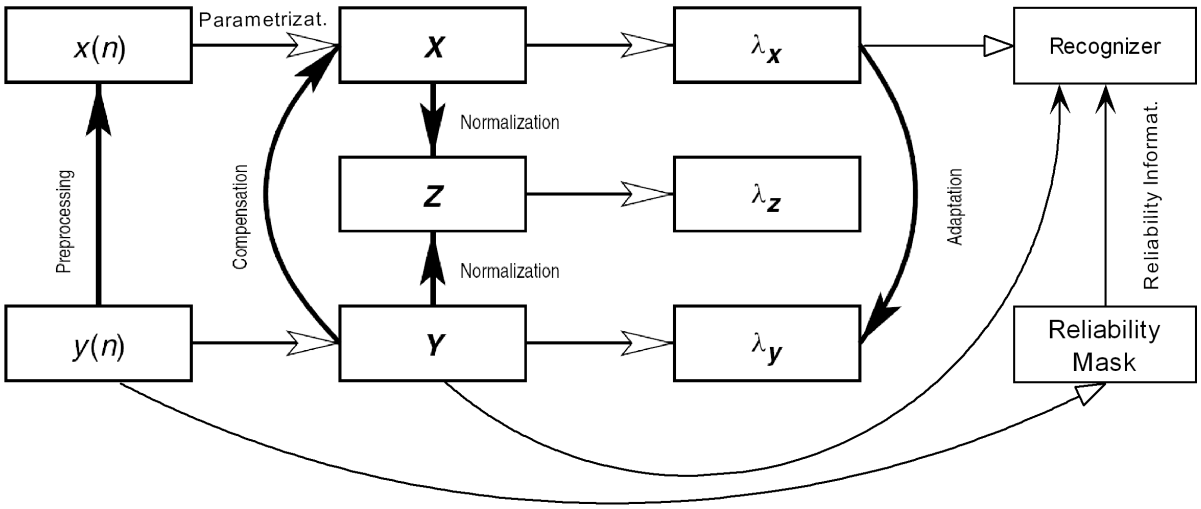



Figura B.1: ([121] adapted) A possible classification of different conventional robust ASR techniques.

## B.3. Conventional and pitch-based robust techniques

### B.3.1. Conventional robust techniques

Conventional robust ASR techniques can be outlined with the scheme of Fig. B.1 as follows:

**Preprocessing**: the noisy signal is cleaned or modified in temporal domain. We can mention offset and pre-emphasis in the ETSI front end [149], windows such as Hamming, SWP [92] and the variants of enhanced Wiener filter (such as in AFE [147]).

**Parametrization**: when a suitable acoustic representation is chosen that is robust to the speech and noise variabilities.

**Compensation**: the noisy features are modified to obtain an estimate of clean ones. We can mention MMSE techniques such as SPLICE [36] and VQ-MMSE Compensation [51], and the variants of SS (Spectral Subtraction) to avoid musical noise [40, 10, 73].

**Normalization**: when both clean and noisy representations are transformed so that the resulting features are less sensitive to noise. We can mention HEQ [34], CMN (Cepstral Mean Normalization) [108] and CTN [146].

**Model adaptation**: when clean models are modified to reduce the mismatch between training and testing conditions. We can mention PMC [47] and MLLR [79] .

**Reliability processing**: when the reliability of the noisy features is considered for recognition. We can mention WVA [11], Soft-Data [121], Multistream Recognition [15],

MD (Missing Data) [27] and SFD [5].

When comparing these conventional techniques, the following conclusion can be made: Only MD technique (and its extension SFD) tends to imitate human hearing. MD does not need (for example, compared to SS) to estimate perfectly the clean or noise signals. It only needs to know the reliability mask, i. e. where speech dominates noise in the acoustic representation and vice versa. However, this technique has the default of transferring the problem to the mask estimator.

### B.3.2. Robust pitch-based techniques

A bibliographic study of the pitch-based robust techniques, leads us to make the next classification:

**Exploitation of harmonic structure based techniques**: They do not use a pitch directly, but only some properties which derive from periodicity. We can especially mention HASE (High-lag Autocorrelation Spectrum Estimation) [142] which multiplies the high coefficients of the noisy OSA (One Side Autocorrelation) by a DDR (Double Dynamic Range) window to estimate the clean spectrum. The first 15 coefficients of the OSA are rejected because they are expected to be very contaminated by white-like noise (not correlated noise). It is also exploited the fact that in a voiced frame, spectral envelope information (short-term information) is preserved at high lags because of periodic repetitions. HASE is suitable for voiced sounds and silences, but it produces a loss of information for unvoiced frames. In order to avoid any possible mismatches, HASE is applied in both training and test. Some of our proposed techniques employ many of the HASE ideas. Another technique which exploits harmonic structure is HF [129].

**Clean estimation techniques**: They employ pitch extraction either to clean the signal (by means of some kind of comb filtering) or to estimate noise (with a tunnelling comb filtering) and compensate the noisy signal. As an example of the first case, WHNM ([138]) can be mentioned. An example of the second case is HT (Harmonic Tunnelling) [38]. This technique first finds the most energetic peaks of the spectrogram related to the pitch. Pitch extraction is carried out together with this peak search. An algorithm searches for the limits of the tunnelling regions which are expected to be dominated by the noise. Then, a noise spectrum estimate can be obtained by interpolating between these regions. This estimate is used in SS to obtain a clean spectral estimate. This technique has the drawback of not taking into account unvoiced frames. Another tunnelling comb techniques are FPM-NE [19] and the Frazier technique [46] which employ filters with

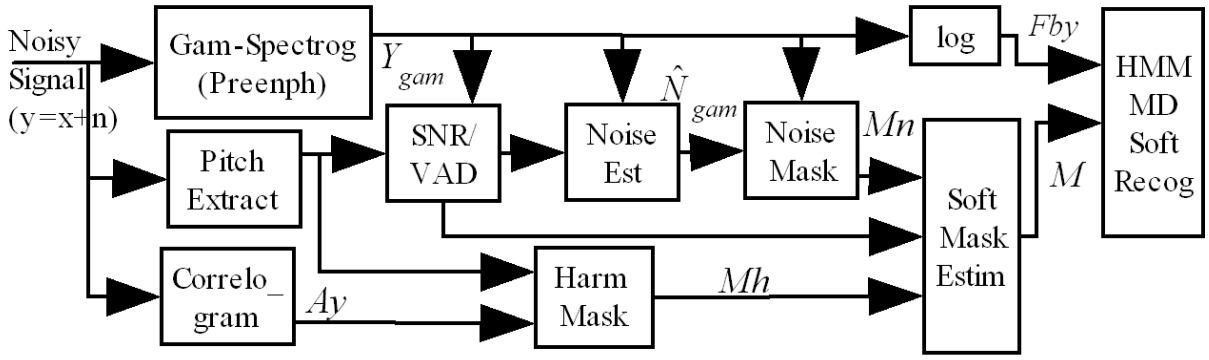



Figura B.2: Adapted recognition system of Barker technique [6] to compare with one of our proposed techniques. Two masks are estimated, $M_n$ based on VAD noise estimation and $M_h$ based on the harmonicity of the correlogram. The final mask $M$ is a combination of both masks.

impulse responses of the type of $h_T(t) = \delta(t) - \delta(t - T)$. Two of our proposed techniques are based on variants of these kind of comb filters.

**Mask estimation techniques**: They also employ pitch extraction to obtain a reliability mask for the considered time-frequency representation (spectrogram or cochleagram). We can especially mention the technique due to Barker [9, 6]. This technique estimates two masks, a noise soft mask $M_n$ based on the local SNR for every time-frequency pixel estimated by means of a ten-first-frame noise estimate (Sec. B.2), and a harmonicity soft mask $M_h$ (based on the harmonicity of each pixel estimated by means of the noisy correlogram and the pitch, Sec. B.2). The final mask is a linear combination of both masks. Fig. B.2 depicts an adaptation of the Barker technique which will be compared with one of our proposed techniques. Other mask techniques have been proposed by Brown [18] and Ma [90]. This last one is based on SFD (Speech Fragment Decoding [5]) to extract the pitch and the mask of a target speaker when the noise is another speaker.

Doing a fair comparison of above pitch-based techniques is a difficult task as we commented in the introduction (Sec. B.1.1). Sec. B.5 is devoted to do it. In addition to these difficulties, pitch-based techniques have others lacks:

- They do not deal with all kind of noises. For example, HASE fails with harmonic noises.
- They do not take into account unvoiced frames. For example, HT may take unvoiced frames as noise.

- They need a fine pitch estimate. For example in the case of comb filtering techniques to estimate clean signal, the spectral harmonics are not exactly located at pitch positions because of quasi-periodicity. Tunnelling comb filtering techniques to estimate the noise do not have this problem because there is «more-space» around tunnelling regions.
- In the case of proposing a pitch extractor, they involve an inaccurate pitch estimate. For example, this is the case of HT.
- They can be complex and not biomimetic. It can be observed that the more biomimetic a technique is the more efficient it is. Ma technique inspired on ASA (Auditory Scene Analysis) does not have this problem but the FPM-SE [19] does.

# B.4. Proposed techniques

## B.4.1. Asymmetric windows

### Introduction

The asymmetric windows technique is explained in detail in a paper accepted with minor changes [107]. This technique tries to do robust ASR with low computational cost. It is inspired by the HASE technique [142] (Sec. B.3.2), which can be interpreted as an asymmetric weighting (or windowing) of the autocorrelation coefficients of the OSA (One Side Autocorrelation). The windowed OSA is employed to obtain a clean spectral estimate and its AMFCC (Autocorrelation Mel-Frequency-Cepstral-Coeficients). Another related techniques are Cyclic-Spectrum [113], OSALPC [60], SMC [93] and LSMYWE [94] which are based on employing high-lag autocorrelation coefficients to estimate the spectrum since these coefficients are usually less contaminated by noise (Sec. B.3.2). Another related technique which also employs asymmetric windows is that of [131], although these windows are applied in the time domain. We will only compare our asymmetric windows with HASE because HASE surpasses the other related techniques.

### Recognition system

Fig. B.3 shows the proposed ASR system to evaluate our asymmetric windows. Its front end uses very similar parameters to the ETSI FE [149]: 23 Log-Mel channels, 13-statics $(C_0, ..., C_12)$ + 13-velocity + 13-acceleration cepstral coefficients, etc.. It takes a

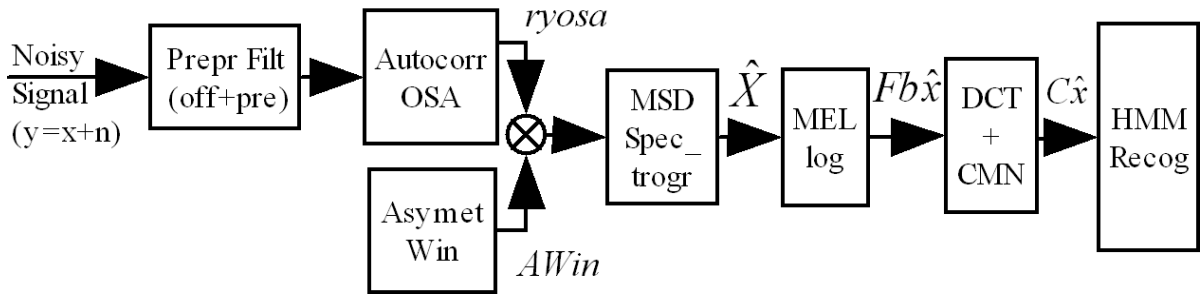



Figura B.3: ASR system based on OSA autocorrelation with the asymmetric windows.

noisy signal $y$, filters offset and enhances high frequencies, obtains the OSA of every frame and multiplies it by an asymmetric window, obtains a clean estimate of MSD (Magnitude Spectral Density) $\hat{X}$, the Log-Mel spectrum $Fb_{\hat{x}}$ and the AMFCC ($C_{\hat{x}}$). CMN (Cepstral Mean Normalization) is applied to each AMFCC and the resulting AMFCC vector is submitted to an HMM (Hidden Markov Model) recognizer. The parameters of recognizer are those of the Aurora-2 framework [120] (3 Gaussians per state, etc.). The proposed asymmetric windows are applied to both training and test in order to avoid any mismatch.

### Proposed asymmetric windows

The set of proposed asymmetric windows noted as $DDR_{c,w}$ depends on two parameters: $c$ and $w$ (center and width in number of samples). This set is:

$$DDR_{c,w}(k) = \begin{cases} DDR_w(\frac{w}{2} - (c+1) + k) & c - \frac{w}{2} < k \leq c + \frac{w}{2} \\ 0 & otherwise \end{cases} \quad (k = \{0, ..., L-1\}) \tag{B.3}$$

where $DDR_w$ is a Double Dynamic Range Hamming window [142] and $L$ is the total window length (in number of samples) (which corresponds to OSA length). Fig. B.4 shows an example of a $DDR_{50,250}$ applied to the OSA of a voiced frame with pitch 50 samples.

An interesting feature of the proposed windows is that they allow a variable contribution of the first autocorrelation coefficients (without discarding them completely as HASE does). Also it applies more weight to the most important coefficients by centering the window on them. Our hypothesis is that the most important coefficients for robust speech recognition are those around the pitch (or its multiples) lags because they are more energetic and less affected by the noise. In addition, they also carry spectrum enve-

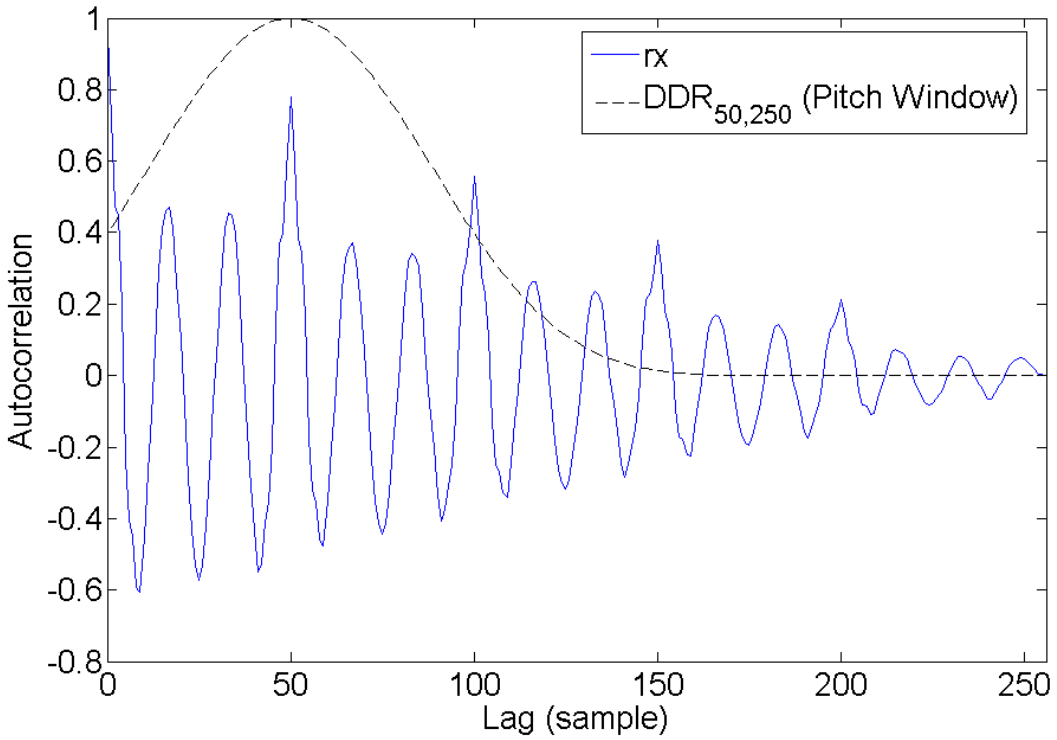



Figura B.4: Example of a $DDR_{50,250}$ window applied to the OSA of a voiced frame with a pitch value of 50 samples.

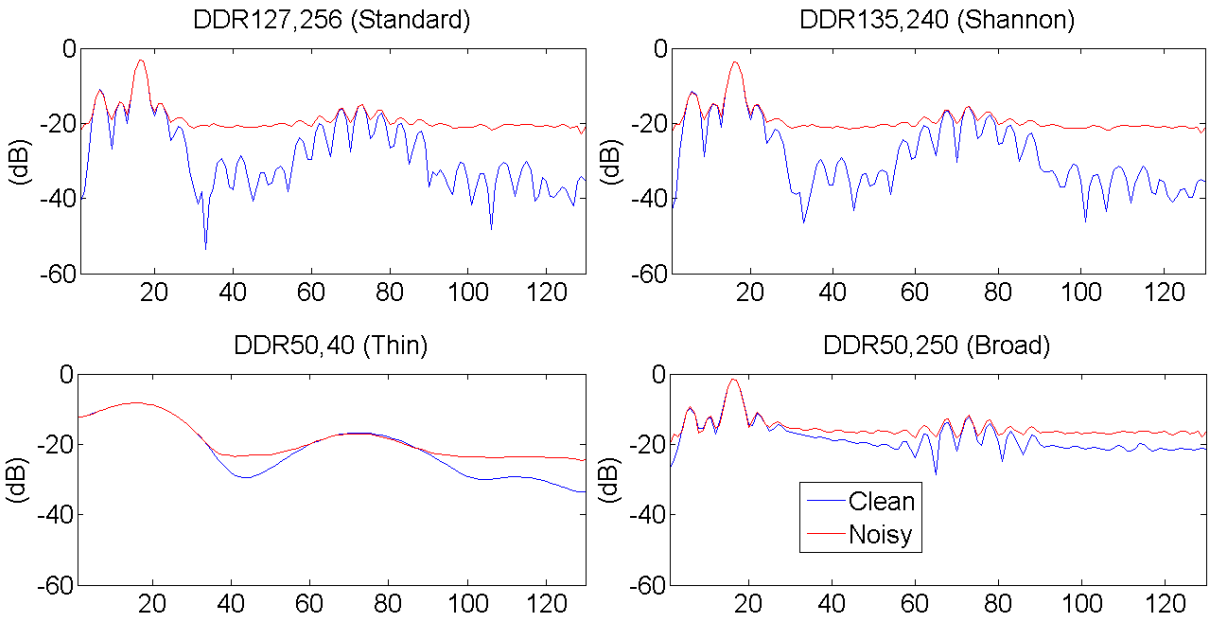



Figura B.5: Averaged spectra of four different windows applied to a vocal with pitch=50 samples contaminated with white noise.

lope information. In Fig. B.4 the asymmetric window is centered over the first pitch (lag 50). It must be taken into account that the HASE Shannon window is equivalent to our $DDR_{135,240}$.

### Spectral analysis of the windows and application to unvoiced frames

Fig. B.5 shows the clean and noisy (contaminated with white noise) spectrum of a voiced frame for four different $DDR_{c,w}$ windows. We can conclude that $DDR_{50,40}$ and $DDR_{50,250}$ have very short dynamic range (i.e. the window has not enough spectral range to cover the 80 dB necessary for speech). In spite of its short dynamic range, $DDR_{50,250}$ is quite similar to the best window for Aurora-2 that will be later obtained.

In order to avoid non homogeneous signal analysis, the same window will be applied to all types of frames (voiced, unvoiced and silence). For voiced sounds and silences, it is clear that this is always beneficial. For unvoiced it could be thought that, since lower lag coefficients (which exclusively carry the spectral envelope information) are deleted or little weighted, the use of a constant window could be harmful.

The experimental results will show that the above mentioned problems do not have effect over the system performance. In order to understand this, it is important to notice

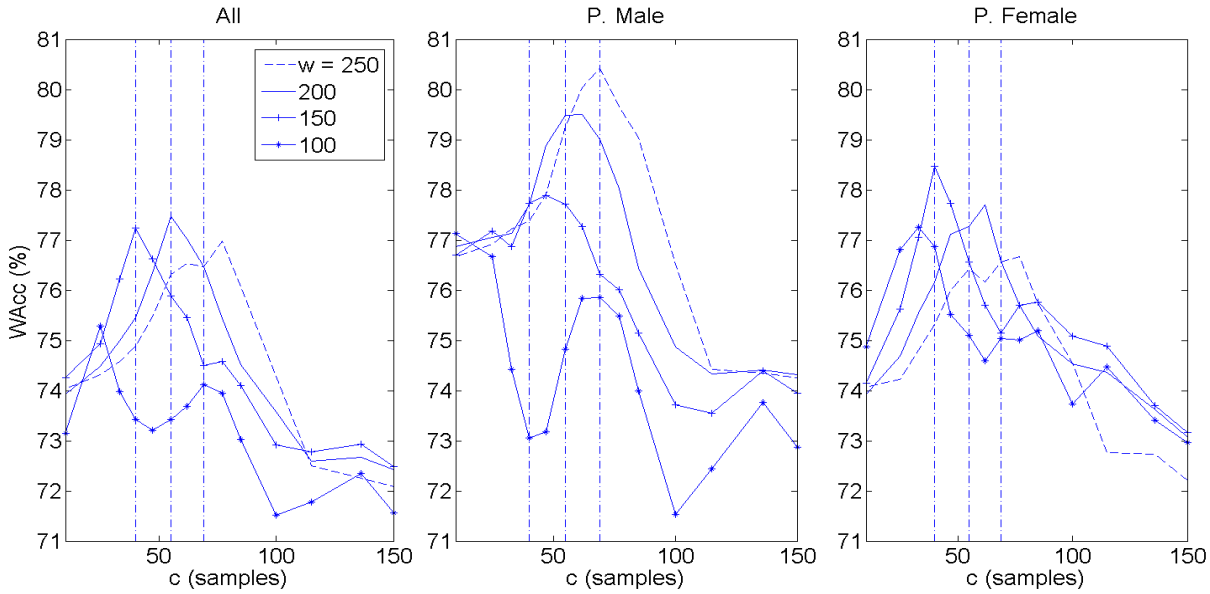



Figura B.6: WAcc ( %) for the whole Aurora-2 (0-20 dB) when all, male pitch and female pitch utterances are employed in training-test stages, againts $c$ (center) and $w$ (width of window). The three vertical lines correspond to the female, mean and male pitches (40, 55 and 69 samples).

that the same asymmetric window is applied in both training and testing.

### Experimental results

In order to confirm the hypothesis that the most important OSA coefficients for robust speech recognition are the pitch lag (or its multiples), a gender-dependent recognition experiment has been carried out:

Taking into account that the histogram of the average pitch per sentence (in Aurora-2 Set A) shows a mean pitch of 55 samples and two different modes for male and female speakers with pitch values at 69 and 40 samples, respectively, training and test utterances of the whole Aurora-2 (Aurora-2 Set A, B, C and clean training) are separated into three groups. These groups are: *All* (without separation depending on pitch), *P. Male* (with pitch greater than 55 samples) and *P. Female* (with pitch lower than 55 samples). A search (applying the same window in both, training and testing) for the the best window of each group is carried out by changing $c$ and $w$. The WAcc (Word Accuracy in %) average (0-20 dB) results are depicted in Fig. B.6.

It can be observed that the best windows for *All*, *P. Male* and *P. Female* groups are $DDR_{55,200}$ with 77.47 %, $DDR_{69,250}$ with 80.43 % and $DDR_{40,150}$ with 78.47 % respectively.

## B. SUMMARY OF THE THESIS: PITCH-BASED ROBUST SPEECH RECOGNITION TECHNIQUES

| Window | Clean | 20 dB | 15 dB | 10 dB | 5 dB | 0 dB | -5 dB | Mean (20-0 dB) |
|---|---|---|---|---|---|---|---|---|
| Hamming (FE) | 99.14 | 97.21 | 92.57 | 76.72 | 44.28 | 22.99 | 13.00 | $66{,}76 \pm 0{,}80$ |
| $DDR_{135,240}$ (HASE) | **99.15** | **97.47** | **94.37** | 84.26 | 58.35 | 27.69 | 14.72 | $72{,}43 \pm 0{,}76$ |
| $DDR_{55,200}$ (Mean Pitch) | 98.85 | 96.12 | 93.21 | **85.91** | **70.00** | **42.09** | **18.07** | $\mathbf{77{,}47 \pm 0{,}71}$ |

Tabla B.1: WAcc (Word Accuracies %) results obtained by different windows tested with Aurora-2 (Set A, B and C) for diferent SNR values.

From these results the following conclusions can be extracted:

1. For the whole Aurora-2 our proposed $DDR_{55,200}$ window with 77.47 % gives better results than the HASE window ($DDR_{135,240}$) with only 72.43 %.

2. The optimum window centers of each group just coincide with the mean pitch of each group: 55, 62 and 40 (are indicated with dashes vertical lines in the figure). This confirms our hypothesis that the most important coefficients are those around the pitch (or its multiple) values.

Tab. B.1 shows the results obtained by the different windows tested for Aurora-2 (Set A, B and C) for different SNR values. Sec. A.3 explains how the confidence intervals of the mean results are obtained. These intervals show that our results are reliable and will be only shown here and in the next table in order to avoid overloading the rest of the tables. It can be concluded that $DDR_{55,200}$ obtains better results than Hamming (very similar to ETSI FE [149]) and HASE. It can also be concluded that both the short dynamic range of the proposed windows and its application to unvoiced frames are not very harmful in clean conditions as results show.

Tab. B.2 shows the results obtained by the different windows applied to Aurora-3 Spanish (real noise) [4]. WM, MM and HM mean well, medium and high mismatch, respectively. It can be concluded that the proposed window surpasses HASE results mainly at high mismatch which is the worst condition.

| Window | WM | MM | HM | Mean |
|---|---|---|---|---|
| Hamming (FE) | 89.08 | 82.15 | 64.51 | $78{,}58 \pm 0{,}64$ |
| $DDR_{135,240}$ (HASE) | 89.76 | **83.16** | 76.39 | $83{,}10 \pm 0{,}58$ |
| $DDR_{55,200}$ (Mean pitch) | **89.85** | 82.87 | **80.15** | $\mathbf{84{,}29 \pm 0{,}57}$ |

Tabla B.2: WAcc results obtained by the different windows applied to Aurora-3 Spanish (real noise). WM, MM and HM mean well, medium and high mismatch, respectively.

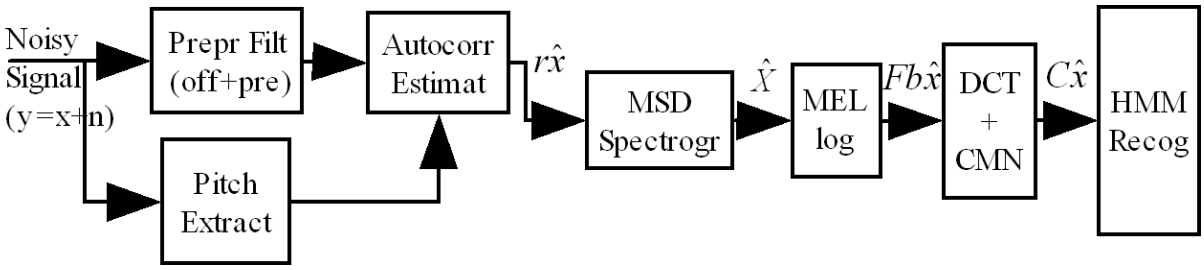



Figura B.7: Recognition system based on the use of pitch-based clean autocorrelation estimates.

## B.4.2. Averaging and sifting autocorrelation

### Introduction

Averaging and sifting autocorrelation estimators are explained in detail in [106]. These techniques try to estimate the clean autocorrelation of every frame by employing its pitch value. The resulting estimates are employed to obtain AMFCC features.

The averaging estimator is very related to techniques which can be reduced to a comb filter (i. e. sampling noisy spectrum at pitch harmonics). These kind of techniques are those of Kuroiwa [77], WHNM [138], etc. It is also very related to HASE [142] in the sense of supposing that the noise usually is concentrated in the first autocorrelations coefficients. We will compare our proposals with HASE.

### Recognition system

Fig. B.7 shows the proposed ASR system to evaluate different AMFCC techniques. It is very similar to that employed to evaluate asymmetric windows B.4.1. A pitch extractor is needed to estimate the clean autocorrelation and instead of windowing the OSA, the

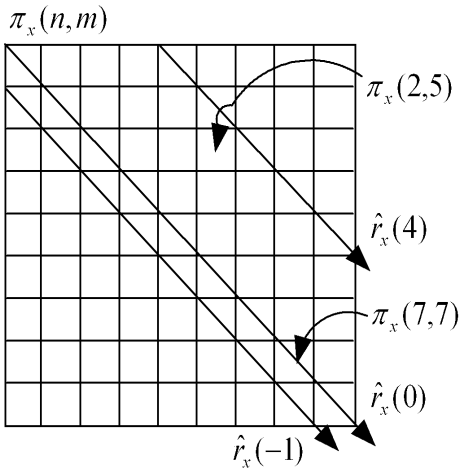



Figura B.8: Product table for a frame $x(n)$ with 9 samples. Some products are illustrated and the diagonal arrows indicate the elements which have to be summed in order to obtain the different autocorrelation coefficients.

whole (negative and positive side) the autocorrelation is employed to obtain the MSD. The window applied to this autocorrelation will be the DDR.

The pitch extractor employed here and in the following will be that presented in [106]. This pitch extractor takes the pitch provided by the ETSI xFE pitch extractor [148] and applies a smoothing processing. This smoothing is needed because the pitch provided by xFE has many errors at lows SNRs.

### Product table and biased autocorrelation

The biased autocorrelation of a segment $x(n)$ is defined as,

$$\hat{r}_x(k) = \frac{1}{N}\sum_{n=k}^{N-1} x(n)x(n-k) \quad (0 \leq k < N) \tag{B.4}$$

It can be reformulated by means of a «product table» $\pi_x(n,m) = x(n)x(m)$, $(n,m = 0,\ldots,N-1)$ (Ec. B.5).

$$\hat{r}_x(k) = \frac{1}{N}\sum_{n=k}^{N-1} \pi_x(n,n-k) \quad (k = 0,\ldots,N-1) \tag{B.5}$$

We see that the biased coefficients can be obtained by summing diagonals of the table. Fig. B.8 shows an example of it for a frame $x(n)$ with 9 samples. This table formulation

will be useful later to better understand the proposed autocorrelation estimators.

Let's suppose now that we have a noisy signal $x(n) = p(n) + d(n)$ which is the sum of a perfect periodic clean signal $p(n)$ (which approximately represents the voiced signal) and a distortion $d(n)$ (which accounts for non-periodic components and, mainly, additive acoustic noise). If we are interested in estimating the clean periodic autocorrelation $r_p(k)$ from the noisy signal, it can be easily demonstrated that the biased estimator is not suitable because its expected value is:

$$E[\hat{r}_x(k)] = w_B^N(k)\,(r_p(k) + r_d(k)) \tag{B.6}$$

where $w_B^N$ is a Barlett window of length $N$. This estimator is not robust because its error is equal to $r_d(k)$. Fig. B.9 shows how far the noisy biased estimate is from the clean biased estimate in both, autocorrelation and spectrum domain. This illustrates the need for finding a better autocorrelation estimator.

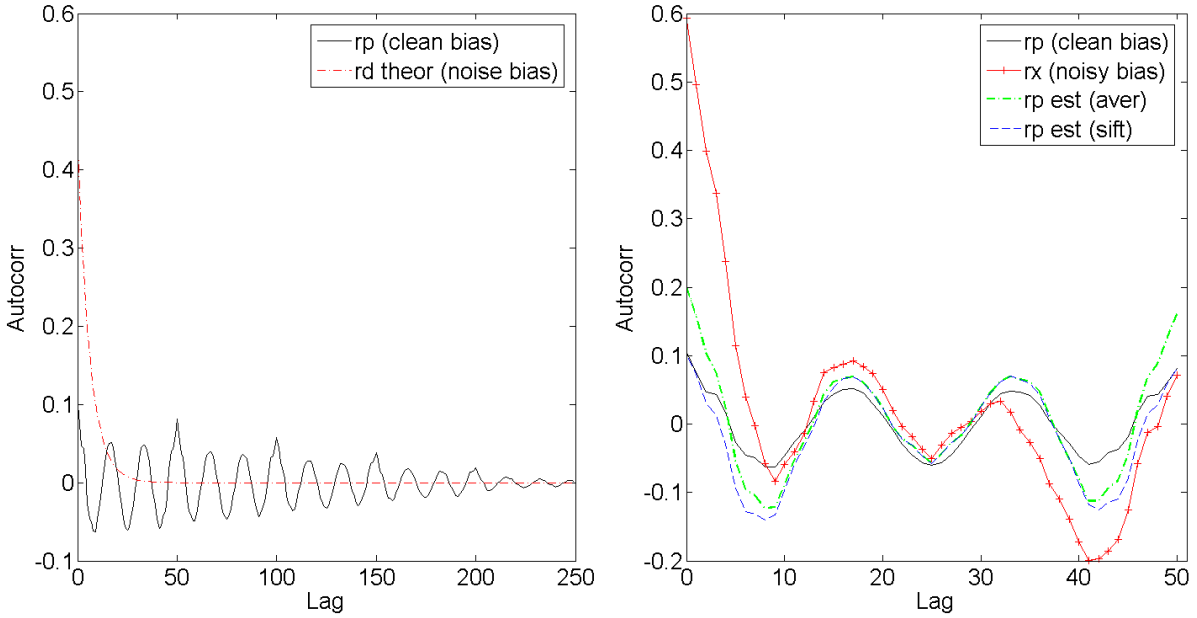
### Averaging autocorrelation

It must be noticed that if the distortion $d(n)$ was null the table would be perfect periodic and many products would be repeated. On the left of Fig. B.10 the repeated products are marked with $X$ for a 9-sample signal with period $T = 3$ samples. Taking this into account an estimate of the clean table can be obtained by averaging the repeated products as follows:

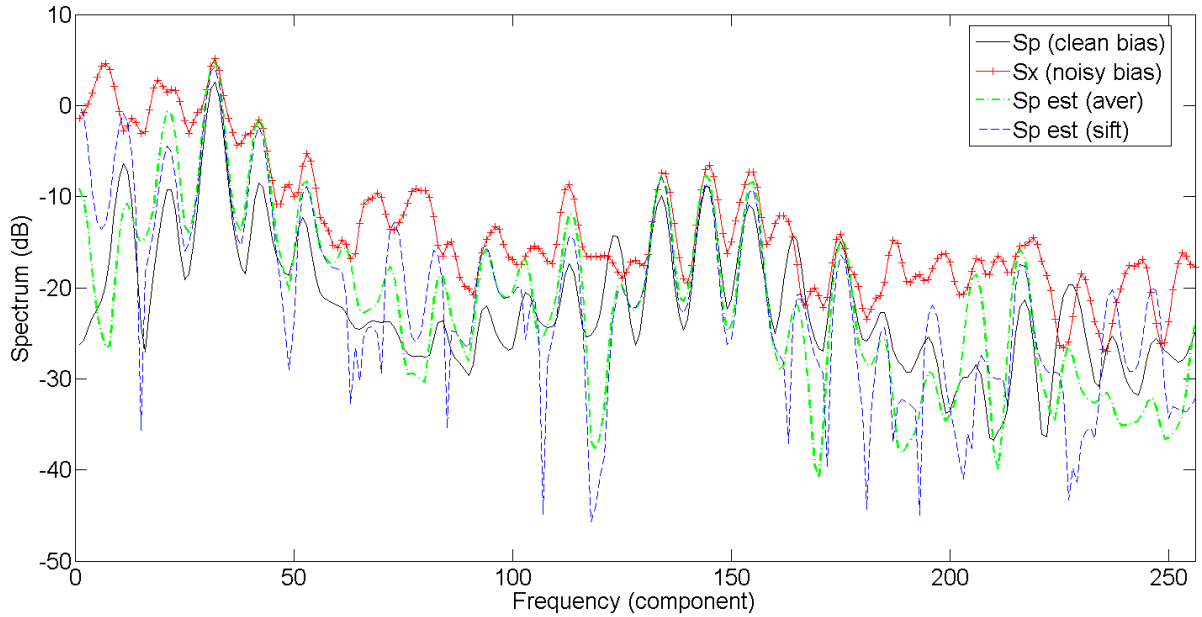
$$\pi_p(n,m) \approx \bar{\pi}_x(n,m) = \frac{1}{N_p^2} \sum_{i=0}^{N_p-1} \sum_{j=0}^{N_p-1} \pi_x(iT + \underline{n}, jT + \underline{m}) \tag{B.7}$$

where, for the sake of simplicity, it is supposed that there is an integer number of periods ($N = N_p{*}T$), $\underline{\mathrm{n}}$ is the remainder of $n/T$, and each averaging product $\bar{\pi}_x(n,m)$ is estimated using the idea that each clean product $\pi_p(n,m)$ is affected by a mean zero error. Fig. B.10 shows an example of how to obtain these products. Finally, the proposed averaging autocorrelation estimator of the periodic clean signal is:

$$r_p(k) \approx \bar{r}_x(k) = \frac{1}{N} \sum_{n=k}^{N-1} \bar{\pi}_x(n, n-k) \tag{B.8}$$




(a) Left, biased autocorrelation of the clean signal ($r_p$) and true AR noise autocorrelation ($r_d$ theor) employed to contaminate it. Right, clean biased, noisy biased, averaging and sifting ($\delta = 16$) autocorrelations.




(b) Spectrums derived from clean, averaging and sifting autocorrelations.

Figura B.9: Top, Comparison of the proposed autocorrelations for a vowel with *pitch* = 50 samples contaminated with an AR noise. Bottom, the corresponding spectra.

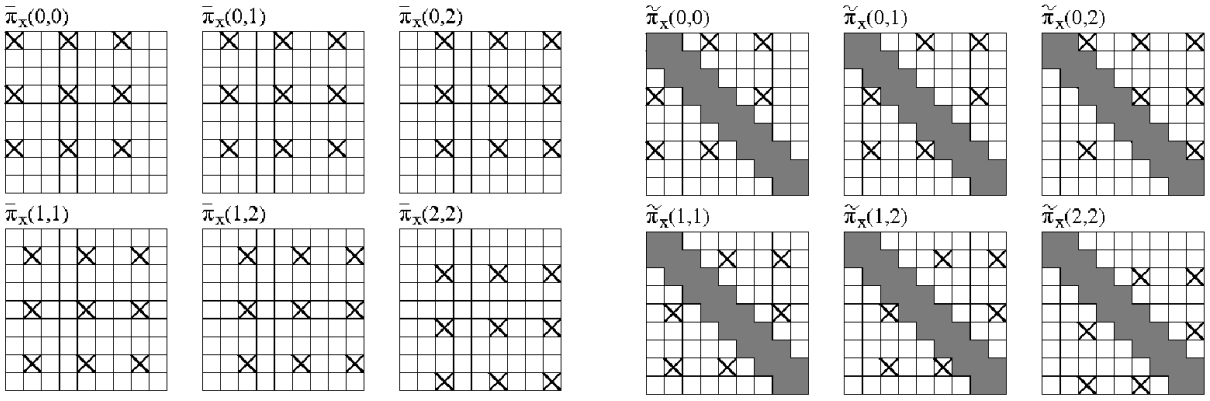



Figura B.10: Product tables $\pi_x(n,m)$ (12 times repeated) of a $x(n)$ signal with $N = 9$ and period $T = 3$ samples. Left, computation of the different products $\bar{\pi}_x(n,m)$ for the averaging autocorrelation. Right, computation of the different products $\tilde{\pi}_x(n,m)$ for the sifting autoc. with $\delta = 2$.

It can be demonstrated that its expected value is:

$$E[\bar{r}_x(k)] = w_B^N(k)\left(r_p(k) + \frac{N_1(k)\bar{s}_d(\underline{k}) + N_2(k)\bar{s}_d(\underline{k}-T)}{N-k}\right) \tag{B.9}$$

where $\bar{s}_d(\underline{k})$ depends on $r_d(k)$ [106]. This estimator is better than the biased one because the additive error term is lower than the whole autocorrelation distortion $r_d(k)$. In particular, it can be shown that the SNR can be increased up to a factor equal to the number of available periods $N_p$. Fig. B.9 shows that this estimate is closer to the clean biased autocorrelation than the biased estimate from noisy signal.

One important issue of the averaging estimation is that it can also be shown that it is equivalent to a sort of comb filtering. Then, this estimator has the advantage (with respect to the biased one) of removing the noise between the gaps or tunnels placed at the middle regions of the pitch spectrum harmonics, although it does not remove noise placed at harmonics.

### Sifting autocorrelation

Averaging estimation can be improved taking into account the HASE idea that white-like noise mainly affects to the lower lag autocorrelation coefficients. The corresponding products of these coefficients (a $\delta$ interval around the main diagonal) can be rejected or

sifted to obtain a better estimate of the clean table as follows:

$$\pi_p(n,m) \approx \tilde{\pi}_x(n,m) = \frac{1}{N_\delta(\underline{n},\underline{m})} \sum_{(i,j)\in S_\delta(\underline{n},\underline{m})} \pi_x(iT+\underline{n}, jT+\underline{m}) \tag{B.10}$$

where $\delta$ is the so-called «sifting interval» and $N_\delta(\underline{n},\underline{m})$ is the number of pairs $i,j = 0,\ldots,N_p-1$ which belong to the set $S_\delta(\underline{n},\underline{m})$ (which contains the surviving index pairs). Fig. B.10 shows how to obtain the different sifting products $\tilde{\pi}_x(n,m)$ for a $\delta = 2$.

The proposed sifting autocorrelation estimate can be obtained as:

$$r_p(k) \approx \tilde{r}_x(k) = \frac{1}{N}\sum_{n=k}^{N-1} \tilde{\pi}_x(n,n-k) \quad (k=0,\ldots,N-1) \tag{B.11}$$

It can be shown that its expected value is that of Ec. B.9 but replacing $\bar{s}_d(\underline{k})$ by its sifted version $\tilde{s}_d(\underline{k})$ (see [106]). It can also be shown that if the noise autocorrelation is fully contained inside the sifting interval, then this estimation gives exactly the biased autocorrelation of the periodic clean signal $\hat{r}_p(k)$. Also it can be seen that sifting is the same as averaging in the interval $\delta \le k \le T-\delta$ and that sifting removes more noise than averaging in the $0 \le k < \delta$ and $T-\delta \le k < T$ intervals [106]. These intervals are just representative of the important information for ASR, i. e. the spectral envelope. Also, it can be easily seen that sifting with $\delta = 0$ becames the averaging estimator. Fig. B.9 shows how sifting is closer to clean than averaging and that they coincide in the $\delta \le k \le T-\delta$ interval.

The important thing about the proposed estimator is that it has the advantages of the averaging (removing noise between the tunnels) plus those of the HASE technique (removing white-like noises).

### Extension of sifting to silence and unvoiced frames

Sifting has been developed to estimate the clean speech autocorrelation on voiced frames. In order to avoid the use of a VAD (Voice Activity Detector) and a different estimator in silence and unvoiced frames, it will be supposed that they have a fictitious pitch of 55 samples which corresponds to the average human pitch (preliminary experiments showed that this is not a critical parameter of the system). In silence frames, the application of sifting is clearly suitable, but for unvoiced frames we could reasonably argue that it is not helpful but even harmful.

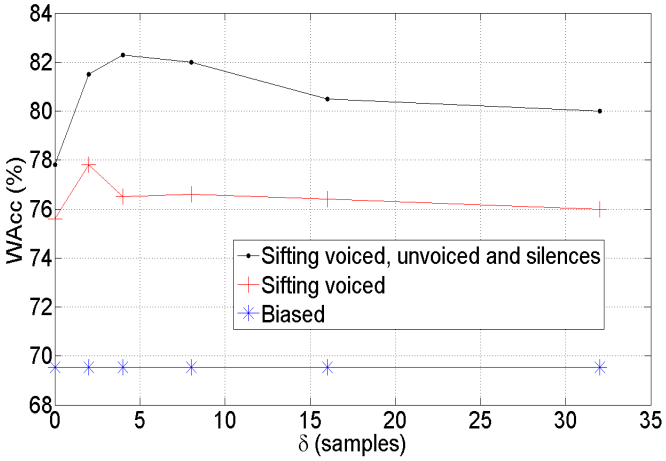



Figura B.11: WAcc of Set-A versus the sifting interval $\delta$ when the biased autocorrelation is used for all frames ($*$), when sifting is only applied to voiced ($+$) and when sifting autocorrelation is applied to all frames $\bullet$ (voiced, unvoiced and silence).

.

However, and due to similar reasons as those employed for asymmetric windows B.4.1 the experimental results will show that this approach (the extension of sifting to types of frames) is suitable.

### Experimental results I: suitable sifting interval

Now, we will search for a suitable $\delta$ interval. Fig. B.11 shows the WAcc (20-0 dB) results obtained for Aurora-2 Set-A versus the sifting interval for three cases: biased autocorrelation applied to all frames, sifting applied only to voiced frames and sifting applied to all (voiced, unvoiced and silence) frames. The following conclusions can be drawn:

- The sifting estimator obtains better results than the biased and the averaging ($\delta = 0$) estimators.
- It is better to apply sifting to all kind of frames than only to voiced frames. This justifies the extension of sifting to silence and unvoiced frames.
- The optimum $\delta$ is 8 samples. This value is both, large enough to reject enough contaminated products and small enough to avoid rejecting much speech information.

In what follows, $\delta = 8$ will be taken as our optimum sifting interval.

| Technique | Clean | 20 dB | 15 dB | 10 dB | 5 dB | 0 dB | -5 dB | Mean (20-0 dB) |
|---|---|---|---|---|---|---|---|---|
| A. Bias (FE) | 99.06 | 97.65 | 94.74 | 84.06 | 55.30 | 26.53 | 13.63 | 71.65 |
| HASE ($\delta = 15$) | 99.15 | 97.47 | 94.37 | 84.26 | 58.35 | 27.69 | 14.72 | 72.43 |
| A. Aver ($\delta = 0$) | 99.36 | 97.99 | 95.85 | 89.98 | 72.36 | 36.55 | 12.94 | 78.55 |
| A. Sift ($\delta = 8$) | 98.63 | 96.69 | 94.50 | 89.39 | 76.30 | 44.60 | 14.75 | 80.30 |
| A. Sift Ideal ($\delta = 8$) | 98.63 | 97.06 | 95.48 | 91.84 | 82.52 | 61.00 | 29.93 | 85.58 |
| AFE | 99.11 | 97.72 | 96.05 | 91.84 | 82.19 | 59.91 | 28.87 | 85.54 |

Tabla B.3: WAcc results obtained by the different techniques tested with Aurora-2 (Set A, B and C) for different SNR values.

| Technique | WM | MM | HM | Mean |
|---|---|---|---|---|
| A. Bias (FE) | 84.03 | 62.15 | 37.85 | 61.34 |
| HASE ($\delta = 15$) | 85.91 | 64.69 | 43.34 | 64.65 |
| A. Sift ($\delta = 8$) | 76.80 | 50.14 | 39.11 | 55.35 |
| A. Sift Ideal ($\delta = 8$) | 84.52 | 71.47 | 61.44 | 72.48 |

Tabla B.4: WAcc results obtained by different techniques tested with Aurora-3 Danish (real noise).

### Experimental results II: Aurora 2 and 3

Tab. B.3 shows the results for the different autocorrelation estimators, HASE and the ETSI AFE front-end [147] over Aurora-2. It can be observed that the application of sifting to unvoiced frames is not very harmful as clean results show. In general, sifting surpasses all except AFE results because this is a more sophisticated front-end which brings together different robust techniques. Sifting with ideal pitch (i. e. pitch extracted from the corresponding clean signal) could perform as well as AFE as shows in the *A. Sift Ideal* row.

Tab. B.4 shows the results obtained over the real noise database Aurora-3 (Danish). It can be observed that sifting would require a better pitch extractor to improve the HASE results. In this case, sifting could surpass HASE in more than 18 % of WAcc (*A. Sift Ideal* experiment).

### Experimental results III: dynamic sifting

Tab. B.5 shows the WAcc over Aurora-2 depending on the type of noise. It is observed that sifting surpasses averaging for all noises except for *Restaurant* and *Airport*. There

| Technique | Set A | | | | Set B | | | | Set C | | Mean (20-0 dB) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Subw | Babb | Car | Exhi | Rest | Stre | Airp | Trai | Subw MIRS | Stre MIRS | |
| A. Aver ($\delta = 0$) | 79.19 | 80.14 | 77.36 | 76.54 | 81.03 | 79.08 | 80.73 | 78.73 | 75.63 | 77.01 | 78.55 |
| A. Sift ($\delta = 8$) | 83.62 | 81.96 | 80.56 | 80.80 | 78.45 | 82.15 | 80.16 | 80.63 | 76.16 | 78.47 | 80.30 |
| A. Sift ($\delta = Ideal$) | 89.07 | 87.49 | 86.68 | 86.88 | 85.03 | 88.07 | 85.92 | 86.03 | 85.17 | 85.96 | 86.63 |

Tabla B.5: WAcc results obtained by the different techniques tested with Aurora-2 (Set A, B and C) for different SNR values.

are several reasons for this shortcoming such as errors in pitch extraction or a unsuitable $\delta$. Another experimental results have shown that with other $\delta$ values (not 8), this shortcoming with *Restaurant* and *Airport* can be sorted out.

This points out the need of applying sifting with a dynamic value for $\delta$(that is, a suitable value for each instant or utterance). *A. Sift* ($\delta = Ideal$) is an oracle experiment which selects the best $\delta$ for each utterance. It shows the limits of improving the results by means of a dynamic delta for each utterance. Thus, dynamic sifting is a possible future reasearch line.

## B.4.3. Pitch-based noise estimation

### Introduction

Our proposed pitch-based noise estimation technique is explained in detail in [105]. Noise estimation is an important issue in robust speech recognition and there exit many approaches to do it. If you want to perform this task, taking into account the spectral masking effect [155], the only way to do it is by interpolating noise from regions where it is known. VAD noise estimators [121] do this and are suitable for stationary noises. Other techniques, such as those which can be reduced to a comb filtering of noise, can be employed in order to obtain more regions of noise and to face non-stationary noises. **HT** (Harmonic Tunnelling) [38] is an example of these kind of comb techniques which require a pitch extractor. Here we propose a noise estimate which combines VAD estimates and a modification of HT noise estimates by means of the pitch extraction. In addition to the modifications applied to HT (such as avoiding overestimation and not including unvoiced

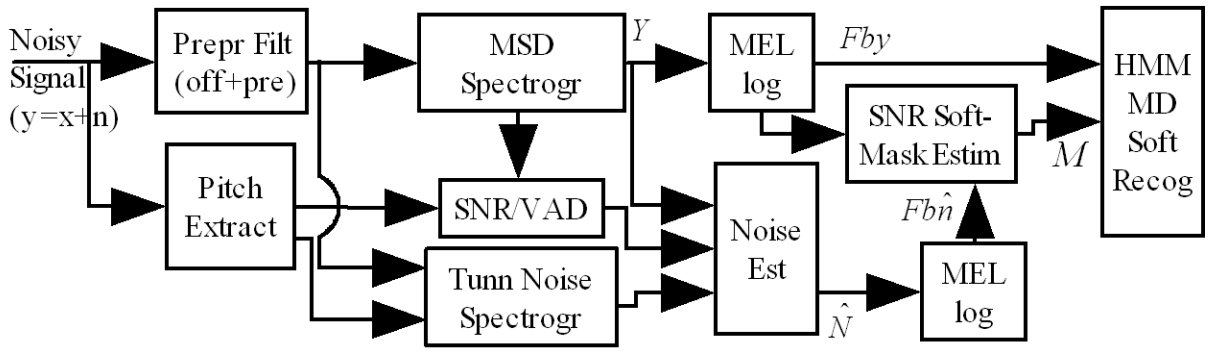



Figura B.12: Proposed recognition system to evaluate MD ASR from pitch-based noise estimation.

.

frames as noise) the important contribution of our proposal is that it fully exploits pitch information to perform robust ASR as we will see in Sec. B.5.

The proposed noise estimate will be evaluated on SS (Spectral Subtraction) and MD (Missing Data) [27]. It will be also compared with a VAD noise estimate and with an adaptation of the Barker's technique [6] which also employs MD and pitch.

### Recognition system

Fig. B.12 shows the proposed MD system to evaluate the proposed noise estimation in ASR. It is very similar to that employed for sifting B.4.2.

The *SNR* (global Signal to Noise Ratio estimator of the utterance) and *VAD* block take as inputs the noisy MSD (Magnitude Spectral Density) $Y$ and the pitch. The *Tunnelling Noise Spectrogram* block estimates the noise in voiced frames using a modification of the HT technique which makes use the of noisy signal and the pitch estimates. Our center block *Noise Estimator* takes $Y$, *SNR*, *VAD* and the tunnelling noise estimate to provide a spectrogram noise estimation $\hat{N}$. $Y$ and $\hat{N}$ are the inputs to the MEL filter bank and the log compresor (which yields $Fb_y$ and $Fb_{\hat{n}}$). These two last outputs are used to estimate an SNR of every frequency-time pixel and then the corresponding soft mask $M$. Finally, $M$ and $Fb_y$ are employed by the *MD Soft Recognizer* [7]. The parameters of the recognizer are those commonly employed over Aurora-2 for ASR with spectral features (9 Gaussians per state, [6]).

Now we will describe the most important blocks of the proposed system. Note that the different parameters were determined through preliminary experiments performed over a set of training (not testing) sentences of Aurora-2 contaminated with noise.

### VAD based on pitch

The proposed VAD is based on the «main source model» of speech (Sec. B.2) because once the pitch (main source) is located, the remaining speech sounds can be localized too.

Our VAD detects three different classes of frames: voiced, unvoiced and silences. Frames labeled as voiced correspond to frames where the pitch extractor gives a valid pitch. Unvoiced frames are searched in an interval of 20 frames before or after a sequence of voiced frames and identified when the instantaneous SNR of high frequencies is greater than 3 dB:

$$\hat{SNR}^{HF}(t_k) = 10 * log_{10}(E^{HF}_{\hat{X}}(t_k)/E^{HF}_{\hat{N}}(t_k))E_{\hat{N}}(t_k)) \tag{B.12}$$

$$\text{where } E_S(t_k) = \sum_{j=j_{1,8KHz}}^{j_{4KHz}} |S(\omega_j, t_k)|^2 \tag{B.13}$$

The reasons for this condition is that unvoiced sounds never occur in isolation and their energies are mainly between 1800 and 4000 Hz (sample frequency) [134]. The clean spectrogram $\hat{X}$ is estimated through the noise estimate $\hat{N}$ based on the 10 first-last noisy frames. Subsequent experiments have also shown that at low SNRs, this unvoiced estimation takes many noise frames as unvoiced. So when the estimate of the global SNR is less than 10dB, it is assumed that unvoiced signals are mixed with noise and no detection of unvoiced frames is carried out. This global SNR is estimated by means of $\hat{X}$ and $\hat{N}$.

Silence frames are those which have been classified neither as voiced nor unvoiced.

### VAD Noise Estimate

*NVAD* (VAD noise) is estimated by interpolating the noise from silence (noisy) frames. An averaging of the noisy MSD $Y$ of the closest 10 silence frames gives the estimate in each voiced or unvoiced frame.

### Harmonic Tunnelling Noise Estimate

The continuous MSD of a noisy signal $y(n)$ with $N$ samples at frequency $\omega$ is:

$$Y(\omega) = |\frac{\sum_{n=0}^{N-1} y(n)w(n)e^{-i\omega n}}{\sqrt{N}}| \tag{B.14}$$

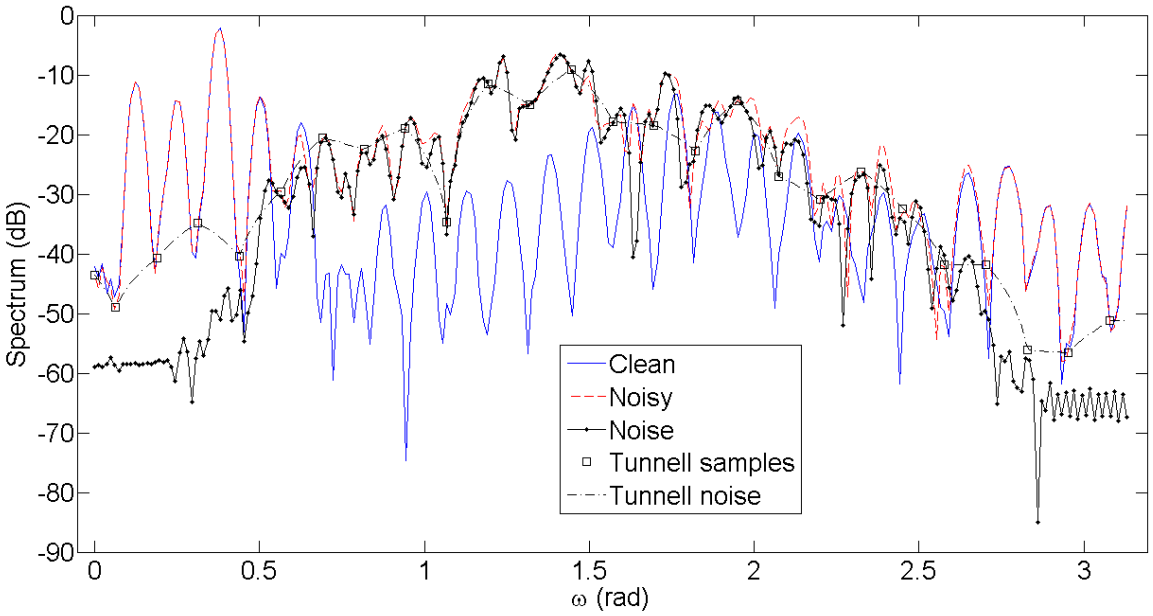



Figura B.13: Example of tunnelling noise estimation on a voiced noisy frame with pitch $\omega_0 = 0.126$ rad..

where $w(n)$ is the Hamming window. Then, the discrete **NTun** (a variation of harmonic tunnelling noise) is estimated by interpolating tunnelling samples $Y(\omega_l)$ which are obtained from the pitch frequency ($\omega_0$) as follow:

$$\hat{N}_{tun}(\omega_j) = Interp(\boldsymbol{\omega_l}, Y(\boldsymbol{\omega_l}), \omega_k) \tag{B.15}$$
$$\omega_l = \omega_0(l + \frac{1}{2}),\ l = \{-1/2, 0, 1, 2, .., ceil(\pi/\omega_0)\}$$
$$\omega_j = \frac{2\pi j}{NFT},\ j = \{0, .., NFT/2 - 1\}$$

Figure B.13 shows an example of tunnelling noise estimation. *NTun* has the problem of overestimation mainly at high SNRs (more than 10dB) because of the spectral window (as shown in the figure at low/high frequencies).

### VAD+Tun Noise Estimate

The final noise estimate is *NVAD* but corrected, depending on global SNR estimate, at voiced frames as follows:

- If global $SNR < 10dB$: NVAD is replaced by *NTun*.

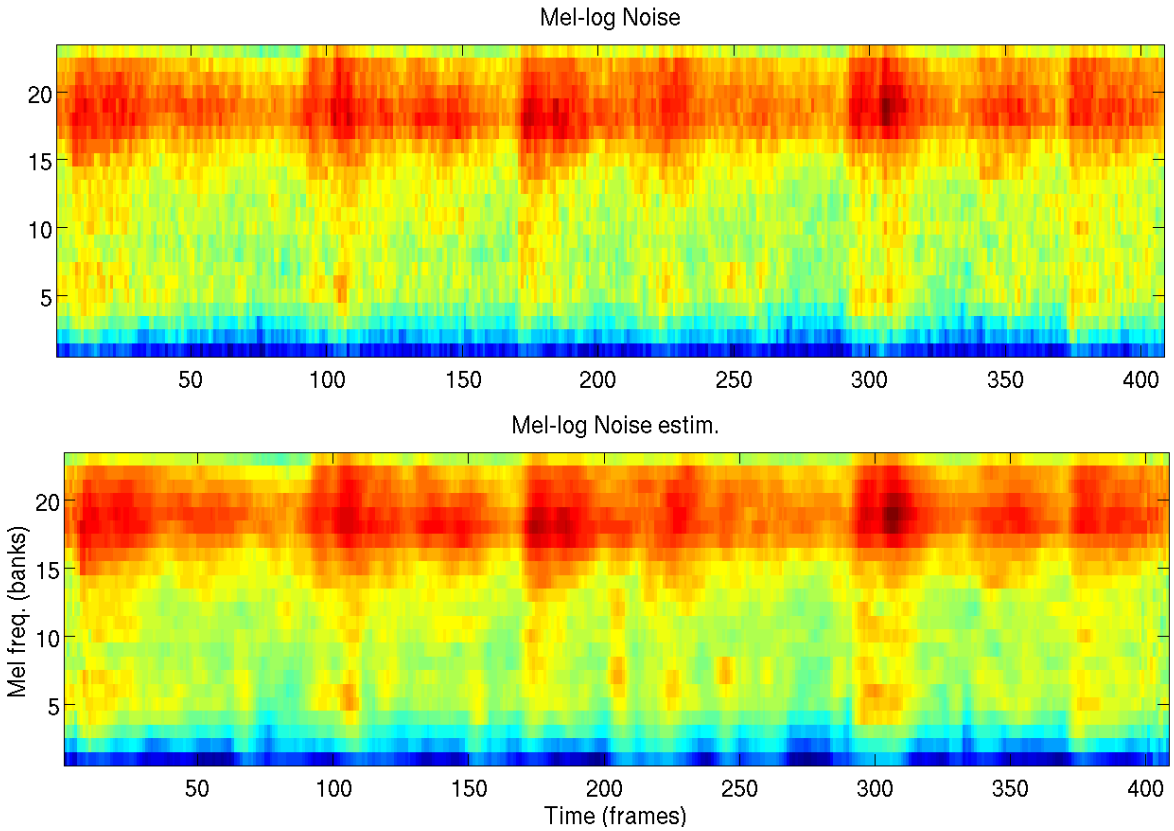



Figura B.14: Subway Mel-log noise and its estimation from Aurora-2 utterance 4460806 at 0dB

- Otherwise: *NTun* is used as an upper bound for *NVAD*.

The reason for using *NTun* only as an upper bound when $SNR \geq 10dB$ is that overestimation is more likely in this case. Also, real noises tend to be more stationary at high SNRs [89]. The final noise spectrogram *NVADTun* is smoothed and its $Fb\hat{n}$ spectrogram (Filter bank Mel-Log representation) is obtained. Fig. B.14 depicts a comparative example.

### Mask Estimation

The clean spectrogram $Fb\hat{x}$ is estimated subtracting $Fby$ and $Fb\hat{n}$ and then the local SNR of every pixel (mel filter $ch_j$ at time $t_k$) can be obtained as:

$$\hat{SNR}(ch_j, t_k) = 20 * log_{10}(e^{Fb\hat{x}(ch_j,t_k)}/e^{Fb\hat{n}(ch_j,t_k)}) \tag{B.16}$$

| System | Clean | 20 dB | 15 dB | 10 dB | 5 dB | 0 dB | -5 dB | Mean (20-0 dB) |
|---|---|---|---|---|---|---|---|---|
| FE (Ceps) | 99.14 | 97.21 | 92.57 | 76.72 | 44.28 | 22.99 | 13.00 | 66.76 |
| N. VAD+Tun, SS (Ceps) | 99.36 | 96.66 | 92.09 | 81.84 | 64.09 | 37.06 | 9.72 | 74.35 |
| A. Sift (Ceps) | 98.63 | 96.69 | 94.50 | 89.39 | 76.30 | 44.60 | 14.75 | 80.30 |
| AFE (Ceps) | 99.11 | 97.72 | 96.05 | 91.84 | 82.19 | 59.91 | 28.87 | 85.54 |
| N. VAD+Harm (MD, Cocl) | 98.67 | 96.18 | 92.67 | 84.17 | 74.21 | 50.41 | 17.65 | 79.53 |
| N. VAD (MD) | 98.76 | 96.19 | 93.38 | 88.42 | 77.92 | 49.52 | 15.56 | 81.09 |
| N. VAD+Tun (MD) | 98.78 | 95.79 | 92.04 | 86.66 | 78.03 | 54.43 | 18.40 | 81.39 |
| N. VAD+Tun Ideal (MD) | 98.78 | 95.97 | 92.81 | 88.57 | 84.24 | 74.43 | 55.83 | 87.21 |

Tabla B.6: WAcc results obtained by different systems tested with Aurora-2 (Set A, B and C) for different SNR values.

This is passed through a sigmoid function to obtain the soft mask estimate $M$ (reliability values between [0, 1]). The threshold and the slope of the sigmoid are -3 dB and 0.2 respectively and they have been determined empirically.

### Experimental results

Tab. B.6 shows the WAcc results with Aurora-2. The first four systems use the cepstrograms with CMN (*Ceps*). *FE* stands for a cepstrum obtained from the spectrogram *Fby* and provides a very similar result to the ETSI front-end [149], *AFE* is the ETSI front-end [147], and *A. Sift* is the sifting autocorrelation (Sec. B.4.2) which is an example of pitch-based robust technique. *N. VAD+Tun, SS* is when the proposed noise estimate is used in an Cepstral SNR-dependent SS (Spectral Subtraction) scheme which parameters have been optimized to avoid musical noise.

The next four systems estimate a soft mask to recognize (*MD*). *N. VAD*, *N. VAD+Tun* and *N. VAD+Tun Ideal* use our proposed noise estimates. *Ideal* means that pitch is obtained from corresponding clean signal. These three systems employ a 23-channel spectrogram as acoustic representation. However, *N. VAD+Harm*, which is an adaptation of Barker's technique explained in Sec. B.3.2 especially developed to compare with our technique, employs a 23-channel cochleagram (Cochl). Its VAD is the same as the one we have previously proposed but adapted to the cochleagram representation. The values of threshold and slope of the sigmoid functions of $M_n$ and $M_h$ are (-6 dB, 0.8) and (0.8,70) respectively, and they have been determined empirically.

The following conclusions can be drawn:

- *N. VAD+Tun* performs better in Spectral MD than in Cepstral SNR-dependent SS. This is because SS is more sensitive to errors of noise level. This is the reason why MD is preferred instead of the SS approach as HT does.
- If we compare *N. VAD* with *N. VAD+Tun*, we see that the addition of *NTun* provides benefits, mainly at low SNRs. However, we also see that tunnelling is not beneficial at higher SNRs. This can be understood if we take into account that Aurora-2 mainly consists of (quite) stationary noises. On the other hand, we think that our technique can be more helpful for non-stationary or sporadic noises.
- If we compare *N. VAD+Harm* with *A. Sift* and *N. VAD+Harm Cocl*, it seems that the proposed noise estimate makes a better use of the pitch information than the other two. However, this can not be concluded definitively as several causes can be influencing on this. Among others, that *A. Sift* and *N. VAD+Harm Cocl* can be more sensitive to pitch errors or that their parameters are not optimally tuned. This kind of problems shows the need of determining which technique makes a better use of the pitch information. The answer to this question will be addressed in Sec. B.5.
- *N. VAD+Tun Ideal* show that with a better pitch estimation, results could be considerably improved (overcoming *AFE*). In future work (Sec. C.3) different possibilities to improve the pitch estimation are discussed.

## B.5. Equivalences and limits of the pitch-based techniques

### B.5.1. Basic mechanisms and equivalences

#### Voiced basic mechanisms

In previous sections we have studied and proposed different pitch-based techniques for robust ASR. Now, we will compare them in a fair way by means of using some equivalences. In principle, they can be supposed as different if we only pay attention to some specific details (pitch extractor, processing of unvoiced and silence frames, etc.). However, they can be reduced to one of these four basic mechanisms which depend on the robust method applied to voiced frames:

1) **Exploitation of the harmonic structure**: these mechanisms do not require a pitch extraction but only some properties which can be derived from periodicity. SWP

[92], HASE [142] and Asymmetric Windows (Sec. B.4.1) try to «clean» the signal using these properties. HF [129] estimates the noise by exploiting the spectral harmonic shape.

2) **Comb estimation of clean signal**: these mechanisms use the pitch frame to apply some kind of comb filtering, i. e. some kind of algorithm which can be reduced to a sort of removing noise between the gaps (or tunnels) which are in the middle between the pitch spectrum harmonics. The resulting clean signal can be recognized from its cepstral representation. WHNM [138], PHCC [52] and Sifting (Sec. B.4.2) use these mechanisms.

3) **Tunnelling estimation of noise**: these mechanisms are the opposite of the preceding ones and estimate noise (tunnelling noise) employing tunnelling samples, that is, the spectral gaps between the harmonics. The resulting noise estimate can be employed in SS, MD, etc.. HT [38], FPM-NE [19] and Pitch-based Noise Estimation (Sec. B.4.3) use these mechanisms.

4) **Harmonicity mask estimation**: this mechanism estimates the mask of each frequency-temporal pixel by means of the correlogram and the pitch. Cochleagram techniques related with ASA, such as the adaptation of Barker's technique (Sec. B.3.2) and the Ma's technique [90] employ this mechanism.

Taking into account these mechanisms we can investigate about which is the best one and whether they fully exploit the pitch information to improve the recognition in voiced frames. These questions are answered in Sec. B.5.2.

### Comparing tunnelling and harmonicity masks

It can be shown that the mask derived from tunnelling noise is similar to that derived from harmonicity measures if similar channel numbers and a suitable selection of thresholds are applied.

Fig. B.15 can help to understand this similarity. The clean and tunnelling noise estimate, which indicates where the mask should be 1 or 0, are on top of the picture along with the 10 Mel filter bank, employed in tunnelling estimation. The outputs of the 10 gammatone channels of the correlogram employed to estimate harmonicity mask are in the middle plot. The two mask estimates (*Harmonicity and Tunnelling Mask*) are overlapped at the bottom of the picture along with the Log-Mel spectra employed to estimate the tunnelling mask, showing the strong similarity of both estimates. We can conjecture that both masks will yield similar recognition results (hypothesis H1).

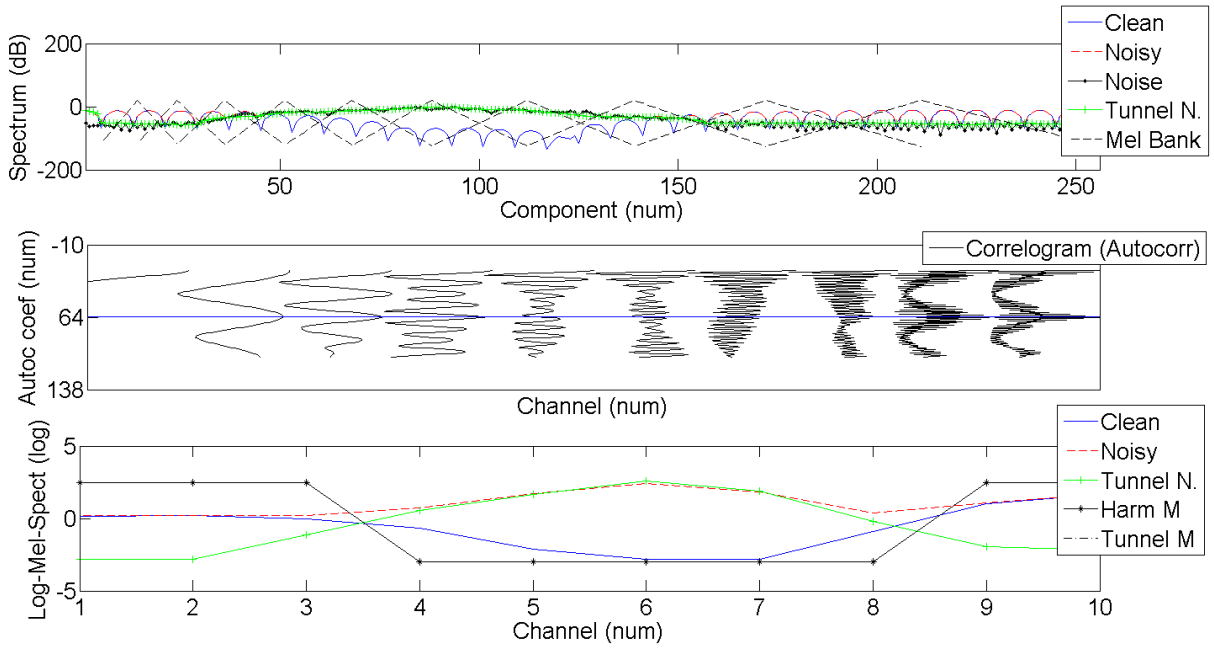



Figura B.15: Comparison of the mechanisms to estimate a tunnelling mask and a harmonicity mask. Both masks are shown in the Log-Mel Spectrum plot

## B.5.2. Optimum voiced mechanisms

### Optimum pitch-based noise estimation

Let's suppose that we have a noisy signal $x(n)$ of length $N$ which is the sum of a pure periodic clean signal $p(n)$ and a distortion $d(n)$. $T$ (or $\omega_0$ in radians) is the period of $p(n)$ and, for the sake of simplicity, we also suppose that we have an integer number of periods $N_p$ ($N = N_p * T$). Its complex discrete noisy spectrum is:

$$X(\omega_k) = P(\omega_k) + D(\omega_k) \qquad (k = 0, ..., N-1) \tag{B.17}$$

Taking into account the periodicity of $p(n)$, the above equation can be expressed as follows:

$$X(\omega_k) = \begin{cases} P(\omega_k) + D(\omega_k) & \text{if} \quad \omega_k = \omega_0 m \\ D(\omega_k) & \text{otherwise (tunnelling samples)} \end{cases} \tag{B.18}$$

where $m = 0, 1, .., T-1$. From this equation, we can deduce that only a percentage $(Np-1)/Np$ of the $N$ noise spectral samples can be recovered if we only know the pitch period $T$, no matter how the noisy signal is transformed. The remaining noise frequency samples are mixed with the speech harmonics and can not be recovered, although they

can be estimated by applying some type of interpolation.

We can consider that the noise spectrum estimates obtained from tunnelling samples and interpolation are optimal in the sense that minimal assumptions about the noise are required (only an interpolation model). In practice, it must be also taken into account that the resulting noise estimation has some problems like non perfect periodicity or unavoidable time-window which also widens the harmonics. The reason of only taking one tunnelling sample (between the harmonics) in the proposed Pitch-based Noise Estimation technique is this widening.

### Optimum voiced mechanisms

Let us consider the following three points:

1. Tunnelling noise estimate is theoretically optimum (just argued above).
2. The similarity between tunnelling and harmonicity masks (Sec. B.5.1).
3. MD (with ideal mask) provides much better results than other techniques which employ a noise estimate (such as SS) (Sec. B.3.1).

From these three considerations, we can say that mask estimation mechanisms based on tunnelling or harmonicity, along with MD recognition, provide a very solid framework for pitch-based recognition of voiced frames, and that in ideal conditions these can be considered as an optimum mechanisms (hypothesis H2).

## Experimental results

In order to compare the robustness of the four basic mechanisms for voiced frames, WAcc results in spectrogram (or cochleagram) domain, with ideal pitch and with oracle mask in unvoiced and silence frames for different techniques (representative of each mechanism) are shown in Tab. B.7.

*FE* is used as baseline (no robust). $DDR_{55,200}$ corresponds to the asymmetric window (Sec. B.4.1) and represents the mechanisms based on exploiting the harmonic structure. *A. Sift* corresponds to the sifting autocorrelation technique (Sec. B.4.2) and represents the mechanisms based on comb estimation of the clean signal. *N. VAD+Harm* is the adaptation of Barker's technique (Sec. B.3.2) and represents the mechanisms based on harmonicity mask estimation. *N. VAD+Tun* is the tunnelling mask (Sec. B.4.3) and represents the mechanism based on tunnelling noise estimation.

| Technique | Mean (20-0 dB) [0 dB] | | |
|---|---|---|---|
| | Technique «per se» (without oracle) | Oracle mask unvoc. and sil. | Oracle mask all |
| FE (Spectr.) | 33.30 [7.66] | 64.25 [25.04] | 95.01 [90.18] |
| $DDR_{55,200}$ (Spectr.) | 35.84 [5.84] | 73.16 [37.98] | 90.35 [82.75] |
| A. Sift ($\delta = 8$) (Spectr.) | 36.61 [8.09] | 77.92 [47.72] | 93.36 [88.94] |
| N. VAD+Harm (Cocl.) | 85.95 [72.21] | 89.15 [73.13] | 95.11 [89.40] |
| N. VAD+Tun (Spectr.) | 87.21 [74.43] | 90.87 [79.46] | 95.01 [90.18] |

Tabla B.7: WAcc results for the whole Aurora-2 (Set A, B and C) obtained by four techniques which represent the four basic voiced mechanisms. 0 dB result is shown in bracket. Ideal pitch is employed.

The first column shows the results obtained by these techniques (all-ones mask has been employed for the first three techniques). The second column shows the same experiments but applying oracle masks to unvoiced frames and silences (this shows the success of the voiced mechanisms), and third column shows oracle mask results. The soft-mask threshold and slope of *N. VAD+Harm* and *N. VAD+Tun* have been re-optimized to improve the results in the second column.

It can be concluded that the best voiced mechanisms are the two last ones, i. e. harmonicity and tunnelling mask estimations. Their results are quite similar although tunnelling is a bit better. This increment can be due to the difference between the Mel scale of the spectrogram and the ERB scale of the cochleagram. Except for this difference, it can be said that these mechanisms are similar and that they are best ones. This confirms many of the previous statements made in this section (hypothesis H1 and H2).

## B.5.3. Limits in pitch-based recognition

### Performance limits

If we compare the first and second columns of Tab. B.7 for the proposed technique *N.VAD+Tun* and it is taken into account that second column contains an approximation to the best performance that we can obtain with the pitch-based techniques (because unvoiced and silence frames have oracle mask and voiced frames have one of the optimum voiced mechanisms) we can conclude that the proposed pitch-based noise estimation technique (first column) is almost optimum because its results are not very far from this upper boundary results (second column).

Let us compare now the second and third columns of the table. Although the results of the second column are not very far from those of the third one (oracle masks for all frames), we can see that the pitch-based mask estimation methods will never perform as well as the oracle masks (this is specially clear at 0 dB), independently of the accuracy of the pitch extractor employed. This points out that in order to obtain further improvements, more information than that extracted from the pitch trajectories would be required to approximate the performance of the oracle masks. This extra information could be obtained from the noise itself or accurate speech models.

### Recognition of speech without pitch

This thesis has been devoted to the recognition of speech as it is usually uttered, that is, with vibration of the vocal folds. However, speech can be sometimes emitted without pitch (whispered speech, [159]) or with multiple pitch values (vocal harmony, in music). Humans can recognize these voices even in noise conditions. This can create the illusion that pitch is not an important cue in robust speech recognition. However, as it is explained in the introduction section, although we consider the pitch as an important cue, it is not the only one. We consider the ASR of whispered speech as an important field for future work which we are willing to study. To do that, the following ideas could be considered (most of them extracted from this Thesis):

- Design of a VAD detector similar to that developed in Sec. B.4.3, taking into account the main source model of speech. In this Thesis, the main source is associated to pitch. Now, the main source could be localized where instantaneous SNR is higher (whispered) or multiple pitches rise at the same time (vocal harmony).
- Adaptation and improvement of the models for this type of speech, taking into account that now it has a flatter spectrum, with less energy (whispered), etc. [159, 67].
- Application and adaptation of the MD (or SFD [5]) techniques to this type of speech.

# Apéndice C

# Conclusions, Contributions and Future Work

## C.1. Conclusions

The present work is motivated by the need of proposing and carrying out a comparative study of robust speech recognition techniques based on pitch (not including robust pitch extraction). The main conclusions are summarized below:

- Taking into account that the message of a speech signal is coded by means of three kind of elements (voiced sounds, unvoiced and silences) and the way they are combined, we can say that the speech signals «mainly» consists of voiced sounds which are surrounded by the unvoiced sounds. This has been referred to as «main source model» which is a simplify definition of speech that it has been employed to develop a VAD (Sec. B.4.3). This model is also suitable for whispering speech if a noise is taken as the main source.

- The state of the art of conventional techniques for robust ASR leads to the conclusion that MD (Missing Data) techniques can obtain very high performances (close to human) without the need of perfectly estimating the noise or the clean signal. However, this transfers the problem to the mask estimation block.

- The comparative study of the pitch-based techniques found in the bibliography (exploitation of harmonic structure, clean signal estimation and mask estimation techniques) is a difficult task because each author employs a different pitch extractor, each technique uses extra techniques and sometimes it is not clear if the author is

proposing a new pitch-based robust technique or a robust pitch extractor itself. Even so, we have tried to establish some equivalences between the different techniques and the recognition limits of the pitch-based techniques.

- A set of asymmetric windows called $DDR_{c,w}$ has been proposed which extends the HASE technique [142] that is employed to perform robust feature extraction by means of the OSA in white-like noises (contained in the first autocorrelation lags). It has been shown that the highest robustness is obtained by windows centered around the pitch values because these are the most energetic autocorrelation lags (have more SNR) and preserve the formant information. The coefficients which should be less weighed are the first ones because they are the most affected by the noise.

- A clean autocorrelation estimation method called *sifting* (based, in turn, on another proposed estimator, which was referred to as *averaging* estimator) has been proposed. It uses the pitch and depends on the sifting parameter $\delta$ which informs about the amount of autocorrelation products which are rejected because they are supposed to be more contaminated by noise. It has been shown that, taken a suitable $\delta$ value, which includes the first (more energetic) autocorrelation coefficients of a white-like noise, the estimate can be equal to the clean signal autocorrelation under certain assumptions.

- Taking into account that for $\delta = 0$ sifting is a sort of comb filtering (a spectral sampling of noisy signal at the pitch harmonics) and that many of the pitch-based techniques can be reduced to a comb filtering, we can concluded that sifting is an extension of many of these comb techniques. Sifting has the advantages of the comb techniques (eliminating the noise placed between pitch harmonics) and HASE (eliminating white-like noises).

- The extension to unvoiced frames of both the $DDR_{c,w}$ windows and sifting could degrade the performance (mainly at clean conditions) because the information of unvoiced sounds is mainly contained in the first autocorrelation coefficients, which tend to be removed. Nevertheless, this problem can be avoided by applying the same technique in both, training and test stages.

- Techniques such as HT [38] or that of Frazier [46], based on estimating the noise spectrum in voiced frames by means of tunnelling samples (spectral samples which

are between the pitch harmonics), have the problem of including as noise unvoiced frames (VAD is not used) and of overestimating it, degrading the performance as they also employ SS (Spectral Subtraction) which is very sensitive to these overestimations. In order to avoid these problems a recognition system, which includes a VAD+Tunnelling noise estimation and MD instead of SS, has been proposed.

- The proposed VAD uses the pitch location in order to locate the rest of the speech elements taking into account the *main source model* of speech. The tunnelling estimate also uses the pitch so we have finally proposed a *noise estimation based completely on pitch*.

- If we do not consider some elements of the pitch-based techniques, such as the pitch extractor, treatment of the unvoiced and silence frames, etc., it can be concluded that they employ one of these four basic mechanisms in voiced frames: exploitation of the harmonic structure, comb estimation of the clean signal, tunnelling noise estimation (or anti-comb-filtering) which can be employed for SS (HT) or for mask estimation (as in our proposal) and harmonicity mask estimation.

- The maximum number of noise spectral samples which can be recovered in a noisy voiced frame by means of the pitch are (in ideal conditions) the $N(N_p - 1)/N_p$ tunnelling samples, where $N$ is the frame length and $N_p$ the number of periods of the voiced signal. From this it can be deduced that, in order to estimate noise, it is necessary to add more information about the noise and it is just what tunnelling estimation (HT, FPM-NE or our proposal) does when the noise is interpolated by using theses tunnel samples. It can be concluded that (ideally) this kind of techniques achieve optimum noise estimation based on pitch and employing very little information about the noise (the interpolation model).

- It can be shown that mask estimation by means of both tunnelling noise and harmonicity mechanisms yields similar masks. Taking into account that tunnelling noise is optimum (at least, under certain conditions) and the advantages of MD (as compared to SS), we can conclude that the mask estimation mechanisms based on tunnelling or harmonicity, along with MD recognition, provide a very solid framework for pitch-based recognition of voiced frames and that, in ideal conditions, these can be considered as an optimum mechanisms. The experimental results, employing oracle masks, support this assertion.

- Taking into account the optimum voiced mechanisms and the experimental results with oracle masks (in unvoiced and silence frames), we can conclude that the proposed pitch-based noise estimation technique performs reasonably well (with ideal pitch) because its results are close to the limits of the pitch-based ASR techniques (using the minimal noise information). Besides, these results are not very far from the oracle mask results. In order to reach these oracle results it would be necessary to add more information (about noise or speech) in the mask estimation.
- Some ideas presented in this work, such as employing MD or the main source model to obtain a VAD, can be exploited to recognize whispered speech (without pitch).

## C.2. Contributions

The main contributions of this Ph.D. dissertation can be summarized as follows:

- We propose a set of asymmetric windows which are applied to the OSA in order to carry out robust feature extraction with low computational cost [107].
- We propose a clean autocorrelation estimator which employs the pitch and can deal with harmonic (not related with pitch) and white-like noises. This estimator is the sifting estimator [106].
- We propose a VAD and a pitch-based noise estimator from a simplify voiced model (main source model) which solves many of the problems of similar techniques [105].
- We study different pitch-based techniques, classify them, show their equivalences and point out the limits of the pitch-based recognition, showing that the proposed pitch-based noise estimation technique is close to these limits.

## C.3. Future Work

Many of the experiments developed in the Thesis (such as those with ideal pitch) point out possible future work. They can be summarized as follows:

- Regarding **asymmetric windows**, robust feature extraction employing windows centered on the mean pitch speaker could be carried out in order to improve performance as experimental results of Sec. B.4.1 show.

- Regarding **sifting autocorrelation** a dynamic $\delta$ could be applied in order to improve the results (experiments with oracle $\delta$ show this, Sec. B.4.2). The idea of sifting could even be extended, in the sense of not deleting only the products around the main diagonal but also those around other diagonals or other table positions more affected by noise.

- Regarding **pitch-based noise estimation** we can say that the main point is to improve the pitch extraction as shown by the ideal pitch results. If this was done, the technique would almost reach the limits of pitch-based techniques as Tab. B.7 points out (without the necessity of improving the VAD). One solution could be to consider several pitch candidates at each frame, and each candidate could result in a different noise estimation hypothesis. These parallel hypotheses could be evaluated separately by using missing data marginalization and employing the mask derived from a hypothesized noise estimate. The pitch which gave the highest likelihood would be chosen. This is similar to the SFD (speech fragment decoding) idea which uses top-down speech models to resolve bottom-up signal ambiguity.

- Another interesting work which is pointed out by table B.7 is trying to reach the oracle mask limits mainly at low SNRs. As we have seen, we can not reach these limits only by means of the pitch. The way to do that would be adding more information about the noise (or speech) to the mask estimator. This information could be dynamically updated in time from silence regions.

- Finally, recognition of speech without or even with multiples pitch values (whispered or vocal harmony speech) is a very interesting line as it is discussed in Sec. B.5.3.

# Bibliografía


[1] S. Ahmed and Volker Tresp. *Advances in neural information processing systems*, chapter Some Solutions to the Missing Feature Problem in Vision. 1993. 4.2.1

[2] A. Albiol-Colomer, V. Naranjo-Ornedo, and J. Prades-Nebot. *Tratamiento digital de la señal: teoría y aplicaciones*. Universidad politécnica de Valencia, 2007. 3.1.3

[3] Aurora-3-Danish. Aurora-3, aurora project database: Subset of speechdat-car, danish database. Technical report, ELRA (European Language Resources Association), 2001. A.2

[4] Aurora-3-Spanish. Aurora-3, aurora project database: Subset of speechdat-car, spanish database. Technical report, ELRA (European Language Resources Association), 2001. A.2, B.4.1

[5] J. Barker, M. Cooke, and D. Ellis. Decoding speech in the presence of other sources. *Speech Communication*, 45:5–25, 2005. 5.1.6, 7.3.2, B.3.1, B.3.2, B.5.3

[6] J. Barker, M. Cooke, and P. Green. Robust asr based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise. In *Eurospeech*, pages 213–216, 2001. (document), 4.2.3, 5.2.3, 5.5, 6.3.4, 7.1.1, B.2, B.3.2, B.4.3, B.4.3

[7] J. Barker, L. Josifovski, M. P. Cooke, and P. D. Green. Soft decisions in missing data techniques for robust automatic speech recognition. In *ICSLP*, 2000. 4.2.3, B.4.3

[8] J. Barker, N. Ma, A. Coy, and M. Cooke. Speech fragment decoding techniques for simultaneous speaker identification and speech recognition. computer speech and language,. *Speech Commun.*, 24 (1):94–111, 2010. 4.2.2, 5.2.3

[9] J. Barker, P.Green, and M.P. Cooke. Linking auditory scene analysis and robust asr by missing data techniques. In *WISP Stratford-upon-Avon*, 2001. 5.2.3, B.3.2

[10] J. Beh and H. Ko. A novel spectral subtraction scheme for robust speech recognition: spectral subtraction using spectral harmonics of speech. In *Proc. IEEE ICASSP*, volume 1, pages 648–651, 2003. 5.1.4, 6.3.4, B.3.1

[11] A. Bernard and A. Alwan. Low-bitrate distributed speech recognition for packet-based and wireless communication. *IEEE Trans. on Speech and Audio Processing*, 10(8):570–579, 2002. 5.1.6, B.3.1

[12] A.D. Berstein and I.D. Shallom. An hypothesized wiener filtering approach to noisy speech recognition. In *ICASSP*, 1991. 5.1.4

[13] G. V. Békésy. The variation of phase along the basilar membrane with sinusoidal vibrations. *The Journal of the Acoustical Society of America*, 1947. 2.2.2, 2.2.3

[14] Steven F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics Speech and Signal Processing*, 27 (2):113–120, 1979. 5.1.4

[15] Herve Bourlard and Stephane Dupont. A new asr approach based on independent processing and recombination of partial frequency bands. In *ICSLP*, 1996. 4.2.1, 5.1.6, B.3.1

[16] Albert Bregman. *Auditory Scene Analysis*. MIT Press, Cambridge MA, 1990. 2.2.6

[17] G. Brown, J. Barker, and D. Wang. A neural oscillator sound separator for missing data speech recognition. In *Int. Joint. Conf. on Neural Networks*, 2001. 5.2.3

[18] Guy Brown and Martin Cooke. Computational auditory scene analysis. *Comput. Speech. Lang.*, 8 (4):297–336, 1994. 5.2.3, 7.1.1, B.3.2

[19] L. Buera, J. Droppo, and A. Acero. Speech enhancement using a pitch predictive mode. In *ICASSP*, 2008. 5.2.2, 5.2.2, 7.1.1, 2, B.3.2, B.3.2, B.5.1

[20] Luis Buera, Eduardo Lleida, Antonio Miguel, Alfonso Ortega, and Óscar Saz. Cepstral vector normalization based on stereo data for robust speech recognition. *IEEE Trans. Audio Speech Lang. Process.*, 15 (3):1098–1113, 2007. 5.1.4

[21] R. Carlyon and T. Shackleton. Comparing the fundamental frequencies of resolved and unresolved harmonics: evidence for two pitch mechanisms? ,. *J. Acoust. Soc. Am.*, 95:3541–3554, 1994. 3.3

[22] José L. Carmona. *Reconocimiento de Voz Codificada sobre Redes IP*. PhD thesis, Universidad de Granada, 2009. 5.1.1

[23] Dan Chazan, Meir Tzur, Ron Hoory, and Gilad Cohen. Efficient periodicity extraction based on sine-wave representation and its application to pitch determination of speech signal. In *EUROSPEECH*, 2001. 3.4.1

[24] A De Cheveigné. Speech f0 extraction based on licklider's pitch perception model. In *ICPhS*, 1991. 3.4.1

[25] Alain De Cheveigné and Hideki Kawahara. Multiple period estimation and pitch perception model. *Speech Communication*, 27:175–185, 1999. 3.4.1

[26] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111 (4):1917–1930, 2002. 3.4.1, 6.2.5

[27] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication*, 34:267–285, 2001. 3.2.2, 4.2.1, 4.2.3, 4.2.3, 4.2.3, 5.1.4, 5.1.6, B.3.1, B.4.3

[28] M. Cooke, A. Morris, and P. Green. Missing data techniques for robust speech recognition. In *ICASSP*, 1997. 4.2.2, 4.2.3

[29] Martin Cooke. *Modelling auditory processing and organisation*. PhD thesis, University of Sheffield (Also published by Cambridge University Press), 1993. 2.2.3, 5.2.3

[30] A. Coy and J. Barker. A multipitch tracker for monaural speech segmentation. In *Interspeech*, 2006. 5.2.3

[31] A. Coy and J. Barker. An automatic speech recognition system based on the scene analysis account of auditory perception. *Speech Commun.*, 49 (5):384–401, 2007. 5.2.3

[32] Malcolm J. Crocker. *Encyclopedia of acoustic*. John Wiley and Sons, Inc., 1996. 2.1

[33] C. J. Darwin. Perceptual grouping of speech components differing in fundamental frequency and onset-time. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 33 (2):185–207, 1981. 1.1.1, 1

[34] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio. Histogram equalization of speech representation for robust speech recognition. *IEEE Trans. Speech Audio Process*, 13:355–366, 2005. 5.1.3, B.3.1

[35] Dimitrios Dimitriadis, Jose C. Segura, Luz Garcia, Ros Potamianos, Petros Maragos, and Vassilis Pitsikalis. Advanced front-end for robust speech recognition in extremely adverse environments. In *Interspeech*, 2007. 3

[36] J. Droppo, L. Deng, and A. Acero. Evaluation of the splice algorithm on the aurora2 database. In *EUROSPEECH*, 2001. 5.1.4, B.3.1

[37] Jasha Droppo and Alex Acero. A fine pitch model for speech. In *INTERSPEECH*, 2007. 3.4.1, 5.2.2

[38] D. Ealey, H. Kelleher, and D. Pearce. Harmonic tunnelling: tracking non-stationary noises during speech. In *EUROSPEECH*, pages 437–440, 2001. (document), 5.1.4, 5.2.2, 5.4, 6.3.1, 6.3.3, 6.3.3, 6.3.3, 7.1.1, 8.1, B.3.2, B.4.3, B.5.1, C.1

[39] D. Ellis. *Prediction-Driven Computational Auditory Scene Analysis*. PhD thesis, MIT, 1996. 3.4.2

[40] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32 (6):1109–1121, 1984. 5.1.4, B.3.1

[41] D. Erro, I. Sainz, I. Saratxaga, E. Navas, and I. Hernáez. Mfcc+f0 extraction and waveform reconstruction using hnm: Preliminary results in an hmm-based synthesizer. In *FALA ('Jornadas en Tecnología del Habla' and 'II Iberian SLTech')*, 2010. 4.1.2

[42] Nicholas W. D. Evans, John S. Mason, and Key Words. Lpc-based, temporal-lateral noise estimation evaluated on the aurora corpus. In *IASTED SPPRA*, 2002. 5.1.4, 5.2.1

[43] Gunnar Fant. *Acoustic Theory of Speech Production*. Mouton: The Hague, 1960. 2.1.3

[44] H. Fletcher. *Speech and hearing in communication.* Van Nostrand Co., New York, 1953. 4.2.2

[45] Harvey Fletcher. Auditory patterns. *Rev. Mod. Phys.*, 1940. 2.2.2

[46] Ronald H. Frazier, Siamak Samsamt, Louis D. Braida, and Alan V. Oppenheim. Enhancement of speech by adaptive filtering. In *ICASSP*, 1976. 3.4.1, 5.2.2, 8.1, B.3.2, C.1

[47] M. Gales and S. Young. Robust continuous speech recognition using parallel model combination. *IEEE T. Speech. Audi. P.*, 4 (5):352–359, 1996. 5.1.5, B.3.1

[48] L. García, S.Umesh, C. Benítez, and J. C. Segura. Combining speaker and noise feature normalization techniques for automatic speech recognition. In *ICASSP*, 2011. 5.1.3

[49] B. Glasberg and B. Moore. Derivation of auditory filter shapes from notched noise data. *Hearing Res.*, pages 103–138. 2.2.2

[50] Julius L. Goldstein. An optimum processor theory for the central formation of the pitch of complex tones. *Journal of the Acoustical Society of America*, 54:1496–1516, 1973. 2.2.5

[51] J. A. González, A. M. Peinado, A. M. Gomez, J. L. Carmona, and J. A. Morales-Cordovilla. Efficient vq-based mmse estimation for robust speech recognition. In *ICASSP*, 2010. 5.1.4, B.3.1

[52] L. Gu and K. Rose. Perceptual harmonic cepstral coefficients for speech recognition in noisy environment. In *ICASSP*, 2001. 5.2.2, 7.1.1, B.5.1

[53] S. Harding, J. Barker, and G. Brown. Mask estimation for missing data speech recognition based on statistics of binaural interaction. *IEEE T. Audio. Speech.*, 14 (1):58–67, 2006. 3.2.3

[54] William Hartmann and Eric Fosler-Lussier. Investigations into the incorporation of the ideal binary mask in asr. In *ICASSP*, 2011. 1.1.1

[55] William M. Hartmann. *Signals, Sound, and Sensation (Modern Acoustics and Signal Processing).* AIP Press, Springer, 1998. 2.2.1

[56] H. Hermansky. Perceptual linear predictive (plp) analysis for speech. *J. Acoust. Soc. Am.*, 87(4):1738–1752, 1990. 5.1.2

[57] H. Hermansky. Recognition of speech in additive and convolutional noise based on rasta spectral processing. In *EUROSPEECH*, 1993. 5.1.2, 5.1.3

[58] H. Hermansky. Should recognizers have ears? *Speech Communication*, 25:3–27, 1998. 4.2.2

[59] Hynek Hermansky, Sangita Tibrewala, and Misha Pavel. Towards asr on partially corrupted speech. In *ICSLP*, 1996. 4.2.1, 5.1.6

[60] J. Hernando and C. Nadeu. Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition. *IEEE Transactions on Speech and Audio Processing*, 5 (1):80–84, 1997. 5.2.1, 6.1.1, B.4.1

[61] H. G. Hirsch and C. Ehrlicher. Noise estimation techniques for robust speech recognition. In *ICASSP*, 1995. 5.1.4, 5.2.1

[62] Hans Günter Hirsch and David Pearce. Automatic speech recognition: Challenges for the next millennium. In *ISCA ITRW ASR2000*, Paris, France, September 18-20 2000. 6.3.4, A.2

[63] John Holdsworth, Ian Nimmo-Smith, Roy Patterson, and Peter Rice. Implementing a gammatone filter bank. *Tecnical report, MRC Applied Psychology*, 1988. 2.2.3

[64] G. Hu and D. Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE T. Neural. Networ.*, 15:1135–1150, 2004. 5.2.3

[65] X. Huang, A. Acero, and H. Hon. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. 2001. 2.1.1, 2.1.3, 3.1.3, 1, 5.1.4, 5.1.4, B.1.1

[66] C. H. Hurst. A new theory of hearing. *Transaction of the liverpool biological society*, 1895. 2.2.5

[67] T. Itoh, K. Takeda, and F. Itakura. Analysis and recognition of whispered speech. *Speech Communication*, 45:139–152, 2005. 7.3.2, B.5.3

[68] Roman Jakobson, Gunnar Fant, and Morris Halle. *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. The MIT Press, 1961. 2.1.1, 4.1.1

[69] P. I. M Johannesma. The pre-response stimulus ensemble of neurons in the cochlear nucleus.the pre-response stimulus ensemble of neurons in the cochlear nucleus. In *Symposium on Hearing Theory, pages 58-69, IPO, Eindhoven, Netherlands*, 1972. 2.2.3

[70] C. Jutten and J. Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991. 5.1.1

[71] Nelson Y. Kiang. *Discharge patterns of single fibers in the cat's auditory nerve*. M.I.T. Press (Cambridge, Mass), 1965. 2.2.1

[72] D. Y. Kim, C. K. Un, and N. S. Kim. Speech recognition in noisy environments using first order vector taylor series. *IEEE Transactions on Signal Processing*, 5 (3):57–59, 1998. 5.1.4, 5.2.2

[73] Hyoung Gook Kim, Markus Schwab, Nicolas Moreau, and Thomas Sikora. Speech enhancement of noisy speech using log-spectral amplitude estimator and harmonic tunneling. In *Structure*, 2003. 5.1.4, B.3.1

[74] A. Klapuri. *Signal Processing Methods for the Automatic Transcription of Music*. PhD thesis, Tampere, 2002. 3.3, 5.2.3

[75] A. Klapuri. *Signal Processing Methods for Music Transcription*, chapter Auditory-Model Based Methods for Multiple F0 Estimation. Springer, New York,, 2006. 3.4.2

[76] T. Kristjansson and J. Hershey. High resolution signal reconstruction. In *IEEE Workshop ASRU*, 2003. 5.2.1

[77] Y. Kuroiwa and T. Shimamura. An improvement of lpc based on noise reduction using pitch synchronous addition. In *IEEE Int. Symp. Circuits and Systems*, volume 3, pages 122–125, 1999. 5.2.2, 6.2.8, B.4.2

[78] Mireille Lavigne, R. Pujol, S. Blatrix, T.Pujol, and V. Reclar-Enjalbert. *Promenade around the cochlea*. CRIC, University Montpelli. (document), 2.4

[79] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Comput. Speech. Lang.*, 9:171–185, 1995. 5.1.5, B.3.1

[80] Victor R. Lesser, S. Hamid Nawab, and Frank I. Klassner. Ipus: An architecture for the integrated processing and understanding of signals. *Artificial Intelligence*, 77, 1995. 4.1.1

[81] M. C. Liberman. The cochlear frequency map for the cat: labeling auditory-nerve fibers of known characteristic frequency. *J. Acoust. Soc. Am.*, 1982. 2.2.2

[82] J. C. R. Licklider. A duplex theory of pitch perception. *Experimentia*, 1951. 2.2.5, 3.3

[83] J. Lim. *Speech enhancement*. Prentice-Hall, 1983. 5.1.1, 5.1.2, 5.2.2

[84] F. H. Liu, R. M. Stern, X. Huang, and A. Acero. Efficient cepstral normalization for robust speech recognition. In *ARPA Speech and Natural Language Workshop,*, 1993. 5.1.3

[85] C. Llamas-Bello and V. Cardeñoso-Payo. *Reconocimiento Automático del Habla*. Universidad de Valladolid., 1997. 4.1.1

[86] Ramón López-Cózar and Zoraida Callejas. Asr post-correction for spoken dialogue systems based on semantic, syntactic, lexical and contextual information. *Speech Communication*, 50:745–766, 2008. 4.1.1

[87] R. Lyon. A computational model of filtering, detection, and compression in the cochlea. In *ICASSP*, 1982. 3.1.2

[88] R. Lyon. Computational models of neural auditory processing. In *ICASSP*, 1984. 3.3

[89] N. Ma, J. Barker, H. Christensen, and P. Green. Distant microphone speech recognition in a noisy indoor environment: combining soft missing data and speech fragment decoding. In *ISCA Tutorial and Research Workshop on Statistical And Perceptual Audition*, 2010. 5.1.6, 5.2.3, 6.3.3, B.4.3

[90] N. Ma, P. Green, J. Barker, and A. Coy. Exploiting correlogram structure for robust speech recognition with multiple speech sources. *Speech Communication*, 49:874–891, 2007. 3.4.1, 3.4.3, 5.1.6, 5.2.3, 6.2.5, 7.1.1, B.3.2, B.5.1

[91] Ning Ma. *Informing Multisource Decoding in Robust Automatic Speech Recognition.* PhD thesis, The University of Sheffield, Department of Computer Science, 2008. (document), 2.2.4, 3.1.2, 3.2.2, 3.2.3, 3.3, 4.3

[92] D. Macho and Yan Ming Cheng. Snr-dependent waveform processing for improving the robustness of asr front-end. In *ICASSP*, 2001. 5.1.2, 7.1.1, B.3.1, B.5.1

[93] D. Mansour and B.H. Juang. The short-time modified coherence representation and noisy speech recognition,. *IEEE Trans. Audio Speech and Signal Processing*, 37:795–804, 1989. 5.2.1, 6.1.1, B.4.1

[94] S. L. Marple. *Digital Spectral Analysis with Applications.* Prentice Hall. New Jersey, 1987. 5.2.1, B.4.1

[95] P. Martin. Comparison of pitch detection by cepstrum and spectral comb analysis. In *ICASSP*, 1982. 3.4.1

[96] R. Meddis. Simulation of mechanical to neural transduction in the auditory receptor. *J Acoust Soc Am.*, 79 (3):702–711, 1986. 3.1.2

[97] R. Meddis. Simulation of auditory-neural transduction: further studies. *J Acoust Soc Am.*, 83(3):1056–1063, 1988. 3.1.2

[98] R. Meddis and M. Hewitt. Modeling the identification of concurrent vowels with different fundamental frequencies. *J. Acoust. Soc. Amer.*, 91 (1):233–245, 1992. 3.3, 3.4.1

[99] Ray Meddis, Michael J. Hewitt, and Trevor M. Shackleton. Implementation details of a computation model of the inner hair?cell auditory?nerve synapse. *J. Acoust. Soc. Am.*, 87 (4):1813–1816, 1990. 3.1.2, 3.1.2

[100] J. Ming and F. Smith. A probabilistic union model for sub-band based robust speech recognition. In *ICASSP*, 2000. 5.1.6

[101] Brian C. J. Moore. *Encyclopedia of acoustic: Frequency analysis and pitch perception*, chapter 116, pages 1447–1460. John Wiley and Sons, Inc., 1997. 2.2.2, 2.2.5, 2.2.5

[102] Brian. C. J. Moore. *An Introduction to the Psychology of Hearing, Fifth Edition.* 2003. (document), 2.2.1, 2.2.2, 2.5, 2.2.3, 2.2.4

[103] Juan A. Morales-Cordovilla. Dos nuevas técnicas para el reconocimiento robusto de la voz. ventana asimétrica y autocorrelación por entremezclado. Dea (diploma de estudios avanzados), Univ. de Granada. Dpto. Teoría de la Señal, Telemática y Comunicaciones, 2008. 6.2.3

[104] Juan A. Morales-Cordovilla, Timo Bauman, José L. Pérez, Antonio M. Peinado, and Ángel M. Gómez. Implementación de un reconocedor distribuido de voz en tiempo real sobre ip. In *Actas de las IV Jornadas en Tecnologías del Habla (Zaragoza)*, 2006, Octubre. 5.1.1

[105] Juan A. Morales-Cordovilla, Ning Ma, Victoria Sánchez, Jose L. Carmona, Antonio M. Peinado, and Jon Barker. A pitch based noise estimation technique for robust speech recognition with missing data. In IEEE, editor, *ICASSP (International Conference on Acoustic, Speech and Signal Processing)*, pages 4808–4811, Mayo, 22-27 2011. 6.3.1, 7.1.1, 8.2, B.4.3, C.2

[106] Juan A. Morales-Cordovilla, Antonio M. Peinado, Victoria Sánchez, and José A. Gonzalez. Feature extraction based on pitch-synchronous averaging for robust speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 19(3):640–651, Marzo 2011. 3.4.1, 3.4.3, 6.2.1, 6.2.3, 6.2.3, 6.2.3, 6.2.5, 6.3.4, 7.1.1, 8.2, B.4.2, B.4.2, B.4.2, B.4.2, C.2

[107] Juan A. Morales-Cordovilla, Victoria Sánchez, Antonio M. Peinado, and Ángel. Gómez. On the use of asymmetric windows for robust speech recognition. *Circuits, Systems and Signal Processing (Springer)*, 2011, Abril (aceptado con cambios). 6.1.1, 7.1.1, 8.2, B.4.1, C.2

[108] P. Moreno. *Speech Recognition in Noisy Environments*. PhD thesis, Carnegie Mellon University, 1996. 5.1.3, 5.1.4, B.3.1

[109] Ángel de la Tore, Antonio M. Peinado, and Antonio J. Rubio. *Reconocimiento Automático de Voz en Condiciones de Ruido*. Monografias del Dpto. de Electrónoca Nº 47, Univ. de Granada, 2001. (document), 1.1.1, 1.1, 2.1, 2.2, 2.3, 5.1.2, 5.1.4

[110] A. M. Noll. Cepstrum pitch determination. *Journal of the Acoustical Society of America*, 41:293–309, 1995. 3.4.2

[111] Douglas O'Shaughnessy. *Speech Communications, Human and Machine, 2nd Edition*. IEEE Press, 2000. 2.1.3, 5.1.1, 5.1.2, 5.2.2

[112] Douglas O'Shaughnessy. Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, Volume 41, Issue 10, October 2008. 2.1.1

[113] Kuldip K. Paliwal and Yoshinori Sagisaka. Cyclic autocorrelation-based linear prediction analysis of speech. In *EUROSPEECH*, 1997. 5.2.1, B.4.1

[114] S. E. Palmer. *Vision Science*. MIT Press., Cambridge MA, 1999. 2.2.6

[115] K. Palomaki, G. Brown, and J. Barker. Techniques for handling convolutional distortion with missing data automatic speech recognition. *Speech Commun.*, 2004:123–142, 43. 3.2.3, 6.3.4

[116] Thomas W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *J. Acoust. Soc. Am.*, 60, Issue 4,:911–918, 1976. 3.4.1, 5.2.2

[117] S. Parveen and P. Green. Speech enhancement with missing data techniques using recurrent neural networks. In *ICASSP*, 2004. 3.2.3

[118] R. Patterson and B. Moore. *Auditory filters and excitation patterns as representations of frequency resolution.*, pages 123–177. Academic Press Ltd., London, 1986. 2.2.3

[119] R. D. Patterson. Auditory filter shapes derived with noise stimuli. *J Acoust Soc Am.*, 1976. 2.2.2, 2.2.2

[120] D. Pearce and H. G. Hirsch. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ICSLP*, volume 4, pages 29–32, 2000. 5.1.5, A.1, A.2, B.4.1

[121] Antonio M. Peinado and Jose C. Segura. *Speech Recognition over Digital Channels*. Wiley, 2006. (document), 4.2.1, 4.2.3, 5.1.1, 5.1, 5.1.4, 5.1.6, B.1.1, B.1, B.3.1, B.4.3

[122] James O. Pickles. *An Introduction to the Physiology of Hearing, Third Edition*. Emerald, 2008. 2.2.1

[123] Dimitris G. Proakis and John Manolakis. *Tratamiento digital de señales ($3^a$ Ed.)*. 2000. 3.1.3

[124] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice-Hall, 1993. 4.1.1

[125] Lawrence R. Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 25 (1), 1977. 3.4.1

[126] Lawrence R. Rabiner. A tutorial on hidden markov models and select aplication in speech recognition. In *IEEE*, 1989. 4.1.2, 4.1.2

[127] Bhiksha Raj, Michael L. Seltzer, and Richard M. Stern. Reconstruction of missing features for robust speech recognition. *Speech Communication*, 43 (3):275–296, 2004. 1.1.1, 4.2.3, 5.1.4

[128] P. Renevey and A. Drygajlo. Introduction of a reliability measure in missing data approach for robust speech. In *EUSIPCO*, 2000. 4.2.3, 5.2.3

[129] C. Ris and S. Dupont. Assessing local noise level estimation methods: application to noise robust asr. *Speech Communication*, 34 (2):141–158, 2001. (document), 5.1.4, 5.3, 5.2.1, 6.3.1, 7.1.1, B.3.2, B.5.1

[130] L. Robles and M. A. Ruggero. Mechanics of the mammalian cochlea. *Physiol. Rev.*, 2001. 2.2.1

[131] Robert Rozman and Dusan M. Kodek. Using asymmetric windows in automatic speech recognition. *Speech Communication*, 2007. B.4.1

[132] M. A. Ruggero. Responses to sound of the basilar membrane of the mammalian cochlea. *Curr. Opin. Neurobiol.*, 1992. 2.2.4

[133] M. A. Ruggero and N. C. Rich. Furosemide alters organ of corti mechanics: evidence for feedback of outer hair cells upon the basilar membrane. *J. Neurosci.*, 1991. 2.2.1

[134] J. Ryalls. *A basic introduction to speech perception.* Speech Science Series, 1997. 2.1, 2.1.1, 2, 2.1.1, 6.3.3, B.4.3

[135] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney. Gammatone features and feature combination for large vocabulary speech recognition. In *ICASSP*, 2007. 5.1.2

[136] JF Schouten. The residue and the mechanism of hearing. *J. Acoust. Soc. Am*, 1940. 2.2.5

[137] M. R. Schroeder. Period histogram and product spectrum: New methods for fundamental-frequency measurement. *J. Acoust. Soc. Am.*, 43, (Issue 4):829–834, 1968. 3.4.1

[138] M. Seltzer, J. Droppo, and A. Acero. A harmonic-model based front end for robust speech recognition. In *EUROSPEECH*, 2003. 5.2.2, 7.1.1, B.3.2, B.4.2, B.5.1

[139] M. Seltzer, B. Raj, and R. Stern. A bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech Commun.*, 43 (4):379–393, 2004. 5.2.3

[140] S. Seneff. Pitch and spectral estimation of speech based on auditory synchrony model. In *ICASSP*, 1984. 3.4.1

[141] Stephanie Seneff. A joint synchrony/mean-rate model of auditory speech processing. *Journal of phonetics*, 16:55–76, 1988. 3.1.2

[142] B. Shannon and K. K. Paliwal. Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition. *Speech Communication*, 48, no. 1:1458–1485, 2006. 5.2.1, 6.1.1, 6.2.1, 6.2.3, 8.1, B.3.2, B.4.1, B.4.1, B.4.2, B.5.1, C.1

[143] M. Slaney and R. F. Lyon. A perceptual pitch detector. In *ICASSP*, 1990. 3.3

[144] Stanley Smith Stevens, John Volkman, and Edwin Newman. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8:185–190, 1937. 2.2.2

[145] James V. Stone. *Encyclopedia of Statistics in Behavioral Science*, chapter Independent Component Analysis, pages 907–912. John Wiley and Sons, Ltd, Chichester, 2005. 5.1.1

[146] Y. H. Suk, S. H. Choi, and H. S. Lee. Cepstrum third-order normalisation method for noisy speech recognition. *IEE Electronic Letters*, 35(7):527–528, 1999. 5.1.3, B.3.1

[147] v1.1.1 ES 202 050. *Advanced front-end feature extraction algorithm.* ETSI, 2002. 5.1.2, 6.3.4, B.3.1, B.4.2, B.4.3

[148] v1.1.1. ES 202 211. *Distributed speech recognition; Extended front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm.* ETSI, July 2001. 3.4.1, 6.2.5, B.4.2

[149] v1.1.3 ES 201 108. *Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithms.* ETSI, April 2003. 3.1.3, 3.1.4, 5.1.2, 6.1.6, 6.3.4, A.1, B.3.1, B.4.1, B.4.1, B.4.3

[150] A. Varga and R. Moore. Hidden markov model decomposition of speech and noise. In *ICASSP*, 1990. 2.2.4, 5.1.5

[151] S. V. Vaseghi and B. P. Milner. Noisy speech recognition based on hmms, wiener filters and re-evaluation of most likely candidates. In *ICASSP*, 1993. 5.1.2

[152] Saeed V. Vaseghi. *Advanced Digital Signal Processing and Noise Reduction.* John Wiley and Sons, LTD, 2000. 3.1.3

[153] Hermann von Helmholtz. *On the sensations of tone as a physiological basis for the theory of music. (English Edition, translated by Alexander J. Ellis, 1877).* Dover, New York, 1885. 2.2.5

[154] Paul J. Walmsley, Simon J. Godsill, and Peter J. W. Rayner. Bayesian graphical models for polyphonic pitch tracking. In *Diderot Forum*, 1999. 3.3

[155] DeLiang Wang and Guy. J. Brown. *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications.* 2006. (document), 1.1.1, 2.2.6, 2.2.6, 3.1, 3.1.2, 3.1.5, 3.2.3, 3.6, 3.4.1, 4.2.2, 5.1.6, 5.2.3, 6.3.3, B.1.1, 1, 3, B.4.3

[156] Richard M. Warren. *Auditory Perception: A New Analysis and Synthesis.* Cambridge University Press, 1999. 2.1.1, 2.2.5

[157] M. Weintraub. The grasp sound separation system. In *ICASSP*, 1984. 3.3

[158] M. Weintraub. *A theory and computational model of auditory monaural sound separation.* PhD thesis, Stanford, 1985. 3.4.1, 5.2.2

[159] S. J. Wenndt, E. J. Cupples, and R. M. Floyd. A study on the classification of whispered and normally phonated speech. In *ICSLP, Denver*, 2002. 2.1.3, 7.3.2, B.5.3

[160] S. Windmann and R. Haeb-Umbach. Modeling the dynamics of speech and noise for speech feature enhancement in asr. In *ICASSP*, 2008. 5.1.4

[161] Stephen Wolfram. *A New Kind of Science.* Wolfram Media, Inc., 2002. 2.2.2

[162] Mingyang Wu and Deliang Wang. A multipitch tracking algorithm for noisy speech. *IEEE Transactions on Speech and Audio Processing*, 11:229–241, 2003. 3.4.1

[163] Qin Yan, Saeed Vaseghi, Esfandiar Zavarehei, Ben Milner, Jonathan Darch, Paul White, and Ioannis Andrianakis. Kalman tracking of linear predictor and harmonic noise models for noisy speech enhancement. *Speech Communication*, 22 (1):69–83, 2008. 5.1.4, 5.2.1

[164] T. Yoshioka, T. Nakatani, and H.G. Okuno. Noisy speech enhancement based on prior knowledge about spectral envelope and harmonic structure. In *ICASSP*, 2010. 5.2.1

[165] A. T. Yu and H. C. Wang. New speech harmonic structure measure and its applications to speech processing,. *Journal Acoustical Society of America*, 120(5):2938–2949, 2006. 5.2.2

[166] E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *J. Acoust. Soc. Am.*, 1961. 2.2.2

[167] E. Zwicker. Analytical expressions for critical band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.*, 68:1523–1525, 1980. 2.2.2