UNIVERSIDAD CARLOS III DE MADRID

DEPARTAMENTO DE TEORÍA DE LA SEÑAL Y COMUNICACIONES

TESIS DOCTORAL

CONTRIBUCIONES AL RECONOCIMIENTO ROBUSTO DE HABLA EN REDES DE COMUNICACIONES MEDIANTE TRANSPARAMETRIZACIÓN

Directores:
Prof. Dr. Fernando Díaz de María
Prof. Dra. Carmen Peláez Moreno

Leganés, Noviembre de 2011

TESIS DOCTORAL

CONTRIBUCIONES AL RECONOCIMIENTO ROBUSTO DE HABLA EN REDES DE COMUNICACIONES MEDIANTE TRANSPARAMETRIZACIÓN

Autor:
Diego Ferney Gómez Cajas

Directores:
Prof. Dr. Fernando Díaz de María
Prof. Dra. Carmen Peláez Moreno

Tribunal nombrado por el Mgfco. y Excmo. Sr. Rector de la Universidad Carlos III de Madrid, el día ___ de _____ de _____.

Presidente

Vocal

Vocal

Vocal

Secretario

Realizado el acto de defensa y lectura de la Tesis el día ___ de _____ de _____ en _____.

Calificación:

EL PRESIDENTE EL SECRETARIO LOS VOCALES

Agradecimientos

Doy gracias a Dios por darme su ayuda, la fortaleza y la inspiración para afrontar los desafíos de esta tesis; a mis tutores Fernando Díaz de María y Carmen Peláez Moreno, por su paciencia, su amabilidad, su permanente motivación, su dedicación y confianza, incluso en los momentos más difíciles.

A mis padres Rafael y Leonor, a mi hermana Sandra, por su ayuda y cariño incondicional, por su ejemplo de superación, de valores y virtudes, porque aún en la distancia siempre han estado presentes en cada palabra y cada gesto de aliento.

A toda mi familia por su permanente apoyo, siempre los tengo presentes.

A Dianey, porque su amor y ternura fueron siempre un motivo de esperanza y de aliento.

A mis profesores del Departamento de Teoría de la Señal y Comunicaciones, mi enorme gratitud porque me han aportado no solo la valía de sus conocimientos, sino también su ejemplo de vida. A Ascensión Gallardo, Matilde Sánchez y Harold Molina por sus aportes constantes a mi trabajo.

Y por su puesto a mis compañeros y amigos, que siempre estuvieron atentos a escucharme y brindarme su ayuda, a todos, mis más sinceros agradecimientos.

A la Universidad Carlos III de Madrid - España; al Ministerio de Tecnologías de la Información y las Comunicaciones, Colciencias, Colfuturo y la Universidad de Pamplona en Colombia, por su apoyo en la financiación de este doctorado.

Resumen

La creciente influencia de las redes de comunicaciones en todos los ámbitos de la vida moderna hace que cada vez sean más los servicios que se ofrecen a través de ellas, y dado que la comunicación oral es la forma más natural de comunicación humana, las tecnologías del habla juegan un rol importante en nuestra sociedad. Por este motivo, en esta tesis planteamos una serie de contribuciones al reconocimiento de habla en entornos de redes de comunicaciones, utilizando la técnica reconocimiento mediante transparametrización [113] (RMT) sobre los dos tipos de redes que más cobertura tienen hoy en día: Internet y la telefonía celular.

En particular, mejoramos la robustez ya demostrada de la técnica RMT frente a la distorsión por codificación y los errores de transmisión, y extendemos el análisis a casos con ruido de ambiente. En primer lugar, proponemos un procedimiento mejorado de estimación de la energía. En segundo lugar, aplicamos una técnica complementaria al RMT consistente en un filtrado del espectro de modulación, demostrando su eficacia en el entorno Internet.

Además, y específicamente para el entorno UMTS proponemos una extensión de parámetros fundamentada en la protección que realiza el codificador de canal normativo y que consigue hacer un uso eficaz de los parámetros más protegidos por el codificador de canal, en beneficio de la robustez del sistema de reconocimiento.

Abstract

Nowadays, the modern communication networks play an outstanding role in our everyday life and the number of services offered through them is continuously increasing. As the interfaces to these services become more natural, they tend to embed speech technologies so that the human-to-machine communication mimics (to some extent) the human-to-human communication. In this context, this thesis tackles the problem of automatic speech recognition (ASR) in communication-centered environments. In particular, our contributions focus on the bitstream-based approach to ASR, which has already proved to be robust, in two of the most relevant communication scenarios: Internet and universal mobile telecommunication system (UMTS) networks.

In this thesis we propose some techniques to improve the robustness of the ASR systems against the distortions resulting from the source coding and the transmission errors. For the voice over IP scenario, we propose an improved method for energy estimation and an additional technique based on filtering the modulation spectrum so that we are able to jointly deal with communication-related distortions and background noise.

For the UMTS scenario, besides an improved energy estimation method, in this thesis we propose an extended feature vector that relies on the unequal error protection mechanism implemented in the channel codec. This extended feature vector makes an effective use of the most protected parameters in the bitstream to provide the ASR system with an enhanced robustness.

Tabla de Contenido

Listado de Figuras

Listado de Tablas

Capítulo 1

Introducción

El uso de la voz como medio natural de comunicación ha hecho que las tecnologías del habla tengan un gran espacio en los ámbitos de investigación. Por otro lado, el despliegue de las redes de comunicaciones ha incrementado el desarrollo de aplicaciones de voz, tanto así que ahora ésta puede ser transmitida a través de un amplio abanico de redes, desde la tradicional Red Telefónica Básica (RTB), hasta las últimas generaciones de redes de telefonía móvil y, últimamente con un gran auge, sobre redes IP, (por mencionar algunas de las redes que han alcanzado mayor cobertura e impacto en la sociedad).

Podemos observar que, tanto en el caso de Internet como en el de las redes de telefonía móvil, la cobertura y capacidad de transmisión ha aumentado significativamente, permitiendo de esta manera el acceso masivo a la información y de una forma casi ubicua [133]. Es por ello que surge la necesidad de brindar nuevos servicios y aplicaciones que permitan sacar provecho a estas nuevas capacidades. No obstante, esto implica dar acceso fácil y natural a la información desde los terminales existentes (teléfonos móviles, tablets, agendas electrónicas, consolas de entretenimiento y todo tipo de dispositivos portátiles con capacidad de conectividad), teniendo en cuenta tanto sus potencialidades como sus limitaciones, siendo precisamente estas últimas (tamaño, consumo de energía, capacidad de procesamiento, etc.), las generadoras de grandes retos a la hora de facilitar el acceso óptimo a la información [138].

Por tanto, las tecnologías del habla proveen diversas soluciones a nuestras necesidades de comunicación modernas, desde sistemas de diálogo basados en reconocimiento y síntesis de habla, hasta sistemas de identificación y verificación de locutor. A continuación, a modo de ilustración, mencionamos algunas de ellas.

En el propio terminal, podemos tener:

- Marcación por voz, llamada a aplicaciones, acceso a documentos, información personal, contenidos pregrabados, etc.
- Interacción con aplicaciones de los terminales, facilidades para discapacitados, aplicaciones de dictado, etc.

Por otro lado, a través del terminal se puede tener acceso a servicios suministrados por aplicaciones centralizadas, y que utilizan una red de comunicaciones para el envío de la información al terminal, por ejemplo:

- Acceso a información que requiere actualización permanente: guías telefónicas, de restaurantes, hoteles, gasolineras; información de tráfico, del tiempo, etc.

- Algunos terminales que disponen de aplicaciones avanzadas pueden dar acceso a páginas web, catálogos en línea, venta de entradas, sistemas de reservas, billetes de transporte, pagos en línea, acceso interactivo a cámaras de vigilancia, servicios domóticos, etc.

Mediante el uso de interfaces vocales, se puede tener acceso a los anteriores servicios y aplicaciones, facilitando la interacción, especialmente en los dispositivos móviles, y en general en terminales limitados en tamaño, sin teclados, o en situaciones en donde se tienen las manos o vista ocupados.

No obstante la gran variedad de soluciones para el acceso a la información, existen también diferentes restricciones que se deben tener en cuenta para conseguir una calidad mínima de los diferentes servicios; en particular, cuando deseamos transmitir voz sobre una red de comunicaciones, existen limitaciones importantes tales como el ancho de banda del canal, y en este sentido, el proceso de codificación es un paso obligado, pues consigue (entre otras ventajas) una mayor eficiencia en el uso del canal. Sin embargo, este proceso de codificación implica retardos algorítmicos originados por el alto coste computacional que deben soportar los terminales, y más importante aún, cierto nivel de distorsión debido al proceso de codificación. Esto que no es un gran problema para algunos servicios como el de telefonía (pues subjetivamente esta compresión de los datos no afecta la calidad del servicio [117]), sí lo es cuando queremos realizar una tarea de Reconocimiento Automático de Habla (RAH) [114].

Por otro lado, existen otras condiciones adversas en un proceso de comunicación, tales como los efectos del canal en una red de telefonía móvil o la pérdida de paquetes en una red IP, que indudablemente generan problemas para los reconocedores de habla y, por tanto, se convierten en desafíos por resolver.

Es por ello que una aplicación que involucra el RAH sobre una red de comunicaciones implica, no solo resolver las dificultades propias del sistema de reconocimiento, sino también las introducidas por el paso de la voz a a través de la red. Estas dificultades, nos obligan a buscar soluciones que aborden de una manera conjunta dichos problemas, lo que ha dando origen a la presente tesis.

A continuación, se expondrán las motivaciones y objetivos concebidos para el desarrollo de esta tesis.

1.1. Motivación

En la introducción de este capítulo, explicamos la forma en que las redes de comunicaciones brindan un amplio abanico de servicios a nuestra vida diaria, pues están presentes en todos los ámbitos de la sociedad, en particular, las redes de telefonía móvil e Internet, sugieren grandes desafíos, bien sea para la creación de servicios o para adaptar los ya existentes a los nuevos entornos. Dichos servicios aportan valor añadido a una red de comunicación; ejemplo de estos son la telefonía IP, los sistemas de dialogo utilizando RAH, etc.

En el caso concreto del RAH, han surgido algunas estrategias que permiten prestar este servicio sobre los distintos tipos de redes; sin embargo, como mencionamos anteriormente, para cada entorno se presentan diferentes dificultades que disminuyen el rendimiento general del servicio; por ello queremos analizar en esta tesis, diversas maneras de abordar de forma conjunta estos problemas y brindar mayor robustez al RAH frente a los factores que lo afectan en dichos entornos, tales como la distorsión de codificación, el ruido, los errores de transmisión o la pérdida de paquetes.

1.2. Objetivos

Como objetivo principal de esta tesis, se ha planteado brindar una solución robusta a la problemática del reconocimiento de voz sobre las actuales de redes de comunicaciones, en particular, las redes de telefonía móvil e Internet, y que dicha solución abarque los principales problemas intrínsecos de cada entorno. Para este propósito, se ha utilizado el codificador ITU-T G.729 [68] para la simulación de las tareas de reconocimiento de voz sobre redes IP, y el codificador ETSI AMR-NB (Adaptive Multi-Rate Narrow Band) [4] para las simulaciones en la Red de Telefonía Móvil de tercera generación UMTS.

Dentro de las diferentes alternativas de reconocimiento de voz codificada, se pretende demostrar la eficacia de la técnica conocida como transparametrización o basada en el flujo digital binario (*bitstream-based*) [113], comparándola con la aproximación más habitual consistente en el reconocimiento de voz decodificada. Para ello se utilizarán diferentes entornos y condiciones que permitan probar la robustez de dichas técnicas frente a los problemas más frecuentes en una tarea de RAH sobre una red de comunicaciones.

Sin embargo, si bien existen algunos problemas comunes para los diferentes tipos de redes, como la distorsión por codificación o el ruido de ambiente, también existen algunos problemas que se deben tratar de forma particular en cada tipo de red. Así en una red IP, los errores en la comunicación se presentan en forma paquetes perdidos en ráfagas, mientras que en una comunicación móvil, el efecto principal cuando existen errores en el canal, se presenta en forma de pérdida de bits, siendo también lo más nocivo la pérdida de bits en ráfagas.

Cabe destacar que, aunque hay muchas soluciones dedicadas al ruido, en muchos casos específicas, en esta tesis hemos centrado nuestro análisis en la manera de compatibilizar la

técnica de la transparametrización con la existencia del ruido de ambiente y los errores de transmisión de forma simultánea.

1.3. Estructura de la Tesis

A continuación, en el Capitulo 2 se hará una introducción a los subsistemas típicos de un sistema de RAH y se proporcionará una breve revisión de sus fundamentos.

El Capítulo 3, se dedica a la exploración de similitudes y diferencias entre los procesos de codificación de voz y RAH, para explicar la convergencia que se puede dar entre ellos y las soluciones producto de esta convergencia. En este capítulo describimos el modelo de producción de voz humana, explicando el modelo fuente-filtro y sus aplicaciones tanto en codificación como en reconocimiento.

En el Capítulo 4, se aborda la problemática del reconocimiento de voz codificada, limitándonos a realizar una introducción a las distorsiones más importantes. Por otro lado, en este capítulo realizamos también una introducción a las soluciones propuestas según tres arquitecturas diferentes, a saber: *Reconocimiento Local, Distribuido y Remoto*.

El Capítulo 5 explica el estado de la técnica en el reconocimiento mediante transparametrización, centrando nuestro análisis en las soluciones que más se acercan a nuestros entornos de estudio, es decir, en redes de telefonía móvil y redes IP.

En el Capítulo 6 se describe a nivel algorítmico, nuestra propuesta de reconocimiento robusto frente a los problemas del RAH en redes de comunicaciones.

El Capítulo 7 explica el modelado del entorno experimental realizado para probar la robustez de la solución planteada. Básicamente, se han implementado dos modelos: uno para transmisión de voz sobre una red IP, y otro para simular la transmisión de voz sobre una red de telefonía móvil, utilizando para ello el estándar UMTS.

En el Capítulo 8 se exponen los resultados experimentales obtenidos con las técnicas de referencia explicadas en el Capítulo 4 y se describen los logros alcanzados con nuestras propuestas (descritas en el Capítulo 6).

Por último, en el Capítulo 9 se resumen las conclusiones, contribuciones y trabajos futuros.

Capítulo 2

Reconocimiento de Habla en Redes de Comunicaciones

2.1. Introducción

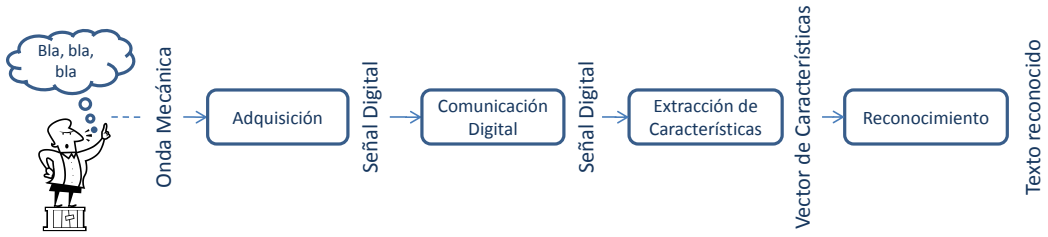


Figura 2.1: Etapas de un sistema de RAH en una red de comunicaciones.

2.2. Adquisición de la Señal de Voz

En un sistema de reconocimiento, lo primero que hay que hacer es convertir la onda mecánica que transporta la voz a un formato adecuado para su posterior procesamiento. Así, la primera etapa consiste en convertir onda mecánica en una señal eléctrica, para lo cual se utiliza un sistema transductor que puede constar de uno o varios micrófonos. Además de la distorsión inherente a la propia transducción, se captan diferentes tipos de ruido externos que influyen en mayor o menor medida en las prestaciones del sistema de reconocimiento.

A este tipo de ruido se denomina comúnmente ruido de ambiente, y es así como se le denominará en esta tesis a partir de ahora.

Una vez realizado el proceso de conversión de señal mecánica a señal eléctrica, se procede a la transformación de esta última en una señal digital. Para ello, se utilizan tres procesos conocidos como: muestreo, cuantificación y codificación. La forma empleada más habitualmente para implementar estos procesos es la descrita en la recomendación ITU-T G.711 [63], que involucra un muestreo a una frecuencia de 8 KHz, utilizando cuantificación no uniforme y codificación con 8 bits/muestra (aunque en la recomendación G.711.1 ya se han estandarizado frecuencias y tasas binarias más altas [64]). Lo anterior genera una señal digital binaria de 64 Kbps que es una velocidad estándar para transmisión digital de voz. A la conversión analógica-digital realizada según el estándar G.711, también se la conoce con el nombre de codificación de forma de onda, pues cada palabra binaria representa la amplitud de la señal de voz en cada instante muestreado.

2.3. Codificación y Transmisión

Si bien en algunos tipos de redes la transmisión de la señal de voz se hace a la tasa binaria generada con el codificador G.711, esta tasa implica el uso de un elevado ancho de banda, y dado el costo de éste, hoy es muy común el uso de codificadores de fuente que reducen de forma importante la tasa binaria a la cual finalmente se transmite la voz. De esta forma, existen diferentes tipos de codificadores dependiendo el tipo de red a utilizar, en los cuales la eficiencia de codificación es muy alta, obteniendo tasas desde 2400 bps [157][56]. El flujo digital binario producto de esta codificación de fuente es comúnmente llamado *bitstream*.

No obstante, el uso de tasas muy bajas implica una distorsión apreciable en la señal de voz y por tanto, una disminución tanto de la satisfacción del cliente de telefonía, como del rendimiento de un sistema de RAH [36]; por ello resulta más común el uso de codificadores a tasas medio altas, que garantizan una mejor calidad en la señal reconstruida.

Por otro lado, dependiendo del tipo de red utilizada, se añade o no una protección adicional del flujo binario (bitstream). Esto se realiza mediante la denominada *Codificación de Canal*, la cual busca proteger la integridad de los datos contenidos en el bitstream, de las diferentes adversidades presentes en el canal de comunicaciones. De este modo, en una red inalámbrica, el uso de codificación de canal es imperativo, pues debido las características inherentes de este tipo de red, se presenta un amplio abanico de distorsiones, desvanecimientos, multitrayectos, interferencias, ruido, etc., que deterioran de forma significativa la calidad de la señal. En el receptor, una vez realizada la decodificación de canal, obtenemos nuevamente el bitstream generados por la codificación fuente.

En el caso de una red cableada, los errores más comunes se deben a la pérdida de paquetes, los cuales son generados por problemas de congestión, fallos en los equipos o en el medio de transmisión, etc. Así, en el caso de una red IP, el problema más común es el de pérdida de paquetes, especialmente dañina cuando ésta se produce en ráfagas [113].

En este punto, la señal de voz ha atravesado diversos sistemas que han introducido diferentes distorsiones, empezando por el ruido de ambiente, los efectos de la codificación y las condiciones adversas del canal (los cuales serán analizados con más detalle en la sección 4.2). Por tanto, es necesario obtener una representación óptima de ésta, que maximice la eficiencia general del sistema de reconocimiento.

2.4. Parametrización

Del bitstream obtenido en la sección anterior, debemos obtener un vector de características que contenga una representación compacta de la voz. Sin embargo, dado que el bitstream es producto del codificador de fuente, tenemos dos alternativas para hacerlo. La primera es sintetizar la voz haciendo uso del decodificador de fuente, y por tanto, obtener el vector de características a partir de la señal de voz reconstruida. La segunda consiste en obtener el vector de características directamente de los parámetros contenidos en el bitstream.

La primera opción da origen a los procedimientos tradicionales de extracción de características, y la segunda alternativa es denominada *Transparametrización*, que sirve de base para los trabajos de esta tesis, habiendo sido elegida por la robustez demostrada frente a los diferentes tipos de distorsión introducidas por redes IP y GSM [113, 81, 61, 52]. Las dos alternativas serán descritas con detalle en los capítulos siguientes.

Independientemente del procedimiento para obtener el vector de características, éste debe contener una representación con las propiedades más relevantes de la señal de voz, de tal forma que facilite la tarea de reconocimiento.

Para poder asumir un comportamiento estacionario, cada vector de características es extraído a intervalos regulares de tiempo, observando una ventana temporal corta. Dado que la voz puede considerarse aproximadamente estacionaria para intervalos inferiores a 30 ms, el análisis espectral utilizado para producir este vector, se suele hacer sobre ventanas de 10, 20 o 30 ms, siendo lo más habitual hacerlo cada 10 ms.

Existen diferentes procedimientos para representar de forma compacta la señal de voz para propósitos de reconocimiento; uno de los más habituales es el Cepstrum [21], pues resulta muy robusto frente a diversos tipos de distorsiones. De este proceso, se obtienen los denominados coeficientes cepstrales, a los cuales usualmente se les acompaña con la energía de trama. A este primer conjunto de características acústicas se le denomina comúnmente *Parámetros Estáticos*. Además, es común añadir a este vector de características estáticas, otro conjunto de parámetros denominados *Parámetros Dinámicos*, calculados como las primeras y segundas derivadas de la evolución temporal de los parámetros estáticos (Véase la Sección 3.4.5).

Finalmente, con los anteriores parámetros se constituye el vector de características que representa cada trama de voz y que será el utilizado por el sistema de reconocimiento.

2.5. Reconocimiento

Una vez se tiene el vector de características, se procede a realizar la tarea de reconocimiento. A continuación se hará una breve descripción de los diferentes tipos de sistemas de reconocimiento en función de la tarea.

2.5.1. Fundamentos del Reconocimiento Automático de Habla

El problema de reconocimiento se puede plantear en términos estadísticos como la determinación de la palabra (o secuencia de palabras) $W = [W_1, \cdots, W_M]$ que corresponde a una secuencia de observaciones acústicas $X = [X_1, \cdots, X_N]$ [88].

Por tanto, si aplicamos la regla de decisión de Máximo A Posteriori (MAP), la secuencia de palabras W reconocida a partir de las observaciones X será:

$$\hat{W} = \arg \max_{\omega} P(W|X) \tag{2.1}$$

siendo $P(W|X)$ la probabilidad de la secuencia de palabras W dada la observación X. Y dado que:

$$P(W|X) = \frac{P(X|W)P(W)}{P(X)} \tag{2.2}$$

podemos expresar la Ecuación (2.1) como:

$$\hat{W} = \arg \max_{\omega} P(X|W)P(W) \tag{2.3}$$

Donde $P(X|W)$ se obtiene del Modelo Acústico, y $P(W)$ es la probabilidad a priori de W proporcionada por el Modelo de Lenguaje.

En este proceso existen dos problemas por resolver [119][31]:

1. Estimación o entrenamiento: se obtienen los valores óptimos para los parámetros de cada modelo. En concreto, si el modelo es un HMM (*Hidden Markov Model*) [120] se deben obtener las probabilidades de transiciones entre estados, así como las probabilidades iniciales y las probabilidades de emisión. El algoritmo utilizado es el de Baum-Welch.

2. Decodificación: busca la secuencia de palabras más verosímil resolviendo la Ecuación (2.3). Para ello $P(X|W)$ se calcula teniendo en cuenta todas las secuencias posibles de estados:

$$P(X|W) = \sum_{s} P(X, S|W) \tag{2.4}$$

o más habitualmente, de forma aproximada mediante el algoritmo de Viterbi [42][93]:

$$P(X|W) \approx \max_{s} P(X, S|W) \tag{2.5}$$

2.5.2. Modelo Acústico y Modelo de Lenguaje

En un sistema de RAH se deben realizar dos tipos de modelado: Acústico y de Lenguaje. El modelado acústico consiste en utilizar un conjunto de locuciones de una determinada unidad acústica, con el fin de obtener el mejor modelo que la represente. Para ello, se debe buscar el mayor número de locuciones de dicha unidad acústica, para conseguir una representación estadística que nos permita establecer los parámetros óptimos del modelo.

Por otro lado, el modelado del lenguaje hace uso de las reglas gramaticales de un determinado lenguaje para establecer las posibles combinaciones de palabras o unidades acústicas que serán reconocidas. Es por esto que en este modelo se tienen en cuenta tanto la sintaxis como la semántica del lenguaje.

Existen diferentes aproximaciones al problema del modelado de un sistema de reconocimiento, sin embargo la forma más extendida es la de representación estadística mediante Modelos Ocultos de Markov (Hidden Markov Models - HMM), debido, entre otras ventajas, a su facilidad a la hora de modelar locuciones con distinta duración, o su flexibilidad para ser combinados con otros HMM [109]. De otro lado, es importante destacar que en un HMM se puede incorporar conocimiento a priori, y conseguir con ello, un incremento en el desempeño del reconocedor [119][17].

Capítulo 3

Codificación y Reconocimiento de Voz: Similitudes y Diferencias

3.1. Introducción

Debido a la creciente demanda de aplicaciones de reconocimiento de voz codificada, se han explorado diversas maneras de brindar soluciones adecuadas a dicho entorno. Así el primer problema con el que nos encontramos, es la diferencia entre los procedimientos usados para codificación y los que se utilizan en reconocimiento, pues el objetivo de cada procedimiento es de naturaleza diferente, sin embargo, también hay coincidencias. Por ello, a continuación realizaremos una descripción de los dos procedimientos atendiendo a sus similitudes y diferencias. No se pretende exponer de forma exhaustiva esta descripción, sino introducir los conceptos clave que intervienen en el problema que nos hemos planteado.

Para empezar, realizaremos una exposición de los procesos comunes que se pueden utilizar para llegar a una convergencia entre la codificación de fuente y la parametrización para reconocimiento. En concreto, como veremos más adelante, podemos destacar la relación existente entre los parámetros usados para modelar la información espectral en codificación, con sus similares en reconocimiento. En codificación son muy usuales los LSP (*Line Spectrum Pairs*) [82] y en reconocimiento robusto, típicamente se utilizan los MFCC (*Mel-Frequency Cepstral Coefficients*) [72]).

3.2. Modelo Fuente Filtro

El modelo Fuente-Filtro, es una representación compacta del Sistema de Producción de Voz Humano mostrado en las Figuras 3.1 y 3.2. En éste, el aire generado en los pulmones y que pasa, primero a través de las cuerdas vocales en la laringe, y después a la faringe, produce una señal acústica que resuena en las cavidades del tracto vocal y nasal. Por tanto, para obtener una representación físico-matemática de éste, se utiliza el modelo fuente-filtro de la Figura 3.2. Este modelo es comúnmente utilizado en diferentes procesos de análisis de señales de voz. La idea básica es por un lado, simular el comportamiento del tracto vocal humano a través de un filtro lineal variable en el tiempo, y por otro lado modelar la fuente

de energía acústica generada en la laringe, a través de una señal llamada *excitación*. La convolución de la excitación $e[n]$ con la respuesta del filtro $h[n]$, produce la señal de voz sintetizada $s[n]$ (ver Ecuación (3.1)), es por esto que a este filtro se le conoce también como *Filtro de Síntesis*. Por otro lado, la fuente o excitación asumida con espectro plano, se puede modelar a su vez, utilizando una componente periódica y una aperiódica, como lo muestra la Figura 3.2[90]. La aparición de este modelo se desprende de los estudios realizados por Johannes Müller en 1848 [99], Fant en 1960 [39] (re-impreso en 1970 [40]) y Lieberman en 1984 [89][90].

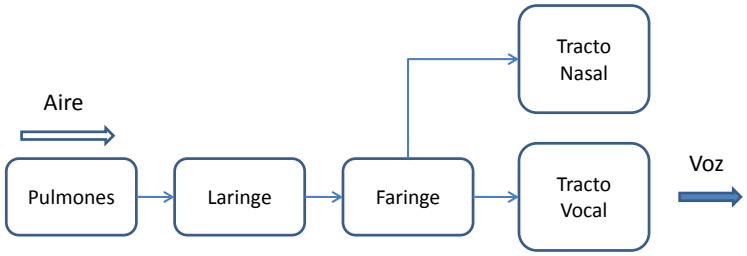


Figura 3.1: Sistema de Producción de Voz Humano.

Una de las suposiciones utilizadas para este modelo, es la de independencia entre la fuente y el filtro, por lo que podemos deconvolucionar las dos componentes para utilizarlas en os so *ıte* *es* ón se la pri ıás gr os pe la Fi

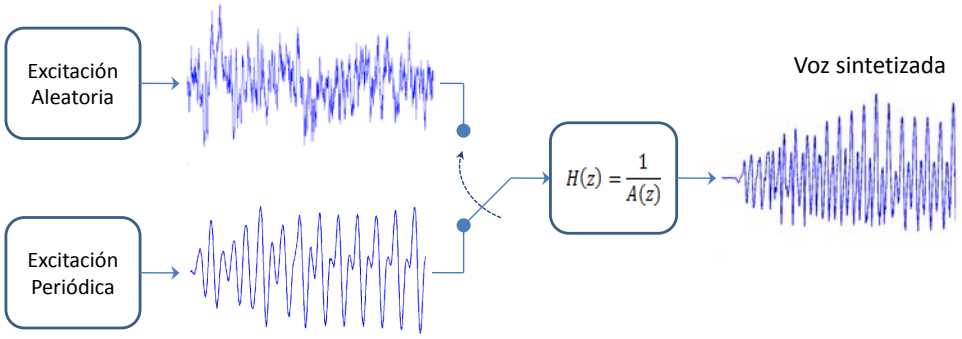


Figura 3.2: Modelo Fuente Filtro.

Para el cálculo de los parámetros del modelo, se utiliza el conocimiento fonético del sistema de producción de voz humano, y la explotación de sus propiedades, depende del objetivo a cumplir. Por ejemplo, en síntesis y codificación, se busca obtener una señal sintetizada que perceptualmente sea lo más parecida posible a la voz humana, para lo cual es muy importante modelar tanto la envolvente espectral como la excitación. Sin embargo en reconocimiento, lo que se busca es que la parametrización obtenida, permita una clara identificación de cada unidad acústica a reconocer, para lo cual se suele utilizar solo la información de la envolvente espectral, aunque no en todos los casos, como veremos en el Capitulo 5.

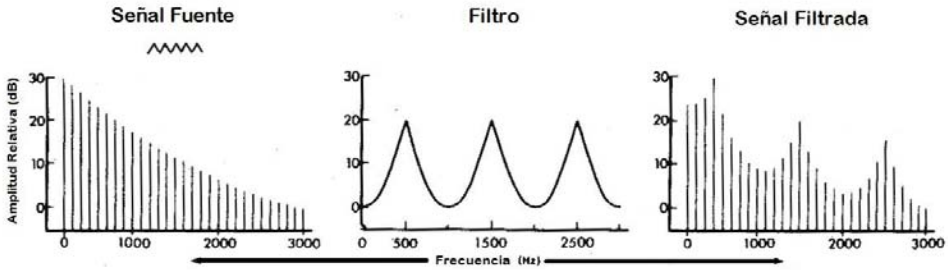


Figura 3.3: Efecto espectral del Modelo Fuente Filtro.

3.2.1. Características de la fuente

La fuente generadora de energía acústica en el sistema de producción de voz humano puede tener diferente naturaleza en función de los órganos que intervengan y de la forma en que el flujo de aire impacta sobre ellos. Así podemos tener los siguientes tipos [92]:

- Periódica: cuando se produce vibración de las cuerdas vocales.
- Aperiódica continua: cuando se produce por fricción del aire.
- Aperiódica impulsional: cuando hay explosión.
- Mixta: efecto combinado de las anteriores.

Sin embargo, para abreviar el modelo, la señal de excitación se suele calcular solo como la combinación de una componente periódica y una aperiódica:

Componente periódica

También llamada componente determinística, puede ser construida utilizando una combinación de señales periódicas, tal como un tren de impulsos en el tiempo (equivalente a una combinación de armónicos en el dominio frecuencial), o una señal real periódica modificada en amplitud a través de una ganancia adaptativa.

Componente aperiódica

También llamada estocástica, pues se asume que tiene un comportamiento aleatorio, usualmente considerada como una variable aleatoria gausiana. Puede ser modelada como ruido, adaptando su media y varianza, o a través de un banco de señales previamente construidas.

3.2.2. Características del filtro

Su función es modelar el tracto vocal, moldeando los valles y las resonancias que se producen en el espectro cuando se genera un sonido. Este filtro, modifica la respuesta en frecuencia de la fuente, acentuando su magnitud en las regiones de los formantes y atenuando su respuesta en los valles. El filtro utilizado más comúnmente es todo polos, cuyos coeficientes son obtenidos a través de un *análisis de predicción lineal*[94], que busca minimizar el error cuadrático medio entre la señal original y la señal estimada.

3.2.3. Deconvolución de las componentes del modelo

Como ya se mencionó, una de las propiedades más importantes del modelo fuente-filtro, es la de asumir la independencia de la fuente y el filtro, por tanto, es posible separarlas (utilizando un método de deconvolución) para utilizarlas a conveniencia en función de la tarea que deseamos realizar. De este modo, surgen dos alternativas para ejecutar la separación. La primera se basa en un modelo de predicción lineal, que obtiene por un lado la señal de excitación como residuo de predicción, y por otro lado los coeficientes del denominado filtro de síntesis. La segunda opción separa la información de la fuente y el filtro en el dominio denominado *cuefrencial*, utilizando el método de *Deconvolución Homomórfica*. A continuación se explicarán las dos técnicas.

Análisis de Predicción Lineal

El análisis de Predicción Lineal (Linear Prediction - LP) se basa en la idea de que la voz puede ser modelada utilizando un sistema de predicción lineal. De esta forma, la señal de voz $s[n]$ puede ser reconstruida sumando su predicción $\hat{s}[n]$, con el error de predicción $e[n]$.

$$s[n] = e[n] + \hat{s}[n] \tag{3.2}$$

donde la predicción $\hat{s}[n]$ es expresada como una combinación lineal de p muestras pasadas de $s[n]$, utilizando los coeficientes de predicción a_k:

$$\hat{s}[n] = \sum_{k=1}^{p} a_k s[n-k] \tag{3.3}$$

Transformando las anteriores ecuaciones al dominio z:

$$S(z) = E(z) + \sum_{k=1}^{p} a_k S(z) z^{-k} \tag{3.4}$$

Por tanto:

$$S(z) = E(z)\,\frac{1}{A(z)} \tag{3.5}$$

donde:

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{3.6}$$

Por otro lado, si llevamos al dominio z la Ecuación (3.1) del modelo fuente-filtro, obtenemos una ecuación similar a la Ecuación (3.5).

$$S(z) = E(z)H(z) \tag{3.7}$$

donde:

$$H(z) = \frac{1}{A(z)} \tag{3.8}$$

Por tanto, la función de transferencia del filtro que modela el tracto vocal $H(z)$ es caracterizado por el polinomio y coeficientes del predictor lineal de la Ecuación (3.6).

Una vez se ha obtenido $H(z)$, podemos conseguir $E(z)$, y así deconvolucionar $h[n]$ y $e[n]$ de la Ecuación (3.1).

Este análisis de predicción lineal será ampliado en la sección 3.3.1 en el contexto de la codificación de voz.

Deconvolución Homomórfica

Otra forma de obtener la deconvolución de $h[n]$ y $e[n]$ de la Ecuación (3.1), es mediante el método conocido como *Deconvolución Homomórfica*, que describiremos a continuación. Sin embargo, antes debemos exponer los fundamentos del *Análisis Cepstral*, que nos llevará al dominio en el que se realiza la separación de la fuente y el filtro.

Análisis Cepstral: El análisis cepstral se puede ver como un conjunto de técnicas de procesamiento digital, que utilizan una transformación no lineal a una determinada secuencia $x[n]$, para obtener su Cepstrum $\tilde{x}[n]$. En nuestro caso de estudio, la transformación utilizada es la siguiente:

$$\tilde{x}[n] = \mathscr{Z}^{-1}\Big\{ log\Big(\mathscr{Z}\{x[n]\}\Big)\Big\} \tag{3.9}$$

donde $\mathscr{Z}$ y $\mathscr{Z}^{-1}$ es la transformada Z y su inversa.

Y si aplicamos la anterior fórmula a la Ecuación (3.1), obtenemos:

$$\tilde{s}[n] = \mathscr{Z}^{-1}\Big\{ log\Big(\mathscr{Z}\{e[n] * h[n]\}\Big)\Big\} \tag{3.10}$$

Y por las propiedades del logaritmo y la Transformada Z, $\tilde{s}[n]$ la podemos expresar como:

$$\tilde{s}[n] = \mathscr{Z}^{-1}\Big\{ log(E(z)) + log(H(z))\Big\} \tag{3.11}$$

$$\tilde{s}[n] = \mathscr{Z}^{-1}\Big\{ log(E(z))\Big\} + \mathscr{Z}^{-1}\Big\{ log(H(z))\Big\} \tag{3.12}$$

Por tanto:

$$\tilde{s}[n] = \tilde{e}[n] + \tilde{h}[n] \tag{3.13}$$

donde $\tilde{e}[n]$ y $\tilde{h}[n]$ son los Cepstrum de $e[n]$ y $h[n]$ respectivamente.

No obstante, la intervención del logaritmo en la transformada, hace que el dominio de $\tilde{s}[n]$ no sea el tiempo, sino el denominado *Dominio Cuefrencial*, y es en éste en donde se puede hacer la separación de $e[n]$ y $h[n]$, pues podemos obtener $\tilde{h}[n]$ tomando las primeras cuefrencias de $\tilde{s}[n]$ [130]. En concreto, es generalizado el uso de las primeras 12 cuefrencias para obtener una buena representación del Cepstrum de $h[n]$. De este modo, se consigue separar $e[n]$ y $h[n]$ utilizando la Deconvolución Homomórfica.

Este método de deconvolución será estudiado con más detalle en la Sección 3.4 dedicada a la parametrización en RAH.

3.3. Parametrizaciones para Codificación

Como se expuso en el Capitulo 2, el uso de la codificación de fuente es imperativo en las redes de comunicación actuales, y en particular, en el caso de comunicación de voz, la codificación fuente permite un ahorro significativo en el ancho de banda de transmisión.

En la actualidad, la codificación de voz es usada principalmente en los ámbitos de telefonía móvil y voz sobre IP (VoIP), y dado que este tipo de codificadores, utilizan el modelo de producción de voz humano expuesto en la Sección 3.2. En el proceso de codificación se debe obtener un conjunto de parámetros que caractericen tanto la información del filtro como la de excitación, para luego reconstruir la voz en el proceso de decodificación. Así, desde el punto de vista de la parametrización, la codificación de voz persigue fundamentalmente dos objetivos:

- Obtener parámetros que modelen la voz con la menor pérdida de calidad perceptual posible.

- Reducir la tasa binaria requerida para la transmisión de los parámetros obtenidos, y con ello optimizar el uso del ancho de banda.

Existen otros aspectos importantes que se deben de tener en cuenta en un sistema de codificación óptimo, tales como la complejidad y el coste computacional, el consumo de energía, el retardo algorítmico, robustez, etc., sin embargo, ahora nos centraremos en las características que consiguen una alta calidad perceptual, pues son las parametrizaciones que resultan de más interés en esta tesis.

Actualmente, la mayoría de los codificadores que operan en el rango de tasas binarias bajas (requeridas para aplicaciones de transmisión inalámbricas e IP), son codificadores tipo CELP (Code Excited Linear Prediction) [65][67][68][66][4]. Éstos modelan la envolvente espectral realizando un análisis de predicción lineal, con el cual obtienen los denominados *Coeficientes de Predicción Lineal* (*Linear Prediction Coefficients - LPC*). Sin embargo, los LPC no son muy adecuados para ser transmitidos, y el codificador los transforma en LSP (Line Spectrum Pairs), los cuales, entre otras ventajas, permiten ser cuantificados e interpolados de forma más eficiente, al igual que facilitan el análisis de estabilidad del filtro que caracterizan.

Dado que la información contenida en estos parámetros es primordial para la reconstrucción de la voz, éstos suelen ser codificados utilizando un alto porcentaje de bits, respecto del total de bits asignados para cada trama de voz codificada. Más aún, cuando existe codificación de canal, los LSP suelen tener asignada la más alta prioridad para ser protegidos [58].

Por otro lado, como se verá en la Sección 3.4 la información de la envolvente espectral contenida en los LSP, es la base para la parametrización de reconocimiento de voz codificada. Por este motivo en la parte final de esta sección, centraremos nuestro análisis en este tipo de parámetros.

3.3.1. Codificación CELP

El codificador CELP es un algoritmo de codificación de voz propuesto por Schroeder y Atal en 1985 [128], y que está enmarcado en lo que se denomina *Codificación Híbrida*. Aunque en la actualidad el término CELP es utilizado para todo un conjunto de codificadores que mantienen las características básicas del algoritmo inicial. Es de destacar, que el CELP es un codificador orientado a tasas medias-bajas de funcionamiento y por ello presenta una alta eficiencia de codificación. No obstante, a pesar de conseguir reducciones importantes en la tasa binaria, mantiene una alta calidad subjetiva de la voz. Es de anotar que tanto el codificador G.729 como los codificadores AMR-NB utilizados en esta tesis pertenecen a esta clase.

En términos generales, para lograr el objetivo de comprimir la voz, un codificador se basa en dos operaciones intrínsecas:

- Eliminar la redundancia.

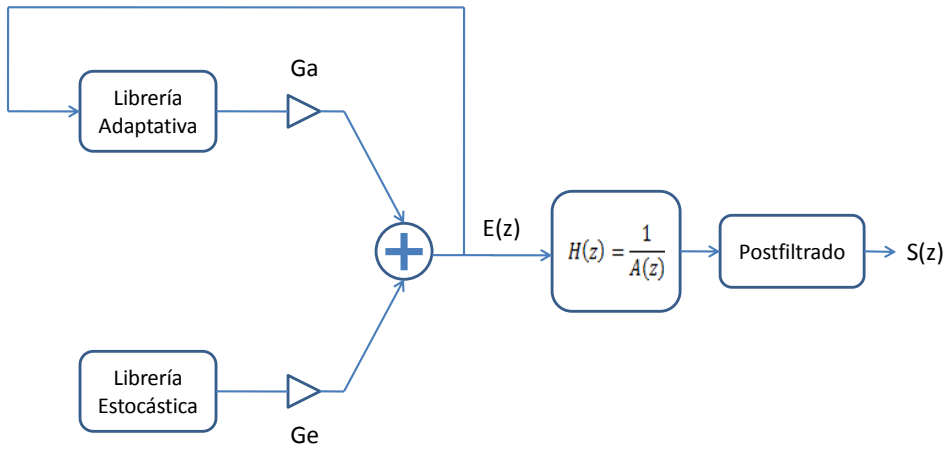


Figura 3.4: Elementos de un Codificador CELP.

- Eliminar la irrelevancia.

Para conseguir lo primero, se realizan diferentes procesos tales como el análisis LPC que logra eliminar algunas componentes redundantes en la señal, quitando la alta correlación existente en las muestras de voz. Por otro lado, para eliminar la irrelevancia, se cuantifican los parámetros obtenidos en el modelado Fuente-Filtro, de tal forma que se obtenga una representación compacta de ellos intentando eliminar solo la información irrelevante. Uno de los métodos con mayor eficiencia es el de *Cuantificación Vectorial* [84].

Como ya se mencionó en la sección anterior, un codificador CELP realiza un análisis LPC (llamado de corto plazo), para modelar la información de la envolvente espectral. De este proceso se obtienen los coeficientes de predicción lineal (LPC) que caracterizarán el filtro de síntesis.

De otro lado, el modelado de la excitación se realiza a partir del residuo obtenido en el análisis LPC. Este residuo se construye con la suma de una componente periódica y una no periódica. En el caso de la representación periódica (determinística), se utiliza un análisis de largo plazo que obtiene el período o pitch, que junto con el cálculo de la ganancia (G_a) se construye la denominada *Librería Adaptativa*. La componente aperiódica (o estocástica), se modela utilizando una ganancia (G_a) y una *Librería Estocástica*.

Finalmente, la voz sintetizada se obtiene sumando las dos componentes de la excitación, y pasando ésta a través del filtro de síntesis de corto plazo (véase la Figura 3.4).

Es de destacar que en el CELP se utiliza un procedimiento llamado de *Análisis por Síntesis* para obtener una secuencia óptima de la excitación. En este procedimiento, el error de predicción entre la voz sintetizada y la original, se minimiza de acuerdo a una medida de distorsión ponderada perceptualmente.

Por tanto, las etapas que se pueden distinguir en el proceso de codificación CELP son las siguientes:

- Pre-procesamiento.
- Enventanado
- Análisis de Corto Plazo.
- Ponderación Perceptual
- Análisis de la Excitación.
- Postfiltrado.

Preprocesado

Antes de realizar el procesamiento que parametriza el modelo fuente-filtro, se utiliza una etapa de preprocesado en la cual se depura la señal para robustecer el proceso de codificación y eliminar componentes no deseadas. Así, en esta etapa se realiza un escalado de la señal para reducir su margen dinámico y evitar posibles errores en la cuantificación. A continuación se utiliza un filtro paso alto, para suprimir componentes de baja frecuencia, tales como ruido de línea (inducido por las líneas eléctricas de 50 o 60 Hz) o la componente de corriente continua. Lo anterior se realiza a través de un filtro como el descrito por la Ecuación (3.14).

$$H_{prep}(z) = \frac{0,46363718 - 0,92724705z^{-1} + 0,46363718z^{-2}}{1 - 1,9059465z^{-1} + 0,9114024z^{-2}} \tag{3.14}$$

En este caso, el escalado es de $1/2$ y la frecuencia de corte es de 140 Hz [68].

Enventanado

Para realizar el cálculo de todos los parámetros del codificador, primero se debe realizar una segmentación de la señal en tramos (o tramas) que usualmente tienen un tamaño de 10 ms (80 muestras cuando se ha muestreado a 8 KHz). Durante este tiempo la señal de voz se considera de naturaleza estacionaria y por tanto se mantienen aproximadamente constante su envolvente espectral. Sin embargo, para obtener los parámetros de cada trama se utiliza una ventana llamada de *análisis* como la mostrada en la Figura 3.5 [68].

Un tamaño de ventana grande (típicamente de 30 ms) permite conseguir una mayor resolución en el dominio de la frecuencia; y por otro lado, introduce una alta correlación entre los parámetros obtenidos entre una trama y sus aledañas. Esta correlación permite cumplir, entre otros objetivos, suavizar la transición entre parámetros obtenidos para una trama y la siguiente, lo que lleva a obtener una mejor calidad perceptual en la voz sintetizada. También es importante destacar, que dadas las restricciones de ancho de banda asumidas para este codificador, es necesario prestar especial atención a los métodos de cuantificación, siendo muy comunes los métodos de cuantificación diferencial, muy efectivos cuando los parámetros a cuantificar tienen una alta correlación entre tramas consecutivas; o vectoriales,

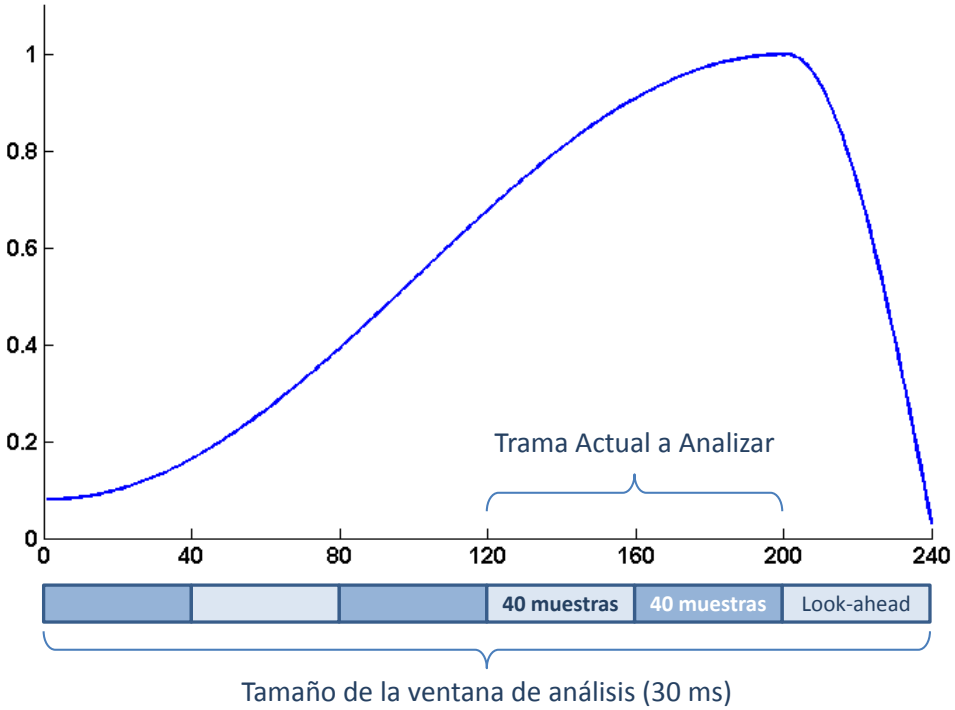


Figura 3.5: Ventana de Análisis.

que permiten representaciones más compactas.

No obstante las ventajas expuestas, el uso de una ventana de análisis es también causa de retardo, pues para el análisis de una determinada trama, se utilizan muestras correspondientes a tramas futuras (*look-ahead*). Sin embargo si se utiliza una ventana asimétrica como la mostrada en la Figura 3.5 se puede disminuir dicho retardo reduciendo el look-ahead a 5 ms, en este caso.

Análisis de Corto Plazo

Después del proceso de enventanado, se procede a caracterizar la envolvente espectral (véase la Figura 3.6). Esta se obtiene a partir de un análisis LPC denominado de corto plazo (como el que se describió en la Sección 3.2.3). En este análisis, se calculan p coeficientes (usualmente 10) de predicción una vez por trama (80 muestras).

Con los p coeficientes se construye el filtro de síntesis todo polos cuya función de transferencia está dada por la ecuación:

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.15}$$

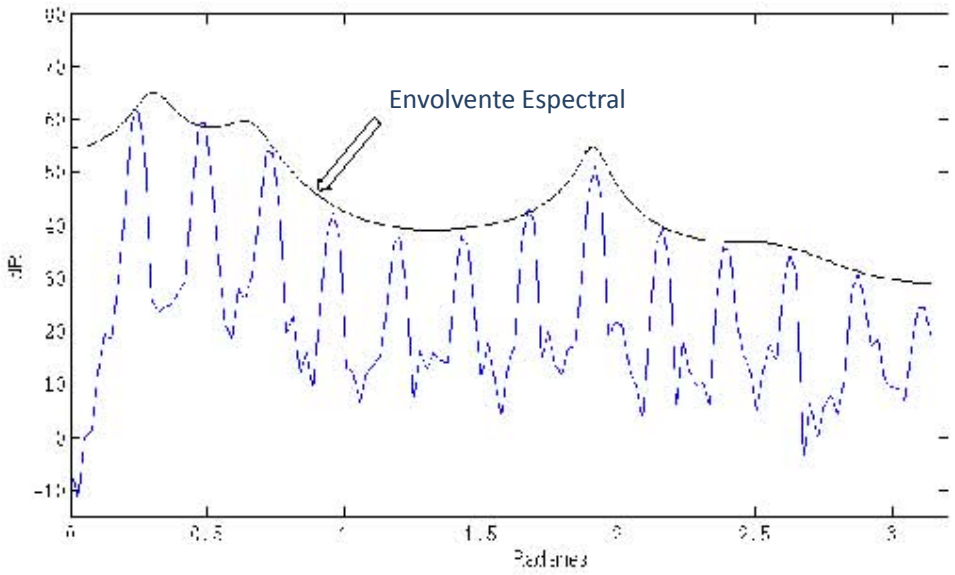


Figura 3.6: Envolvente Espectral en una Señal de Voz Sonora.

que expande la Ecuación (3.8).

donde a_k son los coeficientes de predicción lineal. De la Ecuación (3.2), el error de predicción $e[n]$ se puede expresar como:

$$e[n] = s[n] - \hat{s}[n] \tag{3.16}$$

Clásicamente se utilizan dos métodos para el cálculo de los coeficientes a_k: el Método de la Covarianza o el Método de Autocorrelación, sin embargo el método de covarianza no da garantía de estabilidad en el filtro, mientras que el método de autocorrelación si lo hace y por tanto es el más utilizado para el cálculo de los LPC del filtro de síntesis.

Para el cálculo por el método de autocorrelación, debemos minimizar el error cuadrático medio de la predicción (*Mean Square Error - MSE*) definida como:

$$MSE = E\{e^2[n]\} = E\left\{\left[s[n] - \sum_{k=1}^{p} a_k s[n-k]\right]^2\right\} \tag{3.17}$$

Por tanto, si derivamos con respecto a los coeficientes a_j:

$$\frac{\partial MSE}{\partial a_j} = E\left\{\left[s[n] - \sum_{k=1}^{p} a_k s[n-k]\right]\ s[n-j]\right\} = 0 \quad para\ \ j = 1, ..., p \tag{3.18}$$

Podemos calcular los coeficientes de autocorrelación que luego serán convertidos en coeficientes LP (LPC) utilizando el algoritmo de Levinson-Durbin. Este procedimiento no será detallado en esta tesis, por ser un algoritmo ampliamente explicado en la literatura existente [84].

Transformación de LPC a LSP, Cuantificación e Interpolación de los LSP: Una vez obtenidos los LPC, estos deben ser transformados en LSP (Linear Spectral Pairs) para ser cuantificados e interpolados. Sin embargo, dada la importancia de los coeficientes LPC y LSP en desarrollo de esta tesis, en la Sección 3.3.2 será descrita con detalle esta transformación.

Por otro lado, la información de la envolvente espectral codificada en los parámetros LSP, no es transmitida directamente al canal, pues primero los LSP deben ser cuantificados e interpolados para obtener una representación más compacta de estos. Por tanto, los LSP obtenidos de los LPC se cuantifican utilizando *Cuantificación Vectorial* para aprovechar la relación intra-trama presente en ellos. Sin embargo, dado que el cálculo de los LSP se hace una vez por cada trama, y el cálculo de los parámetros de la excitación se hace una vez por cada subtrama (ver Sección 3.3.1); los LSP sin interpolar obtenidos en el análisis de corto plazo, suelen ser utilizados para la síntesis de la segunda subtrama, y por ello se utiliza una versión interpolada de éstos (correspondientes a la trama actual y a la trama previa) para la primera subtrama.

Transformación de LSP a LPC y verificación de estabilidad: Finalmente los LSP cuantificados e interpolados, son transformados nuevamente a LPC para construir con ellos el filtro de síntesis $\hat{a}_i$:

$$\hat{H}(z) = \frac{1}{\hat{A}(z)} = \frac{1}{1 + \sum_{10}^{i=1} \hat{a}_i z^{-k}} \tag{3.19}$$

Para verificar la estabilidad del filtro de síntesis construido, se utilizan los LSP cuantificados, atendiendo al hecho de que los picos en la envolvente espectral suceden cuando dos LSP están muy próximos. Por tanto, para prevenir resonancias indeseadas en la decodificación, se utiliza un algoritmo que garantiza una separación mínima entre LSP consecutivos, de tal forma que no se produzca inestabilidad del filtro.

Ponderación Perceptual

Debido al efecto de enmascaramiento frecuencial presente en el oído humano, en las zonas de máxima energía (formantes) se produce un enmascaramiento del ruido mayor que en los valles. Por tanto, es necesario modificar las características frecuenciales de la señal a sintetizar, concediendo más importancia al modelado de las zonas de frecuencia en las que el oído es más sensible y menos importancia a las zonas menos sensibles.

De esta manera, el filtro llamado de *Ponderación Perceptual*, pretende resaltar las zonas frecuenciales más sensibles de la envolvente espectral, para que sean mejor modeladas por el codificador. Para conseguir este objetivo, se construye el filtro de ponderación perceptual $W(z)$, utilizando los mismos coeficientes no cuantificados a_k del filtro de síntesis, aunque utilizando por un factor de escala γ, como lo muestra la Ecuación (3.20):

$$W(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k (z/\gamma)^{-k}} \tag{3.20}$$

Análisis de la Excitación

Durante este análisis, se busca modelar la excitación tanto en su componente determinística como en la estocástica, utilizando para ello librerías adaptativa y estocástica. Dichas librerías utilizan un índice identificativo para cada código que es luego transmitido (en lugar del código) para obtener una representación más compacta de los parámetros del codificador.

La componente periódica se modela caracterizando la librería adaptativa (también se le conoce como *Vector de Códigos Adaptativos*) $V_p[n]$ y la componente no periódica a través de la librería estocástica (o *Vector de Códigos Fijos*) $V_c[n]$ (véase la Figura 3.4).

$$u[n] = G_p u[n - N_0] + G_c V_c \tag{3.21}$$

donde $u[n]$ es la excitación estimada; G_p y G_c las ganancias de las librerías adaptativa y estocástica respectivamente y T_0 es el periodo fundamental de la componente periódica. Estos parámetros de la excitación se calculan una vez por cada subtrama.

Para obtener los parámetros de la componente adaptativa se realiza primero un análisis en bucle abierto por cada trama, cuyo objetivo es la búsqueda de un candidato del periodo fundamental (T_0). A continuación, un análisis en bucle cerrado realizado para cada subtrama, busca el mejor índice de la librería adaptativa y su respectiva ganancia. Dicha búsqueda se realiza alrededor del período fundamental encontrado en el análisis de bucle abierto.

La componente estocástica a modelar se obtiene eliminando la componente adaptativa de la excitación, para luego proceder a la búsqueda del índice de la librería estocástica y su ganancia. La combinación óptima se busca en la librería estocástica, de tal forma que se minimice el MSE (Mean Square Error) entre la señal ponderada original y la señal ponderada reconstruida.

Postfiltrado

El postfiltro, como se explicará a continuación, busca aumentar la calidad subjetiva de la voz sintetizada, especialmente bajo condiciones de ruido y bajas tasas de codificación. La siguiente descripción, corresponde al codificador G.729 [125], sin embargo los procedimientos aquí descritos son comunes a otros otros codificadores tipo CELP, objeto de nuestro estudio. Las 3 etapas que constituyen el proceso de postfiltrado son: *Postfiltrado Adaptativo*, *Control de Ganancia* y *Filtrado Paso Alto*. A continuación se detalla cada una de ellas.

Postfiltro Adaptativo: Chen et al [19], propusieron un Postfiltro Adaptativo (*Adaptive Post-Filter - APF*) para reducción de ruido en codificadores LPC de análisis por síntesis. Lo anterior se hace posible atenuando las componentes en los valles espectrales, reduciendo con ello, el nivel de ruido en las zonas del espectro más sensibles al oído humano [84]. En el codificador G.729, el APF es la primera etapa del postfiltro, y consiste en el efecto combinado de tres filtros en cascada. A continuación se describe cada uno de ellos.

Postfiltro de Largo Plazo: este filtro acentúa los armónicos de la frecuencia fundamental con el fin de atenuar los valles espectrales entre armónicos [155][19]. La función de transferencia de este filtro está determinada por la Ecuación (3.22):

$$H(z) = \frac{1}{1+\gamma_p g_l} \left(1+\gamma_p g_l z^{-T}\right) \tag{3.22}$$

donde $\gamma_p = 0,5$, T es el pitch y g_l su ganancia.

Postfiltro de Corto Plazo: esta sección del APF modela la envolvente espectral con el ánimo de reducir el efecto del ruido en los valles espectrales.

$$H_f(z) = \frac{1}{g_f}\frac{\hat{A}(z/\gamma_n)}{\hat{A}(z/\gamma_d)} = \frac{1}{g_f}\frac{1+\sum_{i=1}^{10}\gamma_n^i \hat{a}_i z^{-i}}{1+\sum_{i=1}^{10}\gamma_d^i \hat{a}_i z^{-i}} \tag{3.23}$$

donde $\gamma_n = 0,55$, T es el pitch y $g_f = \sum_{n=0}^{19} |h_f(n)|$

Filtro de compensación de pendiente: dado que la sección todo polos del filtro anterior, produce un efecto paso-bajo en la pendiente espectral, se utiliza el filtro descrito por la Ecuación (3.24) para compensar dicho efecto.

$$H_t(z) = \frac{1}{g_t}\left(1+\gamma_t k_1' z^{-1}\right) \tag{3.24}$$

donde $k_1' = -\frac{r_h(1)}{r_h(0)}$ y $r_h(i) = \sum_{j=0}^{19-i} h_f(j)h_f(j+i)$ y si k_1' es negativa, $\gamma_t = 0,9$, y si k_1' es positiva, $\gamma_t = 0,2$

El valor de g_t compensa la atenuación producida por g_f en el Postfiltro de Corto Plazo anterior.

Control de Ganancia Adaptativo: En la segunda etapa del postfiltro, se calcula la ganancia entre la señal reconstruida $\hat{s}(n)$ y la postfiltrada $sf(n)$, para que la señal postfiltrada tenga el mismo nivel de potencia de la señal reconstruida.

$$G = \frac{\sum_{n=0}^{39} |\hat{s}(n)|}{\sum_{n=0}^{39} |sf(n)|} \tag{3.25}$$

Por tanto, la señal obtenida después de esta etapa es:

$$sf'(n) = g^{(n)} sf(n) \tag{3.26}$$

donde $g^{(n)} = 0,85g^{(n-1)} + 0,15G$ con $n = 0, \ldots, 39$

Filtrado Paso-Alto: En esta última etapa, se utiliza un filtro paso-alto para eliminar componentes de baja frecuencia indeseadas en $sf'(n)$. Su frecuencia de corte es de 100 Hz.

$$H_{h2}(z) = \frac{0,93980581 - 1,8795834z^{-1} + 0,93980581z^{-2}}{1 - 1,9330735z^{-1} + 0,93589199z^{-2}} \tag{3.27}$$

Finalmente la señal filtrada se multiplica por 2 para restablecer el nivel de la señal original de entrada.

Efecto del Postfiltro

Los efectos del postfiltro en el RAH serán analizados con detalle en la Sección 8.2.2, sin embargo, de la anterior descripción se puede inferir que las etapas de filtrado previstas en el postfiltro ayudan a eliminar o por lo menos disminuir el efecto del ruido aditivo, pues de un lado el postfiltro de corto plazo inicial atenúa las componentes de alta frecuencia en la envolvente espectral y por tanto, también las componentes de ruido localizadas en la parte alta del espectro. Por otro lado, el filtro paso-alto de la última etapa, ayudaría a reducir las componentes de baja frecuencia del ruido.

3.3.2. Transformación de LPC a LSP

Como se comentó en la Sección 3.3.1, los LPC no son adecuados para ser transmitidos directamente al canal, por tanto deben ser transformados a una representación que permita cuantificarlos e interpolarlos de forma óptima. En este sentido han habido históricamente diferentes tipos de parámetros que buscaban dicha representación.

Las premisas que debían cumplir los parámetros resultantes de la transformación son entre otras: mantener la estabilidad del filtro reconstruido después de los procesos de cuantificación y transformación realizados a los parámetros LP. De otro lado, la transformación aplicada a los LPC debía ser invertible. Por último, los parámetros fruto de

la transformación debían ser robustos frente a los errores en la transmisión.

Algunas de las transformaciones planteadas a través de la historia han sido: los coeficientes de reflexión, coeficientes LAR (Logarithmic Area Ratio), y el arco-coseno de los coeficientes de reflexión. Sin embargo en los años 80, aparecen los denominados *Pares de Lineas Espectrales* o LSP (Line Spectral Pairs, por su nomenclatura en inglés), que resultan ser más eficientes a la hora de ser cuantificados (tanto escalar como vectorialmente).

La transformación en LSP, también denominados LSF (Line Spectral Frequencies) fue introducida en [62] como una representación alternativa de los LPC. Sus principales ventajas son:

- Poca sensibilidad frente al ruido de cuantificación.
- Pueden ser fácilmente interpolados.
- La estabilidad del filtro es comprobada directamente en los LSP.
- Cuantificación robusta, tanto escalar como vectorial.
- Su transformación es invertible.

Proceso de Transformación

Para obtener los parámetros LSP a partir de los LPC, partimos del hecho de que el filtro inverso $A(z)$ del análisis LPC (ver Ecuación (3.6)) satisface la siguiente recursión [4]:

$$A_k(z) = A_{k-1}(z) - r_k z^{-k} A_{k-1}(z^{-1}) \ , \quad k = 1, ..., p \tag{3.28}$$

donde p es el número de coeficientes LPC de $A(z)$, r_k es el coeficiente de reflexión de orden k y, $A_0 = 1$.

De igual modo, la recursión para $p + 1$ será:

$$A_{p+1}(z) = A_p(z) - r_{p+1} z^{-(p+1)} A_p(z^{-1}) \tag{3.29}$$

De la cual podemos obtener dos casos particulares:

$$P'(z) = A(z) + z^{-(p+1)} A(z^{-1}) \tag{3.30}$$

$$Q'(z) = A(z) - z^{-(p+1)} A(z^{-1}) \tag{3.31}$$

donde $P'(z)$ corresponde al modelo del tracto vocal cuando existe un cierre completo de la glotis ($r_{p+1} = 1$), y $Q'(z)$ cuando existe una apertura completa ($r_{p+1} = -1$). El polinomio $P'(z)$ es simétrico, mientras que $Q'(z)$ es asimétrico y puede demostrarse que $A(z)$ es de fase mínima y por tanto se puede garantizar la estabilidad de $H(z)$.

Por otro lado, las raíces de $P'(z)$ y $Q'(z)$ ocurren en pares simétricos $\pm\omega$ (de ahí el nombre de Pares de Lineas Espectrales) y se encuentran dentro de la circunferencia unidad.

Y dado que $P'(z)$ tiene una raíz en $z = -1$ ($\omega = \pi$) y $Q'(z)$ una raíz en $z = 1$ ($\omega = 0$), para eliminar estas dos raíces se definen dos nuevos polinomios:

$$P(z) = \frac{P'}{(1 + z^{-1})} \tag{3.32}$$

$$Q(z) = \frac{Q'}{(1 - z^{-1})} \tag{3.33}$$

Cada polinomio tiene $p/2$ raíces complejas conjugadas en la circunferencia unidad ($e^{\pm j\omega_i}$) que son alternadas. Por tanto los anteriores polinomios pueden ser escritos así:

$$P(z) = \prod_{i=1,3,...,p-1} (1 - 2q_i z^{-1} + z^{-2}) \tag{3.34}$$

$$Q(z) = \prod_{i=2,4,...,p} (1 - 2q_i z^{-1} + z^{-2}) \tag{3.35}$$

donde $q_i = cos(\omega_i)$ son los LSP del dominio coseno, y ω_i son los LSF en el dominio de la frecuencia. Estas últimas satisfacen la propiedad de ordenación:

$$0 < \omega_1 < \omega_2 < ... < \omega_p < \pi$$

Los detalles de esta transformación se pueden encontrar en la descripción del codificador AMR-NB [4].

3.4. Parametrizaciones para Reconocimiento

En un problema de reconocimiento de habla, desde el punto de vista del modelado acústico, unos de las etapas más importantes es la de la *parametrización*.

Del análisis hecho en la Sección 2.4, la parametrización consiste en extraer una secuencia de vectores de características a partir de las muestras de voz. Lo anterior, con el ánimo de reducir su variabilidad no lingüística y obtener así una representación compacta. Por tanto, una buena parametrización debe de un lado, capturar la información más relevante para el sistema de reconocimiento, y de otro lado, descartar la información no importante o que distraiga el proceso de reconocimiento.

Una vez obtenida una adecuada parametrización, se procede al modelado acústico basado en la información contenida en los vectores de características encontrados. Sin embargo, existen diferentes tipos de fuentes que introducen variabilidad lingüística en la voz, y que se deben tener en cuenta antes de escoger los parámetros que serán utilizados para conformar el vector de características definitivo. Entre las fuentes más importante podemos destacar:

- Variabilidad en el lenguaje: cada lengua tiene su propia entonación, además de un conjunto diferente de palabras y reglas gramaticales.

- Variabilidad Acústica: cada persona imprime un carácter propio a sus locuciones, con diferencias relativas a la tonalidad (grave o aguda), acento, etc. Además, una misma unidad acústica (fonema, palabra, etc.) es pronunciada de forma diferente dependiendo del contexto en el que se utilice.

- Variabilidad en los sistemas de adquisición (o grabación): todos los sistemas de captura (micrófonos) tienen un sistema de transducción diferente.

- Condiciones Ambientales: ruido de fondo, acústica del lugar, distancia al micrófono, etc.

Por tanto, no es fácil establecer un conjunto de parámetros óptimos para una tarea de RAH determinada, pues depende en gran medida de la tarea que se quiera modelar. Sin embargo, como veremos a continuación, la información contenida en la envolvente espectral ha sido tradicionalmente utilizada para obtener una representación compacta de la voz. No obstante, existen otros tipos de parámetros que pueden ser añadidos directamente al vector de características, siendo muy usual agregar la energía o un valor equivalente de ésta. En el Capítulo 5 se explicarán otros tipos de parámetros que pueden ser añadidos y que resultan muy efectivos para mejorar la robustez de un sistema de RAH en condiciones de ruido y/o errores de transmisión.

Los parámetros con los que se describe habitualmente la envolvente espectral, son los cepstrum o parámetros cepstrales y sus variantes. Sin embargo, como se explicará más adelante, existe una modificación de estos parámetros inspirada en el sistema auditivo humano: la escala Mel (ver Sección 3.4.2) que es también muy comúnmente utilizada.

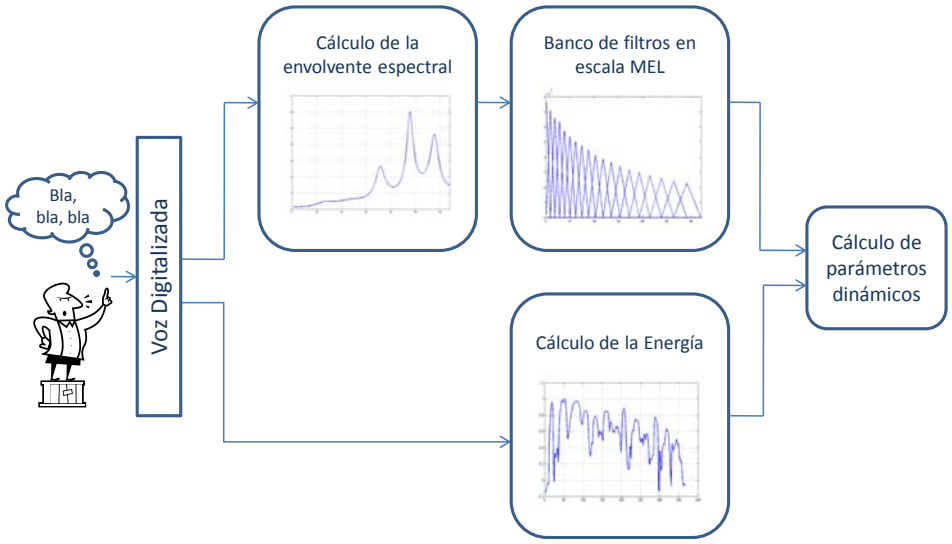


Figura 3.7: Parametrización típica de un sistema de RAH.

Teniendo en cuenta que utilizamos HMM para el modelado acústico en nuestro estudio, resulta muy útil utilizar los coeficientes cepstrales, pues se puede asumir que están incorrelados, y por tanto simplificar la estima de los parámetros de covarianza en las probabilidades de emisión de los estados del HMM.

No obstante lo anterior, es de destacar que los coeficientes cepstrales son muy sensibles a las variaciones introducidas entre locutores, y por tanto es necesario utilizar un gran número de locutores en la etapa de entrenamiento si se quiere conseguir un sistema de RAH con independencia de los mismos.

3.4.1. Extracción de la envolvente espectral

Como su nombre indica, la envolvente espectral es una curva que describe la envolvente del módulo de la respuesta en frecuencia de una señal en una ventana de tiempo dada. En este sentido, podemos ver a la envolvente espectral como una versión suavizada del espectro, pero que mantiene a su vez, la información de picos y valles tal como se ilustra en la Figura 3.6.

Por tanto, bajo un esquema fuente-filtro, la información de la envolvente espectral, estará contenida en el filtro, pues se asume, es el que contiene las variaciones más suaves en el dominio de la frecuencia.

De esta manera, retomando la Ecuación (3.1) podemos extraer $\hat{h}[n]$ a partir de $\hat{s}[n]$, utilizando alguno de los dos métodos descritos en la Sección 3.2.3 para obtener la envolvente espectral:

- Deconvolución Homomórfica, ó
- Predicción Lineal

A continuación se detalla el uso de éstos métodos, para construir el vector de características de un sistema de RAH.

Método de la Deconvolución Homomórfica

Para obtener el Cepstrum $\tilde{s}[n]$, tendríamos que seguir los pasos descritos en la Sección 3.2.3. Sin embargo, desde el punto de vista práctico se utiliza la siguiente transformación [130]:

$$\tilde{s}[n] = DFT^{-1} \log(|S[k]|) \tag{3.36}$$

donde $S[k]$ es el espectro de $s[n]$ obtenido utilizando la Transformada Discreta de Fourier (Discrete Fourier Transform - DFT), y dado que se ha utilizado el logaritmo sólo sobre el módulo de $S[k]$, el resultado es el *Cepstrum Real* de $s[n]$.

Una característica importante a destacar, es que en el cepstrum obtenido para una señal de voz $\tilde{s}[n]$, la componente que corresponde a la envolvente espectral ($\tilde{h}[n]$), decae

rápidamente en el dominio cepstral, y por tanto para obtener su cepstrum real podemos realizar un liftrado (filtrado en el dominio cepstral) de los primeros valores de $\tilde{s}[n]$. Para ello utilizamos una ventana paso-bajo $\mathscr{W}$ en el dominio cepstral, así:

$$\tilde{h}[n] = \tilde{s}[n] \cdot \mathscr{W}[n] \tag{3.37}$$

donde $\mathscr{W}$ está definida como:

$$\mathscr{W}[n] = \begin{cases} 1 & \text{si } |n| \leq n_c \\ 0 & \text{si } |n| > n_c \end{cases} \tag{3.38}$$

donde n_c es el número de *coeficientes cepstrales* (Cepstral Coefficients - CC) que serán utilizados para obtener la envolvente espectral. El valor típico de n_c es 12.

Finalmente, la envolvente espectral puede ser obtenida aplicando la DFT sobre $\tilde{h}[n]$:

$$H[k] = DFT(\tilde{h}[n]) \tag{3.39}$$

No obstante, el paso anterior no es necesario en RAH puesto que se utilizan directamente los n_c primeros valores de $\tilde{h}[n]$.

Finalmente, es de anotar que para el cálculo del cepstrum de $s[n]$, solamente se ha utilizado el módulo de su transformada de Fourier: $|S[k]|$, y por tanto, la transformación es irreversible. Es decir, en general no se puede obtener $s[n]$ a partir de su cepstrum real. Sin embargo, si $s[n]$ es de fase mínima (y por tanto su inversa es causal y estable), si que podemos obtener $s[n]$ a partir de su cepstrum. Por otro lado, si $s[n]$ es real, su cepstrum $\tilde{s}[n]$ será también real y simétrico.

Método de Predicción Lineal

En este caso la envolvente espectral se calcula del análisis LPC descrito en la Sección 3.2.3. En ésta se obtienen los coeficientes de predicción lineal a_k del filtro todo polos H de orden p definido en la Ecuación (3.15).

Y una vez obtenida $H(z)$ podemos obtener su espectro para conseguir la envolvente espectral.

Sin embargo, es posible calcular directamente el cepstrum a partir de los coeficientes de predicción lineal utilizando la siguiente recursión [9][113][61]:

$$\tilde{h}[n] = a_n + \sum_{i=1}^{n-1} (n-i) a_{n-i} \tilde{h}[i] \quad \text{para } n = 1, ..., n_s \tag{3.40}$$

donde $a_0 = 1$

Es de notar que, por este procedimiento, si $n_c > p$, es decir: sí el número de coeficientes cepstrales a calcular, es mayor que el número de coeficientes de predicción lineal, los

coeficientes cepstrales superiores a p podrán ser calculados, pero no aportaran información, pues para $n > p$, los coeficientes a_n son nulos [60].

Finalmente, dado que este procedimiento parte del análisis LPC, a este tipo de cepstrum se le denomina cepstrum LP, para diferenciarlo del anteriormente obtenido a través de la deconvolución homomórfica.

En las siguientes secciones, se explicarán los diferentes tipos de parámetros que se utilizan para conformar el vector de características que será finalmente utilizado en una tarea de RAH. Empezando por la escala Mel y el uso de bancos de filtros, que se basan en información que imita hasta cierto punto el comportamiento del *Sistema Auditivo Humano* para transformar la información espectral y moldearla para obtener una representación más apropiada y eficiente. A continuación haremos énfasis en la importancia de incluir el parámetro de energía para mejorar el rendimiento de un sistema de RAH y que jugará un papel importante en esta tesis. Finalmente, expondremos los denominados parámetros dinámicos, que introducen información de la evolución en el tiempo que resulta beneficiosa para un sistema de RAH.

3.4.2. Escala Mel

La *escala Mel* fue propuesta por Stevens, Volkman y Newmann en 1937 [135], y busca representar de una manera más fidedigna el mecanismo de percepción humano. El nombre *mel* hace referencia a la palabra *melodía*, como una manera de hacer explícito el hecho de que la escala Mel representa una escala perceptual de la altura de un sonido [18].

Para obtener una representación en la escala Mel, se utiliza una transformada no lineal en el dominio de la frecuencia. Dicho procedimiento fue obtenido de forma empírica dando como resultado la siguiente ecuación [14] [158]:

$$m(mels) = 2595 \log_{10} \left(1 + \frac{f(Hz)}{700} \right) \tag{3.41}$$

También se puede utilizar la siguiente aproximación [113]:

$$m(mels) = f + \arctan \left(\frac{0,45 sen(f(Hz))}{1 - 0,45 cos(f(Hz))} \right) \tag{3.42}$$

donde las unidades de origen de la señal a transformar están en Hz, y las de la señal transformada resultante estarán en *mels* (véase la Figura 3.8 [1]).

Cuando aplicamos la escala Mel en el cálculo de los coeficientes cepstrales expuestos en las anteriores secciones, se obtiene un mejor desempeño de las tareas de reconocimiento como se pueden observar en los trabajos realizados por [14] [113]. Sin embargo, dado que la transformación se realiza en el dominio de la frecuencia, los coeficientes espectrales deben ser transformados a esta escala antes de que se les aplique la DFT de la Ecuación (3.36).

[1] Tomada de Wikipedia

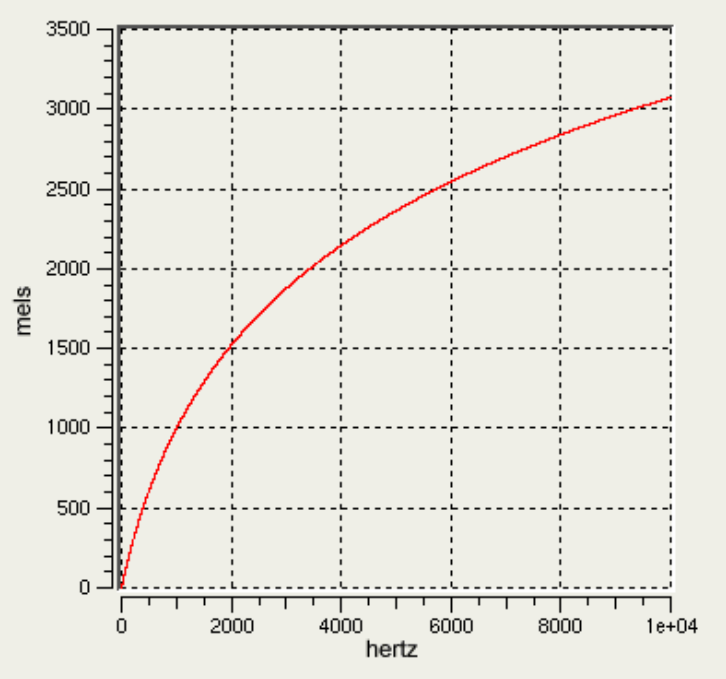


Figura 3.8: Escala Mel.

Para ello, lo habitual es utilizar *bancos de filtros* como explicaremos en la Sección 3.4.3.

Finalmente los coeficientes cepstrales transformados a la escala Mel son llamados *Coeficientes Mel-Cepstrales* o MFCC (Mel Frequency Cepstral Coefficients).

Es de destacar, que tanto en el cepstrum obtenido a través de la deconvolución homomórfica, como en el obtenido a través del análisis LPC se puede aplicar la transformación Mel, pues en ambos casos se puede calcular el espectro y por tanto es posible aplicar la escala Mel en el dominio frecuencial. Sin embargo, la transformación directa de la Ecuación (3.40) no permite el uso de esta escala.

3.4.3. Banco de Filtros

Del análisis LPC expuesto en la sección 3.4.1, podemos observar que es posible obtener los coeficientes cepstrales a partir de los coeficientes del polinomio que caracteriza el filtro de síntesis $H(z)$ utilizando la Ecuación (3.40). Sin embargo, este procedimiento no permite la transformación Mel. No obstante, existe un procedimiento alternativo para obtener el cepstrum a partir del espectro de potencia del filtro de síntesis que admite dicha transformación. A continuación explicaremos dicho procedimiento que se ilustra en la Figura 3.9.

En primer lugar, se obtiene la función de transferencia del filtro de síntesis $H(z)$ mediante análisis LPC. A continuación se calcula el módulo de la respuesta en frecuencia

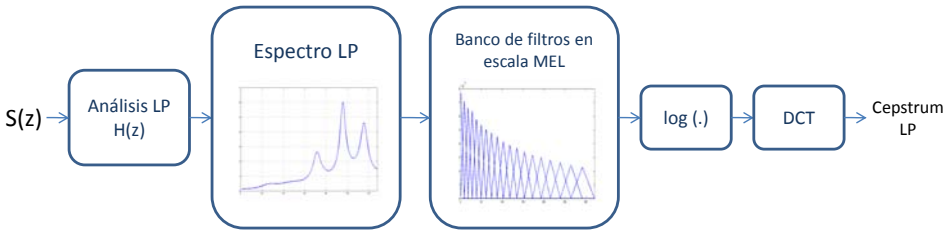


Figura 3.9: Banco de filtros para el cálculo del Cepstrum LP.

del filtro, utilizando para ello la DFT (por este motivo, el espectro obtenido es comúnmente llamado espectro LP). Sin embargo, dado que es un espectro en el dominio discreto de la frecuencia, contiene un número finito de muestras, que en general es una potencia de 2. Siendo los valores más habituales 128, 256 o 512.

Acto seguido, el espectro LP pasa a través de un banco de filtros que ponderan el espectro en bandas o canales, consiguiendo entre otros aspectos reducir el número de coeficientes que caracterizan la información espectral. Habitualmente utiliza la escala Mel para establecer las características de cada filtro, de tal manera que las bandas quedan espaciadas logarítmicamente en la frecuencia. Con ello se consigue imitar hasta cierto punto, el comportamiento del sistema auditivo humano.

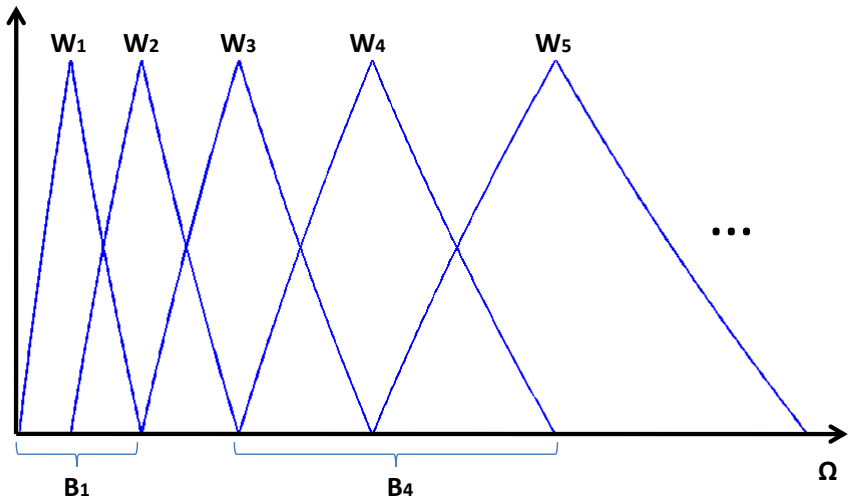


Figura 3.10: Banco de filtros en escala Mel.

La respuesta en frecuencia de cada filtro, es de forma triangular, y tienen un ancho de banda B_i.

Por otro lado, existe un solapamiento entre dos filtros consecutivos que es igual a la mitad del ancho de banda del anterior como lo indica la Figura 3.10.

La salida del banco de filtros es un conjunto de valores cuyo número es igual al número de filtros del banco (usualmente se utilizan entre 30 a 40 filtros), ya que para cada filtro se calcula la energía del espectro filtrado. A esta salida así obtenida se le conoce con el nombre de espectro Mel (o Mel-Spectrum por su nomenclatura en inglés).

A la salida del banco de filtros se aplica el logaritmo y la DCT, para obtener finalmente el cepstrum LP del filtro $H(z)$ de entrada.

En 1980 Davis y Mermelstein, demostraron la gran robustez de los coeficientes mel-cepstrales obtenidos a partir de las salidas de un banco de filtros en la escala Mel. En particular, demostraron que el cepstrum así obtenido permite suprimir variaciones espectrales insignificantes en bandas de alta frecuencia. También demostraron que las representaciones suavizadas de los espectros de magnitud, capturan mejor la información acústica relevante para una tarea de RAH [25].

3.4.4. Energía

En nuestro caso, queremos hacer especial énfasis en el estudio de la energía, pues disminuye significativamente las tasas de error en el reconocimiento, como será explicado en el Capítulo 5. Para obtenerla, es común en algunos codificadores de voz, que se incluya como un parámetro adicional, pero en otras ocasiones no es así, y por tanto es necesario obtenerla, bien sea a partir de las muestras de la señal decodificada o a partir de los parámetros que envía el codificador y que están contenidos en el bitstream. Para el cálculo de la energía a partir de las muestras se suele utilizar la Ecuación (3.43), la cual se obtiene una vez por cada trama de L muestras:

$$E = \sum_{n=0}^{L} s^2[n] \tag{3.43}$$

Sin embargo, a la hora de incluir la información de la energía en el vector de características, no se suele incluir la obtenida en la Ecuación (3.43), sino la denominada *log-energía*:

$$E_{log} = 10 \log_{10} E \tag{3.44}$$

De otro lado, en la literatura existen soluciones para la estima de la energía a partir de los parámetros del bitstream y que se han mostrado muy robustas en RAH. Por tanto, nos hemos propuesto estudiar también una forma de obtener una estima robusta de energía, frente a los problemas típicos del reconocimiento de voz codificada, esto es: ruido de ambiente, pérdida de bits o paquetes, etc. Para ello, se han utilizado trabajos previos [114][113][115][45], que serán adaptados a los entornos que son objeto de nuestro análisis (codificadores G.729 y AMR-NB). Estos procedimientos de estima serán presentados con detalle en la Sección 6.4.

3.4.5. Parámetros Dinámicos

Al conjunto de parámetros conformado por los coeficientes cepstrales (o melcepstrales) más la energía, se le denomina comúnmente *Conjunto de Parámetros Estáticos*, pues la información contenida en ellos es relativa a una trama de voz y no a la evolución de éstas en el tiempo. Sin embargo, dado que el tamaño de las unidades acústicas que se deben identificar en un sistema de RAH, es mayor al de una trama; es importante añadir parámetros que incorporen información relativa a la transición de los parámetros estáticos entre tramas contiguas. Por este motivo, en 1986 Furui et al [44] plantea la inserción de características dinámicas que reflejen la trayectoria temporal de los parámetros estáticos en el contorno de una trama. Para ello planteo tres tipos de parámetros: valor medio, pendiente y curvatura.

Para el cálculo de la pendiente (delta o primera derivada), a cada característica (en este caso coeficiente cepstral) $\tilde{h}[n]$ de la trama k, se le construye la secuencia en el tiempo $\tilde{h}[n,k]$.

De esta manera, la pendiente $\delta[n]$ para cada trama k es calculada utilizando la ecuación:

$$\delta[n,k] = \frac{\sum_{i=1}^{K} \left(\tilde{h}[n,k+i] - \tilde{h}[n,k-i]\right) i}{2\sum_{i=1} K i^2} \tag{3.45}$$

donde K es el número de muestras posteriores. Esta misma recursión se puede utilizar para cualquier coeficiente cepstral obtenido de cualquiera de los procedimientos expuestos en la Sección 3.4.1. Similarmente, la doble-delta o segunda derivada se puede obtener utilizando la misma Ecuación (3.45), sobre los parámetros calculados en la primera derivada, en lugar de los parámetros estáticos.

Capítulo 4

Problemática del Reconocimiento de Voz Codificada

4.1. Introducción

En este capítulo se describirán las principales distorsiones introducidas en la voz a su paso por la red de comunicaciones. A continuación, se procede a explicar las diferentes arquitecturas que surgen en función de la distribución de los procesos del sistema de RAH y que intentan solucionar alguno de los problemas anteriores. Sin embargo, y dado que nuestro foco de interés apunta al reconocimiento remoto, se detallarán las alternativas bajo dicho entorno, dando especial importancia a la Transparametrización.

4.2. Principales Tipos de Distorsión

En esta sección, se discutirán los problemas más relevantes que afectan la calidad del reconocimiento sobre una red de comunicaciones. Bajo este entorno, cobran especial importancia los debidos a la distorsión de codificación, los errores de transmisión y el ruido de ambiente.

4.2.1. Distorsión por Codificación

En el RAH sobre una red de comunicaciones, un problema importante surge del hecho de que existen algunas restricciones inherentes a la naturaleza de una comunicación oral, es decir, problemas derivados de las exigencias que implican mantener activa una comunicación natural, fiable y continua. Entre otras exigencias, una red de comunicaciones debe ser robusta frente a las siguientes limitaciones:

Retardos en la comunicación, que pueden estar presentes por el elevado procesamiento que implican las tareas de codificación, especialmente de los codificadores que buscan la optimización de la tasa binaria. El retardo variable (jitter) que se introduce en una red de paquetes, etc.

Limitación del ancho de banda, que en general, es una limitación presente en cualquier servicio de comunicaciones, pero que en el caso de la comunicación de voz, tiene especial relevancia, dadas las altas exigencias que se piden a éste, pues se debe garantizar la entrega continua y a tiempo de cada trama enviada por el codificador, porque de lo contrario el retardo presente podría afectar gravemente la calidad de la comunicación o incluso interrumpirla de forma abrupta, generando inconformidades en el servicio.

Por lo tanto, los algoritmos de codificación son diseñados para que la comunicación de voz cumpla con las exigencias antes descritas. Sin embargo, el diseño de los codificadores de voz está orientado a la recepción humana y, por tanto, los esfuerzos se concentran en gran medida en la optimización de la calidad perceptual de la voz codificada. Por este motivo y por la capacidad de procesamiento del cerebro, los seres humanos se muestran muy tolerantes a la distorsión de codificación. Esto no ocurre, sin embargo, en el caso de los reconocedores automáticos de habla [136].

Uno de los primeros estudios sobre la influencia de la codificación en el reconocimiento se presenta en [36]. En dicho estudio se utilizaron codificadores con tasas entre 4,8 Kbps y 64 Kbps, siendo la tasa de 4,8 Kbps la que presenta peores resultados tanto en las pruebas de reconocimiento de palabras aisladas (independiente del locutor), como en las de verificación de locutor. Esto es previsible, pues con una menor tasa binaria, el ruido de cuantificación es mayor, produciendo una pérdida perceptible en la calidad de la señal reconstruida.

En el mismo estudio, también se analiza el efecto de la codificación en la desajuste de las condiciones de entrenamiento y de prueba, y se concluye que las tasas de reconocimiento de voz codificada, mejoran cuando el reconocedor de voz se entrena usando voz procesada con el mismo algoritmo de codificación.

Dado que se considera factible, que en el extremo decodificador se conozca el algoritmo de codificación empleado y que el número de estos codificadores sea normalmente bajo para estas aplicaciones, a lo largo de esta tesis presentaremos resultados en los que no existe desajuste por codificación.

4.2.2. Errores de Transmisión

Otro de los problemas que afecta el reconocimiento de voz codificada, es el de los errores de transmisión, que en el caso de una red de telefonía móvil, se manifiesta como pérdida de bits y en el caso de Internet, la pérdida de información se produce en paquetes. Es, sin embargo, especialmente dañino el hecho de que tanto los bits como los paquetes se pierdan en ráfagas.

Errores en Redes de Telefonía Móvil:

Debido al ruido y los desvanecimientos presentes en los canales de transmisión, la señal de radiofrecuencia y, por lo tanto, los flujos binarios contenidos en ella, se ven alterados y por tanto llegan al receptor con errores. Sin embargo, la codificación de canal, tiene previsto este tipo de problemas y utiliza procedimientos para reducirlos o eliminarlos. En las redes de

telefonía móvil actuales, se utilizan codificadores convolucionales, turbocódigos u otro tipo de codificación para conseguir el objetivo antes mencionado. No obstante, la codificación de canal no puede corregir todos los errores y por tanto, éstos llegan al reconocedor afectando su desempeño. Entre los diferentes errores producidos por efectos del canal, se destaca la pérdida de bits en ráfagas, pues éstas (si son especialmente largas) no pueden ser corregidas por el codificador de canal y por tanto pueden producir una pérdida importante de bits en las tramas enviadas por el codificador de voz. En los casos más extremos, las ráfagas erróneas producen a su vez la pérdida de tramas completas o incluso de tramas consecutivas, siendo esto último, lo que más deteriora la calidad del reconocimiento [122].

Errores en Redes IP:

Similar al problema presente en las redes de telefonía móvil, en general para el caso de una red de conmutación de paquetes (como Internet), es la pérdida continua de éstos [116] la que afecta de forma grave el funcionamiento de los reconocedores. Las principales causas se explican a continuación:

Para el caso concreto de la transmisión de datos sobre una red IP, se utilizan comúnmente dos protocolos: TCP (Transmission Control Protocol) y UDP (User Datagram Protocol).

El TCP compensa la pérdida de paquetes re-transmitiéndolos cuando se detecta alguna pérdida, sin embargo este tipo de control introduce retardos considerables que no son adecuados para el uso de servicios de transmisión en tiempo real, tal como lo exige una transmisión de voz. En definitiva, las retransmisiones de paquetes perdidos de TCP no solucionan dicha pérdida, pues llega tarde para su reproducción, resultando en una carga más que una ayuda.

Por otro lado, el protocolo UDP no es un protocolo orientado a conexión y por tanto, no introduce mecanismos de recuperación de paquetes. Lo anterior hace que el UDP sea igual de fiable para las aplicaciones que nos ocupan, pero menos pesado que el TCP.

De esta manera, el protocolo UDP es el mas comúnmente usado para transmisiones en tiempo real, como en el caso de VoIP [98], pero esto hace que cuando exista pérdida de paquetes en la red, estos no sean retransmitidos y por tanto, se produzca una pérdida definitiva de ellos.

Además de los problemas generados por la falta de mecanismos de recuperación de paquetes del protocolo UDP, la congestión en la red de transporte hace que los sistemas de conmutación desechen paquetes, incrementando por tanto la tasa de paquetes perdidos.

El resultado final es que la pérdida de paquetes en ráfagas es la principal causa del deterioro del desempeño de un reconocedor, pues esta pérdida no es mitigada por la red de transporte y tampoco se puede corregir con procesos de recuperación presentes en el decodificador de voz, ya que estos últimos se dedican fundamentalmente a repetir tramas pasadas, para reemplazar las que han llegado erróneas [3]. Es de destacar, que en la

transmisión de voz por Internet, es usual transmitir varias tramas de voz por paquete para mejorar la eficacia de la transmisión, lo cual afecta aún más la tasa de reconocimiento, pues se multiplica el número de tramas perdidas, produciendo además ráfagas de errores más largas. En [98], se analiza esta problemática del reconocimiento de voz sobre Internet (bajo un esquema de reconocimiento de voz decodificada) utilizando el estándar G.723.1 [66].

4.2.3. Ruido de Ambiente

El efecto del ruido en un sistema de reconocimiento, es un factor importante a la hora de buscar alternativas robustas, en particular, en entornos móviles, puesto que precisamente debido a la movilidad de los terminales, la comunicación se ve expuesta a diferentes entornos, y muchos de éstos son perjudiciales para el reconocimiento. En particular, el ruido de ambiente merece atención aparte, por las dificultades que su presencia implica.

Existen diferentes tipos de ruido ambiente, que en términos generales, pueden agruparse entorno a tres factores: densidad espectral de ruido, que además provoca el efecto Lombard [20] (como reacción del hablante a dicho ruido); reverberación (caracterizada por la respuesta impulsional del canal entre la boca y el micrófono) y, el eco producido por el acoplamiento acústico entre el altavoz y el micrófono del teléfono. Estos problemas no sólo perjudican el reconocimiento automático sino también al humano.

4.3. Arquitecturas de RAH en una Red de Comunicaciones

Cuando queremos realizar tareas de reconocimiento de voz transmitida por una red de comunicaciones, se pueden dar varios enfoques en función de la distribución de las tareas que lo componen, así, para nuestro análisis podemos retomar el esquema planteado por [29] que separa los procesos de extracción de características (Front-end), y de decodificación (Back-end).

Por otro lado, si consideramos una arquitectura cliente-servidor, las tareas del reconocimiento pueden estar localizadas del lado del cliente, del lado del servidor, o de forma distribuida entre ellos. Es así como nacen tres formas de hacer reconocimiento [113].

4.3.1. Reconocimiento Local

Se da cuando todo el proceso de reconocimiento se realiza en el terminal (Front-end y Back-end). Dicho terminal puede ser un teléfono móvil, o cualquier dispositivo que soporte algún tipo de conectividad de red y captura de voz.

En la Figura 4.1[113] se puede observar que las tareas de extracción de características y reconocimiento se realizan en el terminal. Esta opción resulta útil en aplicaciones que requieren poca complejidad, tales como marcación por voz o reconocimiento de palabras aisladas, debido a que, si bien es cierto que en los últimos años ha habido un avance espectacular, las limitaciones intrínsecas de los terminales (procesador, batería, memoria, etc.), hacen difícil la ejecución de tareas de reconocimiento más complejas, necesarias en

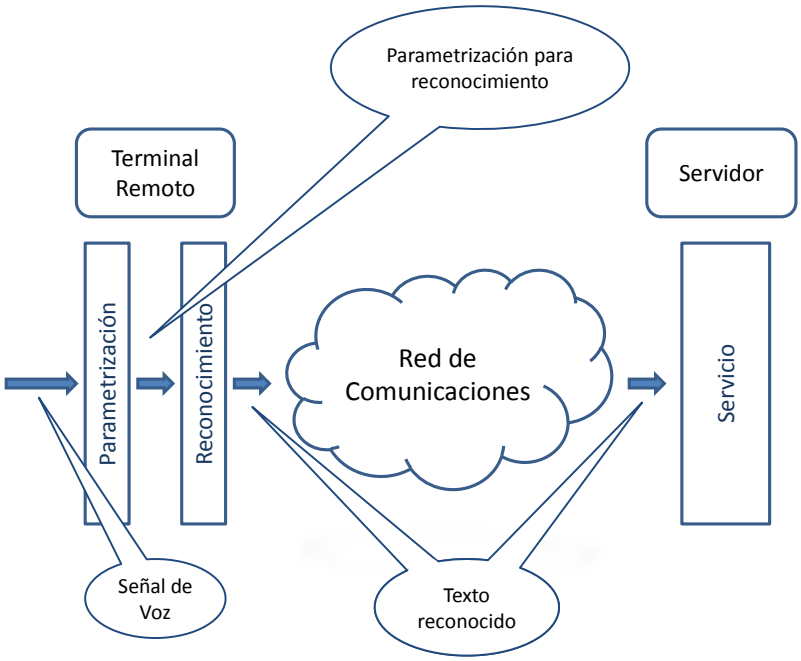


Figura 4.1: Reconocimiento Local.

aplicaciones de dictado o sistemas de diálogo.

Lo más notable en este caso, es que se evitan las distorsiones de codificación y de transmisión.

4.3.2. Reconocimiento Distribuido

En este caso, el proceso de extracción de características (llamado también parametrización) se realiza en el terminal, pues comúnmente requiere de poca potencia de procesamiento. Por otro lado, las tareas de reconocimiento se desarrollan en el servidor remoto [34] (véase la Figura 4.2[113]).

Una de las ventajas principales de esta configuración, es que sólo es necesario transmitir al servidor, los parámetros (o características) para el reconocimiento, utilizando para ello, un ancho de banda considerablemente inferior al necesario para enviar la señal de voz.

Sin embargo, cuando se utilizan codificadores a bajas tasas binarias, el ancho de banda utilizado para transmitir parámetros de codificación es casi tan pequeño como el utilizado para la transmisión de parámetros de reconocimiento. Ésto se puede apreciar más aún, si en el codificador se activa un Sistema de Transmisión Discontinua (conocido como DTX [149]), el cual es muy común en los codificadores actuales, y que puede conseguir una reducción significativa en el ancho de banda, pues en una comunicación se pueden encontrar silencios que suman incluso más de un 60 % del tiempo de conversación[87].

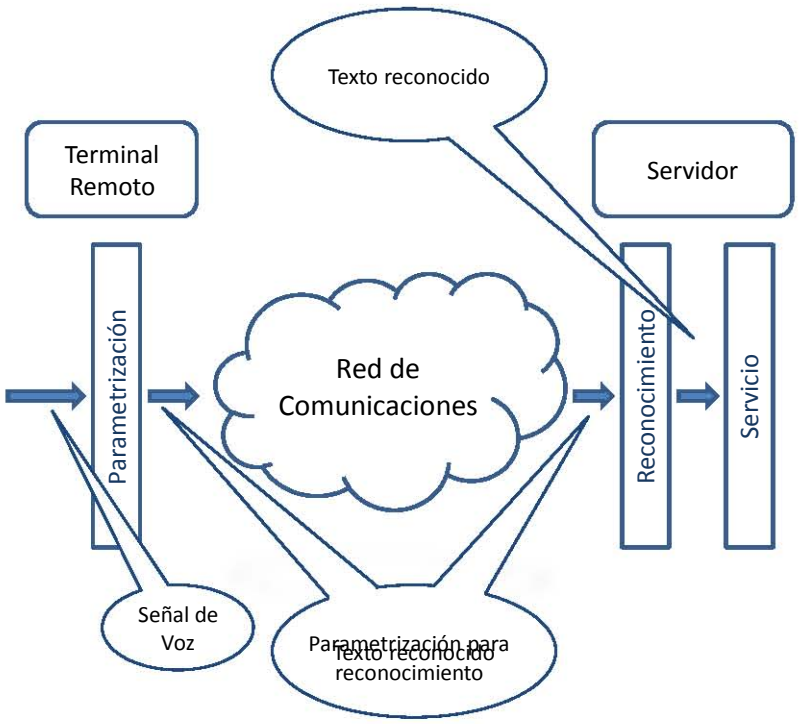


Figura 4.2: Reconocimiento Distribuido.

Por otra parte, es importante destacar, que debe existir un acoplamiento entre el Front-End y el Back-End, es decir, debe existir un procedimiento estándar para la transmisión de los parámetros, en donde se establezcan los mecanismos de protección, cuantificación, etc. para conseguir una comunicación eficiente. En este sentido, se han desarrollado diferentes esfuerzos para conseguir un estándar apropiado, entre ellos podemos destacar el estándar ETSI AURORA [34].

4.3.3. Reconocimiento Remoto

Bajo esta configuración, tanto el Front-end como el Back-end se ejecutan en el servidor remoto, por tanto, el terminal no participa de estos procesos, y no es necesario que este sea un terminal diseñado para reconocimiento. Esta es una de las principales ventajas de esta configuración, pues amplia enormemente las posibilidades de terminales que se pueden utilizar, dado que éstos no tienen que estar diseñados para soportar procesado específico que les permita acceder a un servicio de reconocimiento. Esta configuración, permite el uso de cualquier terminal con capacidad para transmitir voz, como un teléfono móvil, un terminal de VoIP, una consola de juegos, etc. (Véase la Figura 4.3[113]).

Es de destacar, que en este entorno, el ancho de banda disponible por el terminal es reducido y por tanto es necesario que la voz sea codificada para optimizar el ancho de banda disponible. Lo anterior, sin embargo, introduce distorsión de codificación que provoca una

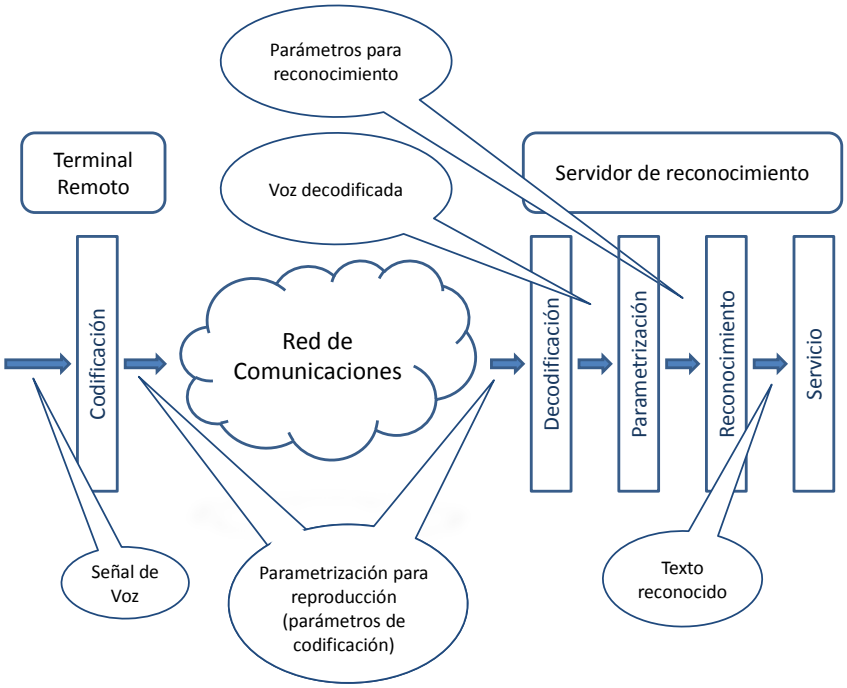


Figura 4.3: Reconocimiento Remoto.

disminución del desempeño del sistema de RAH [91].

4.3.4. Ventajas del Reconocimiento Remoto

De las anteriores arquitecturas, en esta tesis hemos escogido la alternativa de reconocimiento remoto por varias razones, entre las cuales podemos destacar:

- Dado que el reconocimiento se realiza en el servidor, tenemos la disponibilidad de escoger el front-end que mejor se adapte a una aplicación determinada, y además actualizarla cuando sea necesario. De este modo, las nuevas investigaciones en front-end's pueden ser inmediatamente aplicadas. En particular, la mayoría de los front-end robustos diseñados para Reconocimiento Distribuido pueden ser incluidos, permitiendo con ello más técnicas de adaptación de ambiente o incluso pueden ser organizadas en cascada o combinarlas con otros tipos de front-end's con el ánimo de minimizar los efectos del canal de comunicaciones mencionados al principio de éste capítulo (Véase Sección 4.2.1).

- Debido a que el terminal no realiza el reconocimiento, éste no tiene imposiciones restrictivas en sus capacidades, y por tanto no crea necesidades para configuraciones especiales o acuerdos entre en el terminal remoto y el servidor.

- Este modelo mantiene los requerimientos de ancho de banda en la transmisión, y la compatibilidad con las aplicaciones existentes de voz basadas en estándares. Por esto, no requiere ningún cambio en los protocolos de comunicación para manejar

conmutación dinámica entre transmisiones de voz y datos, y por lo tanto, es más adecuado para desarrollos en donde se requiere que la red de comunicaciones sea transparente al sistema de reconocimiento.

- Podemos recuperar la forma de onda de la señal de voz, con la calidad provista por el codificador de voz empleado. Y por otro lado, dado que la información del codificador usado es siempre enviada en la señalización del sistema de comunicaciones, podemos usar modelos adaptados para mejorar el proceso de reconocimiento.

- Los codificadores de voz AMR aplicados tanto en UMTS como en GSM (también se ha contemplado su uso en la red de telefonía móvil de cuarta generación *Long Term Evolution - LTE* [70]), equilibran apropiadamente la cantidad de ancho de banda empleado, por un lado, en la transmisión de voz y por el otro, en la protección de dicha transmisión, teniendo en cuenta para ello las condiciones presentes en el canal. Mientras que esas cantidades permanecen fijas en el caso de Reconocimiento Distribuido. Además, los sistemas TFO (Tandem Free Operation) limitan la distorsión de codificación a una etapa. Sin embargo, es muy usual encontrar 2 ó 3 etapas de transcodificación si no se aplican.

4.4. Técnicas de Reconocimiento Remoto

Una vez decidido el modelo de reconocimiento remoto, debemos explorar las diferentes alternativas que se pueden utilizar para conseguir una parametrización óptima bajo dicho modelo [11].

Tal como se expuso en la Sección 3.4, existen diferentes opciones a la hora de obtener una parametrización óptima. Sin embargo, en el entorno de reconocimiento remoto existen algunas limitaciones, pero también ventajas que pueden ser aprovechadas para decantarse por uno u otro esquema de parametrización. Dichos esquemas están enmarcadas dentro de los procedimientos descritos en el Capitulo 2, y que ahora pueden ser descritos con más detalle.

Para abordar las las diferentes soluciones, se hará una división en dos grupos:

- Por un lado aquellas que utilizan voz decodificada para realizar la extracción de los parámetros de reconocimiento y que describiremos en la Sección 4.5.

- Por otro lado, aquellas que obtienen directamente los parámetros de reconocimiento a partir de los parámetros transmitidos por el codificador, transformando estos últimos para obtener los primeros, sin necesidad de sintetizar la forma de onda de la voz por el procedimiento de codificación. Esta opción la describiremos en la Sección 4.6.

A la primera opción le podemos denominar *Reconocimiento a Partir de Voz Decodificada* y, a la segunda *Reconocimiento Mediante Transparametrización* (Véase Sección 4.6).

4.5. Reconocimiento a Partir de Voz Decodificada

Realizar la decodificación de la voz antes de la extracción de características para el reconocimiento, es el procedimiento habitual. Sin embargo, dentro de este esquema se pueden utilizar diferentes alternativas para obtener el cepstrum, siendo las más usuales las explicadas en la Sección 3.4.1 y que se diferencia en la forma en que se obtiene la envolvente espectral.

De un lado el *Método Tradicional* obtiene la envolvente espectral utilizando deconvolución homomórfica y de otro lado, el método que denominamos *Reconocimiento Suavizado* utiliza predicción lineal para extraerla. A continuación se explicarán con más detalles los dos procedimientos.

4.5.1. Método Convencional

Es nuestra referencia a la hora de comparar los resultados con los demás procedimientos, y como se ya explicó, el cepstrum se obtiene utilizando la envolvente espectral extraída por el método de deconvolución homomórfica a partir de la señal de voz reconstruida por el decodificador.

Una de las bondades de esta técnica, es que el proceso de reconocimiento toma directamente la voz decodificada y esto hace que el front-end sea, en un principio, transparente al codificador utilizado. Además, se pueden calcular directamente de la voz decodificada, otros parámetros que pueden ayudar a la tarea de reconocimiento (p.e. la energía). Por otro lado, la señal decodificada se puede usar para otros propósitos que pueden complementar el servicio de reconocimiento de voz, como puede ser la identificación de locutor.

En contrapartida, es de destacar que esta técnica se torna poco robusta debido a la distorsión que introduce el proceso de codificación - decodificación, y frente a los errores de transmisión, en comparación con otras técnicas. Esto se ha demostrado en varios estudios existentes en la literatura, además del efecto de otro tipo de distorsiones. Sin embargo éstos trabajos serán detallados en el Capitulo 5.

Procedimiento

En la Sección 3.4.1 se explicó la forma de obtener los coeficientes cepstrales utilizando el método de deconvolución homomórfica. En este caso, el cepstrum es obtenido partiendo de la señal de voz $s[n]$ después del proceso de decodificación. Sin embargo, dado que el ámbito de estudio de esta tesis, es el RAH sobre redes de comunicaciones, asumimos que la voz ha sido codificada, y por lo tanto todo el vector de características se obtiene a partir de la muestras de voz de la señal reconstruida en el proceso de decodificación (véase la Figura 4.4).

Lo anterior, implica que todos los parámetros que conforman el vector de características arrastran consigo las distorsiones asociadas al proceso de codificación-decodificación, a los

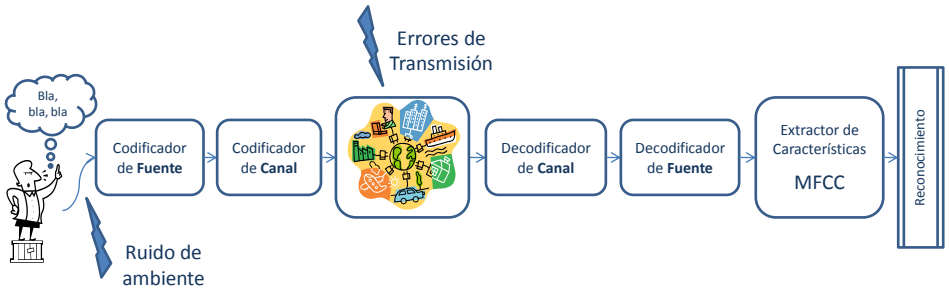


Figura 4.4: Reconocimiento de Voz Decodificada.

errores producidos en la red de comunicaciones, y al ruido de ambiente, entre otros. Ver Sección 4.2.

Las etapas que conlleva el proceso de conformación del vector de características son las siguientes:

1. *Cálculo de la Envolvente Espectral*:

   En este caso, los parámetros que contienen la información de la envolvente espectral son los coeficientes cepstrales obtenidos por los métodos de deconvolución homomórfica y escalado Mel, descritos en las Secciones 3.4.1 y 3.4.2. Sin embargo, como se reseñó antes, el cepstrum se obtiene partir de la señal decodificada, y por tanto, éste será una aproximación al cepstrum de la señal original, pues la señal reconstruida en el receptor, lleva consigo las distorsiones introducidas por la red, a lo largo de todo el proceso de comunicación (Véase la Sección 4.2).

2. *Obtención de la Energía de Trama*:

   En este caso, dado que la señal ha sido decodificada, podemos procesar directamente las muestras de voz para el cálculo de la energía, y por tanto, utilizar el procedimiento descrito en la Sección 3.4.4.

   Esta energía concatenada a los MFCC constituyen el conjunto de los denominados parámetros estáticos.

3. *Cálculo de los Parámetros Dinámicos*:

   Una vez se han obtenido los coeficientes mel-cepstrales y la energía, se procede a calcular los deltas (parámetros dinámicos) para los parámetros anteriores. Para dicho cálculo se puede hacer uso del procedimiento descrito en la Sección 3.4.5, dando como resultado un vector de parámetros que puede ser finalmente utilizado como vector de características para la tarea de RAH.

4.5.2. Espectro Suavizado

Cuando se realiza un análisis LPC como el descrito en la Sección 3.4.1, se obtiene la función de transferencia del filtro de síntesis $H(z)$ y a partir de ella, podemos extraer la información de la envolvente espectral. Por tanto, si aplicamos al espectro de $H(z)$ un banco de filtros y los demás procesos descritos en la Sección 3.4.3 podemos obtener el cepstrum de $H(z)$.

Es de anotar, que tal como se expuso en la Sección 3.4.3, el cepstrum conseguido será un cepstrum LP. Sin embargo, existen algunas particularidades en el procedimiento utilizado para su cálculo en esta técnica de reconocimiento, pues de forma similar a lo visto en la Sección 4.5, hemos partido del hecho de que la voz utilizada para obtener el vector de características es la decodificada.

Lo anterior, implica que la señal de voz original, ha sido sometida a dos procesos de predicción lineal; el primero realizado por el codificador fuente (ver Sección 3.3.1), y el segu

A ı de
reco

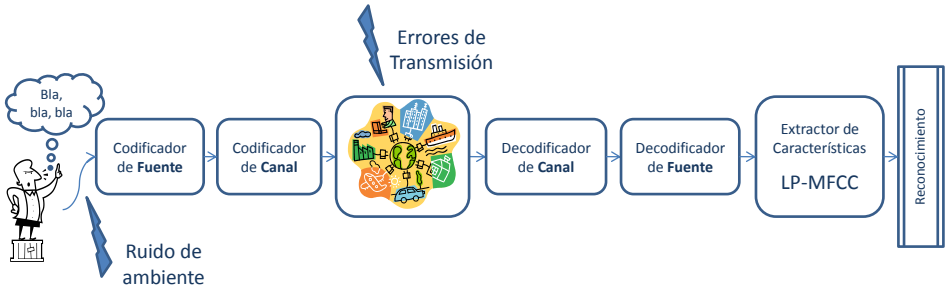


Figura 4.5: Reconocimiento de Voz Decodificada con Suavizado Espectral.

Las etapas que conlleva el proceso de conformación del vector de características son las siguientes:

1. *Cálculo de la Envolvente Espectral*:

   Dado que el espectro es obtenido a partir de voz decodificada, primero se realiza un análisis LPC sobre las muestras de voz reconstruidas en el proceso de decodificación, para así obtener los coeficientes LP (LPC). Con estos coeficientes podemos calcular el espectro LP y continuar el proceso descrito en la Sección 3.4.3 para obtener el cepstrum LP en la escala Mel.

2. *Obtención de la Energía de Trama y Parámetros Dinámicos*:

Puesto que la extracción de características se realiza sobre voz decodificada, el procedimiento para obtener la energía es el mismo que el utilizado en la Sección 3.4.4. De igual forma, los parámetros dinámicos son calculados como en la Sección 4.5.

4.6. Reconocimiento a partir de los Parámetros del Bitstream

Las técnicas anteriores, han utilizado voz decodificada para obtener el vector de características necesario para el sistema de RAH. Sin embargo, como se expuso en la Sección 3.3, el análisis LPC realizado para conseguir los coeficientes de predicción en la aproximación de reconocimiento suavizado (Sección 4.5.2), también se realiza en la inmensa mayoría de codificadores que trabajan en los rangos de tasa binaria de nuestro interés, entre los cuales se encuentran los de la familia CELP. Por tanto, los parámetros que representan la envolvente espectral se pueden extraer directamente del bitstream enviado por el codificador para evitar la repetición del proceso de predicción.

Como se explicó en la Sección 3.3.2, la información relativa al filtro de síntesis es enviada a través del bitstream utilizando LSP en lugar de LPC. Por tanto, es necesario utilizar, o bien un proceso de transformación de parámetros LSP a LPC, o bien un procedimiento para el cálculo del espectro LP, a partir de los LSP directamente [115]. De esta manera, a partir del espectro LP se podrían obtener los coeficientes cepstrales utilizando el procedimiento expuesto en la Sección 3.4.3, evitando así la distorsión de decodificación.

Por tanto, surge la de idea de buscar un proceso de transformación de los parámetros LSP enviados por el codificador de voz, en parámetros cepstrales para reconocimiento. A este procedimiento se le ha denominado: *Transparametrización*[113] o también "basado en el flujo digital binario" (*bitstream based*)[81], y dado que es la técnica central en el análisis de esta tesis, se realizará una explicación más amplia y detallada de ésta a continuación.

Por otra parte, también se han desarrollado algunos trabajos que proponen algoritmos de codificación basados en parámetros de reconocimiento (concretamente MFCC) tal como se muestra en [86], aunque esto implica modificar el sistema de codificación y por tanto no sería transparente a la red de comunicaciones como pretendemos en nuestra solución.

A continuación, se referencian algunos estudios previos que dan origen a la Transparametrización, así como los procedimientos específicos requeridos para su implementación. Finalmente, se exponen las ventajas que esta ofrece frente a las otras técnicas de referencia.

4.6.1. Estudios Previos

En los sistemas de reconocimiento basados en voz decodificada, se ha observado que el proceso de codificación-decodificación disminuye significativamente la tasa de reconocimiento, disminuyendo aún más, conforme se disminuye la tasa de codificación [38] (para un análisis más detallado Véase la Sección 5.2.1). Debido a este inconveniente, han

surgido algunas alternativas que pretenden evitarlo o al menos reducirlo.

Una forma de disminuir el efecto de la distorsión por decodificación, es construir un Front End que obtenga los parámetros de reconocimiento, a partir de la secuencia de bits que envía el codificador. Para ello es necesario encontrar una transformación que a partir de los parámetros enviados en dicha secuencia (orientados a la reconstrucción de la voz) pueda calcular los parámetros orientados al reconocimiento.

De otro lado, dentro de las posibilidades de utilizar los parámetros de codificación directamente en una tarea de RAH, Choi et al [22][24][23], Zheng et al [159] utilizaron los LSP como vector acústico de entrada al sistema de reconocimiento, pues la información contenida en ellos está presente en el dominio frecuencial y representan la envolvente espectral. Sin embargo, estos parámetros usados directamente, obtienen un desempeño muy por debajo de los obtenidos con los parámetros cepstrales, razón por la cual no son utilizados para la conformación del vector de características en un sistema de RAH.

4.6.2. Características principales de la Transparametrización

A través del uso de la transparametrización se pueden obtener de la señal codificada, sólo las características que interesan para el reconocimiento, sin que intervenga la decodificación y por consiguiente la distorsión asociada a ella [36][91] (véase la Figura 4.6). Es de destacar que en el caso de existir un codificador de canal, los parámetros más importantes y los que

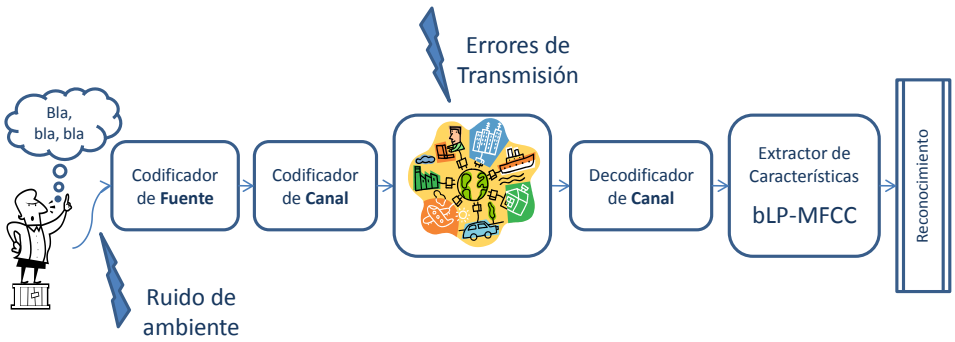


Figura 4.6: Reconocimiento de Voz por Transparametrización.

Sintetizando, algunas ventajas que podemos destacar del uso de esta técnica, son las siguientes:

- Por un lado, dado que el reconocimiento se realiza a partir de la secuencia de bits, se

evita la distorsión asociada al proceso de decodificación de la voz. Un análisis más detallado se puede ver en la Sección 5.2.1.

- De otro lado, Choi y Kim [24] destacan que el procedimiento de conversión de parámetros, reduce el coste computacional necesario para una tarea de reconocimiento de voz codificada. Dicha reducción se obtiene minimizando el proceso de extracción de características, pues la envolvente espectral es calculada directamente a partir de los LSP. Además, se omite gran parte del proceso de decodificación utilizado en el método tradicional de reconocimiento de voz reconstruida (véase Sección 4.5). Una vez se reducen los costes computacionales, se pueden también reducir los tiempo de procesamiento y con ellos el retardo generado.

- Por ultimo, bajo un entorno con errores de transmisión, los mecanismos de recuperación de errores provistos por los codificadores estándar pueden ser mejorados [117][154], adaptándolos al problema de reconocimiento automático de habla [97]. Lo anterior se puede conseguir, flexibilizando las restricciones impuestas al proceso de codificación, tales como el retardo máximo, o los métodos de interpolación ponderada, entre otros aspectos a considerar.

Para lograr lo anterior, podemos recurrir a los estudios desarrollados por Atal [9], Sugamura [136], Kim [78], Peláez-Moreno [113]; para obtener el cepstrum directamente a partir de los LSP, derivando en una solución muy robusta, no solo frente a la distorsión asociada a la codificación, sino también frente a otros tipos de distorsión (Véase Sección 5). Es sin embargo notable, que en ninguno de estos trabajos se considere el efecto del ruido de ambiente.

4.6.3. Procedimiento de Conversión

La información sobre la envolvente espectral, puede ser representada por varios tipos de parámetros dependiendo del codificador usado; sin embargo, la mayoría provienen de un análisis LPC. Por tanto, el primer paso que hay que realizar es el de análisis de la estabilidad del filtro definido por los parámetros recibidos; para ello, debemos transformarlos de tal forma que se pueda hacer esta verificación. De esta manera, si se han recibido parámetros LSP, no es necesaria dicha transformación puesto que se puede verificar la estabilidad directamente [81], y en su caso, hacer las correcciones oportunas.

Y dado que a partir del bitstream podemos obtener los parámetros LSP, podremos utilizar éstos para calcular los coeficientes LP cepstrales.

Primero extraemos los LSP (que han sido obtenidos una vez por cada trama) y los utilizamos para el cálculo de la envolvente espectral, que según el procedimiento establecido por Sugamura [136] y que explicaremos a continuación, modelamos primero un filtro todo-polos a partir de los LSP. Luego aplicamos un banco de filtros de escala Mel a la envolvente espectral y realizamos una transformación discreta de coseno a las energías en bandas obtenidas a la salida del banco de filtros, para obtener finalmente los coeficientes LP-Mel-cepstrales. (Véase la Sección 3.4.3).

Epectro LP a partir de LSP: Sugamura et al [136], utilizaron un esquema de filtro todo polos para conseguir señales sintéticas de gran calidad a bajas tasas binarias (por debajo de 9600 bps). Concretamente, lograron caracterizar un filtro todo polos utilizando LSP en lugar de los tradicionales LPC. La Ecuación (4.1) ilustra la función de transferencia del filtro utilizado:

$$|\hat{H}(\Omega)|^2 = \frac{2^{-p}}{sen^2\left(\frac{\Omega}{2}\right)\ T_o(\Omega) + cos^2\left(\frac{\Omega}{2}\right)\ T_e(\Omega)} \tag{4.1}$$

donde,

$$T_o(\Omega)| = \prod_{i=1}^{\frac{1}{2}p}(cos\ \Omega - cos\ \omega_{2i-1})^2 \tag{4.2}$$

y,

$$T_e(\Omega)| = \prod_{i=1}^{\frac{1}{2}p}(cos\ \Omega - cos\ \omega_{2i})^2 \tag{4.3}$$

Siendo ω_i el i-ésimo LSP, y p el número de LSP. En este caso, p debe ser par. Para valores impares de p, se proporcionan las ecuaciones correspondientes en [136]. Es de resaltar, que la función de transferencia obtenida a partir de los LSP: $\hat{H}(\Omega)$, es una aproximación de la original $H(\Omega)$, pues la primera es calculada utilizando los LSP transmitidos por el codificador, y por tanto están cuantificados.

En [113] se ha comprobado la efectividad de la anterior transformación, comparando además los resultados de reconocimiento, partiendo de LSP y de coeficientes LP. En el Capítulo 5, se analizarán los resultados obtenidos utilizando esta transformación de parámetros, para el caso de los codificadores G.729 y AMR-NB.

4.6.4. Estimación de la Energía

Dado que en la Transparametrización no se realiza el proceso de decodificación de la voz, no es posible obtener las muestras de la señal reconstruida y por tanto tampoco es posible calcular la energía con el procedimiento descrito en la Sección 3.4.4.

Es por esto que debemos realizar una estimación de la energía, a partir de la envolvente espectral y otros parámetros enviados en el bitstream por el codificador (idealmente aquellos mas protegidos o menos afectados por las distorsiones existentes) [113][45]. Sin embargo, dada la importancia de este parámetro en la tasa de reconocimiento, y debido a las particularidades inherentes a los dos estándares de codificación que se analizan en esta tesis, el procedimiento detallado será descrito en el Capítulo 6.

4.6.5. Reconocimiento a partir de Pseudo-Cepstrum

Una transformación alternativa fue propuesta por Kim et al [79][78] y Choi et al [23], para obtener los coeficientes cepstrales a partir de LSP.

$$\hat{\tilde{h}}[n] = \frac{1}{2n}\left(1 + (-1)^n\right) + \frac{1}{n}\sum_{i=1}^{p} cos(n\omega_i) \quad 1 \leq n \leq L \tag{4.4}$$

Al igual que en la Ecuación (4.1), en esta transformación hemos utilizado $\hat{h}$ para hacer referencia a que la envolvente espectral obtenida a partir de los LSP cuantificados, pues es una aproximación de la envolvente espectral original h. De igual modo, el pseudo-cepstrum $\hat{\tilde{h}}[n]$ representa una aproximación de $\tilde{h}[n]$.

Sin embargo, aunque es una aproximación computacionalmente más eficiente, el cepstrum obtenido consigue una tasa de reconocimiento inferior a la obtenida con otros métodos. En [113] [115] se puede observar los resultados que han sido obtenidos utilizando esta técnica. En el Capítulo 8 de esta tesis, se analizarán algunos resultados obtenidos comparando esta técnica de reconocimiento, comparándola con otras aproximaciones, bajo un entorno de pérdida de paquetes.

Capítulo 5

Reconocimiento Mediante Transparametrización: estado de la técnica

5.1. Introducción

En este capítulo se analizan las soluciones descritas en la literatura que atacan los problemas expuestos en el capítulo anterior. Sin embargo, dado que nuestra propuesta se centra en los sistemas de reconocimiento mediante transparametrización (RMT), la revisión del estado del arte que se presenta en este capítulo atiende fundamentalmente a lo publicado en este ámbito; no obstante, también se describen algunas soluciones relativas a las arquitecturas de reconocimiento local y/o distribuido que tienen relación con la propuesta que planteamos en el Capítulo 6.

De otro lado, para limitar nuestro análisis, nos restringiremos a abordar sólo los dos tipos de redes planteados para el desarrollo de esta tesis: IP, UMTS.

5.2. Técnicas Robustas frente a la Distorsión de Codificación y Decodificación

El uso de la codificación de voz en las redes de comunicaciones se ha orientado a la comunicación entre seres humanos, y por tanto los esfuerzos se concentran en criterios tales como reducir la tasa binaria maximizando la calidad perceptual. Por este motivo, los seres humanos resultan muy tolerantes a la distorsión de codificación, no siendo así en en el caso de los reconocedores automáticos [136], pues como lo demuestran diversos estudios, el efecto de la codificación puede ser muy nocivo para la tasa de reconocimiento, especialmente cuando se codifica a tasas binarias bajas.

No obstante, existen algunos codificadores que utilizan predicción no lineal para conseguir un incremento significativo en la calidad de la voz reconstruida a bajas tasas de codificación [27][26]), aunque en la literatura no se encuentran estudios que indiquen si son

menos nocivos desde el punto de vista de reconocimiento automático de habla.

A continuación, se exponen algunos estudios realizados sobre el efecto de la codificación en los sistemas de reconocimiento automático y las soluciones propuestas.

5.2.1. Efectos de la Codificación de Fuente

La codificación de fuente genera diferentes problemas a la tarea de reconocimiento de habla. Entre los efectos más notables se encuentran la distorsión derivada de la reducción de la tasa binaria, el desajuste entre los datos de entrenamiento y de test, la cuantificación de los parámetros, la reconstrucción de la voz a partir de bits afectados por el ruido o los errores de transmisión, etc.. A continuación se describirán con más detalle estos problemas.

Distorsión por Reducción de la Tasa Binaria

Uno de los primeros análisis sobre la influencia de la codificación de voz en el reconocimiento fue realizado por Euler et al (1994) [36]. En dicho análisis utilizaron diversos codificadores con tasas entre 4,8 Kbps y 64 Kbps, siendo la tasa de 4,8 Kbps la que dio lugar a peores resultados, tanto en las pruebas de reconocimiento de palabras aisladas (independiente del locutor), como en las de verificación de locutor. Estos resultados son previsibles, pues con una menor tasa binaria de codificación, la distorsión de codificación tiende a incrementarse, produciendo una pérdida de calidad en la voz reconstruida.

De otro lado en [91] también se analiza el efecto de la tasa binaria en el desempeño de un sistema de reconocimiento. Para ello utilizan 6 codificadores con tasas que van desde 4,8 Kbps a 40 Kbps, llegando a la misma conclusión de Euler, pues a pesar de utilizar diferentes algoritmos de codificación, la tendencia indica que a menor tasa de reconocimiento, menor es el desempeño del reconocedor. En este mismo análisis, se concluye también que el efecto de la codificación en cascada perjudica más a los codificadores que funcionan a tasas bajas y medias, como los tipo CELP, que a los de forma de onda que funcionan a tasas altas.

Dafour et al [30] también evalúan el efecto de la distorsión debida a las bajas tasas binarias de los codificadores utilizados, llegando a la misma conclusión.

En general, para codificadores que utilizan tasas inferiores a 16 Kbps, se aprecia una degradación en la tasa de reconocimiento [110].

Efecto de la Cuantificación de Parámetros

En 1998, Choi y Kim [24] examinan el efecto de la cuantificación de los LSP (realizada por el codificador en el terminal cliente) en la precisión del reconocedor. Concluyen que la cuantificación de los LSP introduce distorsión en la información espectral que finalmente repercute en la tasa de reconocimiento. Sin embargo, si a lo anterior se suma la degradación introducida por el decodificador en el proceso de reconstrucción de la voz, el desempeño del reconocedor disminuye aún más.

Choi y Kim [24] plantean además un procedimiento para medir la degradación de la calidad de la voz reconstruida con el ánimo de determinar la distorsión de la información espectral introducida por el codificador. En particular, proponen el uso de la Distancia Media Espectral, descrita en [105].

Huerta et al [61] realizan una comparación del cepstrum calculado a partir de los parámetros LAR (cuantificados y no cuantificados) obtenidos por el codificador Full Rate de GSM [32]. Los resultados nos muestran el efecto nocivo de la cuantificación de los parámetros LAR en la tasa de reconocimiento, siendo éste más perceptible para el caso de voz contaminada con ruido.

Por otro lado, Turunen et al [148] realizan una comparativa de 9 configuraciones de codificación, estudiando la importancia de la parametrización utilizada en cada codificador, y concluyen que si la información del tracto vocal se codifica de forma adecuada, el funcionamiento del sistema de RAH se verá poco afectado aunque se presente una reducción importante en la tasa de codificación. Lo cual no contradice las conclusiones anteriores, pues los coeficientes cepstrales se calculan a partir de la envolvente espectral de la voz.

Lo anterior se puede observar también en la distribución binaria que utilizan los codificadores modernos, tales como el codificador AMR-NB [4] (Véase Tabla 5.1 y Anexo B Tabla B.1). En este codificador, los parámetros que contienen la información de la envolvente espectral (en este caso los LSP), concentran una apreciable cantidad de bits. Sin embargo, para las tasas más altas, el número de bits que aumenta significativamente es el que corresponde a los parámetros de la excitación. Esto no es de extrañar, pues la información contenida en la excitación es la que aporta más naturalidad a la voz reconstruida, mientras que la información de la envolvente espectral contribuye más a su inteligibilidad.

Modo	**Asignación binaria por parámetro**			
	LSP	**Pitch**	**Ganancias**	**Librería Estocástica**
AMR_12.2	38	30	36	140
AMR_10.2	26	26	28	124
AMR_7.95	27	28	36	68
AMR_7.40	26	26	28	68
AMR_6.70	26	24	28	56
AMR_5.90	26	24	24	44
AMR_5.15	23	20	24	36
AMR_4.75	23	20	16	36

Tabla 5.1: Asignación binaria del Codificador AMR-NB para todos los modos de trabajo.

Por tanto, aunque la cuantificación de los parámetros enviados por el codificador de voz es un factor que podría generar dificultades importantes en un sistema de reconocimiento, este tipo de distorsión se puede compensar con el uso de codificadores que utilicen una

parametrización adecuada de los parámetros que caracterizan la envolvente espectral. No obstante, el proceso de reconstrucción llevado a cabo en el decodificador puede introducir distorsiones importantes debido a que incluye información de la excitación, que usualmente se protege poco y es muy vulnerable a los errores de transmisión (como se verá en la Sección 7.4.2).

Desajuste entre los Conjuntos de Entrenamiento y Test

En [36], Euler estudia el efecto de la codificación sobre el desajuste entre las condiciones de entrenamiento y de reconocimiento, y concluye que las tasas de reconocimiento de voz codificada mejoran cuando el reconocedor de voz se entrena usando voz procesada con el mismo algoritmo de codificación.

Debido a lo anterior, Euler et al plantean el uso de de un clasificador gausiano que identifica el tipo de algoritmo de codificación, y por tanto ayuda a reducir el desajuste entre los datos utilizados para realizar el entrenamiento de los modelos y los usados para realizar las pruebas de reconocimiento.

No obstante, en algunas ocasiones es factible conocer el algoritmo de codificación a partir de información a priori, o incluso inferirlo a partir de la información contenida en la señalización, y por tanto no es necesario el uso del anterior algoritmo.

De otro lado, Huerta et al [61] concluyen también que tanto en condiciones limpias como en voz contaminada por ruido, se obtiene una mejor tasa de reconocimiento cuando los datos de entrenamiento presentan las mismas condiciones que las de reconocimiento.

Este desajuste aparece no sólo por el efecto de la codificación de fuente, sino también por otros factores externos, tales como la distorsión convolutiva que introduce un micrófono o un canal de comunicaciones, y que modifican la respuesta en frecuencia de la voz. En este sentido, Mokbel et al [100] estudian un conjunto de técnicas que buscan robustecer el reconocimiento de voz en presencia de desajustes entre las condiciones de entrenamiento y de prueba. Para su estudio dividen las técnicas en dos clases, por un lado las técnicas de pre-procesamiento y, por el otro, la adaptación de parámetros de los Modelos Ocultos de Harkov (HMM). Su análisis está basado en el entorno GSM y en la red de telefonía publica conmutada.

De otra parte, es de destacar que cuando la voz está contaminada por ruido, y podemos ajustar las bases de datos de entrenamiento con las de test, se pueden obtener mejores resultados, respecto de no realizar dicho ajuste, tal como lo exponen Wet et al en [156].

En términos generales es mejor realizar una adaptación de las condiciones de entrenamiento con las de reconocimiento; sin embargo, en la práctica no siempre se puede llevar a cabo, pues en ocasiones no se dispone de la información necesaria para hacerlo. Más aún, si el tipo de distorsiones es variable, como en el caso de una comunicación móvil, el proceso de adaptación puede ser inviable.

Dado lo anterior, para el caso de voz contaminada con diferentes tipos de ruido, lo aconsejable es realizar el entrenamiento utilizando voz limpia. No obstante, sí podemos realizar ajustes en cuestiones como el tipo de codificación, o el procesado que se le aplique a la señal de voz antes de ser codificada.

Efecto de la Decodificación

La decodificación puede ser nociva para un sistema de reconocimiento cuando se utilizan bajas tasas de codificación, pues los parámetros que usualmente se ven más afectados son los que describen las componentes de la excitación y, por tanto, si se utiliza una técnica de reconocimiento basada en voz decodificada, ésta asumirá las distorsiones introducidas por la excitación en la voz reconstruida.

Sin embargo, el proceso de decodificación puede resultar beneficioso debido al procesamiento realizado después de la extracción de parámetros contenidos en el bitstream. En la Sección 3.3.1 se describen las etapas del postfiltrado que se utilizan en los codificadores CELP empleados en esta tesis. En ellas, se incluyen algunos procedimientos que resultan útiles frente al ruido, y por tanto pueden favorecer los resultados de reconocimiento.

A pesar de lo anterior, el proceso de decodificación está orientado a conseguir una buena reconstrucción de la voz y no a los sistemas de reconocimiento. Por tanto, en muchos casos puede ser mejor desechar este procedimiento [113].

5.2.2. Soluciones Robustas frente a la Distorsión por Codificación

En respuesta a los problemas analizados en la sección anterior se han publicado diferentes soluciones que buscan mitigar los efectos que introduce la distorsión por codificación en un sistema de reconocimiento. A continuación se exponen los más destacados.

Reconocimiento Local o Distribuido

Quizás la solución más inmediata consiste en realizar reconocimiento local o distribuido [50], pues con ello se eliminan la codificación de fuente y, por tanto, las distorsiones asociadas a ella. Algunos ejemplos destacados de este tipo de soluciones se describen en Milner et al [98] o en el estándar ETSI AURORA [34], que hacen uso de de una arquitectura de reconocimiento distribuido para eliminar el uso de los codificadores de voz, aunque haya que utilizar otro sistema de codificación para una adecuada transmisión de los parámetros extraídos para el reconocimiento. Sin embargo, por los motivos ya expuestos en la Sección 4.3.4, centraremos nuestro análisis en las soluciones enmarcadas en la arquitectura de reconocimiento remoto que son motivo de análisis en esta tesis.

Coeficientes Cepstrales a partir del Bitstream: Transparametrización

Dentro de las soluciones enmarcadas en el ámbito de reconocimiento remoto, una alternativa robusta que previene en gran medida los efectos de la distorsión por codificación es la Transparametrización (Véase Sección 4.6). Con esta técnica, se pueden obtener los coeficientes cepstrales a partir de los parámetros calculados, codificados y enviados por el codificador de voz.

Sin embargo, antes de entrar a detallar las aplicaciones concretas de la Transparametrización en diversos tipos de redes y codificadores, citaremos a continuación algunos de los trabajos que permitieron transformar los parámetros producto del análisis LPC (LPC y LSP) en coeficientes cepstrales; cuyos detalles ya presentamos en la Sección 3.4.1.

De esta manera, en 1974 Atal et al [9] sugieren un método por medio del cual se pueda obtener los cepstra a partir de los coeficientes LPC, como una alternativa al método clásico de deconvolución homomórfica [103], como explicamos en la Sección 3.4.1.

Por otro lado, en 1986 Sugamura et al. [136] en su trabajo sobre métodos de análisis y síntesis de voz presentan una serie de principios e interpretaciones físicas que dan origen a la representación del espectro a partir de los LSP.

Como resultado de los anteriores trabajos, finalmente se pueden obtener los cepstra a partir del análisis LPC, calculando primero el espectro a partir de los LPC o los LSP. El espectro así obtenido se le denomina espectro LP y, por tanto, los cepstra calculados a partir de este espectro, se denominan LP cepstra.

En 1993, Kim et al [78] sugieren una alternativa a los cepstra obtenidos por deconvolución homomórfica o predicción lineal. En su propuesta, establecen un método por el cual se puede calcular los denominados Pseudo-Cepstra directamente a partir de los LSP. Dichos pseudo-cepstra son en realidad una aproximación a los cepstra, sin embargo resultan una alternativa computacionalmente más eficiente.

5.2.3. Discusión General de las Soluciones Existentes frente a la Distorsión por Codificación

De las soluciones anteriores podemos concluir que las bajas tasas binarias de codificación son una de las principales causas de distorsión en la señal de voz reconstruida, y que afectan de forma significativa el desempeño de los sistemas de RAH.

Una forma de reducir este efecto podría ser utilizando codificadores con mayores tasas de codificación que ofrezcan preferencia a los parámetros que contienen la información de la envolvente espectral, no solamente asignando una alta cantidad de bits sino también una alta prioridad a la hora de brindarles protección. Debido a esto, en la Sección 8.3 se exponen los resultados obtenidos en las pruebas de RAH utilizando el codificador AMR-NB bajo el ámbito de UMTS, que no solo permite el uso de una tasa binaria medianamente alta (12,2

Kbps), sino que además prioriza la protección de los bits que contienen la información más relevante para el codificador y que como veremos en esa misma sección, podemos aprovechar para el RAH.

De otro lado, el desajuste entre los datos de entrenamiento y de test pueden reducir aún más el desempeño del sistema de RAH. Por tanto, en los experimentos desarrollados en esta tesis, tanto los datos de entrenamiento como los de test se han ajustado para que utilicen las mismas condiciones de codificación (aunque no de ruido) que consideramos previsible y asumibles sus variantes.

Otra forma de prevenir el efecto de la distorsión por codificación es mediante el uso de técnicas de pre-procesado de la señal de voz con el ánimo de conseguir una mejor robustez en los parámetros obtenidos por el codificador de fuente: sin embargo, este tipo de procedimientos no son transparentes a los terminales y por tanto no hemos desarrollado propuestas en este sentido en esta tesis.

Por último, aunque la arquitectura de reconocimiento distribuido (*Distributed Speech Recognition - DSR*, por su nomenclatura en inglés) puede conseguir una reducción importante en la distorsión por codificación utilizando esquemas de parametrización y codificación orientados al RAH [106][132][161], esta arquitectura no es transparente a los terminales de la red de comunicaciones, mientras que el RMT si lo es, y dado que ésta última ha demostrado ser una aproximación buena para abordar los problemas derivados de la distorsión por codificación, ha sido la alternativa que hemos seleccionado en esta tesis para el desarrollo de nuestra propuesta de reconocimiento robusto en una red de comunicaciones.

Como veremos a continuación, existen diferentes esquemas de parametrización enmarcados dentro del RMT, y que pueden ser utilizados para enfrentar no sólo los diferentes problemas asociados a la distorsión por codificación, sino también a los generados por los errores de transmisión y/o el ruido.

5.3. Técnicas Robustas frente a Errores de Transmisión

Otro de los problemas que afecta el reconocimiento de voz codificada es el de los errores de transmisión, que se manifiesta de diferentes maneras en función de la red de comunicaciones que se utilice. Para el caso de una red de telefonía móvil, los errores producidos en la red, se evidencian como errores a nivel de bit, siendo especialmente problemáticos, los ocurridos en ráfagas, debidas entre otros aspectos, al desvanecimiento de la señal de radiofrecuencia [113]. De otro lado, para el caso de una red de conmutación de paquetes (como Internet), los errores de transmisión que mas deterioran el desempeño del sistema de reconocimiento, se deben a la pérdida de paquetes en ráfagas [116][129]; estos errores de transmisión deterioran la información contenida en las tramas de voz, de tal forma que puede ser necesario el descarte de una o más tramas consecutivas, generando una reducción importante en las prestaciones del sistema de reconocimiento.

5.3.1. Reconocimiento Mediante Transparametrización

El RMT también se muestra eficaz a la hora de combatir los errores de transmisión, como se demuestra en [113] con dos escenarios de aplicación que verifican la robustez de la solución propuesta bajo diferentes condiciones. A continuación se describen los trabajos desarrollados en su investigación, así como los resultados obtenidos, ya que son el punto de partida de nuestra investigación, para luego pasar a describir en otras subsecciones, otros trabajos relacionados.

Para empezar, hay que destacar que el RMT no sólo implica un procedimiento de transformación de parámetros con miras al cálculo de los coeficientes cepstrales (que a pesar de tener un esquema pre-establecido, debe ser adaptado a las condiciones de cada entorno de aplicación, tal como se describió en la Sección 4.6), sino que además introduce algunas metodologías con el fin de adaptar la conformación del vector de características a los requerimientos y limitaciones propias de la transparametrización. Esto implica sacar provecho de algunas ventajas disponibles en el sistema de reconocimiento, pero que no están disponibles en el codificador, pues las limitaciones de retardo y coste computacional presentes en la parametrización llevada a cabo por codificador, no son las mismas que tiene un reconocedor.

Por lo anterior, el RMT ajusta no solo el proceso de transformación de parámetros que conducen a los MFCC, sino también el procesado que conlleva la construcción del vector de características, ya que aparte de los MFCC, a menudo se incluyen otros parámetros que dan robustez al sistema de RAH. En este sentido, uno de los parámetros mas frecuentemente utilizados para acompañar a los coeficientes cepstrales, es el de la energía, que, aunque en algunos casos se pueda obtener directamente del bitstream, hay ocasiones en las cuales debe ser estimada, y para ello se utilizan no solo los parámetros que describen la envolvente espectral, sino también los que modelan las componentes de la excitación. Por lo tanto, el RMT plantea diversas alternativas para sacar provecho de todos los parámetros codificados en el bitstream y de las condiciones favorables del reconocedor para su posterior procesamiento.

Para evaluar las prestaciones del RMT, en [113] utilizan un reconocedor de dígitos aislados (IDR - Isolated Digit Recognizer) y un reconocedor de habla continua (CSR - Continuous Speech Recognition), en ambos casos independiente de locutor. Por otra parte, el sistema de reconocimiento contempla dos escenarios que diversifican las condiciones que pueden surgir en un entorno real: VoIP y GSM.

El modelo de transmisión de voz sobre IP [114] utiliza el codificador G.723.1 y un esquema de simulación con diferentes patrones de pérdida de paquetes basado en estadísticas reales de tráfico de voz en Internet. Asimismo, el entorno GSM [45] utiliza los codificadores Half Rate y Full Rate, modelando también diferentes configuraciones de canales, para lo cual se implementan los codificadores de canal respectivos para cada codificador de fuente.

Los dos escenarios no solo contemplan las distorsiones presentes por los codificadores

de cada modelo, sino que además contemplan el efecto tandem, es decir, a la distorsión introducida en la señal de voz, por el efecto acumulado de dos o más esquemas de codificación en cascada [45].

Efecto de la Codificación

Las pruebas relativas al efecto de la distorsión por codificación se realizan utilizando voz decodificada y las mismas condiciones de entrenamiento y reconocimiento (sin desajuste). Bajo este escenario se verifica la distorsión provocada por varios estándares de codificación de voz, destacando especialmente las pérdidas generadas en la tarea de habla continua. En cuanto a la robustez de los codificadores, es el estándar Full Rate el que consigue mejores resultados, lo anterior debido a su alta tasa binaria (apróx. el doble del Half Rate).

En cuanto al codificador G.723.1, las conclusiones son similares, pues la distorsión por codificación se nota especialmente en la tarea de habla continua.

Los anteriores resultados son compatibles con los obtenidos en otros estudios y que fueron analizados en la Sección 5.2.1.

Efecto de los Errores de Transmisión en GSM

Para modelar este escenario, se utilizó un modelo del canal de GSM con 6 niveles diferentes de error, y debido a las ráfagas de errores, la tasa de reconocimiento disminuye drásticamente en las dos tareas de reconocimiento (dígitos aislados y habla continua), aunque es mayor el efecto en la tarea de habla continua.

De otra parte, al realizar la comparación del RMT con el procedimiento tradicional de reconocimiento de voz decodificada, se puede apreciar la mayor robustez ofrecida por el RMT. Lo anterior se observa en las dos tareas de reconocimiento, pero especialmente en la tarea de habla continua.

Es de destacar que el entrenamiento ha sido realizado bajo condiciones de voz limpia, aunque bajo las mismas condiciones de codificación.

Los resultados del RMT son concluyentes y demuestran con claridad la robustez del RMT frente a los errores de transmisión.

Efecto de los Errores de Transmisión en IP

De igual modo que en el caso de los errores de transmisión en GSM, en el escenario de VoIP y pérdida de paquetes, el RMT se muestra más robusto que la aproximación tradicional [110]. Sin embargo, y a pesar de que la diferencia es más notable en la tarea de habla continua, no hay una diferencia tan marcada (como la vista en GSM) entre el RMT y la aproximación convencional. Lo anterior se puede explicar atendiendo a que el RMT saca partido de la potencia del decodificador de canal de GSM, que no está presente en IP.

Efecto de la Interpolación de Tramas

Otro de los aportes realizados al RMT consiste en un método alternativo de interpolación de tramas para adaptar las tasas de trama del sistema de reconocimiento a la tasa de codificación.

Los resultados muestran que el procedimiento de interpolación propuesto consigue mejorar la tasa de reconocimiento. Lo anterior se demuestra en las dos tareas de reconocimiento y utilizando tanto RMT como la aproximación convencional, aunque los mejores resultados se obtienen utilizando el RMT.

La propuesta de interpolación presentada utiliza una mayor longitud de interpolación que la utilizada por el codificador G.723.1 y eso justifica la ganancia en la tasa de reconocimiento, pues, además, los parámetros que resultan más relevantes para el reconocimiento (los LSP) tienen una alta correlación intertrama y por tanto son muy adecuados para su estima a través de la interpolación.

La mayor longitud utilizada por la interpolación propuesta en el reconocedor no incurre en un retardo excesivo para una tarea de reconocimiento, aunque si lo sería para el codificador.

Efecto de la Estima de la Energía

Otra de las ventajas de la extracción de los parámetros contenidos en el bitstream es que se pueden utilizar los que se consideren adecuados para ser incluidos en la conformación del vector de parámetros acústicos que utiliza el sistema de reconocimiento, bien sea de forma directa o por medio de una transformación.

Debido a lo anterior, y dado que la energía contribuye de forma significativa al desempeño del sistema de reconocimiento (como se corrobora en la Sección 8.2.1), el procedimiento de estima propuesto por Peláez-Moreno et al resulta muy relevante. Lo anterior se puede verificar en los resultados obtenidos por la solución de RMT que utiliza una estima de la energía y que consiguen una mayor tasa de reconocimiento respecto de la solución de RMT que no la utiliza.

Efecto de la Transcodificación

En este caso se mide la reducción en la tasa de reconocimiento producida por la codificación en cascada a la que puede ser sometida la señal de voz a su paso por una red de comunicaciones [75].

Los resultados muestran un efecto mayor en la tarea de reconocimiento de habla continua respecto del obtenido en la tarea de reconocimiento de palabras aisladas. De otro lado, la solución de RMT consigue ser más robusta, a pesar de que la transparametrización sólo es aplicada en la ultima etapa (dado que no es factible obtener el bitstream en las etapas

de codificación intermedias).

Para el caso del escenario con errores de transmisión se mantiene la tendencia anterior, resaltando aún mayor la robustez del RMT sobre la aproximación tradicional.

5.3.2. Otros Trabajos en Redes de Telefonía Móvil

A continuación se resumen algunas propuestas enmarcadas dentro del RMT. Con el ánimo de presentar la información estructurada con cierta coherencia, las diversas contribuciones se han organizado de acuerdo con el modo en que se extrae la información de la envolvente espectral.

Pseudo-Cepstrum a partir de LSP en QCELP

En 1998, Choi et al [24] fueron unos de los primeros en plantear un sistema de reconocimiento utilizando diferentes conjuntos de parámetros obtenidos a partir de LSP cuantificados y extraídos del bitstream enviado por un codificador de voz, en este caso el QCELP (Qualcomm Code-Excited Linear Prediction)[2]. En dicho trabajo demuestran la eficiencia del método de conversión, utilizando una tarea de reconocimiento de palabras aisladas, comparando el denominado pseudo-cepstrum [78] obtenido a partir de tres conjuntos de LSP (cuantificados, no cuantificados y reconstruidos).

Naturalmente, los mejores resultados se obtuvieron utilizando el pseudo-cepstrum obtenido a partir de los LSP no cuantificados, pues no llevan consigo la distorsión por codificación. Sin embargo el pseudo-cepstrum calculado a partir de los LSP cuantificados consiguió una mejor tasa de reconocimiento, respecto del obtenido a partir de los LSP reconstruidos (véase la Sección 4.5.2).

Cepstrum a partir de parámetros LAR en GSM Full Rate

De otro lado, también en 1998, Gallardo et al [46], proponen un sistema de reconocimiento para el estándar de telefonía móvil GSM en el cual se obtienen los coeficientes cepstrales a partir de los parámetros LAR (enviados en cada trama de 20 ms por el codificador Full Rate de GSM [32]). Por tanto, el cepstrum calculado utiliza solo la información de la envolvente espectral contenida en los parámetros LAR que transmite el codificador. Es de destacar que en el vector de parámetros acústicos utilizan además de los coeficientes cepstrales, un parámetro de energía calculado a partir de voz decodificada.

En las pruebas realizadas en un escenario con errores de transmisión (producidos de forma aleatoria y/o a ráfagas), demuestran que la aproximación propuesta obtiene una considerable robustez frente a la aproximación tradicional (de reconocimiento a partir de voz decodificada), especialmente en la medida en que la tasa de error de bit (Bit Error Rate - BER) aumenta.

Similares resultados se obtienen para el codificador Half-Rate de GSM [33] en [47][48].

Cepstrum a partir de LSP en IS-641

Kim et al [76][80][81] exploran también el reconocimiento a partir de voz codificada, esta vez sobre la red de comunicación celular digital IS-136, utilizando para ello el codificador IS-641 [59]. En su propuesta plantean un front-end que convierte la información espectral cuantificada en el bitstream en información cepstral para reconocimiento. No obstante, además de los coeficientes cepstrales, el vector de características utiliza una combinación de la información de sonoridad obtenida también del bitstream. En sus resultados consiguen una mayor precisión que la obtenida cuando se realiza el reconocimiento sobre voz decodificada, y muestran que es comparable con la precisión obtenida cuando el reconocimiento lo hacen sobre el cepstrum de la voz sin codificar. Las ganancias de las librerías estocástica y adaptativa se usan como información de sonoridad de la voz, y actúan como un mecanismo de decisión blanda en la clasificación de sonoridad de una trama de voz.

Cepstrum a partir de LSP en GSM

Huerta et al [60][61] también presentan una comparación entre el reconocimiento de voz decodificada y el reconocimiento de voz usando características cepstrales derivadas de los parámetros enviados por el codificador. Para ello utilizan el codificador GSM Full-Rate de 13 Kbps [32]. En su estudio, los resultados muestran que se puede obtener igual o mayor precisión en el reconocimiento realizado directamente a partir de los parámetros del codificador, respecto del reconocimiento realizado desde la señal reconstruida-decodificada. Es de destacar que utilizan información tanto de la envolvente espectral como de la señal de excitación, para el cálculo de los parámetros cepstrales.

De otro lado, en [52] Gómez et al proponen un método para obtener los parámetros de reconocimiento de Aurora [35], a partir de los parámetros extraídos del codificador EFR (*Enhanced Full Rate*) de GSM [37][71]. En este procedimiento, los coeficientes cepstrales se calculan a partir de los LSP extraídos del bitstream que transmite el codificador EFR. De igual manera la energía se estima utilizando tanto los parámetros de la excitación como los LSP extraídos del bitstream.

5.3.3. Otros Trabajos en Voz sobre Redes IP

Existen diversas soluciones que abordan este problema, que como se describe en la Sección 4.2.2, se produce por la congestión en la red y la falta de mecanismos de recuperación (tanto en los protocolos de transporte, como en los codificadores), entre otros.

Las siguientes soluciones están dirigidas tanto al sistema de reconocimiento, como a la recuperación de tramas perdidas; no obstante, están enmarcadas dentro del esquema de RMT y, por tanto, su clasificación se hará, como en la subsección anterior, de acuerdo a la envolvente espectral utilizada para el cálculo del cepstrum.

Cepstrum a partir de LSP de G.723.1

Falavigna et al [38] realizan un análisis sobre reconocimiento automático del habla con diferentes características acústicas obtenidas del codificador G.723.1 en un entorno IP. En particular, se analiza la influencia que tienen en un sistema de reconocimiento la ventana de análisis y la cuantificación de los parámetros acústicos del codificador (en [7][6], Alonso et al realizan también un análisis del tamaño óptimo de la ventana de tiempo en relación con la calidad de la voz). Los experimentos son realizados con un reconocedor de dígitos conectados y sin modelo de lenguaje. Los resultados obtenidos muestran el efecto negativo de la corta ventana de análisis que utiliza el codificador, y concluye que el efecto de la cuantificación de los parámetros acústicos es menor. Por otro lado, muestra que a bajas tasas de trama, la información de las componentes de la excitación es fundamental. Sin embargo, debido a que su propuesta plantea introducir modificaciones al codificador, puede hacer inviable su implementación en una red de comunicaciones ya establecida como es el caso de las redes de nuestro interés.

Cepstrum a partir de LSP de iLBC

En [16] Carmona et al presentan una aproximación basada en la transparametrización, en la cual obtienen el vector de características descrito por el estándar ETSI-AURORA [34] a partir de los parámetros enviados por el codificador iLBC [8] operando a una tasa de 15,2 Kbps. Sin embargo, el codificador iLBC en este modo de operación transmite sólo un conjunto de LSP por cada trama de 20 ms, mientras que el estándar AURORA lo hace cada 10 ms; por lo tanto, recurren a un procedimiento de interpolación para conseguir los dos conjuntos de LSP.

Por otro lado, la señal de excitación es reconstruida para obtener a partir de esta, una envolvente espectral adicional. De esta manera, la envolvente espectral final se obtiene como el producto de las dos envolventes espectrales: la calculada a partir de los LSP interpolados y la calculada a partir de la excitación reconstruida. La envolvente espectral resultante, es utilizada para el cálculo de 13 coeficientes cepstrales.

Finalmente, se obtiene un parámetro de energía calculado a partir de la envolvente espectral y de una ganancia derivada del la información del residuo.

Frente al problema de pérdida de paquetes, plantean un esquema de recuperación basado en la interpolación lineal de los parámetros derivados de la excitación decodificada (LSP y energía) y la repetición de los LSP interpolados.

En las pruebas realizadas comparan el desempeño de la aproximación propuesta, con otras aproximaciones basadas en el reconocimiento de voz decodificada (utilizando los codificadores iLBC a 15,2 Kbps, G.729 a 8 Kbps y AMR a 12,2 Kbps) y con el estándar AURORA de reconocimiento distribuido. Sus resultados consiguen obtener una tasa de reconocimiento muy cercana a la obtenida por AURORA y superior a las demás aproximaciones basadas en voz decodificada. En estas pruebas se utiliza un reconocedor de dígitos conectados basado en [108] en diferentes condiciones de pérdida de paquetes.

Es destacable el desempeño alcanzado con la aproximación propuesta, que a diferencia de otras vincula la información de la excitación a la envolvente espectral utilizándola para el cálculo de los MFCC y la energía. No obstante, el uso de todos los parámetros de la excitación puede ser contraproducente en otros ámbitos en donde algunos de estos parámetros (por ejemplo, los que describen la componente estocástica) son poco protegidos por la red como se explica en la Sección 7.4.2 de esta tesis. De otro lado, en los resultados expuestos no se ha tenido en cuenta el efecto del ruido de ambiente en la tasa de reconocimiento.

Pseudo-Cepstrum a partir de LSP de G.723.1 y G.729

En [115] Peláez-Moreno et al realizan una comparación de front-ends basados en la extracción de características para el reconocimiento automático del habla a partir del bitstream en un entorno IP. Los procedimientos comparados son el denominado pseudo-cepstrum [79] y el procedimiento de cálculo exacto del cepstrum. En este estudio concluyen que el pseudo-cepstrum es preferible cuando las condiciones de la red son buenas o cuando se tienen bajos recursos computacionales, mientras que el procedimiento exacto es mejor cuando las condiciones de la red llegan a ser más adversas. Para realizar el estudio, se utilizan los codificadores ITU G.723.1 [66] y G.729 [68].

Recuperación de Tramas Perdidas

En [112][111] Pelaez-Moreno et al, exponen un método de reconstrucción de paquetes perdidos basado en SVM (Support Vector Machine) y lo comparan con el método tradicional de reconstrucción de paquetes empleado por los codificadores estándar, obteniendo resultados satisfactorios.

De otra parte, en [98] Milner et al analizan el problema de reconocimiento de voz sobre Internet (bajo un esquema de reconocimiento remoto, utilizando voz decodificada con el estándar G.723.1). Sus observaciones indican que bajo condiciones de perdida de paquetes con una tasa inferior al 10 %, la reducción en el desempeño del reconocedor no es significativa, sin embargo en condiciones de pérdida de paquetes a ráfagas o a tasas de pérdida superiores al 10 % el efecto es perjudicial.

La solución propuesta consiste en un detector y estimador de tramas perdidas. Para el proceso de detección añaden un elemento de conteo de tramas, que también utilizan para reordenar la secuencia de tramas recibidas. De otro lado, el proceso de estimación se realiza utilizando interpolación polinómica, concretamente utilizando polinomios de Lagrange. Los autores concluyen que, con el esquema planteado, se consigue mantener una alta tasa de reconocimiento (cercana al 90 %) a pesar de incrementar la pérdida de paquetes a niveles del 50 %.

Por otro lado, Zhong et al [160] examinan el efecto de la pérdida de paquetes en un entorno de reconocimiento de voz codificada con el algoritmo G.729. Para realizar su análisis comparan el algoritmo de recuperación de tramas que posee el codificador, con

tres diferentes esquemas de protección: codificación repetitiva, codificación por diversidad en el tiempo y codificación por compensación temporal. Sus resultados muestran que el algoritmo de recuperación de tramas que utiliza el G.729 se ve muy afectado cuando se producen altas tasas de pérdida de paquetes; sin embargo utilizando los esquemas de protección mencionados, se puede mitigar en gran manera dicho problema. En sus resultados, concluyen que el esquema de codificación por diversidad en el tiempo es el que presenta la mejor protección de los tres métodos comparados (bajo sus condiciones experimentales).

De otra forma, en [53] Gómez et al proponen un método de entrelazado por bloques con mínima latencia para enfrentar el problema de la pérdida de paquetes. Si bien su propuesta está enmarcada bajo la arquitectura de reconocimiento distribuido, podría también ser utilizado en reconocimiento remoto bajo el entorno de voz sobre IP, modificando el terminal del cliente para que realice el procedimiento de entrelazado antes de enviar el bitstream a la red. De esta manera, se podría imitar el procedimiento de protección desigual (UEP) utilizado en UMTS para proteger solo los parámetros más relevantes para el reconocimiento con lo cual se consigue que la latencia no aumente de forma significativa. Lo anterior por supuesto, aunque no es un procedimiento transparente a la red, bajo el entorno IP es viable hacerlo teniendo en cuenta que si se utiliza un terminal software, éste puede ser actualizado con facilidad desde el sistema proveedor del servicio.

De igual manera, en [69] James et al realizan una comparación de tres métodos de estimación de vectores de características perdidos. Los métodos estudiados realizan un proceso de entrelazado con lo que distribuyen las ráfagas de paquetes de larga duración en series de pequeñas ráfagas en la secuencia de bits. Para su análisis utilizan diversas condiciones de canal, y muestran que se puede conseguir una alta ganancia en la precisión del reconocimiento usando técnicas de estimación cuando se ha conseguido reducir el tamaño de las ráfagas. Sin embargo, el uso de los métodos de entrelazado aumenta significativamente el retardo.

Carmona et al en [15], proponen una técnica de *Encubrimiento de Pérdida de Paquetes* (*Packet Loss Concealment - PLC*) basada en el método de estimación *MMSE* (*Minimum Mean Square Error*), que busca de un lado reemplazar los vectores de parámetros perdidos y, por otro lado, reconstruir aquellos que han sido afectados por la propagación de error. Para comprobar la robustez de su propuesta, utilizan los codificadores G.729 y AMR-NB (a una tasa de 12,2 Kbps) mediante diferentes técnicas de reconocimiento, alcanzando tasas cercanas a las obtenidas por el estándar AURORA de reconocimiento distribuido [34]. Algunas de las ventajas de su propuesta radican en que de un lado puede ser utilizada en cualquier técnica de reconocimiento de voz codificada, dado que la solución planteada se aplica directamente sobre el vector de características de reconocimiento y, de otro lado, se mejora la robustez del sistema de RAH mediante el proceso de reconstrucción de los vectores afectados por los errores de propagación, estos últimos generados por la alta correlación inter-trama presente en los codificadores CELP.

De modo similar, en [51] Gómez et al exponen un método para mitigar la distorsión

debida a la propagación de error que ocurre con la pérdida de paquetes en la red. En este caso plantean una técnica de corrección de errores (*Forward Error Correction - FEC*) orienta a codificadores tipo CELP. Con esta técnica se busca reducir la dependencia inter-trama presente en este tipo de codificadores, debida a la predicción de largo plazo o al uso de librerías adaptativas en la codificación de la excitación. El método FEC propuesto transmite (adicionalmente a los parámetros estándar del codificador) una señal multipulso que reemplaza la excitación de la trama previa cuando ésta se ha perdido. Debido a lo anterior, debe ser añadido al bitstream original un conjunto de bits que representan la excitación multipulso codificada, aumentando por tanto la tasa binaria equivalente del codificador. No obstante, este tipo de técnicas introduce modificaciones en el terminal y por lo tanto no son transparentes a éstos.

5.3.4. Soluciones en Otros Tipos de Redes

Ganapathiraju et al [49] presentan dos métodos para mejorar la precisión de un sistema de RAH en un entorno de voz codificada con el algoritmo CVSD (Continuously Variable Slope Delta) [127][83], el cual ha sido usado particularmente en aplicaciones militares y ahora ha sido adoptado en Bluetooth.

Por un lado, realizan un estudio de las características de los parámetros extraídos por el sistema de RAH, y buscan la relación con los parámetros calculados a partir de voz codificada con PCM. Esto lo hacen con el fin de corregir y mejorar la precisión los parámetros de reconocimiento obtenidos a partir de la voz codificada con CVSD.

En segundo lugar, comprueban que se obtiene una buena precisión en el reconocimiento hecho a partir de los parámetros extraídos directamente del bitstream, y mejoran aún mas cuando se combina con la corrección que ellos proponen.

Sin embargo, los codificadores tipo CELP presentan buena calidad de señal a tasas bajas y medias (entre 2,4 y 13 Kbps), mientras que CVSD requiere una gran tasa binaria (en torno a 16 Kbps) para conseguir similar calidad [49].

Por otro lado, dado que el codificador CVSD no pertenece a la familia de los codificadores CELP, y dadas sus diferencias fundamentales en cuanto al proceso de codificación (complejidad, retardo, y calidad), no podemos realizar una comparación objetiva con los demás trabajos; sin embargo, se ha citado en este texto porque su solución se encuadra en reconocimiento a partir de voz no decodificada.

5.3.5. Discusión General de las Aproximaciones Existentes frente a los Errores de Transmisión

Similar a lo expuesto en la sección dedicada a la distorsión por codificación, la cuantificación de los parámetros introduce una disminución en la tasa de reconocimiento cuando además existen errores de transmisión, tal como lo concluye [24]. Sin embargo, lo anterior se puede solucionar en gran medida utilizando codificadores con mayores tasas binarias, como es el caso del AMR-NB que utilizamos en nuestra propuesta de

reconocimiento bajo el entorno de UMTS, el cual además de disponer de tasas binarias medianamente altas, brinda una mayor protección a los parámetros más relevantes para la codificación y que podemos aprovechar también para el RAH, tal como se expone en la Sección 8.3.1.

De otro lado, es importante destacar en los trabajos anteriormente expuestos el uso de otros parámetros tales como energía y/o parámetros que contienen información de la excitación, los cuales se han utilizado para dar mayor robustez al sistema de RAH frente a los errores de transmisión. Ejemplos de lo anterior se pueden observar en [46][110][81][52][16][15] que utilizan la energía en el vector de características para obtener mayor robustez frente a errores bien sea, por el efecto del canal inalámbrico o bien, por la pérdida de paquetes. En este sentido, nuestra propuesta incluye también la energía como un parámetro robusto, resaltando el aporte que ésta brinda a la tasa de reconocimiento y mejorando el procedimiento de estima propuesto en [113] (el cual ha servido de base para otros procedimientos de estima como el expuesto en [15]). Estos resultados se pueden observar en la Sección 8.2.1 para el entorno de VoIP, y en la Sección 8.3 para el entorno de UMTS.

De igual manera, la información de la excitación también se ha utilizado para aumentar el desempeño del sistema de reconocimiento, bien sea para utilizarla en el procedimiento de estima de la energía como en los trabajos antes descritos, bien para incorporarla al cálculo de la envolvente espectral [16], o bien para calcular nuevos parámetros como en [76][80][81][77]. En estos últimos, se añaden dos parámetros que contienen información de las ganancias adaptativa y estocástica que brindan mayor robustez tanto con errores de transmisión como sin ellos. También en [61] Huerta et al utilizan la información de la excitación para calcular un conjunto adicional de coeficientes cepstrales que combinado con los obtenidos a partir de los parámetros LAR consiguen una mejora en el desempeño del sistema de reconocimiento.

De otra parte, existen también algunos trabajos que si bien no utilizan el RMT, si incluyen información del pitch u otros parámetros que pueden ser fácilmente calculados a partir de la señal de voz. En este sentido, Thomson et al [147] y Stephenson et al [134] añaden directamente el pitch al vector de características consiguiendo incrementar la tasa de reconocimiento. En [104] O'Shaughnessy utilizan la información de sonoridad de la voz para conseguir mayor rapidez en la tarea de reconocimiento.

Debido a lo anterior, se concluye que la información contenida en la excitación puede ser utilizada de diversas maneras para acompañar o contribuir al cálculo de los coeficientes cepstrales, resultando muy útil para brindar mayor robustez al sistema de RAH, no solo en condiciones libres de errores sino también en presencia de ellos. Por este motivo, nuestra solución propuesta para UMTS utiliza dicha información en el vector de características, buscando aprovechar no solo la información que contienen los parámetros que la representan, sino también hacer uso de la alta protección que UMTS aplica a algunos de los parámetros mediante el codificador de canal. Esta contribución se explica en detalle en la Sección 8.3.

Existen otros trabajos que plantean algunas soluciones con miras a la recuperación de las tramas perdidas [112][98][160][69]. Sin embargo, dado que estas podrían ser utilizadas como complemento a nuestra propuesta, no planteamos una solución propia de tal forma que se puede utilizar alguna de las existentes.

No obstante lo anterior, en nuestra propuesta utilizamos el filtrado de los MSP como mecanismo robusto frente a los errores de transmisión, el cual puede ser combinado con las anteriores soluciones y aumentar las posibilidades de robustez frente a dicho problema.

De otra parte, de la comparativa que se nos presenta en [115] entre el pseudo-cepstrum (calculado como una aproximación al cepstrum) y el cepstrum tradicional, bajo condiciones de pérdida de paquetes utilizando los codificadores G.723.1 y G.729; podemos observar que el cepstrum consigue un mejor desempeño que el pseudo-cepstrum, aunque este último es más eficiente computacionalmente. Es de destacar, que los resultados obtenidos para el codificador G.729 han sido obtenidos por los autores en el desarrollo de esta tesis y se pueden observar en la Sección 8.2.1. Sin embargo, dado que nuestra propuesta prioriza la robustez frente a los errores de transmisión, las demás pruebas experimentales las hemos desarrollado utilizando el cepstrum tradicional.

Por último, es de destacar que el RMT puede ser extendido a otros tipos de redes como se hace referencia en [49], aunque la adaptación sea mucho más fácil hacerla en codificadores tipo CELP. En este sentido, el estándar de telefonía móvil de cuarta generación LTE (Long Term Evolution) [118], seguirá haciendo uso de este esquema de codificación de voz, pues continuará empleando el AMR-NB [4](AMR de banda estrecha utilizado en UMTS) y también el AMR-WB [12] (AMR de banda ancha, estandarizado por la ITU-T como G.722.2 [151]) que posibilita transmitir voz con altas tasas binarias de hasta 23,85 Kbps [70].

5.4. Soluciones frente al Ruido en el Ámbito de RMT

El efecto del ruido en un sistema de reconocimiento es un factor importante a la hora de buscar alternativas robustas, en particular en entornos móviles, puesto que precisamente debido a esta movilidad la comunicación se puede ver afectada por una gran variedad de tipos de ruido. En general, estos ruidos pueden agruparse en torno a tres factores: densidad espectral de potencia, que además provoca el efecto Lombard como reacción del hablante a dicho ruido, la reverberación (caracterizada por la respuesta impulsional del canal entre la boca y el micrófono) y el eco producido por el acoplamiento acústico entre el altavoz y el micrófono del teléfono. Estos problemas no sólo perjudican el reconocimiento automático, sino también al humano.

Dado lo anterior, en esta sección solamente se describirán los trabajos que abordan el problema de robustez frente al ruido en el ámbito del reconocimiento mediante transparametrización, pues obviamente, son muchas las propuestas generales que abordan el problema del ruido y la mayoría de ellas son complementarias al RMT.

5.4.1. RMT frente al Ruido

En [78], Kim et al utilizan el pseudo-cepstrum obtenido a partir de parámetros LSP como una solución robusta frente al problema de reconocimiento de voz contaminada con ruido blanco gausiano. En su propuesta plantean un método para calcular el denominado Cepstrum Ponderado (Weighted Cepstrum) a partir de los LSP. Las pruebas son realizadas en presencia de ruido blanco utilizando un sistema de reconocimiento de sílabas dependiente de locutor. Los resultados demuestran que el Cepstrum Ponderado resulta más robusto que el cepstrum sin ponderar, especialmente para SNR bajas.

De otro lado, en [81] Kim et al los parámetros característicos se obtienen a partir de información de la envolvente espectral y de información específica del codificador. Sus resultados muestran que en condiciones de ruido ambiente, disminuye la calidad de la voz generada por el codificador y por ende el rendimiento del reconocedor. Por ello, concluye que realizando un pre-procesado a la voz ruidosa, antes de ser codificada, el rendimiento del reconocedor aumenta. Es más, también exploran la posibilidad de utilizar (de forma combinada) características provenientes de la voz decodificada, proponiendo un algoritmo de selección de características. La información específica del codificador utilizada por el front-end, es optimizada re-estimando las ganancias de las librerías estocástica y adaptativa, y la energía residual a partir de la señal residual mejorada. Sus resultados muestran que en los experimentos de reconocimiento de dígitos aislados y en los de gran vocabulario, se mejora el rendimiento del reconocimiento al aplicar el algoritmo de optimización de características, tanto en condiciones de voz limpia, como para SNR bajas.

5.4.2. Procesado de la Voz antes de ser transmitida por la Red de Comunicaciones

Dado que el ruido contamina la señal de voz antes de que ésta sea enviada a través de la red de comunicaciones, describiremos algunas soluciones que atacan el problema mediante pre-procesado de la voz. Este tipo de soluciones están orientadas a la reducción del ruido de ambiente y/o convolutivo.

En [80], Kim et al proponen una solución robusta frente al ruido y el efecto de la codificación. Su propuesta se basa en el hecho de que bajo condiciones de ruido de ambiente, el modelo de generación de voz (usado en el codificador) puede fallar y, por tanto, puede verse comprometido el cálculo de los parámetros de codificación, especialmente los que describen las componentes de la excitación.

Por tanto, para que el codificador de voz realice un análisis espectral más preciso, utilizan un algoritmo de estimación de la envolvente espectral basado en la reducción del Error Cuadrático Medio (Mean Square Error - MSE), que intenta corregir las distorsiones introducidas por el ruido. Los resultados obtenidos muestran un apreciable incremento en la tasa de reconocimiento cuando se aplica el algoritmo propuesto. La experimentación se realiza utilizando dos tipos de ruidos: de coche (“car”) y de voces (“babble”) y en ambos casos se consigue un incremento en la tasa de reconocimiento. En particular, el desempeño del sistema mejora especialmente frente al ruido de coche.

Martin et al [95] también plantean una serie de soluciones de pre-procesado de señales ruidosas con el ánimo de mejorar la estimación de los parámetros cepstrales, antes de ser sometidas a codificación con regímenes binarios bajos (la experimentación la realizan utilizando un codificador MELP a 2400 bps [96]).

5.4.3. Soluciones que utilizan Filtrado del Espectro de Modulación

A continuación se describen algunas de las técnicas más usadas en RAH para brindar mayor robustez al sistema de RAH frente a diversos tipos de distorsiones. No obstante, centraremos nuestro análisis en aquellas en las que su fundamento radica en el filtrado del espectro de modulación (*Modulation Spectrum - MS*), el cual es obtenido como el espectro de las trayectorias temporales de los coeficientes que describen la envolvente espectral. Por tanto, podemos calcular el MS a partir de las trayectorias temporales de los coeficientes cepstrales o de las salidas del banco de filtros, etc.

En la literatura se encuentran diferentes técnicas que utilizan el procesado de las trayectorias temporales de los coeficientes utilizados para describir la envolvente espectral. Entre ellas se pueden destacar CMN (*Cepstral Mean Normalization*)[43] que resta la media cepstral, CMS (*Cepstral Mean Subtraction*)[9], RASTA (*RelAtive SpecTrA*)[57] que utilizan un filtrado paso banda y en general aquellas que utilizan el filtrado del MS para obtener sus mejores bandas de frecuencia [74]. No obstante, a continuación nos centraremos en la descripción de CMN y algunas propuestas que utilizan el filtrado paso banda.

Cepstral Mean Normalization

Una de las técnicas más famosas de normalización de parámetros de reconocimiento, fue presentada por Furui en 1981 [43]. Esta se basa en el hecho de que las distorsiones convolutivas en el tiempo, se convierten en distorsiones aditivas en el dominio cepstral, así:

Cuando la voz pasa por un canal de transmisión se genera una convolución entre la señal de voz y la función de transferencia del canal, lo cual implica una distorsión convolutiva en el dominio del tiempo; sin embargo, en el dominio espectral es multiplicativa y al aplicar un logaritmo la podemos convertir en aditiva.

Por tanto, dado que los MFCC se obtienen como la Transformada de Fourier Inversa del logaritmo del espectro, las distorsiones convolutivas en el dominio temporal, se convierten en aditivas en el dominio cepstral.

De esta manera, si al cepstrum le sustraemos la media cepstral (calculada a partir de la evolución temporal del cepstrum), podemos eliminar de la señal de voz la distorsión convolutiva introducida por el canal . A este procedimiento se le denomina CMN (*Cepstral Mean Normalization*), y es uno de los métodos más utilizados para obtener parametrizaciones robustas frente a las distorsiones convolutivas. Finalmente, es de destacar que la media cepstral se suele calcular a partir de locuciones de frases completas [152][28].

En nuestros experimentos incluimos siempre esta técnica clásica, ya que hemos comprobado que mejora de forma ostensible los resultados en entornos ruidosos.

Filtrado Pasa Banda

La aplicación del filtrado sobre las trayectorias temporales de los parámetros de reconocimiento, tiene su origen en la correlación existente entre los parámetros extraídos de tramas adyacentes de voz, buscando por tanto con ello, conseguir las bandas de frecuencia que contribuyan a obtener la información de inteligibilidad de la voz, ya que esta es la que contiene la mayor parte de la información lingüística.

En [74] se plantea el uso de un filtro pasa-banda al MS para seleccionar las frecuencias que contienen la información lingüística más útil para una tarea de RAH. Después de diferentes configuraciones de parámetros y reconocedores, concluyen que el rango de frecuencias más útiles está entre 1 y 16 HZ, siendo 4 Hz, la componente dominante.

De otro lado, en [152] se concluye que el mejor filtrado se obtiene mediante una combinación de la sección paso alto del filtro RASTA, con la sección paso bajo de un filtro FIR de orden 20 y frecuencia de corte igual a 12 Hz.

Similares resultados se obtienen en otros trabajos presentes en la literatura, de modo que en nuestra propuesta utilizamos un filtrado de las trayectorias temporales de las salidas del banco de filtros para conseguir mayor robustez, tal como se expone en la Sección 8.2.2.

5.4.4. Discusión general de las aproximaciones existentes frente al ruido

En términos generales, es de destacar que en cuanto a las soluciones que realizan un pre-procesado de la señal de voz para mejorar el desempeño del sistema de RAH sobre una red de comunicaciones, bien sea frente al problema de la distorsión por codificación, los errores de transmisión o el ruido, su implementación implica introducir modificaciones en el terminal del cliente con los problemas que esto implica.

No obstante, frente al problema del ruido, la mayoría de soluciones presentes en la literatura están orientadas al pre-procesamiento y por tanto se exponen a continuación sus principales limitaciones:

En un entorno de telefonía móvil supone o bien desarrollar un terminal específico con el pre-procesado incorporado o bien agregar un componente hardware y/o software al terminal. En el primer caso, implicaría una renovación de terminales de todos los usuarios, con el alto coste que esto produce, no solo en el ámbito económico sino logístico. De otro lado, el uso de un hardware y/o software adicional involucra también costes y dificultades añadidas.

Para el caso de reconocimiento sobre una red IP, habría que introducir los cambios pertinentes en el software/hardware que sirve de terminal VoIP, o bien, desarrollar una aplicación de terminal específica que incorpore la solución de pre-procesado. No obstante,

es menos costoso introducir el pre-procesado bajo el entorno IP, que en un entorno de telefonía móvil.

Dado lo anterior, la principal desventaja de este tipo de soluciones, es que el sistema de reconocimiento implica modificaciones en los terminales o las redes y, por tanto, puede resultar muy costoso llevarlas a la práctica. Es por este motivo que no se ha desarrollado en esta tesis una propuesta en este sentido.

A pesar de lo anterior, existen otras soluciones que buscan reducir el efecto del ruido combinando la información de la voz reconstruida con la de los parámetros extraídos del bitstream [81]. Sin embargo, cuando se utiliza la voz reconstruida para obtener el vector de características, esta puede introducir distorsiones añadidas producto del uso de parámetros deteriorados y mal protegidos por la red de comunicaciones en el proceso de decodificación, ya que en un entorno real los errores de transmisión, los efectos nocivos de la codificación y el ruido, se combinan deteriorando de forma conjunta el desempeño del sistema de RAH.

Por otro lado, en la propuesta vista en [78] no se contempla el uso de otros parámetros que pueden robustecer el sistema de reconocimiento frente al ruido tales como las ganancias o el "pitch" que suelen estar bien protegidos tanto por el codificación de canal (si este existe) como por el codificador de fuente (que otorga una alta proporción de bits a dichos parámetros proporcionando una gran precisión.

De esta manera, en la Sección 6.6 se propone el uso de parametrizaciones extendidas en donde se busca aprovechar los parámetros mejor protegidos por la red de comunicaciones (en este caso UMTS) y que pueden ser incluidos de forma directa en el vector de características, o de forma indirecta para la estima de la energía. Dicha solución busca robustecer de manera conjunta al sistema de RAH frente a los problemas de ruido, errores de transmisión y distorsión por codificación.

Finalmente, es de destacar que en el RMT se pueden utilizar también otras técnicas genéricas que se aplican directamente al vector de características, para brindar una mayor robustez al sistema.

Por este motivo, nuestra propuesta incluye el uso de CMN como una alternativa clásica y robusta frente a las distorsiones convolutivas, además de una aproximación que utiliza un filtrado paso-bajo del espectro de modulación para conseguir mayor robustez frente al ruido y los errores de transmisión.

Capítulo 6

Reconocimiento Mediante Transparametrización: propuesta

6.1. Introducción

En los trabajos realizados hasta hoy, se han realizado análisis de redes particulares y algunos efectos debidos a la distorsión que sufre la voz en tales redes, así como la influencia de otros factores importantes como los errores de transmisión y el ruido. No obstante, cada uno de estos análisis abarca uno o dos de los diversos aspectos relevantes. En otras palabras, no hay en la literatura un trabajo que abarque de forma global estos factores de distorsión, que hemos considerado como los más importantes sobre las redes modernas de comunicaciones.

6.2. Propuesta de Solución Integrada y Robusta

Por lo anteriormente expuesto, en nuestro estudio, hemos analizado de forma integral las diversas maneras de abordar los problemas que afectan el reconocimiento de voz, en los nuevos entornos de las redes de comunicaciones. En particular, nos hemos centrado en los siguientes problemas:

- Distorsión de codificación.
- Ruido aditivo.
- Errores de transmisión a nivel de bit.
- Pérdida de paquetes.

Como entornos de trabajo, hemos escogido el de telefonía móvil y el de Internet, por ser dos de las redes con mayor crecimiento y despliegue. Para el caso de la telefonía móvil, centramos nuestro análisis en UMTS, por ser la tecnología sucesora de GSM y que gracias a su amplia capacidad de servicios, es hoy una de las redes de tercera generación más extendidas en el mundo [58]. Por otro lado, la transmisión de voz sobre IP está cobrando cada día mayor importancia, por lo que resulta razonable plantear soluciones robustas para

un sistema de reconocimiento de habla en este entorno.

Para realizar nuestro estudio, continuamos el trabajo desarrollado en [113] que utiliza el método llamado transparametrización, como una técnica robusta para abordar el problema de la distorsión de codificación, y que también se muestra robusta frente a los errores de transmisión, como se puede observar en los resultados obtenidos en su estudio.

Nuestro trabajo extiende los trabajos previos de reconocimiento mediante transparametrización [113] a otros codificadores de actual relevancia utilizados por la redes de telefonía móvil e IP: AMR-NB y G.729, cuyo funcionamiento se describe en el Capítulo 7.

Con respecto a la transparametrización, buscamos evitar la distorsión del proceso de codificación-decodificación, realizando una selección de los parámetros más relevantes para el reconocimiento a partir del bitstream, y con ello minimizar la probabilidad de que el procedimiento de extracción de características se vea influenciado por errores de transmisión que afectan a parámetros innecesarios para el reconocimiento.

Para completar la transparametrización como un mecanismo robusto de reconocimiento de voz codificada, es importante estudiar también la manera de obtener la energía a partir de los parámetros entregados por el codificador, pues no siempre es enviada como un parámetro de codificación. En particular, ninguno de los codificadores empleados en esta tesis y mencionados arriba, la transmiten como tal.

Por otro lado, debido a que los mecanismos de recuperación frente a errores provistos en los codificadores tienen limitaciones inherentes al proceso de codificación, tales como el retardo máximo o los métodos de interpolación, pretendemos mejorarlos adaptándolos a las características propias de un entorno de reconocimiento, que en general son más flexibles, permitiéndonos con ello, explorar nuevas alternativas para dar robustez al sistema.

6.3. Descripción del Procedimiento de Transparametrización

En [113] se explica el procedimiento conocido como Transparametrización, sin embargo, por resultar crucial en el desarrollo de esta tesis se describe brevemente a continuación. Para ello, partimos del esquema ilustrado en la Figura 6.1, en donde se pueden observar las etapas que conlleva el proceso.

1. *Extracción de Parámetros del Bitstream*:

   Para calcular la parametrización que se usará en las tareas de reconocimiento, primero se deben obtener del bitstream los parámetros codificados a partir de los cuales se obtendrá dicha parametrización. La Tabla 6.1, muestra los parámetros codificados y enviados en el bitstream por el G.729.

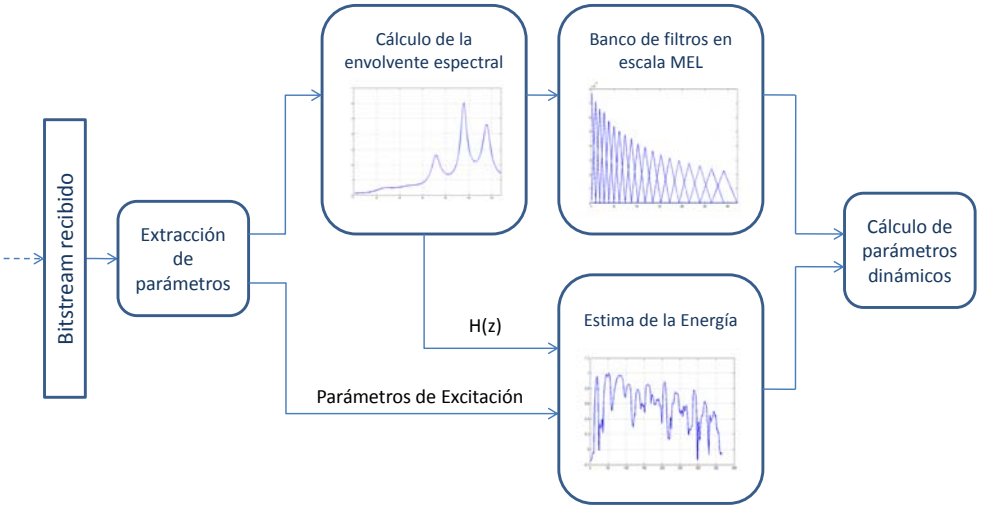


Figura 6.1: Etapas de la Transparametrización.

Parámetros enviados por el G729	
LSP	10 parámetros por trama
Pitch (periodo fundamental)	2 veces por trama
Ganancia Adaptativa	2 veces por trama
Ganancia Estocástica	2 veces por trama
Vector de códigos estocásticos	2 veces por trama

Tabla 6.1: Parámetros codificados por el G.729.

El codificador G.729 obtiene los parámetros codificados una vez cada 10 ms. Para ello, segmenta las muestras de la señal de entrada en tramas de 80 muestras (10 ms), las cuales son divididas a su vez en dos subtramas (ST1 y ST2) de 40 muestras cada una (5 ms), tal como se ilustra en la Figura 6.2.

A continuación el codificador realiza un análisis LPC una vez por cada trama, utilizando una ventana de análisis de 240 muestras (30 ms). Dicha ventana está constituida por 120 muestras de las 2 tramas anteriores, 80 muestras de la trama actual y 40 muestras de la trama siguiente (véase la Figura 6.2).

Como producto de este análisis, se obtiene un conjunto de 10 LSP (*Linear Spectrum Pairs*) por cada trama de voz codificada y que representan la envolvente espectral de dichas tramas.

Por otro lado, debido a que el anterior análisis LPC se realiza cada 80 muestras utilizando una ventana de análisis de 240 muestras, existe un solapamiento entre

ventanas que introduce un efecto de correlación entre los parámetros de codificación obtenidos para cada trama y sus adyacentes.

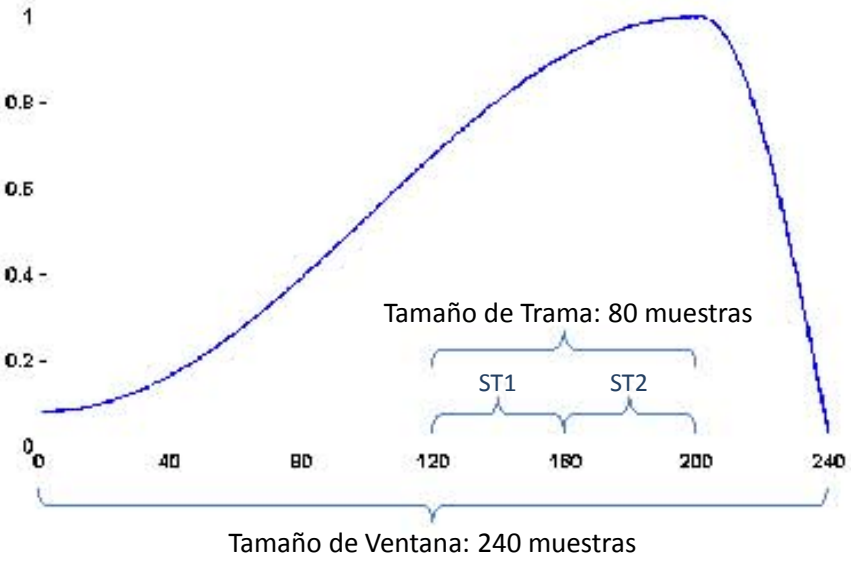


Figura 6.2: Ventana de Análisis del Codificador G.729.

De modo similar, el codificador AMR-NB (para el modo 12,2 Kbps) calcula los parámetros de codificación, segmentando la señal de entrada en tramas de 160 muestras (20 ms), las cuales son divididas a su vez, en 4 subtramas de 40 muestras (5 ms) cada una (ST1, ST2, ST3 y ST4) (véase la Figura 6.3).

En este caso, el análisis LPC se realiza dos veces por cada trama, utilizando una ventana diferente para cada análisis (W1 y W2). El tamaño de cada ventana es de 240 muestras, tal como se ilustra en la Figura 6.3.

La primera ventana (W1) se utiliza para el análisis LPC de las primeras 80 muestras de la trama (muestras 81-160 en la ventana de análisis). De igual manera, la segunda ventana (W2) se utiliza para el análisis LPC de las siguientes 80 muestras de la trama (muestras 161-240 en la ventana de análisis).

De este modo, se generan dos conjuntos de LSP (10 LSP por conjunto), que de forma similar al caso del G.729, son obtenidos como representación de la envolvente espectral para cada intervalo de 10 ms de voz.

Por otra parte, el conjunto de parámetros que se utilizan para modelar la excitación, se calculan cada 5 ms en ambos codificadores, es decir, se obtiene un conjunto de parámetros de excitación para cada subtrama de 40 muestras.

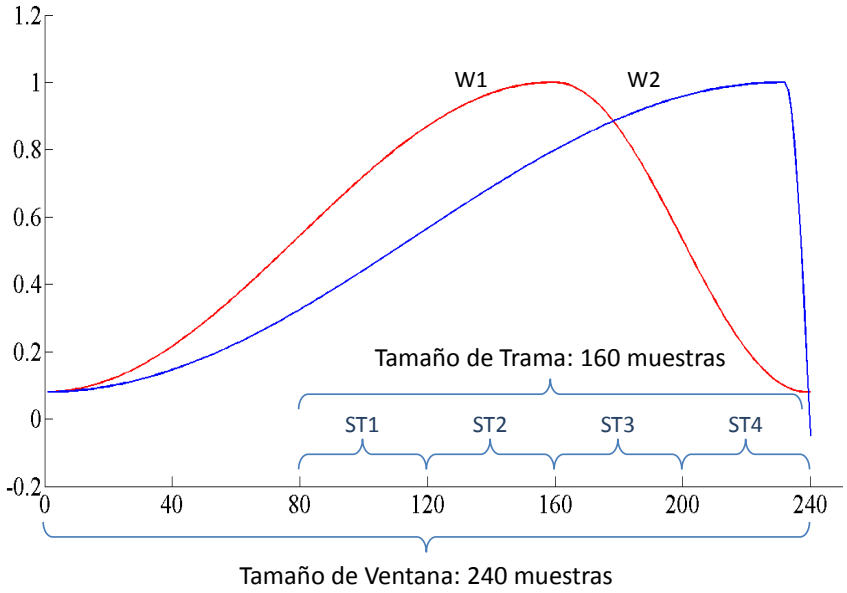


Figura 6.3: Ventanas de Análisis del Codificador AMR-NB.

Los LSP obtenidos en el análisis LPC, son cuantificados vectorialmente para ser transmitidos en el bitstream. Por tanto, en el receptor estos se extraen y se convierten en coeficientes LP para su uso posterior.

De otro lado, aunque los LPC extraídos son diferentes de los obtenidos en el análisis de predicción lineal llevado a cabo en el codificador (debido a la cuantificación y al proceso de transmisión), esta diferencia no es relevante para el desarrollo matemático que haremos a continuación, y por lo tanto no introduciremos diferencias en su notación.

2. *Cálculo de la Envolvente Espectral*:

$H(\Omega)$ se calcula cada 10 ms, a partir de los a_r extraídos del bitstream, utilizando 256 puntos para todo el espectro, y por tanto 128 puntos para la parte positiva que nos interesa, así:

$$H(k) = H\left(\Omega = \frac{2\pi k}{N}\right) = \frac{1}{1 - \sum_{r=1}^{P} a_r e^{\frac{-j2\pi kr}{N}}}, \quad N = 256, \quad k = 0, ..., 127 \quad (6.1)$$

donde $r = 1, ..., P$ siendo P el orden del filtro de síntesis, es decir, $P = 10$.

3. *Banco de Filtros Mel*:

   A continuación, el módulo de la envolvente espectral se pondera utilizando 40 filtros simétricos triangulares espaciados según la escala Mel. Después, calculamos el logaritmo a la salida de los filtros y finalmente se aplica una DCT (*Discrete Cosine Transform*) que decorrela los coeficientes. de los cuales nos quedamos con los 12 primeros.

4. *Estima de la Energía de Trama*:

   En paralelo, se realiza un procedimiento de estimación de la energía de la trama a partir de los parámetros del bitstream; sin embargo, dada la importancia de este procedimiento, este será explicado con más detalle en la Sección 6.4.

5. *Cálculo de los Parámetros Dinámicos*:

   Una vez se han obtenido los coeficientes MFCC (12 parámetros) y la estima de la energía (1 parámetro), se procede a calcular los deltas (parámetros dinámicos) para los 13 parámetros anteriores. En dicho cálculo se hace uso del procedimiento convencional utilizado por HTK, dando como resultado un vector de 26 parámetros que será utilizado finalmente en el RAH.

6.4. Estima de la Energía

En algunos codificadores, la energía es enviada como parámetro en el bitstream [113], sin embargo en los codificadores utilizados en esta tesis (G.729 y AMR-NB), no está incluida, y por tanto es necesario obtenerla, pues resulta muy relevante desde el punto de vista de reconocimiento. Sin embargo, a pesar de que se puede calcular la energía a partir de la forma de onda de la voz decodificada, debemos encontrar un procedimiento alternativo, pues una de las ventajas de la transparametrización radica en que no es necesario decodificar para obtener los parámetros de reconocimiento. Por tanto debemos realizar una estima de ésta a partir de un conjunto reducido de parámetros enviados por el codificador.

En [45] se describe un procedimiento de estima de la energía a partir de los parámetros enviados por los codificadores GSM Half Rate y Full Rate. En este caso utilizando dos tareas de reconocimiento, de un lado una de dígitos aislados y de otro lado una de reconocimiento de habla continua.

A continuación se presenta la adaptación al codificador G.729, del procedimiento propuesto en [113][45] para la estima de la energía. El procedimiento seguido para el caso del AMR-NB se omite por ser muy similar.

6.4.1. Estima de la Energía en el Codificador

La estima de la energía se hace utilizando las contribuciones de la excitación y de la envolvente espectral; sin embargo, dado que la información de la envolvente espectral se calcula cada 10 ms, la estima de energía de esta componente será realizada una vez por trama. Por otro lado, la estima de la energía de la excitación será realizada por subtramas,

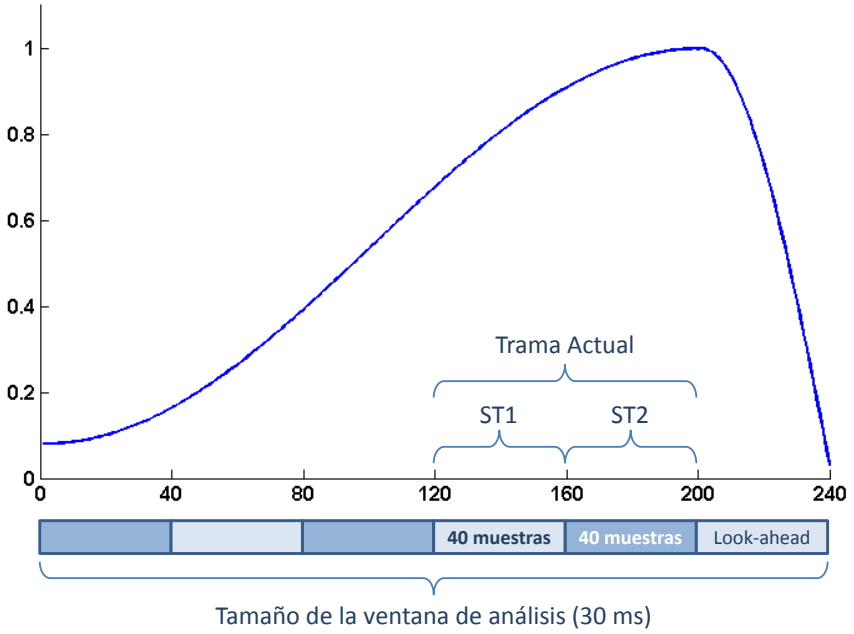


Figura 6.4: Ventana de análisis y distribución de tramas y subtramas en el codificador G.729.

De acuerdo al Teorema de Parseval, la potencia de una señal se puede obtener alternativamente en el dominio del tiempo o de la frecuencia [103]:

$$P_x = E\left\{x^2\left[n\right]\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_{xx}\left(\Omega\right) d\Omega \tag{6.2}$$

donde $\Phi_{xx}(\Omega)$ es la densidad espectral de potencia de la voz sintética.

Por otro lado, del modelo fuente-filtro explicado en la Sección 3.2, la excitación $e[n]$ para una señal de voz codificada se puede modelar como ruido blanco Gausiano de media cero y, por tanto, $\Phi_{xx}(\Omega)$ se puede expresar así:

$$\Phi_{xx}(\Omega) = \sigma_e^2 \left|H(\Omega)\right|^2 \tag{6.3}$$

donde σ_e^2 es la densidad espectral de potencia de la excitación y $H(\Omega)$ la respuesta en frecuencia del filtro de síntesis. De ésta manera, la potencia de la señal de voz se puede calcular cómo:

$$P_x = \frac{\sigma_e^2}{2\pi} \int_{-\pi}^{\pi} \left| H\left(\Omega\right) \right|^2 d\Omega \tag{6.4}$$

Sin embargo, dado que cuando se usa la transparametrización no se decodifica la voz, el proceso de síntesis de la excitación tampoco se realiza. Por tanto, debemos realizar una estima de la potencia de la excitación a partir de los parámetros que envía el codificador. Como ya hemos mencionado, los parámetros de la envolvente espectral se calculan para cada trama de 10 ms, mientras que los parámetros de la excitación se obtienen cada 5 ms. Así, la estima de la potencia media de la subtrama i correspondiente a la trama k, se puede expresar como:

$$\widehat{P}_x[k,i] = \widehat{\sigma}_e^2[k,i]\widehat{E}_h[k] \quad 0 \leq i \leq N_{st} - 1 \tag{6.5}$$

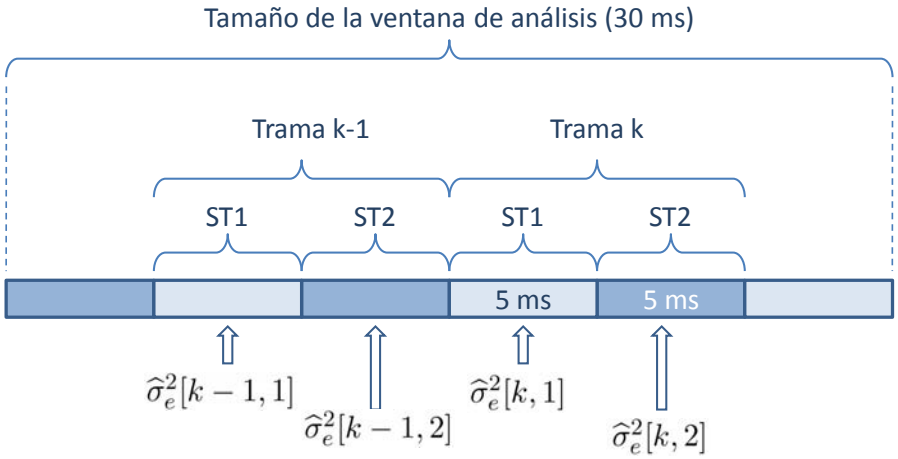


Figura 6.5: Notación usada para la potencia de cada una de las subtramas en el codificador G.729.

La energía asociada a la envolvente espectral $\widehat{E}_h[k]$ se puede expresar de forma general como:

$$\widehat{E}_h = \frac{2}{N} \sum_{r=0}^{\frac{N}{2}-1} \left| H\left[\frac{2\pi}{N} r\right] \right|^2 \quad N = 256 \; puntos \tag{6.6}$$

donde N es el número de puntos empleados para el cálculo de la FFT. En nuestro caso, utilizamos 256 por compatibilidad con el procedimiento de extracción de los MFCC utilizados como referencia.

Por otro lado, para realizar el cálculo de la potencia media de la excitación, se puede expresar $e[n]$ como la suma de una componente estocástica $\nu[n]$ y una determinista (o adaptativa) $\mu[n]$:

$$e[n] = \nu[n] + \mu[n] \tag{6.7}$$

Y suponiendo que dichas componentes no están correladas, podemos expresar la potencia media de la excitación como la suma de las potencias medias de cada componente, esto es:

$$\widehat{\sigma}_e^2 = \widehat{\sigma}_\nu^2 + \widehat{\sigma}_\mu^2 \tag{6.8}$$

Para calcular $\widehat{\sigma}_\mu^2$ recurrimos al análisis del esquema CELP [123], donde la componente determinista (o periódica) $\mu[n]$ puede escribirse como:

$$\mu[n] = G_a e[n - L_m] \tag{6.9}$$

es decir, $\mu[n]$ es proporcional a una versión retardada de la excitación, donde L_m es el retardo expresado en muestras y G_a la ganancia de la librería adaptativa, estando tanto L_m como G_a disponibles en el bitstream para cada subtrama. El número de muestras de cada subtrama es L_{st}, y para el caso del G.729 $L_{st} = 40$. Los límites de L_m están definidos por el estándar y en este caso son: $L_0 = 20$ el límite inferior y $L_f = 143$ el límite superior, esto es:

$$0 \leq n \leq L_{st} - 1 \quad , \quad L_0 \leq L_m \leq L_f$$

Por lo tanto la potencia media de la componente determinista de la excitación, para una determinada subtrama i de la trama k, se puede expresar como:

$$\widehat{\sigma}_\mu^2[k,i] = \frac{1}{L_{st}} \sum_{n=0}^{L_{st}-1} (G_a[k,i] e[n - L_m])^2 \tag{6.10}$$

donde $G_a[k,i]$ es la ganancia de la librería adaptativa para la subtrama i de la trama k, y $e[n - L_m]$ es la parte de excitación correspondiente a la trama k y la subtrama i retrasada L_m muestras.

Sin embargo, dado que en la transparametrización no es necesario decodificar la voz, tampoco se realiza la reconstrucción de la excitación $e[n - L_m]$ y por tanto no podemos estimar esta potencia media de este modo. No obstante, podemos estimarla en términos de $\widehat{\sigma}_e^2[k, i - I]$, donde I el desplazamiento en subtramas equivalente a L_m muestras; así:

$$\widehat{\sigma}_\mu^2[k,i] = G_a^2[k,i] \widehat{\sigma}_e^2 \Big[k, i - I \Big] \tag{6.11}$$

Pero en este caso, dado que $\widehat{\sigma}_e^2$ se calcula para cada subtrama (40 muestras), el retardo equivalente a L_m expresado en subtramas vendría dada por el coeficiente: $\frac{L_m}{40}$. Sin embargo, este último puede o no ser un valor entero, y por tanto para obtener un valor equivalente de $\widehat{\sigma}_e^2$ desplazada $\frac{L_m}{40}$ muestras atrás, debemos ponderar la potencia de las subtramas pasadas a las que se refiere L_m (véase la Figura 6.6); así:

$$\widehat{\sigma}_e^2\Big[k, i-I\Big] \approx (\alpha)\widehat{\sigma}_e^2[k-K-\Delta_{K1}, i-(l+1)] + (1-\alpha)\widehat{\sigma}_e^2[k-K-\Delta_{K2}, i-l] \quad (6.12)$$

donde:

$$\alpha = frac\left\{\frac{L_m}{40}\right\} \quad (6.13)$$

$$K = int\left\{\frac{L_m}{80}\right\} \quad (6.14)$$

$$l = int\left\{\frac{res\left\{\frac{L_m}{40}\right\}}{40}\right\} \quad (6.15)$$

siendo int, la parte entera del cociente, $frac$ la parte fraccionaria, y res su residuo.

De otro lado:

$$\begin{cases} 0 & \text{si } i - (l+1) > 0 \end{cases}$$

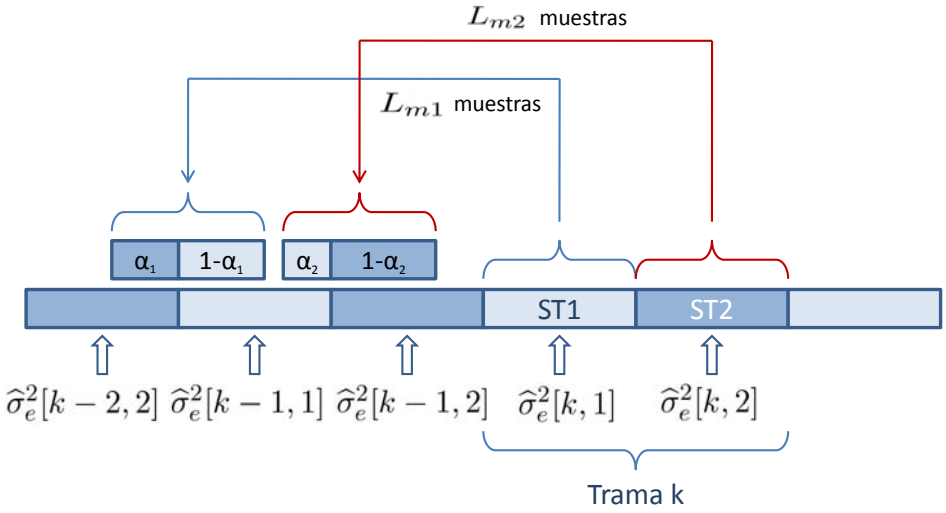


Figura 6.6: Ponderación de la potencia media de la excitación según la subtrama pasada.

Para el cálculo de potencia de la contribución estocástica $\widehat{\sigma}_\nu^2$, podemos utilizar la Ecuación 6.2, y dado que $\nu[n]$ se construye utilizando sólo 4 pulsos de amplitud unitaria en cada subtrama de 40 muestras (el resto son cero), $\widehat{\sigma}_\nu^2$ puede estimarse así:

$$\widehat{\sigma}_{\nu}^{2} = \frac{G_{e}^{2}[k,i]}{L_{st}} \sum_{j=1}^{4} P_{j}^{2} \tag{6.18}$$

donde $P_j = \pm 1$ y $G_e^2[k,i]$ es la ganancia de la librería estocástica de la trama k y la subtrama i; por tanto

$$\widehat{\sigma}_{\nu}^{2} = \frac{4G_{e}^{2}[k,i]}{L_{st}} \tag{6.19}$$

De esta manera, sustituyendo (6.12) en (6.11) y sumando (6.19), obtenemos la estima de la potencia de la excitación $\widehat{\sigma}_e^2$ para la subtrama i de la trama k:

$$\widehat{\sigma}_{e}^{2}[k,i] = G_{a}^{2}[k,i]\,(\alpha\widehat{\sigma}_{e}^{2}[k-K-\Delta_{K1}, i-(l+1)]+(1-\alpha)\widehat{\sigma}_{e}^{2}[k-K-\Delta_{K2}, i-l]) + \frac{4G_{e}^{2}[k,i]}{L_{st}} \tag{6.20}$$

Finalmente, dado que la potencia de la excitación ha sido calculada para cada subtrama, y la potencia del filtro de síntesis se calcula para cada trama, se realiza un promediado de la potencia de las dos subtramas para obtener un solo valor ($\widehat{\sigma}_e^2[k]$) para cada trama que utilizamos en (6.5) para obtener la potencia total.

Una vez hemos calculado la potencia total de cada trama, la transformamos en energía:

$$\widehat{E}_{x}[k,i] = L_{st}N_{st}\hat{P}_{x}[k,i] \tag{6.21}$$

$$\widehat{E}_{x}[k,i] = L_{st}N_{st}(\widehat{\sigma}_{e}^{2}[k]\widehat{E}_{h}[k]) \tag{6.22}$$

Aplicamos logaritmo a la Ecuación (6.22) para insertar esta log-energía en el vector de características estáticas que junto con los MFCC, serán utilizado para el reconocimiento.

$$log(\widehat{E}_{x}) = log(L_{st}N_{st}) + log(\widehat{\sigma}_{e}^{2}) + log(\widehat{E}_{h}) \tag{6.23}$$

Finalmente se utiliza el procedimiento de normalización por defecto definido por el HTK para el cálculo de la energía decodificada [158, p. 65]HTK. Este procedimiento fija los valores de la log-energía en el rango -Emin ... 1 (cuando se fija el parámetro ENORMALISE a “true”, el cual es el valor por defecto).

La implementación del proceso de normalización se realiza substrayendo el máximo valor de la log-energía (calculado para cada frase a reconocer) y añadiendo 1,0 a cada valor.

De igual manera, se puede establecer un valor mínimo de log-energía (Emin) para cada frase utilizando el parámetro SILFLOOR. Este parámetro determina una proporción entre los valores máximo y mínimo de la energía y se expresa en dB. Su valor por defecto es 50 dB.

Finalmente, la log-energía en conjunto se puede escalar utilizando el parámetro de configuración ESCALE. Su valor por defecto es 0,1.

En nuestro caso hemos implementado los anteriores procedimientos utilizando los valores por defecto sugeridos por el HTK.

6.4.2. Procedimiento Mejorado de Estima de la Energía

Si bien el método anterior ha demostrado una gran robustez al reconocimiento mediante transparametrización, en comparación con el uso de la potencia calculada a partir de la voz decodificada; se ha planteado en esta tesis una mejora en el procedimiento, el cual consiste en ponderar la estima de la potencia de la excitación, de acuerdo con el peso que da la ventana de análisis del codificador a cada subtrama.

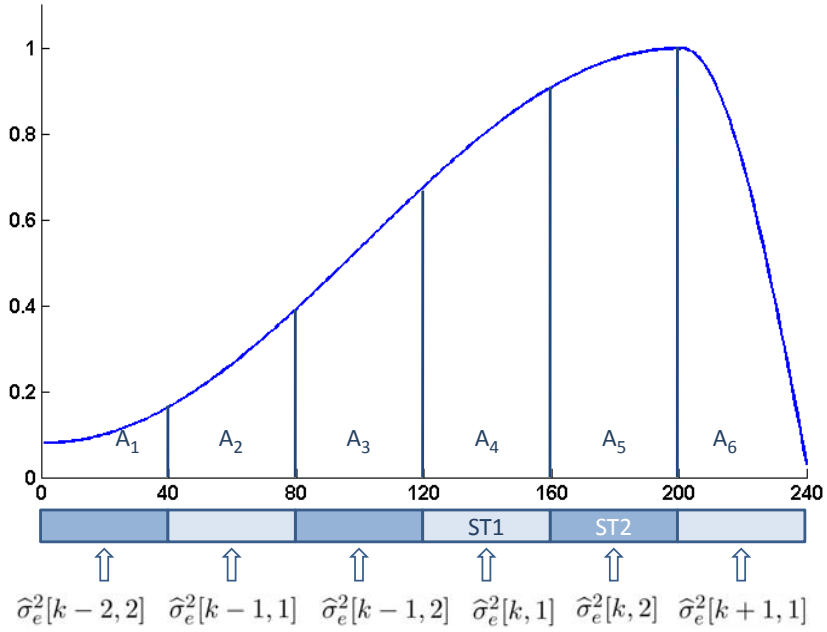


Figura 6.7: Ponderación de la potencia media de la excitación de acuerdo al peso asignado por la ventana de análisis del codificador a cada subtrama.

Para realizar la ponderación, primero se debe calcular el área bajo la curva que le corresponde a cada subtrama dentro de la ventana de análisis, como se muestra en la Figura 6.7. Por tanto si $W[n]$ representa la curva de la ventana, el área de cada subtrama A_j, para $1 \leq j \leq 6$ será:

$$A_j = \sum_{n=(j-1)L_{st}+1}^{(j)L_{st}} W[n] \tag{6.24}$$

Y por tanto, el peso w_j que la ventana de análisis asigna a cada subtrama j, según la Figura 6.7 es:

$$w_j = \frac{A_j}{\sum_{j=1}^{6} A_j} \tag{6.25}$$

A continuación, realizamos el cálculo de la potencia de la excitación para cada trama k, utilizando los pesos w_j y las potencias $\widehat{\sigma}_e^2[k,i]$ calculadas para cada subtrama i de cada trama k (Ver Ecuación (6.20)).

$$\widehat{\sigma}_e^{2\prime}[k] = \sum_{j=0}^{5} w_j \widehat{\sigma}_e^2[k, j-3] \tag{6.26}$$

De modo que la Ecuación (6.5) para cada trama k queda:

$$\widehat{P}_x[k] = \widehat{\sigma}_e^{2\prime}[k]\widehat{E}_h[k] \tag{6.27}$$

Y la log energía para cada trama:

$$log(\widehat{E}_x[k]) = log(L_{st}N_{st}\widehat{P}_x[k]) \tag{6.28}$$

6.5. Filtrado del Espectro de Modulación

Como se ha descrito en el capitulo anterior, el espectro de modulación (Modulation Spectrum - MS) no es otra cosa que el espectro generado por la evolución temporal de los parámetros utilizados para la construcción del vector de características. Y, usualmente, los parámetros más utilizados son los coeficientes cepstrales y la energía. Sin embargo, en el proceso de construcción del vector de características, se obtienen una serie de parámetros intermedios que contienen información de la envolvente espectral, bien sea representada en el dominio frecuencial o no. Ejemplo de estos parámetros son los LSP, los LPC, o las salidas del banco de filtros referidas como MELSPEC o MSP (Linear Mel-Filter Bank Channel Outputs) [158] A.1.

El filtrado del espectro de modulación busca, por un lado, preservar las componentes frecuenciales con información lingüística más relevante y, por otro, eliminar o atenuar las componentes que introducen distorsión en el proceso de reconocimiento, o que no contribuyen al mismo.

Teniendo en cuenta lo anterior, se propone un procedimiento de *Filtrado del Espectro de Modulación* que ayude a brindar mayor robustez al RMT frente a las distorsiones presentes en un sistema de RAH sobre una red de comunicaciones. En nuestro caso se realizaron diferentes esquemas para obtener el espectro de modulación, utilizando LSP, MSP y MFCC; sin embargo, como se expondrá en el Capítulo 8, los mejores resultados se obtuvieron al utilizar los MSP (salidas del banco de filtros). Por otro lado, el filtro que proponemos utilizar es un FIR paso-bajo de cuarto orden con frecuencia de corte de 6 Hz y que aplicado al espectro del modulación construido con las trayectorias temporales de los MSP, brinda

mayor robustez frente al ruido y los errores de transmisión, tal como se expondrá en la Sección 8.2.2 que explica las ventajas de su utilización.

6.6. Parametrizaciones Extendidas

Generalmente se asume que el residuo de predicción lineal contiene información no relevante para una tarea de reconocimiento independiente de locutor [121] [61]. La información relativa al residuo se representa mediante parámetros como el pitch (periodo fundamental), decisión sordo/sonora, ganancias, etc. Sin embargo, en la literatura se pueden encontrar diversos estudios que utilizan dicha información para la construcción del vector de características usado en una tarea de RAH. A continuación se describen algunos de ellos.

Huerta et al [61] utilizan la información del residuo para obtener el cepstrum a partir de los parámetros enviados en el bitstream por el codificador GSM Full Rate [32]. Si bien los resultados son inferiores a los obtenidos por el cepstrum calculado a partir de los parámetros LAR, es destacable la información contenida en el residuo.

En nuestro escenario de reconocimiento de voz sobre UMTS, en la Sección 7.4.2 se analiza la forma en que la codificación de canal bajo el esquema UEP (Unequal Error Protection) confiere una especial importancia a ciertas clases de bits (los clase A), en detrimento de las otros clases (clases B y C), y más aún respecto de los bits clase C. De otro lado, si nos centramos en la observación de los parámetros que más importancia relativa tienen en cuanto a protección, podemos concluir que son los LSP y el pitch a los que más bits clase A se les asigna (35,80 % y 34,57 % respectivamente), seguidos de las ganancias de las librerias estocástica y adaptativa (14,81 % cada una). Sin embargo a los parámetros relativos a la librería estocástica, solo se le asignan bits tipo B y C, siendo estos últimos los menos protegidos (pues no utilizan codificación convolucional) y, más se protegen en la codificación de canal, tal como se expone en la Tabla 6.2.

Parámetro	**Clase A (%)**	**Clase B (%)**	**Clase C (%)**
LSP	35,80	8,74	0,00
Periodo Fundamental (T)	34,57	0,00	3,33
Gp	14,81	2,91	1,67
Gc	14,81	7,77	0,00
Códigos Fijos	0,00	80,58	95,00
Total	**81 (100%)**	**103 (100%)**	**60 (100%)**

Tabla 6.2: Protección desigual aplicada a los parámetros del codificador AMR-NB a una tasa de 12,2 Kbps.

Por tanto, con el ánimo de sacar ventaja de la alta protección brindada al pitch y las ganancias de las librerías estocástica y adaptativa, se propone su utilización en la

construcción del vector de parámetros acústicos del reconocedor.

En la Sección 8.3.1 se pueden comprobar los efectos positivos en la tasa de reconocimiento, cuando en presencia de errores de transmisión, se incluyen este tipo de parámetros en el vector acústico del sistema de reconocimiento.

Capítulo 7

Marco Experimental: Modelos de Simulación

7.1. Introducción

Entre los factores más importantes que se han tenido en cuenta para la construcción de nuestro *Modelo de Simulación*, podemos destacar; el codificador fuente, la codificación de canal (en UMTS), el modelado del canal, las distorsiones propias de cada entorno y la implementación de las técnicas de reconocimiento, entre otros.

A continuación se realizará una descripción de las características comunes utilizadas por los modelos de simulación, para luego entrar a detallar las características específicas de cada entorno.

7.2. Protocolo de Experimentación

En esta sección se describirán los elementos comunes que han sido utilizados para realizar los experimentos planteados en esta tesis. Para empezar se describe el sistema de reconocimiento y a continuación la base de datos utilizada, haciendo énfasis en los aspectos más relevantes para el desarrollo de las pruebas descritas en el Capítulo 8.

7.2.1. Reconocedor y Base de Datos

Para el sistema de RAH, se utilizó una tarea de reconocimiento de habla continua CSR (Continuous Speech Recognition) independiente de locutor haciendo uso del software HTK (Hidden Markov Model Toolkit)[158] y empleando modelos acústicos de izquierda a derecha de 3 estados, dependientes de contexto (concretamente, trifonemas). Para el modelo del lenguaje, se utilizó una gramática sencilla de pares de palabras proporcionada en la distribución de la base de datos. Además, para realizar la síntesis de los trifonemas que no están presentes en el conjunto de entrenamiento, se utiliza un árbol de decisión de agrupamiento (Clustering) de estados [113].

La base de datos utilizada es la RM1 (Resource Management Part 1) [1], que utiliza un vocabulario de 991 palabras. Para la implementación de una tarea de reconocimiento independiente de locutor, se ha utilizado de un lado un grupo de entrenamiento compuesto por 3990 oraciones provenientes de 109 locutores y, de otro lado, un grupo de validación de 1200 oraciones, estas últimas procedentes de 40 locutores que corresponden a la compilación de los cuatro primeros conjuntos de validación oficial. Para la transcripción de los datos, se ha utilizado el diccionario SRI Resource Management incluido en la distribución.

Esta base de datos ha sido grabada a 16 KHz en condiciones limpias; sin embargo, para el desarrollo de nuestra experimentación se ha realizado un submuestrado a 8 KHz con el fin de adaptarla a los codificadores empleados en nuestros entornos de simulación (G.729 y AMR-NB).

Para las pruebas de reconocimiento de habla contaminada con ruido, se han generado 5 conjuntos de validación (uno por cada tipo de ruido), utilizando para ello una versión submuestrada de la base de datos NOISEX-92 [150]. La contaminación se realiza añadiendo cada tipo de ruido al conjunto de voz limpia del grupo de validación de la RM1, generando de esta manera 5 grupos de validación contaminados con: “Babble” (ruido de voces), “Factory” (ruido de fábricas), “Pink” (ruido rosa), “White” (ruido blanco) y “Volvo” (ruido de coches). Esta contaminación se generó utilizando una *Relación Señal a Ruido* (Signal to Noise Ratio - SNR) de 15 dB.

7.3. Reconocimiento de Habla en Internet con el G.729

Las etapas que conlleva el proceso de reconocimiento mediante Transparametrización en una red IP, se pueden observar en la Figura 7.1. En nuestro caso, el codificador de fuente modelado es el G.729, y dada la importancia de éste en todo el proceso de reconocimiento, la primera parte de esta sección se dedica a la explicación de las características más relevantes de este codificador a la hora de establecer el modelo de simulación.

De otro lado, en la segunda parte se describe el modelo adoptado para la simulación de errores de transmisión, centrando dicha descripción en la pérdida de paquetes.

Las demás etapas que intervienen en el proceso de reconocimiento y que no son propias del entorno IP, serán descritas en la Sección 7.2.

7.3.1. Codificador G.729

El codificador de voz ITU-T G.729 [68][126] es uno de los más usados en el entorno IP, y la ITU lo define como un codificador de voz mediante predicción lineal con excitación por código algebraico de estructura conjugada, (Conjugate Structure-Algebraic Code Excited Linear Prediction (CS-ACELP)[123][124]. Este codificador trabaja a dos tasas (incluyendo el Anexo E del estándar): 8 y 11,8 Kbps, en nuestro caso lo usamos a 8 Kbps, por ser una

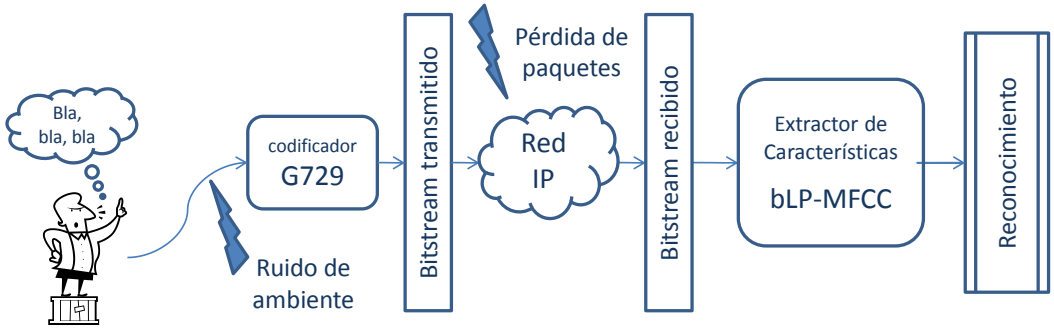


Figura 7.1: Modelo de transmisión de voz en una red IP utilizando Reconocimiento Mediante Transparametrización.

tasa más estandarizada. La señal de entrada al codificador debe ser una señal digital MIC lineal (Modulación por Impulsos Codificados, o PCM - Pulse Code Modulation, por su sigla en inglés) de 16 bits, muestreada a 8000 Hz y, la codificación se realiza en tramas de 10 ms empleando un look-ahead (o anticipación) de 5 ms.

Dado que el G.729 es un codificador CELP, éste realiza un análisis mediante predicción lineal tomando las tramas vocales correspondientes a 10 ms (80 muestras), para extraer los parámetros del modelo CELP así:

La información de la envolvente espectral se codifica en 10 parámetros LSP. Sin embargo, esta información no se transmite directamente al canal sino que se codifica diferencialmente utilizando un predictor MA (Moving Average) de orden 4 (en realidad se puede optar entre dos predictores y la elección se transmite al decodificador con un bit indicador), de forma que lo que se envía es el residuo de esta predicción. El esquema de cuantificación vectorial que se emplea para representar este residuo consta de dos etapas: la primera selecciona un vector de dimensión 10 de una librería con 128 entradas (7 bits) con el objetivo de minimizar el error entre el vector objetivo y el correspondiente cuantificado, la segunda divide el vector en dos de dimensión 5 que acceden a librerías de 32 entradas (5 bits); en este último caso el objetivo es minimizar un error cuadrático medio con una ponderación adaptativa que depende de los coeficientes LSP actuales sin cuantificar.

En cuanto a la excitación, está compuesta por una librería adaptativa y una estocástica. Cada trama se divide en 2 subtramas de 5 ms (40 muestras) cada una, luego para estas subtramas se calcula el periodo del predictor de largo plazo, permitiendo valores fraccionarios del mismo de hasta $1/3$ de la resolución original. Por su parte, la librería estocástica es de tipo ACELP con 4 pistas y solo se permite un pulso por cada pista que toma valor (+1, -1). Para las 3 primeras pistas existen 8 posiciones posibles para cada pulso y por tanto se necesitan 3 bits para su cuantificación, más un bit para el signo. Sin embargo la última pista consta de 16 posibles posiciones, con lo que es necesario un bit adicional es necesario para codificarla haciendo un total de 17 bits por subtrama. La cuantificación de ambas ganancias (la estocástica y la adaptativa) se hace de forma conjunta mediante

cuantificación vectorial con una estructura conjugada (Conjugate Structure - CS) que consta de dos etapas: la primera con un cuantificador vectorial de 3 bits donde la ganancia de la librería estocástica tiene un rango de valores mayor y por tanto sesga la selección del código que se busca y la segunda con un cuantificador de 4 bits dividido en dos, donde de la misma forma que antes la selección está sesgada hacia la ganancia adaptativa.

La Tabla 7.1 resume la distribución de bits utilizada para representar cada parámetro enviado en el bitstream.

Parámetro	**Subtrama 1**	**Subtrama 2**	**Bits por trama**
LSP			18
Periodo Fundamental (T)	8	5	13
Bit de paridad del Periodo Fundamental	1		1
Índice de la Librería Estocástica	13	13	26
Bits de signo de la Librería Estocástica	4	4	8
Ganancias de las Librerías (Parte 1)	3	3	6
Ganancias de las Librerías (Parte 2)	4	4	8
Total			**80**

Tabla 7.1: Asignación binaria de los parámetros del codificador G.729.

De otro lado, es de destacar que codificador G.729 prevé mecanismos para paliar los efectos de los errores de transmisión. Sin embargo, el hecho de que no se transmitan los parámetros espectrales directamente, sino el residuo de la predicción, hace que el codificador dependa no sólo de la trama que se está procesando, sino también de su propio estado (definido por las tramas anteriores). Esto hace que aún cuando las tramas comienzan a llegar correctamente después de haber tenido lugar una pérdida, el codificador tarde en recuperarse.

El comportamiento general del codificador G.729 frente a los errores de transmisión es el siguiente: repite los parámetros LSP de la trama anterior y también las ganancias de las librerías adaptativa y estocástica (aunque con una atenuación progresiva). La clasificación sonora/sorda proviene también de la realizada en la trama anterior; en el primer caso, sólo se tiene en cuenta la excitación adaptativa (con el pitch de la trama anterior) mientras que en el segundo, sólo se reproduce la contribución de la librería estocástica. Como ya hemos mencionado, este codificador tiene la dificultad añadida de que los parámetros espectrales (en este caso los LSP) se codifican de forma diferencial, y por lo tanto la pérdida de paquetes no sólo afecta a la trama o tramas perdidas, si no que se prolonga hasta que el codificador consigue resincronizarse.

A continuación pasaremos a explicar el modelo utilizado para describir el entorno de VoIP, y los errores de transmisión que se producen. Para conseguirlo, se ha utilizado un modelo probabilístico, que basado en observaciones reales, nos permite moldear diferentes escenarios de pérdida de paquetes utilizando diferentes configuraciones de canales.

7.3.2. Modelo de Errores de Transmisión en VoIP

Dado que Internet es una red altamente heterogénea, es muy difícil modelar todos los posibles escenarios de transmisión de voz que se pueden dar. Sin embargo, Peláez-Moreno et al [113] plantean el uso de los resultados experimentales descritos por Borella et al [13], junto con el modelo de Gilbert de dos estados de la Figura 7.2 [73][85] para obtener un modelo de canal, que permite simular diversos escenarios y condiciones en una transmisión de VoIP. Po lo anterior, este mismo modelo será el que utilizamos en esta tesis.

Como hemos explicado en capítulos anteriores, las pérdidas de paquetes en una red IP se producen generalmente en ráfagas. Por tanto el modelo de Gilbert de dos estados de la Figura 7.2 es idóneo para se utilizado en el modelado de errores de transmisión en ráfagas, como explicaremos a continuación [73].

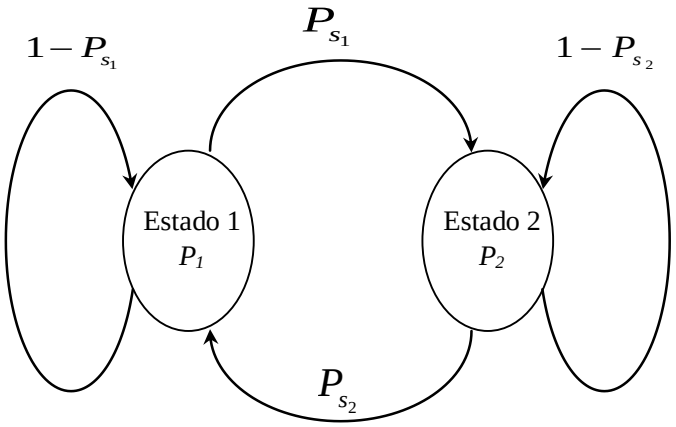


Figura 7.2: Modelo de Gilbert de 2 estados.

Cuando el modelo se encuentra en el estado 1 podemos considerar que la probabilidad de pérdida de paquetes P_1 es baja y, por tanto, representar este estado como un *estado bueno* del canal. De otro lado, si la probabilidad de pérdida de paquetes en el estado 2 P_2 es alta, este puede representar un *estado malo* del canal. De esta manera se debe cumplir que: $P_1 << P_2$.

De otro lado, los saltos entre un estado y otro vienen dadas por las probabilidades de transición P_{s_1} y P_{s_2}. Donde P_{s1} es la probabilidad de transición del estado 1 al 2 y P_{s2} es la probabilidad de transición del estado 2 al 1. Por tanto, la probabilidad de pérdida de paquetes la podemos calcular como:

$$P_e = \frac{P_1 P_{s_2} + P_2 P_{s_1}}{P_{s_1} + P_{s_2}} \quad (1)$$

Así, cuando $P_{s_1} << (1 - P_{s_2})$ se producirán ráfagas de mayor longitud, ya que de esta forma, es muy poco verosímil caer en el estado "malo", pero una vez que se ha caído es también poco probable salir de él.

$$P_e = \frac{P_1 P_{s2} + P_2 P_{s1}}{P_{s1} + P_{s2}} \quad (7.1)$$

Así, cuando $P_{s1} << (1 - P_{s2})$ se producirán ráfagas de larga duración, pues la probabilidad de caer en el *estado malo* es muy pequeña, pero una vez se cae en dicho estado, la probabilidad de cambiar al estado bueno es baja y por tanto la permanencia en el *estado*

Características de los canales IP				
Canales	P_{s_1}	P_{s_2}	P1	P2
A	0,001	0,3	0,001	0,85
B	0,002	0,25	0,005	0,85
C	0,005	0,25	0,01	0,85
D	0,005	0,20	0,015	0,85
E	0,010	0,25	0,025	0,90
F	0,010	0,20	0,001	0,90

Tabla 1; Características comunes de los canales utilizados

malo es larga.

Para establecer el valor de los cuatro parámetros que determinan el Modelo de Gilbert (P_1, P_2, P_{s1} y P_{s2}) se han tenido en cuenta las conclusiones realizadas por Borella et al [13] que resumimos a continuación:

Gc	14,81	7,77	0,00
Códigos Fijos	0,00	80,58	95,00
Total	**81 (100 %)**	**103 (100 %)**	**60 (100 %)**

- La tasa de pérdida de paquetes (PLR -Packet Loss Rate-) debe situarse entre el 0,5 % y el 3,5 %. Sin embargo, debido a las diferencias observadas por Paxon [107] entre las PLR de Estados Unidos y Europa (en Europa se presentan PLR más altas), se ha modelado PLR un poco mayores.
$$H(z) = \frac{1}{A(z)}$$
- La longitud media de las ráfagas (MBL -Mean Burst Length-) es de 6,9 paquetes. Sin embargo, en Borella se explica que esto es debido a la aparición de ráfagas excepcionalmente largas, aunque muy poco frecuentes. Por este motivo, aquí hemos utilizado un valor más pequeño en dichas ráfagas, pues de lo contrario la tasa de reconocimiento resultaría prácticamente inútil.

Canal	Total Paquetes	Total Tramas Erróneas	Total Ráfagas de Error	Tasa Pérdida de Paquetes (PLR)	Long. Media de Ráfaga (MBL)
A	200.485	1.478	402	0,37%	3,67
B	200.485	4.646	1.454	1,16%	3,20
C	200.485	10.656	3.130	2,67%	3,40
D	200.485	14.852	4.210	3,72%	3,53
E	200.485	23.542	6.817	5,90%	3,45
F	200.485	17.478	2.994	4,38%	6,74

De lo anterior, se han establecido 6 canales con diferentes características de PLR y MBL simulando 2 tramas por paquetes. Las estadísticas de estos se pueden observar en el resumen de la Tabla 7.2. En cuanto a MBL, el canal B es el menos malo (1,60) y el canal F el peor (3,37). De otro lado, el canal E presenta la tasa más alta de PLR (5,87 %, mientras que el canal A tiene la más baja (0,37 %). En general, el canal A es el menos malo y los canales E y F resultan ser los más agresivos.

Canal	Total Paquetes	Total Ráfagas de Error	Total Paquetes Erróneos	Tasa de Error de Paquetes (PLR %)	Long. Media de Ráfaga (MBL)
A	200.485	402	739	0,37	1,84
B	200.485	1.454	2.323	1,16	1,60
C	200.485	3.130	5.328	2,66	1,70
D	200.485	4.210	7.426	3,70	1,76
E	200.485	6.817	11.771	5,87	1,73
F	200.485	2.594	8.739	4,36	3,37

Tabla 7.2: Estadísticas de los canales utilizados para el modelado de pérdida de paquetes en una red IP.

A continuación se describirán las condiciones que han sido modeladas para el entorno de reconocimiento en UMTS.

7.4. Reconocimiento de Habla en UMTS

Con el propósito de realizar una experimentación realista bajo el entorno UMTS, el desarrollo de nuestro modelo de simulación se ha basado en la descripción de la capa física del estándar 3GPP [142][145]. Lo anterior nos proporciona suficiente flexibilidad a la hora de simular diferentes escenarios de distorsiones y efectos adversos en la comunicación. Esto

último, claro está, dentro de las limitaciones implícitas de un modelo de simulación.

La necesidad del desarrollo de un modelo propio, se debe a que los trabajos de simulación disponibles en la literatura sobre UMTS están orientados hacia aplicaciones específicas y en ellos no se describen en detalle las características de dichas simulaciones [5], generalmente debido a la complejidad del sistema UMTS, pues al tener flexibilidad para soportar servicios a diferentes velocidades y configuraciones, hace que el trabajo de implementar un simulador de propósito general sea muy largo y tedioso. Por tal motivo la literatura disponible y que explica el comportamiento de la capa física de UMTS [140], se concentra en mostrar ejemplos concretos con parámetros específicos [139].

Por otra parte, a pesar de que existan trabajos de simulación de UMTS de propósito general [101], estos no son de uso público y el modo de funcionamiento no es el más adecuado para nuestro objetivo, pues su desarrollo está orientado a realizar experimentos que se apoyan en resultados obtenidos por el simulador, más que a tener la posibilidad de incluir el simulador como un bloque más del prototipo de experimentación de nuestro sistema de reconocimiento. Por este motivo, el simulador implementado no es de propósito general, aunque contempla un esquema flexible a la hora de modelar diferentes configuraciones de transmisión de voz.

Sin embargo, es de destacar que en esta sección nos centraremos en la descripción de los procedimientos del modelo que tienen relación directa con el reconocimiento de voz, en particular, los parámetros y características más importantes de los algoritmos utilizados (una exposición más detallada se encuentra en [143]).

El esquema básico del sistema UMTS se muestra en la Figura 7.3, en el se encuentran los terminales móviles, el canal de comunicaciones (en modo Downlink: DL y Uplink: UL), las antenas que forman parte de la Red de Acceso Radio (Radio Access Network - RAN), las estaciones base (Base Station - BS), llamadas Nodos B en UMTS, [153] y el Núcleo de Red (Core Network - CN).

Sin embargo, para modelar las componentes principales de UMTS nos hemos centrado en el modelado de tres partes fundamentales de la capa física de UMTS: la codificación fuente, la codificación de canal y, el uso de un modelo de canal.

De esta manera, en la Figura 7.4 podemos observar las etapas que hemos utilizado para modelar nuestro sistema de RMT bajo el entorno de UMTS.

Respecto de la codificación de fuente el codificador utilizado es el AMR-NB descrito en la especificación técnica TS 26.090 del 3GPP (3rd Generation Partnership Project) [4].

De otro lado, la codificación de canal tiene diferentes esquemas de protección que dependen del modo de acceso al medio, el tipo de enlace y la tasa de bits a la que se va a transmitir.

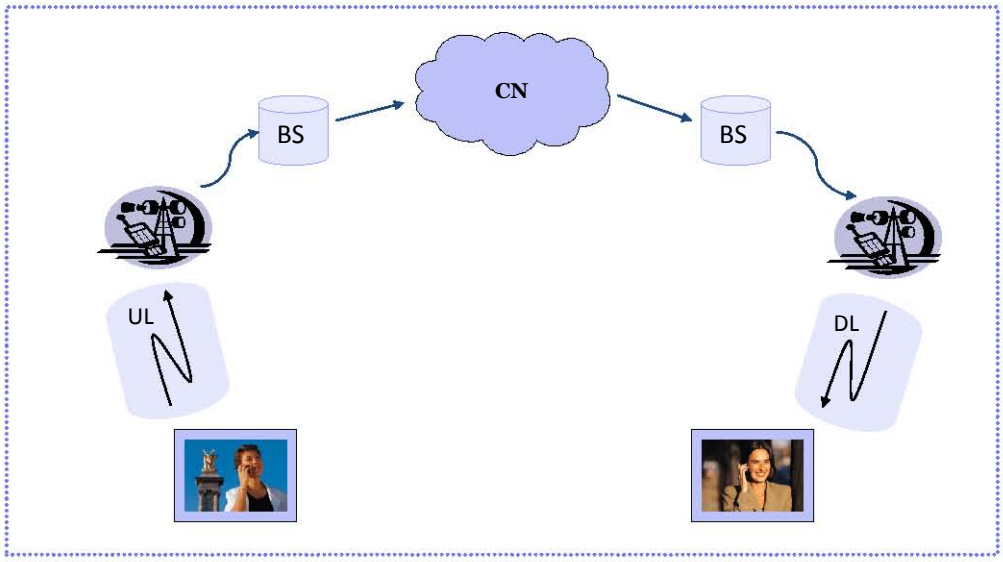


Figura 7.3: Diagrama general de la transmisión de voz en UMTS.

En UMTS el esquema de acceso a la capa física es W-CDMA, y dentro de este, se contemplan dos sistemas de acceso: FDD-UMTS (Frequency Division Duplex) [143] y TDD-UMTS (Time Division Duplex)[144], ambos estandarizados por la ETSI. No obstante, el modo FDD terminó siendo el sucesor de GSM, debido entre otros aspectos, a la cantidad del espectro asignado [153].

De otro lado, debemos tener en cuenta que en una comunicación de voz a través de una red de telefonía móvil, existen dos tipos enlaces: Ascendente (uplink) y Descendente (downlink), sin embargo, para nuestro propósito solo nos interesa modelar el enlace ascendente, pues nuestro interés se centra en reconocimiento remoto de voz, y por tanto el sistema de reconocimiento no está localizado en el terminal, sino en el lado del servidor. Esto posibilita que bajo este esquema, el servidor pueda tener comunicación directa con la red sin necesidad de utilizar un acceso a través de un enlace inalámbrico descendente y por

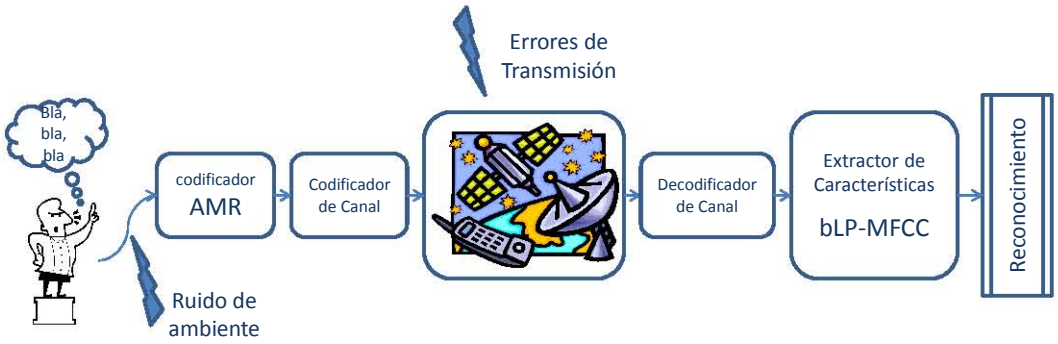


Figura 7.4: Modelo de transmisión de voz en UMTS, utilizando RMT.

tanto podemos omitirlo en nuestro modelo.

Como se explicará a continuación, el esquema de codificación de canal utilizado en UMTS es el resultado de una combinación de técnicas de detección y corrección de error. Entre las partes más importantes del proceso de codificación de canal, se destacan la codificación convolucional, los Códigos de Redundancia Cíclica (CRC), la Adaptación de Tasa (Rate Matching), y el entrelazado.

Finalmente, el modelo del canal físico utilizado en nuestra simulación fue el desarrollado por Sánchez et al [131].

A continuación pasamos a describir los elementos principales del modelo de simulación utilizado para describir UMTS.

7.4.1. Codificador AMR

Como ya se ha expuesto, el codificador utilizado en UMTS para transmitir voz es el estándar ETSI AMR-NB [4]. Este dispone de varias tasas de transmisión, que van desde 4,75 Kbps hasta 12,2 Kbps, utilizando un esquema de clasificación binaria tal como se puede observar en la Tabla 7.3.

No obstante, para nuestra simulación hemos contemplado solo la tasa de 12,2 Kbps y por tanto la descripción realizada para esta y las futuras secciones corresponde a esta tasa de transmisión. En el codificador AMR-NB existen dos formatos de trama que se pueden utilizar para la conformación del bitstream a transmitir (AMR IF1/2 - AMR Interface Format 1/2). En nuestro caso usaremos el formato IF1, pues éste determina la clasificación binaria de acuerdo a la importancia relativa que tiene cada bit en el conjunto de parámetros codificados por el AMR-NB (véase la Tabla 7.3) [129].

Tipo de trama	Modo AMR	Total de bits	Clase A	Clase B	Clase C
0	4,75	95	42	53	0
1	5,15	103	49	54	0
2	5,90	118	55	63	0
3	6,70	134	58	76	0
4	7,40	148	61	87	0
5	7,95	159	75	84	0
6	10,2	204	65	99	40
7	12,2	244	81	103	60

Tabla 7.3: Asignación binaria por clases en los diferentes modos de operación del codificador AMR-NB.

De otro lado, y de forma similar al estándar G.729, el codificador AMR-NB es también del tipo CELP y utiliza un filtro de predicción lineal de orden 10. El tamaño de trama para el modo de 12,2 Kbps es de 20 ms, por lo que el análisis LPC es desarrollado dos veces por trama, utilizando para ello dos ventanas de análisis asimétricas de 30 ms (sobre las mismas

muestras)(véase la Figura 6.3). Lo anterior busca obtener dos conjuntos de parámetros LP, que luego son convertidos en LSP para ser cuantificados conjuntamente usando una matriz SMQ (Split Matrix Quantization) de 38 bits (ver Tabla 8.1). Para el modelado de la excitación, cada trama de voz se divide en 4 subtramas de 5 ms, a partir de las cuales se calculan los parámetros de las librerías estocástica y adaptativa. Las ganancias de la librería adaptativa (Gp) y estocástica (Gc), se cuantifican con 20 y 16 bits respectivamente. El periodo fundamental (o pitch - T) es cuantificado utilizando 30 bits. No obstante la anterior asignación de bits, en UMTS se ha establecido un sistema de protección binario que establece una clasificación de bits de acuerdo a su importancia relativa (véase la Sección 7.4.2). Por otro lado, es importante resaltar aquí, que el hecho de reducir la tasa binaria influye casi exclusivamente en el modelado de la excitación, pues los bits asignados a la información de la envolvente espectral (LSP), varían muy poco entre unas tasas y otras (véase Anexo B, Tabla B.1).

Por otro lado, dentro de los procedimientos establecidos por UMTS para la protección frente a errores, merece especial importancia destacar el esquema que se describe a continuación.

7.4.2. Unequal Error Protection

UMTS utiliza un esquema de protección denominado Protección Desigual de Error (Unequal Error Protection - UEP), que clasifica de forma diferenciada los datos de entrada, basado en la importancia de estos en el destino final. De esta forma, para el caso de transmisión de voz, los bits producto del codificador fuente son agrupados en tres clases, llamadas A, B y C, siendo la clase A, la que contiene los datos de mayor importancia (y por tanto los más protegidos) y la clase C los de menor importancia.

Por ello, no todos los bits asignados a cada parámetro tienen la misma prioridad, pues su clasificación depende del tipo de parámetro al que pertenecen y de la importancia que tienen a la hora de recuperar la información en ellos almacenada. La asignación por clases realizada a los bits del codificador AMR-NB, de acuerdo al tipo de parámetro se muestra en las Tablas 7.3 y 8.1.

De la Tabla 8.1 se puede inferir, que los LSP y el pitch son los parámetros a los que más bits se le asignan, pero también los que más bits clase A utilizan y por tanto, los que serán más protegidos por la codificación de canal. En esta tabla, se pueden observar también que respecto a los bits utilizados para describir la librería estocástica, si bien utilizan la mayor parte de los bits transmitidos (140), no se protegen mucho, pues a ninguno se le asigna el nivel de mayor protección (clase A).

7.4.3. Codificación de Canal

Como se expuso al principio de esta sección, la Figura 7.4 expone las componentes principales del modelo de simulación utilizado para el reconocimiento de voz en UMTS. Sin embargo dado que la codificación de canal reviste una especial importancia por los

mecanismos de protección que utiliza, empezaremos la descripción del modelo con esta importante etapa.

Para una explicación detallada de la codificación de canal, se debería estudiar primero la definición de los canales de transporte y los canales lógicos usados en UMTS. Sin embargo, dado que no es el propósito fundamental de esta memoria, referimos para una mayor profundización a [141][142].

No obstante, y con el ánimo de ilustrar el proceso de codificación de canal de forma compacta, nos limitaremos a explicar las etapas mas importantes para nuestro trabajo, utilizando como ejemplo, el esquema de transmisión de voz sobre un canal ascendente en el modo MR122 (12,2 Kbps) del codificador AMR-NB, y que se ilustra en la Tabla 7.4).

Es de notar que los parámetros mencionados en la Tabla 7.4 se determinan en las capas altas de UMTS. De esta manera, la configuración de los Formatos de Transporte (Transport Format - TF)[141] se define en primera instancia por el tamaño del *payload*, el cual dependiendo del tipo de trama que transmita el codificador AMR-NB, puede ser: trama de voz, trama de silencio, o trama nula.

La combinación de los bloques binarios que surgen de cada tipo de trama, se especifica por los llamados Conjuntos de Combinación de Formatos de Transporte (Transport Format Combination Sets - TFCS) y, en este caso pueden ser:

- Trama de voz: (TF2, TF1, TF1).
- Trama de silencio: (TF1, TF0, TF0).
- Trama nula: (TF0, TF0, TF0).

Donde cada TFCS está compuesto por tres TFI (Indicadores de Formatos de Transporte, o Transport Format Indicator, por su nomenclatura en inglés): TF0, TF1 y TF2, los cuales representan el numero de bits por cada clase que debe tener cada combinación. Por ejemplo, en nuestro caso cuando se transmite una trama de voz, se deben utilizar los bits clase A del TF2, en conjunto con los bits clase B del TF1 y los bits clase C del TF1:

- TF0 : 1x0=0 bits clase A, 0x103=0 bits clase B y 0x60=0 bits clase C.
- TF1 : 1x39=39 bits clase A, 1x103=103 bits clase B y 1x60=60 bits clase C.
- TF2 : 1x81=81 bits clase A.

Por tanto, para una trama de voz se utilizarán: 81 bits clase A, 103 bits clase B y 60 bits clase C; para una de silencio sólo 39 bits clase A y, para una trama nula, 0 bits (aunque sólo en el payload, pues en el canal de señalización serán incluidos los bits que determinan el tipo de trama transportada). En la misma tabla, se específica la máxima tasa binaria, que para el AMR-NB es de 12.200 bps.

Capas Altas	RAB/Señalización RB		**RAB Subflujo #1 (Clase A)**	**RAB Subflujo #2 (Clase B)**	**RAB Subflujo #3 (Clase C)**
RLC (Radio Link control)	Tipo de canal lógico		DTCH (Canal de tráfico dedicado)		
	Tamaño de Payload, bit		0 39 81	103	60
	Max Tasa de datos, bps		12 200		
MAC	Cabecera MAC, bit		0		
	Multiplexado MAC		N/A		
Capa 1	Tipo de TrCH (Canal de Transp.)		DCH (Canal Dedicado)	DCH (Canal Dedicado)	DCH (Canal Dedicado)
	Tamaño de TB (Bloque de transporte), bit		0 39 81	103	60
	TFS	TF0, bits	0x81	0x103	0x60
		TF1, bits	1x39	1x103	1x60
		TF2, bits	1x81	N/A	N/A
	TTI, ms		20	20	20
	Coding type		CC 1/3	CC 1/3	CC 1/2
	CRC, bit		12	N/A	N/A
	Max número de bits/TTI después de la cod. de canal		303	333	136
	Uplink: Número Max de bits/Radio trama antes de Adaptación de velocidad		152	167	68
	RM attribute		180-220	170-210	215-256
TFCS size	3				
TFCS	(RAB subflujo#1, RAB subflujo#2, RAB subflujo#3)= (TF0, TF0, TF0), (TF1, TF0, TF0), (TF2, TF1, TF1),				

Tabla 7.4: Configuración de parámetros utilizada para la codificación de canal de un enlace ascendente y una tasa de 12,2 Kbps en el codificador AMR-NB.

Una vez establecidos los Formatos de Transporte, se determinan las variables que serán utilizadas para la codificación de canal. En este caso, se utiliza un Canal Dedicado (Dedicated CHannel - DCH) para cada clase de bits, 3 tamaños de Bloques de Transporte (Transport Block): uno para cada tipo de trama, dos tasas de Codificación Convolucional (Colvolutional Coding - CC): $R = 1/2$ y $R = 1/3$, 12 bits de CRC (Cyclic Redundancy Check) para los bits clase A y, el atributo de Ajuste de Tasa (Rate Matching - RM). A continuación se explicará con más detalle los anteriores procedimientos a aplicar en la codificación de canal.

Cálculo de CRC

Para asegurar la integridad de los datos contenidos en los Bloques de Transporte, se utilizan 12 bits de Comprobación de Redundancia Cíclica (Cyclic Redundancy Check - CRC). Sin embargo, no todos los bloques son protegidos por este método, sino sólo los que contienen la información más relevante. De esta forma, en una transmisión de voz los Bloques de Transporte protegidos son sólo los que transportan bits clase A (ver Tabla 7.3). Para conseguir lo anterior, se utiliza el siguiente polinomio generador: (véase la Figura 7.5):

$$G_{CRC12}(D) = D^{12} + D^{11} + D^3 + D^2 + D + 1 \tag{7.2}$$

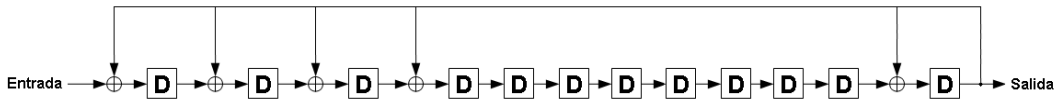


Figura 7.5: Código de Redundancia Cíclica para bits clase A.

Dado que el procedimiento de CRC es el primero que se aplica a los Bloques de Transporte elegidos (en nuestro caso, sólo el correspondiente a los bits clase A), los 12 bits calculados en esta etapa se concatenan a cada bloque para pasar a la etapa de codificación convolucional. En el extremo receptor, la integridad de estos bloques se verifica detectando posibles alteraciones (debidas fundamentalmente a errores en el canal) que no han sido corregidas por el codificador convolucional y las demás etapas de la codificación de canal. En el caso de que un bloque se muestre corrupto se pude decidir si es descartado o no.

Codificación Convolucional

Como se expuso en la Sección 7.4.2 la protección que utiliza UMTS en una transmisión de voz, es diferente para cada clase de bits emitidos por el codificador AMR-NB, brindando la mayor protección a los bits de clase A. Lo anterior lo realiza utilizando dos esquemas de codificación convolucional, asignando una tasa de codificación de 1/3 para los bits clase A y B, y de 1/2 para los bits clase C en los modos 12,2 Kbps y 10,2 Kbps y, ninguna en el resto de modos. Para proteger aún más la integridad de los bits clase A se agregan los 12 bits de CRC explicados en el apartado anterior (véase la Tabla 7.5).

La escasa protección de los bits clase C es debida a que la información que transportan, se utiliza fundamentalmente para dar mayor naturalidad a la voz reconstruida y por tanto,

Modo AMR	Tasa de Cod. Clase A	Tasa de Cod. Clase B	Tasa de Cod. Clase C
AMR_4.75	1/3	1/3	-
AMR_5.15	1/3	1/3	-
AMR_5.90	1/3	1/3	-
AMR_6.70	1/3	1/3	-
AMR_7.40	1/3	1/3	-
AMR_7.95	1/3	1/3	-
AMR_10.20	1/3	1/3	1/2
AMR_12.20	1/3	1/3	1/2

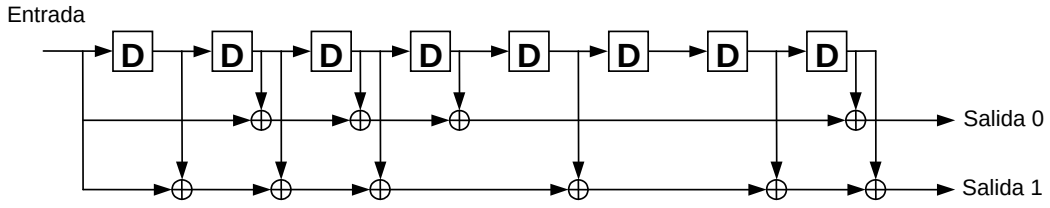
Para UEP

Tabla 7.5: Tasa de codificación por clases, para cada modo del codificador AMR-NB.

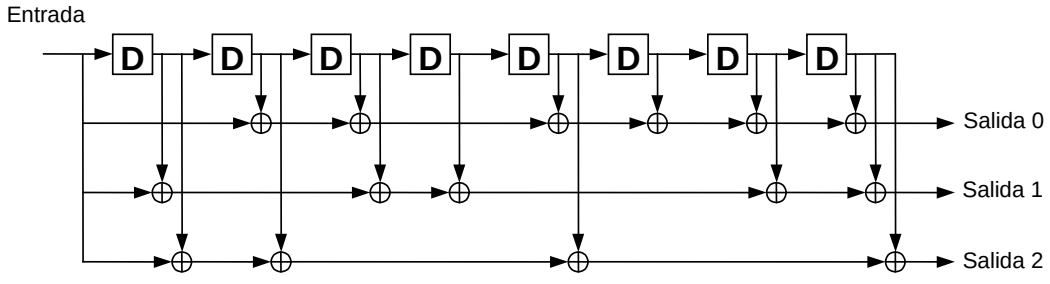
Parámetro	Clase A (%)	Clase B (%)	Clase C (%)
LSP	35,80	8,74	0,00
Periodo Fundamental (T)	34,57	0,00	3,33
Gp	14,81	2,91	1,67
Gc	14,81	7,77	0,00
Códigos Fijos	0,00	80,58	95,00
Total	**81 (100 %)**	**103 (100 %)**	**60 (100 %)**

no compromete la inteligibilidad de la voz.

En la Figura 7.6 se describen los detalles de los codificadores convolucionales implementados para las tasas de codificación de $R = 1/2$ y $R = 1/3$ utilizadas para la protección de los bits del codificador AMR-NB.



(a) Codificador Convolucional, R=1/2



(b) Codificador Convolucional, R=1/3

Figura 7.6: Codificador Convolucional para $R = 1/2$ y $R = 1/3$.

Y sus polinomios generadores, para $R = 1/2$:

$$G_{a0}(D) = D^8 + D^4 + D^3 + D^2 + 1 \tag{7.3}$$

$$G_{a1}(D) = D^8 + D^7 + D^5 + D^3 + D^2 + D + 1 \tag{7.4}$$

Y para $R = 1/3$:

$$G_{b0}(D) = D^8 + D^7 + D^6 + D^5 + D^3 + D^2 + 1 \tag{7.5}$$

$$G_{b1}(D) = D^8 + D^7 + D^4 + D^3 + D + 1 \tag{7.6}$$

$$G_{b2}(D) = D^8 + D^5 + D^2 + D + 1 \tag{7.7}$$

Como se puede observar, el número de registros (o retardos) es 8 y, por tanto, antes de ingresar cada bloque de bits al codificador convolucional se deben añadir 8 bits de cola a cada bloque. Por lo tanto, el tamaño de cada bloque de transporte después de incluir los bits de CRC y de la codificación convolucional se puede observar en la Tabla 7.6.

Modo AMR	**Clase A** **(N_A+12+8)*3**	**Clase B** **(N_A+8)*3**	**Clase B** **(N_A+8)*2**
4,75	186	183	0
5,15	207	186	0
5,90	225	213	0
6,70	234	252	0
7,40	243	285	0
7,95	285	276	0
10,2	255	321	96
12,2	303	333	136

Tabla 7.6: Bits a la salida del codificador convolucional (con bits de CRC añadidos), para cada modo del codificador AMR-NB.

Para la decodificación, se utiliza el algoritmo de Viterbi con decisión dura (Hard Decision) con un diagrama de 256 estados (2^8) y utilizando como métrica para encontrar la secuencia óptima, la distancia de Hamming.

Entrelazado

Un procedimiento relevante para la protección frente a errores, es el entrelazado. Este procedimiento se realiza a nivel de Canales de Transporte (Transport CHannel - TCH) y brinda protección adicional a los bits transmitidos, específicamente frente a la pérdida de bits en ráfagas. La especificación establece un entrelazado de bloque, realizado mediante permutaciones entre filas y columnas. La profundidad de entrelazado depende de los Formatos de Transporte y se establece en el estándar.

Rate Matching

Una vez se ha realizado el entrelazado para cada Canal de Transporte (un canal por cada clase de bits), se procede a la segmentación de éstos en tramas de radio de 10 ms (debido a que el codificador AMR-NB genera tramas de 20 ms, y el tamaño de las tramas de radio en UMTS es de 10 ms).

A continuación se realiza un procedimiento que reviste especial importancia en la multiplexación y codificación de canal, se trata del módulo de ajuste de tasa (Rate Matching). Este es el encargado de proveer la prestación de servicios a diferentes velocidades de tráfico sobre un conjunto limitado de velocidades en los canales físicos. De esta manera, se adapta la velocidad de los canales de transporte a la velocidad de los canales físicos. Esto se hace por repetición o eliminación de bits, de acuerdo a parámetros establecidos por las capas altas del sistema. En nuestro caso, este ajuste de tasa ha sido realizado de acuerdo a los formatos de transporte establecidos por el estándar para el codificador AMR-NB [146].

7.4.4. Canal

Los formatos de transporte son mapeados en el canal físico, de acuerdo al modo de acceso utilizado. En nuestro caso (como ya se explicó al principio de esta sección) el modo utilizado es el FDD. De otro lado, la asignación del canal físico depende a su vez del tipo de enlace. Por tanto, para el enlace ascendente que hemos simulado, el canal físico será definido por el procedimiento de modulación (en este caso QPSK), el factor de ensanchado (en nuestro caso 64), etc. En un enlace ascendente el canal de transporte puede ir sobre la componente en fase o sobre la componente en cuadratura de la modulación QPSK, por lo tanto, es necesario incluir también la información de fase relativa; sin embargo, los canales de transporte que contienen la información relativa la voz son asignados a la componente en fase de QPSK [131].

Para nuestro modelo de simulación, hemos utilizado un modelo de canal desarrollado en [131], adaptándolo a los canales de transporte obtenidos a partir del bitstream generado por el codificador AMR-NB.

7.5. Parametrización en las Técnicas de Reconocimiento

A continuación se describen los detalles de implementación utilizados en la parametrización de cada una de las técnicas de reconocimiento utilizadas en los entornos modelados en esta tesis.

7.5.1. Parametrización en la Técnica Decodificada

La *Técnica Decodificada* o Método Tradicional tal como se define en la Sección 4.5, se basa en el reconocimiento a partir de voz decodificada y por tanto, extrae los parámetros de reconocimiento a partir de la señal de voz reconstruida, en este caso a partir de la señal decodificada por el G.729 o el AMR-NB. Las etapas que conllevan el proceso de parametrización llevado a cabo para utilizando la herramienta HTK son las siguientes:

- Para empezar, se define el tamaño de la ventana que será aplicada a las muestras de la señal de voz reconstruida $s[n]$. En nuestro caso, hemos establecido éste en 25 ms, de tal forma que a una frecuencia de muestreo de 8000 muestras por segundo (frecuencia de la señal reconstruida), la señal enventanada $s_w[n]$ tenga un total de $N = 200$ muestras.

- A continuación se realiza un procedimiento de pre-énfasis descrito por la ecuación:

$$s'_w[n] = s_w[n] - k_p s_w[n-1] \tag{7.8}$$

donde $k_p = 0,97$ es el coeficiente de pre-énfasis, para $n = 2, ..., N$

- Para $n = 1$ se utiliza la ecuación:

$$s'_w[1] = s_w[1](1 - k_p) \tag{7.9}$$

- A continuación se aplica una ventana de Hamming:

$$s''_w[n] = s'_w[n] \left\{ 0,54 - 0,46cos\left(\frac{2\pi(n-1)}{N-1}\right)\right\} \tag{7.10}$$

- A partir de $s''_w[n]$ se calcula el llamado Espectro Mel descrito en la sección 3.4.3, el cual se caracteriza con 40 coeficientes $M(i)$ (salidas de igual número de filtros). Para $i = 1, ..., 40$.

- Después se calculan 12 coeficientes cepstrales mfc_h^{12}, aplicando la DCT a los coeficientes $M(i)$ y se obtiene 1 parámetro de energía (en este caso a partir de la voz decodificada). Finalmente se calculan los parámetros dinámicos como se describe en la Sección 3.4.5

Para referirnos a este procedimiento de parametrización se utilizará el acrónimo *MFCC (Mel Frequency Cepstral Coefficients)* de acuerdo a la notación establecida en la Tabla A.1 de los Anexos.

7.5.2. Parametrización en la Técnica del Suavizado

Similar al procedimiento descrito para la Técnica Decodificada, la parametrización de la Técnica del Suavizado parte de voz decodificada, sin embargo para obtener la envolvente espectral se realiza primero un análisis LPC tal como se describe en la Sección 4.5.2. Los detalles de las etapas de este procedimiento se describen a continuación; no obstante, los procedimientos previos de de enventanado, pre-énfasis y ventana Hamming son similares a los descritos en la Técnica del Suavizado, por lo tanto en esta explicación se han omitido.

- Una vez se realiza el procesado previo, se calculan 10 coeficientes LPC a partir de la señal $s''_w[n]$.

- A continuación, se obtiene el espectro LP utilizando los LPC del apartado anterior (de acuerdo al procedimiento descrito en la Sección 3.4.1) y se calculan los 40 coeficientes $M(i)$ (véase la Sección 3.4.3).

- Finalmente se calculan 12 coeficientes cepstrales mfc_h^{12}, aplicando la DCT a los coeficientes $M(i)$ y se obtiene 1 parámetro de energía decodificada. A continuación se obtienen los parámetros dinámicos (véase la Sección 3.4.5). Una alternativa al anterior procedimiento, es el denominado *Suavizado Extendido* que incluye junto a la energía y los 12 coeficientes cepstrales, el pitch (periodo fundamental) extraído del bitstream codificado.

Para referirnos a los dos procedimiento de parametrización aquí descritos, utilizaremos el acrónimo *LP-MFCC (Linear Prediction - MFCC)* para el procedimiento tradicional de la Técnica del Suavizado, y el acrónimo *XLP-MFCC (Extended LP-MFCC)* la versión extendida del procedimiento (véase la Tabla A.1).

Por otro lado, en los resultados que se exponen en la Sección 8.2.2 se utiliza una versión modificada de la Técnica del Suavizado, en donde se omite el procedimiento de postfiltrado llevado a cabo por el decodificador de voz. A esta aproximación se le ha dado el acrónimo *LP-MFCC**.

7.5.3. Parametrización en RMT

En este caso, para obtener los parámetros utilizados por el sistema de reconocimiento se utiliza la transformación descrita en la Sección 4.6. Sin embargo de acuerdo a las propuestas realizadas en el Capítulo 6, existen algunos procedimientos adicionales al procedimiento clásico que describiremos a continuación.

Transparametrizado Clásico

La siguiente descripción corresponde al procedimiento clásico de transparametrización propuesto por Peláez-Moreno et al en [113].

- El primer paso consiste en extraer los LSP a partir del bitstream enviado por el codificador.
- A continuación, similar a la Técnica del Suavizado, se obtiene el espectro LP a partir de los LSP del apartado anterior (véase la Sección 3.4.1) y se calculan los 40 coeficientes $M(i)$ (véase la Sección 3.4.3).
- Finalmente se calculan 12 coeficientes cepstrales mfc_h^{12}, aplicando la DCT a los coeficientes $M(i)$ y se obtiene 1 parámetro de energía estimada a partir de los parámetros extraídos del bitstream de acuerdo al procedimiento descrito en [113]. A continuación se obtienen los parámetros dinámicos (véase la Sección 3.4.5).

El acrónimo utilizado para este procedimiento es *bLP-MFCC* (véase la Tabla A.1).

PseudoCepstrum

Esta alternativa de reconocimiento utiliza también el procedimiento de transparametrización para el cálculo de los coeficientes cepstrales; sin embargo, dicho procedimiento no obtiene una representación exacta del cepstrum sino una aproximación, tal como se explica en la Sección 4.6.5.

- Similar al procedimiento clásico de transparametrización, el primer paso consiste en extraer los LSP a partir del bitstream enviado por el codificador.

- A continuación, se obtienen directamente los coeficientes cepstrales mfc_h^{12} (12 en nuestro caso) a partir de los LSP del apartado anterior, aplicando el procedimiento descrito en la Sección 4.6.5. La energía se calcula a partir de la voz decodificada, para obtener finalmente los parámetros dinámicos (véase la Sección 3.4.5).

Para esta variación del procedimiento de transparametrización se utilizará el acrónimo *pLP-MFCC* (véase la Tabla A.1).

Parametrizaciones alternativas dentro del RMT

Como se expuso en el Capítulo 6, con el ánimo de introducir mayor robustez al RMT frente a las distorsiones de ruido y/o errores de transmisión, se han propuesto en esta tesis algunas variantes al procedimiento clásico de transparametrización propuesto en [113].

Una de las propuestas consiste en utilizar el Procedimiento de Estima Mejorado descrito en la Sección 6.4. Por lo tanto, para diferenciar esta aproximación del procedimiento clásico de RMT se utilizará el acrónimo *bLP-MFCC+*.

De otro lado, cuando se aplica en el RMT el procedimiento de Filtrado del Espectro de Modulación descrito en la Sección 6.5, se utilizará el acrónimo *FbLP-MFCC* para referirnos a esta aproximación. De igual manera, cuando la anterior aproximación utiliza el Procedimiento de Estima Mejorado de la Sección 6.4 el acrónimo *FbLP-MFCC+* será el utilizado para referirnos a esta nueva aproximación.

Similar a la aproximación del Suavizado, en la Sección 6.6 se propone el uso de parámetros extra a los utilizados por la aproximación clásica de RMT. En este caso el acrónimo para describir esta parametrización extendida es *XbLP-MFCC: Xtended bLP-MFCC*. De igual manera, cuando la anterior aproximación utiliza el Procedimiento de Estima Mejorado de la Sección 6.4, se utilizará el acrónimo *XbLP-MFCC+*.

Capítulo 8

Resultados Experimentales

8.1. Introducción

A continuación se expondrán los resultados obtenidos en los diferentes entornos descritos en las secciones anteriores de esta tesis. Se describe en primer lugar los logros alcanzados en términos de reducción de la tasa de error de palabra (Word Error Rate - WER) en el entorno IP, y luego en el entorno UMTS.

Cada entorno a su vez se ha dividido en tres partes. En el caso de la transmisión de voz sobre IP (VoIP), primero se muestran los resultados considerando pérdida de paquetes y a continuación, bajo condiciones de ruido de ambiente. Finalmente se exponen los resultados obtenidos bajo el efecto combinado del ruido de ambiente y pérdida de paquetes.

De forma similar, para el caso de la transmisión de voz sobre UMTS primero se explican los resultados alcanzados con errores de transmisión y ruido, y a continuación teniendo en cuenta ambas cosas: ruido de ambiente y errores de transmisión.

La nomenclatura utilizada para describir las diferentes aproximaciones de reconocimiento se puede consultar en la Tabla A.1 de los Anexos.

8.1.1. Medidas de Confianza

Para determinar si las mejoras obtenidas resultan o no estadísticamente significativas, definimos un *Intervalo de Confianza (IC)* alrededor de cada tasa de reconocimiento de forma que podamos afirmar, en función del tamaño de las bases de datos, cómo de fiables son las conclusiones que extraigamos.

En particular, el intervalo de confianza que establecemos se calcula utilizando la siguiente ecuación [113]:

$$IC = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(100-p)}{n}} \tag{8.1}$$

donde p es el número de palabras correctamente reconocidas en la realización de un experimento, n el número de palabras que componen la base de datos (10288 en RM1) y $Z_{1-\frac{\alpha}{2}}$ es el cuantil de la distribución normal. De esta manera, si deseamos obtener un nivel de confianza del 95 % (o sea $\alpha = 0,5$), el valor del cuantil será $Z_{1-\frac{\alpha}{2}} = 1,96$.

En la Figura 8.1 se muestran los niveles de los Intervalos de Confianza en función de la tasa de error de palabra - WER para obtener un nivel de confianza del 90, 95 o 99 %.

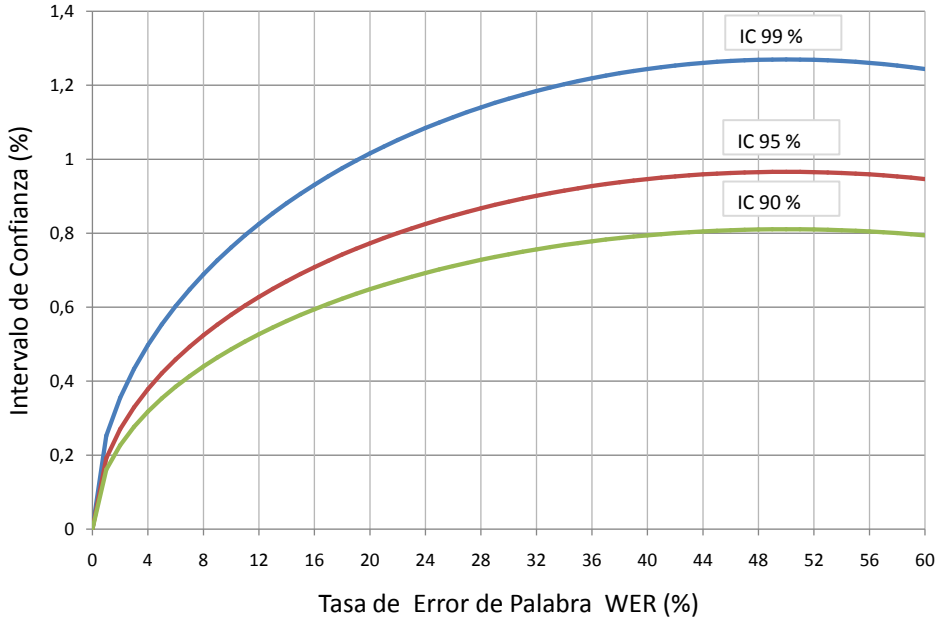


Figura 8.1: Curvas de Intervalos de Confianza de 90, 95 y 99 % para la base de datos RM1, en función de la WER.

En las figuras incluidas en este capítulo, los intervalos de confianza se han calculado con un nivel de confianza del 95 %, y aparecen en forma de "I" en las curvas de los resultados.

8.2. Reconocimiento de Habla sobre Redes IP

La Figura 8.2 muestra las etapas que conlleva el proceso de reconocimiento de habla sobre una red de paquetes. Sin embargo, como se explica en el capítulo anterior, hay algunas etapas que pueden ser omitidas cuando se utilizan soluciones basadas en reconocimiento mediante transparametrización, tal es el caso del proceso de decodificación de fuente (en este caso con el estándar G.729). En particular, en los siguientes resultados expondremos dos técnicas de este tipo: la aproximación bLP-MFCC y la pLP-MFCC de acuerdo a la nomenclatura usada para describir el procedimiento clásico de Transparametrización y el denominado Pseudo-Cepstrum respectivamente [115](Véase la nomenclatura usada para referirse a las diferentes aproximaciones de reconocimiento en la Tabla A.1). De otro lado, el procedimiento de transparametrización será comparado con dos procedimientos propios de reconocimiento de voz decodificada: los referidos como LP-MFCC y MFCC,

que corresponden a la Técnica del Suavizado y a la del Decodificado, respectivamente (ver el resumen de la nomenclatura usada en el Anexo A: Tabla A.1).

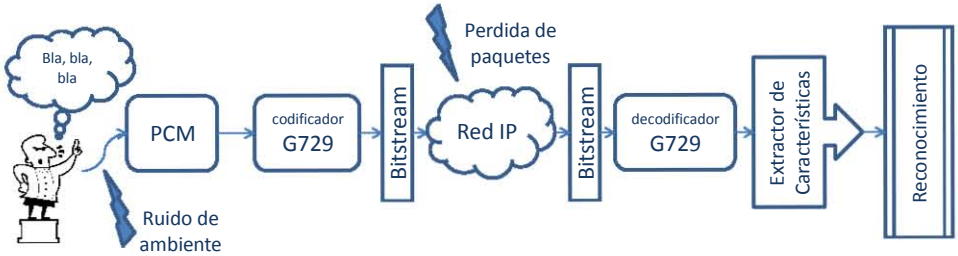


Figura 8.2: Etapas de un sistema de RAH sobre una red IP.

Como se ha detallado en el capitulo anterior, cuando se aborda la tarea de reconocimiento de voz en el entorno IP, surge una multitud de problemas que deterioran el funcionamiento del sistema. Sin embargo, como se ha demostrado en trabajos previos [113][115][45], los errores de transmisión introducen la degradación más importante en el proceso de reconocimiento. Por este motivo, y teniendo en cuenta que en este entorno los errores de transmisión se manifiestan como pérdida de paquetes, empezamos describiendo la experimentación realizada asumiendo este tipo de distorsión.

8.2.1. Pérdida de Paquetes

Las condiciones de simulación son las descritas en el capítulo anterior, en el que se explicó que se ha utilizado un modelo de Gilbert de dos estados para simular el estado del canal. La Tabla 7.2 resume las estadísticas de simulación para los canales modelados en los experimentos realizados. En este caso se ha simulado el efecto de dos tramas por paquete, sin embargo se han realizado pruebas utilizando tamaños más grandes de paquetes obteniendo resultados similares.

Efecto de la Energía en el RAH

Para empezar esta sección, primero vamos a exponer la importancia que tiene la energía dentro del proceso de extracción de características. Para ello, en la Figura 8.3 se pueden observar los resultados obtenidos cuando ésta se utiliza para la construcción del vector de características. En este caso, la técnica empleada para la comparativa es la Transparametrización. Hemos de mencionar también, que la energía utilizada en estos experimentos sigue el Procedimiento de Estima Mejorada descrito en el capitulo anterior y que se utiliza en la Transparametrización en su versión bLP-MFCC+ (ver Tabla A.1).

En esta comparativa se puede observar una importante reducción en la WER cuando se utiliza la energía. Esta reducción se produce en todos los canales simulados, siendo un poco más notable en los canales de peores condiciones (E y F), que a su vez se nota también en sus intervalos de confianza que están ampliamente separados. De esta manera, la reducción

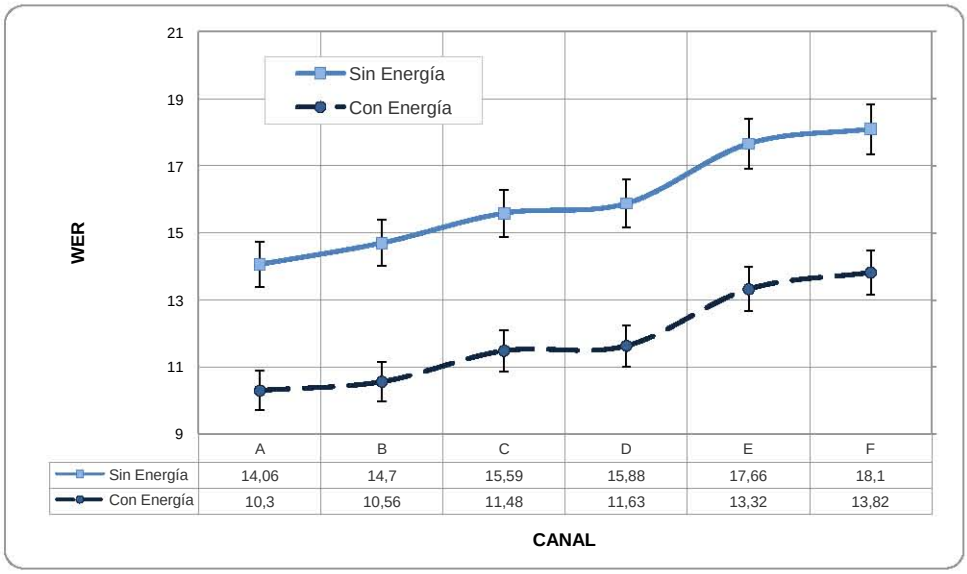


Figura 8.3: Reducción de la WER cuando se incluye la energía en el vector de características utilizando el método de Transparametrización para reconocimiento de voz sobre IP con pérdida de paquetes.

media en la WER para los seis canales simulados es de 4,16 puntos porcentuales.

De otro lado, si tenemos en cuenta que en general es más difícil reducir la WER cuando ésta es más pequeña, podemos obtener una tasa de reducción relativa (porcentaje de reducción de la WER obtenido por una técnica de reconocimiento, con respecto a la otra que está siendo comparada). Por tanto, si tomamos como referencia la media de la WER para todos los canales, en el caso de la tarea de reconocimiento sin energía la media de la WER es 15,99 % y con energía 11,85 %, lo que implica que la reducción de la WER es de un 25,9 %, cifra muy significativa teniendo en cuenta que la energía solo aporta 2 parámetros (la energía propiamente y su delta), de los 26 que se han utilizado para la construcción del vector de características.

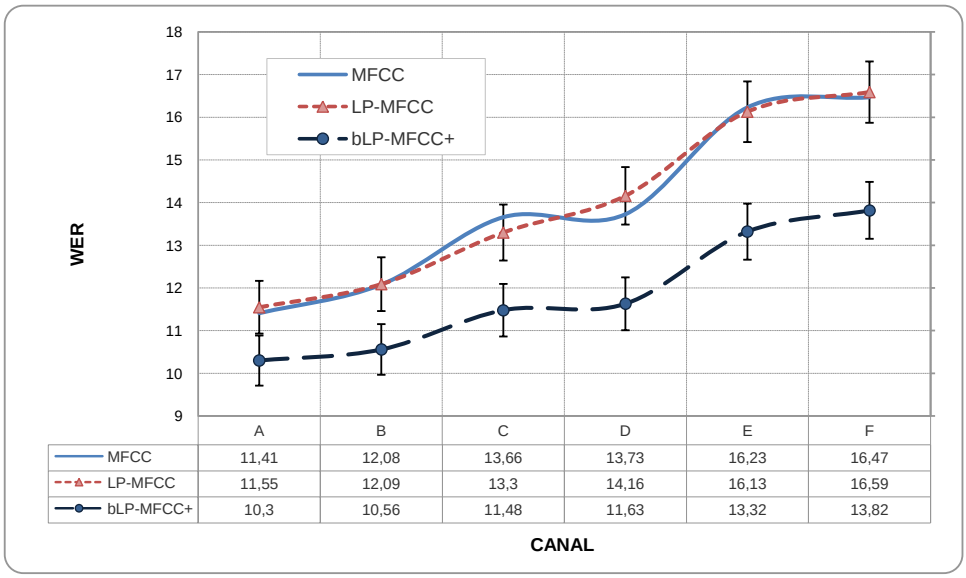
Por otro lado, si bien en los resultados expuestos en la Figura 8.3 se ha limitado la comparativa al procedimiento mediante transparametrización, a lo largo del desarrollo de esta tesis, se han realizado otras pruebas utilizando otras técnicas de reconocimiento, obteniendo también reducciones importantes en la WER. Es de destacar, que en las otras técnicas analizadas la energía utilizada para la comparación es la energía decodificada.

No obstante lo anterior, y como se puede inferir de los resultados expuestos, el procedimiento de estima de la energía tiene una especial relevancia en el reconocimiento mediante transparametrización, pues no sólo aporta una reducción significativa en la WER sino que también permite la posibilidad de obtener un procedimiento más robusto frente a los problemas intrínsecos de la Transparametrización, tal como se ha explicado en el capitulo anterior (ver Procedimiento de Estima Mejorado).

Finalmente, mencionar que en la Sección 8.2.2 se realiza una comparación de procedimientos de estima de energía en presencia de ruido de ambiente.

Comparación de Técnicas de Reconocimiento

A continuación se realiza una comparativa del reconocimiento mediante Transparametrización con las técnicas denotadas como Suavizado y Decodificado. Para el caso de la Transparametrización, la energía ha sido obtenida utilizando el Procedimiento de Estima Mejorada, y para las otras dos técnicas se ha utilizado la energía decodificada (véase la Sección 4.4).



	A	B	C	D	E	F
MFCC	11,41	12,08	13,66	13,73	16,23	16,47
LP-MFCC	11,55	12,09	13,3	14,16	16,13	16,59
bLP-MFCC+	10,3	10,56	11,48	11,63	13,32	13,82

Figura 8.4: Comparación de técnicas de reconocimiento en presencia de pérdida de paquetes.

De esta comparativa podemos observar en la Figura 8.4, que la técnica de Transparametrización consigue un desempeño superior en términos de la WER, respecto de las otras dos técnicas de reconocimiento bajo este esquema de pérdida de paquetes. Esto no es nuevo, pues similares resultados se han obtenido en trabajos previos [113][55][54]. Sin embargo, también es cierto que los resultados de la Transparametrización que aquí se presentan, tienen la novedad de incluir el Procedimiento de Estima Mejorada. En la Sección 8.2.2 se expondrán los resultados que demuestran la contribución del procedimiento de estima mejorado frente al procedimiento de estima original.

De otro lado, también se puede observar que la reducción en la WER es mayor a medida que empeoran las condiciones del canal. Es decir, la Transparametrización se torna muy robusta frente a las otras dos técnicas comparadas. En cuanto a los intervalos de confianza (IC), también se aprecia una mayor separación para los canales con más pérdidas (en esta gráfica no se han dibujado los IC del Decodificado, debido a que son similares a los del Suavizado).

Si calculamos la reducción relativa de la WER, encontramos que para el canal A (el mejor), la Transparametrización consigue una reducción de un 10,82 % respecto de la técnica del suavizado; mientras que para el canal F (el peor), la reducción llega hasta un 16,69 %. Una conclusión similar aplica a los resultados relativos la técnica decodificada, pues son muy cercanos a los del suavizado.

Pseudo Cepstrum

Como se expuso en el Capitulo 5, existe una alternativa computacionalmente más eficiente para obtener la envolvente espectral a partir de los parámetros enviados en el bitstream. Se trata del Pseudo-Cepstrum, referido como pLP-MFCC, y es una forma alternativa de realizar el proceso de transparametrización. En la siguiente figura se compara esta aproximación con el Transparametrizado clásico, referido como bLP-MFCC (véase la Figura 8.5) [115].

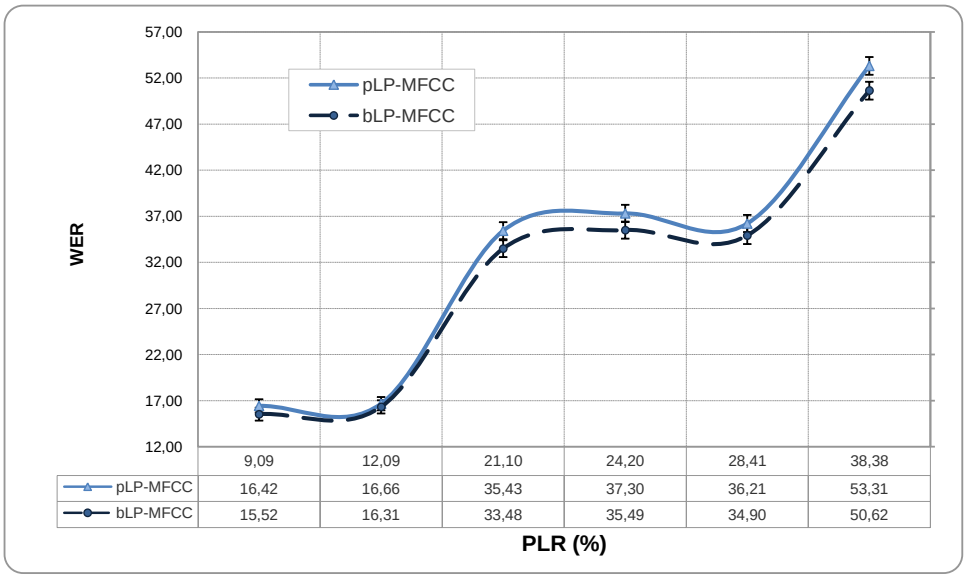


Figura 8.5: Comparativa de soluciones de Reconocimiento Mediante Transparametrización en presencia de pérdida de paquetes.

Las dos aproximaciones comparadas utilizan el procedimiento de estima de energía descrito en [114][45]. En estos resultados se han utilizado una serie de canales más agresivos que los anteriores, con el ánimo de aumentar la WER a niveles más elevados y conseguir de esta forma, ilustrar la robustez que presenta el procedimiento clásico de transparametrización frente al pseudo-cepstrum, especialmente en los efectos de canal mas adversos. En los IC de las dos curvas se puede apreciar también una mayor separación para los canales con mayor PLR [1], expresado en porcentaje.

[1] Packet Loss Rate

De los anteriores resultados podemos concluir que aunque la aproximación Pseudo-Cepstrum es computacionalmente más eficiente, la aproximación clásica de transparametrización es más robusta, pues la ganancia relativa obtenida es de un 5,48 % para la PLR más baja (9,09 %), y un valor similar se obtiene para PLR altas (5,04 % para una PLR de 38,38 %). Por tanto, en condiciones malas del canal el procedimiento clásico de transparametrización es el más recomendado, sin embargo si se tienen limitaciones en cuanto a los recursos computacionales (p.e. terminales móviles) la aproximación del Pseudo-Cepstrum es una buena alternativa.

8.2.2. Ruido de Ambiente

A continuación se expondrán los resultados obtenidos cuando se añade el ruido de ambiente a la voz original (de acuerdo a lo expuesto en la Sección 7.2.1).

Para empezar, se ilustra la ganancia obtenida en la Transparametrización cuando se utiliza el Procedimiento de Estima Mejorado de la energía, continuando con una comparativa de la Transparametrización con las técnicas del Suavizado y el Decodificado. Finalmente, se evalúa la influencia del proceso del postfiltrado que se realiza en el decodificador de fuente y la utilidad del filtrado paso-bajo del Espectro de Modulación frente al ruido de ambiente.

Procedimiento de Estima Mejorado en Presencia de Ruido de Ambiente

Como se estableció en la Sección 8.2.1, la energía juega un papel muy importante en una tarea de reconocimiento, pero especialmente en la Transparametrización, en donde el procedimiento de estima puede ayudar a mejorar en la robustez de la técnica. La Figura 8.6 ilustra dicha contribución, en este caso, en condiciones de ruido de ambiente y comparando el Transparametrizado utilizando el procedimiento original de estima de energía (bLP-MFCC), con el procedimiento de estima mejorado (bLP-MFCC+) (véase el detalle de esta notación en la Tabla A.1).

Los resultados nos muestran que el aporte del Procedimiento de Estima Mejorado en presencia de ruido es relevante, pues la reducción relativa alcanzada en la WER es de un 9 % para ruido de voces (Babble), 8,89 % para ruido de fábricas (Factory) y un 13,58 % para ruido rosa (Pink). Y aunque para ruido blanco no haya una mejora significativa, e incluso para ruido de coches (volvo) haya un retroceso del 1,9 %, es destacable la reducción alcanzada en los primeros tres tipos de ruido, pues ésta se obtiene sobre un procedimiento de estima ya de por sí robusto, como se ha puesto de manifiesto en [113][114][45].

De otro lado, en la Figura 8.7 se pueden observar los resultados de comparar la Transparametrización utilizando el Procedimiento de Estima Mejorado, con las técnicas del Suavizado y Decodificado que utilizan energía decodificada. De estos resultados se puede inferir, por un lado, que el reconocimiento mediante transparametrización obtiene una considerable reducción relativa de la WER frente a la técnica Decodificada (11-14 % aprox.), con la única excepción del ruido blanco para el que se obtienen resultados de WER similares.

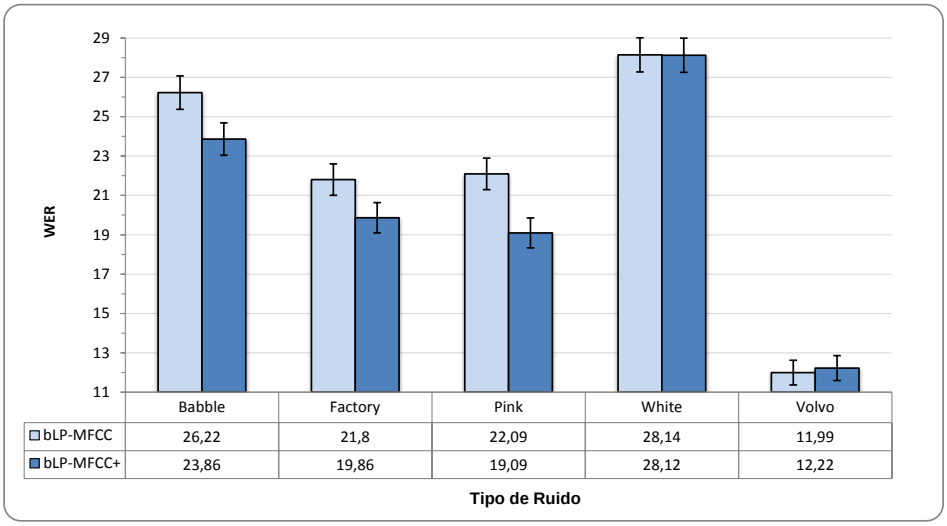


Figura 8.6: Disminución de la WER obtenida por el Procedimiento de Estima Mejorado de la energía, en este caso en presencia de ruido de ambiente.

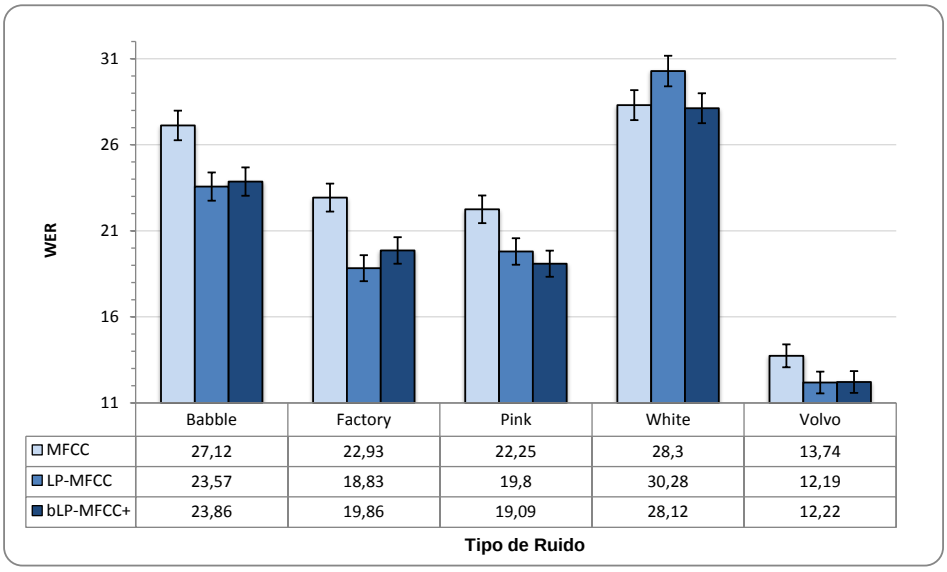


Figura 8.7: Comparativa de técnicas de reconocimiento en presencia de ruido de ambiente.

De otro lado, frente a la técnica del Suavizado, la Transparametrización se muestra más robusta para los ruidos blanco y rosa, en donde logra una reducción relativa de la WER de 7,13 % y 3,58 %, respectivamente. Sin embargo, para el ruido de coches la WER obtenida en ambas técnicas es similar, y para los ruidos de fábrica y de voces, la técnica del Suavizado consigue una WER inferior a la del Transparametrizado en un 5,18 y 1,21 % respectivamente.

De los resultados anteriores y de los observados en la Sección 8.2.1 podemos concluir que el reconocimiento mediante transparametrización consigue una mayor reducción en la WER respecto de la técnica Decodificada tanto en condiciones de pérdida de paquetes como en presencia de ruido de ambiente.

De otro lado, frente a la técnica del Suavizado, la Transparametrización resulta claramente más robusta cuando existe pérdida de paquetes, no siendo así ante la existencia de algunos tipos de ruido de ambiente. Obviamente, cuando combinamos ambos efectos (ruido y pérdida de paquetes) la técnica de RMT ofrece resultados netamente superiores. Lo anterior, nos condujo a un análisis más detallado de las diferencias existentes en los procedimientos de extracción de características de las dos técnicas de reconocimiento.

Efecto del postfiltro en RAH

Como se detalló en el Capítulo 4, la técnica de Suavizado obtiene la envolvente espectral a partir de los LPC, de forma similar a como lo hace la Transparametrización, que obtiene la envolvente a partir de los LSP o LPC. Sin embargo, la técnica del Suavizado calcula los LPC a partir de la voz decodificada, mientras que la Transparametrización lo hace a partir del bitstream, es decir sin involucrar el proceso de decodificación.

En el decodificador, una vez se han extraído los parámetros contenidos en el bitstream, se procede a la reconstrucción de la señal de voz. Sin embargo a continuación de esta etapa existe un post-procesado que introduce cambios en la envolvente espectral de la señal reconstruida. Este post-procesado involucra una serie de filtros que buscan mejorar la calidad perceptual de la voz. A este proceso realizado por el conjunto de filtros se le denomina *postfiltrado* [4].

Por lo tanto, dado que la envolvente espectral obtenida por la Transparametrización no se ve modificada por el efecto del postfiltrado, quisimos comprobar el efecto que éste produce sobre la envolvente espectral utilizada para reconocimiento. Para ello, realizamos dos experimentos utilizando la técnica del Suavizado que implica la decodificación y por tanto el postfiltrado. En la Figura 8.8 se exponen los resultados de la comparación del Suavizado (LP-MFCC) con y sin la etapa del postfiltro (LP-MFCC*).

De estos resultados podemos observar que cuando se elimina el postfiltro en la decodificación, la WER aumenta, y por tanto, podemos ver al proceso de postfiltrado como un procedimiento que ayuda a disminuir la WER en el Suavizado. En particular, la reducción relativa de la WER sería de un 3,79 % para ruido de voces, 5,71 % para ruido de

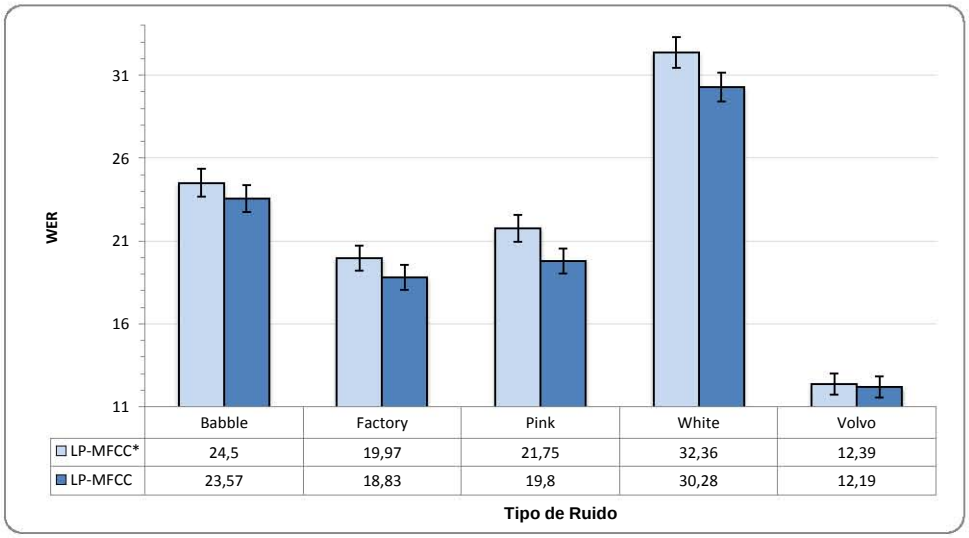


Figura 8.8: Reducción de la WER debida al efecto del postfiltro en la Técnica del Suavizado.

fábrica, 8,96 % para ruido rosa, 6,42 % para ruido blanco y 1,61 % para ruido de coches.

Por otro lado, si consideramos la WER obtenida con la técnica de Suavizado sin postfiltrado y la comparamos con la obtenida por la Transparametrización, podemos observar que la Transparametrización presenta un mejor desempeño para todos los tipos de ruido (véase Figura 8.9).

Estas observaciones nos llevan a concluir que es razonable incorporar los efectos del postfiltrado en el cálculo de la envolvente espectral y en la estima de energía que se realiza en la Transparametrización. Sin embargo, el análisis de las etapas del postfiltro que más afectan el proceso de reconocimiento, y su incorporación a la Transparametrización se plantean como trabajo futuro.

Filtrado Paso-Bajo del Espectro de Modulación

Utilizando el procedimiento descrito en la Sección 6.5, a continuación se presenta una comparativa de los resultados obtenidos por el RMT utilizando el Procedimiento de Estima Mejorado (bLP-MFCC+) con uno, que además de lo anterior, utiliza el Filtrado del Espectro de Modulación (FbLP-MFCC+) (véase Sección 7.5.3). Estos resultados se ilustran en la Figura 8.10.

Los resultados nos muestran que el filtrado paso-bajo del espectro de modulación consigue una reducción relativa de la WER en niveles que van desde un 3,26 % para el ruido de voces, hasta un 8,17 % para el ruido rosa. Para el caso de ruido de coches, no se obtuvo una reducción significativa. Sin embargo, es de destacar que a pesar de la naturaleza distinta de cada tipo de ruido, esta solución resulta muy eficaz en todos los casos. En la Figura 8.11 se ilustra la comparativa del RMT utilizando tanto el procedimiento de estima mejorado,

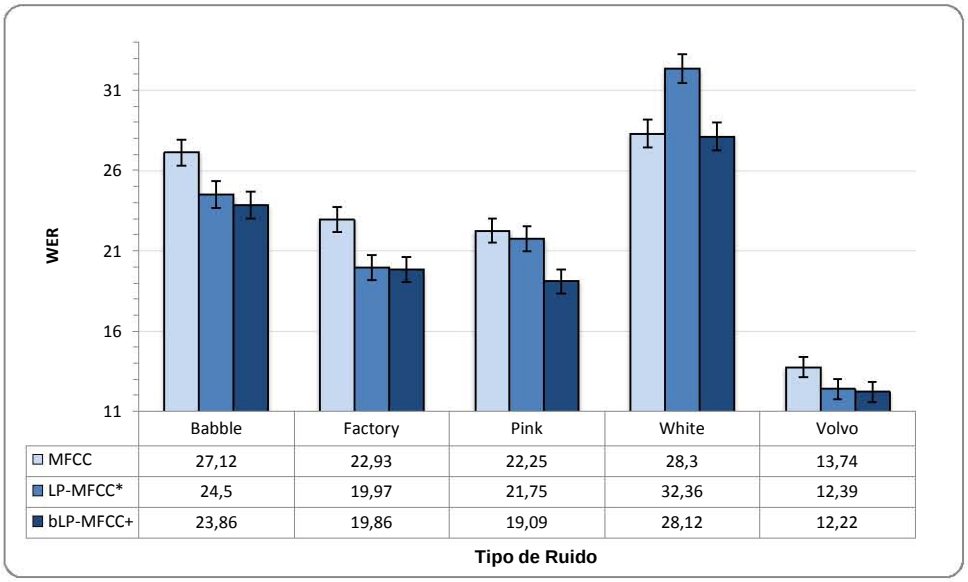


Figura 8.9: Comparación de la Técnica del Suavizado sin postfiltro con el RMT y la Técnica Decodificada.

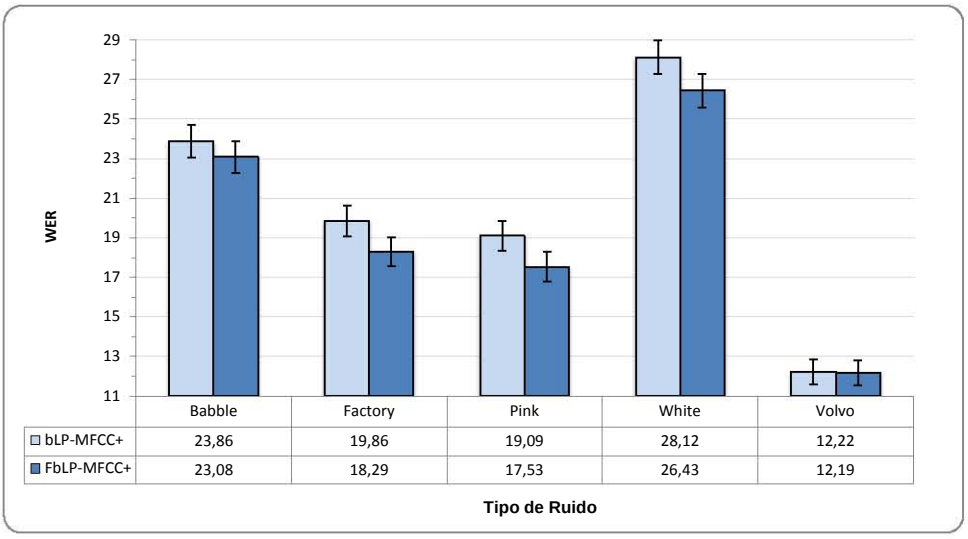
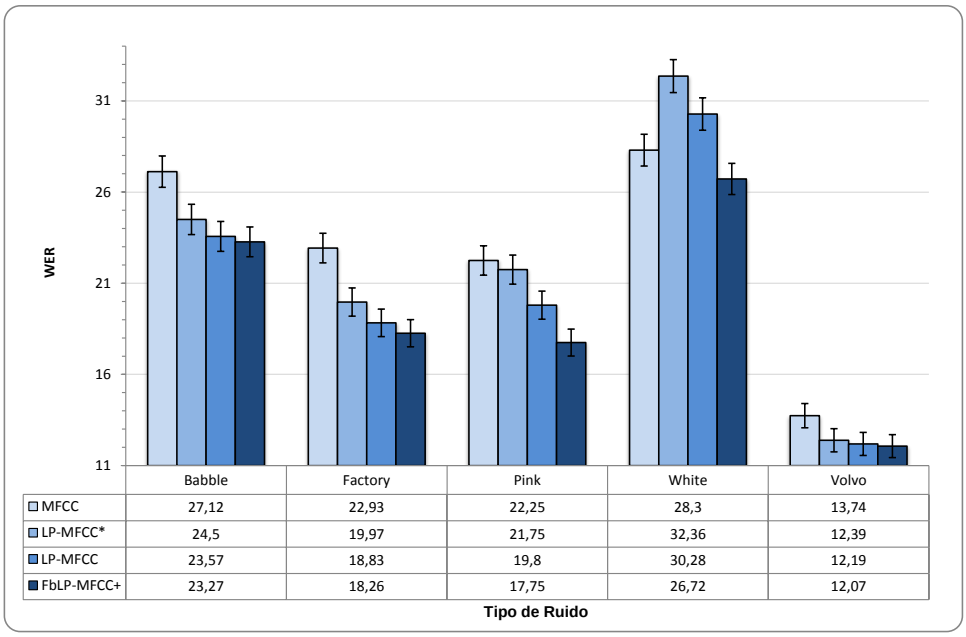


Figura 8.10: Efecto del Filtrado paso-bajo al Espectro de Modulación construido con la evolución temporal de los MSP.

como el filtrado del espectro de modulación; con las técnicas de referencia, incluido el Suavizado sin el postfiltro.



	Babble	Factory	Pink	White	Volvo
MFCC	27,12	22,93	22,25	28,3	13,74
LP-MFCC*	24,5	19,97	21,75	32,36	12,39
LP-MFCC	23,57	18,83	19,8	30,28	12,19
FbLP-MFCC+	23,27	18,26	17,75	26,72	12,07

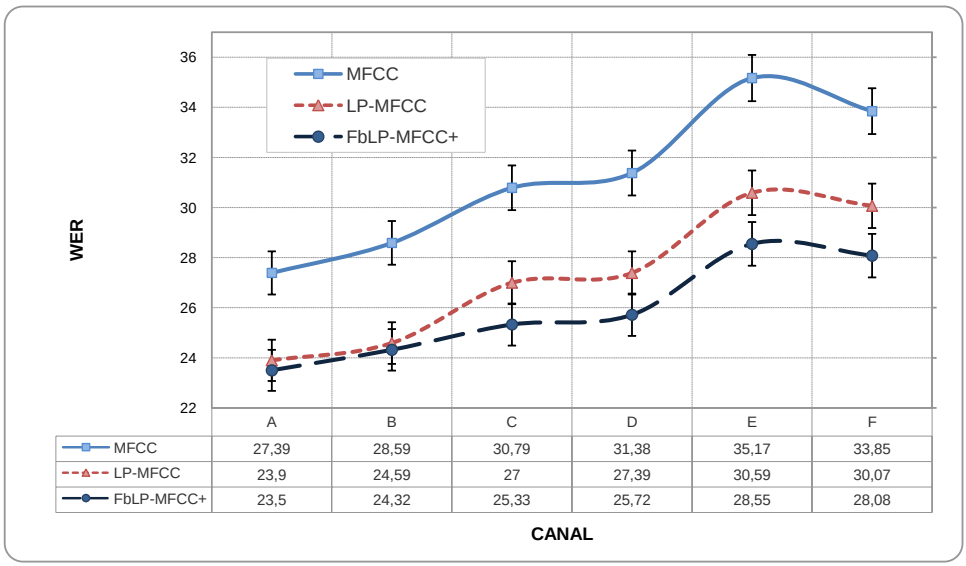
Figura 8.11: Comparación del RMT utilizando el Filtrado paso-bajo al Espectro de Modulación, con las demás técnicas de referencia.

8.2.3. Efecto Combinado del Ruido de Ambiente y Pérdida de Paquetes

Finalmente, las Figuras 8.12 - 8.16 ilustran los resultados obtenidos cuando se combinan los efectos del ruido de ambiente y la pérdida de paquetes. En éstas se puede apreciar que el RMT alcanza una mayor robustez frente a las dos técnicas de referencia, principalmente en los canales con mayores pérdidas, utilizando para ello tanto el procedimiento de estima mejorado como el filtrado paso-bajo del espectro de modulación.

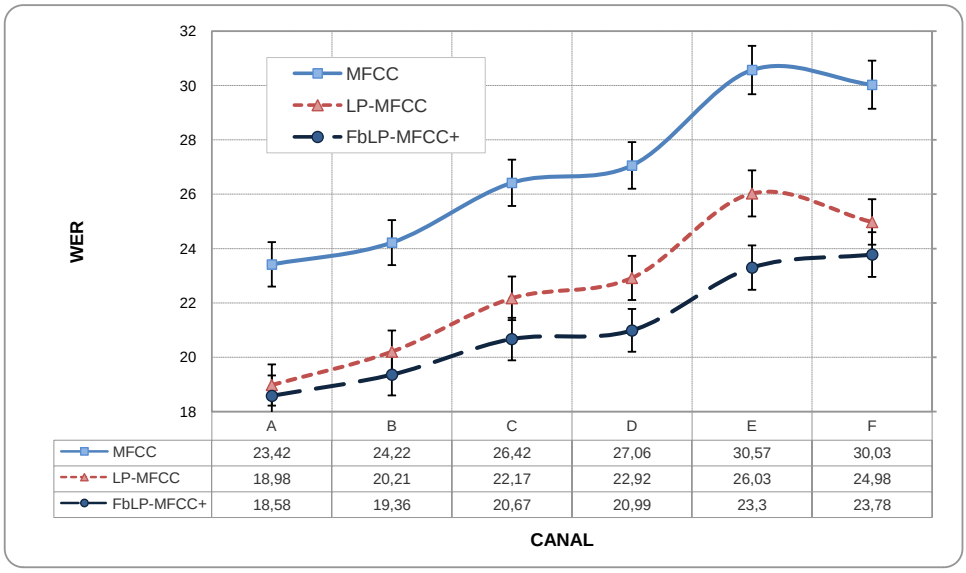
De las anteriores gráficas, se puede observar que el RMT es especialmente robusto frente a los ruidos blanco (“White”) y rosa (“Pink”), en donde la diferencia en la WER con respecto al Suavizado se mantiene con pocas variaciones para todos los canales simulados. De otro lado, en términos generales la técnica del Suavizado obtiene un mejor desempeño que la del Decodificado, siendo solamente inferior en el caso de ruido blanco.

Con relación a los canales, la mayor diferencia en la WER entre el RMT y el Suavizado se obtiene en el canal E (el que tiene la PLR más alta). No obstante, para el canal F (el que tiene la MBL más alta), si bien el RMT obtiene también una WER más baja que las técnicas de referencia, ésta es inferior a la obtenida con el canal E. Lo anterior puede ser ocasionado por el filtro paso-bajo que introduce un efecto de suavizado temporal de las trayectorias temporales de las salidas del banco de filtros, y que ayuda al proceso de reconstrucción de



	A	B	C	D	E	F
MFCC	27,39	28,59	30,79	31,38	35,17	33,85
LP-MFCC	23,9	24,59	27	27,39	30,59	30,07
FbLP-MFCC+	23,5	24,32	25,33	25,72	28,55	28,08

Figura 8.12: Comparación del RMT con las Técnicas del Suavizado y Decodificado, en presencia de ruido “babble” y pérdida de paquetes.



	A	B	C	D	E	F
MFCC	23,42	24,22	26,42	27,06	30,57	30,03
LP-MFCC	18,98	20,21	22,17	22,92	26,03	24,98
FbLP-MFCC+	18,58	19,36	20,67	20,99	23,3	23,78

Figura 8.13: Comparación del RMT con las Técnicas del Suavizado y Decodificado, en presencia de ruido “Factory” y pérdida de paquetes.

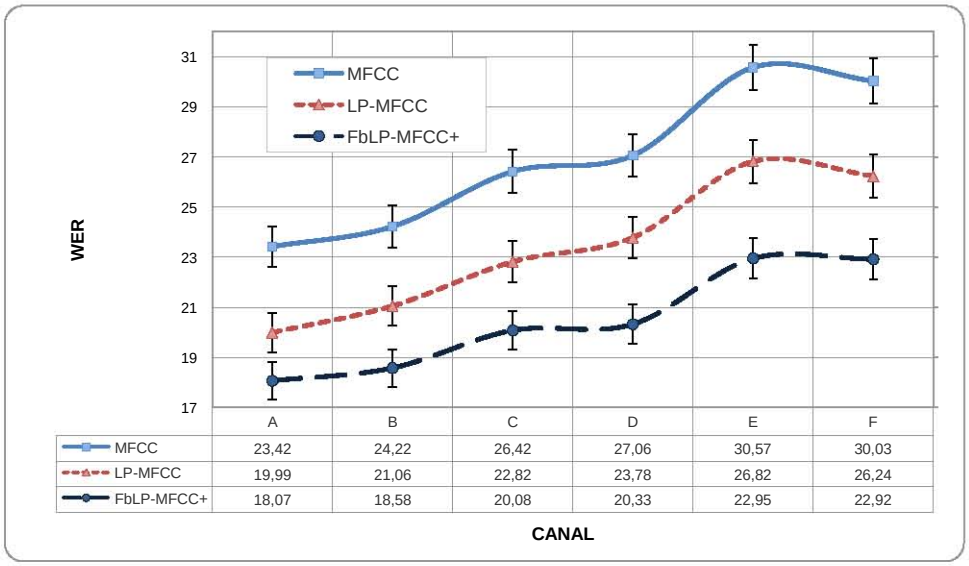


Figura 8.14: Comparación del RMT con las Técnicas del Suavizado y Decodificado, en presencia de ruido “Pink” y pérdida de paquetes.

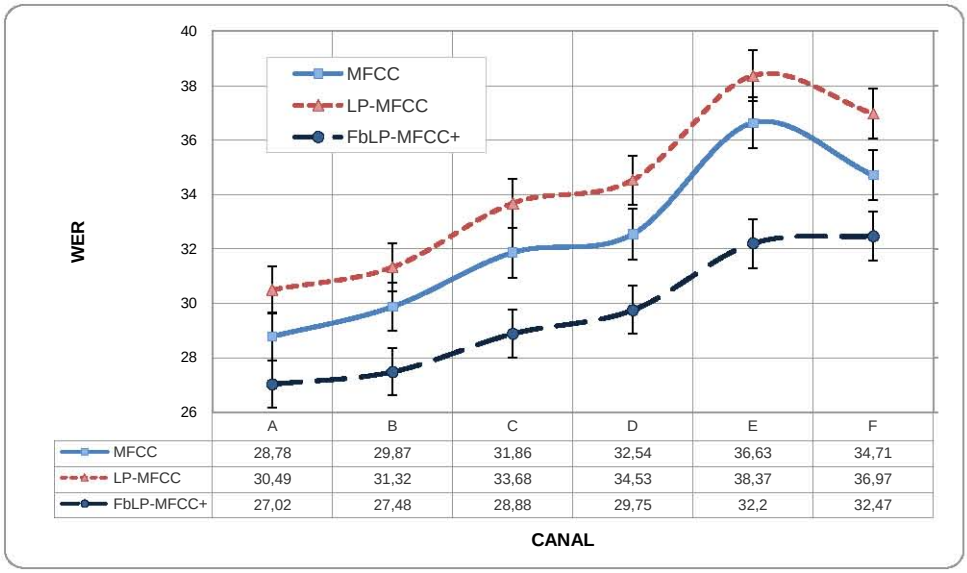


Figura 8.15: Comparación del RMT con las Técnicas del Suavizado y Decodificado, en presencia de ruido “White” y pérdida de paquetes.

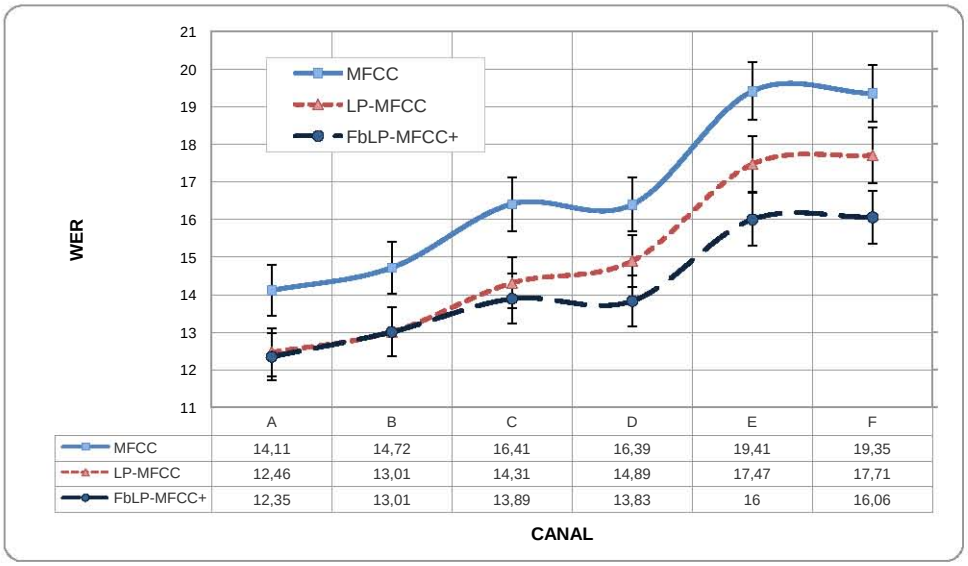


Figura 8.16: Comparación del RMT con las Técnicas del Suavizado y Decodificado, en presencia de ruido "Volvo" y pérdida de paquetes.

la envolvente espectral cuando se produce la pérdida de paquetes (aunque la tasa de pérdida sea alta como en el caso del canal E). Sin embargo, cuando se produce pérdida de paquetes en ráfagas largas (como en el canal F), la ayuda que podría prestar el filtrado paso-bajo es inferior a la que obtendría con canales de bajas longitudes de ráfaga (canales C, D y E), pues el proceso de reconstrucción de la envolvente espectral es más difícil cuanto se produce una ráfaga larga, que en una corta.

De los resultados expuestos bajo el entorno de voz sobre IP, se ha verificado la eficacia del RMT frente a la distorsión por codificación y la pérdida de paquetes, resaltando la contribución de la energía en el proceso de reconocimiento. Debido a lo anterior, el procedimiento de estima mejorado se plantea como una solución que contribuye a disminuir el efecto del ruido de ambiente en el desempeño del RMT. Por otra parte, el procedimiento de filtrado del espectro de modulación se presenta como una buena solución para el efecto combinado del ruido de ambiente y la pérdida de paquetes.

A continuación se evaluarán las contribuciones planteadas en el capítulo anterior para el entorno de UMTS.

8.3. Reconocimiento de Habla sobre redes UMTS

El segundo entorno que ha sido modelado en esta tesis es el de UMTS. En la Figura 8.17 se puede observar un esquema del proceso de reconocimiento de voz para el caso de UMTS. En primer lugar, se describirán los factores más relevantes a la hora de exponer y analizar los resultados (no obstante, para una descripción más detallada del modelado del sistema, consultar la sección dedicada a este modelo en el Capitulo 7).

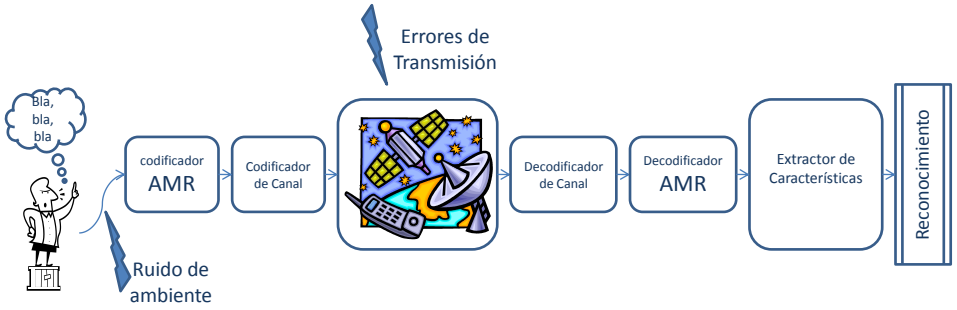


Figura 8.17: Etapas de un sistema de RAH sobre UMTS.

Bajo este esquema, la distorsión asociada al proceso de transmisión se presenta en forma de errores de transmisión a nivel de bit. En particular, el problema que más afecta el desempeño del reconocedor es la pérdida de bits en ráfagas. De otro lado, el ruido de ambiente también está presente y por tanto también será motivo de análisis en este entorno.

Dentro de las etapas presentes en la Figura 7.4, la codificación de canal merece una mención especial, y dentro de esta, la denominada Protección Desigual frente a Errores (Unequal Error Protection - UEP) [102].

Tal como se expuso en la Sección 7.4.2, el esquema de protección desigual discrimina por clases los bits a proteger, y por tanto brinda mayor protección a unos bits en detrimento de otros. La Tabla 8.1 resume los bits asignados a cada parámetro enviado por el codificador de fuente (AMR-NB a 12,2 Kbps) y a la clase a la cual pertenece [4]. Siendo los bits clase A los más protegidos y los bits clase C los menos protegidos.

Parámetro	Clase A	Clase B	Clase C	Total (20 ms)
LSP	29	9	0	38
Periodo Fundamental (T)	28	0	2	30
Ganancia Adaptativa (Ga)	12	3	1	16
Ganancia Estocástica (Ge}	12	8	0	20
Librería Estocástica	0	83	57	140

Tabla 8.1: Asignación de bits por parámetro según el esquema UEP.

Como se puede observar, los parámetros más protegidos son los LSP, seguidos del pitch (T) y las ganancias (Ga y Gf). Lo anterior nos hace pensar que el reconocimiento mediante transparametrización será el más beneficiado, pues justamente los parámetros más protegidos son los que se usan para el cálculo de la envolvente espectral. Además, los parámetros menos protegidos se verán muy deteriorados cuando existan errores en el canal, y por tanto trasladarán esta distorsión a la voz reconstruida en la decodificación que utilizan las otras dos técnicas de referencia.

No obstante lo anterior, y a pesar de que los LSP utilizan el 35,8 % de los bits clase

A (29 de 81 bits), el porcentaje restante no es despreciable y por tanto, surge la idea de incorporar otros parámetros fuertemente protegidos para utilizarlos en la conformación del vector de características.

De esta manera planteamos una nueva aproximación referida como XbLP-MFCC o Transparametrización Extendida, que añade al vector de características tradicional (conformado por los coeficientes cepstrales y la energía), una combinación de parámetros extra. A continuación se expondrán los detalles de esta aproximación, así como los resultados obtenidos tanto en presencia de errores de transmisión, como en presencia de ruido de ambiente.

Por otro lado, es de destacar el papel que juega la energía en el entorno IP, y como se verá en la Sección 8.3.3, juega un papel también muy relevante frente al efecto combinado de las distorsiones de codificación, errores de transmisión y ruido de ambiente, en este caso bajo el entorno de UMTS.

8.3.1. Errores de Transmisión

El modelado de la capa física de UMTS, explicado en el capítulo anterior, se utiliza ahora para evaluar el desempeño de la Transparametrización en presencia de errores de transmisión.

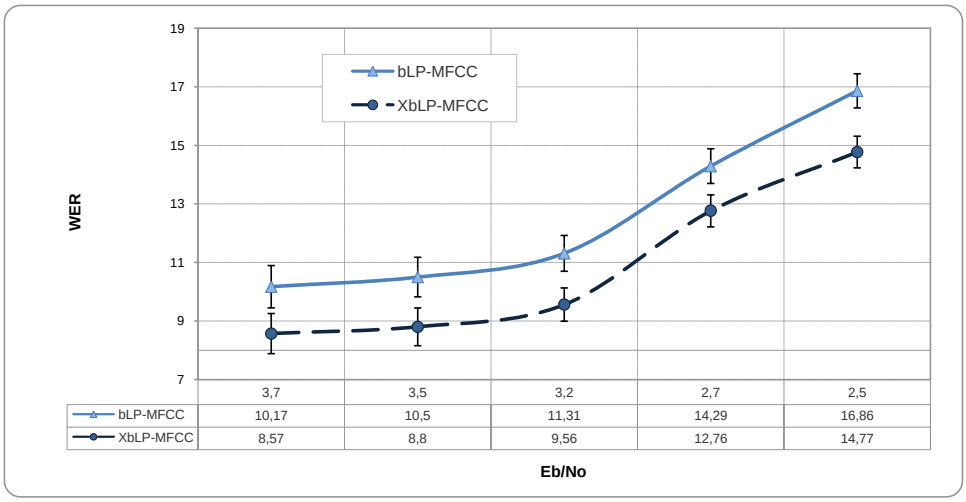
De acuerdo con lo expuesto en la sección anterior, la Transparametrización Extendida se obtiene añadiendo una combinación de parámetros altamente protegidos por la codificación de canal, siguiendo un esquema de protección desigual. Por tanto, hemos añadido dos de los parámetros de la excitación que resultan más protegidos: el pitch (T) y la ganancia de la librería adaptativa (Gp), que junto con los parámetros MFCC (12) y la energía, conforman el vector de características utilizado para la tarea de reconocimiento.

En la técnica del suavizado también se ha añadido el pitch (T) como complemento a los coeficientes cepstrales y la energía decodificada, pues, a pesar de que el Suavizado no accede a los parámetros del bitstream para conformar el vector de características, el pitch puede ser calculado con procedimientos estándar para ser utilizado en lo que hemos denominado Suavizado Extendido y que es referido con el acrónimo XLP-MFCC. Ver Tabla A.1.

El método decodificado no utilizará parámetros extra al conjunto de parámetros que describen el Cepstrum (12 mfcc y la energía decodificada - Edec). Esta comparativa se ha hecho así, teniendo en cuenta que el Decodificado (MFCC) es el procedimiento tradicional para el reconocimiento de voz codificada y en su forma original y más extendida, no hace uso de parametrizaciones añadidas.

La Figura 8.18 expone los resultados de la reducción relativa en la WER obtenida por la Transparametrización Extendida (XbLP-MFCC) frente al procedimiento clásico de transparametrización (bLP-MFCC). Como se puede inferir de los anteriores resultados, la

aproximación XbLP-MFCC consigue una reducción relativa entre un 12 % y un 16 % en la WER, respecto de la aproximación bLP-MFCC, tal como se esperaba dada la alta protección que reciben tanto los LSP (utilizados para el cálculo de los MFCC), como el pitch y la ganancia adaptativa, en la codificación de canal en UMTS a través de la protección desigual (UEP) descrita en la Sección 7.4.2. Y teniendo en cuenta que los dos parámetros añadidos pueden ser directamente extraidos del bitstream, es decir, no se requiere un procedimiento adicional para su cálculo, el coste computacional añadido es relativamente bajo si se tienen en cuenta los logros alcanzados por esta aproximación.



	3,7	3,5	3,2	2,7	2,5
bLP-MFCC	10,17	10,5	11,31	14,29	16,86
XbLP-MFCC	8,57	8,8	9,56	12,76	14,77

Figura 8.18: Transparametrización Extendida bajo un entorno de errores de transmisión en UMTS.

A continuación, en la Figura 8.19 se ilustran los resultados de reconocimiento mediante Transparametrización Extendida (XbLP-MFCC), frente a los obtenidos por el Suavizado Extendido (XLP-MFCC) y el Decodificado (MFCC).

A la vista de estos resultados, podemos concluir no solo una amplia diferencia obtenida por la Transparametrización Extendida respecto del Suavizado y el Decodificado, sino también una gran robustez frente al deterioro de las condiciones del canal. Pues la reducción relativa en la WER obtenida para los canales con menos pérdidas está entre un 32 % y un 35 % y para los canales con mayores pérdidas la reducción relativa ronda un 43 % y 47 %.

8.3.2. Ruido de Ambiente

Similar al entorno de voz sobre IP, el procedimiento de estima mejorado también se puede utilizar al entorno de UMTS, en este caso al codificador AMR-NB a 12,2 Kbps. En la Figura 8.20 se puede observar el efecto de dicho procedimiento sobre el desempeño del RMT en condiciones de ruido.

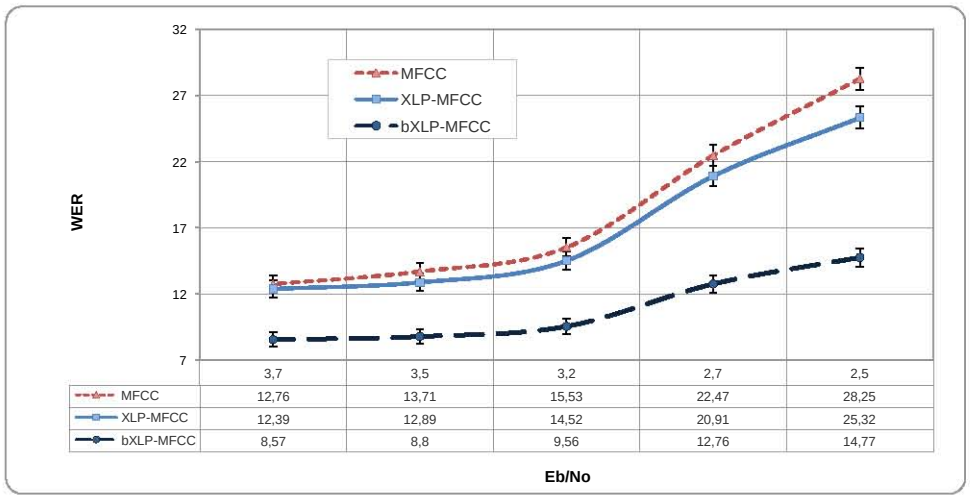


Figura 8.19: Efecto de los errores de transmisión de UMTS, en tres técnicas de reconocimiento.

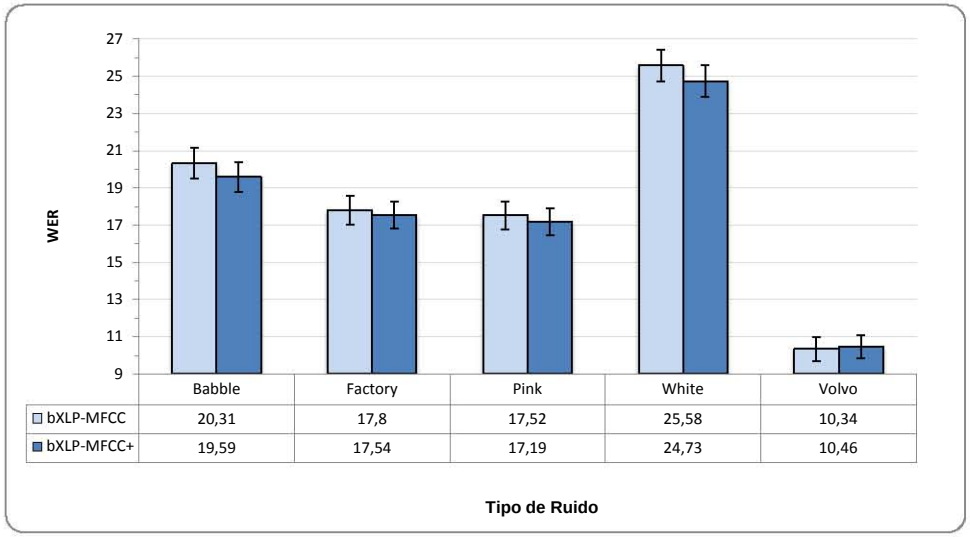


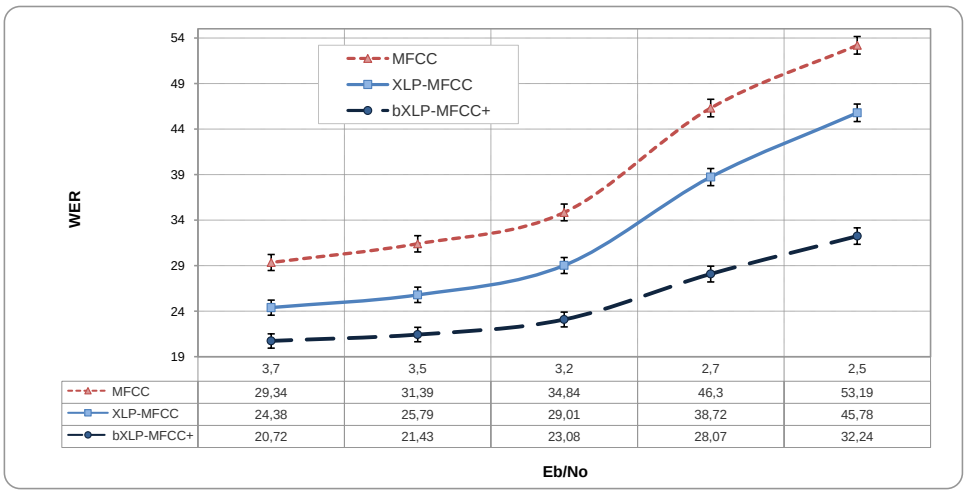
Figura 8.20: Efecto del Procedimiento de Estima Mejorado en presencia de ruido de ambiente en la técnica de RMT Extendida (XbLP-MFCC).

Si bien el procedimiento de estima mejorado aplicado al codificador AMR-NB consigue una menor reducción en la WER respecto de la obtenida con el codificador G.729; es de destacar que en este caso la comparación se ha realizado utilizando la versión extendida del RMT (bXLP-MFCC) y, por lo tanto, ya incluye una reducción importante en la WER, respecto del procedimiento clásico de RMT. No obstante, se puede observar que la reducción se consigue para los 4 primeros tipos de ruido ("Babble", "Factory", "Pink" y "White"), similar al comportamiento observado en el codificador G.729.

En otras pruebas realizadas, se ha comprobado también la disminución en la WER debida al Procedimiento de Estima Mejorado en el RMT, especialmente bajo el efecto combinado de errores de transmisión y ruido de ambiente, más aún en presencia de ruido blanco.

8.3.3. Efecto Combinado del Ruido de Ambiente y los Errores de Transmisión

Finalmente en las Figuras 8.21-8.25 se ilustran los resultados obtenidos por el RMT frente a las dos técnicas de referencia, considerando el efecto combinado del ruido y los errores de transmisión. Como se puede observar, las tendencias son similares a las obtenidas en el entrono IP.

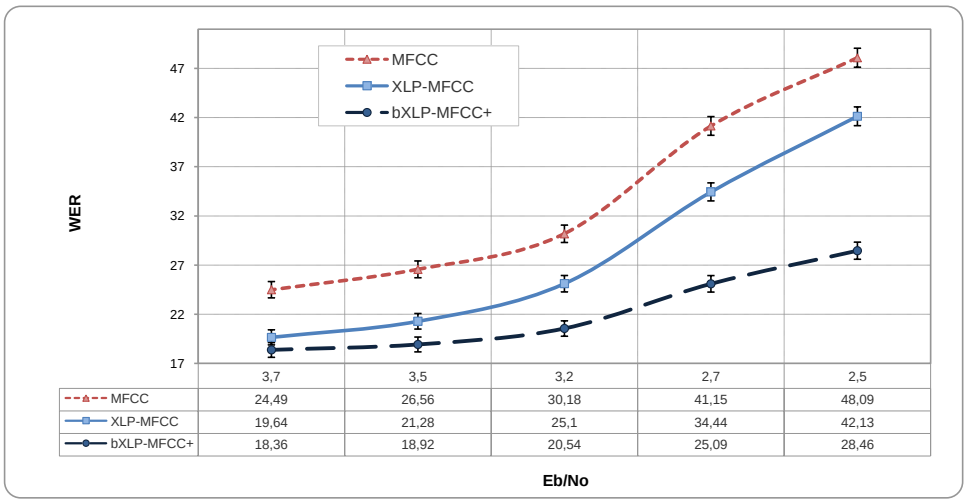


	3,7	3,5	3,2	2,7	2,5
MFCC	29,34	31,39	34,84	46,3	53,19
XLP-MFCC	24,38	25,79	29,01	38,72	45,78
bXLP-MFCC+	20,72	21,43	23,08	28,07	32,24

Figura 8.21: Comparativa del RMT Extendido utilizando el Procedimiento de Estima Mejorado (XbLP-MFCC+), con el Suavizado Extendido (XLP-MFCC) y el Decodificado (MFCC), bajo el efecto combinado de los errores de transmisión y ruido "Babble".

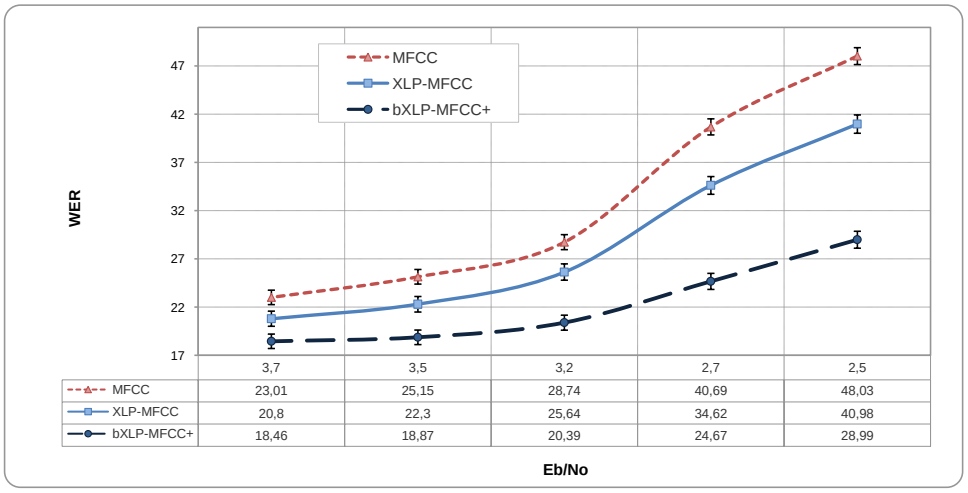
En las anteriores figuras se compara la Transparametrización Extendida con el Procedimiento de Estima Mejorada bajo el acrónimo XbLP-MFCC+ (ver Tabla A.1), con la Técnica del Suavizado Extendido (XLP-MFCC) y el Decodificado (MFCC); éstas dos últimas utilizando la energía decodificada.

En este efecto combinado, si bien la diferencia entre la Transparametrización Extendida y el Suavizado Extendido se estrecha con respecto a los resultados obtenidos bajo el efecto



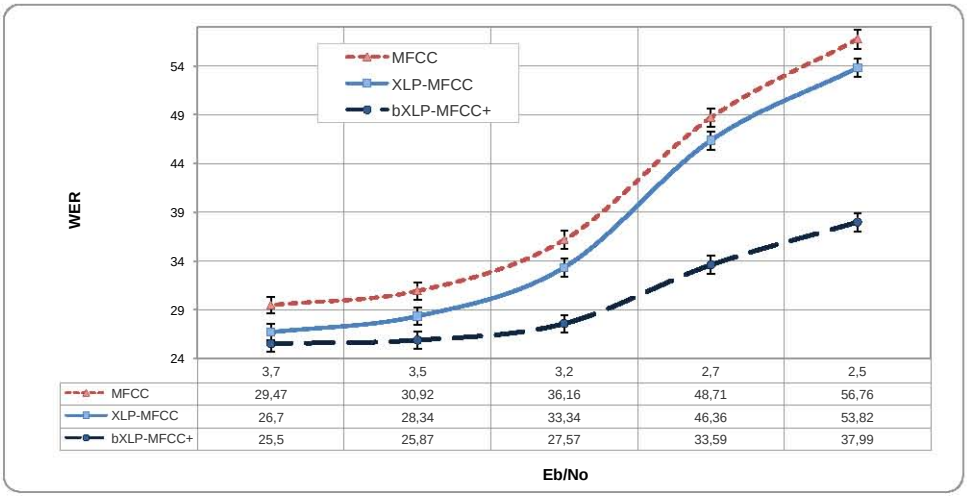
	3,7	3,5	3,2	2,7	2,5
MFCC	24,49	26,56	30,18	41,15	48,09
XLP-MFCC	19,64	21,28	25,1	34,44	42,13
bXLP-MFCC+	18,36	18,92	20,54	25,09	28,46

Figura 8.22: Comparativa del RMT Extendido utilizando el Procedimiento de Estima Mejorado (XbLP-MFCC+), con el Suavizado Extendido (XLP-MFCC) y el Decodificado (MFCC), bajo el efecto combinado de los errores de transmisión y ruido “Factory”.



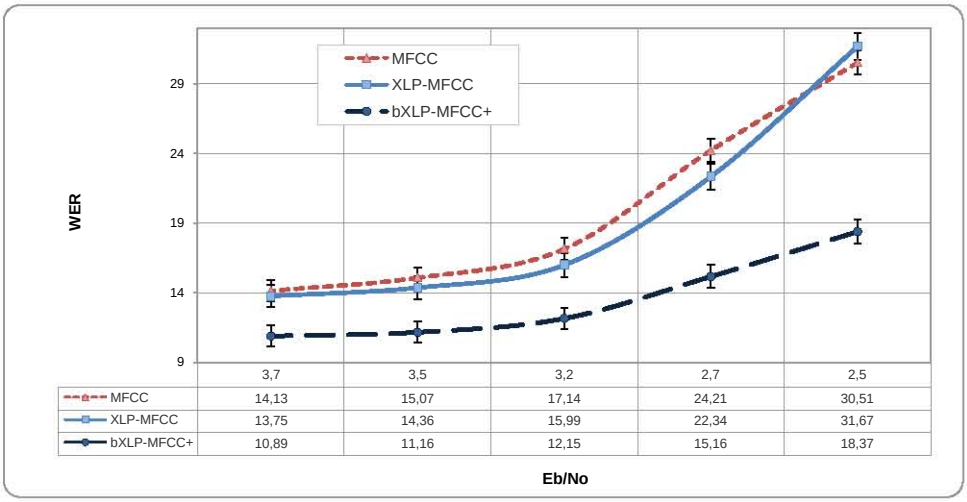
	3,7	3,5	3,2	2,7	2,5
MFCC	23,01	25,15	28,74	40,69	48,03
XLP-MFCC	20,8	22,3	25,64	34,62	40,98
bXLP-MFCC+	18,46	18,87	20,39	24,67	28,99

Figura 8.23: Comparativa del RMT Extendido utilizando el Procedimiento de Estima Mejorado (XbLP-MFCC+), con el Suavizado Extendido (XLP-MFCC) y el Decodificado (MFCC), bajo el efecto combinado de los errores de transmisión y ruido “Pink”.



	3,7	3,5	3,2	2,7	2,5
MFCC	29,47	30,92	36,16	48,71	56,76
XLP-MFCC	26,7	28,34	33,34	46,36	53,82
bXLP-MFCC+	25,5	25,87	27,57	33,59	37,99

Figura 8.24: Comparativa del RMT Extendido utilizando el Procedimiento de Estima Mejorado (XbLP-MFCC+), con el Suavizado Extendido (XLP-MFCC) y el Decodificado (MFCC), bajo el efecto combinado de los errores de transmisión y ruido “White”.



	3,7	3,5	3,2	2,7	2,5
MFCC	14,13	15,07	17,14	24,21	30,51
XLP-MFCC	13,75	14,36	15,99	22,34	31,67
bXLP-MFCC+	10,89	11,16	12,15	15,16	18,37

Figura 8.25: Comparativa del RMT Extendido utilizando el Procedimiento de Estima Mejorado (XbLP-MFCC+), con el Suavizado Extendido (XLP-MFCC) y el Decodificado (MFCC), bajo el efecto combinado de los errores de transmisión y ruido “Volvo”.

de los errores de transmisión (solamente) de la Figura 8.19; se mantiene un notable margen entre las WER logradas por las dos técnicas, especialmente cuando se presenta un gran deterioro en las condiciones del canal. Concretamente, para la Figura 8.24 que ilustra los resultados del ruido blanco, la reducción relativa en la WER pasa de un 4,49 % para el canal con menos pérdidas (Eb/No = 3,7 dB) a una de 29,41 % en el canal con mayores pérdidas (Eb/No = 2,5 dB). Con respecto a la técnica Decodificada, la reducción relativa en la WER es de 13,47 % para el canal con menos pérdidas y de 33,06 % para el canal con más pérdidas. Similares tendencias se presentan para los demás tipos de ruido, siendo más notable la reducción de la WER conseguida por el RMT en los rudios de coches ("Volvo") y de voces ("Babble").

Lo anterior verifica la conveniencia de utilizar la Transparametrización Extendida con el Procedimiento de Estima Mejorado en un entorno con distorsiones asociadas al efecto combinado del ruido de ambiente y los errores de transmisión, pues se consigue un apreciable incremento en la robustez frente a las técnicas de referencia, más aún cuando existe una alta degradación en las condiciones del canal.

Capítulo 9

Conclusiones y Trabajo Futuro

9.1. Conclusiones

En un sistema de reconocimiento de habla que opera sobre una red de comunicaciones existen tres problemas fundamentales: 1) las distorsiones derivadas del proceso de codificación-decodificación; 2) las debidas a los errores de transmisión; y 3) la ocasionadas por la presencia de ruido de ambiente. Como ya se había puesto de manifiesto en [113], y en esta tesis se ha corroborado, extendido y mejorado, la técnica conocida como transparametrización (consistente en derivar la parametrización empleada en el reconocedor a partir de la empleada por el codificador, siempre presente en las actuales redes de comunicaciones), resulta una excelente alternativa para abordar dichos problemas eficazmente.

Tanto las ventajas de la transparametrización sobre la aproximación convencional (consistente en decodificar la señal de voz y luego reconocer), como las mejoras aportadas en esta tesis respecto a la transparametrización de referencia [113], se han demostrado experimentalmente en dos entornos de especial interés y relevancia: VoIP y UMTS.

En particular, en el entorno IP se han extendido los resultados previos al codificador G.729, rediseñando toda la algorítmica necesaria para la estimación de la energía (parámetro fundamental desde el punto de vista del reconocedor, pero que no es directamente codificado y transmitido) y aportando mejoras significativas al proceso de estimación. También se ha corroborado experimentalmente la influencia que la etapa de post-filtrado habitualmente presente en los codificadores de voz tiene un efecto positivo relevante sobre el reconocimiento de habla posterior. Y, finalmente, se demostrado que la transparametrización puede emplearse de manera complementaria a muchas de las técnicas típicamente empleadas para enfrentarse al ruido aditivo: concretamente, se han hecho experimentos combinando esta técnica con el filtrado del espectro de modulación, resultando en ganancias acumulativas.

En el entorno UMTS, aparte de la extensión de las técnicas originales para el codificador AMR-NB, se ha propuesto una transparametrización extendida que se apoya en la técnica *Unequal Error Protection* empleada por el codificador de

canal. Los resultados experimentales en este entorno arrojan una conclusión clara: la transparametrización extendida mejora significativamente la robustez de la propuesta original y muy sustancialmente la de la aproximación convencional (tanto más cuanto peores son las condiciones de canal). Asimismo, como se hizo en el caso IP, se ha demostrado que la transparametrización resulta superior a la aproximación convencional en un escenario realista que incluye tanto errores de transmisión como ruido aditivo.

9.2. Contribuciones

Las contribuciones de esta tesis se enumeran seguidamente:

- Se ha extendido el trabajo original sobre reconocimiento de habla mediante transparametrización [113] para codificadores y entornos de comunicaciones más actuales: G.729 para VoIP y AMR-NB para UMTS [54, 55].

- Se ha adaptado el procedimiento de estimación de la energía a partir de los parámetros incluidos en el bitstream para los dos codificadores antes mencionados.

- Se ha extendido el análisis original, que sólo contemplaba errores de transmisión, considerando ruido aditivo en ambos escenarios, IP y UMTS.

- Se han implementado diferentes aproximaciones de transparametrización para comparar su desempeño en los escenarios propuestos [115].

- Se ha mejorado el procedimiento de estimación de la energía para conferirle una mayor robustez.

- Se ha propuesto una parametrización extendida para el entorno UMTS que incorpora nuevos parámetros disponibles en el bitstream a la parametrización típica. En particular, dicha mejora saca partido de la protección desigual que brinda el codificador de canal, sugiriendo el uso en reconocimiento de los parámetros de la excitación que van más protegidos.

- Se ha verificado experimentalmente la influencia positiva sobre el reconocimiento de habla de la etapa de postfiltrado habitualmente presente en los codificadores de voz.

- Se ha demostrado experimentalmente que la transparametrización puede combinarse con técnicas convencionales para combatir el ruido aditivo: en particular, se han hecho experimentos que combinan transparametrización y filtrado del espectro de modulación.

En resumen, se mejorado la técnica original de transparametrización y se ha evaluado en condiciones más realistas que combinan errores de transmisión y ruido en entornos IP y UMTS.

9.3. Trabajos Futuros

La investigación realizada nos ha permitido identificar las líneas que, a nuestro juicio, merecen estudio posterior:

- Una vez demostrada la influencia positiva del postfiltro sobre los resultados de reconocimiento de habla, se propone como línea futura de trabajo la incorporación de un mecanismo equivalente en el ámbito de transparametrización. Es de esperar que la ventaja que aporta esta técnica sea mayor en este ámbito ya que, como ocurre en otros casos, estaríamos reduciendo la influencia de los errores de transmisión en la voz decodificada.
- En esta tesis no se han planteado soluciones basadas en el preprocesado de la señal de voz con el fin de que el sistema de reconocimiento resulte completamente transparente respecto al sistema de comunicaciones. No obstante, consideramos que en el entorno VoIP sería perfectamente factible desarrollar un terminal software (embebido en una pagina web, por ejemplo) que incluyera diversos tipos de preprocesado orientados a reducir las distorsiones introducidas por el ruido ambiente y/o el canal.

Además de lo anterior, resulta obvio que todo tipo de estudio que, o bien acerque aún más el escenario simulado al escenario real (como pudiera ser la inclusión de del mecanismo de transmisión discontinua (DTX), o bien complete el estudio realizado (como la incorporación de los modernos codificadores de banda ancha), resultaría de indudable interés.

Bibliografía

[1] The Resource Management Corpus Part 1 (RM1). Tech. rep., National Institute of Standards and Technology (NIST) (distribuidor), 1992.

[2] Speech service option standard for wideband spread spectrum digital cellular system, QCELP coder. Tech. rep., TIA/EIA Interim Standard-96, April 1993.

[3] 3GPP TECHNICAL SPECIFICATION GROUP SERVICES AND SYSTEM ASPECTS. Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding Functions, (3GPP TS 26.090 Release 8), 2008.

[4] 3GPP TECHNICAL SPECIFICATION GROUP SERVICES AND SYSTEM ASPECTS. Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Frame Structure, (3GPP TS 26.101 Release 8), 2008.

[5] ALESANCO, A., GÁLLEGO, J., CANALES, M., VALDOVINOS, A., LAGUNA, P., AND GARCÍA, J. Análisis de los errores producidos durante la transmisión de ECGS sobre un canal UMTS.

[6] ALONSO, J. B., DÍAZ-DE MARÍA, F., TRAVIESO, C. M., AND FERRER, M. *Nonlinear Analyses and Algorithms for Speech Processing*, vol. 3817 of *Lecture Notes on Computer Science*. Springer-Verlag, 2005, ch. Optimal size of time window in nonlinear features for voice quality measurement, pp. 206–218.

[7] ALONSO, J. B., DÍAZ-DE MARÍA, F., TRAVIESO, C. M., AND FERRER, M. Optimal size of time window in nonlinear features for voice quality measurement. *Nonlinear Analyses and Algorithms for Speech Processing 3817* (2005), pp. 206–218. PT: S; CT: International Conference on Non-Linear Speech Processing; CY: APR 19-22, 2005; CL: Barcelona, SPAIN.

[8] ANDERSEN, S., KLEIJN, W., HAGEN, R., LINDEN, J., MURTHI, M., AND SKOGLUND, J. ilbc - a linear predictive coder with robustness to packet losses. In *Speech Coding, 2002, IEEE Workshop Proceedings.* (oct. 2002).

[9] ATAL, B. S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America 55*, 6 (1974), pp. 1304–1312.

[10] BENYASSINE, A., SHLOMOT, E., SU, H.-Y., MASSALOUX, D., LAMBLIN, C., AND PETIT, J.-P. ITU-T Recommendation G.729 Annex B: a silence compression

scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. *Communications Magazine, IEEE 35*, 9 (sep 1997), pp. 64–73.

[11] BERNARD, A., AND ALWAN, A. Source and channel coding for remote speech recognition over error-prone channels. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on* (2001), vol. 4, pp. 2613–2616 vol.4.

[12] BESSETTE, B., SALAMI, R., LEFEBVRE, R., JELINEK, M., ROTOLA-PUKKILA, J., VAINIO, J., MIKKOLA, H., AND JARVINEN, K. The adaptive multirate wideband speech codec (AMR-WB). *Speech and Audio Processing, IEEE Transactions on 10*, 8 (2002), pp. 620–636.

[13] BORELLA, M. Measurement and interpretation of internet packet loss. In *Journal of Communication and Networks* (June 2000), vol. 2, pp. 93–102.

[14] BOU-GHAZALE, S., AND HANSEN, J. A comparative study of traditional and newly proposed features for recognition of speech under stress. *Speech and Audio Processing, IEEE Transactions on 8*, 4 (jul 2000), pp. 429–442.

[15] CARMONA, J., PEINADO, A., PÉREZ-CÓRDOBA, J., AND GÓMEZ, A. MMSE-Based Packet Loss Concealment for CELP-Coded Speech Recognition. *Audio, Speech, and Language Processing, IEEE Transactions on 18*, 6 (aug. 2010), pp. 1341–1353.

[16] CARMONA, J., PEINADO, A., PÉREZ-CÓRDOBA, J., GÓMEZ, A., AND SÁNCHEZ, V. iLBC-Based Transparametrization: A Real Alternative to DSR for Speech Recognition Over Packet Networks. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on* (april 2007), vol. 4, pp. IV–961 –IV–964.

[17] CARMONA-MAQUEDA, J. *Reconocimiento de voz codificada sobre redes IP*. PhD thesis, Universidad de Granada, 2009.

[18] CÁDIZ, R. *Introducción a la Música Computacional (libro online)*. http://www.rodrigocadiz.com/imc, 2008. [Online].

[19] CHEN, J.-H., AND GERSHO, A. Adaptive postfiltering for quality enhancement of coded speech. *Speech and Audio Processing, IEEE Transactions on 3*, 1 (jan 1995), pp. 59–71.

[20] CHI, S.-M., AND OH, Y.-H. Lombard effect compensation and noise suppression for noisy Lombard speech recognition. In *Spoken Language, Fourth International Conference on* (oct 1996), vol. 4, pp. 2013–2016 vol.4.

[21] CHILDERS, D., SKINNER, D., AND KEMERAIT, R. The cepstrum: A guide to processing. vol. 65, pp. 1428–1443.

[22] CHOI, S. H., KIM, H. K., AND LEE, H. S. LSP weighting functions based on spectral sensitivity and mel-frequency warping for speech recognition in digital communication. In *Acoustics, Speech, and Signal Processing, 1999. ICASSP*

'99. Proceedings., 1999 IEEE International Conference on (mar 1999), vol. 1, pp. 401–404.

[23] CHOI, S. H., KIM, H. K., AND LEE, H. S. Speech recognition using quantized LSP parameters and their transformations in digital communication. *Speech Communication 30*, 4 (2000), 223–233.

[24] CHOI, S. H., KIM, H. K., LEE, H. S., AND GRAY, R. Speech recognition method using quantised LSP parameters in CELP-type coders. *Electronics Letters 34*, 2 (jan 1998), pp. 156–157.

[25] DAVIS, S., AND MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on 28*, 4 (aug 1980), pp. 357–366.

[26] DÍAZ-DE MARÍA, F., AND FIGUEIRAS-VIDAL, A. R. Nonlinear prediction for speech coding using radial basis functions. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on* (1995), vol. 1, pp. 788–791 vol.1.

[27] DÍAZ-DE MARÍA, F., AND FIGUEIRAS-VIDAL, A. R. Improving CELP coders by backward adaptive non-linear prediction. *International Journal of Adaptive Control and Signal Processing 11*, 7 (NOV 1997), pp. 585–601.

[28] DE-VICENTE-PEÑA, J. *Contribuciones al reconocimiento robusto de habla*. PhD thesis, Universidad Carlos III de Madrid, 2007.

[29] DIGALAKIS, V., NEUMEYER, L., AND PERAKAKIS, M. Quantization of cepstral parameters for speech recognition over the World Wide Web. *Selected Areas in Communications, IEEE Journal on 17*, 1 (jan 1999), pp. 82–90.

[30] DUFOUR, S., GLORION, C., AND LOCKWOOD, P. Evaluation of root-normalised front-end (RN LFCC) for speech recognition in wireless GSM network environments. In *Proceedings of the Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings., 1996 IEEE International Conference - Volume 01* (Washington, DC, USA, 1996), ICASSP '96, IEEE Computer Society, pp. 77–80.

[31] EISNER, J. An Interactive Spreadsheet for Teaching the Forward-Backward Algorithm. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching NLP and CL* (2002), D. Radev and C. Brew, Eds., pp. 10–18.

[32] ETSI EUROPEAN DIGITAL CELLULAR TELECOMMUNICATIONS SYSTEM. Full Rate Speech Transcoding (GSM 6.10).

[33] ETSI EUROPEAN DIGITAL CELLULAR TELECOMMUNICATIONS SYSTEM. Half Rate Speech Transcoding (GSM 6.20).

[34] ETSI SPEECH PROCESSING, TRANSMISSION AND QUALITY ASPECTS (STQ). Distributed Speech Recognition; Front-End feature extraction algorithm; Compression algorithms, (ES 201 108 Ver. 1.1.3), 2003.

[35] ETSI SPEECH PROCESSING, TRANSMISSION AND QUALITY ASPECTS (STQ). Distributed Speech Recognition; Advanced Front-End feature extraction algorithm; Compression algorithms, (ES 202 050 Ver. 1.1.1), 2004.

[36] EULER, S., AND ZINKE, J. The influence of speech coding algorithms on automatic speech recognition. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on* (apr 1994), vol. 1, pp. I/621–I/624.

[37] EUROPEAN TELECOMMUNICATIONS STANDARDS INSTITUTE, ETSI. Digital cellular communications system; enhanced full rate (EFR) speech transcoding (GSM 06.60). *Tech. Rep. ETS 300* (1997), 726.

[38] FALAVIGNA, D., MATASSONI, M., AND TURCHETTI, S. Analysis of different acoustic front-ends for automatic voice over IP recognition. In *Automatic Speech Recognition and Understanding, IEEE Workshop on* (nov.-3 dec. 2003), pp. 363–368.

[39] FANT, G. *Acoustic Theory of Speech Production*. Mouton, 1960, second printing, 1970.

[40] FANT, G. *Acoustic theory of speech production with calculations based on X-ray studies of Russian articulations*. Mouton, 1970.

[41] FLANDRIN, P., RILLING, G., AND GONCALVES, P. Empirical mode decomposition as a filter bank. *Signal Processing Letters, IEEE 11*, 2 (2004), pp. 112–114.

[42] FORNEY, G.D., J. The viterbi algorithm. *Proceedings of the IEEE 61*, 3 (march 1973), 268 – 278.

[43] FURUI, S. Cepstral analysis technique for automatic speaker verification. *Acoustics, Speech and Signal Processing, IEEE Transactions on 29*, 2 (apr 1981), 254 – 272.

[44] FURUI, S. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *Acoustics, Speech and Signal Processing, IEEE Transactions on 34*, 1 (feb 1986), pp. 52–59.

[45] GALLARDO-ANTOLIN, A., PELÁEZ-MORENO, C., AND DÍAZ-DE MARÍA, F. Recognizing GSM digital speech. *Speech and Audio Processing, IEEE Transactions on 13*, 6 (nov. 2005), pp. 1186–1205.

[46] GALLARDO-ANTOLÍN, A., DÍAZ-DE MARÍA, F., AND VALVERDE-ALBACETE, F. Recognition from GSM digital signal. *Proc. ICSLP 4* (1998), pp. 1443–1446.

[47] GALLARDO-ANTOLÍN, A., DÍAZ-DE MARÍA, F., AND VALVERDE-ALBACETE, F. Avoiding distortions due to speech coding and transmission errors in GSM ASR tasks. In *Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference - Volume 01* (Washington, DC, USA, 1999), IEEE Computer Society, pp. 277–280.

[48] GALLARDO-ANTOLÍN, A., DÍAZ-DE MARÍA, F., VALVERDE-ALBACETE, F., AND BRAVO-MENÉNDEZ-RIVAS, R. Reconocimiento de voz procedente de teléfonos móviles digitales. *Telecom I+D, Madrid* (1998), pp. 379–387.

[49] GANAPATHIRAJU, M., BALAKRISHNAN, N., AND REDDY, R. Improving Recognition Accuracy on CVSD Speech under Mismatched Conditions. *School of Computer Science, Carnegie Mellon University Vol. 2* (2003), pp. 887–892.

[50] GARCÍA-MORAL, A. I., GALLARDO-ANTOLÍN, A., DÍAZ-DE MARÍA, F., AND PELÁEZ-MORENO, C. Reconocimiento de habla distribuído en redes IP. In *Actas de las XIII JORNADAS de I+D en Telecomunicaciones (TELECOMI+D03)* (Madrid, España, 2003).

[51] GÓMEZ, A., CARMONA, J., PEINADO, A., AND SÁNCHEZ, V. A Multipulse-Based Forward Error Correction Technique for Robust CELP-Coded Speech Transmission Over Erasure Channels. *Audio, Speech, and Language Processing, IEEE Transactions on 18*, 6 (aug. 2010), 1258 –1268.

[52] GÓMEZ, A., PEINADO, A., SÁNCHEZ, V., AND RUBIO, A. Recognition of coded speech transmitted over wireless channels. *Wireless Communications, IEEE Transactions on 5*, 9 (september 2006), 2555 –2562.

[53] GÓMEZ, A., PEINADO, A., SÁNCHEZ, V., AND RUBIO, A. On the ramsey class of interleavers for robust speech recognition in burst-like packet loss. *Audio, Speech, and Language Processing, IEEE Transactions on 15*, 4 (2007), pp. 1496–1499.

[54] GÓMEZ-CAJAS, D. F., PELÁEZ-MORENO, C., AND DÍAZ-DE MARÍA, F. Reconocimiento robusto de habla en entornos IP. In *Proceedings of the International Conference on Internet Technologies* (Popayán, Colombia, 2003).

[55] GÓMEZ-CAJAS, D. F., PELÁEZ-MORENO, C., AND DÍAZ-DE MARÍA, F. Reconocimiento robusto de habla en redes IP. In *Actas de las XIII JORNADAS de I+D en Telecomunicaciones (TELECOMI+D+03)* (Madrid, España, 2003).

[56] HARISH, D., AND RAMASUBRAMANIAN, V. Comparison of segment quantizers: VQ, MQ, VLSQ and unit-selection algorithms for ultra low bit-rate speech coding. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* (31 2008-april 4 2008), pp. 4773–4776.

[57] HERMANSKY, H., AND MORGAN, N. RASTA processing of speech. *Speech and Audio Processing, IEEE Transactions on 2*, 4 (1994), pp. 578–589.

[58] HOLMA, H., AND TOSKALA, A. *WCDMA for UMTS: radio access for third generation mobile communications*. John Wiley & Sons, 2004.

[59] HONKANEN, T., VAINIO, J., JARVINEN, K., HAAVISTO, P., SALAMI, R., LAFLAMME, C., AND ADOUL, J. Enhanced full rate speech codec for IS-136 digital cellular system. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on* (1997), vol. 2, IEEE, pp. 731–734.

[60] HUERTA, J. *Speech Recognition in Mobile Environments*. PhD thesis, Department of ECE, Carnegie Mellon University, Pittsburgh, PA, 2000.

[61] HUERTA, J., AND STERN, R. Speech recognition from GSM codec parameters. In *Proc. ICSLP* (1998), vol. 4, pp. 1463–1466.

[62] ITAKURA, F. Minimum prediction residual principle applied to speech recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on 23*, 1 (feb 1975), pp. 67–72.

[63] ITU-T RECOMMENDATION G.711. Pulse code modulation (PCM) of voice frequencies. *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)* (nov 1988).

[64] ITU-T RECOMMENDATION G.711.1. Wideband embedded extension for G.711 pulse code modulation. *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)* (2008).

[65] ITU-T RECOMMENDATION G.718. Frame error robust narrowband and wideband embedded variable bit-rate coding of speech and audio from 8-32 Kbps. *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)*.

[66] ITU-T RECOMMENDATION G.723.1. Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s. *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)*.

[67] ITU-T RECOMMENDATION G.728. Coding of speech at 16 kbit/s using low-delay code excited linear prediction. *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)*.

[68] ITU-T RECOMMENDATION G.729. Coding of Speech at 8 kbit/s using Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP). *International Telecommunication Union-Telecommunication Standardisation Sector (ITU-T)* (1996).

[69] JAMES, A., AND MILNER, B. An analysis of interleavers for robust speech recognition in burst-like packet loss. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on* (2004), vol. 1, IEEE, p. 853.

[70] JÄRVINEN, K., BOUAZIZI, I., LAAKSONEN, L., OJALA, P., AND RÄMÖ, A. Media coding for the next generation mobile system LTE. *Computer Communications 33*, 16 (2010), pp. 1916–1927.

[71] JARVINEN, K., VAINIO, J., KAPANEN, P., HONKANEN, T., HAAVISTO, P., SALAMI, R., LAFLAMME, C., AND ADOUL, J. GSM enhanced full rate speech codec. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on* (1997), vol. 2, IEEE, pp. 771–774.

[72] JIANG, Z., HUANG, H., YANG, S., LU, S., AND HAO, Z. Acoustic Feature Comparison of MFCC and CZT-Based Cepstrum for Speech Recognition. In *Natural Computation, Fifth International Conference on* (aug. 2009), vol. 1, pp. 55–59.

[73] KANAL, L., AND SASTRY, A. Models for channels with memory and their applications to error control. *Proceedings of the IEEE 66*, 7 (1978), pp. 724–744.

[74] KANEDERA, N., ARAI, T., HERMANSKY, H., AND PAVEL, M. On the relative importance of various components of the modulation spectrum for automatic speech recognition. *Speech Communication 28*, 1 (1999), pp. 43–55.

[75] KANG, H., KIM, H., AND COX, R. Improving the transcoding capability of speech coders. *Multimedia, IEEE Transactions on 5*, 1 (2003), pp. 24–33.

[76] KIM, H., AND COX, R. Bitstream-based feature extraction for wireless speech recognition. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on* (2000), vol. 3, IEEE, pp. 1607–1610.

[77] KIM, H., AND COX, R. Feature enhancement for a bitstream-based front-end in wireless speech recognition. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on* (2001), vol. 1, IEEE, pp. 241–244.

[78] KIM, H., KIM, K., AND LEE, H. Enhanced distance measure for LSP-based speech recognition. *Electronics Letters 29*, 16 (aug. 1993), pp. 1463–1465.

[79] KIM, H. K., CHOI, S. H., AND LEE, H. S. On approximating line spectral frequencies to LPC cepstral coefficients. *Speech and Audio Processing, IEEE Transactions on 8*, 2 (mar 2000), pp. 195–199.

[80] KIM, H. K., AND COX, R. A bitstream-based front-end for wireless speech recognition on IS-136 communications system. *Speech and Audio Processing, IEEE Transactions on 9*, 5 (jul 2001), pp. 558–568.

[81] KIM, H. K., COX, R., AND ROSE, R. Performance improvement of a bitstream-based front-end for wireless speech recognition in adverse environments. *Speech and Audio Processing, IEEE Transactions on 10*, 8 (nov 2002), pp. 591–604.

[82] KLEIJN, W., BACKSTROM, T., AND ALKU, P. On line spectral frequencies. *Signal Processing Letters, IEEE 10*, 3 (mar 2003), 75 – 77.

[83] KOHLER, M. A comparison of the new 2400 bps MELP federal standard with other standard coders. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on* (1997), vol. 2, IEEE, pp. 1587–1590.

[84] KONDOZ, A. *Digital speech: coding for low bit rate communication systems*. Wiley, 2004.

[85] KUMAR, A. Comparative performance analysis of versions of tcp in a local network with a lossy link. *IEEE/ACM Transactions on Networking (TON) 6*, 4 (1998), pp. 485–498.

[86] LEE, G. H., YOON, J. S., OH, Y. R., AND KIM, H. K. Design of a speech coder utilizing speech recognition parameters for server-based wireless speech recognition. In *Intelligent Signal Processing and Communication Systems, 2004. ISPACS 2004. Proceedings of 2004 International Symposium on* (nov. 2004), pp. 159–163.

[87] LEE, I., STERN, H., AND MAHMOUD, S. A voice activity detection algorithm for communication systems with dynamically varying background acoustic noise. In *Vehicular Technology Conference, 48th IEEE* (may 1998), vol. 2, pp. 1214–1218 vol.2.

[88] LEE, K. *Automatic speech recognition: the development of the SPHINX system*. No. 62. Kluwer Academic Pub, 1989.

[89] LIEBERMAN, P. *The biology and evolution of language*. Harvard Univ Pr, 1984.

[90] LIEBERMAN, P., AND BLUMSTEIN, S. *Speech physiology, speech perception, and acoustic phonetics*. Cambridge studies in speech science and communication. Cambridge University Press, 1988.

[91] LILLY, B., AND PALIWAL, K. Effect of speech coders on speech recognition performance. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on* (oct 1996), vol. 4, pp. 2344–2347 vol.4.

[92] LLISTERRI, J. *El modelo de la fuente y el filtro en la síntesis del habla (libro online)*. http://liceu.uab.es/ joaquim, 2010. [Online].

[93] LOU, H.-L. Implementing the viterbi algorithm. *Signal Processing Magazine, IEEE 12*, 5 (sep 1995), 42 –52.

[94] MARKEL, J., AND GRAY, A. *Linear prediction of speech*. Springer-Verlag New York, Inc., 1982.

[95] MARTIN, R., AND COX, R. New speech enhancement techniques for low bit rate speech coding. In *Speech Coding Proceedings, 1999 IEEE Workshop on* (1999), IEEE, pp. 165–167.

[96] MCCREE, A., TRUONG, K., GEORGE, E., BARNWELL, T., AND VISWANATHAN, V. A 2.4 kbit/s MELP coder candidate for the new US Federal Standard. In *icassp* (1996), IEEE, pp. 200–203.

[97] MILNER, B. Robust speech recognition in burst-like packet loss. *Acoustics, Speech, and Signal Processing, IEEE International Conference on 1* (2001), pp. 261–264.

[98] MILNER, B., AND SEMNANI, S. Robust speech recognition over IP networks. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on* (2000), vol. 3, pp. 1791–1794 vol.3.

[99] MÜLLER, J., AND BALY, W. *The Physiology of the Senses, Voice, and Muscular Motion, with the Mental Faculties...* Taylor, Walton & Maberly, 1848.

[100] MOKBEL, C., MAUUARY, L., KARRAY, L., JOUVET, D., MONNÉ, J., SIMONIN, J., AND BARTKOVA, K. Towards improving ASR robustness for PSN and GSM telephone applications. *Speech communication 23*, 1-2 (1997), pp. 141–159.

[101] MORENO-GONZÁLEZ, J. A. AND, M.-S. J. L., DÍAZ-GUERRA-VICO, M. A., BARANDALLA-TORREGROSA, I. E., AND LORCA-HERNANDO, F. J. Simulador de enlaces para el sistema umts en modo fdd. *Comunicaciones de Telefónica I+D*, 24 (2002).

[102] NOKIA TSG R1-99B85. Effect of EEP and UEP on channel coding for AMR.

[103] OPPENHEIM, A., AND SCHAFER, R. Homomorphic analysis of speech. *Audio and Electroacoustics, IEEE Transactions on 16*, 2 (jun 1968), 221 – 226.

[104] O'SHAUGHNESSY, D., AND TOLBA, H. Towards a robust/fast continuous speech recognition system using a voiced-unvoiced decision. In *Acoustics, Speech, and Signal Processing, 1999. ICASSP'99. Proceedings., 1999 IEEE International Conference on* (1999), vol. 1, IEEE, pp. 413–416.

[105] PALIWAL, K., AND ATAL, B. Efficient vector quantization of LPC parameters at 24 bits/frame. *Speech and Audio Processing, IEEE Transactions on 1*, 1 (jan 1993), 3 –14.

[106] PALIWAL, K., AND SO, S. Scalable distributed speech recognition using multi-frame GMM-based block quantization. In *Eight International Conference on Spoken Language Processing* (2004).

[107] PAXON, V. *Measurements and Analysis of End-to-End Internet Dynamics*. PhD thesis, University of California, Berkeley, 1997.

[108] PEARCE, D., AND HIRSCH, H. G. Aurora project: Experimental framework for the performance evaluation of distributed speech recognition front-ends. *ICSLP-2000 4* (2000), pp. 29–32.

[109] PEINADO, A., AND SEGURA, J. *Speech recognition over digital channels: robustness and standards*. John Wiley, 2006.

[110] PELÁEZ-MORENO, C., AND GALLARDO-ANTOLIN, A. Recognizing voice over IP: A robust front-end for speech recognition on the World Wide Web. *Multimedia, IEEE Transactions on 3*, 2 (2001), pp. 209–218.

[111] PELÁEZ-MORENO, C., GALLARDO-ANTOLÍN, A., PARRADO-HERNÁNDEZ, E., AND DÍAZ-DE MARÍA, F. SVM-based lost packets concealment for ASR applications over IP. *XI European signal Processing Conference - EUSIPCO III* (2002), pp. 529–532.

[112] PELÁEZ-MORENO, C., PARRADO-HERNÁNDEZ, E., GALLARDO-ANTOLÍN, A., ZAMBRANO-MIRANDA, A., AND DÍAZ-DE MARÍA, F. An application of SVM to lost packets reconstruction in voice-enabled services. *Artificial Neural Networks, ICANN 2002* (2002), pp. 793–793.

[113] PELÁEZ-MORENO, C. *Reconocimiento de habla mediante transparametrización: una alternativa robusta para entornos móviles e IP*. PhD thesis, Universidad Carlos III de Madrid, 2002.

[114] PELÁEZ-MORENO, C., GALLARDO-ANTOLÍN, A., AND DÍAZ-DE MARÍA, F. Recognizing voice over IP: a robust front-end for speech recognition on the world wide web. *Multimedia, IEEE Transactions on 3*, 2 (jun 2001), pp. 209–218.

[115] PELÁEZ-MORENO, C., GALLARDO-ANTOLÍN, A., GÓMEZ-CAJAS, D. F., AND DÍAZ-DE MARÍA, F. A comparison of front-ends for bitstream-based ASR over IP. *Signal Processing 86*, 7 (JUL 2006), pp. 1502–1508.

[116] PERKINS, C., HODSON, O., AND HARDMAN, V. A survey of packet loss recovery techniques for streaming audio. *Network, IEEE 12*, 5 (sep/oct 1998), pp. 40–48.

[117] PERKINS, M., EVANS, K., PASCAL, D., AND THORPE, L. Characterizing the subjective performance of the ITU-T 8 kb/s speech coding algorithm-ITU-T G.729. *Communications Magazine, IEEE 35*, 9 (sep 1997), pp. 74–81.

[118] PERSSON, F. Voice over IP Realized for the 3GPP Long Term Evolution. In *Vehicular Technology Conference, 2007. VTC-2007 Fall. 2007 IEEE 66th* (30 2007-oct. 3 2007), pp. 1436–1440.

[119] RABINER, L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 2 (feb 1989), pp. 257–286.

[120] RABINER, L., AND JUANG, B. An introduction to hidden markov models. *ASSP Magazine, IEEE 3*, 1 (1986), pp. 4–16.

[121] RABINER, L., AND JUANG, B. *Fundamentals of speech recognition*. Prentice Hall signal processing series. PTR Prentice Hall, 1993.

[122] ROSE, R., PARTHASARATHY, S., GAJIC, B., ROSENBERG, A., AND NARAYANAN, S. On the implementation of ASR algorithms for hand-held wireless mobile devices. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on* (2001), vol. 1, pp. 17–20.

[123] SALAMI, R., LAFLAMME, C., ADOUL, J.-P., KATAOKA, A., HAYASHI, S., MORIYA, T., LAMBLIN, C., MASSALOUX, D., PROUST, S., KROON, P., AND SHOHAM, Y. Design and description of CS-ACELP: a toll quality 8 Kbps speech coder. *Speech and Audio Processing, IEEE Transactions on 6*, 2 (mar 1998), pp. 116–130.

[124] SALAMI, R., LAFLAMME, C., ADOUL, J.-P., AND MASSALOUX, D. A toll quality 8 Kbps speech codec for the personal communications system (PCS). *Vehicular Technology, IEEE Transactions on 43*, 3 (aug 1994), pp. 808–816.

[125] SALAMI, R., LAFLAMME, C., BESSETTE, B., AND ADOUL, J. Description of ITU-T Recommendation G. 729 Annex A: reduced complexity 8 kbit/s CS-ACELP codec. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on* (1997), vol. 2, IEEE, pp. 775–778.

[126] SALAMI, R., LAFLAMME, C., BESSETTE, B., AND ADOUL, J. ITU-T G. 729 Annex A: reduced complexity 8 kb/s CS-ACELP codec for digital simultaneous voice and data. *Communications Magazine, IEEE 35*, 9 (1997), pp. 56–63.

[127] SAMBUR, M. Adaptive noise canceling for speech signals. *Acoustics, Speech and Signal Processing, IEEE Transactions on 26*, 5 (1978), pp. 419–423.

[128] SCHROEDER, M., AND ATAL, B. Code-excited linear prediction(CELP): High-quality speech at very low bit rates. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.* (apr 1985), vol. 10, pp. 937–940.

[129] SJOBERG, J., WESTERLUND, M., LAKANIEMI, A., AND XIE, Q. RFC3267, Real-time transport protocol (RTP) payload format and file storage format for the adaptive multi-rate (AMR) and adaptive multi-rate wideband (AMR-WB) audio codecs.

[130] SMITH, J. O. *Spectral Audio Signal Processing, October 2008 Draft*. http://ccrma.stanford.edu/ jos/sasp/, 2008. [Online].

[131] SÁNCHEZ-FERNÁNDEZ, M. P. *Contribución al estudio de las prestaciones de esquemas de codificación basados en turbo códigos para sistemas de comunicaciones móviles de tercera generación*. PhD thesis, Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, Tesis Doctoral, 2001.

[132] SO, S., AND PALIWAL, K. K. Improved noise-robustness in distributed speech recognition via perceptually-weighted vector quantisation of filterbank energies. In *INTERSPEECH* (2005), pp. 941–944.

[133] SOLDANI, D., AND DIXIT, S. Wireless relays for broadband access [radio communications series]. *Communications Magazine, IEEE 46*, 3 (march 2008), pp. 58–66.

[134] STEPHENSON, T., ESCOFET, J., MAGIMAI-DOSS, M., AND BOURLARD, H. Dynamic Bayesian network based speech recognition with pitch and energy as auxiliary variables. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on* (2002), IEEE, pp. 637–646.

[135] STEVENS, S. S., VOLKMANN, J., AND NEWMAN, E. B. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America 8*, 3 (1937), pp. 185–190.

[136] SUGAMURA, N. ITAKURA, F. Speech analysis and synthesis methods developed at ECL in NTT - from LPC to LSP. *Speech Communications 5* (1986), pp. 199–215.

[137] SUN, L., WADE, G., LINES, B., AND IFEACHOR, E. Impact of packet loss location on perceived speech quality. In *2nd IP-Telephony Workshop* (2001), pp. 114–122.

[138] TAN, Z., AND LINDBERG, B. *Automatic speech recognition on mobile devices and over communication networks*. Springer-Verlag New York Inc, 2008.

[139] TECHNICAL SPECIFICATION GROUP RADIO ACCESS NETWORK. Channel coding and multiplexing examples, (3GPP TR 25.944), 2004-06.

[140] TECHNICAL SPECIFICATION GROUP RADIO ACCESS NETWORK. Physical layer - General description, (3GPP TS 25.201), 2004-06.

[141] TECHNICAL SPECIFICATION GROUP RADIO ACCESS NETWORK. Physical channels and mapping of transport channels onto physical channels (FDD), (3GPP TS 25.211), 2004-06.

[142] TECHNICAL SPECIFICATION GROUP RADIO ACCESS NETWORK. Mapping of Transport Channels Onto Physical Channels (FDD), (3GPP TS 25.211 Release 6), 2003.

[143] TECHNICAL SPECIFICATION GROUP RADIO ACCESS NETWORK. Multiplexing and channel coding (FDD), (3GPP TS 25.212), 2004.

[144] TECHNICAL SPECIFICATION GROUP RADIO ACCESS NETWORK. Multiplexing and channel coding (TDD), (3GPP TS 25.222), 2003-12.

[145] TECHNICAL SPECIFICATION GROUP RADIO ACCESS NETWORK. Services provided by the physical layer, (3GPP TS 25.302), 2004-06.

[146] TECHNICAL SPECIFICATION GROUP TERMINALS. Common test environments for User Equipment (UE); conformance testing, (3GPP TS 34.108), 2004-06.

[147] THOMSON, D., AND CHENGALVARAYAN, R. Use of periodicity and jitter as speech recognition features. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on* (may 1998), vol. 1, pp. 21–24 vol.1.

[148] TURUNEN, J., AND VLAJ, D. A study of speech coding parameters in speech recognition. In *Seventh European Conference on Speech Communication and Technology, EUROSPEECH-2001* (2001), pp. 2363–2366.

[149] VAHATALO, A., AND JOHANSSON, I. Voice activity detection for GSM adaptive multi-rate codec. In *Speech Coding Proceedings, IEEE Workshop on* (1999), pp. 55–57.

[150] VARGA, A., AND STEENEKEN, H. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication 12*, 3 (1993), pp. 247–251.

[151] VARGA, I., DE LACOVO, R., AND USAI, P. Standardization of the AMR wideband speech codec in 3GPP and ITU-T. *Communications Magazine, IEEE 44*, 5 (may 2006), 66 –73.

[152] VICENTE-PEÑA, J., GALLARDO-ANTOLÍN, A., PELÁEZ-MORENO, C., AND DÍAZ-DE MARÍA, F. Band-pass filtering of the time sequences of spectral parameters for robust wireless speech recognition. *Speech Communication 48*, 10 (OCT 2006), pp. 1379–1398.

[153] WALKE, B., SEIDENBERG, P., AND ALTHOFF, M. *UMTS: the fundamentals*. Wiley, 2003.

[154] WANG, J., AND GIBSON, J. Parameter interpolation to enhance the frame erasure robustness of CELP coders in packet networks. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on* (2001), vol. 2, pp. 745–748 vol.2.

[155] WANG, J., HE, H., AND KUANG, J. Quality enhancement of coded transient audio with a post-filter in frequency domain. In *Signal Processing (ICSP), 2010 IEEE 10th International Conference on* (oct. 2010), pp. 506–509.

[156] WET, F., CRANEN, B., VETH, J., AND BOVES, L. A comparison of LPC and FFT-based acoustic features for noise robust ASR. In *Seventh European Conference on Speech Communication and Technology* (2001), pp. 865–868.

[157] YANG, M. Low bit rate speech coding. *Potentials, IEEE 23*, 4 (oct.-nov. 2004), 32 – 36.

[158] YOUNG, S., EVERMANN, G., KERSHAW, D., MOORE, G., ODELL, J., OLLASON, D., POVEY, D., VALTCHEV, V., AND WOODLAND, P. The HTK book (for HTK version 3.2). *Cambridge University Engineering Department* (2002).

[159] ZHENG, F., SONG, Z., LI, L., YU, W., ZHENG, F., AND WU, W. The distance measure for line spectrum pairs applied to speech recognition. In *Fifth International Conference on Spoken Language Processing* (1998), vol. 98, pp. 1123–1126.

[160] ZHONG, X., ARROWOOD, J., MORENO, A., AND CLEMENTS, M. Multiple description coding for recognizing voice over IP. In *Digital Signal Processing Workshop, 2002 and the 2nd Signal Processing Education Workshop. Proceedings of 2002 IEEE 10th* (2002), IEEE, pp. 383–386.

[161] ZHU, Q., AND ALWAN, A. An efficient and scalable 2D DCT-based feature coding scheme for remote speech recognition. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on* (2001), vol. 1, IEEE, pp. 113–116.

Anexo A

Técnicas de Reconocimiento

Resumen de Contribuciones

La nomenclatura de las técnicas a comparar será la siguiente:

Nombre Genérico		Realización	Envolvente Espectral	Acrónimo
Reconocimiento de voz Decodificada	Experimento Decodificado	RAW-FFT-MEL-LOG-DCT	$mfc_h^{(12)}$	MFCC (Mel Frequency Cepstrum Coefficients)
	Experimento Suavizado	RAW-LPC-SPEC-MEL-LOG-DCT	$mfc_{h_{LP}}^{(12)}$	LP-MFCC (Linear Prediction - MFCC)
	Experimento Suavizado Sin Postfiltro	RAW(Sin postfiltro) - LPC-SPEC-MEL-LOG-DCT	$mfc_{h_{LP}}^{(12)}$	LP-MFCC* (Linear Prediction - MFCC)
	Experimento Suavizado Extendido (con Pitch)	RAW-LPC-SPEC-MEL-LOG-DCT	$mfc_{h_{LP}}^{(12)}$	XLP-MFCC (Extended LP-MFCC)
Reconocimiento Mediante Transparametrización	Experimento Transparametrizado	Bitstream-LSP-SPEC-MEL-LOG-DCT	$mfc_{h_{bLP}}^{(12)}$	bLP-MFCC (Bitstream based LP-MFCC)
	Experimento Transparametrizado (con Estima de Energía Mejorada)	Bitstream-LSP-SPEC-MEL-LOG-DCT	$mfc_{h_{bLP}}^{(12)}$	bLP-MFCC+ (Bitstream based LP-MFCC+)
	Experimento Transparametrizado Filtrado (con Estima de Energía Mejorada)	Bitstream-LSP-SPEC-MEL-LOG-FILT-DCT	$mfc_{h_{bLP}}^{(12)}$	FbLP-MFCC+ (Filtred bLP-MFCC+)
	Experimento Transparametrizado Extendido (Con Pitch y Ganancia de Códigos Adaptativos)	Bitstream-LSP-SPEC-MEL-LOG-DCT	$mfc_{h_{bLP}}^{(12)}$	XbLP-MFCC (Extended bLP-MFCC)
	Experimento Transparametrizado Extendido (Con Pitch, Ganancia de Códigos Adaptativos, y Estima de Energía Mejorada)	Bitstream-LSP-SPEC-MEL-LOG-DCT	$mfc_{h_{bLP}}^{(12)}$	XbLP-MFCC+ (Extended bLP-MFCC+)
	Pseudo Cepstrum	Bitstream-LSP-MEL-PCEPS	$\hat{mfc}_{h_{LP}}^{(12)}$	pLP-MFCC (Pseudo LP-MFCC)

Tabla A.1: Nomenclatura usada para describir las técnicas de reconocimiento y sus variantes.

Nombre Genérico	Descripción	Acrónimo
Estima de la Energía	Procedimiento de estima de la energía basado en los parámetros del bitstream	Eest
Estima de la Energía Plus	Estima de la energía, ponderando la energía de la excitación, con la ventana de análisis del codificador (WG729)	$Eest^+$
Energía Decodificada	Energía calculada a partir de voz decodificada	Edec

Nombre	Descripción	Acrónimo
Linear Prediction Coefficients	Son usados para describir la envolvente espectral	lpc

Anexo B

Codificador AMR-NB

Modo	Parametro	Subtrama 1	Subtrama 2	Subtrama 3	Subtrama 4	Total bits por trama
12.2 kbit/s (GSM EFR)	2 conjuntos de LSP					38
	Periodo fundamental	9	6	9	6	30
	Ganancia de pitch	4	4	4	4	16
	Código algebraico	35	35	35	35	140
	Ganancias	5	5	5	5	20
	Total					244
10.2 kbit/s	Conjunto de LSP					26
	Periodo fundamental	8	5	8	5	26
	Código algebraico	31	31	31	31	124
	Ganancias	7	7	7	7	28
	Total					204
7.95 kbit/s	Conjunto de LSP					27
	Periodo fundamental	8	6	8	6	28
	Ganancia de pitch	4	4	4	4	16
	Código algebraico	17	17	17	17	68
	Ganancias	5	5	5	5	20
	Total					159
7.40 kbit/s (TDMA EFR)	Conjunto de LSP					26
	Periodo fundamental	8	5	8	5	26
	Código algebraico	17	17	17	17	68
	Ganancias	7	7	7	7	28
	Total					148
6.70 kbit/s (PDC EFR)	Conjunto de LSP					26
	Periodo fundamental	8	4	8	4	24
	Código algebraico	14	14	14	14	56
	Ganancias	7	7	7	7	28
	Total					134
5.90 kbit/s	Conjunto de LSP					26
	Periodo fundamental	8	4	8	4	24
	Código algebraico	11	11	11	11	44
	Ganancias	6	6	6	6	24
	Total					118
5.15 kbit/s	Conjunto de LSP					23
	Periodo fundamental	8	4	4	4	20
	Código algebraico	9	9	9	9	36
	Ganancias	6	6	6	6	24
	Total					103
4.75 kbit/s	Conjunto de LSP					23
	Periodo fundamental	8	4	4	4	20
	Código algebraico	9	9	9	9	36
	Ganancias	8		8		16
	Total					95

Tabla B.1: Asinación binaria por parámetro codificado para todos los modos del AMR-NB.