TESIS DOCTORAL

RECONOCIMIENTO DE HABLA MEDIANTE TRANSPARAMETRIZACIÓN: UNA ALTERNATIVA ROBUSTA PARA ENTORNOS MÓVILES E IP

Autora:

Carmen Peláez Moreno

Ingeniera de Telecomunicación

Director:

Prof. Dr. Fernando Díaz de María

Prof. Titular del Dpto. de Teoría de la Señal y Comunicaciones

Universidad Carlos III de Madrid

Dpto de. TEORÍA DE LA SEÑAL Y COMUNICACIONES

ESCUELA POLITÉCNICA SUPERIOR

UNIVERSIDAD CARLOS III DE MADRID

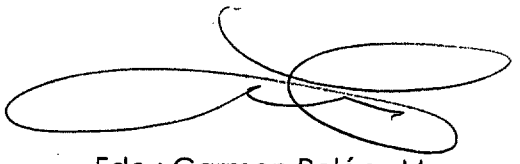
Leganés, 2002

Dña. Carmen Peláez Moreno, con D. N. I. : 33442266 J

AUTORIZA:

A que su tesis doctoral con el título: **"Reconocimiento de habla mediante transparametrización: una alternativa robusta para entornos móviles e IP"** pueda ser utilizada para fines de investigación por parte de la Universidad Carlos III de Madrid.

Leganés, 5 de febrero de 2002



Fdo.: Carmen Peláez Moreno

TESIS DOCTORAL: RECONOCIMIENTO DE HABLA MEDIANTE TRANSPARAMETRIZACIÓN: UNA ALTERNATIVA ROBUSTA PARA ENTORNOS MÓVILES E IP

AUTORA: Carmen Peláez Moreno

DIRECTOR: Prof. Dr. Fernando Díaz de María

El tribunal nombrado para juzgar la tesis arriba indicada, compuesto por los doctores:

PRESIDENTE:

VOCALES:

Fdo. Luis A. Hernández Gómez

Fdo. Fco. Javier Hernando Pericás

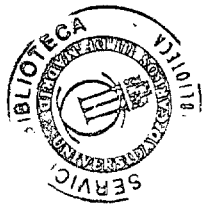
Carmen García Mateo

SECRETARIO:

Belen Ruiz.

acuerda otorgarle la calificación de: SOBRESALIENTE CUM LAUDE (POR UNANIMIDAD.

Leganés, 2002



El Secretario del Tribunal

A mis padres y mis hermanos, por su apoyo.

A Fran, por su cariño.

Agradecimientos

Haciendo balance de los años que han supuesto la elaboración de este trabajo, las dificultades, las dudas de todo tipo que ha sido necesario solucionar y también aquellas que han quedado pendientes, me gustaría, en primer lugar, agradecer a Ascensión Gallardo Antolín su gran colaboración, sin la cual esta tesis no habría podido llevarse a cabo, a Harold Molina Bulla y a Antonio Caamaño Fernández su importante apoyo, tanto profesional como personal, a Maryan Vázquez Castro, sus contribuciones en la simulación del canal GSM y a todo el resto de compañeros de este departamento por compartir los abatares de cada día.

Por supuesto, tengo la obligación de agradecer a Fernando Díaz de María su enorme interés y dedicación, su ilusión y su guía, y a Anibal Figueiras Vidal la posibilidad de realizar la presente tesis en este departamento.

Además, tengo también una deuda con mis familiares, amigos y todas las demás personas que han estado a mi lado durante todo este tiempo.

Resumen

En el panorama actual de las telecomunicaciones, dos son los tipos de redes con mayor éxito en la actualidad: las redes de móviles y las redes de paquetes basadas en el protocolo TCP/IP (*–Transport Control Protocol / Internet Protocol–*). Entre los factores que han llevado al éxito de las primeras en su segunda generación (2G) está su ubicuidad, es decir, gracias al enorme despliegue geográfico de estas redes es posible realizar una llamada telefónica desde casi cualquier localización (en el mundo desarrollado). Por su parte, las redes IP (originalmente diseñadas para el transporte de datos) también están logrando imponer su presencia en detrimento de cualquier otro tipo de red fija y uno de sus puntos fuertes es, sin duda, su capacidad –todavía bastante limitada– para transmitir cualquier tipo de información multimedia.

Uno de los puntos de convergencia entre las dos redes es su objetivo de permitir que todo tipo de información transite por ellas con ciertas garantías de calidad de servicio (QoS *–Quality of Service–*). Esto está motivado por la cantidad de nuevas aplicaciones que pueden crearse a partir de la posibilidad de combinar informaciones de distinto tipo (texto, video, voz, imágenes, música, etc.) y las tecnologías del habla están llamadas a jugar un papel fundamental a través del desarrollo de interfaces más naturales para estas aplicaciones.

Entre estas tecnologías, el reconocimiento de habla está llegando a una fase de madurez que hace cada vez más viables estos desarrollos. De hecho, desde hace algún tiempo se viene prestando mucha atención a la robustez de estos sistemas cuando se trasladan al mundo real, habiéndose desarrollado numerosas técnicas para enfrentarse a problemas tales como: variaciones en el entorno acústico, influencia de los transductores y el canal de transmisión y variaciones en el hablante y la tarea que se aborda.

En esta tesis estudiamos la influencia de dos tipos de canales de transmisión concretos, representantes de los dos tipos de redes que hemos venido introduciendo: el estándar europeo para comunicaciones móviles GSM (*–Global System for Mobile–*, anteriormente *–Group Speciale Mobile–*) y el de las actuales redes basadas en los protocolos TCP/IP. Además, proponemos una solución, que hemos denominado *reconocimiento mediante transparametrización*, con la que mejoramos la tasas de reconocimiento en ambos entornos y que, aunque en un principio, hemos particularizado para dichos entornos, puede ser aplicada en otros.

La característica común de la transmisión de voz a través de estas dos redes es el proceso de codificación que tiene lugar para adecuar su régimen binario reducir. Esta compresión con pérdidas de la señal de voz produce un deterioro de su calidad, que si bien es aceptable en el caso de reconocedores humanos –los codificadores están

diseñados para minimizar la distorsión perceptible–, se traduce en una disminución apreciable de las prestaciones de los reconocedores automáticos.

Por otra parte, los errores de transmisión que se producen en ambos entornos, contribuyen también a la degradación de las prestaciones de los reconocedores. En GSM, estos errores aparecen en forma de ráfagas de bits erróneos producidas por desvanecimientos de la señal de radiofrecuencia, que pueden afectar a una o varias tramas consecutivas, completamente o sólo en parte. El caso de IP es algo distinto ya que, en general, no se suelen producir errores de bit a ráfagas y muy raramente errores aislados, debido a la alta fiabilidad del canal, sino que lo más común es que se produzcan pérdidas de paquetes (a ráfagas) en los nodos de enrutamiento.

En cualquier caso, lo que se pone de manifiesto en esta tesis es que el hecho de que este tipo de errores se produzcan sobre la voz codificada tiene consecuencias que no se pueden tratar de la misma manera que si se produjeran sobre la señal de voz original (por ejemplo, modelando los errores haciendo las hipótesis habituales de ruido convolutivo o aditivo). Es decir, si tenemos en cuenta que el proceso de codificación de la voz consiste, a grandes rasgos, en la extracción de una serie de parámetros que representan distintos aspectos específicos de este tipo de señal (su periodo fundamental, la posición de sus formantes, su característica sonora o sorda, su energía, etc), nos percataremos de que la modificación de cada uno de ellos tiene consecuencias muy distintas sobre la señal vocal reconstruida.

Un reconocedor convencional que recibe una señal de voz codificada, la primera acción que realiza sobre ella es su decodificación y de esa forma, ya puede proceder a realizar la extracción de características o parametrización para reconocimiento. En este proceso, las distorsiones de codificación y de los errores se trasladan a los parámetros a partir de los que se realizará el reconocimiento, produciendo el deterioro de las prestaciones del reconocedor.

Para mejorar esta situación, en esta tesis proponemos el análisis de la parametrización de la señal de voz que lleva a cabo el codificador antes de su decodificación y la transformación de ésta en otra adecuada para el reconocimiento. Esto además, nos permite utilizar métodos de recuperación frente a errores y de transformación de parametrizaciones orientados directamente al reconocimiento, sin limitarnos a los ya previstos en los estándares de codificación, cuyo propósito es recuperar una señal de voz perceptualmente aceptable sujetos a una fuerte restricción de tiempo real.

De esta forma, obtenemos una solución aplicable a ambos entornos (GSM e IP) que reduce la influencia sobre los reconocedores, por una parte, la distorsión de codificación haciendo una selección de la información relevante para reconocimiento, y por otra, el efecto de los errores de transmisión, actuando directamente sobre los parámetros afectados. Resulta notable el hecho de que esta solución sea aplicable tanto a entornos móviles como a redes de tipo IP, ya que puede emplearse cuando existe una combinación de ambas como parece ser la tendencia.

Abstract

Nowadays, two are the most important types of telecommunication networks worldwide: mobile networks and those based on the TCP/IP (*–Transport Control Protocol / Internet Protocol–*) protocol. Among the factors which have led the first ones to success in its second generation (2G) is its ubiquity, i.e., the enormous geographic deployment of these networks makes it possible to place a phone call virtually from any location (in the developed world). By its side, IP networks (originally designed for data transport) are succeeding in progressively substituting other kinds of fixed networks and one of its most remarkable advantages is, no doubt, its ability –still quite limited– for transmitting any kind of multimedia information.

One of the points of convergence between both networks is thus, its aim to provide the means for any kind of information to be sent over it with a certain quality of service (QoS). This is motivated by the huge range of new applications that can be created due to the possibility of combining multiple kind of information types (text, video, voice, images, music, etc.), and here is where speech technologies are called to play a fundamental role by providing more natural interfaces.

Among these technologies, speech recognition is attaining enough maturity to allow for these developments. In fact, during the last years, much attention has been payed to the robustness of these systems when they are transferred to the real world, having many techniques been developed to cope with problems such as: acoustic environment variations, transductors and transmission channels influences and variations in the speaker and tasks.

In this thesis we will analyse the influence of two specific transmission channel types, which represent an example of both types of network introduced above: European mobile communications standard, GSM (*–Global System for Mobile–*, which previously stood for *–Group Speciale Mobile–*) and the present-day TCP/IP-based networks. Besides, we propose an alternative solution which we have named *recognition by transparameterization* by which we attain improvements in the performances of recognizers under both environments conditions and, though tested specifically in those particular networks, also potencially applicable in some other.

A common characteristic of speech transmission over these networks is the previous speech coding process that takes place to adequate the bit rate. This lossy process produces a quality drop which, though not very disappointing for human recognizers –as codecs are designed to minimize perceptual distortion–, sometimes harmful for automatic speech recognizers.

Besides, transmission errors caused by both networks, contribute to the degradation of the recognizers accuracy. Under the GSM environment, these errors appear in the form of bit bursts produced by signal fadings in radio-frequency channels which can affect one or more consecutive frames, just in part or completely. IP-based networks is somehow different because, generally speaking, isolated or burst bit errors are not usually encountered due to the high channel reliability, but on the contrary, it is common for the router nodes to drop packets during congestion situations.

In any case, what is here emphasized is that the fact that when these errors are placed on the coded speech signal produce effects that cannot be treated the same way as if they had been produced on the original one (for example, modelling error sources as convolutive or additive). In other words, if we take into account that the speech coding process consists, roughly speaking, on the extraction of a series of parameters representing the various aspects of this particular kind of signal (its fundamental frequency, formant positions, voicing mode, energy, etc.), it will become apparent that the modification of each of them affects in a different manner the reconstructed speech.

A conventional speech recognizer receiving a coded signal, first proceeds to its decoding which enables the usual feature extraction procedure that recognizers habitually perform. By this process, coding and errors distortions are transferred to the recognition parameterization, finally producing a disminishing of the recognition rates.

To improve this situation, we propose the analysis of the speech codec parameterization before its decoding and its transformation into an adequate one for recognition. This, besides, allows us to use recognition-oriented error recovery and parameterization transformation methods not restricted to the ones provided by the coding standards, whose aim is to recover perceptually acceptable signals subject to very strict real-time restrictions.

Therefore, we have obtained a method applicable to both environments, by which we are capable of reducing the influence on the recognizer by minimizing, on the one hand, the coding distortion by means of making a selection of the relevant recognition information embedded into the speech coding parameterization and, on the other hand, the effect of the transmission errors, by acting straight on the affected parameters. It is worth mention that the fact that this method is effective both under mobile and IP networks environments can represent an advantage for its application in combined situations, which seems to be the trend.

Contenido

Glosario de acrónimos

AbS:	Analysis by Synthesis
ACELP:	Algebraic Code-Excited Linear Predicition
ADPCM:	Adaptive Differential Pulse Code Modulation
AMR:	Adaptive Multi-Rate
APC:	Adaptive Predictive Coding
BER:	Bit Error Rate
BFI:	Bad Frame Indicator
BTS:	Base Transceiver Station
CMS:	Cepstral Mean Substraction
CRC:	Cyclic Redundancy Code
CS:	Conjugate Structure
CSR:	Continuous Speech Recognition
DAM:	Diagnostic Acceptability Measure
DCT:	Discrete Cosine Transform
DFT:	Discrete Fourier Transform
DPCM:	Differential PCM
DRT:	Diagnostic Rhyme Test
DS:	Differentiated Service
DSR:	Distributed Speech Recognition
DTMF:	Dual Tone Multiple Frequency
DTW:	Dynamic Time Warping
ETSI:	European Telecommunications Standards Institute
FEC:	Forward Error Correction

FER:	Frame Error Rate
GSM:	Global System for Mobile
GSM-EFR:	GSM-Enhanced Full Rate
GSM-FR:	GSM-Full Rate
GSM-HR:	GSM-Half Rate
HMM:	Hidden Markov Models
HTK:	Hidden markov model ToolKit
IANA:	Internet Assigned Numbers Authority
IDR:	Isolated Digit Recognition
IETF:	Internet Engineering Task Force
IP:	Internet Protocol
IPNG:	Internet Protocol Next Generation
IPPM:	Internet Protocol Performance Metrics
IS:	Inverse Sine
ISA:	Integrated Services Architecture
ITU:	International Telecommunications Union
LAN:	Local Area Networks
LAR:	Log Area Ratio
LDA:	Linear Discriminant Analysis
LP:	Linear Prediction
LPC:	Linear Prediction Coefficients
LSP:	Line Spectral Pairs
LTP:	Long Term Predictor
MA:	Moving Average
MAP:	Maximum a Posteriori
MBL:	Mean Burst Length
MELP:	Mixed Excitation Linear Prediction
MFCC:	Mel Frequency Cepstral Coefficients
MGCP:	Media Gateway Control Protocol

MLLR:	Maximum Likelihood Linear Regresión
MOS:	Mean Opinión Score
MP-MLQ:	Multi-Pulse Maximum Likelihood Quantization
MS:	Mobile Station
MSE:	Mean Square Error
NSS:	Non-linear Spectral Substraction
PARCOR:	PARtial CORrelation
PCA:	Principal Component Analysis
PCM:	Pulse Coded Modulation
PHB:	Per Hop Behaviour
PLP:	Perceptual Linear Prediction
PLR:	Packet Loss Rate
POF:	Probabilistic Optimum Filtering
QoS:	Quality of Service
RBER:	Residual Bit Error Rate
RC:	Reflexion Coefficients
RECOVC:	Recognition COmpatible Voice Compression
RM1:	Resource Management part 1
RMT:	Reconocimiento Mediante Transparametrización
RN_LFCC:	Root-Normalized Linear Frequency Cepstral Coefficients
RNN:	Recurrent Neural Network
RPE:	Regular Pulse Excitation
RPE-LTP:	Regular Pulse Excitation-Long Term Prediction
RSVP:	Resource reSerVation Protocol
RTCP:	Real-time Transport Control Protocol
RTP:	Real-time Transport Protocol
SA:	Speaker Adaptive
SD:	Speaker Dependent
SI:	Speaker Independent

SIP:	Session Initialization Protocol
TCH/FS:	Traffic CHannel/Full rate Speech
TCH/HS:	Traffic CHannel/Half rate Speech
TCP:	Transport Control Protocol
TFO:	Tandem Free Operation
TIPHON:	Telecommunications and Internet Protocol Harmonization Over Networks
ToS:	Type of Service
UDP:	User Datagram Protocol
UFI:	Unreliable Frame Indicator
UMTS:	Universal Mobile Telecommunication System
VAD:	Voice Activity Detection
VCB:	Vocoder Bypass
VoIP:	Voice over IP. Voice over IP
VSE:	Vector Sum Excitation
VSELP:	Vector Sum Excited Linear Prediction
WER:	Word Error Rate

Capítulo I
Introducción

En el panorama actual de las telecomunicaciones, dos son los tipos de redes con mayor éxito en la actualidad: las redes de móviles y las redes de paquetes basadas en el protocolo TCP/IP (*–Transport Control Protocol / Internet Protocol–*). Entre los factores que han llevado al éxito de las primeras en su segunda generación (2G) está su ubicuidad, es decir, gracias al enorme despliegue geográfico de estas redes es posible realizar una llamada telefónica desde casi cualquier localización (en el mundo desarrollado). Por su parte, las redes IP (originalmente diseñadas para el transporte de datos) también están logrando imponer su presencia en detrimento de cualquier otro tipo de red fija y uno de sus puntos fuertes es, sin duda, su capacidad –todavía bastante limitada– para transmitir cualquier tipo de información multimedia.

Ambas redes, sin embargo, presentan hoy por hoy, serias limitaciones que están tratando de ser subsanadas en sus próximas generaciones. En concreto, la tercera generación de móviles (3G) o UMTS (*–Universal Mobile Telecommunication System–*), persigue el objetivo de proporcionar a cada usuario un ancho de banda superior a la anterior, de forma que puedan transmitirse datos a mayor velocidad y, en definitiva, de permitir la transmisión, de nuevo, de información multimedia.

Por su parte, lo que se ha denominado "Internet Siguiente Generación" (IPNG – *Internet Protocol Next Generation–*) trata de paliar los problemas derivados del crecimiento desmesurado e imprevisto de estas redes y de la asignación de unas nuevas funcionalidades para las que no estaba inicialmente preparada. Así, aunque el detonante de este cambio de generación fue el solucionar el problema del numero insuficiente de direcciones previstas originalmente, no menos importante resulta, hoy, el trabajo que se está llevando a cabo para la adecuación de este tipo de redes al tráfico con requerimientos de tiempo real (voz y vídeo, principalmente), y que ya está empezando a ser aplicado incluso sobre las redes actuales.

En definitiva, ambas redes evolucionan para proporcionar a sus usuarios una calidad de servicio (QoS *–Quality of Service–*) superior y, sobre todo, negociable. Así, en UMTS se definen 4 clases de QoS que cada usuario puede negociar con su proveedor, y en las redes IP, el protocolo RSVP (*–Resource reSerVation Protocol–*) permite el control de los recursos de ancho de banda y la autenticación de los usuarios.

Por otra parte, la gran aceptación que están teniendo estos dos tipos de redes crea una demanda muy fuerte de posibilidades de interconexión entre ambas: ya existen grupos de trabajo y foros que discuten la problemática de la movilidad en el protocolo

IP, la convergencia entre ellas e incluso la conveniencia de la implantación del protocolo IP en las redes de móviles para generar una única red "todo-IP" ("*All-IP*").

Como vemos, uno de los puntos de convergencia entre las dos redes es su objetivo de permitir que todo tipo de información transite por ellas con ciertas garantías de calidad de servicio (QoS –*Quality of Service*–). Esto está motivado por la cantidad de nuevas aplicaciones que pueden crearse a partir de la posibilidad de combinar informaciones de distinto tipo (texto, video, voz, imágenes, música, etc.) y las tecnologías del habla están llamadas a jugar un papel fundamental a través del desarrollo de interfaces más naturales para estas aplicaciones.

Entre estas tecnologías, el reconocimiento de habla está llegando a una fase de madurez que hace cada vez más viables estos desarrollos. De hecho, desde hace algún tiempo se viene prestando mucha atención a la robustez de estos sistemas cuando se trasladan al mundo real, habiéndose desarrollado numerosas técnicas para enfrentarse a problemas tales como:

- variaciones en el entorno acústico,
- influencia de los transductores y el canal de transmisión y
- variaciones en el hablante y la tarea que se aborda.

En esta tesis estudiamos la influencia de dos tipos de canales de transmisión concretos, representantes de los dos tipos de redes que hemos venido introduciendo: el estándar europeo para comunicaciones móviles GSM (–*Global System for Mobile*–, anteriormente –*Group Speciale Mobile*–) y el de las actuales redes basadas en los protocolos TCP/IP. Además, proponemos una solución, que hemos denominado *reconocimiento mediante transparametrización*, con la que mejoramos la tasas de reconocimiento en ambos entornos y que, aunque en un principio hemos particularizado para dichos entornos, puede ser aplicada en otros.

La característica común de la transmisión de voz a través de estas dos redes es el proceso de codificación que tiene lugar para adecuar su régimen binario reducir. Esta compresión con pérdidas de la señal de voz produce un deterioro de su calidad, que si bien es aceptable en el caso de reconocedores humanos –los codificadores están diseñados para minimizar la distorsión perceptible–, se traduce en una disminución apreciable de las prestaciones de los reconocedores automáticos.

Por otra parte, los errores de transmisión que se producen en ambos entornos, contribuyen también a la degradación de las prestaciones de los reconocedores. En GSM, estos errores aparecen en forma de ráfagas de bits erróneos producidas por desvanecimientos de la señal de radiofrecuencia, que pueden afectar a una o varias tramas consecutivas, completamente o sólo en parte. El caso de IP es algo distinto ya que, en general, no se suelen producir errores de bit a ráfagas y muy raramente errores aislados, debido a la alta fiabilidad del canal, sino que lo más común es que se produzcan pérdidas de paquetes (a ráfagas) en los nodos de enrutamiento.

En cualquier caso, lo que se pone de manifiesto en esta tesis es que el hecho de que este tipo de errores se produzcan sobre la voz codificada tiene consecuencias que no se pueden tratar de la misma manera que si se produjeran sobre la señal de voz original

(por ejemplo, modelando los errores haciendo las hipótesis habituales de ruido convolutivo o aditivo). Es decir, si tenemos en cuenta que el proceso de codificación de la voz consiste, a grandes rasgos, en la extracción de una serie de parámetros que representan distintos aspectos específicos de este tipo de señal (su periodo fundamental, la posición de sus formantes, su característica sonora o sorda, su energía, etc), nos percataremos de que la modificación de cada uno de ellos tiene consecuencias muy distintas sobre la señal vocal reconstruida.

Un reconocedor convencional que recibe una señal de voz codificada, la primera acción que realiza sobre ella es su decodificación y de esa forma, ya puede proceder a realizar la extracción de características o parametrización para reconocimiento. En este proceso, las distorsiones de codificación y de los errores se trasladan a los parámetros a partir de los que se realizará el reconocimiento, produciendo el deterioro de las prestaciones del reconocedor.

Para mejorar esta situación, en esta tesis proponemos el análisis de la parametrización de la señal de voz que lleva a cabo el codificador antes de su decodificación y la transformación de ésta en otra adecuada para el reconocimiento. Esto además, nos permite utilizar métodos de recuperación frente a errores y de transformación de parametrizaciones orientados directamente al reconocimiento, sin limitarnos a los ya previstos en los estándares de codificación, cuyo propósito es recuperar una señal de voz perceptualmente aceptable sujetos a una fuerte restricción de tiempo real.

De esta forma, obtenemos una solución aplicable a ambos entornos (GSM e IP) que reduce la influencia sobre los reconocedores, por una parte, la distorsión de codificación haciendo una selección de la información relevante para reconocimiento, y por otra, el efecto de los errores de transmisión, actuando directamente sobre los parámetros afectados. Resulta notable el hecho de que esta solución sea aplicable tanto a entornos móviles como a redes de tipo IP, ya que puede emplearse cuando existe una combinación de ambas como parece ser la tendencia.

Así, el Capítulo II está dedicado a presentar el problema del reconocimiento a partir de voz codificada y a justificar su interés. Para ello introducimos los elementos fundamentales que intervienen en el mismo, a saber: la codificación y el reconocimiento. Establecemos así un marco que permite poner de manifiesto los paralelismos entre ambos mundos y encontrar las similitudes y diferencias entre la parametrización que llevan a cabo cada uno de ellos. Finalmente, evaluamos otras soluciones alternativas para implementar servicios a través de redes de comunicaciones utilizando reconocimiento de habla que no implican la codificación de la misma y exponemos los motivos que nos han llevado a elegir, a pesar de sus inconvenientes, la que incluye dicha codificación.

El Capítulo III profundiza en las dificultades con las que nos enfrentamos a la hora de realizar este reconocimiento en los entornos GSM e IP y las soluciones que proponen otros autores. Estas dificultades, como ya hemos adelantado, proceden de la distorsión producida por la codificación y los errores de transmisión en estos canales.

Así comenzamos presentando la estructura de los codificadores más frecuentemente utilizados en dichos entornos, incluyendo también los codificadores de tasa múltiple adaptativa o AMR (*–Adaptive Multi-Rate–*), que previsiblemente sustituirán a los actualmente utilizados en los sistemas móviles y que además también se contemplan

dentro de los protocolos de transporte en tiempo real (RTP *–Real-time Transport Protocol–*) de las redes IP. Destaca en esta exposición el hecho de que todos ellos se basan en análisis mediante síntesis (AbS *–Analysis by synthesis–*), modelando, en primer lugar, la envolvente espectral y posteriormente escogiendo una excitación que da lugar a una señal de voz lo más próxima posible a la original.

Después analizamos, los efectos que este tipo de codificación produce sobre reconocedores automáticos de habla y los resultados obtenidos por otros autores en este ámbito.

Seguidamente, exponemos los efectos de los errores de transmisión, haciendo hincapié en las características especiales que le confiere el hecho de se produzcan sobre la voz codificada. Así, las soluciones que existen en la literatura se pueden dividir en dos grandes bloques: las que están específicamente diseñadas para reconocimiento y las que no lo están. Entre las segundas podemos además distinguir por un lado las orientadas a evitar que estos errores alcancen el extremo final (codificación de canal, en el caso de GSM, o nuevos protocolos de transporte en tiempo real, en IP), y por otro, las que, una vez que los errores están ahí, pretenden recuperar una señal perceptualmente aceptable.

En el Capítulo IV retomamos el tema de las representaciones de la señal de voz manejadas en los ámbitos de codificación y reconocimiento automático. Analizamos, entonces, los factores que determinan o que influyen en estas dos representaciones para acabar presentando transformaciones de una en la otra, y viceversa. En concreto, desarrollamos el procedimiento de reconocimiento mediante transparametrización, objeto de esta tesis.

El Capítulo V comienza con la descripción del sistema de referencia utilizado en esta tesis para la verificación de las ventajas de nuestra propuesta, que incluye las bases de datos utilizadas y los métodos empleados para la simulación de los canales GSM e IP. Además, planteamos una serie de escenarios que presumiblemente pueden tener lugar en los que se incluyen situaciones tales como transcodificaciones dentro del sistema GSM o interacciones entre GSM e IP, y presentamos los resultados de reconocimiento obtenidos en estos casos.

Como es habitual, cierra la tesis un capítulo de conclusiones y líneas futuras de trabajo.

Capítulo II
Reconocimiento a partir de voz codificada

En este capítulo se describe el contexto en el que surge la necesidad de estudiar el problema del reconocimiento de voz codificada y se presentan los elementos fundamentales que intervienen en el mismo. Una vez hecho esto, estaremos en condiciones de justificar el punto de vista desde el cual se aborda este problema en esta tesis y que consiste en la realización del proceso de reconocimiento (en un servidor remoto) a partir de voz codificada procedente del canal GSM o IP, sin realizar la decodificación de la misma; es decir, mediante la transformación de la parametrización utilizada por los codificadores correspondientes en la parametrización utilizada habitualmente por los reconocedores de habla.

Así comenzamos introduciendo las redes de transmisión de voz, señalando los compromisos que deben adoptarse para lograr esa transmisión (relación entre calidad y ancho de banda utilizados, nivel de protección frente a errores, disponibilidad de recursos computacionales, etc.), su influencia sobre la forma en que la señal de voz es transmitida y las características que le imprime una red particular. Después haremos hincapié en las aplicaciones y servicios posibilitados por el empleo de tecnologías del habla tales como reconocimiento o síntesis. Posteriormente nos centraremos en reconocimiento, objeto de la tesis, comenzando por revisar sus fundamentos. Por último expondremos las alternativas de implementación que permiten configurar este tipo de servicios describiendo sus ventajas e inconvenientes y justificando la elección de una de ellas como tema central de esta tesis.

II.1. Transmisión de voz. Evolución.

Hasta hace pocos años, las grandes redes de comunicaciones se han ocupando principalmente de la transmisión de voz, siendo minoritario el volumen de tráfico de datos. Con este propósito se desplegaron, enormes infraestructuras que transportaban señales de voz analógicas. Un gran hito en la historia de las telecomunicaciones consistió en la introducción de sistemas digitales en estas redes; a partir de ese momento, se fueron adaptando progresivamente las redes analógicas existentes para transportar la voz digitalizada.

De forma similar, las primeras redes de comunicaciones móviles fueron inicialmente analógicas; sin embargo, el gran éxito cosechado por ellas en los últimos años, sobre todo en Europa, ha venido de la mano del sistema digital GSM (*–Global System for Mobile–*). Una de las grandes ventajas del sistema GSM es que ha sido implementado de forma compatible por un gran número de compañías operadoras, lo que permite que su cobertura territorial sea muy elevada. Sin embargo, una limitación inherente de estos sistemas es la escasez de ancho de banda disponible que exige que la señal de voz sea comprimida sustancialmente antes de ser transmitida. La tercera generación de móviles UMTS, por su parte, intenta mejorar el aprovechamiento del espectro para proporcionar la capacidad de transmitir señales multimedia.

Paralelamente a las redes de transmisión de voz, se fueron desarrollando redes para la transmisión de datos que se caracterizaban por utilizar tecnología de conmutación de paquetes en lugar de conmutación de circuitos, característica de las redes de transmisión de voz. Estas redes, sin embargo, tuvieron una importancia marginal hasta que apareció en escena Internet. Esta red, basada en el protocolo TCP/IP ha acabado imponiéndose virtualmente en todo el mundo y su gran éxito ha propiciado que el tráfico de voz, antes dominante, vaya perdiendo su preponderancia y que, en consecuencia, deje de estar justificada la existencia de una red específica para este tipo de transmisión. No por ello, sin embargo, deja de tener dificultades, desde el punto de vista tecnológico, la transmisión de voz sobre redes de conmutación de paquetes, cuando se pretende proporcionar la misma calidad de servicio que proporcionan las de conmutación de circuitos: la naturaleza inelástica de las transmisiones de voz no estaba prevista en el desarrollo de las redes TCP/IP que fueron concebidas para transportar tráfico elástico, es decir, tráfico en el que la integridad de los datos prima sobre el tiempo de su entrega. Uno de los factores limitadores en los sistemas de VoIP (VoIP *–Voice over IP–)*, en estos momentos es, de nuevo, el ancho de banda, imponiéndose una vez más una codificación agresiva que reduzca significativamente estos requerimientos.

De esta forma, se puede observar cómo tanto las redes móviles como las fijas, a pesar de estar concebidas inicialmente para transmitir un único tipo de información, evolucionan para abarcar la transmisión de todo tipo de medios (audio, texto, video, etc.) posibilitando así, el desarrollo de aplicaciones y servicios basados en combinaciones de ellos. Pero además, la combinación de las posibilidades que ofrecen las redes móviles con las de las fijas amplía considerablemente los horizontes de creación de nuevas aplicaciones. No hay que olvidar, sin embargo, que para que todo esto sea viable es necesario que se estudien las consecuencias que estas nuevas formas de transmisión tienen sobre la calidad percibida en el extremo receptor. En concreto, en esta tesis evaluamos el impacto que las distorsiones introducidas por las redes GSM o IP tienen sobre los reconocedores de habla y proponemos soluciones que alivian la degradación de su comportamiento. Efectivamente, estas distorsiones también tienen influencia sobre reconocedores humanos, pero éstos son capaces de tolerar más fácilmente algunos niveles.

II.1.1. Codificación de voz.

La codificación es una tecnología fundamental para la transmisión de voz en los entornos GSM e IP y surge de la necesidad de incrementar la capacidad de tales redes de comunicaciones, aumentando el número de canales de voz, o lo que es lo mismo, disminuyendo el ancho de banda requerido por cada uno de ellos. El objetivo de la codificación es, por lo tanto, encontrar una parametrización de la voz optimizada para su transmisión en un canal determinado (o tal vez para su almacenamiento en un sistema concreto) de forma que la reconstrucción posterior de la misma sea perceptualmente aceptable para un receptor humano.

Así, si partimos de una señal de voz digitalizada, $s[n]$, con $n=1,\cdots,N$, el proceso de codificación consiste en encontrar una secuencia de vectores de parámetros $\mathbf{F}=\mathbf{f}_1,\mathbf{f}_2,\cdots,\mathbf{f}_R$ de tal forma que el error (perceptualmente ponderado) entre la señal original y la reconstruida a partir de los parámetros óptimos $\hat{\mathbf{F}}$, $s_{\hat{\mathbf{F}}}[n]$, sea el mínimo posible, esto es:

$$\hat{\mathbf{F}}=\arg\min_{F}\left\{\sum_{n=1}^{N}\left(p_{\mathbf{F}}[n]*e_{\mathbf{F}}[n]\right)^2\right\} \qquad \textbf{(II-1)}$$

donde $p_{\mathrm{F}}[n]$ es el filtro de ponderación perceptual que generalmente también depende de los vectores de parámetros y

$$e_{\mathbf{F}}[n]=s[n]-s_{\mathbf{F}}[n] \qquad \textbf{(II-2)}$$

es el error cometido en cada muestra al reconstruir la señal a partir de la secuencia $\mathbf{F}$.

Evidentemente, la obtención de la secuencia de parámetros que mejor representan la señal original no es sencillo y está sujeto a una serie de restricciones que dependen principalmente del entorno (en nuestro caso GSM o IP) en el que se utiliza un determinado codificador, haciendo que alguno sea más adecuado que otro dependiendo básicamente (aunque no sólo) de cuatro características:

- Régimen binario, que limita la dimensión de los vectores $\mathbf{f}_i$, el número de bits con los que se cuantifica cada uno de los elementos de dichos vectores y la frecuencia con la que se extraen.
- Complejidad de los algoritmos empleados para extraerlos.
- Retardo algorítmico, que también limita, esta vez en el sentido contrario la frecuencia con la que se calculan.
- Calidad, medida habitualemente utilizando tests subjetivos tales como MOS (–Mean Opinion Score–), DRT (–*Diagnostic Rhyme Test*–) o DAM (–*Diagnostic Acceptability Measure*–).

En concreto, en los sistemas GSM se emplean tres codificadores: el GSM-FR (–*Full Rate–),* el GSM-HR (–*Half Rate*–) y el GSM-EFR (–*Enhanced Full Rate*–). Todos ellos proporcionan un retardo algorítmico similar; sin embargo, el primero y más antiguo y su versión mejorada hacen uso del doble de ancho de banda que el segundo. Además el

EFR proporciona una calidad bastante superior a la del FR a costa de aumentar su complejidad. La próxima generación de codificadores, que también será utilizada en UMTS es la denominada AMR (*–Adaptive Multi-Rate–*); estos codificadores tienen en cuenta además la variabilidad temporal de las características del canal permitiendo una adaptación del régimen binario de una trama a otra en función de su disponibilidad.

Por su parte, en IP, los codificadores más empleados son el G.723.1 y el G.729, aunque también se utiliza en ocasiones el G.726 y los anteriormente citados del sistema GSM. La característica más sobresaliente del primero es su bajo régimen binario, que se proporciona a cambio de un retardo algorítmico, muy superior a los de el resto de los codificadores. El G.729, por el contrario, proporciona un retardo algorítmico mucho menor y mayor calidad empleando un régimen binario bastante superior.

A continuación vamos a revisar la evolución de los algoritmos que han dado lugar a las diferentes variantes en codificación que hemos expuesto.

El primer paso en el proceso de codificación consiste en la cuantificación de las muestras de la señal discreta resultante del proceso de digitalización (codificación PCM *–Pulse Coded Modulation–*). El diseño de un cuantificador adecuado a las características estadísticas de la señal de voz (utilizando la ley A o la ley μ) fue la primera mejora introducida. Debido a que estas características varían con el tiempo, el siguiente paso fue la adaptación temporal de estos cuantificadores. Esta adaptación se puede hacer tanto "hacia delante" ("*forward*") como "hacia atrás" ("*backward*").

Posteriormente se introduce la codificación diferencial, cuyo objetivo es reducir la varianza de la señal, facilitando así su cuantificación (DPCM *–Differential PCM–* y ADPCM *–Adaptive DPCM–*).

Una generalización de esta idea consiste en el filtrado de orden superior de la señal de forma que ya no sólo implique las muestras anteriores, sino un número mayor de muestras previas (APC *–Adaptive Predictive Coding–*). Esto se puede interpretar en términos de modelado de la respuesta impulsional del tracto vocal (predicción corta) y de la periodicidad de la señal de voz (predicción larga), que se suele llevar a la práctica utilizando la técnica de codificación predictiva lineal (LPC *–Linear Predictive Coding–*). En definitiva, el efecto de la predicción es la reducción de la varianza de la señal y blanqueado del espectro del error, concentrando la información de la parte redundante de la señal en los coeficientes del predictor.

Por otra parte, el enmascaramiento del ruido de cuantificación a través del uso del conocimiento del sistema auditivo conduce al uso de filtros de ponderación perceptual que se aplican sobre el error en sistemas de análisis mediante síntesis (AbS *–Analysis by Synthesis–*). La forma de aplicar estos filtros de ponderación suele ser la siguiente: cada trama de la señal es analizada primero para extraer el filtro de síntesis LPC; a partir de este filtro se calcula el filtro de ponderación perceptual correspondiente, que se aplica a la señal original para obtener la señal objeto de la comparación. Se trata, entonces, de encontrar la excitación que consiga un mínimo error cuadrático medio (MSE *–Mean Square Error–*) al atravesar el filtro de síntesis LPC, en cascada con el perceptual.

Los codificadores AbS son esencialmente codificadores de forma de onda, pues tratan de aproximar lo más fielmente posible la señal original (ponderada). Los vocoders, en cambio, generan una excitación en función de la clase a la que pertenece la

trama, que originalmente estaba limitada a la opción sonora o sorda, y más tarde se fue ampliando con mezclas de ambos tipos de excitación (sonora y sorda) (MELP –*Mixed Excitation Linear Prediction*–).

Los codificadores que se utilizan en GSM e IP emplean este modelo de producción de la voz en el que hay una fuente o excitación representando los pulmones y la vibración de las cuerdas vocales, en su caso, y un filtro que representa el tracto vocal, la cavidad bucal, la nasal y la radiación de los labios, básicamente, y que conforma la envolvente espectral de la señal de voz. Esta característica, común a todos ellos, es fundamental a la hora de transformar los parámetros de codificación en parámetros de reconocimiento, ya que es precisamente la información contenida en la parametrización del filtro la que emplean, casi exclusivamente, los reconocedores automáticos de habla.

II.1.2. Protección frente a errores de transmisión

La cantidad de errores que se producen en el canal radio de un sistema de comunicaciones móviles es muy superior a la que se produce en un enlace fijo. Por este motivo, el sistema GSM tiene previsto un mecanismo de protección de los datos transmitidos, en el que se invierte una porción considerable del ancho de banda total disponible para la transmisión.

En las redes IP, como ya hemos señalado, el problema más grave no son los errores de transmisión que se producen en los enlaces, sino la pérdida de paquetes en los nodos de conmutación. Además, en nuestro caso, en el que lo que se transmite es un flujo ("*stream*") de voz en tiempo real, existe otra fuente de error que también contribuye a aumentar el número de paquetes perdidos y es que aquellos paquetes que lleguen después del instante de tiempo en el que debieron ser reproducidos deben considerarse también inservibles.

Sin embargo, dada la relativa novedad de la transmisión de voz sobre el protocolo IP, este problema ha sido tratado muy recientemente y no existe todavía consenso sobre qué protocolo emplear de entre los que se han propuesto para solucionarlo o al menos aliviarlo.

II.1.2.1. Codificación de canal en GSM

El codificador de canal implementado en los sistemas GSM se describe en la recomendación GSM 5.03 del Instituto de Estándares de Telecomunicación Europeo ETSI (–*European Telecommunications Standards Institute*–) [45]. El codificador de canal toma como entrada los datos provenientes del codificador de voz, del terminal de datos, del controlador de la MS (–*Mobile Station*–) o del controlador de la BTS (–*Base*

Transceiver Station–), realiza la codificación, el reordenamiento de bits, el entrelazado y añade el indicador de *stealing*[1].

La especificación para cada canal lógico es distinta y más adelante explicaremos con más detalle la que corresponde a los canales de voz (sección III.3.1.1). La secuencia de operaciones común a todos los canales es la siguiente:

- Los bits de información se codifican con un código bloque, de forma que se obtienen palabras de información + bits de paridad.
- Esta información + bits de paridad se codifica con un código convolucional que da lugar a los bits codificados.
- Finalmente se realiza el reordenamiento de bits, el entrelazado y se añade el indicador de stealing conformando lo que se denomina "bits entrelazados".

Como veremos este tipo de codificación permite proteger mejor ciertos bits (más sensibles) que otros en el caso de los canales vocales. En concreto, parte de los bits con más protección son las que se dedican a codificar la información de la envolvente espectral, mientras que los bits que representan la excitación, quedan más vulnerables. Este es un hecho muy importante, si tenemos en cuenta que para el reconocimiento automático los parámetros más relevantes son precisamente los primeros: la reconstrucción de una señal con el propósito de realizar el reconocimiento resulta desaconsejable ya que ésta se verá afectada por los errores, más verosímiles en la excitación, dando lugar a una voz decodificada distorsionada innecesariamente (desde el punto de vista de reconocimiento) en un mayor número de tramas.

II.1.2.2. Soluciones de red en IP

El problema de la pérdida de paquetes en las redes IP se debe principalmente a que el tipo de tráfico para el cual estaban diseñadas estas redes (el tráfico de datos) no requería que los paquetes en los que se trocea la información llegaran a su destino en un intervalo de tiempo determinado. Así, la pérdida de paquetes en los nodos de conmutación se resuelve pidiendo la retransmisión de los mismos y de ello se encarga el protocolo TCP. Este procedimiento es totalmente inútil en el caso del tráfico vocal, donde un paquete se considera perdido, no solamente si no llega, sino también si llega demasiado tarde para ser reproducido.

Para solucionar este problema se están introduciendo nuevos protocolos capaces de manejar estos flujos de tráfico inelástico, proporcionando una calidad de servicio similar a la que proporcionan en estos momentos las redes de telefonía convencional. El problema de este tipo de solución es la lentitud con la que se producen los cambios y la introducción de nuevos protocolos debido a las enormes magnitudes y la gran

[1] El indicador de *stealing* señala las tramas que no transmiten información de voz o datos y que se dedican a funciones de señalización.

heterogeneidad de sistemas que forman parte de Internet. Más adelante (sección III.3.2) se discutirán estos protocolos con más profundidad.

II.1.3. Mecanismos de recuperación frente a errores

Los mecanismos previstos para evitar errores de transmisión, tanto en GSM como en IP, no son lo suficientemente efectivos y la realidad es que se producen errores a ráfagas en el flujo de bits que se entrega al codificador de voz GSM y pérdida de paquetes en IP. Por ello son necesarios mecanismos que permitan recuperar o reparar las distorsiones debidas a estos errores. Estos métodos pueden bien actuar directamente sobre la forma de onda de la señal de voz, o bien sobre los parámetros del codificador antes de su reconstrucción. Estos últimos son, por tanto, dependientes del tipo de codificador que se emplea y más adelante (secciones III.3.1.2 y III.3.2.2) haremos una revisión de tales métodos valorando su efectividad tanto desde el punto de vista de recuperación de una señal perceptualmente aceptable como desde el del reconocimiento automático.

II.2. Tele-servicios con interfaces basados en tecnologías del habla

A medida que la capacidad de transmisión de las redes de comunicaciones aumenta permitiendo que todo tipo de medios lleguen a los usuarios finales independientemente del lugar en que se encuentran surge la necesidad de nuevas aplicaciones que sean capaces de sacar partido a la posibilidad de acceder a información multimedia desde terminales móviles. Cualquiera que sea la nueva aplicación o servicio que se desarrolle basándose en estas premisas, el uso de una interfaz vocal contribuye a crear un acceso más natural. Para ello es necesario crear sistemas de diálogo basados en tecnologías de reconocimiento, síntesis de habla e identificación y verificación de locutor. Así, es posible pensar en:

- Aplicaciones de acceso vocal a correo electrónico, agendas personales u otro tipo de información personal (bancaria, médica, fiscal, etc.).

- Aplicaciones de acceso a información de tipo genérico (guías de restaurantes, gasolineras, alojamiento, bursátil, tráfico, tiempo atmosférico, etc.) que hagan uso de tecnologías de localización geográfica.

- Se puede habilitar el acceso vocal a catálogos de productos vía web o sistemas de reserva o venta de entradas para espectáculos, billetes de avión, tren, autobús, etc.

- También es posible utilizar sistemas de dictado de forma remota, navegación web o navegación en aplicaciones mediante comandos vocales, lo que resulta especialmente útil cuando el tamaño de los terminales de acceso limita (o

hace incómodo) el uso de teclados, o en situaciones en las que se tienen las manos o la vista ocupados.

II.3. Reconocimiento automático de habla

Los sistemas de reconocimiento de habla actuales están basados en el reconocimiento estadístico de patrones y el problema fundamental se plantea de la siguiente manera: tomamos una secuencia de vectores acústicos $\mathbf{Y} = \mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T$ que han sido extraídos a partir de una señal de voz $s[n]$ mediante un proceso que denominaremos "parametrización", y deseamos averiguar qué secuencia de palabras o símbolos, $W = w_1, w_2, \cdots, w_M$, ha sido emitida. Por lo tanto, queremos obtener la secuencia de palabras $\hat{W}$ que más verosimilmente corresponde con la observación acústica $\mathbf{Y}$, es decir,

$$\hat{W} = \arg\max_W P(W|\mathbf{Y}) = \arg\max_W \frac{P(W)P(\mathbf{Y}|W)}{P(\mathbf{Y})} \qquad \textbf{(II-3)}$$

donde en la segunda parte de la ecuación se ha utilizado la regla de Bayes con el propósito de descomponer este problema en otros más sencillos. De esta manera el problema de encontrar la secuencia W que maximiza la probabilidad a posteriori $P(W|\mathbf{Y})$ equivale a encontrar la que maximiza el producto $P(W) \cdot P(\mathbf{Y}|W)$. El primero de estos términos representa la probabilidad a priori de observar la secuencia *W*, mientras que el segundo expresa la probabilidad de observar la secuencia de vectores acústicos $\mathbf{Y}$ dada una secuencia específica de palabras o símbolos *W.*

Los subproblemas en los que queda descompuesto el problema fundamental de reconocimiento son entonces cuatro (por ejemplo, [35] o [100]):

- Parametrización o generación de los vectores acústicos $\mathbf{Y}$ a partir de la señal de voz $s[n]$: es necesario encontrar una descripción compacta de la señal de voz que contenga todas las características necesarias para reconocimiento y que sea compatible con los modelos acústicos que se empleen. Adelantamos aquí que la parametrización más comúnmente utilizada es la MFCC (–*Mel Frequency Cepstral Coefficients*–). En IV.3 se exponen todos los detalles de la misma.

- Modelado acústico: el problema del modelado acústico es diseñar un método para calcular la verosimilitud de cualquier vector acústico de entrada, $\mathbf{Y}$, dada una secuencia de palabras, *W*, es decir, $P(\mathbf{Y}|W)$. Además debe ser capaz de proporcionar unos modelos de distribuciones que permitan discriminar entre distintas señales a la vez que optimizan el número de parámetros necesarios en función de la disponibilidad de datos de entrenamiento. Los modelos acústicos más empleados son los denominados modelos ocultos de Markov (HMM –*Hidden Markov Models*–) y son los que vamos a emplear en esta tesis, pero también existen soluciones basadas en redes neuronales o DTW (–*Dynamic Time Warping*–).

- Modelado del lenguaje: el objetivo del modelado del lenguaje es encontrar las probabilidades a priori $P(W)$ para cada posible secuencia W. Para ello podemos descomponer esta probabilidad de la siguiente manera:

$$P(W) = \prod_{i=1}^{n} P(w_i | w_1, \ldots, w_{i-1}) \qquad \textbf{(II-4)}$$

de forma que el problema se transforma en la estimación de las probabilidades de cada uno de los elementos de la secuencia w_i dados sus predecesores. Sin embargo, para el cálculo de estas probabilidades, incluso en el caso de tareas con vocabulario moderado, el número de argumentos de los que dependen cada una de ellas es excesivo, de forma que se utilizan simplificaciones tales como las de las gramáticas de pares de palabras que utilizaremos en los experimentos de esta tesis y otras con algo más de profuncidad.

- Decodificación o búsqueda de la hipótesis más verosimil: una vez obtenidos todos los factores que intervienen en la ecuación (II-3) es necesario encontrar la secuencia óptima $\hat{W}$ que maximiza dicha ecuación. Uno de los algoritmos más populares es el de Viterbi que utilizaremos en nuestros experimentos.

El reconocimiento estadístico de patrones está basado en la existencia de suficientes patrones de entrenamiento para la generación de modelos fiables. Por este motivo, es muy importante que estos patrones de entrenamiento estén tomados en las mismas condiciones que los de test o de operación; sin embargo, esto no suele ser lo habitual. El reconocimiento robusto de habla tiene como objetivo contrarrestar la influencia de las posibles fuentes de distorsión; en IV.3.5 se exponen brevemente las distintas técnicas que se utilizan.

II.4. Reconocimiento de habla: implementaciones alternativas

Tanto si se trata de GSM como de IP, el problema se puede enfocar desde tres puntos de vista, en función de cómo se haga la distribución de los procesos que componen la tarea de reconocimiento [38]. Así, si la tarea de reconocimiento se lleva a cabo íntegramente en el terminal local o cliente tendremos lo que denominaremos reconocimiento local; si por el contrario, la tarea completa se desarrolla en el terminal remoto o servidor, lo denominaremos reconocimiento remoto; y finalmente, si parte de la tarea (normalmente la parametrización de la señal de voz) se realiza en el primero pero el resto tiene lugar en el segundo, lo llamaremos reconocimiento distribuido. En esta tesis nos hemos centrado en una variante del reconocimiento remoto y expondremos las razones de tal decisión en II.5. En todo caso es importante destacar que las tres opciones que presentamos a continuación responden a tres enfoques distintos, que resultan ventajosos respecto al resto en determinadas aplicaciones. Con esto queremos decir que aunque aquí hemos elegido el reconocimiento remoto y esta tesis está encaminada a encontrar soluciones que mejoren el comportamiento del

reconocedor bajo estas circunstancias, cualquiera de las otras dos alternativas puede ser preferible bajo otros supuestos.

II.4.1. Reconocimiento local

En el caso de reconocimiento local, todo el proceso de reconocimiento tiene lugar en el propio terminal, que puede ser un móvil, en el caso de GSM, o cualquier otro terminal conectado a Internet con capacidad de capturar voz, en el caso de IP.

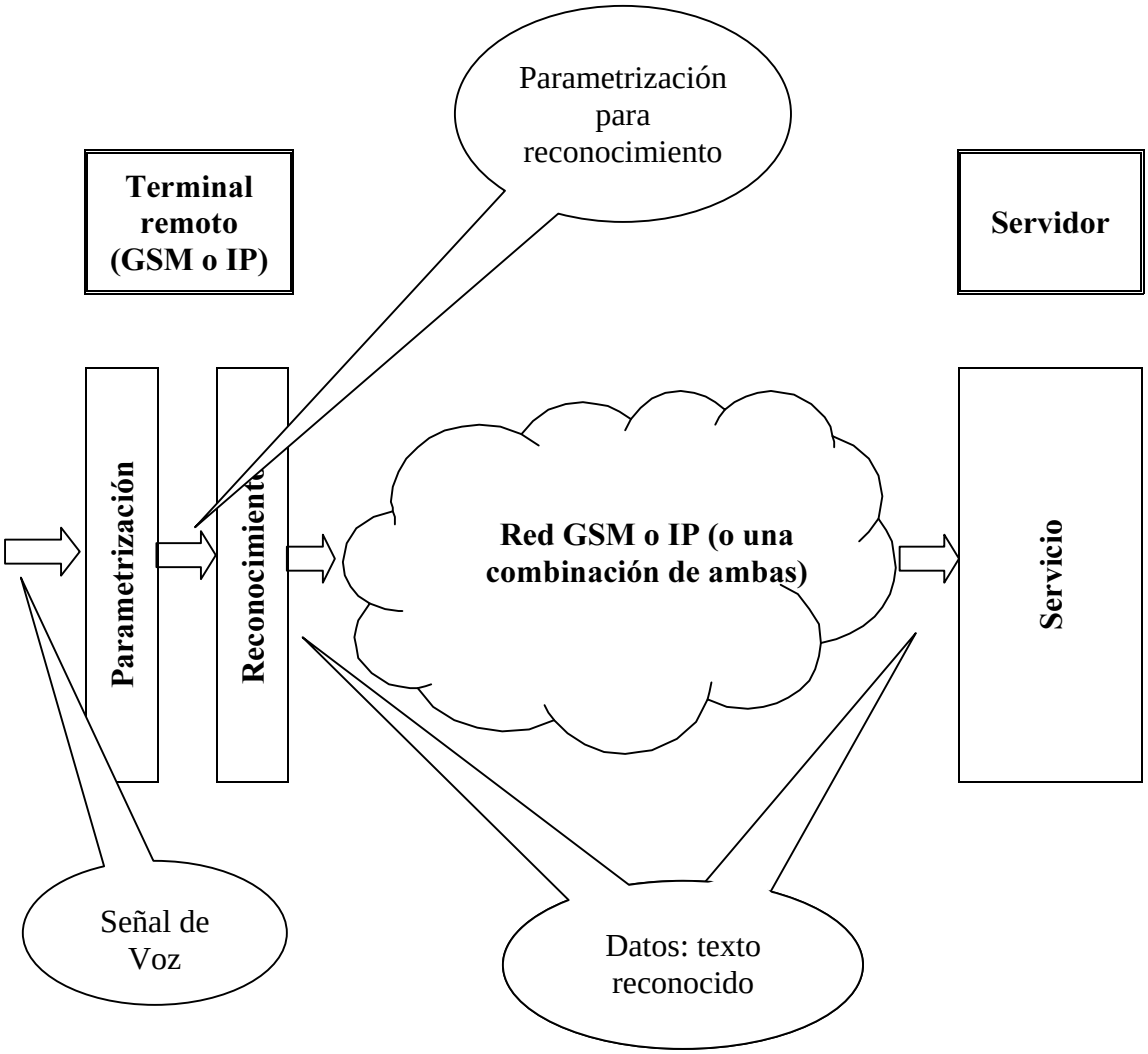


Figura II-1. Configuración de reconocimiento Local.

En la Figura II-1 se puede observar cómo el terminal comienza extrayendo los parámetros adecuados para realizar el reconocimiento, que posteriormente lleva a cabo él mismo. Si, por ejemplo, pretendemos navegar por el menú del móvil no es necesario transmitir ningún tipo de información, pero si, por el contrario, pretendemos acceder

algún tipo de servicio que no resida en el terminal, lo que se envía a través de la red GSM o IP es un flujo de datos en los que se ha codificado la información que se ha extraído a través de este reconocimiento automático y que tiene que ser la adecuada para el servicio concreto al que se acceda en el otro extremo.

La ventaja de esta aproximación es doble: por un lado la voz no necesita ser codificada para adaptarse a un ancho de banda determinado (tan solo la limitación que el sistema de captura de voz imponga); y por otro, el reconocedor se puede integrar de forma efectiva en la interfaz del usuario. La contrapartida es que esta aproximación supone una fuerte demanda de capacidad computacional por parte de los terminales y además, exige que resida en el terminal la aplicación de reconocimiento específica para acceder a un determinado servicio. La variedad de terminales susceptibles de acceder a estos servicios actualmente es muy variada y tal exigencia dificulta el acceso a los mismos. Además, estando en un servidor remoto puede ser más fácilmente mantenido y actualizado.

Efectivamente, sin embargo, ciertas aplicaciones tales como marcación por voz (dependiente del locutor) o reconocimiento de dígitos en móviles sí que son susceptibles de implementarse con esta configuración, obteniendo las ventajas correspondientes [76] pero otras con demandas superiores (de procesadores de alto rendimiento o de memoria) como sistemas de diálogo o dictado, requieren que la aplicación de reconocimiento resida en un servidor remoto.

II.4.2. Reconocimiento distribuido

Cuando el proceso de reconocimiento se hace de forma distribuida, la extracción de características o parametrización, que constituye sólo una pequeña parte del coste computacional del sistema de reconocimiento, se realiza en el terminal cliente, mientras que el resto se realiza en el servidor remoto.

De esta forma se reduce considerablemente el ancho de banda requerido para la transmisión, sin exigir una gran capacidad de computación al terminal. Como se puede observar en la Figura II-2, en este caso lo que viaja por el canal de comunicación es la parametrización adecuada para reconocimiento convenientemente cuantificada y codificada. Así, en el extremo servidor tan sólo tiene que realizar el reconocimiento propiamente dicho.

Sin embargo, cómo cabía esperar, esta alternativa no está exenta de inconvenientes. En primer lugar, es necesario estandarizar el "*front-end*" o "cabecera" del reconocedor, es decir, el reconocedor debe estar preparado para recibir exactamente el tipo de parámetros que el cliente le envíe, independientemente de que sean o no los más adecuados para esa aplicación en concreto. Esto no permite, por ejemplo, el acceso a reconocedores que utilicen una parametrización distinta a la establecida, como podrían ser los preparados para lenguas tonales en las que se utiliza (como parámetro) la frecuencia fundamental de la voz o "*pitch*", a no ser que se desarrolle un protocolo para negociar ese punto entre las dos entidades implicadas. En segundo lugar, la transmisión tiene que dedicarse únicamente al reconocimiento; consecuentemente, en principio, no es posible reconstruir posteriormente la señal de voz, lo que podría resultar útil en determinadas aplicaciones. Por último, es necesario también diseñar cuidadosamente la

cuantificación de la parametrización enviada para que, por una parte, el ancho de banda utilizado sea moderado, y por otra, los errores en el canal tengan una influencia mínima. En todo caso, el reconocimiento distribuido ha generado gran interés en los últimos años y ya se han propuesto soluciones para los problemas que comentamos, como veremos a continuación.

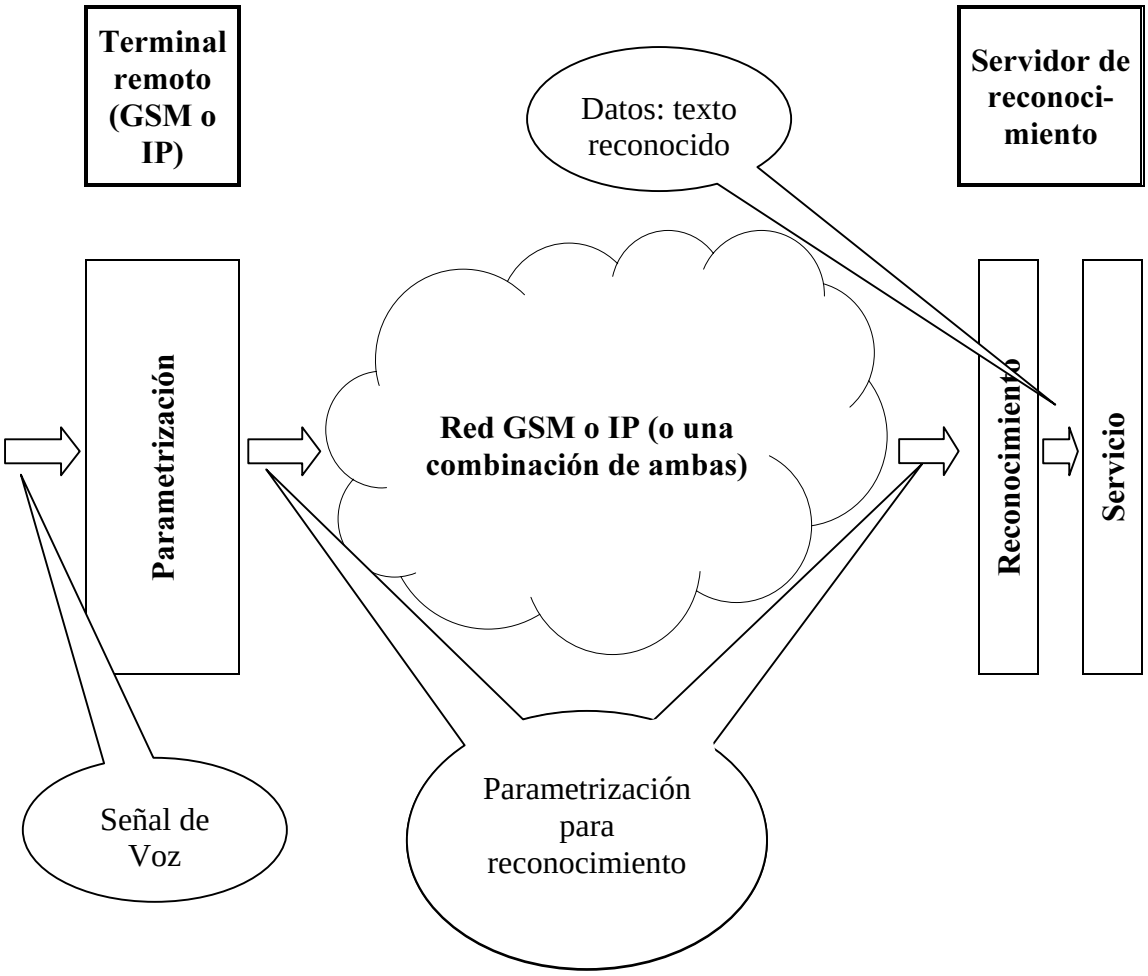


Figura II-2.- Configuración de reconocimiento distribuido.

Los primeros autores que plantearon esta aproximación fueron por una parte Digalakis et al. ([37], [38]) y por otra Ramaswamy et al. ([152]). Básicamente estos autores se concentraban en conseguir una cuantificación eficiente para transmitir los parámetros de reconocimiento. En todos estos casos, los parámetros que se codifican son los 13 primeros MFCC (véase apartado IV.2.1) obtenidos cada 10 ms. En [38] y [37] se ensayan varias estrategias, consiguiendo mejoras sobre el reconocimiento a partir de voz decodificada en lo que a la tasa de reconocimiento se refiere. Estas mejoras dependen del tipo de cuantificación de los parámetros que se escoja y del régimen

binario. En concreto para mejorar la tasa errores de reconocimiento que para el codificador GSM-FR es de 14,5% WER (tasa de errores de palabra –*Word Error Rate*–), hasta un 6,5% WER es necesario utilizar un régimen binario de 3,9 Kb/s, cuando se utiliza un cuantificador escalar no uniforme con un número constante de bits por coeficiente, de 2,8 Kb/s, cuando el número de bits por coeficiente es variable, y de 2 Kb/s, cuando utiliza cuantificación vectorial de código producto. Esta última cuantificación se basa en dividir el vector de coeficientes cepstrales en subvectores, los cuales serán cuantificados vectorialmente por separado. También se sugiere el uso de cuantificación vectorial predictiva, para explotar así la correlación temporal que existe entre los vectores cepstrales. Se advierte, como resulta obvio, que para reducir el ancho de banda requerido para la transmisión, la complejidad de los algoritmos utilizados aumenta, exigiendo mayor capacidad de cómputo en el terminal.

En [152], asimismo, se propone un algoritmo de codificación de los parámetros de reconocimiento que explota, precisamente, la relación temporal entre los parámetros cepstrales a la que aludíamos anteriormente, como primer paso para su codificación. Este paso consiste en una predicción lineal de primer orden, en donde el vector actual de parámetros se compara con el anterior previamente codificado. Se indica, además, que la predicción de orden superior no consigue la suficiente mejora si se considera el coste adicional que supone. El vector de error obtenido tras la predicción lineal se pasa, a continuación, por un cuantificador vectorial de varias etapas: en la primera se utiliza una librería de códigos compuesta de 4096 vectores de dimensión 13; en la segunda, el residuo de la etapa anterior se divide en dos vectores de 6 parámetros cada uno y un escalar correspondiente a la energía de la trama. Cada uno de los 2 primeros vectores de cuantifica con una librería de 4096 vectores de 6 dimensiones y el último se cuantifica con un cuantificador escalar de 16 entradas. Esto se hace así debido a que se considera que este último parámetro es más sensible y de esta forma se dota al sistema de mayor robustez. Con este procedimiento se logran resultados de reconocimiento similares a los que se obtienen extrayendo los parámetros de reconocimiento directamente de la voz original.

Por otra parte, desde 1997, un comité técnico de la ETSI, se ocupa de los aspectos de procesado de voz, transmisión y calidad (–*Speech Processing, transmission and quality aspects*–), y una de las tres áreas principales de las que se ocupa es precisamente el reconocimiento distribuido (bajo el nombre de Proyecto Aurora [41]).

En el marco del proyecto Aurora, que pretende compatibilizar los formatos de parámetros entre el terminal emisor y el reconocedor, se propuso, en Abril de 2000, una “cabecera” basada también en parámetros MFCC que obtiene buenos rendimientos en entornos poco ruidosos [43]. En este documento se describen:

- El algoritmo para la extracción del vector de parámetros, que está formado por 13 parámetros MFCC y 1 parámetro representando la energía para cada trama de 25 ms si se analizan señales de voz muestreada a 8 y 16 KHz o de 23,27 ms si se trata de señales a 11 KHz.

- El algoritmo para comprimir dichos parámetros de forma que se obtengan regímenes binarios de 4,8 y 9,6 Kb/s, basado en la cuantificación vectorial del vector de parámetros dividido en 7 subvectores, de forma que se utilizan 7 librerías diferentes.

- El formato para la transmisión del flujo de bits para su transmisión, que incluye protección frente a errores (codificación de canal),

- La decodificación del flujo de bits para generar de nuevo los vectores de parámetros así como algoritmos para mitigar los efectos de los errores procedentes del canal. Para detectar errores en las tramas se describen dos procedimientos: comprobación del código de redundancia cíclica (CRC –*Cyclic Redundancy Code*–) que envía el emisor y evaluación de la consistencia de los datos por medio de un procedimiento heurístico. También se prevé un método de sustitución de las tramas erróneas que consiste en la repetición de las tramas correctas anteriores y posteriores a la defectuosa.

El próximo hito previsto dentro del proyecto Aurora contempla la creación de algoritmos capaces de actuar en entornos adversos tales como coches, aeropuertos, etc. La publicación de los mismos está prevista para finales del año 2001. Otros desarrollos futuros en este grupo de trabajo incluyen protocolos cliente-servidor que contemplen entre otros, los requerimientos de las aplicaciones, multimodalidad, reconstrucción de la voz ([29]), etc. Por otra parte, también existen estudios para incluir medidas sobre el tono fundamental de la voz o *pitch* dentro de la cabecera ([164]).

Por otra parte, Tucker et al, proponen que el terminal cliente no sólo realice la parametrización previa al reconocimiento, sino que además calcule de las probabilidades a posteriori de los fonemas (Tucker propone trabajar con estas unidades) para cada vector acústico [172]. Esto es posible al utilizar un sistema de reconocimiento híbrido de redes neuronales recurrentes (RNN *–Recurrent Neural Network–)* y HMMs. En este sistema se modela la naturaleza temporalmente variante de la señal de voz con un proceso oculto de Markov, mientras que la estimación de las verosimilitudes de las observaciones se realiza mediante estas redes neuronales recurrentes, cuya ventaja es que actúan como modelos no paramétricos capaces de capturar el contexto acústico. Es por eso que en el caso del sistema utilizado por Tucker, no son necesarios más que 54 modelos de fonema independientes de contexto, lo cual permite (utilizando cuantificación vectorial de estas probabilidades a posteriori) obtener tasas tan bajas como 625 b/s. En todo caso, esto aplica sólo en aquellas situaciones que no requieran una reconstrucción posterior de la voz en el extremo receptor. Cuando esta reconstrucción sí que es necesaria, Tucker propone las siguientes modificaciones en los codificadores basados en el análisis LPC para mejorar el reconocimiento automático:

1. Aumentar la frecuencia a la que se realiza dicho análisis.

2. Mejorar la cuantificación de los parámetros (aunque en caso de necesitar un compromiso entre este punto y el anterior, es preferible el primero).

3. Utilizar un esquema de división en bandas (*–split-band–*) para extender el codificador a banda ancha, ya que para un codificador de muy baja tasa binaria (2,4 Kb/s), la inteligibilidad mejora más con esta modificación que con una cuantificación más fina.

4. Reentrenar el reconocedor sobre la voz codificada.

Una aproximación similar que contempla el problema de la reconstrucción de la señal de voz a partir de la parametrización de reconocimiento es la propuesta por Chazan et al. que denominan "Codificador de Voz compatible con Reconocimiento" (RECOVC *–Recognition COmpatible Voice Compression–)* ([19] y [20]). Según estos autores el interés de este tipo de recuperación se encuentra en aplicaciones tales como corrección de errores en dictado automático, en cuyo caso es posible incluso la reproducción más veloz de la señal de voz o aplicaciones en donde por motivos legales sea conveniente conservar una grabación de la conversación (transacciones comerciales, movimientos bancarios, etc.). El sistema de reconstrucción que proponen está basado en un modelo sinusoidal de la voz en donde las amplitudes, fases y frecuencias de las componentes sinusoidales de la voz sintetizada se determinan a partir de los parámetros MFCC y la frecuencia fundamental o "*pitch*". El sistema consigue regímenes binarios entre 4 y 6,9 Kb/s en función de la dimensión del vector de parámetros MFCC que emplea (13 ó 24, está última para proporcionar robustez en entornos ruidosos) y de la habilitación o deshabilitación de la opción de reconstrucción. Ampliaremos más este procedimiento en la sección IV.4.

II.4.3. Reconocimiento remoto

Finalmente, en la Figura II-3, podemos observar la implementación en la que nos hemos basado en esta tesis. En este caso el reconocimiento se realiza en el servidor remoto, permitiendo así que todo tipo de terminales cliente accedan al servicio, sin imponer restricciones de capacidad computacional ni configuración sobre los mismos.

Por supuesto, debido a que en este tipo de aplicaciones el ancho de banda del que dispone cada cliente suele ser bastante limitado, la voz se transmite codificada, empeorando el comportamiento del reconocedor. Sin embargo, como se puede observar la información que viaja por el canal es la parametrización generada por codificadores estándar, ya sean GSM (FR, HR o EFR) o IP (G.723.1, G.729, etc.). Esto quiere decir que el terminal no tiene por qué ser consciente ni realizar ningún tipo de procesado específico para acceder a un servicio que utilice reconocimiento de habla.

Una vez que se recibe la voz codificada en el servidor pueden seguirse dos estrategias para obtener los parámetros de reconocimiento: decodificar la voz y extraer los parámetros como si se tratara de voz original, o bien, obtener directamente estos parámetros a partir de los extraídos, cuantificados y transmitidos por el codificador. A la primera la denominaremos reconocimiento a partir de voz decodificada y a la segunda reconocimiento a partir de voz codificada. La demostración de las ventajas de esta última frente a la primera, una vez que se ha escogido realizar el reconocimiento de forma remota, constituye el objetivo de esta tesis. Por ello en los próximos apartados nos dedicaremos, respectivamente, a justificar la elección del reconocimiento remoto desde el punto de vista del diseño del sistema y a presentar una visión preliminar de las ventajas del reconocimiento a partir de voz codificada.

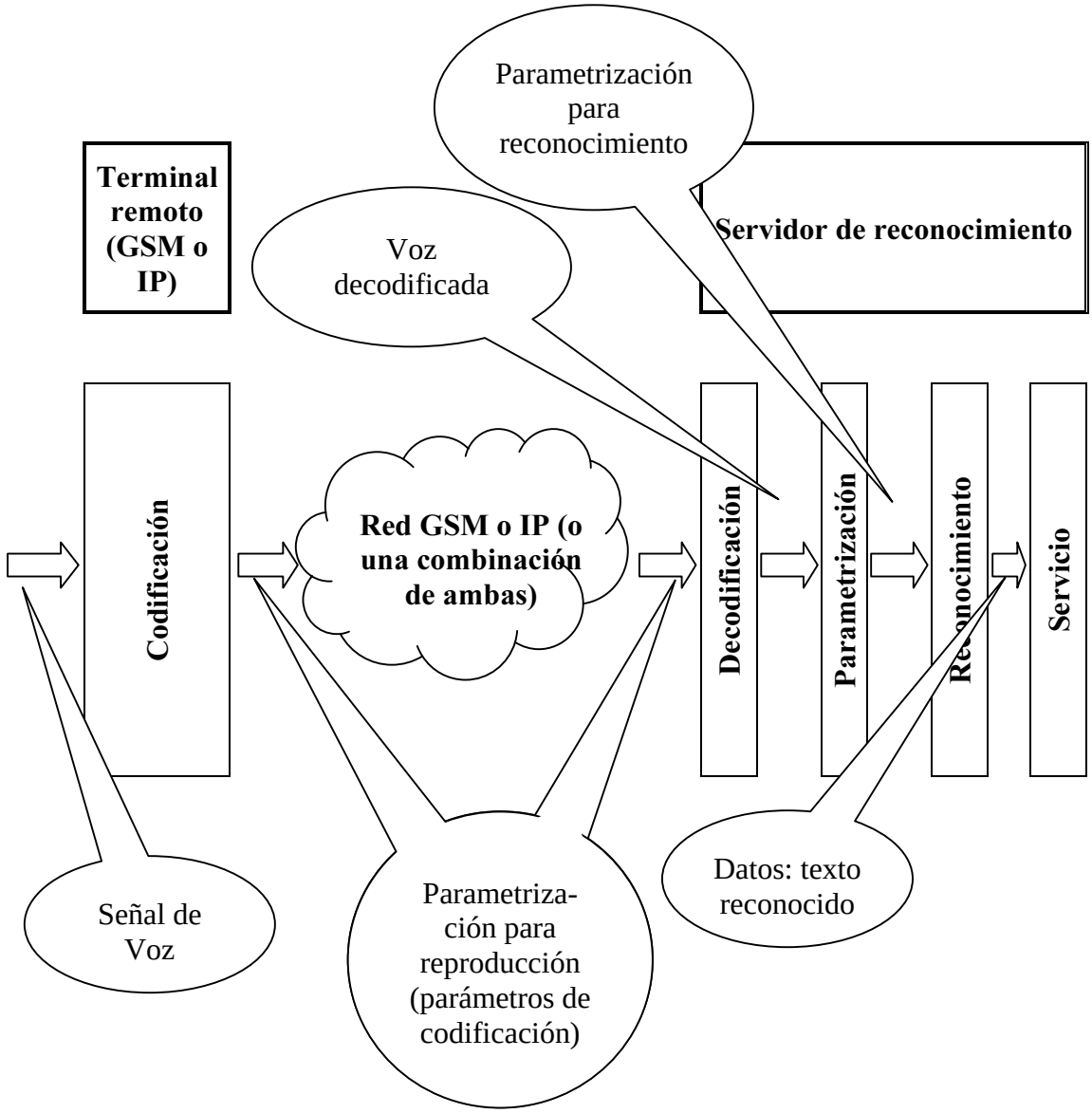


Figura II-3. Configuración de reconocimiento remoto.

II.5. Reconocimiento remoto: justificación

Las razones por las que se ha elegido esta aproximación, y no alguna otra de las alternativas presentadas en II.4, son las siguientes:

1. Al contrario de lo que ocurre en el reconocimiento local, el remoto tiene la ventaja de que no requiere que el terminal incorpore la capacidad de realizar el proceso de reconocimiento. Consecuentemente, no requiere que el terminal esté convenientemente actualizado para poder acceder a determinados servicios ni le exige una gran capacidad computacional. De esta forma es posible poner en marcha un servicio cualquiera que haga uso de la

tecnología de reconocimiento de habla sin establecer ningún requisito adicional sobre el terminal que accede a él (que tan solo tiene que incorporar la capacidad de transmitir voz utilizando un codificador estándar).

2. Respecto al reconocimiento distribuido, la aproximación propuesta tiene dos ventajas fundamentales: en primer lugar, no sólo se transmiten los parámetros necesarios para el reconocimiento, sino la señal de voz completa (aunque codificada), siendo, por tanto, más versátil. Muchas veces, por ejemplo, es necesaria la reconstrucción de la señal de voz en el receptor para otros propósitos, lo cual no es posible desde la aproximación distribuida, a no ser que se envíen parámetros adicionales [29]. Además, en este caso estos parámetros adicionales deberán ser robustos frente a las degradaciones del canal, remitiéndonos de nuevo a los diseños de codificadores estándar.

En segundo lugar, el terminal no tiene por qué ser consciente de que está accediendo a un reconocedor automático de habla con unas determinadas características y no tiene que estar específicamente programado para ello. Además de los inconvenientes que se desprenden de esta especialización exigida al terminal para el desarrollo de servicios y que ya se han expuesto en el caso de reconocimiento local, tiene la ventaja de no requerir una estandarización del "*front-end*" del codificador. Así, por ejemplo, y como ya hemos mencionado, el reconocedor óptimo en el caso de lenguas tonales requiere la transmisión del tono fundamental o "*pitch*" [164]; por lo tanto, esto tiene que estar contemplado de alguna forma en el diseño de la parametrización que se envía y en definitiva en el terminal que accede al servicio. La solución del reconocimiento remoto, aquí propuesta, no presupone nada acerca del terminal cliente, que transmite la voz codificada según los estándares de codificación habituales.

II.6. Reconocimiento mediante transparametrización

La solución que proponemos en esta tesis para el reconocimiento de habla tanto en GSM como en IP se muestra en la Figura II-4:

Esta solución consiste en la adaptación de el "*front-end*" de los reconocedores actuales, para contrarrestar tanto la distorsión que introduce el codificador, como la debida a los errores de transmisión producidos por el canal.

Para ello, hemos estudiado la parametrización que realizan los codificadores estándar utilizados en estos entornos y la relevancia de cada uno de los parámetros que se envían (cuya intención original es la reconstrucción de la voz) para la tarea de reconocimiento. Es decir, pretendemos encontrar la manera de transformar los vectores de parámetros óptimos de la codificación, $\hat{\mathbf{F}}$, en los vectores de observaciones acústicas, $\mathbf{Y}$, más adecuados para el reconocimiento.

Esta configuración se puede observar en la Figura II-4 donde los bloques de decodificación y parametrización de la Figura II-3 se han sustituido por uno sólo de transparametrización.

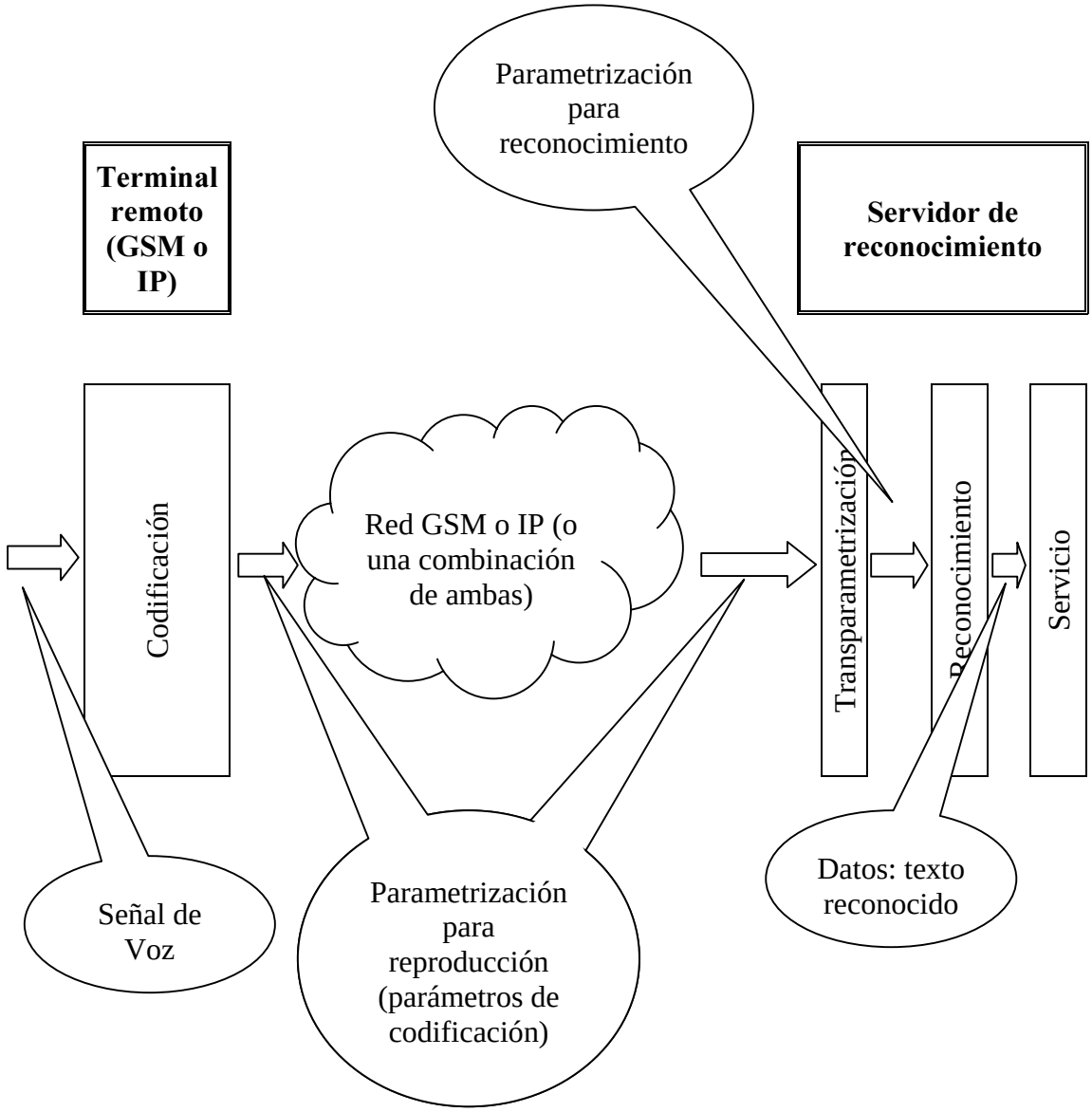


Figura II-4.- Configuración de reconocimiento remoto a partir de voz codificada.

Esta solución que explicaremos con más detalle en el Capítulo IV, resulta más robusta en ambos sistemas de transmisión aquí considerados (GSM e IP), incluso en los casos en los que hay transcodificaciones a lo largo del camino que recorre la señal. En todo caso adelantamos aquí algunas de las razones que hacen que esto sea así:

1. En el caso de GSM, los errores que afectan a los bits que codifican la información de la excitación del filtro de predicción lineal, no afectan a los parámetros de reconocimiento que se obtienen, puesto que éstos no intervienen para nada en el proceso de transparametrización. Sin embargo, en el caso del reconocimiento a partir de voz decodificada, en el que estos parámetros se obtienen tras decodificar la voz, los errores en la excitación dan lugar a una señal de voz distorsionada de la que, en consecuencia, se

extraen vectores de parámetros erróneos. Si comparamos el número de bits con los que se codifica la excitación de cada trama que, en el caso del estándar HR, es de 77, con el que se utiliza para la codificación del filtro de predicción y la energía, que es de 35, nos podemos hacer una idea de la importancia de este punto.

2. En el caso de IP, la ventaja que proponemos en el punto anterior no es aplicable puesto que en este canal la fuente principal de error es la pérdida de paquetes y por lo tanto, cuando se produce un error toda la trama al completo (o varias tramas, según sea el número de tramas que se dispongan en cada paquete IP) desaparece. Sin embargo, cuando en la codificación utilizada se emplean esquemas diferenciales o existen algunos de los parámetros codificados tienen dependencias de las tramas anteriores, utilizar sólo los parámetros estrictamente necesarios de las tramas correctas ayuda a evitar los errores producidos por la pérdia de las tramas anteriores. Por ejemplo, cuando se interpolan los parámetros que representan la envolvente espectral de las subtramas para lograr transiciones suaves entre tramas, la excitación de cada subtrama se calcula para cada envolvente espectral interpolada, y por lo tanto depende la envolvente de tramas anteriores (normalmente se restrige a una trama). Por ello, es desaconsejable la utilización de esta excitación cuando se carece de la información espectral correcta.

3. Tanto en GSM como en IP, esta aproximación evita la distorsión de codificación. Esta distorsión se produce al asumir como modelo de producción de la voz, el modelo fuente-filtro al que hemos hecho alusión en II.1.1. Este modelo, que también se utiliza implícitamente en las parametrizaciones para reconocimiento, es eficaz siempre que ambos elementos estén adecuadamente representados. En el caso de la codificación de voz, cuyo objetivo es la compresión o disminución del número de bits necesario para codificar una señal perceptualmente aceptable, esta reducción se consigue en parte, a base de describir de forma más grosera ambos elementos. Sin embargo, es lo más habitual que la codificación del filtro sea bastante precisa, mientras que la de la excitación lo sea menos, ya que tanto la reconstrucción como el reconocimiento automático resultan más sensibles a la precisión con la que se representa el filtro. Por lo tanto, normalmente, con la aproximación de reconocimiento a partir de voz codificada la distorsión de codificación se reduce a la correspondiente a la cuantificación de los parámetros del filtro, la cual suele ser muy pequeña según ha comprobado Huerta con el codificador FR ([85]) y también Tucker ([172]).

4. En el caso de GSM, tenemos que contar además con el efecto del codificador de canal, que proporciona una protección desigual a los bits en función de su importancia subjetiva. De nuevo, los bits más significativos de los parámetros del filtro LPC resultan más protegidos que el resto, con lo que aumenta la probabilidad de encontrarnos en la situación descrita en el punto 1.

5. Tanto en GSM como en IP, el hecho de sustituir la decodificación y la posterior parametrización con la transparametrización, facilita la realización de un postprocesado adaptado al reconocimiento. Así, es posible aumentar la

tasa de tramas utilizando métodos de interpolación mejores que los incluidos en los codificadores estándar, que suelen ser bastante burdos debido a las restricciones de retardo algorítmico que impone la tarea de reconstrucción. De la misma forma es posible utilizar métodos de recuperación frente a errores que no se restrinjan a los sugeridos para la reconstrucción, cuyo objetivo es el de obtener una señal perceptualmente aceptable.

Capítulo III
Reconocimiento en entornos con necesidades de codificación

III.1. Introducción

En el capítulo anterior dejamos establecida la configuración de red en la que nos proponemos implementar un servicio con una interfaz basada en reconocimiento automático de habla. En tal configuración, que denominamos, de "reconocimiento remoto", la señal de voz debe codificarse para viajar por alguna red de comunicación. Este capítulo que comienza está dedicado a describir los efectos característicos de los dos tipos de redes que hemos considerado en esta tesis: GSM e IP.

Así, empezaremos introduciendo cada una de estas dos redes, haciendo hincapié en las características de los codificadores de voz de los que hacen uso para pasar después a describir el primero de los problemas que se presenta al intentar reconocer la voz en estos entornos: la distorsión de codificación. Expondremos, en primer lugar, porque supone un problema para el reconocimiento y posteriormente revisaremos las soluciones que podemos encontrar en la literatura al respecto.

El segundo problema que contemplamos es el de los errores de transmisión que inevitablemente se producen en estas redes, distinguiendo por una parte, los introducidos por la red GSM (normalmente errores a ráfagas), y por otra, los correspondientes a las redes basadas en el protocolo IP, consistentes en pérdidas de paquetes. Tanto para los primeros como para los segundos, describiremos los métodos diseñados para prevenir este tipo de errores y que se concretan, en el caso de GSM, en los denominados codificadores de canal y en el caso de IP, en los nuevos protocolos que pretenden proporcionar calidad de servicio (*QoS –Quality of Service–)* para tráfico inelástico, como el de voz.. Por otra parte, consideraremos también los métodos de recuperación frente a errores descritos por otros autores, que aunque no hayan sido concebidos específicamente para reconocimiento automático, contribuyen a mejorar su funcionamiento.

Finalmente haremos una pequeña reseña de otro tipo de distorsiones que también aparecen en el ámbito que consideramos pero, que al contrario de las dos anteriores no son específicas de las redes GSM o IP y que por ese motivo, no trataremos en esta tesis.

La solución que proponemos en esta tesis –que describiremos con más detalle en Capítulo IV–, considera de forma integrada ambos problemas, y como consecuencia, disminuye los efectos negativos que tiene sobre los reconocedores de habla la recepción de voz codificada y afectada por errores.

III.1.1. El sistema GSM

El organismo de estandarización ETSI (*–European Telecommunications Standards Institute–*) describe dos tipos de canales de tráfico de voz en GSM: el *Full Rate Speech* (TCH/FS *–Traffic CHannel/Full rate Speech–*) a 22,8 Kb/s y el *Half Rate Speech* (TCH/HS *–Traffic CHannel/Half rate Speech–*) a 11,4 Kb/s. No obstante, debido a la poca fiabilidad del canal radio, una parte importante de esta tasa binaria se dedica a la codificación de canal.

En la fase 1 de la estandarización GSM, se describe un solo codificador de voz: el codificador FR (*–Full Rate–*), que utiliza el canal de voz del mismo nombre. Este codificador emplea 13 Kb/s dejando los restantes 9,8 Kb/s al codificador de canal. En la fase 2 se introducen dos nuevos codificadores: el EFR (*–Enhanced Full Rate–*) para el canal TCH/FS, y el HR (*–Half Rate–*) para el canal TCH/HS. El primero utiliza 12,2 Kb/s y el segundo, 5,6 Kb/s. Por último, recientemente, se ha definido un conjunto de codificadores de tipo Multi-tasa adaptativa (AMR *–Adaptive Multi-Rate–*), que consiguen una mejor calidad de voz distribuyendo los bits dedicados a la codificación de voz y de canal de forma dinámica en función de medidas de fiabilidad del canal. Más concretamente, el sistema AMR se adapta a las condiciones del canal seleccionando el modo apropiado (*full rate* o *half rate*) y la tasa binaria de codificación de voz que proporcione el nivel de protección suficiente frente a errores para determinadas condiciones de canal, pudiendo esta última variarse para adaptarse a los cambios de dichas condiciones.

A continuación se puede encontrar una descripción de cada uno de estos codificadores; no obstante, cabe resaltar precisamente que todos ellos pertenecen a la categoría de codificadores de análisis mediante síntesis que mencionábamos en II.1.1 y utilizan el modelo fuente-filtro del que haremos un análisis más detallado en IV.2.1; consecuentemente, es posible plantear en todos los casos la solución de reconocimiento mediante transparametrización que proponemos. La descripción del codificador de canal asociado a cada uno de ellos se postpone para más adelante (sección III.3.1.1).

III.1.1.1. Full Rate

El codificador FR funciona a 13 Kb/s y utiliza un predictor corto, uno largo y una excitación consistente en pulsos regularmente espaciados (RPE-LTP *–Regular Pulse Excitation-Long Term Prediction–*). Básicamente, la señal de entrada se divide en tramas de 20 ms de duración, y para cada una de ellas se calculan 8 coeficientes de predicción lineal (LP *–Linear Prediction–*) a corto plazo. Estos coeficientes se codifican como parámetros LAR (*–Log-Area-Ratio–*) utilizando 36 bits, de los cuales se asignan mayor cantidad para los iniciales, más importantes desde el punto de vista perceptual

(véase sección IV.2). Hay que resaltar en este punto, que el número de coeficientes LPC en este codificador es significativamente más pequeño que en el resto de los codificadores GSM, que utilizan 10. Si se tiene en cuenta que ésta es prácticamente toda la información utilizada por el reconocedor, se puede intuir que las tasas de reconocimiento alcanzadas seran menores que en otros casos.

A continuación, cada una de las tramas se subdivide en 4 subtramas de 5 ms y para cada una de ellas se obtiene el retardo, *L*, que corresponde con el periodo fundamental, y la ganancia, β, que definen el predictor largo de orden 1 (LTP –*Long Time Predictor*–). Finalmente, cada señal residual de 40 muestras se diezma en tres posibles secuencias de excitación de 13 pulsos cada una, regularmente espaciadas (RPE –*Regular Pulse Excitation*–); se escoge la de mayor energía y se cuantifica la amplitud máxima, $X_{máx}$, con 6 bits, y la amplitud normalizada de cada pulso, $X_{RPE}(p_i)$, con 3 bits. La posición de los pulsos (*grid position*) se codifica con 2 bits, ya que al estar regularmente espaciados sólo es necesario definir la posición de comienzo.

En el decodificador, como es lógico, se realiza el proceso inverso y se reconstruye la señal alimentando con la excitación, primero el sistema de síntesis basado en el predictor largo y después el basado en el corto. Para acabar, se realiza un postfiltrado con el fin de mejorar la calidad subjetiva de la señal reconstruida [48].

Parámetro Codificado	Nº de bits / trama	Parámetro para Reconocimiento mediante Transparametrización
LAR	36	MFCC
X_{max}	24	Energía
Ganancia LTP	8	
Retardo LTP	28	
Pulsos RPE	156	-
Posición de los pulsos	8	
TOTAL	260	

Tabla III-1. -Asignación de bits a los parámetros de codificación y su correspondencia con los parámetros de reconocimiento, para el codificador GSM-FR.

En la Tabla III-1, se puede observar como se distribuyen los 260 bits que componen cada trama entre los distintos parámetros que hemos descrito. Además, en la columna 5 señalamos cuales de estos parámetros son necesarios a la hora de efectuar el reconocimiento, según el método que proponemos (véase el Capítulo IV). De esta forma queda patente que más de la mitad de los bits codificados resultan innecesarios para nuestra tarea y que por lo tanto la utilización de los mismos sólo puede contribuir a distorsionar el vector de parámetros de reconocimiento por un doble motivo: posibles errores de transmisión (sección III.3) y distorsión en la decodificación (sección III.2).

III.1.1.2. Half Rate

El codificador HR emplea predicción lineal con excitación por suma de vectores (VSELP –*Vector Sum Excited Linear Prediction*–) para codificar con 5,6 Kb/s la señal de voz. Para empezar, se extraen 10 coeficientes LPC por cada trama de 20 ms y se transforman en 10 coeficientes de reflexión (RC –*Reflexion Coefficients*–), que se cuantifican vectorialmente (véase sección IV.2). Por tanto, la información espectral que envía este codificador es más precisa que la del FR, que sólo considera 8 coeficientes, aunque sigue siendo peor que la del EFR que como se explicará en la próxima sección, actualiza esta información cada 10 ms. Por otro lado, el GSM-HR contempla un mecanismo de interpolación blanda que pretende suavizar las transiciones entre los coeficientes LPC de dos tramas consecutivas; esta interpolación, de tipo lineal, se puede activar si la señal sintetizada de esta forma resulta mejorada por la interpolación; para ello, se transmite un bit al decodificador que indica la decisión tomada en el codificador [50].

Al igual que en el codificador *full rate* cada trama se subdivide en 4 subtramas de 5 ms. Para cada una de ellas, se utiliza un predictor largo LTP (–*Long Term Prediction*–) seleccionado mediante un procedimiento mixto que contempla una primera aproximación de en lazo abierto y un refinamiento en lazo cerrado. En primer lugar, se escogen una serie de candidatos para el retardo del predictor, *L*; estos valores se utilizan, además, para seleccionar uno de los 4 modos posibles de la excitación: 3 sonoros y 1 sordo. Después y sólo para los modos sonoros se refina el valor del retardo utilizando técnicas de lazo cerrado. Posteriormente se cuantifica *L* para la primera subtrama con 8 bits y para las otras tres de forma diferencial. Finalmente se elige un vector de la librería de excitaciones, formada por 2^9 vectores-código construidos a partir de 9 vectores base en el caso sonoro y de 2 librerías de 7 vectores base (cada una) en el caso sordo. Las ganancias de las librerías y del predictor largo se cuantifican de forma conjunta utilizando 5 bits por trama.

En la Tabla III-2 se pueden observar la distribución de los 112 bits de cada trama entre los parámetros que acabamos de mencionar. En este caso y gracias a que la energía se codifica separadamente y no es necesario extraerla a partir de otros parámetros, casi las dos terceras partes de los bits no son necesarios para reconocimiento (véase el Capítulo IV).

Parámetro Codificado (modo 0)	Nº de bits / trama	Parámetro Codificado (modos 1-3)	Nº de bits / trama	Parámetro para Reconocimiento mediante Transparametrización
PARCOR	28	PARCOR	28	MFCC
R_0	5	R_0	5	Energía
Ganancia excitación {GS,P0}	20	Ganancia excitación {GS,P0}	20	-
Códigos VSE (librería H)	28	Retardo LTP	20	
Códigos VSE (librería I)	28	Códigos VSE (librería I)	36	
Indicador de interpolación	1	Indicador de interpolación	1	
Modo	2	Modo	2	
TOTAL	112	TOTAL	112	

Tabla III-2. -Asignación de bits para cada uno de los parámetros de codificación y su correspondencia con los de reconocimiento, para el codificador GSM-HR.

III.1.1.3. Enhanced Full Rate

El estándar EFR codifica la voz a 12,2 Kb/s utilizando predicción lineal y una excitación algebraica (ACELP *–Algebraic Code Excited Linear Prediction–*), y consigue una calidad alrededor de 4 en la escala MOS.

El algoritmo divide cada trama de 20 ms en 4 subtramas de 5 ms; se extraen 10 coeficientes LPC dos veces por trama, considerando ventanas de 30 ms; la primera ventana pone más énfasis en la segunda subtrama, mientras que la segunda lo hace en la cuarta subtrama. Los parámetros LP se convierten en coeficientes LSP (*–Line Spectral Pairs–*) (véase la sección IV.2). A continuación, se aplica sobre los LSP un predictor de tipo MA (*–Moving Average–*) y se cuantifica el residuo. Los dos conjuntos de LSP se cuantifican con un total de 38 bits. Como se puede observar, la información espectral que se envía en este codificador es muy distinta de la que se envía en el codificador FR: se calculan 10 coeficientes en lugar de 8 cada 10 ms en vez de cada 20 ms.

Para el cálculo de la excitación, al igual que en los anteriores codificadores, se considera por una parte una excitación adaptativa y una fija. La obtención del LTP se realiza en dos pasos: una primera aproximación en lazo abierto y un posterior

refinamiento en lazo cerrado. El retardo se cuantifica con 9 bits para la primera y tercera subtramas y con 6 de forma diferencial para la segunda y cuarta.

Por su parte, la librería algebraica está basada en 5 pistas (*–tracks–*) que determinan (cada una de ellas), las 8 posiciones permitidas de 2 pulsos no nulos en cada subtrama de 40 muestras; las amplitudes de los pulsos pueden ser +1 ó –1, pero también pueden colocarse los dos pulsos de una pista en la misma posición para generar un pulso de amplitud +2 ó –2 [52]. Para la cuantificación de esta librería se utilizan un total de 35 bits por subtrama. En la sección IV.2.3, explicamos con más detalle este tipo de codificación de la excitación.

Parámetro Codificado	Nº de bits / trama	Parámetro para Reconocimiento mediante Transparametrización
LSP	38	MFCC
Ganancia librería estática	20	Energía
Ganancia LTP	16	
Retardo LTP	30	
Códigos algebraicos	140	-
CRC (adicional)	8	
Bits repetidos (códigos de protección)	8	
TOTAL	260	

Tabla III-3. -Asignación de bits para los parámetros de codificación y su correspondencia con los de reconocimiento, para el codificador GSM-EFR.

En la Tabla III-3, se puede observar cómo la distribución de bits entre estos parámetros y su correspondencia con los parámetros de reconocimiento sigue la misma tónica que los anteriores codificadores.

III.1.1.4. Adaptive Multi-Rate

El grupo SMG 11 de ETSI ha presentado recientemente (mayo, 2000) una nueva familia de codificadores llamada GSM AMR (*–Adaptive Multi-Rate–*). El concepto de tasa de variable adaptativa permite una adaptación dinámica del régimen binario

asignado al codificador de canal y de fuente, tanto para el canal TCH/FS como el TCH/HS. Así, la calidad global de la voz mejora aumentando la proporción de bits que se dedican a protección contra errores, cuando la fiabilidad del canal es baja. La velocidad de adaptación se controla mediante un procedimiento "*en banda*" dedicado, es decir, utilizando el mismo canal en el que se transmiten los datos de voz, se transmite la información sobre el modo y tasa de bits utilizados, empleando para ello, unos bits exclusivamente dedicados a esta labor. Esta familia de codificadores no sólo se perfila como una forma de mejorar la calidad de voz en GSM, sino también como parte de la futura generación de comunicaciones móviles UMTS (*–Universal Mobile Telecommunications System–*). Además, también está previsto cómo transportar este estándar dentro del protocolo RTP (*–Real-time Transport Protocol–*) utilizando en el entorno IP.

El codificador AMR es de tipo ACELP, con 8 posibles regímenes binarios: 4,75, 5,15, 5,90, 6,7, 7,4, 7,95, 10,2 ó 12, 2 Kb/s sobre los dos canales (TCH/HS y TCH/FS), aunque sólo los 6 primeros son posibles en el caso del TCH/HS.

Para cualquier regimen binario se extraen 10 coeficientes LPC cada 20 ms. La información sobre estos coeficientes se cuantifica en forma de LSPs. El vector LSP se predice con un predictor de orden 1 y se cuantifica el residuo. El vector global de 10 LSP se divide en subvectores de longitud 3, 3 y 4, y cada subvector se codifica con 7 ó 9 bits, dependiendo de la tasa de bits dedicada al codificador de fuente. En el decodificador, los LSP se interpolan para cada subtrama.

Para el cálculo de la excitación, cada trama de 20 ms se divide en 4 subtramas. Se calcula el residuo para cada una de las subtramas y se determina una ganancia y un desplazamiento para la librería adaptativa (LTP *–Long Time Prediction–*). La librería algebraica es muy similar a la del codificador EFR, aunque utiliza menos pistas para reducir el número de bits requeridos para codificar los índices. Las tasas binarias más bajas incluyen sólo un pulso por pista [54].

Es preciso señalar aquí que independientemente del régimen binario elegido se conserva el número de coeficientes LPC y la frecuencia con la que se extraen, a la vez que se disminuye la precisión con la que se cuantifican. Esto beneficia nuestra aproximación de reconocimiento, ya que, según se desprende de [172] y como se discutirá en 0, la tasa de tramas o la frecuencia con la que se extraen estos coeficientes es crítica a la hora de realizar el reconocimiento. También es conveniente resaltar como a medida que los regímenes binarios bajan, la descripción de la excitación es notablemente más burda. Esto es un factor más que desaconseja la utilización de esta última cuando sea posible evitarlo.

III.1.2. Las redes IP

Tradicionalmente se ha venido distinguiendo entre servicios de voz, proporcionados a través de redes de conmutación de circuitos, y servicios de datos, prestados por las redes de conmutación de paquetes. Esta distinción se basa en los distintos requerimientos de estos dos tipos de tráfico: por una parte, el tráfico de voz requiere un retardo, un *jitter* (o variación del retardo) y un ancho de banda determinados y constantes, aunque es hasta cierto punto tolerante a fallos en la transmisión. Por otra

parte, el tráfico de datos, no tiene restricciones severas en cuanto al retardo y el ancho de banda, pero es muy exigente en lo que se refiere a la integridad de los mismos. El primero se denomina tráfico ineslástico y el segundo, elástico.

Hasta hace unos años, el volumen de tráfico de voz era muy superior al de datos lo que ha motivado el desarrollo de una gran infraestructura especialmente optimizada para el tráfico vocal. Sin embargo, hay que tener en cuenta que el volumen de tráfico de datos está creciendo a velocidades mucho mayores y que, además, crece la demanda de otro tipo de tráfico como el vídeo.

Por otra parte, las redes de paquetes basadas en el protocolo TCP/IP son las que más éxito han tenido, pero si han de tener éxito a la hora de asumir también las funciones de transmisión de voz, deben adaptarse para proporcionar, al menos, una calidad similar a la que ofrecen actualmente las redes exclusivamente dedicadas a este tipo de tráfico.

Con el fin de justificar el gran esfuerzo que se está realizando para dotar a las redes IP de capacidades para transmitir voz de calidad, resulta conveniente evaluar los beneficios que cabe esperar de esta tecnología y que podemos agrupar en cuatro categorías:

1. **Reducción de costes**: a pesar de que la reducción de los costes de las llamadas a larga distancia la hayan hecho popular, todavía no esta claro que la transmisión de voz sobre IP pueda producir ventajas económicas a largo plazo, ya que esta reducción de los precios de las llamadas se debe más a que se evitan costes de acceso y establecimiento, que a una reducción en los costes de los recursos puestos en juego. En todo caso, la compartición de equipos y costes de operación de los usuarios de voz y de datos, puede mejorar la eficiencia de la red.

2. **Simplificación de tareas**: una infraestructura integrada que soporte todas las formas de comunicación permite una mayor estandarización y reduce los costes totales. Esta infraestructura puede soportar optimización dinámica del ancho de banda, ya que los diferentes patrones de tráfico de voz y datos ofrecen oportunidades para mejorar la eficiencia significativamente.

3. **Consolidación**: dado que la mano de obra asociada al mantenimiento es uno de los elementos más costosos en una red, cualquier oportunidad de combinar operaciones, eliminar fallos, y de consolidar sistemas de tarificación es beneficiosa. Por tanto, el uso universal de los protocolos IP para todas las aplicaciones mantiene la promesa tanto de reducir la complejidad, como de aumentar la flexibilidad.

4. **Aplicaciones avanzadas**: a pesar de lo expuesto anteriormente y como ya hemos adelantado en II.2, se espera que los beneficios a largo plazo procedan de aplicaciones multimedia y multiservicio; es decir, la integración de voz y datos en nuevas aplicaciones será el motor económico en el futuro. En este caso, las tecnologías del habla, y en concreto, el reconocimiento, juegan un papel fundamental, ya que posibilitan el desarrollo de interfaces vocales para todo tipo de aplicaciones.

Uno de los grandes problemas que surgieron, hace un lustro, para la puesta en funcionamiento de los sistemas de telefonía IP fue la carencia de un estándar que definiera cómo transmitir voz a través de la red IP. Por ese motivo, en aquellos momentos, fue elegido como estándar "de facto" el protocolo H.323, que como veremos en III.3.2.1.4, fue inicialmente diseñado para soportar tráfico multimedia a través de redes de área local. Así, entre los codificadores de voz más utilizados en estas redes se encuentran el G.723.1 y el G.729, ambos están contemplados en dicha especificación. En las próximas dos secciones describimos las características de estos codificadores.

Más tarde se desarrolló otro protocolo específico para redes IP: el protocolo SIP (*–Session Initialization Protocol–*) (véase sección III.3.2.1.4.2) y actualmente no existe consenso acerca de la conveniencia de uno u otro protocolos. Éste admite todos los codecs registrados en la IANA (*–Internet Assigned Numbers Authority–*), que también incluye los dos anteriormente mencionados.

Por otra parte, a través del proyecto TIPHON (*–Telecommunications and Internet Protocol Harmonization Over Networks–*) de ETSI (*–European Telecommunications Standards Institute–*), también se estudia incluir dentro de H.323 (Anexo H) los codificadores de GSM [57].

III.1.2.1. G.723.1

Esta recomendación especifica un algoritmo de codificación que puede usarse para comprimir voz u otras señales de audio que formen parte de servicios multimedia a bajo régimen binario; este codificador funciona a 5,3 ó 6,3 Kb/s, siendo esta última, obviamente, la que proporciona mayor calidad. Es posible conmutar entre estas dos tasas binarias trama a trama y además, opcionalmente, se pueden obtener tasas menores comprimiendo los espacios de silencio, haciendo uso de un detector de actividad de voz (VAD *–Voice Activity Detection–*).

El G.723.1 está diseñado para producir voz de alta calidad (entre 3,8 y 3,9 en la escala MOS para cada uno de los dos regímenes binarios disponibles) con una complejidad limitada. Codifica la voz por tramas y, como el resto de codificadores que hemos presentado, pertenece al conjunto de codificadores denominados de análisis mediante síntesis.

La envolvente espectral se extrae mediante predicción lineal de orden 10 y posteriormente se transforma en coeficientes LSP (*–Line Spectral Pairs–*) para su transmisión (véase la sección IV.2).

Para el cálculo de la excitación se divide la trama en 4 subtramas de 7,5 ms de duración (60 muestras) y a continuación para cada subtrama, vuelve de nuevo a distinguirse entre una parte adaptativa y otra estocástica. La primera el común para ambos regímenes binarios y está implementada en un predictor largo de orden 5. La segunda, sin embargo depende del régimen binario que se utiliza en cada momento: es de tipo multipulso (MP-MLQ *–Multi-Pulse Maximum Likelihood Quantization–*) cuando funciona a 6,3 Kb/s y está basada en la selección de 6 pulsos para las tramas pares y 5 para las impares, imponiendo un criterio de máxima verosimilitud restringido a posiciones pares o impares de todos los pulsos. Cuando utiliza la tasa menor, por otra parte, la excitación es de tipo CELP algebraico (ACELP *–Algebraic Code-Excited*

Linear-Prediction–), como en el caso del codificador GSM-EFR, que consta de 4 pistas en las que de nuevo sólo son posibles las posiciones pares o las impares (con un bit adicional que señala esta opción). Un máximo de 4 pulsos se codifican utilizado 3 bits para la posición y otro más para el signo que toma valores {+1,–1} aun cuando también es posible generar una excitación con sólo 2 pulsos.

La longitud de la trama es de 30 ms y la ventana de análisis añade además 7,5 ms de la trama posterior, lo que se denomina *look-ahead* o anticipación. Es por esto que el retardo algorítmico es de 37,5 ms. Llama la atención el hecho de que la longitud de trama es relativamente grande si la comparamos con el resto de los codificadores que se utilizan en estas aplicaciones, pero gracias a ello se consiguen regímenes binarios tan bajos. Sin embargo, la pérdida de paquetes es especialmente dañina, puesto que la porción de voz que se pierde es considerable. Es más, muchas veces se opta por enviar varias tramas de voz en un solo paquete para aumentar la eficiencia de la transmisión (bits útiles / bits de cabeceras) y como consecuencia el problema de las pérdidas se agudiza.

Otra característica interesante de este codificador es que ha sido diseñado para ser robusto ante pérdidas de tramas; sin embargo, la estrategia de recuperación frente a errores depende de una indicación externa de cuáles son las tramas perdidas. Esto se puede hacer fácilmente si se utilizan para el transporte protocolos como el RTP (véase III.3.2.1.3). Como se expondrá en III.3.2.2, cuando el decodificador está en modo "recuperación frente a errores", utiliza los LSPs de las tramas previas para predecir los de la trama perdida y genera una excitación sintética sonora o sorda basada en una decisión tomada a partir de las tramas anteriores. Además, va atenuando la voz decodificada cuando se dejan de recibir varias tramas seguidas y la enmudece totalmente tras perderse tres tramas consecutivas.

III.1.2.2. G.729

Esta recomendación especifica un codificador a 8 Kb/s de tipo CS-ACELP (–*Conjugate-Structure Algebraic Code Excited Linear Predictive Coder*–). Opera sobre tramas de 10 ms y emplea, como en el caso de G.723.1, un *look-ahead* o anticipación de 5 ms.

La información de la envolvente espectral, se codifica, como en el caso anterior, en 10 parámetros LSP. Sin embargo, esta información no se transmite directamente al canal sino que se codifica diferencialmente utilizando un predictor 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 sendas 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, como en los casos anteriores, está compuesta por una librería adaptativa y una estocástica. Cada trama es dividida en 2 subtramas de 5 ms (80 muestras) cada una y para cada una de estas subtramas se calcula un retardo del predictor largo permitiendo valores fraccionarios del mismo de hasta $\frac{1}{3}$ de la resolución. Por su parte, la librería estocástica es de tipo ACELP y es similar a la utilizada en GSM-EFR, pero aquí el número de pistas es menor (4) y solo se permite un pulso por cada pista que toma valor {+1, –1}. Cada pista señala 8 posiciones posibles para cada pulso para las 3 primeras pistas y por tanto necesita 3 bits para su cuantificación, más un bit para el signo. Sin embargo la última pista consta de 16 posibles posiciones, con lo que un bit adicional es necesario 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 (CS –*Conjugate Structure*–) 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.

Existen dos versiones de este codificador: G.729 y G.729A. La primera es más compleja y ofrece una calidad ligeramente superior; Se trata de un codificador preparado para obtener un retardo bajo; el hecho de que utilice tramas de 10 ms, frente a las de 30 ms de G.723.1, apunta en esa dirección. Sin embargo, esto podría ser un inconveniente a la hora de utilizarlo en entornos IP, ya que por ser la trama más pequeña, la cantidad de bits de información enviados por paquete resulta ser una porción muy pequeña del total de bits enviados (incluyendo cabeceras), salvo que se envíen varias tramas por paquete, obviamente.

Al igual que G.723.1, este codificador está preparado para actuar frente a pérdidas de paquetes, como veremos en III.3.2.2, pero 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 en estos momentos está procesando sino también de su propio estado (definido por las tramas anteriores). Esto hace que aún cuando las tramas comienzan a llegar correctamenten después de haber tenido lugar un pérdida, el codificador tarde en recuperarse.

III.2. Distorsión de codificación

La elección de un determinado codificador para una aplicación concreta exige un compromiso entre la calidad, el régimen binario, el retardo y la complejidad computacional. Tanto en GSM como en IP, el ancho de banda disponible es un factor determinante. Es por ello que todos los codificadores que hemos revisado en la sección anterior actúan a tasas binarias bajas que se consiguen tras asumir el modelo fuente-filtro para la generación de la voz (sección IV.2.1). Lógicamente, esta simplificación redunda en una pérdida de calidad en la señal reconstruida tanto más cuanto menor sea el régimen binario.

Los seres humanos resultan bastante tolerantes a este tipo de distorsión, pero no ocurre lo mismo con los reconocedores automáticos [172]. Durante los últimos años varios autores han cuantificado los efectos de los distintos codificadores sobre los sistemas de reconocimiento, de forma que podemos concluir que:

1. Si el sistema se entrena con voz codificada a 64 Kb/s, la tasa de error aumenta considerablemente cuando se aplica para reconocer señales codificadas con menor tasa binaria [61].

2. Como es de esperar, se obtienen los mejores resultados cuando el entrenamiento y el test se realizan con la voz codificada con el mismo algoritmo[61].

3. Los codificadores con tasas binarias superiores a 16 Kpbs no influyen significativamente en la tasa de reconocimiento, incluso aunque la señal de voz atraviese varias etapas consecutivas de codificación [126].

4. Si el régimen binario baja de16 Kb/s, los efectos son importantes y tanto más cuanto menor sea dicho régimen [126].

Como se habrá podido deducir, todos los codificadores a los que nos referimos en esta tesis están bajo el supuesto cuarto (13 Kb/s el de más alta tasa binaria –FR– y 5,3 Kb/s el de menor –G.723.1–) y como se pondrá de manifiesto en el Capítulo V la distorsión, aún cuando el reconocedor ha sido entrenado con voz decodificada (supuesto 2), afecta en mayor o menor medida según el codificador del que se trate.

Para paliar los efectos de esta distorsión, Dufour et al. proponen y evalúan una parametrización robusta llamada RN_LFCC (–*Root-Normalized Linear Frequency Cepstral Coefficients*–) con la que substituyen a los tradicionales MFCC (véase la sección IV.3.1). De esta forma consiguen mejoras tanto en el caso de HR como de FR. Además, emplean técnicas de compensación de ruido (NSS –*Non-linear Spectral Substraction*–) y también de canal (CMS –*Cepstral Mean Substraction*–) que interactúan de forma irregular con las parametrizaciones que comparan [40].

Salonidis et al. [162] han estudiado el efecto de varias topologías de red en las que se consideran los efectos del *tandeming* o la disposición en cascada de varios codificadores de voz. En concreto, han evaluado distintas combinaciones del codificador FR (hasta tres etapas) y los codificadores G.711 (codificación PCM) y G.721 y G.723 (codificación ADPCM –32 y 16 Kb/s, respectivamente–), con el objetivo de caracterizar los efectos de la distorsión de codificación en casos tales como llamadas de móvil a móvil que atraviesan la red fija, distintas transcodificaciones dentro de esta última, etc.

Para aliviar los efectos de esta distorsión, se proponen en primer lugar caracterizar el ruido de codificación y utilizan la técnica del Filtrado Probabilístico Óptimo (POF –*Probabilistic Optimum Filtering*–) para transformar o "limpiar" los vectores de parámetros de entrada ruidoso de forma que se asemejen a los originales o no contaminados. Los parámetros que caracterizan este método son el número de gaussianas utilizadas, cada una de las cuales representa una región del espacio acústico de los vectores de características de reconocimiento, y el retardo o número de tramas vecinas de la que se está considerando, que intervienen en la estimación. La conclusión

es que, a pesar de que las tasas de reconocimiento conseguidas utilizando este procedimiento mejoran en alguna medida las originales, no se puede decir que el ruido de codificación siga algún tipo de patrón que pueda ser modelado por este procedimiento, puesto que no hay variaciones en los resultados obtenidos en función del número de gaussianas o el retardo empleados.

Esta conclusión coincide con la de Huerta en [85], que afirma que analizar el efecto de la codificación (en concreto, la de FR y aunque él no considera el caso de varias etapas de codificación) sobre los parámetros cepstrales es una tarea prohibitiva, debido a la naturaleza no estacionaria de la perturbación; esto es así porque la perturbación producida por la codificación varía sustancialmente de trama a trama.

Sin embargo, Huerta sí analiza el efecto promedio sobre las distribuciones de mezcla de Gaussianas generadas por los HMM que caracterizan las observaciones, comparando para un mismo estado de un determinado fonema las distribuciones obtenidas por la mezcla de las gaussianas cuando la señal ha sido codificada (estándar FR) o cuando no lo ha sido. Se concluye que, aunque no hay cambios radicales en la forma de estas distribuciones, las medias de las Gaussianas que las componen aparecen más alejadas entre sí; no obstante, las varianzas no aumentan. Este alejamiento entre los centros de las Gaussianas hace que las distribuciones aparezcan más ensanchadas en el espacio de parámetros, y por lo tanto, la probabilidad de error en la clasificación o solapamiento entre distribuciones de clases distintas aumenta.

Esto entra en contradicción con las conclusiones de Salonidis et al., que al comparar dos métodos de adaptación de los modelos de reconocimiento, uno de los cuales permite la modificación de las varianzas de las gaussianas y el otro no, afirman que la causa de que el primero funcione mejor que el segundo es precisamente esta posibilidad de ensanchar dichas gaussianas. A pesar de ello, en nuestra opinión, lo que están observando ambos autores es un ensanchamiento de las distribuciones de probabilidad debido al ruido, que el primero compensa aumentando las varianzas de las gaussianas y el segundo alejando los centros de las mismas entre sí.

Otra conclusión notable de Salonidis et al. se obtiene cuando intentan generalizar el método de adaptación propuesto para el caso real en el que no se conoce la topología o el historial de codificaciones de la señal obtenida en el destino. En esta situación proponen lo que denominan “transformación *cocktail*”, que consiste en entrenar la adaptación de los modelos con frases procedentes, a partes iguales, de las seis topologías que consideran. El resultado es sólo ligeramente peor que el obtenido particularizando la adaptación para cada topología. Esto se debe, probablemente, a que sólo el codificador FR, de todos los que conforman las diferentes topologías, produce una distorsión de codificación apreciable, como se puede deducir del supuesto 3, que mencionábamos al comienzo de esta discusión.

En la sección V.2.3 presentamos los experimentos que hemos llevado a cabo para verificar la robustez del método de reconocimiento mediante transparametrización también en situaciones de transcodificación o *tandeming*, en donde no sólo consideramos las situaciones de transcodificación dentro del sistema GSM sino también posibles transcodificaciones que resulten de atravesar redes IP ([68] y [69]). En estos experimentos, como es lógico, sólo podemos aplicar esta técnica con los modelos correspondientes al codificador de la última etapa ya que no es posible la “transformación *cocktail*”. Los resultados que hemos obtenido constatan ciertas

pérdidas respecto al caso en el que no hay transcodificaciones, pero en todo caso sigue resultando ventajoso respecto al procedimiento de decodificar la voz.

En [67] y [66], Gallardo et al., y en [86], Huerta et al, introducen paralelamente la idea de evitar o aliviar la distorsión de codificación empleando para reconocer vectores directamente derivados de los parámetros de codificación recibidos del canal de transmisión. Ambos trabajos están orientados a aplicaciones GSM y el estándar utilizado es el FR.

En el primer caso, la idea subyacente es la de evitar una decodificación de la señal de voz para evitar que los parámetros correspondientes a la excitación, que poco tienen que aportar en el proceso de reconocimiento contaminen a los parámetros espectrales fundamentales en dicho proceso. Resulta por ello lógico que los beneficios de esta aproximación se obtengan en condiciones de contaminación por errores de transmisión, como explicaremos con más detalle en III.3.1.3.

En el segundo caso, Huerta et al. proponen extraer también información del residuo o excitación para obtener los parámetros de reconocimiento, a pesar de que también señala que la cuantificación de la excitación es bastante menos precisa que la del filtro. De hecho, más tarde ([85]) comparará los resultados de reconocimiento utilizando estos parámetros cuantificados y sin cuantificar, concluyendo que las prestaciones se degradan bastante más debido a la cuantificación de la excitación que a la del filtro (a pesar de que más del 85 % de los bits se destinan a la cuantificación de la primera –véase la Tabla III-1–). En todo caso, el estudio de Huerta, se concentra en contrarrestar los efectos de ruido aditivo coloreado y no en los errores introducidos por el canal radio, como comentaremos en III.4.

Más tarde, Gallardo et al. ensayarán las mismas ideas sobre el codificador HR ([65]). Como ya hemos advertido en la descripción de los codificadores del sistema GSM (sección III.1.1), este codificador, como todos los demás que hemos considerado y que fueron diseñados posteriormente, utiliza 10 parámetros LPC para describir la envolvente espectral de la señal, a diferencia del más antiguo FR, objeto de los estudios precedentes, que utiliza solamente 8. Este es, en nuestra opinión el motivo de que Huerta et al. encontraran información espectral relevante codificada como parte de la excitación.

En esta misma dirección, que evita decodificar la señal y la distorsión asociada, Kim et al. [112] proponen una solución con la que consiguen resultados de reconocimiento muy cercanos a los logrados a partir de la señal de voz sin codificar utilizando como parámetros adicionales a los MFCC las ganancias de las librerías fija y adaptativa del codificador IS-641, en condiciones limpias (sin contaminación del canal o de otro tipo). Tan solo les separa de obtener estos resultados el proceso de cuantificación que se efectúa sobre los parámetros. La cuantificación de los parámetros es muy diferente en cada codificador estándar, y de hecho, en Huerta et al, que utilizan el codificador GSM-FR, podemos encontrar conclusiones que demuestran que la cuantificación de los parámetros espectrales no tiene consecuencias notables sobre el reconocedor automático.

Es significativo, sin embargo, el hecho de que para obtener el parámetro de energía, que se incluye habitualmente en el vector de parámetros, los citados autores utilicen la energía correspondiente al vector de excitación. Para ello, se procede a la decodificación

de este vector utilizando los parámetros del codificador que lo describen y que, generalmente, suelen ser aquellos cuya cuantificación produce peores consecuencias (véase por ejemplo, la cuantificación que se lleva a cabo en los codificadores de GSM en la la sección III.3.1.1). Además, también son los más vulnerables en el caso de que exista una codificación de canal que los proteja frente a errores de transmisión, como discutiremos en III.3. Por su parte, Huerta et al. ([85], [86]) utilizan como parámetro de energía, únicamente la contribución del filtro representado mediante los parámetros LP. En la aproximación que nosotros proponemos este parámetro de energía es estimado a partir de los parámetros mejor cuantificados y protegidos, como defenderemos en IV.5.1.

Por otra parte, lo que motiva la inclusión de estos dos parámetros que mencionábamos en el vector de parámetros de reconocimiento, es el hecho de que la adición de información sobre la condición sonora o sorda de la trama considerada mejora los resultados de reconocimiento. En estas dos ganancias, afirman los autores, se encuentra implícita esta distinción sonora/sorda, y es el motivo del mejor funcionamiento del reconocedor. Sin embargo, para estas comparaciones se ha optado por igualar el número de parámetros incluidos en cada vector de forma que los dos últimos parámetros MFCC del vector original son sustituidos por las dos mencionadas ganancias. En nuestra opinión sería muy interesante comparar estas dos opciones incluyendo en el vector de parámetros de la voz original uno o varios parámetros de decisión sonora/sorda, de forma que las dos parametrizaciones consideradas contuvieran el mismo tipo de información.

Además, ninguna de las soluciones anteriores, evalúa la distorsión de codificación producida por los codificadores habitualmente empleados en la transmisión de voz sobre las redes IP y que describimos en III.1.2. En [148], poníamos de manifiesto las complicaciones adicionales que supone el empleo de un codificador como el G.723.1 y que provienen en su mayor parte del hecho de que este codificador emplea un periodo de trama muy superior al de los codificadores de GSM –30 ms, frente a los 20 ms de estos últimos–. En [147] proponemos una solución para este problema que consiste en hacer una interpolación de los parámetros espectrales para obtener periodos de trama similares a los que se emplean en reconocimiento. Esta interpolación se ha diseñado teniendo en cuenta que los parámetros así obtenidos van a ser empleados en reconocimiento, distinguiéndose de esta manera de la que suelen llevar a cabo los codificadores, con el propósito de obtener transiciones suaves entre tramas o subtramas consecutivas. Los detalles de esta interpolación los describimos en V.2.2.3.2. Finalmente, en [145] se introduce un método para estimar la energía de una trama utilizando una porción de los parámetros correspondientes a la excitación. Como veremos en IV.5.1, esto produce mejoras en las tasas de reconocimiento así obtenidas si las comparamos con aquellas en las que se utiliza como parámetro de energía el extraído de la voz decodificada.

III.3. Errores de transmisión

El segundo de los problemas específicos de los entornos GSM e IP que hemos considerado en esta tesis es el de los errores de transmisión. Además del deterioro que provoca la codificación de la voz en la tarea de reconocimiento, hay que considerar el

problema que supone que los datos que recibimos codificados no sean fiables. Sin embargo, al contrario que en el caso de la distorsión de codificación que sólo depende del codificador empleado, el tipo de errores que se producen es distinto en función del entorno que se considere.

En esta sección, discutiremos las consecuencias derivadas de errores aislados y errores a ráfagas (en el entorno GSM) y pérdidas de paquetes (en el entorno IP). Después, para cada uno de esos entornos examinaremos las soluciones que se han planteado y distinguiremos entre dos tipos:

1. Las orientadas a evitar la pérdida efectiva de paquetes, como las que consisten en hacer modificaciones de la red o los protocolos que la gestionan; o las basadas en que el emisor realice alguna acción tal como el reenvío de la trama errónea (protocolo TCP), adaptación de los parámetros de la transmisión (régimen binario e incluso tipo de codificado), o envío de información redundante (códigos de protección en la codificación de canal de GSM o esquemas FEC –*Forward Error Correction*– en redes IP) que ayude a reparar los errores; o también aquellas que ayudan a aliviar los efectos de los errores evitando que su distribución influya de forma muy negativa en la decodificación tales como la reordenación de bits o el entrelazado.

2. Las orientadas a aliviar los efectos de los errores desde el receptor, es decir, aquellos que una vez consumado contribuyen a que la voz reproducida tenga una mejor calidad desde el punto de vista perceptual. Así, incluiremos en este grupo todas las estrategias basadas en el receptor tales como inserciones, interpolación, etc.

En [149], [159] ó [1] se pueden encontrar clasificaciones de estos métodos distintas de la aquí expuesta, pero la clasificación anterior está en consonancia con la orientación de esta tesis, que nos llevó en el Capítulo II a escoger la configuración de reconocimiento remoto, en el sentido de que separamos y tenemos en cuenta aquello que forma parte del entorno y sobre lo que no actuamos, y preparamos al receptor para obtener mejores resultados con aquello que encuentre.

Además, los procedimientos anteriormente señalados no están diseñados para realizar un reconocimiento automático posterior y aunque en algunos casos es posible adecuarlos para esta tarea, debemos tener en cuenta que los requerimientos de reconstrucción y de reconocimiento no son idénticos, como discutiremos en IV.1. Por ello, terminamos exponiendo aquellos procedimientos o estudios que han considerado específicamente esta situación y entre los cuales se encuentra la propuesta de esta tesis..

III.3.1. Entorno GSM

El canal radio a través de cual se transmite la voz en los sistemas GSM es considerablemente más propenso a errores que un enlace fijo convencional. Además, estos errores tienden a producirse a ráfagas, es decir, muchas veces los errores afectan a una trama o varias consecutivas. Desde el punto de vista del emisor, el estándar GSM especifica un codificador de canal capaz de detectar y corregir algunos errores, Por otra

parte, en el receptor también se prevén procedimientos para mejorar la naturalidad de la voz a pesar de los errores, aunque estos no son normativos.

III.3.1.1. Protección frente a errores: Codificación de canal

Una porción considerable del ancho de banda disponible en el canal radio se dedica a la codificación de canal; concretamente, 9,8 de los 22,8 Kb/s disponibles en FR, 5,8 de 11,4 Kb/s en HR y 10,6 de 22,8 en EFR se encargan de garantizar en la medida de lo posible, la integridad de los datos recibidos. En el caso de los estándares de codificación más recientes, AMR, esta porción es variable en función de la fiabilidad del canal, lo cual nos da una idea de la importancia de este tipo de codificación en el sistema GSM y también en el futuro UMTS e incluso en IP, que como ya hemos dicho también contempla la posibilidad de transmitir con este estándar a través del protocolo RTP.

En esta sección dedicada a la codificación de canal, queremos destacar las implicaciones que tiene el hecho de que no se proteja de igual forma unos parámetros que otros, para nuestro objetivo de reconocimiento. Es notable el hecho de que los parámetros que mejor se preservan de los errores son precisamente aquellos que necesitamos utilizar para realizar el reconocimiento mediante transparametrización propuesto en esta tesis, lo cual resulta una ventaja evidente sobre las aproximaciones que utilizan además el resto de los parámetros más verosímilmente afectados por errores de transmisión.

III.3.1.1.1. Full Rate

Para proteger la información transmitida a través del canal radio, el codificador de canal de GSM utiliza codificación convolucional y entrelazado. Los bits procedentes de codificador de voz se distribuyen en tres grupos atendiendo a su importancia subjetiva en la reconstrucción de la señal. Así, los 260 bits de una trama se clasifican de la siguiente forma:

Clase Ia: 50 bits (muy sensibles a errores de transmisión).

Clase Ib: 132 bits (moderadamente sensibles).

Clase II: 78 bits (poco sensibles).

Los bits de la case Ia llevan un Código de Redundancia Cíclica (CRC –*Cyclic Redundancy Code*–) de 3 bits para la detección de errores. Estos 53 bits, junto con los 132 de clase Ib y una secuencia de cola de 4 bits (es decir, un total de 189 bits), se introducen en un codificador convolucional. A la salida de este codificador se obtienen 378 bits a los que se añaden los restantes 78 de la clase II, que van sin protección. De esta forma, se obtienen 456 bits cada 20 ms, y para protegerlos frente a los errores a ráfagas, se reordenan y agrupan en 8 bloques de 57 bits. Éstos se transmiten en 8 ráfagas o intervalos de tiempo consecutivos, los cuales llevan 2 bloques de 57 bits cada uno. Así, cada cuatro bloques empieza una nueva trama de voz que se prolonga durante las ocho ráfagas siguientes y cada ráfaga lleva, información de 2 tramas consecutivas [45].

Parámetro Codificado	Nº de bits / trama				Parámetro para Reconocimiento mediante Transparametrización
	Clase Ia	Clase Ib	Clase II	TOTAL	
LAR	14	9	13	36	MFCC
X_{max}	12	8	4	24	Energía
Ganancia LTP	0	8	0	8	
Retardo LTP	24	4	0	28	
Pulsos RPE	0	95	61	156	-
Posición de los pulsos	0	8	0	8	
TOTAL	50	132	78	260	

Tabla III-4. – Distribución de los bits de protección del codificador de canal entre los distintos parámetros codificados en el codificador GSM-FR y su correspondencia con los utilizados en reconocimiento mediante transparametrización.

En la Tabla III-4 se puede observar cómo es la distribución de la protección del codificador de canal en función de los diferentes parámetros codificados. Concretamente los parámetros LAR, X_{max} y los retardos del predictor largo son los más beneficiados, mientras que los pulsos RPE que codifican la excitación se encuentran considerablemente más desprotegidos. En todo caso es importante resaltar cómo, si observamos esta distribución de la protección sobre los parámetros que necesitamos para realizar el reconocimiento mediante transparametrización, éstos se encuentran notablemente más seguros que los que no se utilizan. En la sección IV.4 expondremos cómo se realiza la transformación de unos en los otros, pero como veremos en V.2.1 esto tiene consecuencias que claramente benefician a la aproximación de reconocimiento mediante transparametrización propuesta en esta tesis y desaconsejan la decodificación de la voz con propósitos de reconocimiento en presencia de errores de transmisión, ya que en este proceso se utilizan todos los parámetros, incluyendo más verosímilmente afectados debido a su desprotección.

III.3.1.1.2. Half Rate

La codificación de canal en el estándar HR es muy similar a la que se ha descrito para el codificador FR. Aquí, los 112 bits codificados se dividen en las 3 clases que han descrito en la sección anterior de la siguiente manera:

Clase Ia: 22 bits (muy sensibles a errores de transmisión).

Clase Ib: 73 bits (moderadamente sensibles).

Clase II: 17 bits (poco sensibles).

Parámetro Codificado	Nº de bits / trama				Parámetro para Reconocimiento mediante Transparametrización
	Clase Ia	Clase Ib	Clase II	TOTAL	
PARCOR	9	19	0	28	MFCC
R_0	3	2	0	5	Energía
Ganancia excitación {GS,P0}	8	12	0	20	-
Códigos VSE (librería H)	0	18	10	28	
Códigos VSE (librería I)	0	21	7	28	
Indicador de interpolación	0	1	0	1	
Modo	2	0	0	2	
TOTAL	22	73	17	112	

Tabla III-5. - Distribución de los bits de protección del codificador de canal entre los distintos parámetros codificados en el codificador GSM-HR (modo 0 –sordo–) y su correspondencia con los utilizados en reconocimiento mediante transparametrización.

De la misma forma que antes, la clase Ia se protege con un CRC de 3 bits. Después, estos 25 bits, junto con los 73 de clase Ib y una secuencia de 6 bits de cola (es decir, 121 bits), se introducen en un codificador convolucional. A los 211 bits de salida de este codificador se le añaden los 17 bits sin protección de clase II y se procede a su

reordenación y división en bloques para el entrelazado. Finalmente, los 228 bits se subdividen en 4 bloques de 57 bits que se transmiten en 4 ráfagas consecutivas [45].

Parámetro Codificado	Nº de bits / trama				Parámetro para Reconocimiento mediante Transparametrización
	Clase Ia	Clase Ib	Clase II	TOTAL	
PARCOR	10	18	0	28	MFCC
R_0	3	2	0	5	Energía
Ganancia excitación {GS,P0}	0	20	0	20	-
Retardo LTP	7	13	0	20	
Códigos VSE (librería I)	0	19	17	36	
Indicador de interpolación	0	1	0	1	
Modo	2	0	0	2	
TOTAL	22	73	17	112	

Tabla III-6. - Distribución de los bits de protección del codificador de canal entre los distintos parámetros codificados en el codificador GSM-HR (modos 1-3 –sonoros–) y su correspondencia con los utilizados en reconocimiento mediante transparametrización.

Dos tipos de indicadores determinan la importancia de los errores detectados en el destino: el indicador de trama errónea (BFI –*Bad Frame Indicator*–) y el de trama poco fiable (UFI –*Unreliable Frame Indicator*–). Si el indicador BFI está activado, el decodificador utiliza el procedimiento de substitución y silenciamiento (en el caso en el que haya varias tramas consecutivas erróneas) como veremos en III.3.1.2.

La Tabla III-5 y Tabla III-6 nos muestran de nuevo cómo se encuentra distribuida la protección del codificador de canal, pudiendo extraerse las mismas conclusiones que en el caso del codificador FR. Es reseñable, sin embargo, que a diferencia del codificador anterior, en éste se codifica de forma explícita el parámetro de energía R_0.

III.3.1.1.3. Enhanced Full Rate

La codificación de canal es idéntica a la del FR, salvo por la codificación preliminar en la que se emplean los 0,8 Kb/s adicionales que deja el codificador EFR. Más concretamente, esta codificación preliminar consiste en un CRC de 8 bits incluido 2 veces [45].

Para este codificador, la Tabla III-7 muestra, una vez más, la distribución de la protección del codificador de canal para los diversos parámetros de codificación y su equivalente en parámetros de reconocimiento, donde también los parámetros espectrales y los que, de forma implícita, codifican la energía se transmiten de forma más segura.

Parámetro Codificado	Nº de bits / trama				Parámetro para Reconocimiento mediante Transparametrización
	Clase Ia	Clase Ib	Clase II	TOTAL	
LSP	18	20	0	38	MFCC
Ganancia librería estática	2	18	0	20	Energía
Ganancia LTP	2	14	0	16	
Retardo LTP	28	2	0	30	
Códigos algebraicos	0	70	70	140	-
CRC (adicional)	0	8	0	8	
Bits repetidos (códigos de protección)	0	0	8	8	
TOTAL	50	132	78	260	

Tabla III-7. - Distribución de los bits de protección del codificador de canal entre los distintos parámetros codificados en el codificador GSM-EFR y su correspondencia con los utilizados en reconocimiento mediante transparametrización.

III.3.1.1.4. Adaptive Multi-Rate

La codificación de canal, en este caso, también utiliza la misma secuencia de procedimientos que en los codificadores anteriormente descritos, aunque empleando distintas particiones de los bits de la trama entre las tres clases Ia, Ib y II, ya mencionadas, y también distintos niveles de redundancia, dependiendo de la tasa binaria utilizada. Además, los datos que se envían "en banda" (*–in band–*) para indicar el modo o la tasa binaria empleada también se protegen con un código bloque [45]. Huelga decir que todo lo dicho anteriormente para otros codificadores aplica también a éste.

III.3.1.2. Soluciones orientadas a la recuperación de la señal de voz desde el receptor

Cuando el contenido de una trama es errónea a pesar de las correcciones que haya podido llevar a cabo el codificador de canal, muchas veces es preferible sustituirla por otra sintética, y por ello es preciso que existan procedimientos de recuperación de tramas que especifiquen cuál es la mejor manera de hacerlo.

Muchos codificadores especifican la forma de actuar en estos casos durante el proceso de decodificación; estos procedimientos van encaminados, por una parte, a evitar la reproducción de ruidos o artefactos desagradables, y por otra, a evitar que el decodificador se "desenganche" en los codificadores basados en análisis mediante síntesis. Las estrategias que más se emplean son las siguientes:

- Repetición de los parámetros LPC de la trama inmediatamente anterior ([47], [3], [30], [161]): en algunos casos, como en [123], la memoria del filtro se actualiza de forma que prediga exactamente la trama anterior y en otros, como en el codificador IS-641, los parámetros LSP de la trama errónea se sustituyen por una combinación de los parámetros de la trama anterior y el valor medio de los mismos (obtenido empíricamente a partir de un conjunto de entrenamiento) [112].

- Expansión progresiva del ancho de banda: de forma que si el número de tramas consecutivas erróneas es muy grande, la envolvente tiende a ser plana ([161], [3]).

- Atenuación progresiva de la ganancia: de nuevo pretendiendo que al cabo de un número de tramas consecutivas erróneas el decodificador enmudezca([161], [120], [3], [88], [175]).

- Repetición del periodo fundamental en los casos de excitación sonora, o estimación del mismo a partir de tramas anteriores y generación de un vector aleatorio en los casos de excitación sorda ([161], [120], [3], [88], [123]).

Concretamente, en el caso del codificador FR, si una trama se deteriora sustancialmente, se sugiere su substitución y silenciamiento (esto último en el caso en el

que haya varias tramas consecutivas erróneas), pero no se especifica cómo [49]. En el caso del codificador HR, dos tipos de indicadores determinan la importancia de los errores detectados en el destino: el BFI (*–Bad Frame Indicator–*) y el UFI (*–Unreliable Frame Indicator–*). Si el indicador BFI está activado, el decodificador utiliza un procedimiento de substitución y silenciamiento (de nuevo en el caso de varias tramas consecutivas erróneas). La sustitución de tramas consiste, la mayoría de las veces, en la repetición del último filtro LP válido (quizás con expansión del ancho de banda de los formantes) y la síntesis de una excitación artificial, aunque de nuevo esto último no es normativo. Si el indicador UFI se activa, el decodificador realiza un test de verosimilitud de los parámetros recibidos, es decir, se comparan algunas propiedades de la señal de salida reconstruida con las de la última trama correctamente recibida y en el caso de diferencias notables se modifica la salida para limitarlas [51].

Hay que hacer notar que, como es lógico, los procedimientos que se acaban de describir no están diseñados para mejorar el comportamiento de un reconocedor automático, sino para conseguir una reproducción perceptualmente aceptable a pesar de los errores y con un retardo razonable para la telefonía. Por este motivo no se proponen soluciones de interpolación que consideren tramas posteriores a la perdida, además de las anteriores. Esta solución, por ejemplo, es perfectamente viable en el caso de reconocimiento, como veremos en la sección IV.1, donde la restricción sobre el retardo no es tan fuerte.

En todo caso, nosotros hemos contemplado este tipo de soluciones y nuestro sistema toma como punto de partida estos procedimientos, que también están presentes en nuestro sistema de referencia (sección V.1) de forma que podamos comparar el sistema basado en la decodificación de la señal con el nuestro.

III.3.1.3. Soluciones específicas para reconocimiento

Algunos de los autores que se han ocupado de la distorsión de codificación y que citamos en III.2, también han tratado, en cierta medida, los errores de transmisión que se producen en GSM y su influencia a la hora de realizar reconocimiento automático. Es el caso de Dufour et al.[40], que evalua la parámetrización RN_LFCC (*–Root-Normalized Linear Frequency Cepstral Coefficients–*), considerando los codificadores GSM-HRy GSM-FR (con sus correspondientes codificadores de canal) y errores de transmisión; en particular insertan un dispositivo para introducir errores que simula una situación de un 90% de cobertura, un nivel de potencia de 7dB sobre interferencias de canal y un 8% de errores. En nuestra opinión, sería conveniente realizar una mayor cantidad de experimentos para poder validar y extraer consecuencias claras acerca de los beneficios de esta parametrización en estas condiciones.

Karray et al, por su parte, proponen una estrategia basada en la clasificación de las tramas erróneas mediante un modelo de “basura” (*garbage*), utilizando para ello una base de datos etiquetada manualmente que considera esta eventualidad. Son etiquetadas como “basura” las zonas de la señal en las que ésta se encuentra tan deteriorada que es más conveniente no realizar el reconocimiento (producen inserciones y sustituciones). Para el caso particular de voz procedente de entornos GSM existen, además, “agujeros” (*holes*) que son detectadas automáticamente (siempre sobre la voz decodificada)

basándose en propiedades estadísticas de la señal, tales como la amplitud de la señal, su varianza y la longitud de los segmentos y corresponden, presumiblemente, a los casos en los que está presente una ráfaga de errores en el canal lo suficientemente larga como para que afecte a varias tramas consecutivas, de forma que el codificador renuncie a reparar la trama y enmudezca completamente. Una vez han detectado estos agujeros, los autores proceden a aumentar la probabilidad de que sean clasificados como "basura" y por tanto rechazados a la hora de realizar el reconocimiento (*hole rejection*). De esta forma y para un mismo nivel de "aceptaciones falsas" disminuyen los errores de sustitución y "falso rechazo".

Los trabajos de Gallardo et al. ([67], [65] y [66]) son los primeros que consideran la existencia de errores en el canal GSM cuando se utiliza la aproximación de reconocimiento mediante transparametrización. De hecho, los autores afirman que parte de los beneficios obtenidos en reconocimiento al utilizar este esquema en presencia de errores proviene de que sólo una porción de los bits recibidos se utiliza para el reconocimiento (sólo aquellos que describen la envolvente espectral); si los bits que resultan contaminados no son los que describen la misma, el reconocedor no se ve afectado. Esto no se cumple en el caso extraer los parámetros de reconocimiento de la voz decodificada, puesto que en su decodificación se emplean todos los bits y por lo tanto cualquier error, independientemente del parámetro al que afecta, provoca una distorsión y por tanto un deterioro de las tasas de reconocimiento.

En estos estudios preliminares, sin embargo, no se tienen en cuenta, por ejemplo, los efectos del codificador de canal o un modelo realista del mismo, ya que se insertan de forma aleatoria, cuando se pretende obtener errores aislados, o con un modelo simple de dos estados, como el descrito en V.1.3, cuando el objetivo es obtener errores "a ráfagas". Tampoco es considerada la importancia de los diferentes periodos de trama que imponen los codificadores realizándose siempre el reconocimiento utilizando el periodo de trama que ofrece el codificador de voz. Además, a pesar de los argumentos que se utilizan en contra de la decodificación de la señal, no se proporciona una alternativa para extraer la energía, en el caso del codificador FR que no transmite este parámetro individualmente. En el Capítulo IV, expondremos las soluciones que hemos adoptado para solventar estos problemas.

Otro trabajo posterior que también considera los errores de transmisión cuando se utiliza el esquema de reconocimiento mediate transparametrización es de Kim et al. ([112]). Estos autores consideran que la parametrización que ellos proponen (véase III.2) está lo suficientemente protegida por el codificador de canal, como para no tener que preocuparse de los errores aislados que se puedan producir sobre ella, de forma que sólo consideran los casos en los que el error es tan devastador, que la trama al completo debe ser descartada. Esto no es del todo cierto por dos motivos: en primer lugar, en la parametrización que proponen se incluye un parámetro de energía que se obtiene sintetizando la excitación, cuyos parámetros constituyentes son precisamente los más desprotegidos por el codificador de canal y en segundo lugar, la protección de dicho codificador de canal sólo se extiende a los bits más significativos del resto de los parámetros que consideran, por tanto, a pesar de que estamos de acuerdo en que el codificador de canal beneficia en gran medida a la aproximación de reconocimiento mediante transparametrización, creemos que el efecto de los errores que no llegan a producir el descarte de la trama implicada también debe ser tenido en cuenta.

En todo caso, estos autores proponen dos soluciones para la pérdida de tramas: el uso del algoritmo de extrapolación propuesto en el estándar y que ya describimos en III.3.1.2, y la eliminación de las tramas borradas o descartadas, método que puede ser interpretado en términos de parametrización con frecuencias de trama variables (*–variable frame rate analysis–*). Ambas aproximaciones arrojan resultados de reconocimiento similares.

Como ya habíamos comentando en III.3.1.2 nuestro sistema de reconocimiento también contempla el mecanismo de recuperación que se propone en cada estándar considerado, si bien el mecanismo de extrapolación que utiliza el estándar IS-641 parece algo más sofisticado.

III.3.2. Entorno IP

La falta de adecuación de las redes IP al tráfico con requerimientos de tiempo real, como es el caso de la voz, se traduce, entre otras cosas, en pérdidas de paquetes o borrados de tramas (*–frame erasures–*). Estas pérdidas se deben básicamente a dos motivos:

1. En primer lugar, es habitual que debido a problemas de congestión en los nodos de la red, éstos descarten los paquetes de los que no pueden hacerse cargo. Esto forma parte de la estrategia empleada para regular el flujo de tráfico en estos nodos y la retransmisión de estos paquetes es el procedimiento previsto en el protocolo TCP para su recuperación. Por supuesto, este método no es viable en el caso del tráfico de voz, con estrictos requerimientos de tiempo real, como veremos a continuación.

2. El segundo motivo tiene que ver con el retardo introducido por el transporte a través de estas redes, pero especialmente con la variación –impredecible– de este retardo o jitter que afecta a los distintos paquetes por el hecho de proceder de distintos caminos o seguir rutas diferentes, como ocurre en el caso de las redes IP. Este retardo obliga a la creación de un *buffer* o registro de entrada que almacene y reordene estos paquetes antes de su reproducción, pero desgraciadamente, debido precisamente a la variación del retardo durante la transmisión, no es posible fijar de antemano un valor para el tamaño de dicho registro. Se adopta, por tanto, un compromiso entre el retardo que se está dispuesto a admitir en la aplicación correspondiente y el número de paquetes que no llegarán a tiempo para su reproducción.

Como es lógico, estas pérdidas influyen negativamente no sólo en la precisión de los reconocedores automáticos, sino también en la calidad de la voz decodificada. Es por esto, que para evitar estas pérdidas de paquetes en las transmisiones de tráfico con requerimientos de tiempo real, se están realizando grandes esfuerzos desde el punto de vista de la creación de nuevos protocolos y adecuación de los ya existentes, como veremos en III.3.2.1. Sin embargo, la adopción de estos protocolos en la totalidad de las redes IP es un proceso muy lento, y por este motivo se han diseñado procedimientos para aliviar la distorsión, similares a los que podemos encontrar en GSM y que revisaremos brevemente en III.3.2.2. Sin embargo, como ya apuntábamos entonces, estas técnicas están orientadas a recuperar una señal de voz perceptualmente aceptable,

lo cual no tiene por qué ser exactamente lo mejor para un reconocedor automático; las soluciones específicas para reconocimiento automático las expondremos en III.3.2.3. De hecho, una restricción importante sobre estos mecanismos de recuperación viene impuesta por la necesidad de reproducir la señal con un retardo máximo determinado. Sin embargo, en el caso del reconocimiento automático, pueden aceptarse retardos superiores; y por el contrario, el hecho en sí de la pérdida de paquetes resulta más devastador que para un reconocedor humano. Discutiremos estas diferencias con más detenimiento en IV.1

III.3.2.1. Soluciones orientadas a evitar la pérdida de paquetes

Ya hemos expuesto cómo la pérdida de paquetes en aplicaciones de telefonía IP no sólo se debe a su extravío en los nodos de la red, sino que el hecho de que la información de voz deba ser reproducida con unas exigencias de tiempo determinadas añade, a los paquetes realmente perdidos todos aquellos que por los motivos que sea, no han llegado a tiempo. Así, el servicio de voz sobre redes IP, se enfrenta al reto de proporcionar una calidad del servicio (QoS –*Quality of Service*–) al menos equiparable a la que ofrece actualmente la red telefónica conmutada. Los principales factores que influyen en esta calidad de servicio son:

1. **Retardo extremo a extremo**, es decir, la diferencia entre el instante de tiempo en que un paquete es recibido en el receptor y el instante en el que fue transmitido. Hay que señalar aquí, que aunque el retardo extremo a extremo es un parámetro fundamental para la obtención de una QoS aceptable desde el punto de vista de un usuario de telefonía IP, no es un factor tan crítico en el caso de reconocimiento de habla, ya que habitualmente, los retardos tolerables en este tipo de aplicaciones suelen ser superiores. Discutiremos con más detalle estas diferencias en IV.1.

2. **Variación del retardo o *jitter***, que proviene del hecho de que cada paquete puede seguir una ruta distinta y por lo tanto su retardo verse afectado por un retardo diferente. Esto plantea un problema muy grave puesto que es imposible predecir el valor de este retardo y resulta difícil dimensionar el registro de entrada, previo a la decodificación.

3. ***Throughput*** o fracción del ancho de banda nominal que se usa realmente para transportar tráfico. Si la red es compartida por varios usuarios, cada uno obtiene una fracción del throughput de la red. La forma de repartir este recurso depende del tipo de red, mientras que el ancho de banda requerido para la transmisión de voz depende del tipo de codificador que se utiliza, el número de tramas por paquete y de otros aspectos tales como si se utiliza compresión de las cabeceras RTP o no [34].

4. **Pérdida de paquetes**: además de los fallos ocasionales que pudieran darse en los enlaces, nodos y demás componentes de cualquier red, las redes IP admiten que un nodo descarte paquetes en caso de congestión, confiando en la retransmisión posterior de los mismos (protocolo TCP) porque en esto consiste

su procedimiento de control de congestión. De esta manera se comunica a los emisores esta situación para que actúen en consecuencia. Es evidente que, en el caso de transmisión de voz, donde se considera un paquete irremisiblemente perdido, cuando no ha llegado antes del instante de su reproducción, este procedimiento no resulta adecuado.

5. **Disponibilidad y fiabilidad**: la primera suele medirse en términos del porcentaje de tiempo en que el servicio no funciona por algún motivo; y la segunda delimita el tiempo máximo que transcurre, en caso de fallo, hasta que el servicio se reanuda de nuevo.

6. **Seguridad**, que en este caso tiene que ver con la privacidad, la confidencialidad y la validación de clientes y servidores.

La importancia de cada uno de estos factores es distinta en función de la aplicación de que se trate, y no todos ellos influyen la pérdida de paquetes a la que aludíamos en esta sección y que constituye la principal fuente de distorsión, junto a la de codificación que hemos considerado para reconocimiento.

Sin embargo, a continuación presentaremos los protocolos que se están desarrollando para mejorar el conjunto de todos ellos y que, en todo caso, configuran el entorno en el que se realizará el reconocimiento de habla. Dedicamos, entonces, las próximas secciones a explicar de forma breve las características más sobresalientes de estos protocolos, con el objetivo añadido de dar una idea de la envergadura de esta solución y el espacio de tiempo que requerirá su puesta en funcionamiento.

III.3.2.1.1. Protocolos TCP/IP y UDP/IP

El protocolo de red de internet, IP (–*Internet Protocol*–), se limita a hacer el "mayor esfuerzo" para transmitir los paquetes de datos de un extremo a otro de la red; sin embargo, si algo va mal y el paquete de datos se corrompe o se pierde este protocolo no prevé ningún mecanismo de recuperación.

Para recuperarse de estos fallos el protocolo de transporte utilizado para la transmisión de datos de tipo elástico o sin requerimientos de tiempo real es el TCP (–*Transport Control Protocol*–). Este protocolo reside en las máquinas origen y destino y sus funciones principales incluyen entrega fiable en destino y control de flujo. Para desarrollar la primera función, cuando la máquina destino recibe un mensaje correctamente envía un mensaje de conformidad a la máquina originaria. Si después de un tiempo establecido ésta no recibe tal conformidad, procede al reenvío del paquete en cuestión.

El control de flujo se realiza de la siguiente manera: como hemos dicho, cuando una red se encuentra muy congestionada y los nodos se quedan sin memoria para almacenar los nuevos paquetes que van llegando, sencillamente los descartan. Esto lo detecta la máquina originaria del tráfico y a partir del patrón de mensajes de conformidad recibidos y es capaz de moderar su flujo.

Sin embargo, como se puede deducir de lo anteriormente expuesto, este protocolo aumenta la carga de la red con el propósito de ofrecer una fiabilidad. Esto que resulta ventajoso en el caso del tráfico elástico, cambia radicalmente cuando trata de tráfico

ineslástico, como el generado por la voz; en este caso, el interés en que se recupere una trama está supeditado a que se recupere en el tiempo adecuado; es decir, no se consigue nada recuperando una trama de voz que ya ha tenido que ser reproducida con anterioridad.

Por este motivo, en estos casos, se utiliza como protocolo de transporte UDP (*–User Datagram Protocol–*), que no contempla ninguna función de conformidad ni control de flujo, que si bien no mejora el problema de la pérdida de paquetes en los nodos, resulta más ligero (por la menor longitud de su cabecera).

Estas características de los protocolos TCP/IP o UDP/IP que acabamos de describir, explican por qué estas redes son tan inadecuadas para el transporte de tráfico inelástico. Sin embargo, merece la pena que nos detenengamos a describir uno de los octetos, denominado "Tipo de Servicio" (ToS *–Type of Service–*), que constituyen la cabecera del protocolo de red:

El objetivo original del mismo era precisamente etiquetar los paquetes especificando el tipo de servicio que requerían: tres de los bits estaban destinados a codificar su prioridad y cuatro más a especificar parámetros de servicio que orientaran a los nodos de la red a la hora de encaminar los paquetes, tales como requerimientos de retardo, *throughput*, fiabilidad o coste económico. Sin embargo, muy pocos encaminadores son capaces de actuar en consecuencia, ya que no hay muchas opciones para tratar el tráfico de forma distinta, y además el tráfico de datos, mayoritario hasta hace poco tiempo, no es muy exigente en estos aspectos. Por este motivo, este octeto no fue tenido en cuenta prácticamente en ningún caso; sin embargo, con el incremento del tráfico inelástico en la red, este octeto ha cobrado un nuevo protagonismo y ha sido redefinido por el grupo de trabajo de la IETF (*–Internet Engineering Task Force–*) de "servicios diferenciados" (diffserv *–differenciated services–*) e incluido en la nueva versión del protocolo IP, IPv6. Más adelante nos referiremos a este grupo de trabajo y a la nueva definición de esta etiqueta; pero ahora exponemos otros cambios que se han realizado en esta nueva versión del protocolo de red de cara a proporcionar QoS.

La aparición de la versión 6, está motivada por el agotamiento del espacio de direcciones previsto en la versión 4 debido al crecimiento espectacular de Internet; para evitar que este problema volviera a ocurrir de nuevo, las nuevas direcciones IPv6 tienen una longitud de 128 bits, en vez de las antiguas de 32, y además se ha diseñado un nuevo esquema de direccionamiento que permite "auto-configurar" las direcciones, y por tanto, que estas direcciones pueden sean asignadas de forma dinámica.

Pero además, la nueva versión del protocolo se ha diseñado teniendo en cuenta las demandas del tráfico inelástico, y entre otras cosas incluye un identificador de flujo (24 bits) que, aunque no prevé ningún tipo de acción y presumiblente tenga que ser utilizado en conjunción con el protocolo RSVP, que describiremos más adelante, permite identificar paquetes pertenecientes a un mismo flujo, para su posterior tratamiento. Otra de las modificaciones que puede contribuir a mejorar la QoS es la inclusión del campo de "encaminamiento", que permite especificar la preferencia de una ruta concreta para cada uno de los paquetes. De esta manera se puede especificar el encaminamiento concreto que proporcione menor retardo (por la razón que sea) frente al que pueda existir en las tablas del encaminador. Por último, otra de las mejoras consiste en la

posibilidad de encriptar los datos dentro de este protocolo, lo que ofrece la posibilidad de proporcionar conversaciones privadas de forma segura.

A estas ventajas que proceden de la especial atención al tratamiento del tráfico inelástico que se ha puesto en esta nueva versión, el nuevo diseño, más eficiente, también contribuirá a la mejora del funcionamiento de estas redes. En concreto el número de campos en IPv6 es menor que en IPv4, a pesar de que la longitud total de la cabecera es el doble, y esto acelera el encaminamiento puesto que con menos campos la complejidad decrece. Además, se ha eliminado el campo de opciones, ya que las opciones se implementan como cabeceras opcionales; esto tiene la ventaja de hacer más eficiente el tratamiento de las cabeceras, así como evitar restricciones acerca del tamaño de las opciones. Esto es así porque generalmente los encaminadores no tienen por qué leer todas las opciones; así, por ejemplo, la etiqueta de encaminamiento a la que aludíamos anteriormente se implementa como una cabecera opcional.

III.3.2.1.2. Servicios integrados (Protocolo RSVP) y servicios diferenciados

La "arquitectura de servicios integrados" (ISA –*Integrated Services Architecture*–). es un conjunto de normas que desarrolla la IETF para permitir que las redes IP proporcionen garantías de calidad de servicio y consta de dos componentes principales:

- Funciones de control de la QoS: se basa en encaminadores y servidores capaces de discernir las diferencias entre distintos niveles de servicio. Por ejemplo, el servicio de "carga controlada" utiliza el control de admisión para identificar y priorizar los paquetes de un determinado tipo con el objetivo de proporcionar un servicio equivalente al que daría la red cuando no está congestionada, pero con la filosofía de "mayor esfuerzo", esto es, los valores de retardo y disponibilidad siguen siendo impredecibles. En el servicio "garantizado", existe un compromiso de provisión de un servicio con unas determinadas características.

- Un método mediante el cual los usuarios sean capaces de comunicar a los nodos de la red sus requerimientos, implementado a través del protocolo RSVP (–*Resource reSerVation Protocol*–). Este protocolo se encarga principalmente de hacer reservas de recursos en los nodos que forman parte del camino que deben seguir los paquetes de voz. Sin embago, este protocolo tiene un grave problema de escalabilidad y por ello no ha tenido una muy buena acogida. Este problema proviene de que cada encaminador debe mantener un estado por cada flujo que esté soportando, al mismo tiempo que recibe constantes mensajes de actualización o de refresco de cada uno de esos flujos. Esto hace que la situación sea insostenible, sobre todo en los encaminadores troncales cuando el tamaño de las redes aumenta.

Por su parte, el grupo de trabajo de "Servicios diferenciados" (diffserv –*differenciated services*–) propone aplicar en los nodos un modelo basado en una clasificación de los paquetes, permitiendo agregar flujos de tráfico y evitando así el problema de escalabilidad de RSVP expuesto. Esta solución está basada en la redefinición del octeto ToS al que aludíamos antes, y se define el campo DS (–

Differentiated Service–) tanto para la versión 4 como para la 6, dejando 2 bits sin utilizar. Este campo permite especificar en el paquete IP un dominio que consiste en un número de encaminadores interconectados que manejan una misma definición de "comportamiento por salto" (PHB[1] *–Per Hop Behaviour–*) bien porque pertenezcan a un solo proveedor, o bien porque estén bajo el control de una misma autoridad administrativa. Para delimitar estos nodos existen una serie de nodos frontera, que se encargan de seleccionar y clasificar los paquetes según los designios de dicha autoridad.

III.3.2.1.3. Protocolo RTP/RTCP

El objetivo principal del protocolo RTP (*–Real-time Transport Protocol–*) y su correspondiente protocolo de control RTCP (*–Real-time Transport Control Protocol–*) es el de proporcionar regularidad y predictibilidad a las aplicaciones que hagan uso de tráfico inelástico.

Este protocolo aporta los mecanismos necesarios para soportar tráfico en tiempo real, como los de señalización de tiempo, detección de pérdidas, seguridad e identificación de contenidos. Esto es muy importante, porque al igual que en el entorno GSM se incluía un codificador de canal que permitía determinar la importancia de los errores introducidos por el canal, en este entorno es RTP el que permite determinar si un paquete se ha perdido o ha llegado fuera de secuencia.

Además, como ya se ha mencionado, debido al retardo variable que impone Internet, los tiempos de llegada de los distintos paquetes no suelen ser constantes. Para compensar este inconveniente, los paquetes entrantes se introducen en registros o *buffers* con un ligero retardo y, posteriormente, se entregan a una velocidad constante al software que genera la salida. Para que este esquema funcione, cada paquete ha de etiquetarse a fin de que el receptor pueda ordenar los datos entrantes en la misma secuencia temporal empleada por la fuente.

El protolo RTCP, por su parte, permite controlar los flujos de voz mediante el envío de informes, tanto al emisor como al receptor, acerca de la evolución de parámetros tales como medidas de tiempo entre llegadas consecutivas, *jitter*, número de paquetes perdidos, etc.

III.3.2.1.4. Protocolos H.323, SIP y MGCP

A la hora de realizar una llamada telefónica a través de una red IP, y sobre todo si además esta llamada requiere interactuar con otras redes (como la red telefónica convencional), es necesario que existan procedimientos para realizar varias funciones que van más allá de la mera transmisión de paquetes de voz, como por ejemplo todos

[1] En el modelo de DiffServ la decisión de cómo un paquete debe ser transmitido se toma en cada uno de los nodos, es decir, para cada salto entre encaminadores (*per-hop*). En función del código DS que obtenga cada paquete al entrar en un dominio DS el nodo tendrá con él un PHB distinto. Las definiciones de estos PHB, sin embargo, serán las mismas para todos los nodos pertenecientes a ese dominio.

los aspectos relacionados con el establecimiento y la finalización de la llamada (resolución de direcciones, locacización de pasarelas, etc.), la negociación de los parámetros bajo los cuales transcurrirá esa llamada, la generación y transmisión de la señalización de llamada, etc.

Para realizar estas funciones sobre redes de paquetes, la Unión Internacional de Telecomunicaciones (ITU –*International Telecommunications Union*–) había desarrollado el estándar H.323. Este estándar no se diseñó originalmente para operar sobre redes TCP/IP, pero su oportunidad, ya que era el único disponible cuando el negocio de la telefonía IP comenzó a desarrollarse, hizo que su uso se generalizara, a pesar de las numerosas críticas. IETF desarrolló con posterioridad SIP, cuya ventaja principal es la de estar específicamente diseñado para telefonía IP sobre redesTCP/IP y ser por lo tanto más ligero y también MGCP, más orientado hacia el establecimiento de redes de telefonía IP por grandes operadoras.

No es nuestra intención profundizar en las características de estos protocolos sino proporcionar una idea general de sus aspectos más relevantes.

III.3.2.1.4.1 Protocolo H.323

H.323 fue originalmente concebido para conferencia multimedia sobre redes de área local (LAN –*Local Area Networks*–). Sólo un subconjunto de los protocolos que constituyen esta especificación son necesarios para telefonía Internet, concretamente son aquellos que se ocupan del audio y que incluyen los codificadores de voz G.723.1 y G.729, (véase la sección III.1.2) y los que proporcionan el control de llamada.

Una vez establecida la comunicación, comienza el intercambio de información según haya sido negociado: el transporte de voz se realiza utilizando el protocolo RTP/RTCP sobre UDP y el intercambio de datos mediante el protocolo T.120, siempre entre los dos terminales. La señalización DTMF (–*Dual Tone Multiple Frequency*–) convencional se intercambia utilizando H.245, mientras que la señalización de llamada, utiliza los protocolos Q.931 y H.245.

III.3.2.1.4.2 Protocolo SIP

SIP se desarrolló en la IETF para aligerar el control distribuído de llamadas y las capacidades de negociación y es la principal alternativa a H.323.

Los detractores de H.323 aducen que, a pesar de que este protocolo haya ganado muchos adeptos por haber sido el primero en estar disponible, tiene muchos inconvenientes y carencias. El inconveniente principal que SIP trata de solventar es la gran complejidad y baja escalabilidad de H.323. El diseño de SIP es modular, es decir, las distintas funciones de las que se hace cargo como la señalización, localización y registro de usuarios, QoS, acceso a directorios, búsqueda de servicios, descripción de contenidos de la sesión, etc, están basadas en distintos protocolos ortogonales entre sí. Por el contrario, el diseño de H.323 es monolítico.

Además, los defensores de SIP afirman que el hecho de que este protocolo fuera específicamente diseñado teniendo en cuenta los aspectos relacionados con Internet, como integración con otros servicios de Internet, extensibilidad, modularidad y simplicidad, lo hacen más adecuado para la telefonía Internet. SIP ha sido adoptado,

además, en los estándares de la tercera generación de telefonía celular (–3gpp–) para las futuras redes “todo-IP”.

Al igual que H.323 y MGCP, SIP está preparado para trabajar con pasarelas que provean protocolos de señalización y traducción de medios a través de distintos segmentos de red. Estos segmentos pueden ser tanto la RTC con su señalización SS7 como MGCP o H.323.

III.3.2.1.4.3 Protocolo MGCP

MGCP (–*Media Gateway Control Protocol*–) es el protocolo de señalización en el que la filosofía sobre dónde debe residir la inteligencia de un servicio cambia radicalmente respecto a las otras dos opciones. En este caso la inteligencia reside en unidades mantenidas por el operador. Así, MGCP es un protocolo de control que permite a un controlador central monitorizar eventos en teléfonos IP y pasarelas, e instruirlos para enviar tráfico a direcciones especificadas. Esta opción está respaldada principalmente por varias operadoras de telefonía y cable y permite interoperar con los protocolos anteriores (H.323 y SIP), de forma que resulta beneficiosa para crear grandes redes o sistemas que finalmente utilicen H.323 o SIP para el establecimiento y finalización de la llamada.

En MGCP una pasarela de telefonía IP está formada por dos partes funcionales: una pasarela “tonta”, que traduce datos de audio, y un controlador de pasarelas “inteligente”, que se comunica con el resto del mundo a través de protocolos de señalización. Por supuesto, controlador y pasarela no tienen por qué estar situados en el mismo lugar geográfico.

Los detractores de este protocolo afirman que la filosofía maestro-esclavo de éste, va en contra de los principios de la arquitectura distribuída de internet.

III.3.2.1.5. Estrategias basadas en el emisor

Además de los protocolos que hemos esbozado en la sección anterior, y que por su naturaleza tardarán bastante en estar disponibles ampliamente, se pueden utilizar una serie de procedimientos que no se limitan a reparar los daños causados por la pérdida de paquetes, desde el receptor, como los que expondremos en la próxima sección, sino que interactuan con el emisor de varias maneras:

- Solicitando la restransmisión del paquete perdido.
- Enviando los paquetes entrelazados, es decir, cada grupo de *N* paquetes consecutivos se distribuyen en *M* paquetes que son los realmente transmitidos; si uno de estos *M* paquetes se pierde, sólo una porción de cada uno de los *N* paquetes se pierde y el error se distribuye.
- Empaquetamiento adaptativo en función de las características de la señal de voz. Así, pueden ser empaquetados en porciones más pequeñas los intervalos correspondientes a tramas sonoras, persiguiendo que la pérdida de una de ellas no tenga efectos muy perniciosos y en porciones mayores las trams sordas.

- Enviando información adicional, en los denominados mecanismos de "corrección de errores hacia adelante" o FEC (*–Forward Error Correction–*) basados en codificación redundante de varios tipos: envío de tramas completas en paquetes posteriores, pero codificadas con codificadores de muy bajo régimen binario; los codificadores de canal habituales en otros entornos, como los que describimos en III.3.1.1, que utilizan códigos de paridad, Reed-Solomon o CRC; o también envíos de varias tramas repetidas por distintos caminos de la red.

- Petición de resincronización al emisor; útil en casos como los del codificador G.729, cuyo funcionamiento no sólo depende de la integridad de las tramas recibidas sino del estado en que se encuentra (definido por las tramas anteriores, eventualmente erróneas).

III.3.2.2. Soluciones orientadas a la recuperación de la señal de voz desde el receptor

Sin ánimo de hacer una clasificación exhaustiva de las distintas estrategias de recuperación en el receptor, cuando nos enfrentamos a situaciones de "borrado" o desaparición de tramas (el lector interesado puede recurrir a [1], [10] o [149], por ejemplo) resaltamos a continuación las más empleadas:

- sustitución por silencio,
- interpolación de ruido,
- interpolación de paquetes,
- repetición del último paquete recibido,
- sustitución por porciones de voz de segmentos correctos inmediatamente anteriores y,
- repetición de la última forma de onda del periodo fundamental para segmentos sonoros, o, directamente, la correspondiente al paquete anterior.

En el caso particular del codificador G.723.1 la estrategia de recuperación frente a errores depende de una indicación externa de cuáles son las tramas perdidas. Esto se puede llevar a cabo fácilmente si se utilizan para el transporte protocolos como el RTP (véase III.3.2.1.3). Cuando el codificador está en modo "recuperación frente a errores", utiliza los LSP de las tramas previas para predecir los valores de la trama perdida y genera una excitación sintética, sonora o sorda, basada en una decisión tomada a partir de las tramas anteriores. Además, va atenuando la voz decodificada cuando se dejan de recibir dos tramas seguidas y la enmudece totalmente tras perderse tres tramas consecutivas.

El comportamiento del codificador G.729 en estos casos es similar: repite los parámetros LSP de la trama anterior y también las ganancias de las librerías adaptativa y fija (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 fija (seleccionada aleatoriamente). Sin embargo, 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.

Es importante señalar, sin embargo, que en la aplicación de estas técnicas hay un factor limitante que es el retardo. De hecho, la variación de retardo o *jitter* de la que adolecen estas redes tiene consecuencias muy graves, no sólo, como comentamos en III.3.2, por su contribución al aumento del número de paquetes perdidos, sino también por el retardo que necesariamente hay que introducir para combatirlo. Para contrarrestar estos efectos, se diseñan procedimentos para adaptar dinámicamente el tamaño de los registros o *buffers* que almacenan la información recibida. Esta adaptación se lleva a cabo durante los periodos sin señal que tienen lugar en las conversaciones ([159]). Como discutiremos en IV.1, en el caso de reconocimiento el factor de retardo no es tan restrictivo y por eso merece la pena considerar mecanismos de recuperación específicos que no estén limitados por este parámetros.

En todo caso, y como ya señalamos para los codificadores de GSM, hemos contemplado siempre los procedimientos de recuperación previstos en los estándares, como punto de partida de nuestras aproximaciones; asimismo siempre están presentes en el sistema de referencia con el que comparamos nuestros resultados.

III.3.2.3. Soluciones específicas para reconocimiento

Aunque hay algunos autores que han abordado el tema del reconocimiento a través de las redes IP, tal y como comentamos en II.4.2, existen muy pocas soluciones que consideren de forma explícita el problema de la pérdida de paquetes en lo que a reconocimiento se refiere.

Los estudios preliminares presentados en [148] constatan los problemas que supone para un reconocedor automático, trabajar sobre la base de una señal codificada con G.723.1, no sólo por la distorsión de codificación –bastante considerable, por los regímenes binarios que emplea–, sino sobre todo porque, debido a la tasa de tramas tan pequeña que utiliza (1 trama cada 30 ms), la pérdida de cualquiera de ellas supone un importante deterioro de las prestaciones del reconocedor, por no hablar de los casos en que estas tramas perdidas son consecutivas.

En [147], utilizamos técnicas de interpolación de tramas (véanse los detalles en V.2.2.3.2) para obtener el periodo de trama habitual de 10 ó 15ms (dependiendo del tipo de tarea de reconocimiento) . En el caso en el que una trama se haya perdido, permitimos que el procedimiento de recuperación de paquetes previsto en el estándar sustituya la trama perdida, tanto en el caso del reconocedor que actúa sobre la voz decodificada, como en nuestra implementación mediante transparametrización. Sin embargo, hay que destacar el hecho de que, debido precisamente a que la interpolación que llevamos a cabo no está sometida a las restricciones de retardo del codificador

(véase IV.1), la trama restaurada se suaviza no sólo con información de la trama anterior, sino también con las tramas posteriores. No obstante, en este trabajo sólo se evalúa este procedimiento sobre un sistema de reconocimiento de dígitos aislados (véase V.1) y las mejoras son modestas.

En [145], por otra parte, introducimos el procedimiento de estimación de la energía a partir de los parámetros del codificador (véase IV.5.1), de forma que se evita totalmente la decodificación de la señal de voz. Además, evaluamos satistactoriamente las prestaciones de nuestro sistema de reconocimiento tanto en la tarea de reconocimiento de dígitos aislados mencionada anteriormente, como en una tarea de reconocimiento de habla continua. Además, mejoramos la simulación del canal (véase V.1.3.2), adecuando los parámetros del modelo que utilizamos, de forma que genere patrones de pérdida de paquetes con las características de medidas reales realizadas sobre Internet por Borella ([15]).

Otras propuestas, que posiblemente se puedan aplicar al problema de pérdida de tramas en codificadores son las basadas en la teoría de la imputación (por ejemplo [128], [101] o [31]), aunque en estos casos no se han utilizado directamente sobre pérdidas de tramas completas, sino en situaciones de pérdidas de ciertas bandas frecuenciales o interrupciones temporales relativamente cortas.

III.4. Otras distorsiones

En esta tesis nos hemos ocupado de las distorsiones causadas por los sistemas de transmisión GSM e IP, que hemos identificado como las causadas por la codificación o compresión severa que se aplica para solucionar la escasez de ancho de banda y los errores introducidos por ambos canales de transmisión. Sin embargo, no queremos dejar de señalar que al trabajar en estos entornos, también aparecen otro tipo de distorsiones que afectan en gran medida a los resultados de reconocimiento y que están mayormente relacionadas con el ambiente en el que tiene lugar la generación de la señal de voz.

Así, y sobre todo en entornos móviles, puesto que precisamente debido a esta movilidad la comunicación puede tener lugar en prácticamente cualquier espacio, el ruido ambiente juega un papel fundamental; por ejemplo, los ruidos asociados al interior de vehículos están siendo profusamente investigados. En general, estos ruidos pueden agruparse entorno a tres factores: densidad espectral de ruido, 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; de ahí que se hayan propuesto muchas técnicas de reducción de ruido y de mejora de la señal de voz, como las basadas en deconvolución ciega para eliminar los efectos de un canal desconocido [134], las cuales, en muchos casos, benefician al reconocimiento automático. No obstante, diferentes problemas requieren diferentes soluciones y lo mejor en un caso no tiene porque resultar eficaz en el otro; en particular, es significativo el hecho de que en el reconocimiento automático sea más perjudicial el desajuste entre los datos de entrenamiento y de test, que hasta cierto punto, la calidad (es decir, lo contaminados acústicamente que estén) de los mismos.

Es mucho el esfuerzo dedicado a combatir el ruido, pero dado que en este trabajo no vamos a abordar esta problemática, nos limitaremos a clasificar en dos grandes grupos las técnicas desarrolladas a estos efectos:

- Las que reducen la sensibilidad de las características de la señal de voz a las posibles distorsiones, como parametrizaciones robustas como PLP (–*Perceptual Linear Prediction*–) [82], transformación de parámetros como los métodos de filtrado de parámetros espectrales [77] o la inclusión de parámetros dinámicos [64]. En IV.3 encontramos un análisis más detallado de algunos de estos métodos.

- Las de adaptación y compensación que pretenden reducir el impacto del desajuste entre el proceso de entrenamiento y el de test, ya sea sobre la señal, como en el caso de deconvolución ciega que acabamos de señalar, los parámetros, como el Filtrado Probabilístico Óptimo que proponen Salonidis et al. en [162] y que ya comentamos en III.2, o los modelos, como es el caso de los multi-HMM que proponen Karray et al. en [109]. Estos métodos, generalmente están basados en la existencia de datos de entrenamiento del nuevo entorno al que prentenden adaptarse.

En todo caso, en nuestra opinión muchas de estas técnicas son aplicables a la solución de reconocimiento mediante transparametrización que nosotros proponemos. Por su parte, como ya hemos comentado en III.2, Huerta et al. [86], verificaron el buen funcionamiento de una aproximación también basada en reconocimiento mediante trasparametrización, considerando ruido rosa aditivo. Además, Kim et al., han incorporado con éxito técnicas de mejora de la señal de voz en condiciones de ruido de coche y ruido "*babble*"[2] a su propuesta de reconocimiento mediante transparametrización.

[2] Murmullo, varias personas hablando a la vez como ruido de fondo.

Capítulo IV
Transparametrización de la señal de voz

En el Capítulo II comenzamos planteando el problema de la implementación en un servidor remoto que hiciera uso de la tecnología de reconocimiento automático de habla, y terminamos comparando las tres alternativas que son posibles desde el punto de vista de distribución de los subprocesos que componen el proceso de reconocimiento. A la vista de ellas, seleccionamos la de "reconocimiento remoto" como la más adecuada a nuestros propósitos y adelantamos, siempre ya dentro de esta opción, la mejora que supone respecto a su aplicación convencional, la sustitución de la decodificación y posterior parametrización de la señal por una transparametrización que nos llevara directamente desde los vectores que parametrizan la señal de voz para codificarla hasta los que la parametrizan para reconocimiento.

Dedicamos el Capítulo III a la descripción de los aspectos que conciernen al reconocimiento en los dos entornos –GSM e IP– en los que hemos aplicado este método e hicimos una revisión de las soluciones que otros autores proponen para paliar los problemas que provocan dichos entornos.

El objetivo del presente capítulo es describir cómo puede realizarse esta transformación y para ello comenzamos evaluando cuáles son las exigencias para la representación de la señal de voz que imponen, por una parte, la codificación, y por otra, el reconocimiento. Estos requerimientos configuran las características de los dos tipos de parametrizaciones entre las que pretendemos establecer un nexo, sin olvidar, desde luego, la influencia que tiene sobre la forma en la que se diseña esta caracterización, el entorno en el que van a ser aplicados.

Una vez discutidas las características de cada una de estas parametrizaciones, dedicamos las dos siguientes secciones a analizar las parametrizaciones más habituales en ambos campos, haciendo especial hincapié en los aspectos que tienen que ver con la robustez de las mismas en entornos adversos.

Finalmente, las dos últimas secciones se emplean en describir las transformaciones entre ambas parametrizaciones, primero brevemente, en el sentido de codificación a partir de parámetros de reconocimiento y después, de forma más detallada, puesto que se trata de la transformación que nos interesa, en el de reconocimiento a partir de parámetros de codificación. En este última parte discutiremos los aspectos más relevantes de esta transparametrización, contrastando nuestra opción con las de otros autores en lo que se refiere a la transparametrización de la envolvente espectral y la

energía, la tasa de tramas y la posibilidad de aplicar algún tipo de procesado sobre la parametrización para mejorar su robustez en los entornos en que nos centramos.

IV.1. Requerimientos para codificación y para reconocimiento: comparación

El objetivo de la codificación es encontrar una descripción compacta de la señal de voz que permita su transmisión o almacenamiento de forma más eficiente. Para ello, se utilizan algoritmos de eliminación de redundancia basados en la asunción de un modelo simplificado del mecanismo de producción de voz. Desgraciadamente, para los regímenes binarios en los que se trabaja habitualmente en los entornos de los que nos ocupamos, no sólo se elimina la información estrictamente redundante, sino que se lleva a cabo una codificación con pérdidas (perceptualmente hablando).

En el caso del reconocimiento, el objetivo de la parametrización es encontrar una representación de la señal que, por una parte, retenga la información discriminante necesaria que permita reducir la variabilidad no lingüística propia de un hablante particular, y por otra, se adecue al modelado y a los algoritmos que se utilizan para su entrenamiento. Además, los parámetros de reconocimiento deben ser robustos frente a distorsiones típicas, productos de las condiciones en que funciona el sistema de reconocimiento y se debe adecuar el número de parámetros del modelo a la cantidad de datos de entrenamiento que estén disponibles.

Además de las diferencias en los objetivos de estos dos tipos de procesado, el entorno en el que van a ser aplicados juega un papel fundamental en el diseño de los dos tipos de descripciones. Por ese motivo, exponemos a continuación las similitudes y diferencias de estos dos tipos de parametrización desde varios puntos de vista:

- Adecuación del modelo: ya comentamos en la sección III.2 que la distorsión de codificación se debe básicamente a dos factores; la cuantificación de los parámetros y la necesidad de asumir ciertos modelos, presumiblemente inexactos, para la descripción de la voz.

  En primer lugar, la codificación de voz a regímenes binarios medios a bajos, como los que se utilizan en los entornos que nos ocupan, parte siempre de un análisis por tramas con una duración de entre 10 y 30 ms. Tal modo de proceder obedece a la suposición de que la señal de voz es de naturaleza “quasi-estacionaria”, es decir, dentro de un periodo corto de tiempo de aproximadamente la duración de la trama, podemos considerar que las propiedades estadísticas de la señal se mantienen. Sin embargo, la duración de estos periodos de estacionariedad es altamente variable, lo cual contrasta con la rigidez de análisis por tramas equiespaciadas temporalmente que se lleva a cabo tanto en codificación como en reconocimiento.

  Para suavizar la rigidez de este análisis por tramas que no refleja la distribución no-uniforme de la información en el plano tiempo-frecuencia, existen diversos procedimientos tanto para la codificación, como para el reconocimiento. En el

primer caso, por ejemplo, se utilizan interpolaciones entre los parámetros espectrales de tramas consecutivas de forma que no existan transiciones bruscas que generen sonidos extraños o desagradables al reconstruir la señal. Otro ejemplo puede encontrarse en la predicción a largo plazo en las zonas sonoras, mediante la cual, se implica a tramas previas en la generación de las actuales.

En la parametrización para reconocimiento, sin embargo, el ejemplo más claro de empleo de información entre tramas es el uso de los parámetros denominados "delta" y "doble-delta" (o de "aceleración") (véase la sección IV.2.1), pero también las diversas técnicas de filtrado espectral, encaminadas principalmente a la obtención de parametrizaciones robustas incluyen información de varias tramas en la parametrización de la trama actual.

Otra de las suposiciones más habituales en procesado de voz, es la inspirada en el mecanismo de producción de la voz y que normalmente se denomina modelo "fuente-filtro" (*–source-filter–*); se estiman las contribuciones de la fuente y del tracto vocal (generalmente mediante predicción lineal) y se procede, después al modelado de cada uno de ellos, por separado. Profundizaremos un poco más en ese tema en IV.2.1.

Este modelo, según el cual se separa, la envolvente espectral de la estructura fina o excitación es también la base para la extracción de los vectores acústicos o parametrización, empleada en reconocimiento, donde habitualemente sólo se emplea la envolvente espectral o información a corto plazo.

- Compacidad: en el proceso de codificación de la señal de voz, uno de los parámetros más importantes es el régimen binario. Por ese motivo, los codificadores estándar definen, hasta el último bit, la cuantificación de cada uno de los parámetros que componen su descripción de la señal de voz. Así, es una decisión de diseño de suma importancia, el número de bits (o la porción del total de bits disponible) que se dedican, por ejemplo, a describir la envolvente espectral. Esto limita no sólo el número de parámetros LP con los que se implementa el filtro de síntesis, sino también la forma en la que se cuantifica cada uno de ellos.

En cuanto al proceso de reconocimiento, no es la compacidad un aspecto que tenga gran relevancia; por ello, el número de parámetros que se emplea en el vector acústico puede ser muy variable y el número de bits que se utiliza para cuantificarlos es más un problema de la precisión con la que se puede operar que un parámetro que intervenga en el diseño.

- Retardo: el análisis por tramas que tiene lugar en el proceso de codificación hace que inevitablemente exista un retardo algorítmico, que consiste en el periodo de duración de la trama más, en el caso de que exista, un *look-ahead* o anticipación correspondiente. De nuevo, el retardo vuelve a ser un factor limitante en el diseño de los codificadores de voz de los que nos ocupamos aquí, puesto que el retardo total, que incluye además, el retardo de transmisión y el de procesado, no debe superar ciertas cotas (entorno a 150 ms, considerándose 400 ms intolerable –[94]–) para que una conversación, por ejemplo, se pueda llevar a cabo con naturalidad y sin problemas de eco. Esto limita en gran medida la información de la que puede hacer uso a la hora de recuperar tramas erróneas o perdidas, ya que en esos casos la información perdida podría interpolarse a partir

de la de las tramas circundantes pero a diferencia de lo que ocurre en reconocimiento, donde el retardo admisible es mayor, en los codificadores esta interpolación se tiene que limitar a hacer uso de la trama anterior

- Carga computacional: la diferencia de carga computacional de ambos procesos es muy grande, siendo la de codificación mucho más liviana que la de reconocimiento. Por ese motivo, los codificadores que consideramos aquí se pueden implementar en dispositivos pequeños como terminales móviles, etc. y por ello en su día fueron optimizados también en este sentido de forma que la mayoría de las operaciones que utilizan son bastante sencillas. Por ejemplo, en los casos de interpolación entre tramas que comentábamos antes, los interpoladores suelen ser de tipo lineal. En el caso del reconocimiento, sin embargo, la carga computacional es mucho mayor y requiere dispositivos con una mayor capacidad y por lo mismo, permitiéndonos consecuentemente, la implementación de algoritmos más complejos.

- Robustez: varias de las decisiones sobre los tipos de parámetros que componen tanto la descripción del codificador, como la del reconocedor están motivadas por aspectos de robustez. Así, por ejemplo, los parámetros LP suelen sustituirse a la hora de cuantificarlos por otros tales como los PARCOR, los LAR o los LSP (véase en la sección IV.2 las propiedades de estos parámetros) debido a que estos resultan más robustos, tanto para los errores de cuantificación (donde la estabilidad del filtro de síntesis juega un papel preponderante) como los de transmisión.

Sin embargo, hasta que se introdujo (muy recientemente) la aproximación de reconocimiento distribuido (DSR –véase la sección II.4.2) no ha cobrado importancia la cuantificación de los parámetros de reconocimiento o el desarrollo de codificadores de canal específicos, todo lo contrario que el desarrollo de métodos robustos para reconocimiento cuando la voz de la que se parte está contaminada. En IV.3.5, nos ocuparemos de este tema.

IV.2. Parametrizaciones para codificación

En II.1.1 introdujimos, de forma general, el concepto de codificación como el de obtención, a partir de una señal de voz digitalizada $s[n]$ con $n=1,\cdots,N$, de una secuencia de vectores de parámetros $\mathbf{F}=\mathbf{f}_1,\mathbf{f}_2,\cdots,\mathbf{f}_R$ de tal forma que el error (perceptualmente ponderado la mayoría de las veces) entre la señal reconstruida, $s_{\hat{\mathbf{F}}}[n]$, a partir de los parámetros óptimos $\hat{\mathbf{F}}$, sea el mínimo posible.

Efectivamente, al definir la voz codificada como una secuencia de vectores de parámetros estamos enfatizando el hecho de que para realizar este proceso, se lleva a cabo una partición de la señal de voz en lo que se denominan tramas, es decir, $s[n]$ se enventana de manera que cada uno de los vectores, $\mathbf{f}_i$, que forman parte de esta secuencia codifica un segmento de la señal. Esto no quiere decir, que a la hora de optimizar el parecido entre la señal original y la codificada no se contemple ninguna

relación entre vectores $\mathbf{f}_i$ sucesivos, que de hecho existe en los codificadores que utilizan esquemas diferenciales, en el solapamiento entre ventanas de análisis sucesivas y en los procedimientos de suavizado entre tramas tales como la interpolación de los parámetros espectrales y la energía correspondientes a subtramas.

Sin embargo, salvo en la definición del alcance de la ventana de análisis, que muchas veces contempla un *look-ahead* que anticipa parte de la trama posterior, la mayoría de las relaciones entre tramas se limitan a la trama anterior, es decir, $\mathbf{f}_i$ sólo está relacionado con $\mathbf{f}_{i-1}$. El principal motivo de esta limitación es el retardo máximo aceptable, o en general en el tipo de aplicaciones en las que se utilizan los codificadores de voz. Esto, unido a la necesidad de suponer estacionariedad, que sólo se mantiene dentro de pequeños segmentos de la señal, conduce a que el análisis se realice de forma independiente en cada una de las tramas.

Para este análisis, la mayoría de los codificadores con regímenes binarios medios y bajos utilizan la predicción lineal (LP –*Linear Prediction*–). Por ello, la próxima sección estará dedicada a introducir este análisis que, aunque también de suma importancia en el problema de parametrización para reconocimiento, se introdujo primero como solución para codificación. Mas tarde y una vez explicado cómo se lleva a cabo la separación entre fuente y filtro, pasaremos a ocuparnos de la parametrización de cada una de estas dos componentes.

IV.2.1. Codificación por Predicción Lineal: aproximación fuente-filtro

En la generación de la voz humana intervienen dos elementos principales: una "excitación" producida por los pulmones que, en algunos casos haciendo vibrar las cuerdas vocales (sonidos sonoros) y en otros no (sonidos sordos), atraviesa un "filtro", definido por el tracto vocal y la cavidad nasal (en ocasiones), que conforma la envolvente espectral. Así, según este modelo la envolvente espectral de la señal de voz se puede caracterizar por la función de transferencia de un filtro, $H_i(z)$, que en el caso particular de la aproximación por predicción lineal toma la siguiente forma:

$$H_i(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{j=1}^{p} a_j z^{-j}} \qquad \textbf{(IV-1)}$$

donde G representa una ganancia y los a_j son los coeficientes de predicción lineal. Como se puede observar, este modelo supone que el filtro solamente contiene polos.

Si asumimos que la señal de voz, $s[n]$, es el resultado de excitar el filtro anterior con una cierta excitación, $x[n]$, la ecuación en diferencias correspondiente a la función de transferencia anterior queda como sigue:

$$s[n] = Gx[n] + \sum_{j=1}^{p} a_j s[n-j] \qquad \textbf{(IV-2)}$$

donde observamos que $s[n]$ es el resultado de añadir a la excitación anterior (con un cierto factor de escalado) una combinación lineal de las *p* muestras anteriores.

Por otra parte, nuestro objetivo es encontrar la forma de sintetizar una señal, $\hat{s}[n]$, de tal forma que el error cometido, $e[n] = s[n] - \hat{s}[n]$ sea el menor posible, es decir, obteniendo los coeficientes $\hat{a}_j$ que minimizan:

$$E = E\{e^2[n]\} = E\left\{\left[s[n] - \sum_{j=1}^{p} \hat{a}_j s[n-j]\right]^2\right\} \qquad \textbf{(IV-3)}$$

Calculando ahora las derivadas parciales de este error respecto a cada uno de coeficientes e igualando a cero se obtiene el siguiente sistema de *p* ecuaciones:

$$E\left\{\left[s[n] - \sum_{j=1}^{p} \hat{a}_j s[n-j]\right] s[n-i]\right\} = 0, \quad para \quad i = 1, 2, ..., p \qquad \textbf{(IV-4)}$$

Cada una de estas ecuaciones se puede escribir de la siguiente manera:

$$\sum_{j=1}^{p} \hat{a}_j \phi_n[i, j] = \phi_n[i, 0] \quad para \quad i = 1, ..., p \qquad \textbf{(IV-5)}$$

donde ϕ_n es la función de autocorrelación, es decir,

$$\phi_n[i, j] = E\{s[n-i]s[n-j]\} \quad para \quad i = 1, ..., p \quad y \quad j = 0, ..., p \qquad \textbf{(IV-6)}$$

La resolución de estas ecuaciones pasa por reemplazar la esperanza de la ecuación (IV-6) por una estimación de la forma:

$$\phi_n[i, j] = \sum_{m} s[m-i]s[m-j] \qquad \textbf{(IV-7)}$$

para lo cual es necesario asumir estacionariedad y aplicar estas ecuaciones sólo en un corto periodo de tiempo, es decir, durante el periodo de tiempo en que las propiedades estadísticas se mantienen razonablemente constantes. Para la resolución del sistema de ecuaciones anterior los métodos más habituales son el de la autocorrelación y el de la covarianza, que se pueden consultar, por ejemplo, en [116].

Mediante este procedimiento y una vez elegido un orden de predicción *p,* la señal de voz queda separada en dos señales como habíamos previsto: la primera de ellas, caracterizada por los coeficientes $\hat{a}_j$ y la segunda por una señal residuo, $e[n]$, que corresponde a la señal de excitación ideal, $x[n]$, salvo una constante *G* que algunas veces se codifica como un parámetro individual y otras, como parte de la excitación

(véase la sección IV.5.1). En los dos próximas secciones nos ocuparemos de la parametrización o representación de cada una de ellas.

IV.2.2. Parametrización del filtro

Los coeficientes $\hat{a}_j$ que se obtienen mediante este procedimiento son los denominados "coeficientes de predicción lineal" (LPC –*Linear Prediction Coefficients*–) y son los que se utilizan para describir, en la práctica, el filtro que modela el tracto vocal. Denotaremos el vector de parámetros de predicción lineal como $\hat{\mathbf{a}}^{(p)}$, donde el orden de predicción *p* que se utilizan habitualmente en los codificadores de regímenes binarios como los que nos ocupan aquí es de 10, aunque como hicimos notar en sus descripciones, el más antiguo de ellos, el codificador GSM-FR utiliza tan solo 8. Esta predicción, en la que se consideran un número moderado de muestras, inmediatamente anteriores a la que se está sintetizando en esos momentos, se denomina predicción "a corto plazo" (*short-term prediction*), en contraposición a la que se utiliza para codificar la periodicidad de los tramos sonoros de la señal, como veremos en la próxima sección.

Sin embargo, no son los parámetros LPC los que se codifican directamente, debido a que sus propiedades no son las más adecuadas para su cuantificación y a la dificultad de verificar la estabilidad del filtro con ellos sintetizado. Así, se consideran transformaciones biunívocas que dan lugar a otro tipo de parametrizaciones más adecuadas de las que nos ocupamos a continuación.

Los "coeficientes de correlación parcial" (PARCOR –*PARtial CORrelation*–), también denominados "coeficientes de reflexión" (RC –*Reflexion Coefficients*–) son los que se transmiten en el codificador GSM-FR. La ventaja de estos parámetros estriba en la facilidad con la que se comprueba la estabilidad de filtro ya que,

$$\hat{H}(z) \quad estable \Leftrightarrow |r_i| \le 1 \qquad 1 \le i \le p \qquad \textbf{(IV-8)}$$

donde r_i, son los coeficientes PARCOR y $\hat{H}(z)$ es el filtro "todo-polos" sintetizado por los coeficientes $\hat{a}_j$.

Debido a esta interesante propiedad estos coeficientes son a menudo calculados por los codificadores con el propósito de verificar la estabilidad antes de proceder a la síntesis de dicho filtro. Sin embargo, debido a que su sensibilidad espectral no es plana, es decir, valores de estos coeficientes cercanos a 1 requieren más bits para su cuantificación para obtener la misma precisión espectral, suelen utilizarse variantes de los mismos.

Así, los parámetros LAR (*Logarithmic or Log Area Ratio*) y los IS (*Inverse Sine*) aplican funciones no lineales, logaritmo y seno, respectivamente, sobre estos coeficientes, siendo los primeros los que transmite el codificador GSM-HR. Sin embargo, estos parámetros tienen dos importantes desventajas: en primer lugar, para minimizar la distorsión espectral son necesarios, al menos, 4 bits por coeficiente, lo cual puede resultar excesivo en algunos codificadores, y en segundo lugar, no revelan la correlación que existen entre tramas consecutivas, haciendo muy difícil su predicción.

Finalmente, los coeficientes LSP (*Line Spectral Pairs*) son los que utilizan los restantes codificadores GSM (GSM-EFR y GSM-AMR) y ambos codificadores IP (G.723.1 y G.729), por varios motivos: primero, que estos coeficientes sí que son altamente predecibles entre tramas consecutivas; esto los hace adecuados para una cuantificación predictiva, estrategia que siguen todos los codificadores anteriores, salvo G.723.1. Segundo, debido también a esta transición suave entre los coeficientes de tramas sucesivas, su interpolación con el objetivo de recuperar tramas perdidas o erróneas es más eficaz. Tercero, se trata de una representación en el dominio frecuencial, de forma que sería posible introducir conceptos relacionados con características del sistema auditivo humano de forma más sencilla. Finalmente, otra característica importante de esta parametrización es la posibilidad de realizar la verificación de la estabilidad del filtro de forma sencilla, ya que,

$$\hat{H}(z) \quad estable \;\Leftrightarrow\; 0 \le \omega_1 < \cdots < \omega_i < \omega_{i+1} < \cdots < \omega_p \le \pi \qquad \textbf{(IV-9)}$$

donde ω_i con $1 \le i \le p$ son los parámetros LSP directamente expresados como frecuencias.

Todas las parametrizaciones que hemos descrito aquí son equivalentes en el sentido de que siempre se pueden transformar unas en otras (véanse los algoritmos de transformación en [116], por ejemplo) y de la elección de unas u otras depende, como se desprende de lo anteriormente expuesto, la robustez frente a errores de cuantificación y de transmisión, la eficiencia de su cuantificación y la facilidad de su manipulación.

También es importante observar cómo en los codificadores más modernos se utilizan los parámetros LSP y, de hecho, dos de las transparametrizaciones que presentaremos en IV.5.1 parten directamente de dicha representación. Es preciso recordar asimismo, que son estos parámetros que codifican la envolvente espectral, los que contienen la información más relevante a la hora de realizar el reconocimiento automático, a pesar de su reducido número y de que normalmente se cuantifican con una cantidad relativamente pequeña de bits, si lo comparamos con los que se emplean para representar la fuente.

IV.2.3. Parametrización de la fuente

Para la parametrización de la señal de excitación se han diseñado estrategias muy variadas y aquí sólo pretendemos revisar someramente los procedimientos que se utilizan en los codificadores de los entornos GSM e IP que nos ocupan, con un énfasis especial en plasmar el tipo de información que se codifica y su relevancia en reconocimiento.

En general, el primer paso que se sigue a la hora de abordar este problema suele ser el de la división en subtramas de cada una de las tramas. Así, dependiendo del codificador, las longitudes de subtrama, L_{sf}, que se consideran están entorno a los 5 ms (salvo el G.723.1, cuya subtrama tiene una longitud de 7,5 ms).

Otra característica común en la codificación de la fuente o excitación es la distinción entre una parte periódica, característica de los tramos sonoros de la señal, y otra parte aleatoria, correspondiente a los tramos sordos. La primera de ellas se denomina habitualmente contribución adaptativa y la segunda, estocástica. Sin embargo, lo normal es no restringirse a ninguna de las dos opciones y generar excitaciones compuestas de la suma ponderada de estas dos contribuciones. Son entonces, las ganancias de cada una de estas contribuciones las que indican si el tramo de voz considerado es sonoro o sordo y por ello, Kim et al. ([112]) utilizaron estas ganancias para incluir dicha información en la parametrización acústica para reconocimiento. Los citados autores utilizan el codificador IS-641, que forma parte del estándar de telefonía celular norteamericano IS-136 y estas ganancias están codificadas de forma explícita en dos parámetros individuales, siendo esto lo más habitual, con algunas excepciones: el codificador G.729, por ejemplo, utiliza una estructura conjugada (CS *–Conjugate Structure–*) de dos etapas para cuantificar vectorialmente estos parámetros: en la primera la elección del vector correspondiente está sesgada para obtener una mejor aproximación de la ganancia adaptativa y en la segunda, de la estocástica. Además, en el codificador G.723.1 no existe una sola ganancia adaptativa, puesto que como veremos a continuación, utiliza un predictor de orden 4 para codificar esta contribución.

Para caracterizar la excitación adaptativa, $u[n]$, suele utilizarse un "predictor a largo plazo" (LTP *–Long Term Predictor–*) que, en general, toma la siguiente forma para cada subtrama *i*:

$$u[n] = \sum_{j=0}^{q} \beta_{ij} e[n - L_i - j], \qquad 0 \le n < L_{sf} \qquad \textbf{(IV-10)}$$

donde *q* es el orden de predicción, la mayoría de las veces 1, pero que, por ejemplo, en el codificador G.723.1 es de 4. Así, una vez decidido este orden de predicción, los parámetros que caracterizan esta contribución son, por una parte, el retardo, L_i, o periodo fundamental y las ganancias, β_{ij}, a las que aludíamos en el párrafo anterior.

La contribución estocástica, $v[n]$, suele estar compuesta por una serie de pulsos, de los que se codifican sus posiciones y amplitudes (normalizadas, generalmente, y con un parámetro de ganancia codificado de forma separada). Así, en el caso del codificador GSM-FR, la denominada "excitación por pulsos regularmente espaciados" (RPE – *Regular Pulse Excitation–*), consiste en un diezmado en el que se seleccionan una de cada 3 muestras del residuo o excitación deseada, de forma que una vez elegida la secuencia apropiada sólo es necesario codificar la posición del primero de ellos y las amplitudes de cada uno.

En el codificador GSM-HR, la excitación estocástica se construye como suma de vectores base (VSE *–Vector Sum Excitation–*) seleccionados en una librería de vectores fija. En particular, para los 3 modos sonoros esta librería, está compuesta por 9 vectores base (con los cuales es posible construir 2^9 vectores código de excitación utilizando palabras código de 9 bits). Para el modo sordo, debido a que no se considera la contribución adaptativa, se dispone de más bits para la codificación de esta excitación y por ello se emplean 2 librerías con 7 vectores básicos cada una.

El codificador G.723.1 consigue conmutar entre los dos regímenes binarios para los que está preparado generando un tipo de excitación distinta para cada régimen. En el

más alto la excitación es de tipo multipulso (MP-MLQ –*Multi-Pulse Maximum Likelihood Quantization*–) y se eligen las posiciones y amplitudes de 6 pulsos para las subtramas pares y 5 para las impares. Las posiciones de los pulsos están restringidas a las posiciones pares o las impares y esta opción se transmite con un bit al decodificador.

Cuando se escoge el régimen binario inferior, se utiliza para la codificación de la excitación estocástica se codifica mediante una librería algebraica (ACELP –*Algebraic Code-Excited Linear Predicition*–). Este tipo de excitación se usa también en el resto de codificadores que no hemos descrito todavía, es decir, en el GSM-EFR, los GSM-AMR y el G.729. Estas librerías definen una serie de pistas (*tracks*) que representan las posibles posiciones de un número de pulsos equiespaciados que generalmente toman los valores $\{+1,-1\}$. En general, la posición *k*-ésima de la pista *i* selecciona la posición *n* de la correspondiente subtrama,

$$n = kM + i, \qquad 0 \le n < L_{sf}, \quad 0 \le i < M, \quad 0 \le k < K \qquad \textbf{(IV-11)}$$

donde *M* representa el número de pistas y *K* el de posiciones equiespaciadas entre sí (*M* muestras) que contiene cada pista; así, para que todas las posiciones de la subtrama estén representadas en alguna pista se debe cumplir que $KM \ge L_{sf}$. En el caso de que $KM > L_{sf}$ las posiciones que no corresponden con ningún $n < L_{sf}$ se emplean para codificar la ausencia de pulsos, como ocurre en el G.723.1. Normalmente suele colocarse un solo pulso en cada pista, pero en el codificador GSM-EFR y en los GSM-AMR con regímenes binarios más altos es posible colocar hasta 2, que pueden incluso situarse sobre una misma posición, permitiendo entonces, que el pulso tome valores $\{+2,-2\}$.

Como sabemos, el número de bits que se emplean en codificar la excitación suele ser bastante mayor que el empleado para la envolvente espectral, que puede representarse de forma mucho más compacta. Además, dentro de los parámetros que codifican la excitación también se establecen prioridades a la hora de representar con más o menos precisión alguno de ellos. Concretamente, suelen estar representados de forma más precisa, el periodo fundamental, L_i, y las ganancias de las respectivas contribuciones adaptativa y estocástica. Esto es debido, desde luego, a que estos parámetros contribuyen de forma determinante a la calidad percibida.

Normalmente, aun cuando la inteligibilidad de la señal reconstruida es un objetivo primordial, los codificadores de voz tratan también de preservar aspectos de la señal que contribuyan, por ejemplo, a la identificación del locutor, de su estado anímico y otras características no lingüísticas que mejoran la naturalidad percibida en la reconstrucción. Es el caso, por ejemplo, del periodo fundamental, que aunque resulta ser una característica importante para el reconocimiento de habla en lenguas tonales (por ejemplo, [164]), no es, en general, un parámetro que suela considerarse en los vectores acústicos para reconocimiento. La discriminación entre tramos sonoros y sordos es también un factor que se ha utilizando en alguna ocasión como parte del vector acústico ([112]) y que, como hemos visto, puede considerarse implícitamente representado en las ganancias respectivas de las contribuciones adaptativa y estocástica.

Finalmente la energía, que sí que se encuentra entre los parámetros habituales en los vectores acústicos empleados para reconocimiento, no suele codificarse explícitamente;

de hecho, de entre los codificadores considerados, sólo el GSM-HR codifica la energía como un parámetro individual. En el resto de los casos, todos los parámetros que se codifican contribuyen de forma implícita en la codificación de la energía. En IV.5.1, presentaremos la forma de obtener una estimación de la misma basada tan solo en las ganancias de las contribuciones adaptativa y estocástica y el periodo fundamental, L_i, para los codificadores GSM-FR y G.723.1 (aunque el método es extrapolable a otros codificadores). No es casualidad que sea posible hacer esta estimación a partir, precisamente, de los parámetros mejor codificados entre los correspondientes a la excitación y tampoco que su buen comportamiento en el caso de errores de transmisión, puesto que, como se recordará (III.3.1.1) también son éstos los más protegidos por el codificador de canal GSM.

IV.3. Parametrizaciones para reconocimiento

Cuando abordamos el problema de reconocimiento de habla, el primer paso consiste, como expusimos en II.3, en la extracción de una secuencia de vectores de características, $\mathbf{Y} = \mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_T$ a partir de la señal de voz, $s[n]$, de forma que se reduzca la variabilidad no lingüística de esta señal para proceder después a su decodificación utilizando modelos estadísticos de estos datos acústicos, para determinar qué secuencia de palabras o símbolos $W = w_1, w_2, \cdots, w_M$ ha sido proferida con mayor verosimilitud.

Sin embargo, a la hora de construir los vectores acústicos no está claro qué parámetros son los más relevantes para el reconocimiento o cuáles son los que reducen en mayor medida la variabilidad lingüística, entre otras cosas porque depende de la tarea que se aborda, la lengua, y sobre todo de la adecuación de este tipo de parámetros al modelo que se utiliza: muchas características que podrían ser empleadas o que podrían resultar relevantes no se utilizan debido a las dificultades para incluirlas de forma efectiva dentro del vector de parámetros.

El tipo de modelado que hemos utilizado en esta tesis es el más ampliamente empleado en reconocimiento automático: los modelos ocultos de Markov (HMM); en este caso lo más habitual es asumir que los elementos del vector de características están incorrelados, lo que permite utilizar matrices de covarianza diagonales para la estimación de los modelos gaussianos de cada estado.

Comenzamos esta sección introduciendo los parámetros con los que se describe habitualmente la envolvente espectral: los cepstrum o parámetros cepstrales y sus variantes. A continuación presentamos una modificación de estos parámetros inspirada en el sistema auditivo humano: la escala Mel. Seguidamente, presentamos una extensión del vector de parámetros consistente en los denominados parámetros dinámicos. Después evaluamos otros tipos de parámetros que pueden añadirse al vector de parámetros tales como la energía, indicadores de si la trama es sonora o sorda, periodo fundamental, etc. Finalizaremos presentando una visión general de algunas de las estrategias más utilizadas en reconocimiento robusto, con el objetivo de situar la alternativa de reconocimiento mediante transparametrización que proponemos.

IV.3.1. Parámetros cepstrales

Los parámetros cepstrales son, con gran diferencia los más utilizados a la hora de construir los vectores acústicos de reconocimiento debido a su buen comportamiento cuando se realiza el reconocimiento con HMM. Estos parámetros resultan normalmente decorrelados entre sí, lo que permite utilizar matrices de covarianza diagonales en la caracterización de los símbolos que pueden ser emitidos en cada estado de Markov. Los parámetros cepstrales proporcionan una aproximación al análisis de componentes principales (PCA –*Principal Component Analysis*–) y por lo tanto proporcionan una representación compacta de la variabilidad del espectro de voz, como se requería, aunque son bastante sensibles a las variaciones entre locutores y por ello es necesario entrenar los modelos con gran cantidad de locutores para proporcionar independencia de los mismos [103].

Para el cálculo de estos parámetros se aplica una función no lineal (el logaritmo) al espectro de la señal y a continuación se obtiene la señal correspondiente en el dominio temporal[1] mediante la transformada de Fourier inversa, lo cual resulta una "transformación homomórfica". Concretamente, la transformación de la señal $s[n]$, en su secuencia de cepstra correspondiente, $c_s[n]$, se realiza de la siguiente manera:

$$c_s[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\left|S(\Omega)\right|\right) e^{j\Omega n} d\Omega \qquad \textbf{(IV-12)}$$

donde $S(\Omega)$ es la transformada de Fourier de la señal $s[n]$. Esta secuencia, $c_s[n]$, es a la que nos referimos habitualmente cuando hablamos de "cepstrum" aunque, como se puede observar en la fórmula anterior, para su cálculo sólo se emplea la parte real del logaritmo, motivo por el cual se le denomina "cepstrum real". Es importante resaltar cómo en este caso solamente se tiene en cuenta el módulo de la transformada $S(\Omega)$ y por lo tanto la trasformación no es reversible, es decir, no es posible recuperar $s[n]$ a partir de $c_s[n]$, a no ser que la primera sea de fase mínima. Otra propiedad importante de $c_s[n]$ es que, cuando se parte de señales $s[n]$ reales, y por tanto $\log|S(\Omega)|$ es real y simétrico, la secuencia resultante es real y simétrica.

En reconocimiento de habla el análisis cepstral se emplea para extraer la envolvente espectral de la señal de voz. Concretamente se asume que la señal de voz, $s[n]$, se obtiene como convolución de dos señales, la excitación $\tilde{e}[n]$ y el filtro, $\tilde{h}[n]$, que no tienen necesariamente que corresponder exactamente con las obtenidas mediante el procedimiento de predicción lineal (véase la sección IV.2.1). Es decir en ese caso,

[1] A este dominio también se le denomina "cuefrencial". Con este cambio en el orden de las sílabas de "frecuencia" se pretende expresar que los papeles del tiempo y la frecuencia están intercambiados; así, por ejemplo, las operaciones de "filtrado" que eliminan ciertas bandas de "frecuencia" pasan a ser operaciones de "liftrado" y eliminan ciertas bandas de "cuefrencias".

$$\log|S(\Omega)| = \log|\tilde{E}(\Omega)| + \log|\tilde{H}(\Omega)| \qquad \textbf{(IV-13)}$$

y por tanto,

$$c_s[n] = c_{\tilde{e}}[n] + c_{\tilde{h}}[n] \qquad \textbf{(IV-14)}$$

Debido a las propiedades intrínsecas de la señal de voz, la componente $c_{\tilde{h}}[n]$ decae rápidamente por lo que basta retener las *L* primeras muestras (típicamente 12 en aplicaciones de reconocimiento) para caracterizar $\tilde{h}[n]$, es decir, seleccionando una ventana con las muestras *1* a *L* conseguimos "deconvolucionar" la señal $s[n]$ separando sus dos contribuciones[2]. En otras palabras, si $\mathbf{c}_s^{(L)} = [c_s^1, c_s^2, \ldots, c_s^L]$ denota los *L* primeros coeficientes cepstrales, normalmente,

$$\mathbf{c}_s^{(L)} \approx \mathbf{c}_{\tilde{h}}^{(L)} \qquad \textbf{(IV-15)}$$

donde no se considera habitualmente $n = 0$, que no es otra cosa que el valor medio de $\log|S(\Omega)|$ Veremos más adelante (sección IV.4), sin embargo, como Huerta et al., ante la dificultad de obtener el parámetro de energía a partir de la parametrización de codificación sin recurrir a la decodificación completa de la señal de voz, sustituyen dicho parámetro por $c_{\hat{h}}^0$.

Sin embargo, como hemos visto en IV.2.1, esta no es la única manera de conseguir la deconvolución de estas dos componentes y así es posible hacerlo mediante predicción lineal, obteniendo los *p* coeficientes de predicción lineal, $\hat{a}_j$, que caracterizan el filtro $\hat{h}[n]$, cuya función de transferencia "todo-polos" es la siguiente:

$$H(z) = \frac{1}{1 - \sum_{j=1}^{p} \hat{a}_j z^{-j}} \qquad \textbf{(IV-16)}$$

Además, es posible obtener los coeficientes cepstrales, $\mathbf{c}_{\hat{h}}^{(L)}$ de forma recursiva a partir de los de predicción lineal, de la siguiente manera:

$$c_{\hat{h}}^n = \hat{a}_n + \sum_{i=1}^{n-1} (n-i)\hat{a}_{n-i} c_{\hat{h}}^i \quad \textit{para } n = 1, \cdots, L \qquad \textbf{(IV-17)}$$

con $\hat{a}_0 = 1$ como se desprende de (IV-16). Nótese que en este caso los cepstrum representan al filtro LP, $\hat{h}[n]$ y son distintos de los obtenidos directamente a partir $s[n]$. Los parámetros $\mathbf{c}_{\hat{h}}^{(L)}$ así obtenidos se denominan habitualmente "LPCepstra" y es preciso señalar que, en un principio no hay ninguna restricción en cuanto a la relación

[2] Esta operación se denomina "deconvolución homomórfica" y el procedimiento de enventanado en el dominio "cuefrencial", "liftrado".

entre el número L de cepstra y el orden de predicción p, pero que en el caso de que $L > p$, aunque el cálculo de los cepstra superiores es posible, no aporta información alguna puesto que los correspondientes $\hat{a}_j$ con $j > p$ son nulos [85].

Por otra parte, como ya señalábamos en IV.2.2 los coeficientes de predicción no tienen una interpretación directa en el dominio de la frecuencia (al contrario que los LSP); por este motivo en muchas ocasiones se opta por calcular primero, $\hat{H}(\Omega)$ a partir de $\hat{a}_j$, de forma que se pueda manipular en el dominio de Ω (por ejemplo introduciendo la escala Mel, como veremos a continuación en IV.3.1), y posteriormente obtener $\mathbf{c}_{\hat{h}}^{(L)}$. También es posible calcular $\hat{H}(\Omega)$ directamente a partir de los LSP, como ponen de manifiesto Sugamura et al. ([167]) en el contexto de análisis y síntesis de voz, de la siguiente manera:

$$\left|\hat{H}(\Omega)\right|^2 = \frac{2^{-p}}{\operatorname{sen}^2\left(\frac{\Omega}{2}\right)T_o(\Omega) + \cos^2\left(\frac{\Omega}{2}\right)T_e(\Omega)} \qquad \textbf{(IV-18)}$$

donde

$$T_o(\Omega) = \prod_{i=1}^{\frac{1}{2}p} (\cos\Omega - \cos\omega_{2i-1})^2$$
$$T_e(\Omega) = \prod_{i=1}^{\frac{1}{2}p} (\cos\Omega - \cos\omega_{2i})^2 \qquad \textbf{(IV-19)}$$

donde p ha de ser par como en los casos que nos ocupan (véase [167], para p impares).

Efectivamente, nosotros hemos comprobado que los resultados de reconocimiento obtenidos a partir de ambas descripciones (LSP o LP) son casi idénticos (véase el Capítulo V).

Debemos mencionar, además, que en la mayoría de las implementaciones se sustituye la transformada inversa de Fourier de (IV-12) por una DFT (*–Discrete Fourier Transform–*) que debido a la simetría tanto de $S[k]$ como de $c_s[n]$ se reduce a su vez a una transformada discreta del coseno (DCT *–Discrete Cosine Transform–*) ([140]) de forma que los cepstra obtenidos son de la siguiente manera:

$$c_s^n = \sqrt{\frac{2}{N}} \sum_{k=0}^{N-1} \log|S[k]| \cos\left(\frac{\pi n}{N}\left(k + \tfrac{1}{2}\right)\right) \quad para\ n = 1, \cdots, L \qquad \textbf{(IV-20)}$$

con $S[k] = S\left(\Omega = \frac{2\pi k}{N}\right)$ y N el número de muestras de la misma que se consideran la transformada discreta de Fourier de $s[n]$. Desde luego, esta sustitución se puede hacer de forma análoga para $\hat{h}[n]$. Finalmente, si fuese necesario calcular los c_s^0:

$$c_s^0 = \sqrt{\frac{1}{N}\sum_{k=0}^{N-1} \log|S[k]|} \qquad \textbf{(IV-21)}$$

Además, muchas veces se utiliza un banco de filtros con *M* canales (en realidad, son medias ponderadas del espectro en el entorno de cada una de las *M* frecuencias), en lugar de las *N* muestras individuales de la transformada discreta de Fourier. Este análisis en bandas críticas está inspirado en el sistema auditivo humano y cobra, en realidad, más significado cuando la distribución y anchura de estos filtros es no lineal a lo largo del eje frecuencial, es decir, cuando se aplican escalas de tipo Mel o Bark, como expondremos en la próxima sección. Sin embargo, hemos considerado oportuno incluirla aquí, puesto que se puede ver como una generalización del análisis de Fourier que hemos presentado.

Estos filtros generalmente tienen una respuesta frecuencial, $W(\Omega)$, triangular y un ancho de banda *B* de la siguiente manera:

$$W(\Omega) = \begin{cases} 1 - \frac{2}{B}|\Omega| & |\Omega| \le \frac{B}{2} \\ 0 & resto \end{cases} \qquad \textbf{(IV-22)}$$

A partir de este tipo de filtros, se obtienen una versión ponderada de $S[k]$ que denotamos $M[i]$ para cada uno de los *M* canales de la siguiente manera:

$$M[i] = \frac{1}{B}\int_0^{\pi} |S(\Omega)|W(\Omega - \Omega_i)d\Omega \qquad 0 < i \le M \qquad \textbf{(IV-23)}$$

que sustituirán a los $S[k]$ en la ecuación (IV-20) y donde Ω_i representa las frecuencias centrales en las que se sitúa cada filtro y que están relacionadas de la siguiente manera permitiendo el solapamiento entre dos filtros consecutivos:

$$\Omega_{i+1} = \Omega_i + \frac{B}{2} \quad 0 < i \le M \qquad \textbf{(IV-24)}$$

donde $\Omega_1 = \frac{\pi}{M}$.

Por otra parte y precisamente a raíz de la demanda de reconocimiento a partir de voz codificada, Choi et al. ([24], [25] y [26]) sugieren utilizar directamente los parámetros LSP como vector acústico, para lo cual proponen una serie de distancias ponderadas por la sensibilidad espectral de estos parámetros, aprovechando sus propiedades en el dominio frecuencial. Concretamente proponen tres tipos de ponderación: la primera se basa en el cálculo del espectro de potencia y la ponderación de cada uno de los ω_i según el valor del espectro en esa frecuencia; la segunda tiene en cuenta que la proximidad de dos parámetros consecutivos implica un pico en el correspondiente espectro de potencia y por tanto enfatiza los ω_i en función de su distancia a los dos parámetros adyacentes, ω_{i-1} y ω_{i+1}; y la tercera utiliza una ponderación basada en la sensibilidad de dichos parámetros a la cuantificación. Estos autores, además, aplican a estas ponderaciones la escala Mel, que veremos a

continuación, resultando esta opción la más ventajosa en comparación con la distancia entre LSPs sin ponderar.

Sin embargo, esta parametrización, aún con las modificaciones esbozadas arriba no resulta competitiva si la comparamos con la basada en parámetros cepstrales y por ese motivo Kim et al. ([113], [111]) y Choi et al. ([25]) proponen una aproximación computacionalmente eficiente para obtener unos parámetros cepstrales que denominan "pseudocepstra", $\hat{\mathbf{c}}_{\hat{h}}^{(L)}$, a partir de los LSPs, que toman el siguiente valor,

$$\hat{c}_{\hat{h}}^{n} = \frac{1}{2n}\left(1+(-1)^{n}\right)+\frac{1}{n}\sum_{i=1}^{p}\cos(n\omega_i) \quad 1 \le n \le L \qquad \textbf{(IV-25)}$$

Convienen hacer hincapié en que se trata de una aproximación que se obtiene al despreciar un cierto residuo y por ello, como demostraremos en el Capítulo V, ofrece resultados de reconocimiento inferioresa los obtenidos con otros parámetros.

IV.3.2. Escala Mel

La escala Mel es una transformación no lineal de la frecuencia inspirada en el comportamiento del sistema auditivo humano que aplicada en el dominio frecuencial previo a la transformación a la DCT que da lugar a los parámetros cepstrales resulta muy beneficiosa (véanse, por ejemplo, algunos experimentos en el Capítulo V de esta tesis).

Esta escala fue obtenida de forma empírica y se puede aproximar mediante la siguiente ecuación ([177]):

$$Mel_1(f) = 2595\log_{10}\left(1+\frac{f}{700}\right) \qquad \textbf{(IV-26)}$$

donde f está expresado en Hz. También suele utilizarse esta segunda aproximación ([25], por ejemplo):

$$Mel_2(f) = f + \text{arctg}\left(\frac{0.45\,\text{sen}(f)}{1-0.45\cos(f)}\right) \qquad \textbf{(IV-27)}$$

Ambas funciones dan lugar a curvas similares y los resultados de reconocimiento obtenidos con cada una de ellas son muy parecidos (véase Capítulo V). Los coeficientes cepstrales así obtenidos se denominan generalmente MFCC (–*Mel Frequency Cepstral Coefficients*–). Sin embargo, es importante hacer notar que sólo los procedimientos en los que se calcula el espectro, ya sea de la señal, $s[n]$, o del filtro, $\hat{h}[n]$, es posible aplicar esta escala, con la excepción del "pseudocepstrum" en el que puede aplicarse directamente sobre los parámetros LSP gracias a la posibilidad de su interpretación directa en el dominio frecuencial.

IV.3.3. Parámetros de información dinámica

Ya hemos señalado en IV.1 cómo el análisis por tramas, que tiene su origen en la tecnología de codificación de voz, se ha exportado a la parametrización para reconocimiento con notable éxito, aunque, efectivamente, las unidades que se quiere identificar en reconocimiento son de mayor longitud que las tramas habitualmente consideradas en codificación [83]. Por tanto, parece lógico pensar que la introducción de información referente a la evolución o la variación de los parámetros básicos de reconocimiento en cada una de las tramas, puede resultar de utilidad.

Furui ([64]) introdujo los parámetros de información dinámica con el objetivo de describir las trayectorias temporales de los parámetros estáticos en la vecindad de la trama que se considera en cada momento (entre 50 y 100 ms). Así, en un principio consideró tres tipos de parámetros:

- el valor medio, que más tarde se sustituyó por el propio parámetro en el centro del intervalo considerado,
- la pendiente (delta) y
- la curvatura (doble-delta o aceleración)

de la trayectoria de (en este caso) los parámetros cepstrales. Es decir, los coeficientes de los parámetros para cada trama *k*, $\boldsymbol{\delta}_{\bullet}^{(L)}[k]$ pueden ser obtenidos de la siguiente manera:

$$\delta_{\bullet}^{n}[k] = \frac{\sum_{i=1}^{N_t} i\left(c_{\bullet}^{n}[k+i] - c_{\bullet}^{n}[k-i]\right)}{2\sum_{i=1}^{N_t} i^2} \qquad \textbf{(IV-28)}$$

donde N_t es el número de muestras posteriores y desde luego esto se puede aplicar al cálculo de los parámetros delta de cualquiera de los parámetros cepstrales que hemos presentado en la sección IV.3.1. Esta posibilidad la hemos denotado con "$\bullet$", que puede ser sustituido por, *s*, $\hat{h}$, $\tilde{h}$ o en general, cualquier señal de la que se haya obtenido el cepstrum.

De forma análoga se pueden calcular los parámetros de aceleración, $\dot{\boldsymbol{\delta}}_{\bullet}^{(L)}[k]$, utilizando la misma expresión, esta vez sobre los parámetros $\boldsymbol{\delta}_{\bullet}^{(L)}[k]$. Como se desprende de la expresión anterior, la inclusión de estos parámetros introduce un retardo algorítmico importante: N_t tramas en el caso de incluir tan sólo parámetros delta y $2N_t$ cuando además se utilizan los parámetros de aceleración. Esto, totalmente inadmisible desde el punto de vista de codificación para la transmisión en tiempo real, es habitual en reconocimiento, lo que nos lleva a considerar estrategias de recuperación frente a errores que contemplan ventanas temporales mucho mayores que las de las estrategias utilizadas en codificación.

La utilización de estos parámetros obedece a dos intenciones claras: la primera tiene que ver con proporcionar en cada instante del análisis (es decir, en cada trama) información a más "largo-plazo" por los motivos que señalábamos antes y la segunda, porque estos parámetros son invariantes ante cualquier distorsión lineal (convolutiva) que se mantenga fija a lo largo del tiempo de duración de las $2N_t$. Es decir, cualquier distorsión $d[n]$ que actúe sobre la señal $s[n]$ de forma convolutiva:

$$s_d[n] = s[n] * d[n] \qquad \textbf{(IV-29)}$$

se refleja sobre los parámetros cepstrales de la siguiente manera:

$$\mathbf{c}_{s_d}^{(L)} = \mathbf{c}_s^{(L)} + \mathbf{c}_d^{(L)} \qquad \textbf{(IV-30)}$$

y por tanto, cuando esta distorsión se mantiene a lo largo de las $2N_t$ que abarca el análisis de la ecuación (IV-28) se cancela en cada uno de los sumandos del numerador de dicha ecuación. El mismo razonamiento aplica a los parámetros de aceleración y además, es la misma idea que subyace en el procedimiento de Sustracción de la Media Cepstral (CMS –*Cepstral Mean Substraction*–), que se introduciremos en IV.3.5.

Debemos recordar, sin embargo, que la distorsión de codificación a la que nos enfrentamos es no lineal, puesto que varía considerablemente de trama a trama (véase la sección III.2). En todo caso, es innegable que estos parámetros proporcionan información acerca de las tramas adyacentes en cada una de las tramas, lo cual resulta beneficioso.

IV.3.4. Otros parámetros

Además de los parámetros cepstrales (con la posibilidad de incluir la escala Mel) y sus correspondientes parámetros dinámicos, el vector de características puede incluir otro tipo de informaciones que mejoren el mejor modelado de las unidades acústicas. Sin embargo, en general, la forma de incluir otro tipo de parámetros no es evidente porque no es fácil determinar cuáles de estos parámetros son pertinentes en cada tarea y cuál es la ponderación que conduce a una mejor discriminación.

En todo caso, es muy habitual que se incluya un parámetro que represente la energía de la trama porque contribuye notablemente al aumento de las prestaciones del reconocedor. Sin embargo, en el caso de reconocimiento mediante transparametrización la obtención de este parámetro no resulta evidente, puesto que se evita la decodificación completa de la señal de voz y por tanto se impide el cálculo de su energía a partir de las muestras de la señal reconstruida. En IV.5.1 expondremos la solución que hemos adoptado en esta tesis para incluir esta información, así como las soluciones de otros autores.

Además del parámetro de energía, algunos autores proponen el uso de información acerca del periodo fundamental de la señal (sobre todo cuando los reconocedores se aplican a lenguas tonales [164]), o también sobre la vibración de las cuerdas vocales, usando bien la clasificación sonoro/sordo o bien las ganancias de las librerías

estocástica y adaptativa, esto último en el caso de utilizar reconocimiento mediante transparametrización, como en [112].

IV.3.5. Reconocimiento Robusto

Nos centramos, a continuación, en las técnicas que tienen como objeto contrarrestar la influencia de las fuentes de distorsión que deterioran el comportamiento de los reconocedores. En esta sección, pretendemos proporcionar una visión general de tales técnicas, de forma que se pueda situar adecuadamente la solución que proponemos.

Podemos agrupar las fuentes de distorsión en tres tipos básicos:

- **Hablante y tarea**: son muchas y muy variadas las fuentes de "distorsión" (o desajuste) debidas al hablante y a la tarea; estilo (por ejemplo, de habla espontánea o bien articulada), variaciones en la velocidad de elocución, enfatización, factores emocionales, hablantes no nativos, dialectos, características propias del hablante, (físicas, como la longitud de su tracto vocal, grupo de la población al que pertenece –por ejemplo, su edad–), calidad de la voz (susurrante, tensa, etc.), contexto (profesional, conversación, diálogo hombre-máquina).

- **Transductores y canal de transmisión**: respecto a los transductores, resulta obvio que las características del micrófono (tipo, direccionalidad, con o sin cancelación de ruido, etc) influyen en la calidad del reconocimiento. En cuanto al canal, existen varias fuentes de distorsión: la línea de transmisión, (ecos eléctricos, ruido), los sistemas de adquisición y digitalización (filtrados, conversión A/D) y, en definitiva, todos los procesos asociados a la transmisión de la señal a través de redes de comunicaciones. Como caso particular de este tipo de distorsión, nos encontramos con la distorsión de codificación (véase la sección III.2), presente en aquellas transmisiones que por escasez del ancho de banda disponible, exigen que la voz sea codificada. Además, como ya hemos comentado, los errores de transmisión afectan de forma distinta a la voz codificada a tasas medias-bajas, que a la voz que no ha sufrido este proceso de compresión, puesto que su influencia es distinta en función del parámetro al que afecten.

- **Entorno acústico**: en la propagación local hay que tener en cuenta los ecos, ruido ambiente, reverberación, distancia al micrófono, dirección del emisor respecto al mismo, efecto Lombard.

En general, se suelen distinguir tres clases de sistemas de reconocimiento de habla: los dependientes del locutor (SD –*Speaker Dependent*–), los independientes del locutor (SI –*Speaker Independent*–) y los adaptables al locutor (SA –*Speaker Adaptive*–). Los primeros están diseñados y entrenados para reconocer el habla particular de un determinado individuo o grupo de individuos, mientras que los segundos pretender hacer el reconocimiento con cualquier persona. Por último, los adaptables parten de una serie de modelos independientes del locutor y tienen como objetivo adaptarse al locutor que los utiliza con el fin último de alcanzar la precisión de un reconocedor dependiente del locutor en cuestión.

En un principio, la anterior clasificación trata exclusivamente la variabilidad correspondiente al locutor. Sin embargo, este mismo razonamiento se podría aplicar al resto de las variabilidades no lingüísticas o distorsiones a las que aludíamos antes y así, cuando se trata de sistemas robustos podemos referirnos a sistemas independientes del entorno o, por el contrario, adaptables al entorno.

Así, los sistemas que persiguen la independencia del entorno actúan básicamente sobre:

- La parametrización o extracción de características, que debe proporcionar información fonética fiable y descartar los detalles específicos del locutor. Además debe ser robusta a distorsiones del entorno que, en general, se consideran bien aditivas o bien convolutivas. La mayoría de los aspectos de parametrización para reconocimiento que hemos tratado en las secciones anteriores tienen como objetivo obtener una parametrización robusta en este sentido. Así, además del uso de los parámetros cepstrales junto con la escala Mel, la parametrización PLP (*–Perceptual Linear Prediction–*) pretende obtener unos parámetros más robustos basándose en varias características del sistema auditivo humano; en primer lugar, aplica sobre el espectro (escalado según la frecuencia Mel) un análisis de bandas críticas (ecuación (IV-23)), a continuación utiliza un preénfasis que responde al incremento de la sensibilidad auditiva con la frecuencia y finalmente, un escalado no lineal (raiz cúbica) que trata de reproducir la relación entre la intensidad del sonido y la percibida. Una vez hechas estas transformaciones se aplica, nuevamente sobre el dominio temporal, un análisis LP a partir del cual se obtienen los parámetros cepstrales (ecuación (IV-17)) ([82]).

  Por otra parte, también hemos mencionado cómo los parámetros de información dinámica resultan robustos cuando tratamos distorsiones de tipo convolutivo. Además, la sustracción de la media cepstral (CMS *–Cepstral Mean Substraction–*) también es un método que obedece a las mismas motivaciones. Este procedimiento consiste en eliminar una estimación del valor medio de los parámetros cepstrales, que se calcula bien a partir de la elocución completa, o bien a partir de un número determinado de tramas anteriores, cuando no es posible asumir un retardo elevado; no obstante, además de eliminar la distorsión (suponiendo que es de muy baja frecuencia) se elimina la parte de la señal que varía más lentamente Esta técnica puede verse como una simplificación de la de normalización espectral (*–spectral normalization–*), en la que se trata de estimar la distorsión para eliminarla y aún de la filtrado del espectro de modulación [77].

  Otras aproximaciones son las que utilizan funciones alternativas a la logarítmica para la obtención de los parámetros cepstrales (ecuación (IV-12)) tales como la exponencial, $(\cdot)^{\gamma}$, con $\gamma \in [0,2]$ ([129]), que también se ha aplicado para reducir la distorsión en canales GSM ([40]). Además, el análisis discriminante lineal (LDA *–Linear Discriminant Analysis–*) propone una trasformación lineal del dominio cepstral obtenida mediante la diagonalización de la matriz $W^{-1}B$, donde W es la matriz de covarianza intra-clases y B la de covarianza entre-clases, de forma que se obtiene una representación más compacta que la cepstral.

- El modelado acústico, que debe proporcionar distribuciones discriminativas, empleando un número de parámetros óptimo a la vista de los

datos de entrenamiento disponibles. Los avance más significativos en modelado acústico tienen que ver con el desarrollo de árboles de decisión para modelado de subunidades acústicas dependientes de contexto. Además, en el entrenamiento de estos modelos se introducen variantes para conseguir la independencia del entorno, como el entrenamiento multi-estilo (*–multi-style–*) y el discriminativo. En el primero se entrenan los modelos con voz procedente de diferentes entornos y estilos de habla, de forma que se consigue que disminuya la sensibilidad de los reconocedores a tales entornos y estilos; sin embargo, como es lógico, debido al promediado que tiene lugar, la capacidad de discriminación disminuye. Por eso, en este contexto, se han propuesto también métodos de entrenamiento discriminativo, que consisten, precisamente, en minimizar una función de pérdida de discriminación a través, generalmente, de algoritmos de descenso de gradiente ([103]).

- La calidad de la señal de entrada, que se ha tratado introduciendo técnicas de mejora e igualación, como las utilizadas precisamente en entornos móviles en [112].

En cuanto a los sistemas de adaptación al entorno y compensación, el objetivo es obtener, a partir de un sistema independiente del entorno (en la medida de lo posible), un sistema adaptado a un entorno, un hablante y un estilo particulares, que se comporte después de la adaptación como se hubiera comportado un sistema entrenado específicamente para esa situación. Así, se pueden distinguir los sistemas que se adaptan "en bloque" (*–batch adaptation–*), que generalmente suelen hacerlo de forma supervisada, de los que se adaptan en "tiempo de ejecución" (*–run-time adaptation–*), la mayoría de las veces sin supervisión.

Las técnicas de adaptación pueden clasificarse, a su vez, en función del componente del sistema de reconocimiento sobre el que actúan ([103]):

- Las que adaptan los parámetros del modelo, como las técnicas Bayesianas o MAP (–Maximum A Posteriori–), las de transformación (MLLR –Maximum Likelihood Linear Regresión–), o las de selección de modelos o clasificación del locutor (–speaker clustering–).

- Las basadas en adaptación de las parametrizaciones, como las de normalización del locutor en las que se trata de encontrar una transformación que aproxime los vectores acústicos obtenidos durante el funcionamiento real a los utilizados para el entrenamiento.

- Las que mejoran la señal de voz, utilizan enmascaramento de ruido o igualación espectral y pretenden eliminar las distorsiones basándose en el conocimiento específico de las mismas.

El reconocimiento mediante transparametrización que nosotros proponemos pretende aliviar una distorsión muy concreta: la resultante de la codificación de voz. Como ya expusimos en el Capítulo III, esta distorsión no se puede considerar aditiva o convolutiva, y además han de considerarse los efectos debidos a los errores de transmisión, que al verificarse sobre la parametrización para codificación también necesitan un tratamiento específico. Por ese motivo, esta solución es aplicable exclusivamente en los entornos en los que se lleva a cabo una codificación; sin embargo, es perfectamente posible aplicar, sobre la base de la parametrización que

proponemos, las técnicas que hemos esbozado en esta sección con la salvedad de las que hacen uso de la señal de voz directamente. Por tanto, es posible que pudieran obtenerse sistemas más robustos o que se adapten a determinadas circunstancias partiendo de un sistema mejorado para casos de voz codificada.

De hecho, hemos aplicado muchas de las técnicas descritas para sistemas independientes del entorno, por estar muy establecidas dentro de la tecnología de reconocimiento; es el caso del uso de parámetros dinámicos o la escala Mel, que nos permiten comparar nuestra propuesta de forma realista con las parametrizaciones más ampliamente utilizadas.

IV.4. Codificación a partir de parametrizaciones para reconocimiento

A raíz de la introducción de la aproximación de reconocimiento distribuido que presentamos en II.4.2, se puso de manifiesto la necesidad de reconstruir la señal de voz a partir de una representación originalmente diseñada para reconocimiento. Además, la codificación a partir de parametrizaciones para reconocimiento tiene las siguientes aplicaciones:

- En grabadores digitales y otros dispositivos que registran voz que más tarde será reconocida automáticamente.
- Servicios de mensajería de voz o de acceso a bases de datos vocales que utilizan técnicas de reconocimiento de habla para dar servicios de valor añadido.
- Servicios basados en reconocimiento de habla en los que la reproducción de las grabaciones archivadas puede ser importante por motivos legales.

En [29], se plantean los requerimientos para un nuevo codificador que permita reconstruir la señal de voz a partir de los parámetros para reconocimiento descritos en [43]. El diseño final de este codificador está previsto para agosto de 2002 y destina 1,6 Kb/s para la codificación de los parámetros adicionales necesarios para la reconstrucción, incluyendo la codificación de canal, además de los 4,8 kb/s ya previstos para los parámetros de reconocimiento. Además, este codificador debe exceder la inteligibilidad del estándar federal del departamento de defensa estadounidense a 2,4 Kb/s (FS 1015) y serán considerados excepcionales aquellos que estén cerca o superen el también estándar federal a 2,4 Kb/s, MIL-STD-3005 basado en MELP (*–Mixed-Excitation Linear Prediction–*). Estos dos codificadores de referencia obtienen una puntuación de 2,2 y 3,2 en la escala MOS, respectivamente [78]. Como vemos, la calidad de la voz reconstruida que se persigue es notablemente inferior a la conseguida por codificadores a regímenes binarios similares (por ejemplo, los 3,9 MOS de G.723.1 a 6,3 Kb/s)

En [20], [19] y [131] Chazan et al. presentan lo que denominan "Codificador de Voz compatible con Reconocimiento" (*RECOVC –Recognition COmpatible Voice Compresión–)*. Este codificador propiedad de la empresa IBM, consigue regímenes binarios entre 4 y 6,9 Kb/s en función de la dimensión del vector de parámetros MFCC (13 ó 24, está última para proporcionar robustez en entornos ruidosos) que emplea y de la habilitación o deshabilitación de la opción de reconstrucción. No nos consta ninguna referencia a pruebas estándar que permitan valora la calidad de este codificador (por ejemplo, basandose en escalas reconocidas como la anteriormente mencionada MOS). En todo caso, describimos a continuación, el procedimiento empleado para la reconstrucción de la señal de voz con el objetivo de ilustrar esta alternativa, basada, precisamente, en la transparametrización inversa a la que presentamos en esta tesis.

En esta ocasión, la reconstrucción de la voz está basada en un modelo sinusoidal definido por las amplitudes, $\{A_i\}$, frecuencias, $\{\Omega_i = 2\pi f_i\}$, y fases, $\{\varphi_i\}$ ($i = 0,\ldots,N-1$) de N componentes sinusoidales de la voz, que se obtienen para cada trama. Éstas se obtienen a partir de los parámetros cepstrales, la frecuencia fundamental, F_0, y la información sobre la clasificación sonora o sorda de la trama correspondiente.

La señal de voz sintetizada de esta forma, se puede ver en el dominio frecuencial como (incluyendo sólo las frecuencias positivas del espectro):

$$\widehat{S}(\Omega) = \sum_{i=0}^{N-1} A_i e^{j\varphi_i} H_w(\Omega - \Omega_i) \qquad \textbf{(IV-31)}$$

donde $H_w(\Omega)$ es la transformada de Fourier de la ventana (normalmente tipo Hamming) empleada en la síntesis y que, por conveniencia, de considerará de banda limitada.

Para la obtención de las frecuencias, $\{f_i\}$, si la trama que se está sintetizando es sonora se generan *R* frecuencias múltiplos de la frecuencia fundamental, es decir,

$$f_i = iF_0 \qquad i = 0,\ldots,R-1 \qquad \textbf{(IV-32)}$$

y las $N-R$ sinusoides restantes se añaden de forma arbitraria para que la voz sintetizada resulte más natural. A cada una de las frecuencias armónicas se les asigna un peso $\rho_i = 1$, que más tarde influirá en el cálculo de las amplitudes, mientras que el resto tienen pesos entre 0 y 1 proporcionales a la frecuencia, f_i. Cuando la trama es sorda, las frecuencias se seleccionan de forma equiespaciada, con pesos $\rho_i = 1 \ \forall i$.

En cuanto a las fases, $\{\varphi_i\}$, en el caso de tramas sonoras, se calculan de la siguiente manera:

$$\varphi_i = i\phi_0(kT) + \Phi_i \qquad i = 0,\ldots,N-1 \qquad \textbf{(IV-33)}$$

donde *k* es el número de trama en la que nos encontramos, *T* es la duración de cada trama, Φ_i es una fase adicional optativa y ϕ_0 es una función que se diseña de forma que no haya transiciones bruscas entre tramas.

Para la obtención de las amplitudes, $\{A_i\}$, se aplica el banco de filtros tal como lo definimos en (IV-23) (posiblemente con la aplicación de la escala Mel), de forma que podemos obtener la expresión de los valores $\mu[i]$ a partir de la señal sintetizada de esta forma:

$$\mu[i] = \frac{1}{B}\int_0^{\pi} \left|\widehat{S}(\Omega)\right| W(\Omega - \Omega_i) d\Omega \qquad 0 < i \le M \qquad \textbf{(IV-34)}$$

Si ahora sustituimos la expresión (IV-31) en la ecuación anterior, obtenemos que:

$$\mu[i] = \frac{1}{B}\int_0^{\pi} \left|\sum_{i=0}^{N-1} A_i e^{j\varphi_i} H_w(\Omega - \Omega_i)\right| W(\Omega - \Omega_i) d\Omega \qquad 0 < i \le M \qquad \textbf{(IV-35)}$$

y nuestro objetivo es minimizar $\sum_{i=1}^{M} (M[i] - \mu[i])^2$, donde $M[i]$ son los coeficientes correspondientes a cada uno de los M canales con los que obtuvieron, en su momento, los parámetros cepstrales de los que disponemos. Sin embargo, dado que normalmente $N > M$, el sistema de ecuaciones queda indeterminado; para resolverlo, se definen las amplitudes, $\{A_i\}$, que buscamos, como combinación lineal de una serie de funciones base, $\psi_m(\Omega)$, y los pesos $\{\rho_i\}$ que describimos anteriormente, es decir,

$$A_i \equiv \rho_i \sum_{m=1}^{M} x_m \psi_m(\Omega_i) \qquad \textbf{(IV-36)}$$

donde x_m son los coeficientes (no negativos) de la combinación.

De esta forma podemos escribir de nuevo la señal sintetizada, $\widehat{S}(\Omega)$, de la siguiente manera:

$$\widehat{S}(\Omega) = \sum_{m=1}^{M} x_m \widehat{S}^{(m)}(\Omega) \qquad \textbf{(IV-37)}$$

donde cada uno de los espectros, $\widehat{S}^{(m)}(\Omega)$, responde a la siguiente expresión:

$$\widehat{S}^{(m)}(\Omega) = \sum_{i=0}^{N-1} \psi_m(\Omega_i) \rho_i e^{j\varphi_i} H_w(\Omega - \Omega_i) \qquad \textbf{(IV-38)}$$

Si sustituimos ahora esta nueva expresión de $\widehat{S}(\Omega)$ en la ecuación (IV-34) obtenemos que:

$$\mu[i] = \frac{1}{B}\int_0^{\pi} \left|\sum_{m=1}^{M} x_m \widehat{S}^{(m)}(\Omega)\right| W(\Omega - \Omega_i) d\Omega \qquad 0 < i \le M \qquad \textbf{(IV-39)}$$

Aproximando el valor absoluto de la suma por la suma de los valores absolutos e intercambiando el orden de sumatorio e integral encontramos la siguiente expresión:

$$\tilde{\mu}[i] = \sum_{m=1}^{M} x_m \frac{1}{B} \int_0^{\pi} \left| \widehat{S}^{(m)}(\Omega) \right| W(\Omega - \Omega_i) d\Omega \qquad 0 < i \leq M \qquad \textbf{(IV-40)}$$

podemos ahora escribir $\tilde{\mu}[i]$ de la siguiente manera:

$$\tilde{\mu}[i] = \sum_{m=1}^{M} x_m R_{i,m} \qquad 0 < i \leq M \qquad \textbf{(IV-41)}$$

donde cada uno de estos elementos, $R_{i,m}$, responde a la siguiente expresión:

$$R_{i,m} = \frac{1}{B} \int_0^{\pi} \left| \widehat{S}^{(m)}(\Omega) \right| W(\Omega - \Omega_i) d\Omega \qquad 0 < i \leq M \quad 0 < m \leq M \qquad \textbf{(IV-42)}$$

Por tanto la expresión que queremos minimizar es la siguiente:

$$\sum_{i=1}^{M} \left(\sum_{m=1}^{M} x_m R_{i,m} - M[i] \right)^2 \qquad \textbf{(IV-43)}$$

que podemos escribir en forma matricial como

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x} \geq \mathbf{0}} \|\mathbf{R}\mathbf{x} - \mathbf{M}\| \qquad \textbf{(IV-44)}$$

donde $\mathbf{R} = [R_{i,m}]$, $\mathbf{x} = [x_m]$ y $\mathbf{M} = [M[k]]$ con $i = 1, \ldots, M$ y $m = 1, \ldots, M$ y $\hat{\mathbf{x}}$ el vector óptimo de pesos solución del problema de minimización. Este problema de mínimos cuadrados se puede resolver con cualquier método iterativo de optimización (por ejemplo, descenso de máxima pendiente). Si además se eligen $\psi_m(\Omega)$ con formas similares a las de $W(\Omega)$, la matriz $\mathbf{R}$ se simplifica notablemente y las soluciones son más fáciles de encontrar.

Este procedimiento para reconstruir la señal de voz es muy sensible a la estimación correcta del periodo fundamental, motivo por el cual sus autores proponen métodos para mejorar la estimación del mismo [21].

IV.5. Reconocimiento a partir de parametrizaciones para codificación: Reconocimiento mediante transparametrización

Desde que en 1998, Gallardo et al. y Huerta et al. propusieran una versión preliminar de reconocimiento mediante transparametrización, varios autores han propuesto diversas variantes, bien sea en la composición del vector de parámetros acústicos que se considera o bien en las variables ambientales, distorsiones o entornos que se evalúan.

En esta sección vamos a presentar los aspectos que consideramos relevantes cuando se trata de realizar este reconocimiento mediante transparametrización, presentando las soluciones que hemos adoptado y contrastándolas con las propuestas por otros autores. En la Tabla IV-1 presentamos una visión general de estas soluciones.

Ref.	Codec	Envolvente espectral	Energ.	Tasa de Trama	Distorsión
Gallardo et al. ([65], [66], [67])	GSM-FR (sin cod. canal)	$\mathbf{mfc}_{\hat{h}}^{(12)}$ [3]	σ_s^2	20 ms	Simulación de errores GSM (modelo de Gilbert)
	GSM-HR (sin cod. canal)	$\mathbf{mfc}_{\hat{h}}^{(12)}$	R_0	20 ms	
Huerta et al. ([85], [86])	GSM-FR (sin cod. canal)	$\mathbf{mfc}_{\hat{h}}^{(12)}$	$c_{\hat{h}}^0$	20 ms	Ruido rosa aditivo
		$\mathbf{mfc}_{\hat{h}}^{(12)} + \mathbf{mfc}_{e}^{(12)}$	$c_{\hat{h}}^0 + c_e^0$		
		$\left[mfc_{\hat{h}}^1, \ldots, mfc_{\hat{h}}^j, mfc_e^{j+1}, \ldots, mfc_e^{12}\right]$ [4]	$c_{\hat{h}}^0$		
Choi et al. ([24], [25])	Qualcomm CELP (sin cod. canal)	$\mathbf{mf\hat{c}}_{\hat{h}}^{(12)}$	-	20 ms	-
		$\mathbf{mf\bar{\omega}}^{(10)}$ [5]			

[3] "mf" denota la aplicación de la escala Mel sobre el parámetro correspondiente.

[4] El parámetro *j* óptimo se sitúa entorno a 7 u 8.

[5] Los parámetros $\mathbf{mf\bar{\omega}}^{(10)}$ son una versión ponderada de $\mathbf{mf\omega}^{(10)}$ que tiene en cuenta su sensibilidad espectral.

Ref.	Codec	Envolvente espectral	Energ.	Tasa de Trama	Distorsión
Kim et al. ([112])	IS-641 (incl. cod. canal)	$\left\lfloor \mathbf{mfc}_{\hat{h}}^{(10)}, \overline{ACG}, \gamma\overline{FCG} \right\rfloor$[6]	σ_e^2	10 ms (intp.)	Simulación de pérdida de tramas completas + Ruido de coche y "*babble*"
Peláez et al. ([68], [69], [145], [147], [148])	GSM-FR (incl. cod. canal)	$\mathbf{mfc}_{\hat{h}}^{(12)}$	$\hat{\sigma}_s^2$	10 ms (intp.)	Simulación de errores GSM Transcodificaciones
	GSM-HR (incl. cod. canal)	$\mathbf{mfc}_{\hat{h}}^{(12)}$	R_0		
	G.723.1 (sin cod. canal)	$\mathbf{mfc}_{\hat{h}}^{(12)}$	$\hat{\sigma}_s^2$		Simulación pérdida de paquetes (modelo de Gilbert) Transcodificaciones (GSM→IP)
		$\hat{\mathbf{mfc}}_{\hat{h}}^{(12)}$			

Tabla IV-1. Comparación de las soluciones de reconcimiento mediante transparametrización.

En concreto, los aspectos que vamos a considerar son; el modelado de la envolvente espectral, la estimación o cálculo de la energía, la tasa de trama y el tipo de distorsión que se considera. La notación y significado de cada uno de los parámetros que aparecen en la tabla se irá introduciendo poco a poco en las subsecciones siguientes.

En el Capítulo V presentaremos los resultados que hemos obtenido con los tipos de parametrización que proponemos, comparándolos con los que se obtendrían con la aproximación convencional (decodificación de la voz previa a la obtención de los parámetros de reconocimiento).

[6] Los parámetros $\overline{ACG}$ y $\overline{FCG}$ son las ganancias de las librerías adaptativa y fija, respectivamente, codificadas por IS-641 a las que se les ha aplicado un suavizado de mediana. La variable γ representa una ponderación aplicada sobre $\overline{FCG}$ y toma el valor (determinado empíricamente) de 0,1.

IV.5.1.Envolvente espectral

Como hemos expuesto en las secciones IV.2 y IV.3, en el flujo de bits que recibe el servidor de reconocimiento la información sobre la envolvente espectral de la señal de voz puede estar representada por varios tipos de parámetros, dependiendo del codificador que se utilice en cada caso. Nótese que, en todo caso, estos parámetros provienen de los de predicción lineal, que como ya hemos explicado realizan un modelado "todo-polos" de dicha envolvente. Por lo tanto, el hecho de partir de estos parámetros ya nos impide calcular los parámetros cepstrales, $\mathbf{c}_{\tilde{h}}^{(L)}$. En la Figura IV-1 mostramos las transparametrizaciones necesarias y las distintas posibilidades de obtener parámetros de reconocimiento que hemos contemplado.

El primer paso consiste, entonces, en verificar la estabilidad del filtro definido por los parámetros recibidos; para ello es necesario transformar los parámetros de entrada en parámetros PARCOR, $\mathbf{r}^{(p)}$, o parámetros LSP, $\boldsymbol{\omega}^{(p)}$, que son los que permiten comprobar de forma sencilla dicha estabilidad (condiciones (IV-8) y (IV-9), respectivamente). Es necesario advertir que los parámetros $\mathbf{r}^{(p)}$, $\hat{\mathbf{a}}^{(p)}$ o $\boldsymbol{\omega}^{(p)}$ que recibimos y consecuentemente, todas las parametrizaciones que obtengamos a partir de ellos, serán, en general, distintos de los que se generaron en el origen emisor, ya que estarán eventualmente contaminados por errores de transmisión. Sin embargo, no utilizaremos una notación diferente para señalar este hecho puesto que complicaría, en nuestra opinión, de forma innecesaria las expresiones que manejamos.

A continuación, se puede optar por calcular los parámetros de predicción lineal, $\hat{\mathbf{a}}^{(p)}$, o bien los parámetros LSP, $\boldsymbol{\omega}^{(p)}$; a partir de ambos vectores de parámetros podemos obtener, entonces, una representación espectral del filtro todo-polos, $\hat{H}(\Omega)$. Cuando se utiliza este procedimiento, la elección de unos u otros parámetros no tiene, según hemos comprobado empíricamente (véase el Capítulo V), mayor relevancia, aunque desde luego, en los casos en los que se reciban, precisamente los parámetros $\boldsymbol{\omega}^{(p)}$ es preferible no volver a realizar la transformación hacia parámetros de predicción. En todo caso recordaremos que con los LSP puede verificarse fácilmente la estabilidad del filtro correspondiente, mientras que la otra opción pasa por la obtención de los PARCOR.

Por último, a partir del filtro $\hat{H}(\Omega)$ podemos obtener las energías en cada uno de los canales del banco de filtros, $M[i]$, para acabar obteniendo los parámetros cepstrales, $\mathbf{c}_{\hat{h}}^{(L)}$, o en su caso los melcepstra, $\mathbf{mfc}_{\hat{h}}^{(L)}$.

Sin necesidad de obtener dicho filtro, es también posible calcular una representación "pseudocepstral", $\hat{\mathbf{c}}_{\hat{h}}^{(L)}$, a partir de los parámetros LSP, tal y como describíamos en IV.3.1, y desde luego, después de aplicar la transformación a la escala Mel se consiguen los "pseudomelcepstrum", $\mathbf{mf\hat{c}}_{\hat{h}}^{(L)}$.

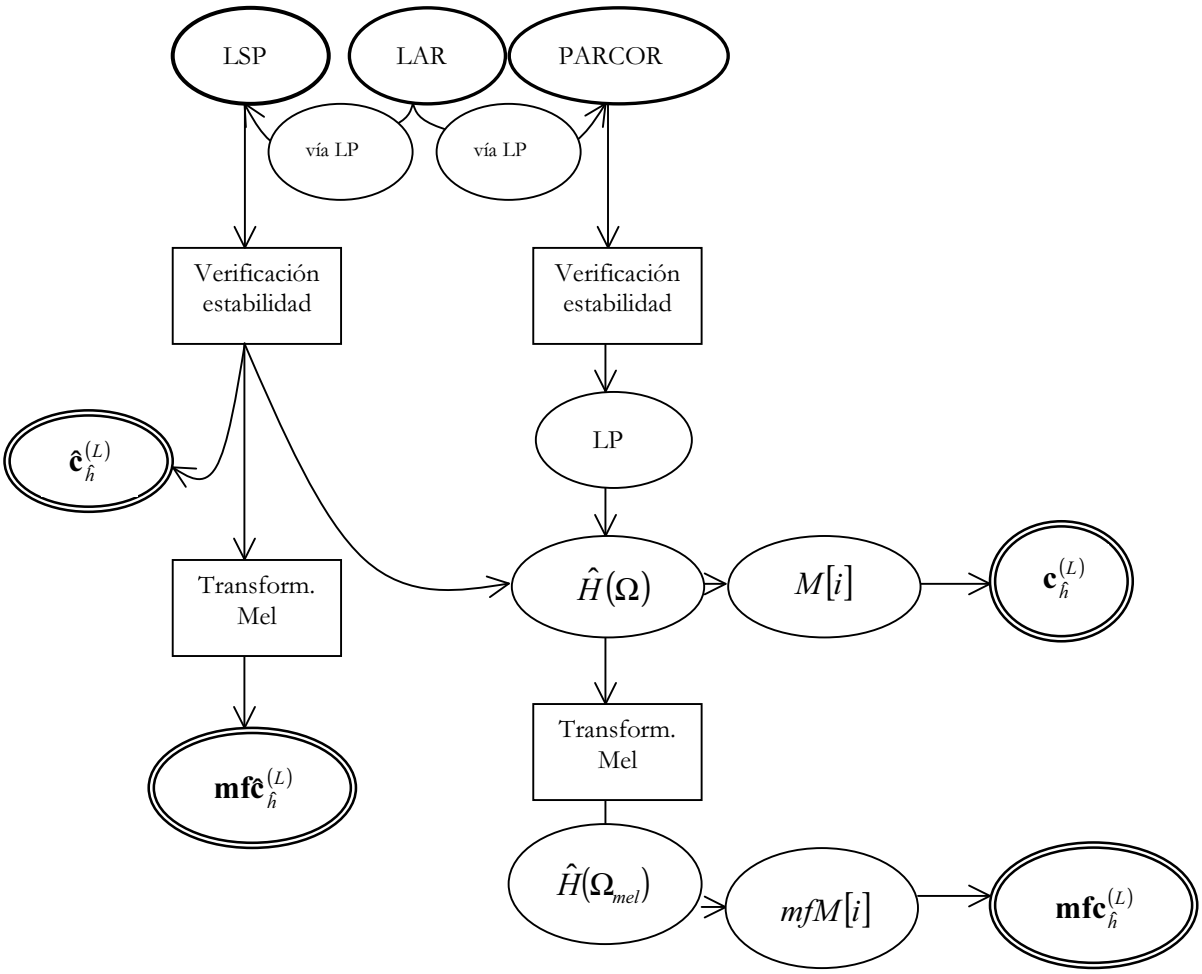


Figura IV-1. Diagrama que muestra las distintas opciones para obtener los parámetros de reconocimiento a partir de los de codificación. En trazo más grueso se muestran los distintos parámetros de partida y con doble trazo las distintas parametrizaciones finales.

Además, Huerta et al. ([85], [86]), proponen extraer también información de la envolvente espectral a partir de la excitación para el caso de GSM-FR. Como se recordará (sección IV.3.1), la excitación, $\tilde{e}[n]$, y el filtro, $\tilde{h}[n]$, se pueden separar ("deconvolucionar") aproximadamente en el dominio cepstral, sin más que tomar las L (entorno a 12) primeras muestras de $c_s[n]$. Sin embargo, cuando el reconocimiento se realiza mediante transparametrización esta deconvolución es realizada por el codificador mediante predicción lineal y se obtienen otras dos señales, $e[n]$ y $\hat{h}[n]$, que modelan las mencionadas componentes.

Debido a que el codificador trata de obtener una parametrización compacta de ambas señales, la exactitud del modelo está bastante limitada. En concreto, la representación de la envolvente espectral es muy sensible al número de coficientes de predicción lineal que se consideran, que en el caso del codificador GSM-FR, es de 8. Es lógico, por tanto, pensar que, puesto que este codificador utiliza el procedimiento de Análisis mediante Síntesis (AbS) parte de la información sobre la envolvente se encuentre en $e[n]$. A partir de la observación de que no hay información nueva en los parámetros cepstrales de orden superior al de predicción cuando los primeros se

calculan desde los segundos (ecuación (IV-17)) no hay información nueva en aquellos parámetros cepstrales de orden superior al de predicción lineal, Huerta et al. introducen dos nuevas alternativas:

$$\mathbf{mfc}_{\hat{h}}^{(12)} + \mathbf{mfc}_{e}^{(12)} \qquad \textbf{(IV-45)}$$

donde las contribuciones de cada una de las componentes se suman una a una (como corresponde a la operación de convolución en el dominio cepstral), o también:

$$\left[mfc_{\hat{h}}^{1}, \ldots, mfc_{\hat{h}}^{j}, mfc_{e}^{j+1}, \ldots, mfc_{e}^{12}\right] \qquad \textbf{(IV-46)}$$

donde se ha tenido en cuenta que son los parámetros cepstrales de orden superior los peor aproximados debido al bajo orden de predicción. De hecho, el valor óptimo del parámetro *j* encontrado por estos autores se sitúa entre 7 y 8, coincidiendo con el orden de predicción.

Efectivamente, en nuestra opinión, el inferior comportamiento en reconocimiento de este codificador si lo comparamos con el resto de los que hemos evaluado en esta tesis, se debe en gran medida a este hecho. Sin embargo, como expusimos en III.1, este es el codificador más antiguo y el único que utiliza tan bajo orden de predicción. Por otra parte, Huerta et al., han verificado el funcionamiento de esta parametrización utilizando ruido aditivo rosa. Por los motivos que hemos venido exponiendo (desigual protección del codificador de canal, distorsión de codificación, desigual probabilidad de error debido al diferente número de bits empleados en la descripción de fuente y filtro, etc.) no nos parece conveniente realizar la decodificación de la excitación cuando contamos con la distorsión añadida de los errores de transmisión, por ello hemos ensayado otras estrategias (introducción de una estimación de la energía, intepolación de los parámetros para obtener tasas de trama inferiores, empleo de los procedimientos de recuperación de tramas, etc.).

Choi et al. ([24], [25]), por su parte, proponen realizar el reconocimiento a partir de modificaciones de los parámetros LSP. Estas modificaciones tienen en cuenta la sensibilidad espectral de cada uno de los parámetros y aunque los resultados que se consiguen mejoran los obtenidos sin esta modificación, quedan todavía bastante por debajo de los conseguidos con los parámetros cepstrales. Por ese motivo, Choi et al. proponen, la utilización de los "pseudomelcepstra". Estos parámetros proporcionan unos resultados de reconocimiento ligeramente inferiores a los obtenidos con los originales, pero permiten su obtención de una forma muy eficiente y directa. En nuestra opinión y dado que la mayor parte del esfuerzo computacional en reconocimiento de habla no se debe a la parametrización, esta solución no tiene gran relevancia práctica.

Choi et al., sin embargo, no consideran ningún tipo de distorsión de canal, motivo por el cual hemos considerado oportuno verificar su comportamiento con el codificador G.723.1 actuando en un entorno con pérdidas de paquetes. Como cabía esperar, las diferencias de resultados entre el pseudocepstrum y el cepstrum convencional aumentan a medida que los efectos del canal son más severos (véase el Capítulo V).

Es necesario advertir además, que la simplificación implícita en el pseudocepstrum evita el cálculo de $\hat{H}(\Omega)$ que, sin embargo, resulta necesario para extraer información sobre la energía de la trama cuando ésta no es codificada y transmitida como parámetro

independiente por el codificador. Por ello, hemos propuesto como solución de complejidad computacional intermedia, la obtención de los parámetros cepstrales a partir del filtro $\hat{H}(\Omega)$, calculado, para el G.723.1, directamente a partir de los parámetros LSP (ecuaciones (IV-18) y (IV-19)). De esta forma los resultados de reconocimiento son muy similares y se evita hacer transparametrizaciones innecesarias (véase la Figura IV-1).

Finalmente, la opción que adoptan Kim et al. ([112]) consiste en añadir parámetros adicionales, que en este caso representan la información sobre el caracter sonoro o sordo de la trama correspondiente, y que se encuentran en la parametrización del codificador (en este caso, las ganancias adaptativa y fija). Esta alternativa resulta muy interesante porque los parámetros que se añaden forman parte del grupo de los más protegidos habitualmente por el codificador de canal; sin embargo, la forma de añadirlos debe tener en cuenta la relevancia de los mismos para el reconocimiento y por este motivo los autores establecen una ponderación sobre el correspondiente a la librería fija. En todo caso, nos parecería conveniente evaluar, los beneficios de añadir tal información (aunque plasmada de otra manera) sobre la opción de referencia (es decir, la que parte de la voz decodificada).

IV.5.2. Energía

Como ya hemos apuntado, la energía de cada trama de voz forma parte normalmente del vector de parámetros de reconocimiento, ya que contribuye notablemente a mejorar las prestaciones de los reconocedores. Sin embargo, este parámetro no siempre está presente como tal en las parametrizaciones de los codificadores. Por ejemplo, en el codificador GSM-HR sí que se envía un parámetro que representa la energía de la trama, mientras que en el GSM-FR y el G.723.1 esto no es así y resulta conveniente obtener una estimación de la misma.

De esta forma conseguimos que sólo los parámetros que contienen algo de información acerca de la energía intervengan en proceso de estimación, y evitamos que tanto la distorsión de decodificación debida a la pérdida de información relativa a la excitación, como los errores de transmisión que afectan a los parámetros que no contienen información relevante, degraden la estimación. No sin motivo, los parámetros que finalmente utilizamos para esta estimación son los más protegidos por el codificador de canal, en el caso de GSM-FR (véase Tabla III-4).

Para ello, nos basamos en el modelo fuente-filtro que hemos descrito en IV.2.1, y asumiremos que la excitación, $e[n]$, de cada subtrama puede modelarse como ruido blanco gaussiano de media nula. Esta premisa es perfectamente asumible si tenemos en cuenta que el propio proceso de obtención del filtro $h[n]$ tiene como objetivo el blanquedo de esta señal residual (aunque en las tramas sonoras $e[n]$ retiene la periodicidad). Así, podemos obtener la potencia media, σ_s^2, de la señal de voz sintetizada correspondiente a una subtrama de la siguiente manera:

$$\sigma_s^2 = \sigma_e^2 \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| H(\Omega) \right|^2 d\Omega \qquad \textbf{(IV-47)}$$

donde σ_e^2 es la varianza de la excitación y $H(\Omega)$ es la respuesta en frecuencia del filtro de síntesis.

Podemos por tanto, calcular la potencia media estimada para cada una de las N_{sf} subtramas i ($0 \le i < N_{sf}$) de la trama k, $\hat{\sigma}_s^2[k,i]$ como:

$$\hat{\sigma}_s^2[k,i] = \sigma_e^2[k,i] \cdot \hat{E}_h[k,i] \qquad \textbf{(IV-48)}$$

donde $\sigma_e^2[k,i]$ y $\hat{E}_h[k,i]$ representan las estimaciones de la varianza de la excitación y la contribución del filtro de síntesis, respectivamente. A partir de ahora, prescindiremos de los índices de trama y subtrama, *k* e *i*, para facilitar la lectura y los consideraremos de nuevo cuando sea necesario.

Para calcular la contribución del filtro, $\hat{E}_h$, podemos aproximar la integral de la ecuación (IV-47) como sigue:

$$\hat{E}_h = \frac{1}{N} \sum_{r=0}^{N-1} \left| H\left(\tfrac{2\pi}{N} r\right) \right|^2 \qquad \textbf{(IV-49)}$$

Además, también es posible calcular $\hat{E}_h$ a partir de los parámetros PARCOR, r_i, sin necesidad de calcular la envolvente espectral haciendo uso de la siguiente identidad:

$$\hat{E}_h = \frac{1}{\prod_{i=1}^{p} (1 - r_i^2)} \qquad \textbf{(IV-50)}$$

donde *p* es el orden de predicción que, en el codificador GSM-FR, es 8.

Para estimar la contribución de la excitación, σ_e^2, hay que examinar la forma en la que cada codificador describe esta componente. En todo caso, lo habitual que de hecho ocurre en todos los codificadores considerados en esta tesis, es que la excitación esté compuesta por una contribución procedente de una librería adaptativa que captura la componente periódica y otra procedente de una librería fija:

$$e[n] = u[n] + v[n] \qquad \textbf{(IV-51)}$$

donde $u[n]$ representa la contribución de la librería adaptativa y $v[n]$ la de la fija.

Ahora, para la estimación de la varianza de la excitación asumiremos que sus contribuciones adaptativa y estática están incorreladas y que por tanto:

$$\hat{\sigma}_e^2 = \hat{\sigma}_u^2 + \hat{\sigma}_v^2 \qquad \textbf{(IV-52)}$$

donde $\hat{\sigma}_u^2$ y $\hat{\sigma}_v^2$ son las estimaciones de las varianzas correspondientes.

Así, en el caso del codificador G.723.1, el primer vector de excitación se construye a partir de un predictor largo de orden 4:

$$u[n] = \sum_{j=0}^{4} \beta_{ij} e'[n+j], \qquad 0 \le n < L_{sf} \qquad \textbf{(IV-53)}$$

donde L_{sf} es la longitud en muestras de la subtrama, β_{ij} es el coeficiente *j*-ésimo del predictor largo para la subtrama *i* ($0 \le i \le 3$) y la señal $e'[n]$ se genera de la siguiente forma:

$$\begin{aligned} &e'[0] = e[-L_i - 2], \\ &e'[1] = e[-L_i - 1], \\ &e'[n] = e[(n \bmod L_i) - L_i - 2], \qquad 2 \le n \le L_{sf} + 3 \end{aligned} \qquad \textbf{(IV-54)}$$

siendo L_i el periodo fundamental obtenido para esa subtrama.

Si nos fijamos de nuevo en el procedimiento para generar la excitación adaptativa de la ecuación (IV-53), podemos obtener una estimación de $\hat{\sigma}_u^2$ como sigue:

$$\hat{\sigma}_u^2[k,i] = \frac{1}{L_{sf}} \sum_{n=0}^{L_{sf}-1} \left(\sum_{j=0}^{4} \beta_{ij} e'[n+j] \right)^2 \qquad \textbf{(IV-55)}$$

Sin embargo, como nuestro objetivo es evitar la reconstrucción de la señal de excitación $e[n]$, para evitar, en la medida de lo posible, que los parámetros (eventualmente afectados por errores de transmisión) que no contienen información sobre la energía perjudiquen su estimación, decidimos recurrir a la siguiente aproximación:

$$\hat{\sigma}_u^2[k,i] \approx \sum_{j=0}^{4} \beta_{ij}{}^2 \begin{cases} \hat{\sigma}_e^2[k, l_j] & l_j \ge 0 \\ \hat{\sigma}_e^2[k-1, N_{sf} - l_j] & l_j < 0 \end{cases} \qquad \textbf{(IV-56)}$$

habiendo despreciado los productos cruzados de la suma cuadrática y siendo l_j un indicador de la subtrama a la que pertenecen la mayoría de las muestras de la señal $e'[n+j]$ con $0 \le n < L_{sf}$ y que podemos expresar como:

$$\begin{cases} l_j = \left\lfloor i + \frac{1}{2} - \dfrac{L_i - j}{L_{sf}} \right\rfloor & L_i - j > \dfrac{L_{sf}}{2} \\ l_j = i - 1 & L_i - j \le \dfrac{L_{sf}}{2} \end{cases} \qquad \textbf{(IV-57)}$$

donde $\lfloor x \rfloor$ denota el mayor entero menor o igual que x.

Para la estimación de la contribución aperiódica, $\hat{\sigma}_v^2$, hemos de considerar el número de pulsos que la conforman, N_p, que en el caso del codificador G.723.1 es de 4, y la ganancia *G* que se aplica sobre la librería estática y entonces:

$$\hat{\sigma}_v^2 = \frac{N_p}{L_{sf}} \cdot G^2 \qquad \textbf{(IV-58)}$$

Si nos fijamos ahora en la interpolación de los vectores LSP que hace el decodificador,

$$\hat{\boldsymbol{\omega}}^{(p)}[k,i] = \begin{cases} 0.75\,\hat{\boldsymbol{\omega}}^{(p)}[k-1,3] + 0.25\,\hat{\boldsymbol{\omega}}^{(p)}[k,3], & i=0 \\ 0.50\,\hat{\boldsymbol{\omega}}^{(p)}[k-1,3] + 0.50\,\hat{\boldsymbol{\omega}}^{(p)}[k,3], & i=1 \\ 0.25\,\hat{\boldsymbol{\omega}}^{(p)}[k-1,3] + 0.75\,\hat{\boldsymbol{\omega}}^{(p)}[k,3], & i=2 \\ \hat{\boldsymbol{\omega}}^{(p)}[k,3], & i=3 \end{cases} \qquad \textbf{(IV-59)}$$

donde $\hat{\boldsymbol{\omega}}^{(p)}[k,i]$ es el vector de LSP utilizado en la reconstrucción de la subtrama *i* de la trama *k*, nos damos cuenta de que la excitación que corresponde al vector de parámetros LSP del que disponemos antes de la decodificación es la correspondiente a la subtrama 3 de cada trama, $\hat{\sigma}_x^2[k,3]$, y por tanto, será esta última la única que estimaremos.

Esto es importante porque el vector de parámetros LSP, $\boldsymbol{\omega}^{(p)}[k,3]$, que representa a toda la trama, está calculado a partir de una ventana centrada en la subtrama 3 (que abarca la subtrama anterior y el *look-ahead*). Los filtros de síntesis de las restantes subtramas, utilizados tanto en el codificador para la obtención de la excitación (procedimiento de análisis mediante síntesis) como en el decodificador, se obtienen mediante interpolación de los parámetros LSP tal y como se expresa en (IV-59). Así, sólo la excitación de la tercera subtrama está preparada para reconstruir correctamente la señal original con dicho vector, $\boldsymbol{\omega}^{(p)}[k,3]$.

Así, este proceso de obtención del parámetro de energía es beneficioso para la alternativa de reconocimiento mediante transparametrización que proponemos, en el entorno IP, principalmente por dos motivos:

- De esta forma estamos evitando que el proceso de interpolación diseñado en el codificador estándar intervenga en el proceso de reconocimiento. Como ya hemos comentado en IV.1, el proceso de codificación y decodificación asume una serie de exigencias sobre el retardo algorítmico que no aplican al caso de reconocimiento. La ecuación (IV-59), pone de manifiesto este punto ya que muestra cómo la interpolación de los LSP, muy conveniente tanto en codificación como en reconocimiento para evitar transiciones bruscas entre tramas, se realiza de forma muy burda y contando tan solo con la información de la trama anterior. En IV.2.1 se expondrán otras interpolaciones que han resultado ventajosas para la tarea de reconocimiento.

- Por otra parte, la pérdida de tramas característica del entorno IP en el que se suele utilizar este codificación, hace que esta estimación de la energía en la que la excitación y el filtro de la tercera subtrama representan la energía de toda la trama sea ventajosa respecto a las que utilizan los vectores LSP interpolados, puesto que las sustituciones que se llevan a cabo con el fin de suavizar el efecto de dichas pérdidas hacen que haya un desajuste entre la excitación recibida en una trama sin errores y el filtro de síntesis interpolado a partir de la trama actual y de

una trama anterior, eventualmente perdida y consecuentemente sustituida. Nótese que este no este desajuste no tiene lugar cuando realizamos la interpolación que proponemos para adecuar la tasa de trama en la próxima sección.

En todo caso, merece la pena resaltar aquí cómo para la estimación de la energía que proponemos, tan sólo hemos añadido a los parámetros LSP necesarios para obtener los MFCC, la ganancia de la librería estática, *G*, el periodo fundamental, L_i y los coeficientes del predictor largo, β_{ij}.

Estudiamos ahora el caso del, en el codificador GSM-FR, el predictor utilizado para la construcción de la excitación adaptativa es bastante más sencillo, ya que utiliza un solo coeficiente β_i:

$$u[n] = \beta_i e[n - L_i], \qquad 0 \le n \le L_{sf} \tag{IV-60}$$

Por tanto, de la misma forma que en el codificador G.723.1, la contribución adaptativa puede calcularse de la siguiente manera:

$$\hat{\sigma}_u^2[k,i] = \frac{1}{L_{sf}} \sum_{n=0}^{L_{sf}-1} (\beta_i e[n - L_i])^2 \tag{IV-61}$$

De nuevo, evitamos decodificar $e[n]$, y de forma análoga a cómo actuamos en el caso anterior, proponemos aproximar $\hat{\sigma}_u^2$ como sigue:

$$\hat{\sigma}_u^2[k,i] \approx \begin{cases} \beta_i^2 \hat{\sigma}_e^2[k,l] & l \ge 0 \\ \beta_i^2 \hat{\sigma}_e^2[k-1, N_{sf} - l] & l < 0 \end{cases} \tag{IV-62}$$

donde *l* es la subtrama donde se sitúa al menos la mitad de la porción de señal $e[n]$ de longitud L_{sf} que constituye la contribución adaptativa de la subtrama que estamos considerando, según indica la ecuación (IV-61) y que se puede obtener como:

$$l = \left\lfloor i + \tfrac{1}{2} - \frac{L_i}{L_{sf}} \right\rfloor \tag{IV-63}$$

donde $\lfloor x \rfloor$ denota el mayor entero menor o igual que x.

Para la descripción de la componente aleatoria de la excitación, hemos aprovechado el hecho de que además de la información que codifica la secuencia concreta de pulsos que constituyen la misma, se codifica y transmite un factor de escala general que afecta a estos pulsos para cada subtrama, $X_{\max_i}$.

Asumiendo que este factor de escala se puede considerar proporcional a la desviación estándar de esta subtrama $\hat{\sigma}_v[k,i]$, o lo que es lo mismo:

$$X_{\max_i} = K \hat{\sigma}_v[k,i] \tag{IV-64}$$

podemos entonces obtener la estimación deseada $\hat{\sigma}_v^2[k,i]$ de la siguiente manera:

$$\hat{\sigma}_v^2[k,i] = \left(\frac{X_{\max_i}}{K} \right)^2 \qquad \textbf{(IV-65)}$$

donde *K* es una constante de proporcionalidad que se puede obtener empíricamente a partir de una reducida base de datos.

Es preciso señalar, de nuevo, que para la estimación de σ_s^2 solamente se han empleado adicionalmente, los parámetros β_i, L_i y $X_{\max_i}$ de cada subtrama.

Este procedimiento conduce a una parametrización más robusta en el entorno GSM debido fundamentalmente a dos razones:

- En primer lugar, cuanto menor sea el número de parámetros empleados, más se reduce la probabilidad de que un error en alguno de ellos distorsione el cálculo de la energía. La inclusión de parámetros que no contienen información acerca de la energía sólo responde al deseo de simplificar su estimación utilizando directamente los procedimientos incluidos en el decodificador estándar, como es el caso de Kim y Cox en [112]. Sin embargo, debido a que la información sobre la energía contenida en estos parámetros es muy baja, solamente pueden contribuir a empeorar la estimación, cuando contienen errores. En el caso de la aproximación de Huerta ([86]), por el contrario, la energía se obtiene directamente del parámetro cepstral, c_0, respondiendo, por tanto, a la energía del filtro LP.

- En segundo lugar, el codificador de canal pone más énfasis en proteger los parámetros perceptualmente relevantes. La energía está ciertamente entre ellos, y la elección de β_i, L_i y $X_{\max_i}$, junto con los parámetros espectrales, para realizar su estimación está plenamente justificada desde este punto de vista, puesto que un 52% de los bits empleados en cuantificar estos parámetros pertenece a la clase Ia (máxima protección), un 30% pertenece a la clase Ib (protección media) y el 18% restante a la clase II (sin protección), mientras que entre los bits que cuantifican el resto de parámetros no hay ninguno perteneciente a la clase Ia, el 63% pertenece a la clase Ib y el restante 37% a la clase II.

Es importante señalar, en este punto, que aunque en esta tesis hemos desarrollado procedimientos de estimación de la energía para los codificadores G.723.1 y GSM-FR, es posible encontrar procedimientos similares para todos aquellos codificadores que tengan una estructura semejante.

IV.5.3. Tasa de tramas

Aunque el análisis por tramas de tamaños tales que permitan hacer la suposición de estacionariedad dentro de la misma está muy arraigado dentro de los sistemas de reconocimiento y tiene su origen, precisamente, en las técnicas de codificación, algunos autores consideran que es necesario incluir información de una escala temporal superior ([81], [83] y [84]). De hecho, como ya hemos comentado, este motivo está detrás del uso de los parámetros dinámicos que introdujimos en la sección IV.3.3 y también de las

estrategias de filtrado del espectro de modulación que mencionábamos en el contexto de reconocimiento robusto en la sección IV.3.5. Sin embargo, se advierte en [103] que el uso de este tipo de suavizado, contribuye a la diseminación de la información de cada trama entre sus adyacentes, de forma que es recomendable, en estos casos, el uso de unidades de reconocimiento de mayor longitud, tales como modelos de trifonemas o de palabra.

En cuanto a la codificación, el uso de información perteneciente a otras tramas se encuentra muy limitado por la restricción de retardo. De hecho, cuando el objetivo es el de suavizar las transiciones de la voz reconstruida a través de la interpolación de los parámetros espectrales, la energía de tramas sucesivas, o optimizar la cuantificación de algunos parámetros con la cuantificación diferencial, tan sólo suele hacerse uso de la trama anterior.

En el caso de reconocimiento mediante transparametrización que nos ocupa, hemos verificado que la tasa de tramas juega un papel muy importante, sobre todo en la tarea de reconocimiento de habla continua. Por este motivo hemos propuesto el uso de interpoladores, tanto el los codificadores GSM –que contemplan tasas de trama de 20 ms– como en el G.723.1 –cuya tasa de trama es de 30 ms–, de forma que se obtengan siempre tasas de 10 ms, más adecuadas para reconocimiento. Estos filtros interpoladores de tipo FIR consideran para la síntesis de las tramas interpoladas de 3 tramas anteriores y posteriores (véase el Capítulo V). Estas longitudes sitúan, el intervalo temporal que influye en cada trama en 220 ms, más cercano a la duración de las sílabas.

Además, el hecho de realizar la interpolación directamente sobre los parámetros del codificador evita la influencia del mecanismo de interpolación tan sencillo propio del codificador que repercute negativamente, en la voz decodificada.

Capítulo V
Experimentos y resultados

En el Capítulo II introdujimos la configuración de reconocimiento remoto en contraposición a las de reconocimiento distribuido y local y justificamos su interés para el desarrollo de servicios basados en la tecnología de reconocimiento de habla.

Esto nos llevó, en el Capítulo III, a discutir las dificultades que entraña este tipo de reconocimiento cuando se realiza, en particular, en entornos móviles y en redes IP. Concretamente, además de otras distorsiones también presentes en otros entornos y que, por ese motivo, no hemos tenido en cuenta aquí, encontramos que las dos distorsiones características que nos atañen son: la de codificación y la producida por los errores de transmisión (en forma de ráfagas de errores de bit, en las redes GSM, y en forma de paquetes perdidos –muchas veces a ráfagas–, en las redes IP).

En el Capítulo IV presentamos nuestra propuesta de reconocimiento mediante transparametrización –al que en adelante nos referiremos como RMT)– como alternativa al reconocimiento convencional –es decir, al reconocimiento a partir de voz decodificada–. Así, analizamos los elementos más importantes para la realización de este procedimiento y expusimos los motivos por los que esta aproximación proporciona mejores resultados que la convencional.

En este capítulo, vamos a exponer los experimentos que hemos llevado a cabo para evaluar esta alternativa y compararla con la convencional, en un entorno de simulación lo más cercano posible a la realidad teniendo en cuenta nuestros medios. En primer lugar, por tanto, presentaremos el sistema de referencia que hemos utilizado: las bases de datos de las que disponemos, la forma en que hemos caracterizado los canales GSM e IP, los codificadores que hemos empleado, la descripción de la parametrización y del reconocedor de referencia y por último, las medidas de confianza que hemos considerado.

Finalmente, presentaremos el análisis de los resultados obtenidos en este entorno de referencia comenzando por los obtenidos en GSM, seguidos de los correspondientes al entorno IP y finalmente, los que incluyen escenarios en los que se producen transcodificaciones entre distintos codificadores, bien sea dentro de la red GSM o bien entre las dos redes consideradas.

V.1. Sistema de referencia

En el diagrama de bloques de la Figura V-1, mostramos los elementos que constituyen nuestro sistema de evaluación de la propuesta de reconocimiento mediante transcodificación. La parte superior del mismo representa la secuencia de procedimientos empleados para evaluar las tasas de reconocimiento del reconocedor convencional (sistema de referencia) y la parte inferior, los bloques que implementan (en sustitución de los correspondientes en la parte superior) la alternativa propuesta.

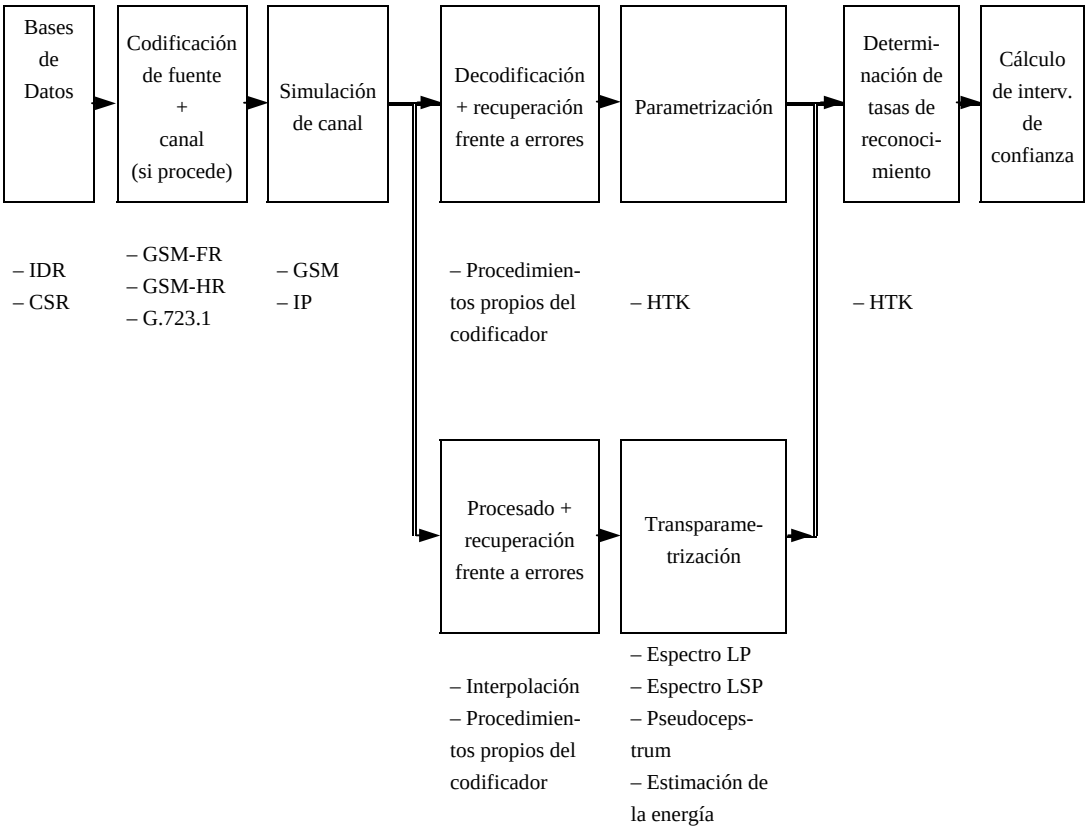


Figura V-1.- Diagrama de bloques del sistema de referencia.

Así, cada una de las siguientes secciones analiza uno de estos bloques que conforman el sistema de referencia:

- En primer lugar, presentaremos las dos bases de datos que hemos utilizado y que nos permiten evaluar nuestra alternativa para dos tareas de reconocimiento distintas: dígitos aislados (IDR –*Isolated Digit Recognition*–) y habla continua (CSR –*Continuous Speech Recognition*–).

- El segundo paso consiste en codificar la voz, para lo cual hemos utilizando los codificadores GSM-FR, GSM-HR (incluyendo sus codificadores de canal correspondientes) y G.723.1.

- A continuación, para cada uno de los dos entornos (GSM e IP), realizamos una simulación de los efectos del canal para una variedad de

condiciones, de forma que obtenemos la voz codificada (es decir, parametrizada para su reconstrucción) y contaminada que se recibiría al otro extremo de cada una de las redes.

- El procedimiento convencional contempla, ahora, la decodificación de esa señal de voz, de manera que lo que se obtiene es la señal de voz reconstruida (incluyendo las correcciones introducidas por los mecanismos de recuperación frente a errores recomendados en los estándares). En RMT, sin embargo, estos procedimientos se aplican directamente sobre los parámetros sin decodificar y además se añade una interpolación con el objetivo de adecuar las tasas de trama de codificador y reconocedor.

- Una vez obtenida la señal decodificada (en el procedimiento convencional), se procede a la parametrización de la misma para reconocimiento, para lo que utilizamos la herramienta HTK (*–Hidden markov model ToolKit–*) [177]. En el caso de RMT, sin embargo, se procede a la transparametrización pudiéndose utilizar todos los procedimientos descritos en el Capítulo IV.

- Una vez obtenidos los vectores acústicos (por ambos procedimientos), determinamos las tasas de reconocimiento de referencia y las de RMT, correspondientes a cada una de las tareas y condiciones de canal consideradas.

- Finalmente, calculamos los intervalos de confianza que nos permitirán valorar cuándo las mejoras introducidas por la solución alternativa que proponemos son estadísticamente significativas.

V.1.1. Bases de Datos

V.1.1.1. Reconocimiento de dígitos aislados

Para la tarea de reconocimiento de dígitos aislados (IDR), hemos utilizado una base de datos consistente en 11 repeticiones de los 10 dígitos en castellano (del “cero” al “nueve”) para un total de 72 locutores, grabados a 8 KHz. en condiciones limpias. Debido a que esta base resulta escasa para obtener resultados fiables para reconocimiento independiente del locutor hemos realizado una validación cruzada dividiendo el total de locutores en 9 grupos de 8 cada uno. Tomando cada uno de estos grupos como conjunto de validación, hemos definido el conjunto de entrenamiento como los restantes 8 y finalmente hemos realizado un promediado entre los resultados de las 9 particiones del conjunto así descritas. Este procedimiento hace posible que consideremos las 7920 elocuciones para el cálculo de los intervalos de confianza estadística.

V.1.1.2. Reconocimiento de habla continua

La base de datos que hemos utilizado para la tarea de reconocimiento de habla continua es la conocida RM1 (*–Resource Management Part 1–*) ([137]), con un

vocabulario de 991 palabras. El corpus de entrenamiento independiente del locutor está compuesto por 3990 oraciones proferidas por 109 locutores. Por su parte, el conjunto de validación se compone de 1200 frases procedentes de 40 locutores correspondientes a la compilación de los cuatro conjuntos de validación oficiales. Esta base se grabó, originalmente, a 16 KHz en condiciones limpias; sin embargo, en nuestros experimentos hemos utilizado una versión diezmada a 8 KHz (para adecuarla a los codificadores estándar que empleamos).

V.1.2. Codificadores

Para la evaluación de nuestra propuesta hemos escogido tres de los codificadores estándar que actúan en los dos entornos que consideramos: para el caso de GSM, hemos dispuesto del GSM-FR y el GSM-HR (incluyendo el codificador y decodificador de canal, en ambos casos) y para el de IP, del G.723.1. Como ya comentados este último codificador tiene la posibilidad de codificar voz a dos regímenes binarios: 5,3 y 6,3 Kb/s. Para los experimentos que hemos realizado, nos hemos centrado en el más bajo de ellos puesto que después de realizar algunos experimentos preliminares con ambos, el comportamiento del sistema era prácticamente idéntico.

Además, para la evaluación de varias etapas de codificación hemos utilizado el codificador de ITU-T G.726, un ADPCM que opera a regímenes binarios de 40, 32, 24 y 16 Kb/s ([96]). De todos ellos, hemos utilizado el denominado en el estándar como "modo primario" a 32 Kb/s por ser el más habitual en las redes troncales de telefonía (dicho modo proporciona calidades por encima de 4 en la escala MOS).

Para la implementación del sistema, hemos partido de las especificaciones estándar de todos los codificadores mencionados, así como de una versión en código C de los mismos.

V.1.3. Simulación de los canales de transmisión

Uno de los objetivos fundamentales en esta tesis era la evaluación de la pérdida de precisión de los reconocedores producida por los errores de transmisión, tanto en los canales móviles, en los que los errores más característicos son los producidos a ráfagas de bits individuales, como en los canales IP, en los que los errores se producen a ráfagas de paquetes completos. Para ello, necesitamos disponer de modelos con los que caracterizar la forma en la que se producen estos errores, que nos permitan generar unos patrones de error sintéticos con los que contaminar las bases de datos codificadas.

V.1.3.1. Caracterización del canal de transmisión GSM

Para la contaminación de las tramas de la señal de voz codificada en el entorno GSM, hemos utilizado los modelos y resultados presentados en [70], donde se propone un método para simular canales radio GSM basadas en una combinación de resultados teóricos, las especificaciones GSM y datos empíricos.

En la Tabla V-1, se pueden observar las características de los distintos canales de tráfico de voz TCH/FS y TCH/HS que hemos utilizado en nuestras simulaciones.

Así, para cada tasa de error de bit (BER –*Bit Error Rate*–) teórica, calculada a partir de una expresión analítica, se extraen los patrones de error de bit utilizando muestras de una distribución lognormal, que producen una variación entorno a esa BER teórica. El resultado de esta simulación produce una tasa de error bruta (BER bruta), que corresponde a la proporción de bits erróneos que hay antes de realizar la decodificación de canal y el desentrelazado.

Una vez se desentrelazan y decodifican los bits procedentes del canal, el decodificador de canal corrige, en la medida de sus posibilidades, los bits erróneos. Sin embargo, algunas de las tramas contienen tal cantidad de errores que son desechadas. La tasa de tramas erróneas es lo que hemos denotado en la tabla como FER (–*Frame Error Rate*–). Finalmente, la tasa de errores residuales RBER (–*Residual Bit Error Rate*–) es el porcentaje de bits erróneos que todavía permanecen después del decodificador de canal excluidas las pertenecientes a tramas desechadas.

Características de los canales GSM								
	TCH/FS				TCH/HS			
CANALES	BER teórica	BER bruta	FER	RBER	BER teórica	BER bruta	FER	RBER
0	0	0	0 %	0	0	0	0 %	0
1	10^{-4}	0,000111	0 %	0,000014	10^{-4}	0,000131	0 %	0,000010
2	10^{-3}	0,001601	0,019 %	0,000372	10^{-3}	0,001635	0,015 %	0,000265
3	$5{\cdot}10^{-3}$	0,010371	0,405 %	0,003199	$5{\cdot}10^{-3}$	0,010672	0,176 %	0,001622
4	10^{-2}	0,016118	1,130 %	0,005033	10^{-2}	0,016465	0,479 %	0,002753
5	$5{\cdot}10^{-2}$	-	-	-	$5{\cdot}10^{-2}$	0,070999	12,333 %	0,023222

Tabla V-1.- Características de los canales GSM empleados en los experimentos para los canales de tráfico de voz correspondientes, a los codificadores GSM-FR y GSM-HR.

Es conveniente observar cómo, a pesar de que la BER bruta es muy similar en ambos canales (los valores están calculados empíricamente sobre las bases de datos codificadas), la cantidad de tramas erróneas FER es muy superior en el caso del canal TCH/FS. Esto es debido a que el canal TCH/HS utiliza tan sólo la mitad del canal TCH/FS y por lo tanto, el entrelazado –que se realiza en bloques del mismo tamaño (véase la sección III.3.1.1) en ambos casos– y el reordenamiento de bits, resultan más efectivos en el segundo a la hora de evitar ráfagas largas.

Por este motivo no hemos considerado en canal número 5 en TCH/FS puesto que la cantidad de errores que resultan hacen totalmente imposible el reconocimiento. Nótese además, que siempre que utilizamos cada uno de los canales volvemos a emplear el mismo patrón de errores, es decir, no se realizamos de nuevo una simulación distinta del mismo (salvo en el caso de la simulación de transcodificaciones, donde simulamos un canal distinto –aunque con las mismas características– para cada etapa de codificación). Esto es así, porque en otro caso, estaríamos obligados a repetir las simulaciones un mayor número de veces para compararlas entre sí.

V.1.3.2. Caracterización del canal de transmisión IP

La caracterización de la dinámica de la red Internet extremo a extremo, es decir, el modelado del comportamiento que un usuario final puede observar después de que los paquetes de información atraviesen toda la serie de nodos que los encaminan a su destino, es un problema excepcionalmente difícil debido a la gran heterogeneidad de dispositivos y redes que conforman dicha red.

Sin embargo, el gran auge de esta red ha provocado un gran interés en este tema, sobre todo por la necesidad de dimensionar las subredes de forma que se pueda garantizar cierta calidad de servicio (QoS). Resulta, por lo tanto, fundamental presentar la metodología empleada para realizar medidas de los parámetros que caracterizan el comportamiento de la red, que determinan sus prestaciones y que además permiten el desarrollo de modelos que permitan predecir su comportamiento. Por este motivo, y con el objetivo de proporcionar una visión general de la complejidad del problema, haremos a continuación una pequeña introducción sobre dicha metodología.

En todo caso, para la evaluación de los reconocedores de habla que proponemos en esta tesis, hemos simulado uno sólo de los aspectos que contribuyen a la QoS: la pérdida de paquetes. Como expusimos en III.1.2, el retardo y la variación del mismo (*jitter*), se reducen, en el caso de la voz, a pérdidas de paquetes, cuando estos superan el máximo retardo permitido por el registro (*buffer*) de entrada.

V.1.3.2.1. Medidas de Calidad de servicio

La realización de medidas sobre los parámetros de QoS tiene dos objetivos principales: el primero y mucho más ambicioso es el de modelar el comportamiento de los elementos que componen el sistema y el segundo, el de proporcionar estimaciones

reales y fiables que ayuden a planificar los recursos de las distintas redes y avalen las prestaciones de un determinado servicio con el fin de facilitar las relaciones comerciales entre cliente y proveedor.

Por lo tanto es fundamental, en ambos casos, que las medidas se realicen de una forma sistemática que permita extraer conclusiones. Para ello, se han desarrollado una serie de directrices por parte de varios organismos de estandarización, gracias a las cuales es posible, comparar las medidas obtenidas por distintos grupos de investigación u organismos.

Son muchos los elementos que intervienen a la hora de proveer una cierta QoS. En ese sentido, la organización ETSI, dentro de su proyecto TIPHON, recomienda una manera de hacer medidas para obtener parámetros de calidad bajo una serie de condiciones que incluyen distintos escenarios en donde las llamadas telefónicas atraviesan diferentes redes ([59] y [60]).

En el caso particular de medidas de parámetros en redes IP exclusivamente, los documentos elaborados por el grupo de trabajo de la IETF denominado IPPM (*–Internet Protocol Performance Metrics–*) [91], o también la recomendación I.380 de ITU-T ([98]), pretenden constituir una guía sobre cómo cuanfiticar retardos, variaciones de retardo y patrones de pérdidas de paquetes.

En general, podemos distinguir entre dos tipos de medidas: las no intrusivas y las intrusivas. Las primeras, se llevan a cabo observando el tráfico natural existente en determinado punto de la red (o incluso en un extremo final). Este tipo de medidas requieren un conocimiento muy amplio del tráfico que circula por el punto de medida para extraer las conclusiones adecuadas; sin embargo, por otra parte, tiene la ventaja de no perturbar el sistema con sus mediciones. Ejemplos de este tipo de análisis se pueden encontrar en [28], donde su autor propone la definición de flujos basándose en ciertas condiciones temporales y de localización del tráfico para un mejor análisis del mismo y además señala las dificultades de la teoría de colas para modelar dicho tráfico.

En las intrusivas, se inyectan de forma controlada paquetes en la red y se recogen, bien en la misma localización (medidas de ida y vuelta *–round-trip–*) o en otra (medidas en un solo sentido *–one-way–*). Este tipo de medidas tiene la ventaja de que permite observar aspectos concretos por medio del control de las características del tipo de tráfico inyectado (por ejemplo, el comportamiento para un tipo de tráfico particular). Sin embargo, tienen el problema de que el propio mecanismo de medida perturba el sistema que se pretende medir, lo cual resulta particularmente nefasto cuando es la propia sonda la que lleva al sistema a la saturación. Bolot et al. fueron pioneros en la realización de medidas de este tipo para el caso de VoIP ([12]); otros autores que han realizado medidas e interpretaciones del tráfico en internet son, por ejemplo, Paxon et al ([143]), que afirman que el tiempo entre las llegadas de paquetes en redes de área extensa no puede ser modelado con procesos de Poisson, salvo en algunos casos muy concretos; Crovella et al., por su parte ([33]), analizan el tráfico debido a transferencias a través de la WWW (*–World Wide Web–*) y afirman que el mismo es “autosimilar”, como en los casos de área local y de área extensa.

Como vemos, dependiendo del tipo de tráfico del que se trate, podemos encontrar resultados muy diferentes. Los estudios de Borella [15] tienen especial relevancia para nosotros porque realizan medidas (de tipo intrusivo) en las que las características del tráfico que se inyecta son las que corresponden a la transmisión de VoIP que nosotros

estamos proponiendo. Concretamente, Borella introduce paquetes UDP de 80 bytes de longitud, con una periodicidad de 30 ms y con una duración de 3 minutos (unos 6000 paquetes). Esta medida se repite cada hora durante 30 días consecutivos, utilizando 3 localizaciones en los Estados Unidos. La conclusión que extrae de estas medidas es que, en general, los paquetes se pierden a ráfagas, siendo las pérdidas de paquetes individuales bastante escasas, y propone una distribución de Pareto para modelar la longitud de estas ráfagas. En todo caso, también hace notar las asimetrías existentes en los enlaces estudiados, así como las diferencias entre enlaces. En la próxima sección expondremos cómo hemos integrado las conclusiones de Borella en el modelo de Gilbert, que es el que hemos utilizado para simular las pérdidas de paquetes y que revisamos seguidamente.

V.1.3.2.2. Simulación del canal IP

Como hemos visto, las perdidas de paquetes en las redes IP, cuando tratamos de transmitir voz, se producen raramente de forma aislada. El modelo de dos estados de Gilbert se ha venido utilizando para simular canales con memoria en otro tipo de redes ([105], [121]); este modelo es el que mostramos en la Figura V-2 y funciona de la siguiente manera: en el primero de los estados (estado 1) la probabilidad de pérdida de paquetes, P_1, es relativamente baja: lo denominamos estado "bueno". Por el contrario la probabilidad de pérdida de paquetes en el estado 2 (estado "malo"), P_2, es considerablemente mayor, es decir, $P_1 << P_2$.

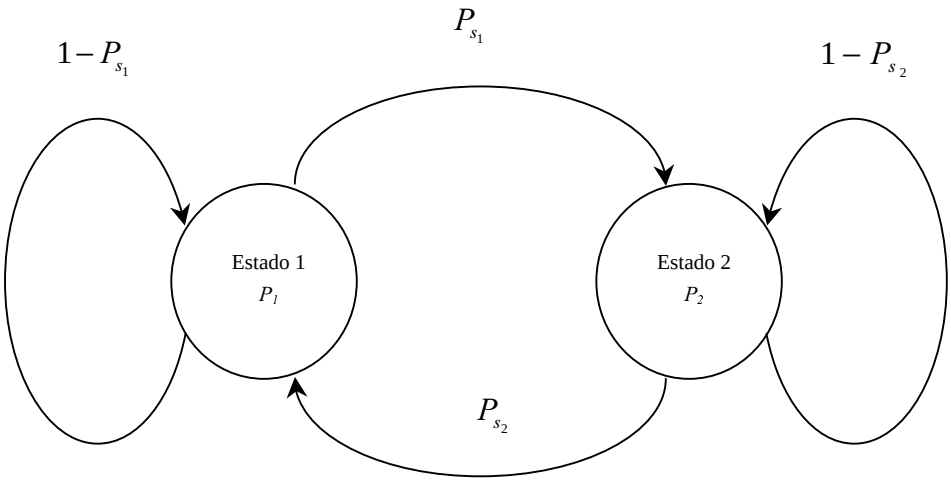


Figura V-2.- Modelo de Gilbert empleado para la simulación de pérdida de paquetes en canales IP.

Los saltos de un estado al otro están gobernados por las probabilidades de transición P_{s_1} y P_{s_2} que indican, respectivamente, la probabilidad de transición del estado 1 al 2 y viceversa. De esta forma, la probabilidad (global) de pérdida de paquetes queda de la siguiente manera:

$$P_e = \frac{P_1 P_{s_2} + P_2 P_{s_1}}{P_{s_1} + P_{s_2}} \qquad \textbf{(V-1)}$$

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

Características de los canales IP							
Canales	P_{s_1}	P_{s_2}	P_1	P_2	PLR	MBL	Percentil 90 longitud de las ráfagas
A	0,001	0,3	0,001	0,85	0,30%	1,76	≤ 3 tramas
B	0,002	0,25	0,005	0,85	1,13%	1,61	≤ 3 tramas
C	0,005	0,25	0,01	0,85	2,54%	1,62	≤ 3 tramas
D	0,005	0,20	0,015	0,85	3,35%	1,63	≤ 3 tramas
E	0,010	0,25	0,025	0,90	5,83%	1,70	≤ 4 tramas
F	0,010	0,20	0,001	0,90	4,11%	3,23	≤ 7 tramas

Tabla V-2.- Características de los canales IP.

Para la configuración de estos cuatro parámetros, P_1, P_2, P_{s_1} y P_{s_2} hemos considerado las conclusiones extraídas del trabajo de Borella ([15]). Concretamente:

- Las tasas de pérdidas de paquetes (PLR –*Packet Loss Rate*–) deben situarse entre el 0,5% y el 3,5%. Sin embargo, dadas las diferencias encontradas entre los diferentes trayectos y las observaciones de Paxon [141] acerca de las diferencias halladas entre Estados Unidos y Europa (con PLRs mayores en esta última), hemos considerado oportuno simular algunos canales con PLRs ligeramente superiores.

- La longitud media de las ráfagas (MBL –*Mean Burst Length*–) es de 6,9 paquetes. Sin embargo, de la observación de los datos proporcionados por Borella, se infiere que ráfagas excepcionalmente largas (que impiden la comunicación totalmente) producen un sesgo muy grande en este valor. Hemos considerado, más conveniente por tanto, reducir considerablemente este valor, puesto que ante este tipo de ráfagas los dos tipos de reconocedores resultarían completamente inútiles.

- El 90% de las ráfagas están compuestas por 3 paquetes o menos, lo que de nuevo nos confirma que aparecen estas ráfagas excepcionalmente largas.

Así, en la Tabla V-2, mostramos las características de los 6 canales que hemos generado distintos siguiendo las directrices anteriores. En primer lugar, los canales que hemos denominado A, B, C y D, exploran distintas PLRs desde el 0,34% hasta el 3,35%, mantenido MBLs entorno a 1,65 y longitudes de ráfaga menores que 3 en el 90% de las ocasiones; en segundo lugar, el canal E está diseñado para explorar PLRs ligeramente superiores (5,83%) teniendo en cuenta los resultados Europeos de Paxon; y finalmente, el canal F tiene en cuenta longitudes de ráfaga muy superiores.

Es preciso señalar aquí, que en los artículos preliminares acerca de reconocimiento mediante transparametrización en entornos IP ([147] y [148]) se han utilizado PLRs bastante superiores (5, 10 y 20 %), que aunque no están basadas en las medidas anteriormente descritas, son también las utilizadas en la recomendación TR 101 329-6 de ETSI, destinada a obtener medidas sobre la calidad de voz en una serie de redes o combinaciones de ellas [58].

V.1.4. Parametrización de referencia

La parametrización empleada en el reconocedor de referencia ha sido la que a nuestro juicio se encuentra más extendida, utilizando para ello la herramienta HTK:

- En primer lugar tomamos ventanas de 25 ms separadas 10 ms entre sí, de forma que dada una frecuencia de muestreo de $f_s = 8000$ muestras/s, la señal enventanada $s_w[n]$ contiene 200 muestras.

- A continuación, aplicamos un preénfasis utilizando la siguiente ecuación en diferencias de orden 1:

$$s'_w[n] = s_w[n] - k_p s_w[n-1] \qquad \textbf{(V-2)}$$

para $n = 2,\ldots,200$ y $k_p = 0.97$. La primera muestra se calcula como:

$$s'_w[1] = s_w[1](1 - k_p) \qquad \textbf{(V-3)}$$

- Después se emplea una ventana de Hamming para obtener:

$$s''_w[n] = \left\{0{,}54 - 0{,}46\cos\left(\frac{2\pi(n-1)}{N-1}\right)\right\} s'_w[n] \qquad \textbf{(V-4)}$$

- El siguiente paso consiste en hallar los coeficientes $M[i]$ (ecuación (IV-23)) utilizando para ello 40 filtros espaciados según la escala Mel (véase la sección IV.3).

- Obtenemos ahora 12 coeficientes cepstrales, $\mathbf{mfc}_s^{(12)}$, utilizando una DCT, y sus correspondientes parámetros delta (véase la sección IV.3).

- Finalmente, añadimos el parámetro de energía también con su correspondiente delta (véase la sección IV.3).

V.1.5. Reconocedor de referencia

Para los reconocedores de referencia de ambas tareas (IDR y CSR) hemos utilizado, como en el caso de la parametrización, las facilidades de la herramienta HTK, configurando ambos reconocedores como sigue:

- Para el desarrollo del reconocedor de dígitos aislados (IDR) independiente del locutor creamos tantos modelos de palabra como dígitos, utilizando para ello HMM (de izquierda a derecha) continuos. El número de estados de cada modelo de dígito es el triple del número de alófonos en la transcripción fonética del mismo y cada uno de ellos está caracterizado por la mezcla de tres gaussianas.

- Para el reconocedor de habla continua (CSR), hemos utilizado modelos acústicos dependientes de contexto, concretamente trifonemas. La síntesis de los trifonemas que no están presentes en el conjunto de entrenamiento se realiza a través de un árbol de decisión de agrupamiento (*–clustering–*) de estados. Como modelo de lenguaje utilizamos la gramática de pares de palabras proporcionada en la distribución de la base de datos.

V.1.6. Medidas de Confianza

Para establecer si las mejoras obtenidas resultan o no estadísticamente significativas, definimos un intervalo de confianza 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.

Así, asumiendo, aunque no es del todo cierto, que la probabilidad de reconocer correctamente una palabra es p (desconocida), constante e independiente de la palabra, el número de palabras correctamente reconocidas, X, es una variable aleatoria binomial caracterizada por dicha probabilidad, p, y el número total de ensayos, n, que corresponde con el número de palabras de la base de datos. Entonces, podemos definir un intervalo de confianza centrado en $\hat{p}$ (estimación de p), que contiene con probabilidad $(1-\alpha)$ la probabilidad, p, que desconocemos.

Para ello, utilizamos un estimador de máxima verosimilitud de p, $\hat{P}$:

$$\hat{P} = \frac{X}{n} \qquad \textbf{(V-5)}$$

cuya media y varianza son las que siguen:

$$E\left(\hat{P}\right)=\frac{1}{n}E\left(X\right)=p$$
$$Var\left(\hat{P}\right)=\frac{1}{n^2}Var\left(X\right)=\frac{p\left(1-p\right)}{n} \qquad \textbf{(V-6)}$$

Por el teorema del límite central, cuando n es suficientemente grande, la distribución de probabilidad de

$$\frac{\hat{P}-p}{\sqrt{\frac{\hat{P}\left(1-\hat{P}\right)}{n}}} \qquad \textbf{(V-7)}$$

tiende a la distribución $N(0,1)$ y por lo tanto la probabilidad del intervalo

$$\left[\hat{P}-z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{P}\left(1-\hat{P}\right)}{n}},\ \hat{P}+z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{P}\left(1-\hat{P}\right)}{n}}\right] \qquad \textbf{(V-8)}$$

es $\left(1-\alpha\right)$, donde $z_{1-\frac{\alpha}{2}}$ es el cuantil $1-\frac{\alpha}{2}$ de la distribución normal estándar. Por ejemplo, para $\alpha=0.05$, es decir, si deseamos obtener un nivel de confianza del 95%, $z_{1-\frac{\alpha}{2}}$ = 1,96.

De acuerdo con lo anterior, el intervalo de confianza que pretendemos establecer se calcula como sigue:

$$\left[\hat{p}-z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}},\ \hat{p}+z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}\right] \qquad \textbf{(V-9)}$$

donde $\hat{p}=x/n$, con x el número de palabras correctamente reconocidas en la realización de un experimento y n el número de palabras que componen la base de datos (7920 en la base de dígitos aislados y 10288 en RM1).

V.2. Sistema de reconocimiento a partir de voz codificada

A continuación comenzamos con la presentación de los resultados obtenidos comparando las opciones de reconocimiento convencional y reconocimiento mediante transparametrización. Así, empezamos con los experimentos realizados en el entorno GSM, continuamos con los correspondientes al entorno IP y finalizamos con los que implican transcodificaciones, bien sea dentro de la propia red GSM o bien entre ésta y la red IP.

En cada una de estas secciones evaluaremos, en primer lugar, la influencia de la codificación sobre el reconocedor automático, en segundo lugar, la influencia de los

errores de transmisión correspondientes a cada entorno particular y finalmente compararemos ambos reconocedores bajo estas dos fuentes de distorsión.

V.2.1. GSM

En la Figura V-3 mostramos un diagrama de bloques en el que están representadas las transformaciones de parámetros necesarias para obtener, por una parte, la parametrización correspondiente a la opción convencional (parte inferior) y la obtenida mediante transformación de los parámetros de codificación en parámetros de reconocimiento (parte superior).

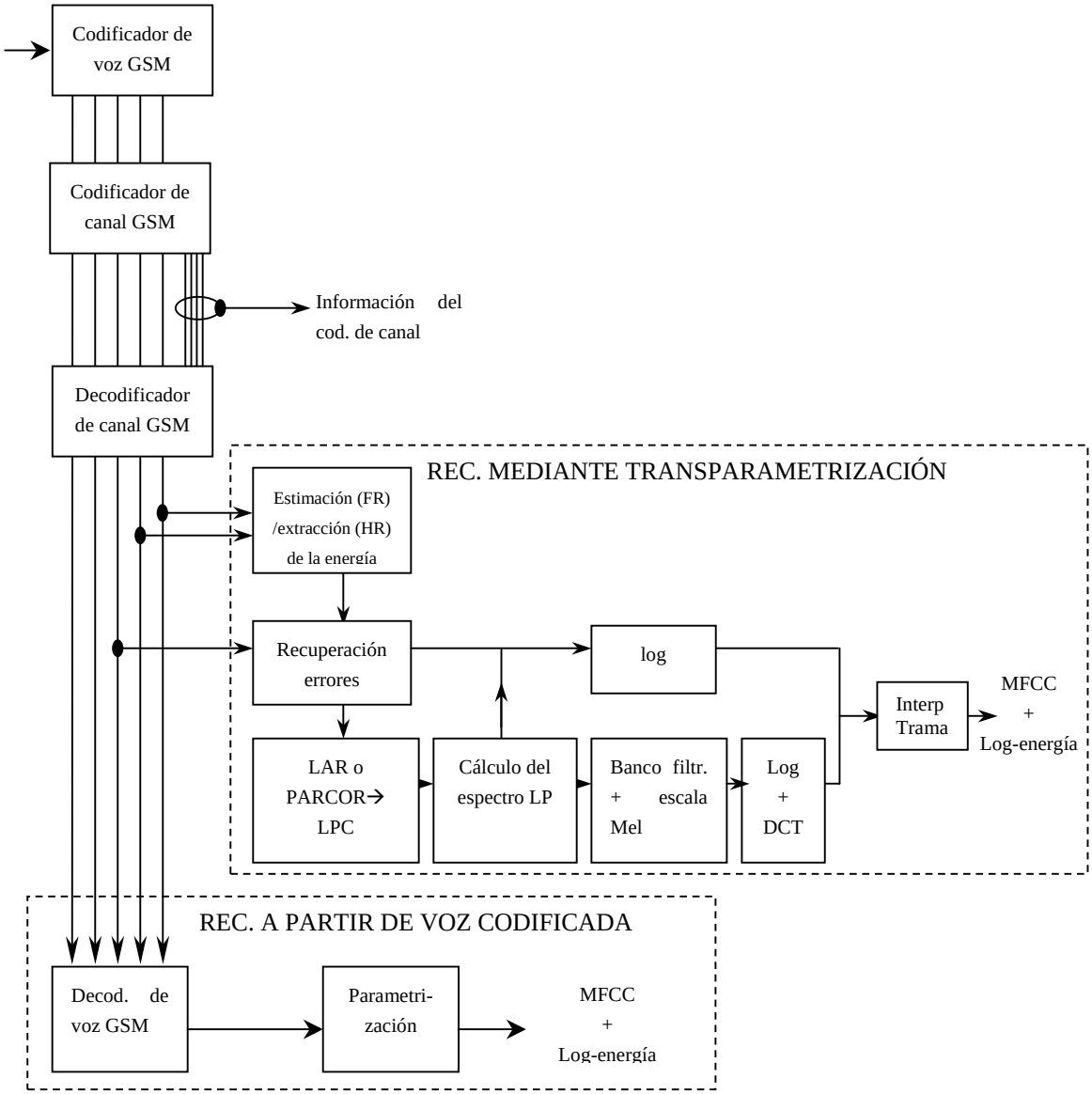


Figura V-3.- Diagrama de bloques que muestra, en la parte superior, las transformaciones necesarias para obtener los parámetros de reconocimiento según el procedimiento que proponemos, y en la parte inferior, el procedimiento convencional, en el caso de considerar la red GSM.

Nótese cómo, a diferencia del diagrama correspondiente al entorno IP (V.2.2), existen dos bloque correspondientes al codificador y el decodificador de canal. Además, en el caso del codificador HR, el parámetro de la energía se extrae directamente de la trama de voz codificada puesto que ésta se codifica explícitamente; mientras que en el caso del codificador FR, es necesario hacer una estimación de la misma a partir de varios de los parámetros de codificación. Sin embargo, como se recordará, estos parámetros están entre los más protegidos por el codificador de canal.

V.2.1.1. Influencia de la distorsión de codificación

En las tablas V-3 y V-4, mostramos los resultados obtenidos para la tarea de reconocimiento de dígitos aislados y de habla continua, respectivamente, comparando los resultados obtenidos utilizando las bases de datos originales (es decir, sin codificar) con las correspondientes a la alternativa convencional (codificando y decodificando posteriormente) utilizando tanto el codificador GSM-FR como el GSM-HR. En ambos casos, el entrenamiento y el test se llevan a cabo en las mismas condiciones (*–matched conditions–*), es decir, ambos con la voz original o ambos con la voz decodificada. Además, se muestran los intervalos de confianza del 95%.

Tarea de reconocimiento de dígitos aislados				
Descripción del experimento (Entrenamiento-Test)	Codificador GSM-FR	Intervalo de confianza del 95%	Codificador GSM-HR	Intervalo de confianza del 95%
Original-Original	99,66 %	(99,53, 99,79)	99,66 %	(99,53, 99,79)
Decodificado-Decodificado	99,51 %	(99,36, 99,66)	99,42 %	(99,25, 99,59)

Tabla V-3- Tasas de reconocimiento correspondientes a la tarea de reconocimiento de dígitos aislados que muestran la influencia de la distorsión de codificación en condiciones idénticas de entrenamiento y test (*–matched conditions–*).

Podemos observar cómo la distorsión de codificación pasa prácticamente inadvertida en la tarea de reconocimiento de dígitos aislados mientras que es bastante notable en la de habla continua. En el primer caso, la tarea resulta demasiado sencilla para el reconocedor que, como vemos, alcanza prestaciones muy. Así, a pesar de que las tasas de reconocimiento disminuyen ligeramente, los intervalos de confianza se solapan completamente, indicando que las diferencias no son significativas estadísticamente.

Sin embargo, si nos fijamos en los resultados de la tarea de reconocimiento de habla continua la conclusión es diferente: la influencia de la distorsión de codificación hace que empeore considerablemente el comportamiento del reconocedor. Como es lógico, los resultados para el codificador GSM-FR son mejores que para el GSM-HR ya que, a

pesar de que se trata de un codificador más antiguo y tecnológicamente inferior, el régimen binario que utiliza es demasiado superior (aproximadamente el doble).

Tarea de reconocimiento de habla continua				
Descripción del experimento (Entrenamiento-Test)	Codificador GSM-FR	Intervalo de confianza del 95%	Codificador GSM-HR	Intervalo de confianza del 95%
Original-Original	90,83 %	(90,27, 91,39)	90,83 %	(90,27, 91,39)
Decodificado-Decodificado	88,10 %	(87,47, 88,73)	85,39 %	(84,71, 86,07)

Tabla V-4.-Tasas de reconocimiento correspondientes a la tarea de reconocimiento de habla continua que muestran la influencia de la distorsión de codificación en condiciones idénticas de entrenamiento y test (–*matched conditions*–).

V.2.1.2. Influencia de los errores de transmisión

En las Tablas V-5 y V-6 se muestra la influencia de los errores de transmisión en los reconocedores convencionales para ambas tareas de reconocimiento y los dos codificadores GSM que hemos evaluado. Para ello hemos utilizado el modelo de canal al que aludíamos en V.1.3.1 y hemos simulado tasas de error que van desde 10^{-4} hasta $5 \cdot 10^{-2}$. El entrenamiento de los modelos siempre se ha realizado sobre un canal limpio (sin errores), aunque utilizando la voz decodificada.

Tarea de reconocimiento de dígitos aislados				
Canales	Codif. GSM-FR	Intervalo de confianza del 95%	Codif. GSM-HR	Intervalo de confianza del 95%
0	99,51 %	(99,36, 99,66)	99,42 %	(99,25, 99,59)
1	99,39 %	(99,22, 99,56)	99,42 %	(99,25, 99,59)
2	94,94 %	(94,46, 95,42)	99,38 %	(99,21, 99,55)
3	88,18 %	(87,47, 88,89)	99,29 %	(99,11, 99,47)
4	84,53 %	(83,73, 85,33)	99,07 %	(98,86, 99,28)
5	38,64 %	(37,57, 39,71)	88,56 %	(87,86, 89,26)

Tabla V-5.- Tasas de reconocimiento correspondientes a la tarea de reconocimiento de dígitos aislados que muestran la influencia de los errores de transmisión sobre un sistema de reconocimiento convencional (es decir, reconocimiento a partir de voz decodificada).

Es conveniente recordar que después de la intervención del codificador de canal, el decodificador de voz se encuentra con 3 tipos de tramas: tramas completamente limpias

(sin ningún tipo de error), tramas completamente erróneas (que el decodificador de canal ha etiquetado como tales) y tramas con errores residuales (véase la Tabla V-1). Las tramas completamente erróneas se sustituyen siguiendo el procedimiento previsto en el estándar y el resto se dejan como están.

Tarea de reconocimiento de habla continua				
Canales	Codif. GSM-FR	Intervalo de confianza del 95%	Codif. GSM-HR	Intervalo de confianza del 95%
0	88,10 %	(87,47, 88,73)	85,39 %	(84,71, 86,07)
1	88,02 %	(87,39, 88,65)	85,32 %	(84,64, 86,00)
2	83,74 %	(83,03, 84,45)	85,28 %	(84,60, 85,96)
3	56,79 %	(55,83, 57,75)	84,11 %	(83,40, 84,82)
4	42,62 %	(41,67, 43,58)	83,10 %	(82,38, 83,82)
5	-	-	55,39 %	(54,43, 56,35)

Tabla V-6.- Tasas de reconocimiento correspondientes a la tarea de reconocimiento de habla continua que muestran la influencia de los errores de transmisión sobre un sistema de reconocimiento convencional (es decir, reconocimiento a partir de voz decodificada).

Como se puede observar, la influencia de los errores de transmisión sobre el reconocedor es muy importante, si bien es mucho más perjudicial en la tarea de habla continua. Además, al contrario que la distorsión de codificación, ésta afecta de forma mucho más notable al reconocedor que utiliza el codificador GSM-FR. Es decir, el codificador GSM-HR es mucho más robusto frente a errores de transmisión que el GSM-FR. Esto es fácilmente explicable si observamos con detenimiento los valores de FER y RBER correspondientes para cada canal a idénticas BER teóricas, y de las cuales puede inferirse que el codificador de canal resulta más efectivo para el canal TCH/HS. Por este motivo, no hemos considerado el canal más severo para el codificador GSM-FR, ya que las tasas de reconocimiento resultantes eran demasiado bajas.

V.2.1.3. Reconocimiento mediante transparametrización

Una vez analizadas las influencias de las dos distorsiones específicas de la transmisión sobre la red GSM, mostramos ahora los resultados obtenidos utilizando la aproximación de RMT. Las tablas V-7 y V-8 muestran para los dos codificadores que evaluamos los resultados obtenidos para la tarea IDR y CSR, respectivamente.

Tarea de reconocimiento de dígitos aislados				
Canales	Tasa de reconocimiento (%) e intervalos de confianza del 95 % (entre paréntesis)			
	GSM-FR		GSM-HR	
	RMT (Periodo de trama de 20 ms)	Decodificado (Periodo de trama de 10 ms)	RMT (Periodo de trama de 20 ms)	Decodificado (Periodo de trama de 10 ms)
0	99,26 % (99,07, 99,45)	99,51 % (99,36, 99,66)	99,39 % (99,22, 99,56)	99,42 % (99,25, 99,59)
1	99,22 % (99,03, 99,41)	99,39 % (99,22, 99,56)	99,39 % (99,22, 99,56)	99,42 % (99,25, 99,59)
2	97,83 % (97,51, 98,15)	94,94 % (94,46, 95,42)	99,38 % (99,21, 99,55)	99,38 % (99,21, 99,55)
3	93,43 % (92,88, 93,98)	88,18 % (87,47, 88,89)	99,32 % (99,14, 99,5)	99,29 % (99,11, 99,47)
4	89,52 % (88,85, 90,19)	84,53 % (83,73, 85,33)	99,04 % (98,83, 99,25)	99,07 % (98,86, 99,28)
5	44,66 % (43,57, 45,75)	38,64 % (37,57, 39,71)	88,50 % (87,80, 89,20)	88,56 % (87,86, 89,26)

Tabla V-7.- Comparación del comportamiento del sistema de RMT y el convencional (decodificado) para los distintos canales GSM y para la tarea de reconocimiento de dígitos aislados.

El entrenamiento de los modelos siempre se ha realizado sobre el canal limpio (sin errores), aunque utilizando bien la voz decodificada, en el caso de la aproximación convencional o bien el procedimiento de transparametrización. En este último caso los parámetros cepstrales se han calculado a partir de espectro LP, utilizando la estimación de la energía y con un periodo de trama de 20 ms para la tarea IDR y 10 ms para la CSR (véase la sección IV.4).

De nuevo hemos preferido no evaluar los reconocedores bajo los efectos del canal $5 \cdot 10^{-2}$ con el codificador GSM-FR, debido a que ambos reconocedores obtienen tasas de reconocimiento excesivamente bajas.

Cuando la tarea que abordamos es la de reconocimiento de dígitos aislados, el reconocedor mediante transparametrización es claramente superior para el codificador GSM-FR (para BERs por encima de 10^{-4}) que, como se recordará, era menos robusto frente a los errores de transmisión. En la V-4, se exponen los porcentajes de mejora obtenidos en esta ocasión; la ventaja aumenta conforme crece la BER. Sin embargo, no observamos diferencias significativas entre las dos alternativas que comparamos para el codificador GSM-HR. Esto es debido al buen comportamiento de este codificador frente

a los errores de transmisión y la sencillez de esta tarea, que no dejan el margen suficiente para una mejora.

Tarea de reconocimiento de habla continua				
Canales	Tasa de reconocimiento (%) e intervalos de confianza del 95 % (entre paréntesis)			
	GSM-FR		GSM-HR	
	RMT (Periodo de trama de 10 ms)	Decodificado (Periodo de trama de 10 ms)	RMT (Periodo de trama de 10 ms)	Decodificado (Periodo de trama de 10 ms)
0	87,64 % (87,00, 88,28)	88,10 % (87,47, 88,73)	88,09 % (87,46, 88,72)	85,39 % (84,71, 86,07)
1	87,60 % (86,96, 88,24)	88,02 % (87,39, 88,65)	88,09 % (87,46, 88,72)	85,32 % (84,64, 86,00)
2	83,45 % (82,73, 84,17)	83,74 % (83,03, 84,45)	88,10 % (87,47, 88,73)	85,28 % (84,60, 85,96)
3	63,12 % (62,19, 64,05)	56,79 % (55,83, 57,75)	87,86 % (87,23, 88,49)	84,11 % (83,40, 84,82)
4	50,07 % (49,10, 51,04)	42,62 % (41,67, 43,58)	87,62 % (86,98, 88,26)	83,10 % (82,38, 83,82)
5	-	-	70,12 % (69,24, 71,00)	55,39 % (54,43, 56,35)

Tabla V-8.- Comparación del comportamiento del sistema de RMT y el convencional (decodificado) para los distintos canales GSM para la tarea de reconocimiento de habla continua.

Para la tarea de habla continua y en lo que respecta a GSM-FR, podemos extraer las mismas conclusiones salvo que en este caso las mejorías del reconocedor que proponemos resultan significativas para BERs a partir de $5 \cdot 10^{-3}$. Sin embargo, para el codificador GSM-HR, las mejoras resultan ahora muy notables y de nuevo aumentando su ventaja conforme el canal empeora su comportamiento (veánse los porcentajes de mejora en la V-4).

Esto es debido a que para esta tarea los errores del canal tienen una influencia mayor sobre el reconocedor (véase la influencia de los errores de transmisión en la sección V.2.1.2) y por tanto las mejoras de una parametrización más robusta se hacen más patentes. Las figuras V-4 y V-5 muestran las tasas de reconocimiento obtenidas por ambas alternativas, para el codificador GSM-FR y GSM-HR, respectivamente.

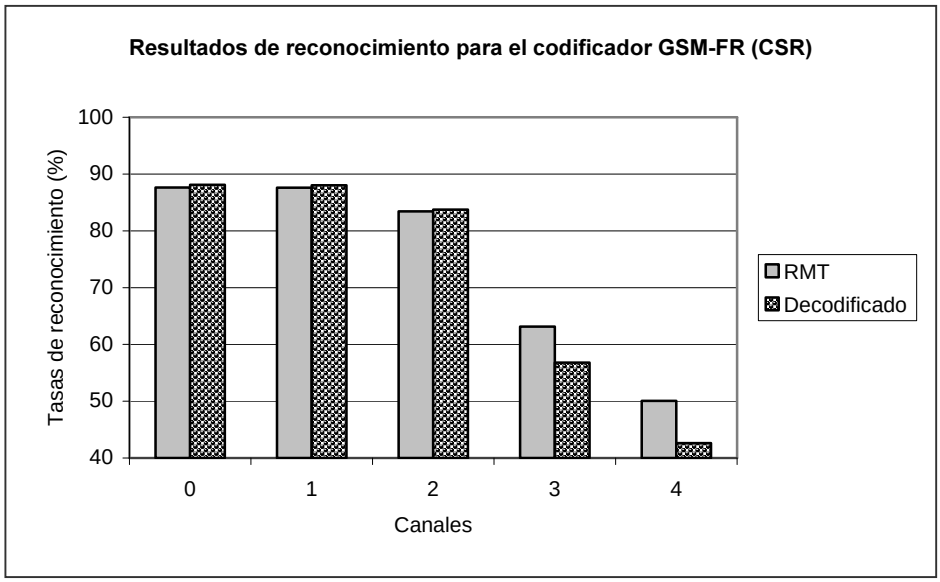


Figura V-4. -Diagrama comparativo entre el sistema RMT y el decodificado para el codificador GSM-FR.

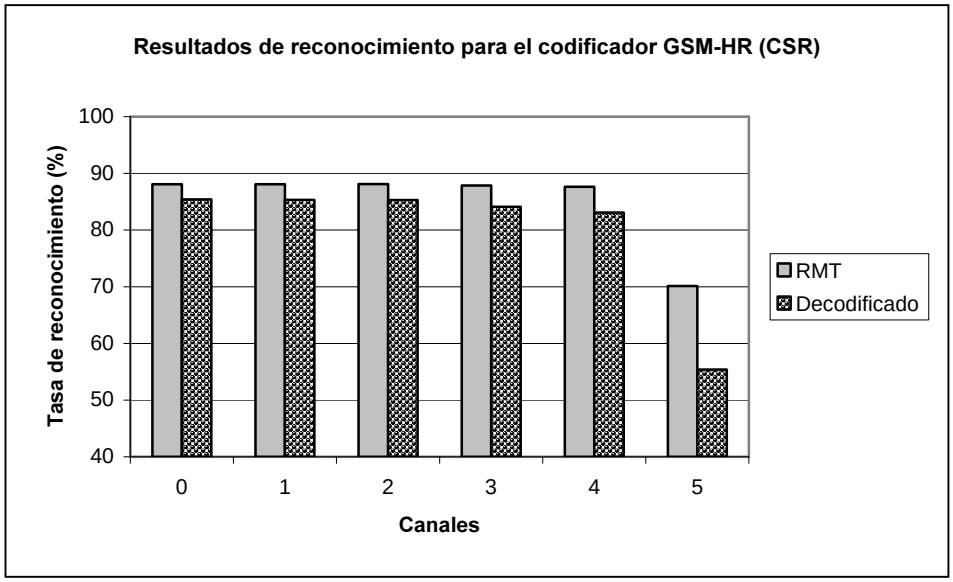


Figura V-5. -Diagrama comparativo entre el sistema RMT y el decodificado para el codificador GSM-HR.

Porcentajes de mejora de la alternativa RMT frente a la convencional				
Canales	IDR		CSR	
	GSM-FR	GSM-HR	GSM-FR	GSM-HR
0	-0,25	-0,03	-0,52	3,07
1	-0,17	-0,03	-0,48	3,14
2	2,95	0,00	-0,35	3,20
3	5,62	0,03	10,03	4,27
4	5,57	-0,03	14,88	5,16
5	13,48	-0,07	-	21,01

Tabla V-9. Porcentajes de mejora de la aproximación RMT sobre la convencional.

V.2.2. IP

En esta sección nos vamos a ocupar de la descripción de los experimentos que hemos llevado a cabo sobre el entorno IP para evaluar las prestaciones del reconocedor que proponemos. En la Figura V-6 presentamos un diagrama de bloques que muestra la secuencia de procesos necesaria para obtener los vectores de parámetros tanto de nuestra propuesta de reconocimiento mediante transparametrización (parte superior), como en el caso convencional (parte inferior). Como vemos, a diferencia del entorno GSM, aquí no existe un codificador de canal asociado a cada codificador de voz. Sin embargo, el canal IP suele ser bastante más fiable (normalmente está compuesto por enlaces fijos), en el sentido de que cuando un paquete llega a su destino la tasa de error de bit es muy baja; como contrapartida la cantidad de paquetes perdidos suele ser elevada.

También a diferencia de los codificadores que consideramos en el entorno GSM, el codificador G.723.1 emplea los parámetros LSP. Esto quiere decir que es posible evitar la transformación a parámetros LP, calculando directamente el espectro LSP; asimismo, es posible calcular el pseudocepstrum en lugar de los MFCC (véase la sección IV.3.1).

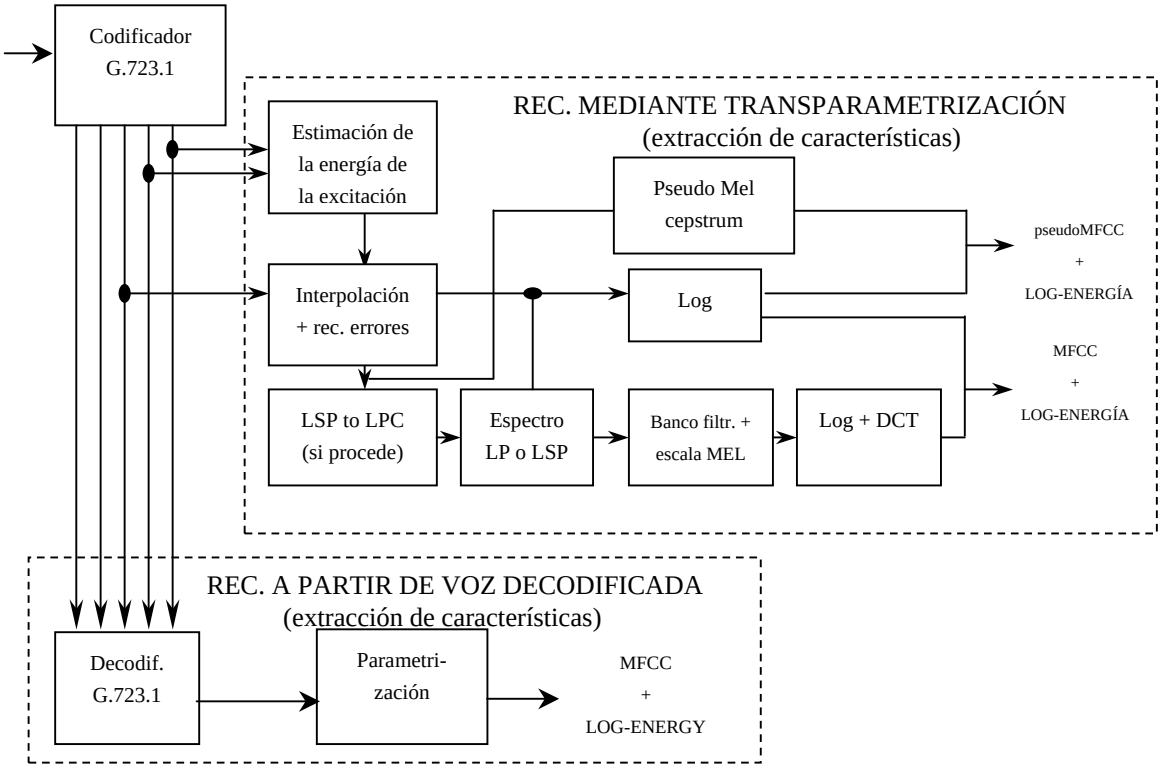


Figura V-6.-Diagrama de bloques que muestra, en la parte superior, las transformaciones necesarias para obtener los parámetros de reconocimiento según el procedimiento que proponemos, y en la parte inferior, el procedimiento convencional, en el caso de considerar la red IP.

Como ya hicimos en el entorno GSM, en primer lugar analizaremos los efectos de la distorsión de codificación y de los errores de transmisión sobre el reconocimiento y posteriormente pasaremos a comparar las prestaciones del reconocedor que proponemos y el convencional. Además, detallaremos los experimentos que hemos llevado a cabo para evaluar las distintas opciones a la hora de realizar la transparametrización que describimos en el Capítulo IV, así como los beneficios derivados de interpolar las tramas para obtener una tasa de trama adecuada para el reconocimiento y de estimar la energía.

V.2.2.1. Influencia de la distorsión de codificación

En la Tabla V-10 mostramos los resultados obtenidos para el codificador G.723.1 para ambas tareas comparando las tasas de reconocimiento obtenidas a partir de la voz original (sin codificación) con las logradas después de una etapa de codificación, siempre realizando el entrenamiento y el test en las mismas condiciones (ambos con la voz original o ambos con la voz decodificada).

Descripción del Experimento (Entrenamiento-Test)	Tarea IDR	Intervalo de confianza del 95%	Tarea CSR	Intervalo de confianza del 95%
Original-Original	99,66 %	(99,53, 99,79)	90,83 %	(90,27, 91,39)
Decodificado-Decodificado	99,33 %	(99,15, 99,51)	87,01 %	(86,36, 87,66)

Tabla V-10.- Tasas de reconocimiento correspondientes a las tareas de reconocimiento de dígitos aislados (IDR) y de habla continua (CSR) que muestran la influencia de la distorsión de codificación en condiciones idénticas de entrenamiento y test (*–matched conditions–*).

Como puede observarse, la pérdida de prestaciones de los reconocedores es estadísticamente significativa para ambas tareas aunque, desde luego, es mucho más notable en la de habla continua. Esto es debido a que la tarea de dígitos aislados es muy sencilla y por lo tanto no se ve muy afectada por este tipo de distorsión.

V.2.2.2. Influencia de la pérdida de paquetes

La Tabla V-11 muestra la influencia de la pérdida de paquetes cuando utilizamos la aproximación de reconocimiento convencional (es decir, reconocemos a partir de la voz decodificada). Como es lógico, el entrenamiento de los modelos siempre se realiza sobre el canal limpio (sin errores) aunque, eso sí, utilizando la voz decodificada. Además, en todos los casos, los mecanismos de recuperación frente a errores previstos para este codificador están activos.

Canales	IDR	Intervalo de confianza del 95%	DSR	Intervalo de confianza del 95%
-	99,33 %	(99,15, 99,51)	87,01 %	(86,36, 87,66)
A	99,20 %	(99,00, 99,40)	86,64 %	(85,98, 87,30)
B	98,91 %	(98,68, 99,14)	85,75 %	(85,07, 86,43)
C	98,70 %	(98,45, 99,95)	84,47 %	(84,18, 85,56)
D	98,13 %	(97,83, 98,43)	83,66 %	(82,95, 84,37)
E	97,53 %	(97,19, 97,87)	81,01 %	(80,25, 81,77)
F	97,12 %	(96,75, 97,49)	81,68 %	(80,93, 82,43)

Tabla V-11.- Tasas de reconocimiento correspondientes a la tarea de reconocimiento de dígitos aislados (IDR) y a la de habla continua (CSR) que muestran la influencia de los errores de transmisión sobre un sistema de reconocimiento convencional (reconocimiento a partir de voz decodificada).

Como expusimos en la sección V.1.3.2.2, la pérdida de paquetes en el entorno IP tiene lugar "a ráfagas", es decir, lo normal es que se pierdan varios paquetes consecutivos. Además, la cantidad de tramas del codificador de voz que se incluyen dentro de un paquete IP es una decisión de diseño de la aplicación de VoIP se trata de un compromiso entre el retardo que se está dispuesto a admitir y la sobrecarga de cabeceras (*overhead*) que supone enviar una sola trama en cada paquete. En general, el problema de la sobrecarga de cabeceras no es tan importante en las redes fijas como en las móviles (véase la sección V.2.3.2) y puesto que el retardo algorítmico de G.723.1 es bastante elevado (30 ms), hemos decidido realizar nuestros experimentos considerando una sola trama por paquete.

Estos resultados ilustran cómo según aumenta la tasa de paquetes perdidos el comportamiento del reconocedor va empeorando para todos los casos en los que la longitud media de las ráfagas se mantiene más o menos constante (canales A a E). Sin embargo, para el canal F, en el que la longitud de ráfagas es considerablemente superior (aunque con una tasa de pérdidas inferior a la del canal E), podemos observar como la tarea IDR se ve afectada de forma más severa que la CSR. Esto es debido a que, en el primer caso, y con las longitudes de ráfagas que consideradas en el canal F, en muchas ocasiones perdemos la totalidad del dígito que queremos reconocer, imposibilitando totalmente su reconocimiento. En el caso de CSR, sin embargo, el modelo de lenguaje puede ayudar en estas ocasiones.

V.2.2.3. Reconocimiento mediante transparametrización

En las tablas V-12 y V-13 ilustramos el comportamiento del reconocedor mediante transparametrización que proponemos en comparación con el del reconocedor convencional, para las tareas IDR y CSR, respectivamente. El entrenamiento de los modelos siempre se ha realizado sobre el canal limpio (sin errores), aunque utilizando el procedimiento de transparametrización o decodificación, según proceda. Los parámetros cepstrales, en el esquema propuesto, se han calculado a partir de espectro LP, utilizando la estimación de la energía y con un período de trama de 15 ms para la tarea IDR y 10 ms para la CSR (véase la sección IV.4).

Para la tarea IDR, hemos considerado un periodo de trama de 15 ms, habiendo interpolado un nuevo vector de parámetros entre cada dos. En el caso de CSR, hemos encontrado más adecuado realizar esta interpolación para 2 de cada 3 vectores, obteniendo así un periodo de trama de 10 ms (véase la sección V.2.2.3.2).

Como vemos, en el caso IDR, aunque la degradación de las prestaciones del reconocedor con la tasa de pérdida de paquetes es menor en el esquema propuesto, la ventaja no llega a ser estadísticamente significativa, debido, como hemos advertido en otras ocasiones, al reducido tamaño de la base de datos que utilizamos y a que la sencillez de la tarea no deja un margen suficiente para la mejora. En la Tabla V-14 se pueden observar los porcentajes de mejora de la aproximación propuesta frente a la convencional.

Tarea de reconocimiento de dígitos aislados.				
Canales	Tasa de reconocimiento (%)			
	RMT (Periodo de trama de 15 ms)	Intervalo de confianza del 95%	Decodificado (Periodo de trama de 10 ms)	Intervalo de confianza del 95%
-	99,29 %	(99,04, 99,42)	99,33 %	(99,15, 99,51)
A	99,17 %	(98,97, 99,37)	99,20 %	(99,00, 99,40)
B	99,03 %	(98,81, 99,25)	98,91 %	(98,68, 99,14)
C	98,84 %	(98,60, 99,08)	98,70 %	(98,45, 98,95)
D	98,59 %	(98,33, 98,85)	98,13 %	(97,83, 98,43)
E	98,09 %	(97,79, 98,39)	97,53 %	(97,19, 97,87)
F	97,54 %	(97,20, 97,88)	97,12 %	(96,75, 97,49)

Tabla V-12.- Comparación del comportamiento del sistema de RMT y el convencional (decodificado) para los distintos canales IP y para la tarea de reconocimiento de dígitos aislados (IDR).

Tarea de reconocimiento de habla continua				
Canales	Tasa de reconocimiento (%)			
	RMT	Intervalo de confianza del 95%	Decodificado	Intervalo de confianza del 95%
-	88,33 %	(87,71, 88,95)	87,01 %	(86,36, 87,66)
A	88,25 %	(87,63, 88,87)	86,64 %	(85,98, 87,30)
B	87,46 %	(86,82, 88,10)	85,75 %	(85,07, 86,43)
C	86,81 %	(86,17, 87,46)	84,47 %	(84,18, 85,56)
D	86,09 %	(85,42, 86,76)	83,66 %	(82,95, 84,37)
E	83,98 %	(83,27, 84,69)	81,01 %	(80,25, 81,77)
F	83,96 %	(83,25, 84,67)	81,68 %	(80,93, 82,43)

Tabla V-13.- Comparación del comportamiento del sistema de RMT y el convencional (decodificado) para los distintos canales IP y para la tarea de reconocimiento de habla continua (CSR).

Para la tarea CSR, sin embargo, la mejora obtenida vía RMT es clara y significativa para todos los canales (incluso en el que no considera pérdidas de paquetes). De nuevo observamos, en el canal F, cómo el aumento de la longitud de las ráfagas de paquetes perdidos tiene una influencia nefasta sobre las tasas de reconocimiento, de forma que éstas resultan similares a las del canal E cuya PLR es superior (5,83% frente a 4,11%).

En todo caso, como se desprende la de Tabla V-14, el porcentaje de mejora aumenta a medida que las condiciones del canal empeoran. Este efecto se puede observar también, en la Figura V-7, para las dos tareas de reconocimiento.

Porcentajes de mejora de la alternativa RMT frente a la convencional		
Canales	IDR	CSR
0	-0,04	1,49
1	-0,03	1,82
2	0,12	1,96
3	0,14	2,7
4	0,47	2,82
5	0,57	3,54

Tabla V-14. Porcentajes de mejora de la aproximación RMT sobre la convencional.

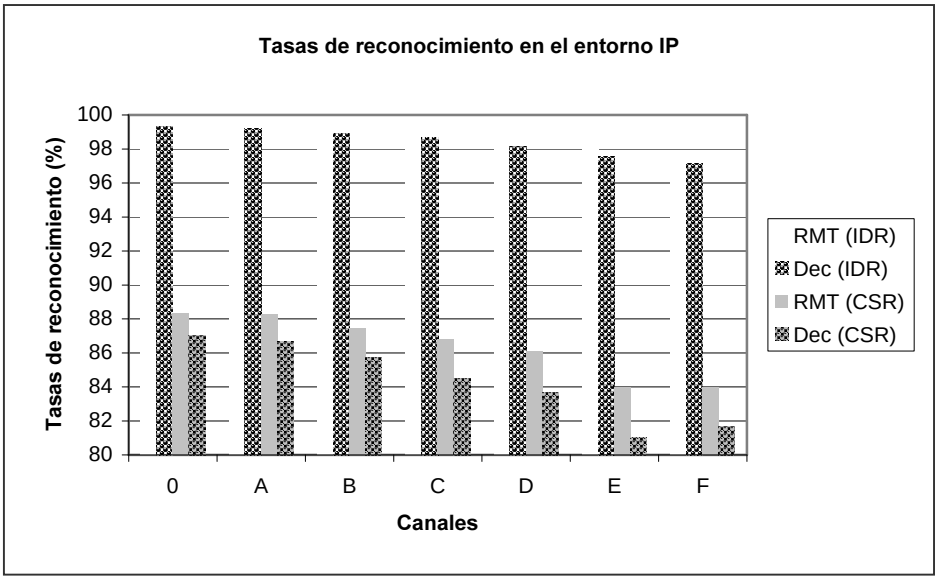


Figura V-7. -Diagrama comparativo entre el sistema RMT y el decodificado para el codificador GSM-FR.

V.2.2.3.1. Cálculo de los parámetros cepstrales

En la sección IV.5.1 presentamos una serie de métodos para estimar la envolvente espectral; a continuación evaluaremos los resultados de reconocimiento que se obtienen con las distintas opciones. En concreto, compararemos el comportamiento de los reconocedores basados en transparametrización que calculan el espectro, $\hat{H}(\Omega_{mel})$, utilizando bien, los parámetros LP, $\hat{\mathbf{a}}^{(p)}$, o bien, los LSP, $\boldsymbol{\omega}^{(p)}$ (ecuación (IV-18)). Además examinaremos la aproximación del pseudocepstrum que nos permite calcular directamente los coeficientes cepstrales, $\hat{\mathbf{c}}_{\hat{h}}^{(L)}$, a partir de los LSP, $\boldsymbol{\omega}^{(p)}$, y también aplicar la escala Mel y obtener $\mathbf{mf\hat{c}}_{\hat{h}}^{(L)}$.

Tarea de reconocimiento de dígitos aislados Tasas de reconocimiento (%) e intervalos de confianza del 95 %					
Canales	Espectro LP	Espectro LSP	Pseudo Mel cepstrum (Ecuación (IV-26))	Pseudo Mel cepstrum (Ecuación (IV-27))	Pseudo cepstrum
–	99,27 (99,08, 99,46)	99,27 (99,08, 99,46)	99,18 (98,98, 99,38)	99,28 (99,09, 99,47)	99,19 (98,99, 99,39)
A	99,17 (98,97, 99,37)	99,18 (98,98, 99,38)	99,09 (98,88, 99,30)	99,17 (98,97, 99,37)	99,12 (98,91, 99,33)
B	99,03 (98,81, 99,25)	98,99 (98,77, 99,21)	98,89 (98,66, 99,12)	98,96 (98,74, 99,18)	98,93 (98,70, 99,16)
C	98,84 (98,60, 99,08)	98,83 (98,59, 99,07)	98,69 (98,44, 98,94)	98,71 (98,46, 98,96)	98,8 (98,56, 99,04)
D	98,59 (98,33, 98,85)	98,54 (98,28, 98,80)	98,23 (97,94, 98,52)	98,38 (98,10, 98,66)	98,38 (98,10, 98,66)
E	98,09 (97,79, 98,39)	98,13 (97,83, 98,43)	97,89 (97,57, 98,21)	97,88 (97,56, 98,20)	97,93 (97,62, 98,24)
F	97,54 (97,20, 97,88)	97,58 (97,24, 97,92)	97,18 (96,82, 97,54)	97,35 (97,00, 97,70)	97,47 (97,12, 97,82)

Tabla V-15.- Comparación entre los distintos métodos para calcular los parámetros cesptrales para la tarea de reconocimiento de dígitos aislados.

Tarea de reconocimiento de habla continua Tasas de reconocimiento (%) e intervalos de confianza del 95 %					
Canales	Espectro LP	Espectro LSP	Pseudo Mel cepstrum (Ecuación (IV-26))	Pseudo Mel cepstrum (Ecuación (IV-27))	Pseudo cepstrum
-	88,1 (87,47, 88,73)	87,96 (87,33, 88,59)	87,29 (86,65, 87,93)	87,51 (86,87, 88,15)	85,71 (85,03, 86,39)
A	87,96 (87,33, 88,59)	87,78 (87,15, 88,41)	87,22 (86,57, 87,87)	87,26 (86,62, 87,9)	85,56 (84,88, 86,24)
B	87,43 (86,79, 88,07)	87,28 (86,64, 87,92)	86,64 (85,98, 87,30)	86,83 (86,18, 87,48)	84,93 (84,24, 85,62)
C	86,51 (85,85, 87,17)	86,35 (85,69, 87,01)	85,28 (84,60, 85,96)	85,46 (84,78, 86,14)	83,85 (83,14, 84,56)
D	85,6 (84,92, 86,28)	85,44 (84,76, 86,12)	85,05 (84,36, 85,74)	84,6 (83,90, 85,3)	82,98 (82,25, 83,71)
E	84,08 (83,37, 84,79)	84,1 (83,39, 84,81)	82,71 (81,98, 83,44)	83,16 (82,44, 83,88)	81,24 (80,49, 81,99)
F	83,05 (82,32, 83,78)	82,94 (82,21, 83,67)	82,87 (82,14, 83,60)	82,7 (81,97, 83,43)	81,06 (80,3 ,81,82)

Tabla V-16.- Comparación entre los distintos métodos para calcular los parámetros cepstrales para la tarea de reconocimiento de habla continua.

Las Tablas V-15 y la V-16 muestran los resultados obtenidos para cada una de estas opciones para la tarea IDR y CSR, respectivamente. Las ligeras diferencias entre esta tabla y los valores de la misma descripción en las tablas de las secciones V.2.2.1, V.2.2.2 y V.2.2.3 se deben al cambio de versión del paquete de software HTK que volvimos a recalcular, en aquellas secciones, para su correcta comparación.

De estos experimentos podemos concluir que no hay diferencias apreciables entre calcular el espectro LP y el LSP, si bien parece que para la tarea CSR hay una muy ligera mejoría en el caso del espectro LP. Esto puede deberse a una mayor sensibilidad del espectro LSP a la precisión aritmética de los cálculos, no obstante, no nos atrevemos a extraer ninguna conclusión.

En cuanto a la utilización del Pseudo-Mel-Cepstrum, observamos un descenso general de las tasas de reconocimiento para todos los canales si las comparamos con las obtenidas con cualquiera de los otros dos espectros; si bien las diferencias no son estadísticamente significativas con un nivel de confianza del 95%. En lo que se refiere a

la utilización de las dos escalas Mel que evaluamos, las diferencias son muy reducidas y desde luego no son significativas. Sin embargo, sí que se observan claramente los beneficios de aplicar esta escala en la tarea CSR, no siendo tan claros en la de IDR [145].

V.2.2.3.2. Beneficios de la interpolación de tramas y la estimación de la energía

En el Capítulo IV identificamos dos elementos relevantes a la hora de realizar el reconocimiento mediante nuestra propuesta de parametrización: la tasa de trama y la energía. Por este motivo realizamos experimentos preliminares utilizando canales distintos a los que posteriormente elegimos como más representativos del comportamiento de la red IP y que hemos venido utilizando en el resto de los experimentos de esta tesis. En concreto, los canales que utilizamos en esta sección, excepcionalmente, son los correspondientes a las PLRs 5%, 10% y 20% empleadas en la recomendación TR 101 329-6 de ETSI ([58]). Dada la importancia de estos dos elementos (tasa de trama y energía) hemos considerado que resultaba ilustrativa la inclusión de estos experimentos que muestran, de forma palpable, ambos efectos a pesar de que no son comparables con los anteriores.

Así, comenzamos analizando, los beneficios que se obtienen al realizar una interpolación para adecuar la tasa de trama del codificador a la necesaria para reconocimiento, sin todavía incorporar la estimación de la energía (utilizamos la proveniente de la señal de voz decodificada).

Tarea de reconocimiento de dígitos aislados					
	Periodo de trama	PLR			
		0 %	5 %	10 %	20 %
RMT: sistema inicial (interpolación lineal)	15 ms	97,84 (97,52, 98,16)	96,97 (96,59, 97,35)	96,41 (96,00, 96,82)	94.12 (93,60, 94.64)
RMT: Interpolador de banda limitada	15 ms	99,07 (98,86, 99,28)	98,78 (98,54, 99,02)	97,75 (97,42, 98,08)	95.69 (95.24 ,96.14)

Tabla V-17.- Comparación entre el interpolador propuesto en el estándar G.723.1 y el interpolador de banda limitada para varias tasas de pérdida de paquetes y la tarea de dígitos aislados (IDR). Entre paréntesis se muestran los intervalos de confianza del 95%.

Para ello, comparamos la interpolación lineal que se propone en el estándar G.723.1 (ecuación (IV-59)) donde los vectores de parámetros LSP para cada subtrama i ($0 \le i < 3$) de la trama k, $\hat{\boldsymbol{\omega}}^{(p)}[k,i]$, se calculan a partir de los de las tramas k y $k-1$,

con la realizada con un filtro interpolador de banda limitada que utiliza para cada vector interpolado los vectores correspondientes a dos tramas anteriores y dos posteriores para la tarea IDR. Los resultados se pueden observar en la Tabla V-17.

Como vemos, el reconocedor que utiliza la interpolación de banda limitada y de mayor longitud que proponemos, genera resultados muy superiores y estadísticamente significativos para todas las tasas de pérdida de paquetes consideradas a las obtenidas por el que utiliza el interpolador implementado por el estándar para codificación, poniendo de manifiesto las diferencias entre los dos procesos (reconocimiento y codificación –véase la sección IV.1–).

Tarea de reconocimiento de habla continua					
	Periodo de trama	PLR			
		0 %	5 %	10 %	20 %
RMT: Interpolador de banda limitada	15 ms	83,81 (83,10, 84,52)	81,51 (80,76, 82,26)	78,47 (77,68, 79,26)	69,48 (68,59, 70,37)
RMT: Interpolador de banda limitada	10 ms	86,08 (85,41, 86,75)	84,07 (83,36, 84,78)	82,08 (81,34, 82,82)	74,49 (73,65, 75,33)
Decodificado	15ms	81,54 (80,79, 82,29)	78,65 (77,86, 79,44)	74,78 (73,94, 75,62)	63,94 (63,01, 64,87)
Decodificado	10ms	87,01 (86,36, 87,66)	83,55 (82,83, 84,27)	79,78 (79,00, 80,56)	70,16 (69,28, 71,04)

Tabla V-18.- Comparación entre los periodos de trama de 10 y 15 ms para varias tasas de pérdida de paquetes y la tarea de reconocimiento de habla continua (CSR). Entre paréntesis se muestran los intervalos de confianza del 95%.

Una vez establecidos los beneficios de esta interpolación comparamos las tasas de trama de 15 ms y 10 ms, tanto para la aproximación RMT que proponemos, como la convencional. Para la tarea IDR, los resultados para ambas son muy parecidos y no encontramos ventajoso utilizar la de 10 ms. Sin embargo, en el caso de la tarea de reconocimiento de habla continua (Tabla V-18) las diferencias son muy grandes, tanto para el reconocedor que proponemos como para el convencional, indicando la gran relevancia de este parámetro para el problema de reconocimiento.

El otro elemento que se ha revelado significativo a la hora de realizar este reconocimiento mediante transparamentrización es la estimación del parámetro de energía que forma parte del vector acústico que utilizamos. En las Tablas V-19 y V-20 mostramos los beneficios que se obtienen de la estimación de la energía por el procedimiento que se detalla en la sección IV.5.1, comparando la opción RMT que utiliza como parámetro de energía el obtenido por el decodificador con el que incorpora

la estimación propuesta. En estos resultados ya hemos incorporado la interpolación más conveniente según hemos analizado anteriormente

Tarea de reconocimiento de dígitos aislados					
	Periodo de Trama	PLR			
		0 %	5 %	10 %	20 %
RMT: Interpolador FIR Sin estimación de energía	15 ms	99,07 (98,86, 99,28)	98,78 (98,54, 99,02)	97,75 (97,42, 98,08)	95.69 (95.24 ,96.14)
RMT: Interpolador FIR Con estimación de energía	15 ms	99,29 (99,04, 99,42)	98,96 (98,74, 99,18)	98,62 (98,36, 98,88)	97,35 (97,00, 97,70)

Tabla V-19.- Beneficios de la estimación de la energía en la tarea de reconocimiento de dígitos aislados para distintas tasas de pérdida de paquetes.

Tarea de reconocimiento de habla continua					
	Periodo de Trama	PLR			
		0 %	5 %	10 %	20 %
RMT: Interpolador de banda limitada Sin estimación de energía	10 ms	86,08 (85,41, 86,75)	84,07 (83,36, 84,78)	82,08 (81,34, 82,82)	74,49 (73,65, 75,33)
RMT: Interpolador de banda limitada Con estimación de energía	10 ms	88,33 (87,71, 88,95)	86,44 (85,78, 87,10)	84,04 (83,33, 84,75)	77,25 (76,44, 78,06)

Tabla V-20.-Beneficios de la estimación de la energía en la tarea de reconocimiento de habla continua para distintas tasas de pérdida de paquetes.

Como podemos observar se obtienen beneficios importantes de la utilización de la estimación de la energía, que aunque en el caso de IDR la base de datos no resulte lo suficientemente grande como para arrojar datos estadísticamente significativos la tarea CSR no deja ninguna duda al respecto. Las razones para esta mejora se expusieron en IV.5.1.

V.2.3. Transcodificación

Desde que la señal de voz parte del terminal emisor hasta que alcanza su destino atraviesa, a menudo, varios tipos de redes; es habitual en estos casos que la señal de voz codificada sufra varias codificaciones sucesivas (transcodificación) para adaptarse a la red que atraviese en esos momentos. Aunque las redes GSM prevén un mecanismo denominado "Operación Libre de Tándem" (TFO –*Tandem Free Operation*–) o también *"Vocoder Bypass"* (VCB) [56] que describe la forma en la que debe realizarse la señalización para evitar que la voz codificada tenga que sufrir codificaciones y decodificaciones sucesivas, lo cierto es que muchas veces sigue llevándose a cabo. Es necesario señalar, sin embargo, que estos mecanismos están recibiendo recientemente mucha atención dentro de los proyectos de investigación relacionados con la nueva generación de móviles ([1], [171]).

El hecho de realizar estas transcodificaciones de la señal de voz tiene consecuencias muy serias sobre la calidad de la señal, a pesar de que, en general, todos los codificadores contemplan este hecho en las etapas de diseño. Es evidente, por tanto, que este deterioro producirá reducciones de las tasas de reconocimiento (véase la sección III.2 y [162]). En esta sección evaluamos la influencia de esta etapas de codificación sobre el reconocedor mediante transparametrización que proponemos y el convencional; en primer lugar, dentro de la red GSM, y más tarde en un escenario en el que se combinan dicha red y la red IP.

V.2.3.1. Dentro de la red GSM

Dentro de las posibles combinaciones de codificaciones en tándem que pueden llevarse a cabo hemos seleccionado varias que implican a los codificadores GSM-FR, GSM-HR y G.726, inspiradas en varios de los escenarios propuestos en [162] y que se muestran, en la V-21. Estos escenarios responden a las combinaciones de llamadas originadas en la red telefónica básica o en la red de móviles y que tienen como destino, de nuevo, cualquiera de estas dos redes.

Al igual que en la descripción de los experimentos en los entornos GSM e IP, comenzaremos con la evaluación de la influencia de la distorsión de codificación, continuaremos evaluando la correspondiente a los errores de transmisión y terminaremos comparando los efectos de ambas sobre nuestra propuesta y la convencional.

V.2.3.1.1. Influencia de la distorsión de codificación

En las Tablas V-21 y V-22 mostramos los resultados de reconocimiento obtenidos utilizando la aproximación convencional para las tareas de reconocimiento de dígitos aislados y de habla continua en varios escenarios con transcodificaciones.

Tarea de reconocimiento de dígitos aislados		
Configuraciones Tándem	Tasa de reconocimiento (%)	Intervalo de confianza del 95%
–	99,66 %	(99,53, 99,79)
FR	99,51 %	(99,36, 99,66)
FR–FR	99,44 %	(99,28, 99,60)
HR–FR	98,31 %	(98,03, 98,59)
G.726 – FR	99,44 %	(99,28, 99,60)
FR– G.726 – FR	99,26 %	(99,07, 99,45)
HR– G.726 – FR	98,23 %	(97,94, 98,52)
HR	99,42 %	(99,25, 99,59)
HR–HR	98,75 %	(98,51, 98,99)
FR–HR	99,27 %	(99,08, 99,46)
G.726 – HR	99,36 %	(99,18, 99,54)
HR– G.726 – HR	98,59 %	(98,33, 98,85)
FR– G.726 – HR	99,17 %	(98,97, 99,37)

Tabla V-21.- Tasas de reconocimiento que ilustran la influencia de varias etapas de codificación para la tarea de reconocimiento de dígitos aislados (IDR).

Podemos observar como la distorsión de codificación que era insignificante para la tarea IDR cuando sólo existía una etapa de codificación, se hace mucho más patente en estos casos y especialmente cuando alguna de esta etapas incluye al codificador GSM-HR que, como ya habíamos notado, es el que mayor distorsión de codificación produce. También es interesante resaltar cómo la influencia del codificador G.726 es prácticamente despreciable, como era de esperar (véase la sección III.2).

Por el contrario, para la tarea CSR, la influencia de la distorsión de codificación es muy importante para cualquier número de etapas que consideremos (desde un 4,6% para G.726-FR hasta un 11,1% para HR-G.726-HR). En este caso, incluso el codificador G.726 produce disminuciones en la tasa de reconocimiento, significativas estadísticamente en algunos casos (G.726-FR, G.726-HR y FR-G.726-HR). Además, de nuevo, las prestaciones son peores cuando el codificador GSM-HR está presente en alguna de las etapas.

Tarea de reconocimiento de habla continua		
Configuraciones Tándem	Tasa de Reconocimiento (%)	Intervalo de confianza del 95%
–	90,83 %	(90,27, 91,39)
FR	88,10 %	(87,47, 88,73)
FR–FR	86,63 %	(85,97, 87,29)
HR–FR	83,14 %	(82,42, 83,86)
G.726 – FR	86,66 %	(86,00, 87,32)
FR– G.726 – FR	85,94 %	(85,27, 86,61)
HR– G.726 – FR	81,82 %	(81,07, 82,57)
HR	85,39 %	(84.71, 86.07)
HR–HR	81.53 %	(80.78, 82.28)
FR–HR	82.98 %	(82.25, 83.70)
G.726 – HR	82.26 %	(81.52, 82.00)
HR– G.726 – HR	80.75 %	(79.99, 81.51)
FR– G.726 – HR	81.21 %	(80.46, 81.96)

Tabla V-22.- Tasas de reconocimiento que ilustran la influencia de varias etapas de codificación para la tarea de reconocimiento de habla continua (CSR).

V.2.3.1.2. Influencia de los errores de transmisión

A continuación evaluamos la influencia de los errores de transmisión cuando nos encontramos con varias etapas de codificación: en las Tablas V-23 y V-24 mostramos los resultados obtenidos para las tareas IDR y CSR, respectivamente. En estos experimentos no hemos considerado etapas de codificación G.726, de menor influencia, con el objetivo de reducir el número de pruebas. Para cada etapa de codificación hemos introducido errores utilizando los canales que describimos en V.1.3.1 generandod nuevos patrones de error para aplicarlos a la segunda etapa del tándem, ya que de lo contrario, los errores aparecerían exactamente en los mismos lugares en los que aparecieron en la primera etapa, falseando de algún modo los resultados.

Tarea de reconocimiento de dígitos aislados						
Configuraciones Tándem	Tasa de reconocimiento (%) e intervalo de confianza del 95%					
	Canales					
	0	1	2	3	4	5
–	99,66 % (99,53, 99,79)	–	–	–	–	–
FR	99,51 % (99,36, 99,66)	99,39 % (99,22, 99,56)	94,94 % (94,46, 95,42)	88,18 % (87,47, 88,89)	84,53 % (83,73, 85,33)	38,64 % (37,57, 39,71)
FR–FR	99,44 % (99,28, 99,60)	99,22 % (99,03, 99,41)	95,62 % (95,17, 96,07)	90,58 % (89,94, 91,22)	83,74 % (82,93, 84,55)	33,52 % (32,48, 34,56)
HR–FR	98,31 % (98,03, 98,59)	98,09 % (97,79, 98,39)	92,66 % (92,09, 93,23)	90,39 % (89,74, 91,04)	74,67 % (73,71, 75,63)	27,37 % (26,39, 28,35)
HR	99,42 % (99,25, 99,59)	99,42 % (99,25, 99,59)	99,38 % (99,21, 99,55)	99,29 % (99,11, 99,47)	99,07 % (98,86, 99,28)	88,56 % (87,86, 89,26)
HR–HR	98,74 % (98,49, 98,99)	98,74 % (98,49, 98,99)	98,78 % (98,54, 99,02)	98,51 % (98,24, 98,78)	98,04 % (97,73, 98,35)	76,28 % (75,34, 77,22)
FR–HR	99,27 % (99,08, 99,46)	99,22 % (99,03, 99,41)	95,69 % (95,24, 96,14)	89,90 % (89,24, 90,56)	85,21 % (84,43, 85,99)	41,74 % (40,65, 42,83)

Tabla V-23.- Influencia de los errores de transmisión en escenarios con varias etapas de codificación para la tarea de reconocimiento de dígitos aislados (IDR).

Como era de esperar, los resultados de reconocimiento disminuyen, dramáticamente en algunos casos, cuando son sometidos a dos etapas de codificación con errores de transmisión. Al igual que observábamos en la sección V.2.1.2 para una sola etapa, cuando está presente el codificador GSM-FR, los resultados de reconocimiento se empobrecen de forma mucho más notable que para el codificador GSM-HR, imposibilitando en muchos casos el proceso de reconocimiento (véase, por ejemplo, el caso FR-FR para los canales 3 ó 4 y la tarea CSR).

Tarea de reconocimiento de habla continua				
Configuraciones Tándem	Tasas de reconocimiento (%) e intervalos de confianza del 95%			
	Canales			
	0	2	3	4
–	90,83 % (90,27, 91,39)	–	–	–
FR	88,10 % (87,47, 88,73)	83,74 % (83,03, 84,45)	56,79 % (55,83, 57,75)	42,62 % (41,67, 43,58)
FR–FR	86,63 % (85,97, 87,29)	73,51 % (72,66, 74,36)	15,61 % (14,91, 16,31)	14,14 % (13,61, 15,21
HR–FR	83,14 % (82,42, 83,86)	75,62 % (74,79, 76,45)	40,01 % (39,06, 40,96)	24,96 % (24,12, 25,80)
HR	85,39 % (84,71, 86,07)	85,28 % (84,60, 85,96)	84,11 % (83,40, 84,82)	83,10 % (82,38, 83,82)
HR–HR	81,53 % (80,78, 82,28)	81,12 % (80,36, 81,88)	81,54 % (80,79, 82,29)	78,27 % (77,47, 79,07)
FR–HR	82,98 % (82,25, 83,70)	77,83 % (77,03, 78,63)	53,54 % (52,58, 54,50)	38,87 % (37,93, 39,81)

Tabla V-24.- Influencia de los errores de transmisión en escenarios con varias etapas de codificación para la tarea de reconocimiento de habla continua (CSR).

V.2.3.1.3. Reconocimiento mediante transparametrización

Las Tablas V-25 y la V-26 muestran los resultados obtenidos para nuestra propuesta de reconocimiento mediante transparametrización en las mismas condiciones que en la sección anterior. En estas situaciones, lógicamente, la alternativa que proponemos sólo puede actuar sobre la distorsión y errores generados en la última etapa, puesto que la parametrización de codificación de la primera etapa no está accesible en el receptor.

Tarea de reconocimiento de dígitos aislados Reconocimiento mediante transparametrización						
Configu-raciones Tándem.	Tasas de reconocimiento (%) Intervalo de confianza del 95 % Canales					
	0	1	2	3	4	5
–	99,66 % (99,53, 99,79)	–	–	–	–	–
FR	99,26 % (99,07, 99,45)	99,22 % (99,03, 99,41)	97,83 % (97,51, 98,15)	93,43 % (92,88, 93,98)	89,52 % (88,85, 90,19)	44,66 % (43,57, 45,75)
FR–FR	99,03 % (98,81, 99,25)	99,00 % (98,78, 99,22)	97,06 % (96,69, 97,43)	92,97 % (92,41, 93,53)	88,95 % (88,26, 89,64)	37,54 % (36,47, 38,61)
HR–FR	98,80 % (98,56, 99,04)	98,72 % (98,47, 98,97)	96,82 % (96,43, 97,21)	96,82 % (96,43, 97,21)	82,98 % (82,15, 83,81)	37,89 % (36,82, 38,96)
HR	99,39 % (99,22, 99,56)	99,39 % (99,22, 99,56)	99,38 % (99,21, 99,55)	99,32 % (99,14, 99,50)	99,04 % (98,83, 99,25)	88,50 % (87,80, 89,20)
HR–HR	99,03 % (98,81, 99,25)	99,03 % (98,81, 99,25)	98,99 % (98,77, 99,21)	98,84 % (98,60, 99,08)	98,40 % (98,12, 98,68)	77,10 % (76,17, 78,03)
FR–HR	99,28 % (99,09, 99,47)	99,26 % (99,07, 99,45)	98,13 % (97,83, 98,43)	94,95 % (94,47, 95,43)	91,81 % (91,21, 92,41)	49,19 % (48,09, 50,29)

Tabla V-25.- Comportamiento del sistema de RMT (para comparar con el convencional véase la Tabla V-23) en escenarios con varias etapas de codificación para la tarea de reconocimiento de dígitos aislados (IDR).

Para la tarea IDR podemos observar cómo, consistentemente con lo que observamos para una sola etapa de codificación, siempre que en alguna de las etapas está presente el codificador GSM-FR el reconocedor mediante transparametrización proporciona ventajas considerables que van desde un 1,51% para la configuración FR-FR con el canal 2, hasta un 38,44% para la configuración HR-FR, el canal 5. Nótese, de nuevo, cómo las mejoras más importantes se consiguen en las combinaciones en donde hay más margen, es decir, donde los errores han sido más dañinos.

Tarea de reconocimiento de habla continua Reconocimiento mediante transparametrización				
Configu-raciones tándem.	Tasas de reconocimiento (%) Intervalo de confianza del 95 %			
	Canales			
	0	2	3	4
–	90,83 % (90,27, 91,39)	–	–	–
FR	87,64 % (87,00, 88,28)	83,45 % (82,73, 84,17)	63,12 % (62,19, 64,05)	50,07 % (49,10, 51,04)
FR–FR	85,58 % (84,90, 86,26)	75,33 % (74,50, 76,16)	21,84 % (21,04, 22,64)	14,11 % (13,44, 14,78)
HR–FR	82,10 % (81,36, 82,84)	75,19 % (74,36, 76,02)	46,32 % (45,36, 47,28)	30,93 % (30,04, 31,82)
HR	88,09 % (87,46, 88,72)	88,10 % (87,47, 88,73)	87,86 % (87,23, 88,49)	87,62 % (86,98, 88,26)
HR–HR	84,22 % (83,52, 84,92)	84,38 % (83,68, 85,08)	84,55 % (83,85, 85,25)	82,46 % (81,73, 83,19)
FR–HR	86,82 % (86,17, 87,47)	83,53 % (82,81, 84,24)	68,07 % (67,17, 68,97)	59,17 % (58,22, 60,12)

Tabla V-26.- Comportamiento del sistema de RMT (para comparar con el convencional véase la Tabla V-24) en escenarios con varias etapas de codificación para la tarea de reconocimiento de habla continua (CSR).

Para la tarea CSR, sin embargo, los errores resultan mucho más críticos (de ahí que no hayamos considerado el canal 5, que producía tasas de reconocimiento inadmisibles) y por lo tanto las mejoras son más abultadas cuando interviene en la segunda etapa (la única sobre la que puede actuar RMT) el codificador GSM-HR, que como indicamos en V.2.1.3 se beneficia más de nuestra propuesta por contar con 10 coeficientes LP. Así, se obtienen mejoras del 5,4% para el canal 4 en HR-HR y de hasta el 52,2% para ese mismo canal en FR-HR.

V.2.3.2. Entre la red IP y la GSM

Cada vez más a menudo, la señal de voz viaja a través de redes de paquetes que se están convirtiendo en la mejor forma de combinar toda clase de medios sobre un mismo soporte de transmisión. Por tanto, es muy probable, que a la hora de instalar servidores de reconocimiento, estos se hallen situados o sean accesibles a través de una red IP. Sin embargo, es también bastante probable que usuarios móviles tengan interés en acceder a este servidor a través de la red de móviles. Nos encontramos, entonces, en la situación de la Figura V-8 y aparece, de nuevo, el problema de la transcodificación. Este tipo de escenarios se contemplan en la iniciativa TIPHON de ETSI, que ya introdujimos en la sección V.1.3.2. y cuyo objetivo es proporcionar interoperatividad entre redes [59].

Es necesario mencionar, sin embargo, que al igual que en el caso de las redes móviles se llevan a cabo esfuerzos por implantar el modo TFO para evitar transcodificaciones, existe gran interés en desarrollar mecanismos y protocolos que permitan utilizar el protocolo IP también sobre redes de móviles (por ejemplo, [107], [108], [17] o [132]). Sin embargo, estas aproximaciones tienen, de momento graves inconvenientes (por ejemplo, la sobrecarga de cabeceras *–overhead–* que produce la utilización del protocolo IP con paquetes de datos tan pequeños como los que generan los codificadores de voz).

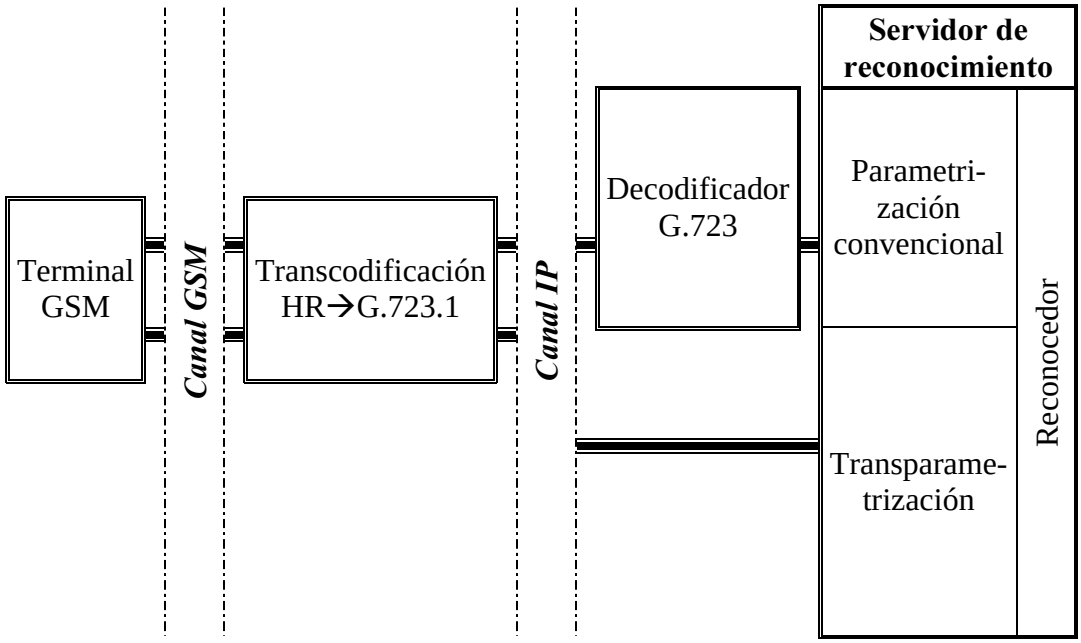


Figura V-8.- Diagrama de bloques que describe los procesos que tienen lugar en un escenario con una etapa de transcodificación donde un terminal móvil GSM requiere los servicios de un servidor de reconocimiento remoto accesible a través de una red IP.

Por ese motivo, hemos considerado oportuno evaluar las prestaciones de nuestro reconocedor en el escenario que planteábamos anteriormente, mostrando los resultados en las Tablas V-27 y V-28, para las tareas IDR y CSR, respectivamente. El codificador que hemos utilizado para la red móvil es el GSM-HR, y para la red IP, el G.723.1.

Así, en primer lugar hemos utilizado el codificador GSM-HR y hemos contaminado la voz codificada haciéndola atravesar distintos canales GSM. A continuación, hemos realizado la decodificación de esa señal de voz y vuelto a codificar con G.723.1. Posteriormente, hemos aplicado las pérdidas de paquetes correspondientes a varios canales IP y el resultado lo hemos introducido en el reconocedor RMT y en el convencional.

Tarea de reconocimiento de dígitos aislados					
		Canales IP			
Canales GSM	Decodificado / RMT	O	B	D	F
	0	97,69 (97,36, 98,02)	97,31 (96,95, 97,67)	96,09 (95,66, 96,52)	94,62 (94,12, 95,12)
		98,43 (98,16, 98,70)	98,22 (97,93, 98,51)	97,41 (97,06, 97,76)	96,28 (95,86, 96,70)
	3	97,53 (97,19, 97,87)	97,13 (96,76, 97,50)	95,88 (95,44, 96,32)	94,61 (94,11, 95,11)
		98,33 (98,05, 98,61)	98,14 (97,84, 98,44)	97,35 (97,00, 97,70)	96,05 (95,62, 96,48)
	4	97,41 (97,06, 97,76)	96,93 (96,55, 97,31)	95,77 (95,33, 96,21)	94,37 (93,86, 94,88)
		98,03 (97,72, 98,34)	97,82 (97,50, 98,14)	97,13 (96,76, 97,50)	95,83 (95,39, 96,27)

Tabla V-27.- Tasas de reconocimiento (%) para la tarea de reconocimiento de dígitos aislados (IDR) en el escenario tándem de la Figura V-8.

Como vemos, en todos los casos se obtienen mejoras estadísticamente significativas al aplicar el procedimiento RMT, si bien estas mejoras resultan más acusadas para la tarea CSR. Así, para la tarea IDR obtenemos mejoras de hasta un 1,52% (canales 4 y F) mientras que en CSR para estos mismos canales obtenemos mejoras del 3,46%.

Tarea de reconocimiento de habla continua					
		Canales IP			
	Decodificado / RMT	O	B	D	F
Canales GSM	0	81,14 (80,38, 81,90)	80,06 (79,29, 80,83)	78,41 (77,61, 79,21)	74,61 (73,77, 75,45)
		81,30 (80,55, 82,05)	80,30 (79,53, 81,07)	79,56 (78,78, 80,34)	75,68 (74,85, 76,51)
	3	80,15 (79,38, 80,92)	78,74 (77,95, 79,53)	77,68 (76,88, 78,48)	74,24 (73,39, 75,09)
		81,43 (80,68, 82,18)	79,97 (79,20, 80,74)	79,42 (78,64, 80,20)	76,54 (75,72, 77,36)
	4	79,88 (79,11, 80,65)	78,47 (77,68, 79,26)	77,38 (76,57, 78,19)	73,08 (72,22, 73,94)
		80,55 (79,79, 81,31)	79,21 (78,43, 79,99)	78,66 (77,87, 79,45)	75,70 (74,87, 76,53)

Tabla V-28.- Tasas de reconocimiento (%) para la tarea de reconocimiento de habla continua (CSR) en el escenario tándem de la Figura V-8.

Capítulo VI
Contribuciones, conclusiones y futuras líneas de trabajo

VI.1. Conclusiones

En esta tesis hemos analizado la influencia de los entornos GSM e IP sobre los reconocedores automáticos de habla. Concretamente, las distorsiones propias estos entornos y que afectan a las prestaciones de los reconocedores son, por una parte, la distorsión de codificación, y por otra, los errores producidos en la transmisión. Estos errores aparecen, como ráfagas de errores de bit, en el entorno GSM, y como ráfagas de paquetes perdidos, en el entorno IP.

Además existen, por supuesto, otro tipo de distorsiones que afectan al reconocimiento en estos entornos y que provienen de variaciones en el entorno acústico, el hablante, la tarea o los efectos de los transductores empleados, pero que también se hayan presentes en otro tipo de escenarios de reconocimiento y han sido tratados por otros autores.

Así, para mejorar el comportamiento de los reconocedores en presencia de las dos distorsiones antes mencionadas, hemos propuesto una técnica que denominamos "Reconocimiento mediante transparametrización" y hemos comparado sus prestaciones con las de un reconocedor convencional que actúa a partir de la voz decodificada.

Los resultados de la comparación nos permiten concluir que:

- El reconocimiento mediante transparametrización proporciona mejoras sobre el convencional (salvo en algunos casos excepcionales, en los que la diferencia entre las dos opciones no es estadísticamente significativa) tanto en el entorno GSM (incluyendo escenarios con transcodificaciones) como en el IP, y también en un escenario que contempla una situación mixta.

- Las mejoras respecto al sistema convencional aumentan conforme las dos distorsiones que afectan en ambos entornos tienen una mayor influencia sobre las prestaciones del reconocedor. Es decir, los beneficios de aplicar nuestra técnica son mayores cuanto peores son las condiciones impuestas por los canales.

- En el entorno GSM, los beneficios obtenidos con el codificador GSM-FR, en lo que se refiere a la distorsión de codificación son menores que los logrados para el GSM-HR porque la descripción que el primero hace de la envolvente espectral es peor (utiliza 8 coeficientes LP en lugar de 10). Sin embargo, el GMS-FR es más sensible a los errores de transmisión, por lo cual los beneficios obtenidos del mejor tratamiento de esta distorsión son más relevantes para este codificador.

- Es fundamental transformar la tasa de trama procedente del codificador en otra más adecuada para la tarea de reconocimiento, especialmente en los casos en los que éstas son muy dispares (por ejemplo, cuando utilizamos el codificador G.723.1 –30 ms– en la tarea de reconocimiento de habla continua –10 ms–). En este sentido, nuestra propuesta, consiste en emplear métodos de interpolación con condicionamientos de tiempo real menos restrictivos que los que afectan al decodificador y por tanto a la aproximación convencional.

- La estimación, a partir de los parámetros de codificación, del valor de energía incluido en la parametrización produce mejoras muy significativas respecto a su obtención a partir de la voz decodificada.

- La obtención de los parámetros cepstrales puede hacerse por varias vías, siendo la del pseudocepstrum ([113], [111] o [25]) menos costosa computacionalmente, pero con peores resultados en presencia de las distorsiones que hemos considerado en el entorno IP.

- La distorsión de codificación se reduce en ambos entornos, ya que los parámetros de reconocimiento se extraen directamente de los de codificación y la información más relevante para reconocimiento suele estar más precisamente descrita en los codificadores. De esta forma, sólo el ruido de cuantificación que afecta a la información relevante perjudica al reconocedor propuesto, habiéndose comprobado que tal perjuicio es muy poco significativo. Por otra parte, hemos constatado que la decodificación puede contribuir a distorsionar la información relevante.

- El codificador de canal en el entorno GSM produce mayores beneficios sobre nuestra propuesta que sobre el reconocedor convencional.

- En el caso de GSM, los errores que alteran los bits que codifican la información no relevante para el reconocimiento no afectan negativamente a las prestaciones del reconocedor propuesto, al contrario que en la solución convencional. Además, el número de bits de información relevante para reconocimiento suele ser considerablemente inferior al de información no relevante, por lo que la probabilidad de error en alguno de los bits pertenecientes al primer grupo es menor que en el segundo.

- En el caso de IP, mediante nuestro procedimiento se independizan los procedimientos de recuperación de tramas perdidas y de suavizado de las transiciones entre tramas establecidas en el codificador, de los correspondientes en el reconocedor (donde el suavizado entre tramas se transforma en adecuación de la tasa de trama). De esta forma, impedimos que algunas de las limitaciones

propias de las técnicas de codificación influyan en nuestra alternativa de reconocimiento.

VI.2. Contribuciones

La principales aportaciones de esta tesis, en orden cronológico, se pueden resumir en:

- Aplicación de la transparametrización de la información espectral propuesta en [66], [67] y [68] para el entorno GSM, al entorno IP, utilizando el codificador G.723.1, para la tarea de reconocimiento de dígitos aislados: identificación de la problemática asociada a este entorno y este codificador ([148]).

- Tratamiento del problema de la disparidad de las tasas de trama del codificador y del reconocedor en las condiciones del punto anterior: sustitución del procedimiento de interpolación entre tramas del codificador por otro orientado a reconocimiento (haciendo uso de varias tramas anteriores y posteriores), justificada tras la discusión de los diferentes requerimentos que limitan una y otra tecnologías ([147]).

- Desarrollo de un método de estimación de la energía para su inclusión en el vector de parámetros de reconocimiento, considerando expresamente que posteriormente se hará uso de procedimientos de interpolación mejores que el utilizado en el codificador estándar y que, por tanto, es beneficioso desligar la obtención de la energía del procedimiento de interpolación de la envolvente espectral en las subtramas del codificador ([145]). Verificación de sus beneficios en dos tareas: reconocimiento de dígitos aislados y de habla continua.

- Introducción de una simulación de los patrones de pérdida de paquetes en Internet que hace uso de medidas sobre tráfico real de voz sobre IP (Borella et al. [15]) con especial énfasis en la longitud de las ráfagas de paquetes perdidos ([145]).

- Aplicación de las anteriores mejoras (interpolación y estimación de energía) al entorno GSM, utilizando una simulación del canal radio más realista ([69]).

- Análisis de los efectos de la codificación de canal sobre la solución propuesta y la convencional ([69]).

- Planteamiento de escenarios con transcodificaciones dentro del sistema GSM ([69]).

- Planteamiento de un escenario con transcodificaciones entre la red GSM y la red IP ([68]).

- Aplicación del procedimiento de obtención de la envolvente espectral a partir de los parámetros LSP al reconocimiento mediante transparametrización y

comparación de los procedimientos de obtención de los parámetros cepstrales con la aproximación denominada “pseudocepstrum” ([145]).

VI.3. Líneas futuras

La finalización de esta tesis deja abiertas las siguientes líneas de investigación:

- Exploración del método de reconocimiento mediante transparametrización en futuras situaciones de implantación de la red IP en entornos móviles. Para ello es necesario conocer los desarrollos y modificaciones que se lleven a cabo en este protocolo para su adaptación a dichos entornos, entre ellos la compresión de cabeceras RTP ([34]), que aumentaría la eficiencia del canal, que es, en estos momentos, muy baja, debido a la sobrecarga de cabeceras en las actuales redes que utilizan RTP/UDP/IP para la transmisión de voz.

- Exploración del método propuesto utilizando la familia de codificadores AMR y estudio de las interacciones que existan entre el codificador de canal y el de fuente. Como hemos visto, los actuales codificadores de canal favorecen al método de reconocimiento que proponemos y por lo tanto es especialmente interesante estudiar esta familia de codificadores, en la que se busca un equilibrio dinámico, precisamente, entre codificador de canal y de fuente.

- También resultaría interesante obtener resultados sobre el codificador G.729.

- El procedimiento de reconocimiento distribuido (DSR) representa otra alternativa al método de reconocimiento que proponemos, en la que (véase la sección II.4.2) los terminales juegan un papel más activo y en donde la parametrización que debe transmitirse al extremo reconocedor está siendo actualmente discutida. Además, también están todavía sin determinar los parámetros de codificación que deben acompañar a esta parametrización para permitir la reconstrucción de la señal de voz cuando ello sea necesario. Así, algunas de las conclusiones de esta tesis pueden ser aplicadas directamente para la selección de esos parámetros, teniendo para ello en cuenta los efectos de los errores que se producen en los dos entornos que hemos considerado aquí. De esta forma podría encontrarse un equilibrio entre ambos tipos de parametrizaciones en función de la QoS requerida en el reconocimiento y en la decodificación.

- Por otra parte, queda abierta la posibilidad de incluir en el vector de parámetros de reconocimiento (tanto en reconocimiento remoto –mediante transparametrización–, como en distribuido) otros de los parámetros disponibles provenientes del codificador.

- Además, pueden explorarse otros procedimientos de obtención de parametrizaciones de reconocimiento robustas, como los expuestos en la sección IV.3.5.

- También pueden ensayarse otras técnicas de recuperación de datos frente a tramas perdidas que tengan en cuenta los anchos de banda de modulación de los parámetros de codificación.

- Sería conveniente ensayar la combinación del método propuesto con aproximaciones al reconocimiento robusto dirigidas a mitigar los efectos del ruido ambiente, tales como la compensación de modelos.

- Por último, queda también abierta la posibilidad de realizar medidas reales sobre el comportamiento de las redes IP para tráfico vocal, que contribuyan a conocer los puntos débiles y aspectos todavía no explorados de esta transmisión, en función de los diversos parámetros configurables de la misma e incluso sus implicaciones en reconocimiento.

Referencias

[1] 3GPP A.S0004-A, *"3GPP2 Tandem Free Operation Specification, version 1.0"*, 2001.

[2] Abreu, M. V., *"Codificación multipulso a tasas variables aplicada a la transmisión fiable de voz sobre redes de datos"* Tesis Doctoral, Dpto. De Tecnoloxías das Comunicacións. E.T.S.E. de Telecomunicación de Vigo, 1999.

[3] Atungsiri, S. A., Soheili, R., Kondoz, A. M., Evans, B. G., *"Effective lost speech frame reconstruction for CELP coders"* Proc. of European Conference on Speech Communication and Technology (Eurospeech), vol. 2, pp. 599-602, 1991.

[4] Avendano, C., *"Temporal Processing of Speech in a Time-Feature Space"*, Tesis Doctoral, Oregon Graduate Institute of Science and Technology, 1997.

[5] Avendano, C., Hermansky, H., *"On the Effects of Short-Term Spectrum Smoothing in Channel Normalization"*, IEEE Transactions on Speech and Audio Processing, Jul. 1997.

[6] Avendano, C., Hermansky, H., *"On the Properties of Temporal Processing for Speech in Adverse Environments"*, Proc. WASPA, Mohonk, Nueva York, 1997.

[7] Baker, F., *"Real Time Services for Router Nets"* http://www.data.com/Tutorials/Real_Time_Services.html, Cisco Systems, 1996.

[8] Beijar, N., *"Signaling Protocols for Internet Telephony"* Helsinki University of Technology, 1998.

[9] Bergmark, D., Keshav, S., *"Building Blocks for IP Telephony"*, IEEE Communications Magazine, vol. 38, no. 4, pp. 88-95, 2000.

[10] Black, U., *"Voice over IP"*, Ed. Prentice Hall ,1999.

[11] Blanch, F., Faúndez, M., *"Reconstrucción de paquetes de voz en comunicaciones de voz por paquetes"*.Proc. URSI, Bilbao, vol. I, pp.705-708, 1997.

[12] Bolot, J. C., *"Characterizing end-to-end packet delay and loss in the internet"*, Journal of High Speed Networks, vol. 2, no. 3, pp. 289-298, Sep. 1993.

[13] Bolot, J., Crepin, H., Vega-Garcia, A., *"Analysis of Audio Packet Loss in the Internet"*, Proc. Workshop on Network and Operaing System Support for Audio and Video, pp. 163-174, 1995.

[14] Bolot, J.-C., Fosse-Parisis, S., Towsley, D., *"Adaptive FEC-based error control for Internet telephony"*, Proc. IEEE Infocom, Piscataway, New Jersey, pp. 1453-60, vol. 3, 1999.

[15] Borella, M. S., *"Measurement and Interpretation of Internet Packet Loss"*, Journal of Communications and Networking, vol. 2, no. 2, pp 93-102, Jun. 2000.

[16] Campbell, A. T., Gomez, J., Kim, S., Valkò, A. G., Wan, C.-Y., Turànyi, Z. W., *"Design, Implementation, and Evaluation of Cellular IP"*, IEEE Personal Communications Magazine, vol. 7, pp. 50-58, Ag. 2000.

[17] Campbell, A. T., Gomez, J., Valko, A. G., *"An Overview of Cellular IP"*, Proc. 1st IEEE Wireless Communications and Networking, Nueva Orleans, pp. 21-24, Sep. 1999.

[18] Canavos, G. C., *"Probabilidad y estadística: aplicaciones y métodos"* Ed. McGraw-Hill, 1988.

[19] Chazan, D., Cohen, G., Hoory, R., Zibulski, M., *"Low Bit Rate Speech Compression for Playback in Speech Recognition Systems"*, Proc. European Signal Processing Conference (EUSIPCO), 2000.

[20] Chazan, D., Cohen, G., Hoory, R., Zibulski, M., *"Speech Reconstruction from Mel-frequency Cepstral Coefficients and Pitch Frequency"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2000.

[21] Chazan, D., Zibulski, M., Hoory, R., Cohen, G., *"Efficient Periodicity Extraction Based on Sine-wave Representation and its Application to Pitch Determination of Speech Signals"*, Proc. of European Conference on Speech Communication and Technology (Eurospeech), Aalborg, Dinamarca ,Sep. 2001.

[22] Chen, J.-H., Cox, R. V., Lin, Y.-C., Jayant, N., Melchner, M. J., *"A low-delay CELP Coder for the CCITT 16 Kb/s speech coding standard"*, IEEE Journal on Selected Areas in Communications, vol 10, no. 5, pp. 830-849, 1992.

[23] Childers, D., Cox, R. V., DeMori, R., Furui, S., Juang, B.H., Mariani, J. J., Price, P., Sagayama, S., Sondhi, M. M., Wischedel, R., *"The Past, Present, and Future of Speech Processing"*, IEEE Signal Processing Magazine, pp. 24-48, May 1998.

[24] Choi, S. H., Kim, H. K., Lee, H. S., *"LSP weighting functions based on spectral sensitivity MEL-frequency warping for speech recognition in digital communications"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 401-404, 1999.

[25] Choi, S. H., Kim, H. K., Lee, H. S., *"Speech recognition using quantized LSP parameters and their transformations in digital communication"*, Speech Communication, vol. 30, pp. 223-233, 2000.

[26] Choi, S. H., Kim, H. K., Lee, H. S., Gray, R. M., *"Speech recognition method using quantised LSP parameters in CELP-type coders"*, Electronics Letters, vol. 34, no. 2, pp. 156-157, 1998.

[27] Chu, W.–C., *"Neural network-based nonlinear prediction model for speech coding : a thesis in electrical engineering"* Tesis Doctoral, Ed. UMI Dissertation Services, 1999.

[28] Claffy. K. C. *"Internet Traffic Characterization"*, Tesis Doctoral, University of California, San Diego, 1994.

[29] Cohen, G., Ramabadran, T., Tucker, R., *"Requirements for speech reconstrucción from the standard cepstral features"* Documento de trabajo, ETSI STQ AURORA DSR WG, En. 2001.

[30] Combescure, P., Schnitzler, J., Fischer, K., Kirchherr, R., Lamblin, C., Le Guyader, A., Massaloux, D., Quinquis, C., Stegmann, J., Vary, P., *"A 16, 24, 32 Kbit/s wideband speech codec based on ATCELP"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 5-8, Phoenix, EE.UU., 1999.

[31] Cooke, M., Green, P., Josifovski, L., Vizinho, A., *"Robust automatic speech recognition with missing and unreliable acoustic data"*, Speech Communication, 34, pp.267-285, 2001.

[32] Cox, R. V., *"Three new speech coders from the ITU cover a range of applications"*, IEEE Communications Magazine, pp. 40-51, 1997.

[33] Crovella, M. E., Bestavros, A. *"Self-similarity in world wide web traffic evidence and possible causes"* Proc. ACM SIGMETRICS, pp. 160-169, Philadelphia, PA, May 1996.

[34] Degermark, M., Hannu, H., Jonsson, L.-E, Svanbro, K., *"Evaluation of CRTP Performance over Cellular Radio Links"*, IEEE Personal Communications Magazine, vol. 7, pp. 26-33, Aug. 2000.

[35] Deller, J. R., *"Discrete-time processing of speech signals"*, Ed. Macmillan Publishing, 1993.

[36] Dettmer, R., *"Packet Phone"*, IEE Review, vol. 44, no. 2, pp. 58-61, 1998.

[37] Digalakis, V. V., Neumeyer, L., Perakakis, M., *"Quantization of cepstral parameters for speech recognition over the world wide web"* Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 1998.

[38] Digalakis, V. V., Neumeyer, L.G, Perakakis, M., *"Quantization of Cepstral Parameters for Speech Recognition Over the World Wide Web"*, IEEE Journal on Selected Areas in Communications, vol. 17, no. 1, pp. 82-90, Jan. 1999.

[39] Douskalis, B., *"IP telephony : the integration of robust VoIP services"*, Ed. Prentice Hall, 2000.

[40] Dufour, S., Glorion, C., Lockwood, P., *"Evaluation of the Root-Normalised Front-End (RN_LFCC) for Speech Recognition in Wireless GSM Network Environments"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Atlanta, EE.UU., vol. 2, pp. 77-80, 1996.

[41] ETSI Aurora Project; Distributed Speech Recognition: www.etsi.org/technicalactiv/dsr.htm

[42] ETSI Speech processing, Transmission and Quality aspects (STQ), *"Distributed Speech Recognition Front-end extension for tonal language recognition and speech reconstruction"* (DES/STQ-00030), 2001.

[43] ETSI Speech processing, Transmission and Quality aspects (STQ) *"Distributed speech recognition (DSR); Front-end feature extraction algorithm; Compression algorithms"* (ES 201 108), Ab. 2000.

[44] ETSI Recommendation GSM 05.01, *"Digital cellular telecommunications system (Phase 2). Physical layer on the radio path. General description"*, Sep. 1994.

[45] ETSI Recommendation GSM 05.03, *"Digital cellular telecommunications systems; Channel Coding"*, Feb. 1992.

[46] ETSI Recommendation GSM 05.04, *"Digital cellular telecommunications system (Phase 2). Modulation"*, Feb. 1992.

[47] ETSI Recommendation GSM 06.02, *"Digital cellular telecommunications system; half rate speech; Half rate speech processing functions"*, Dic. 1995.

[48] ETSI Recommendation GSM 06.10 *"Digital cellular telecommunications system; full rate speech transcoding"*, Feb. 1992.

[49] ETSI Recommendation GSM 06.11, *"Digital cellular telecommunications systems; Substitution and muting of lost frames for full rate speech channels"*, Feb. 1992.

[50] ETSI Recommendation GSM 06.20, *"Digital cellular telecommunications systems; Half Rate speech; Part 2: Half Rate Speech Transcoding"*, Dic. 1995.

[51] ETSI Recommendation GSM 06.21, *"Digital cellular telecommunications systems; Substitution and muting of lost frames for half rate speech channels"*, Dic. 1995.

[52] ETSI Recommendation GSM 06.60, *"Digital cellular telecommunications systems; Enhanced Full Rate (EFR) speech transcoding"*, Mar. 1997.

[53] ETSI Recommendation GSM 06.61, *"Digital cellular telecommunications systems; Substitution and muting of lost frames for Enhance Full Rate (EFR) speech traffic channels"*, Ab. 1997.

[54] ETSI Recommendation GSM 06.90, *"Digital cellular telecommunications systems; Adaptive Multi-Rate (AMR) speech transcoding"*, Ab. 2000.

[55] ETSI Recommendation GSM 06.91, *"Digital cellular telecommunications systems; Substitution and muting of lost frames for Adaptive Multi Rate (AMR) speech traffic channels"*, Ab. 2000.

[56] ETSI Recommendation GSM 08.62, *"Digital cellular telecommunications systems; Inband Tandem Free Operation (TFO) of Speech Codecs; version 7.0.0"*, Jul. 1999.

[57] ETSI Telecommunications and Internet Protocol Harmonization Over Networks (TIPHON); *"Using GSM speech codecs within ITU-T Recommendation H.323"* (TS 101 318), Ag. 1998.

[58] ETSI Telecommunications and Internet Protocol Harmonization Over Networks (TIPHON), *"End to End Quality of Service in TIPHON Systems; Part 6: Actual measurements of networks and their influence on voice quality"* (TR 101 329-6), Jul. 2000.

[59] ETSI Telecommunications and Internet Protocol Harmonization Over Networks (TIPHON) *"Network architecture and reference configurations"* (TS 101 314), Sep. 2000.

[60] ETSI Telecommunications and Internet Protocol Harmonization Over Networks (TIPHON), *"Technology Compliance Specification; Part 5: Quality of Service (QoS) measurement methodologies"* (TS 101 329-5), Nov. 2000.

[61] Euler, S., Zinke, J., *"The Influence of Speech Coding Algorithms on Automatic Speech Recognition"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Australia, vol. 1, pp. 621-624, 1994.

[62] Farvardin, N., Laroia, R., *"Efficient encoding of speech LSP parameters using the discrete cosine transformation"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp.168-171, 1989.

[63] Ferguson, P., Huston, G., *"Quality of Service. Delivering QoS in the Internet and in Corporate Networks"*, Ed. John Wiley and Sons, 1998.

[64] Furui, S., *"Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum"* IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 34, no. 1, pp. 52-59, 1986.

[65] Gallardo-Antolín, A., Díaz-de-María, F., Valverde-Albacete, F., *"Avoiding Distortions Due to Speech Coding and Transmission Errors in GSM ASR Tasks"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Phoenix, Arizona, EE.UU., vol. I, pp. 277-280, 1999.

[66] Gallardo-Antolín, A., Díaz-de-María, F., Valverde-Albacete, F., *"Recognition from GSM Digital Speech"*, Proc. International Conference on Spoken Language Processing (ICSLP) , Sidney, Australia, 1998.

[67] Gallardo-Antolín, A., Díaz-de-María, F., Valverde-Albacete, F., Bravo-Menéndez-Rivas, R., *"Reconocimiento de voz procedente de teléfonos móviles digitales"*, Telecom I+D, Madrid, pp. 379-387, 1998.

[68] Gallardo-Antolín, A., Peláez-Moreno, C., Díaz-de-María, F., *"A robust front-end for ASR over IP and GSM networks: an integrated scenario"* Proc. of European Conference on Speech Communication and Technology (Eurospeech), vol.2, pp. 1103-1106, Aalborg, Dinamarca, Sep. 2001.

[69] Gallardo-Antolín, A., Peláez-Moreno, C., Díaz-de-María, F., *"Recognizing from GSM digital speech"*, IEEE Trans. On Speech and Audio Processing, mayo 2002 (propuesto para publicación).

[70] Gallardo-Antolín, A., Vázquez-Castro, M., Díaz-de-María, F., Valverde-Albacete, F. and Pérez-Fontán, F., *"BER Performance Assessment of the Land Mobile GSM Channel with Application to Automatic Speech Recognition Tasks"*, Proc. 5th Bayona Workshop on Emerging Technologies in Telecommunications. vol. 1, pp. 212-216, 1999.

[71] Goldberg, R., Lance, R., *"A practical handbook of speech coders"*, Ed. CRC Press , 2000.

[72] Goodman, B., *"Internet Telephony and Modem Delay"* IEEE Network, vol. 13, pp. 8-17, 1999.

[73] Goralski, W. J., Kolon, M. C., *"IP telephony"*, Ed. McGraw-Hill, 1999.

[74] Goyal, P., Greenberg, A., Kalmanek, C. R., Marshall, W. T., Mishra, P., Nortz, D., Ramakrishnan, K. K., *"Integration of Call Signaling and Resource Management for IP Telephony age"* IEEE Network, vol. 13, pp. 24-33, 1999.

[75] Grassi, S., Besacier, L., Dufaux, A., Ansorge, M., Pellandini, F., *"Influence of GSM speech coding on the performance of text-independent speaker recognition"*, International Workshop on Intelligent Communication Technologies and Applications, with emphasis on mobile communications, Neuchâtel, Switzerland, May 1999.

[76] Haeb-Umbach, R., *"Robust Speech Recognition for Wireless Networks and Mobile Telephony"*, Proc. of European Conference on Speech Communication and Technology (Eurospeech), pp. 2427-2430, 1997.

[77] Hanson, B., Applebaum, T., Junqua, J.-C, *"Spectral dynamics for speech recognition under adverse conditions"*, en Automatic speech and speaker recognition: advanced topics, Editores: Lee, C.-H, Soong, F. K. y Paliwal, K.K, Ed. Kluwer Academic Publishers, pp. 331-356, 1996.

[78] Hanzo, L., Somerville, F. C. A, Woodard, J. P. *"Voice Compression and Communications Principles and Applications for Fixed and Wireless Channels"* Ed. John Wiley & Sons, 2001.

[79] Hassan, M., Nayandoro, A., Atiquzzaman, M., *"Internet Telephony: Services, Technical Challenges and Products"*, IEEE Communications Magazine, vol. 38, no. 4, pp. 96-103, 2000.

[80] Heinen, S., Adrat, M., Steil, O., Vary, P., Xu, W., *"A 6.1 to 13.3-kb/s Variable Rate CELP Codec (VR-CELP) for AMR Speech Coding"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)'99 (Phoenix/Arizona), vol. I, pp. 9-12, Mar. 1999.

[81] Hermansky, H., *"Human speech perception: some lessons from automatic speech recognition"* Proc of TSD, República Checa, 2001.

[82] Hermansky, H., *"Perceptual Linear Predictive (PLP) analysis of speech"* Journal Acoust. Soc. Am, vol. 87, no. 4, pp. 1738-1752, 1990.

[83] Hermansky, H., *"Speech beyond 10 ms (Temporal Filtering in Feature Domain)"*, Invited Keynote Lecture, in Proc. of the International Workshop on Human Interface Technology, Aizu, Japan, September 1994.

[84] Hermansky, H., Sharma, S., *"TRAPs: classifiers of TempoRAl Patterns"*, Proc. International Conference on Spoken Language Processing (ICSLP), Australia, 1998.

[85] Huerta, J. M., *"Speech Recognition in Mobile Environments"*, Tesis Doctoral, Abril, 2000.

[86] Huerta, J. M., Stern, R. M., *"Speech compression from GSM codec parameters"* Proc. International Conference on Spoken Language Processing (ICSLP) , Sidney, vol. 4, pp. 1463-1466, 1998.

[87] Huitema, C., Cameron, J., Mouchtaris, P., Smyk, D., *"An Architecture for Residential Internet Telephony Service"*, IEEE Network, vol. 13, pp. 50-57, 1999.

[88] Husain, A., Cuperman, V., *"Reconstruction of missing packets for CELP-based speech coders"* Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 245-248, 1995.

[89] IETF *"Differentiated Services (diffserv) Working Group (IETF) homepage"*. http://www.ietf.org/html.charters/diffserv-charter.html, Última modificación: Oct. 2001.

[90] IETF *"IP Telephony (iptel) Working Group (IETF) homepage"*. http://www.ietf.org/html.charters/iptel-charter.html, Última modificación.: Nov. 2001.

[91] IETF RFC 2330 *"Framework for IP Performance Metrics"*, May. 1998.

[92] IETF RFC 2474. *"Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers"*, Dic. 1998.

[93] Iida, K., Kawahara, K., Takine, T., Oie, Y., *" Performance Evaluation of the Architecture for End-To-End Quality-Of-Service Provisioning"*, IEEE Communications Magazine, vol. 38, no. 4, pp. 76-81, 2000.

[94] ITU-T Recommendation G.114, *"Transmission systems and media, general recommendation on the transmission quality for an entire international telephone connection; one-way transmission time"*, Mar., 1993.

[95] ITU-T Recommendation G.723.1, *"Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s"*, Mar, 1996.

[96] ITU-T Recommendation G.726, *"40, 32, 24, 16 kbps adaptive differential pulse code modulation (ADPCM)"*, 1990.

[97] ITU-T Recommendation H. 323, *"Packet-based multimedia communications systems"*, 1998.

[98] ITU-T Recommendation I.380, *"Internet Protocol Data Communication Service – IP Transfer and Availability Performance Parameters"* May, 1999.

[99] Jayant, N. S., *"Digital coding of waveforms : principles and applications to speech and video"*, Ed. Prentice Hall, 1984.

[100] Jelinek, F., *"Statistical methods for speech recognition"*, Ed. MIT Press, 1997.

[101] Josifovski, L., Cooke, M.P, Green, P., Vizinho, A., 1999, *"State Based Imputation of Missing Data for Robust Speech Recognition and Speech Enhancement"*, Proc. of European Conference on Speech Communication and Technology (Eurospeech), vol. 6, pp. 2837-2840, Budapest, 1999.

[102] Julia, L., Chyer, A., Neumeyer, L., Dowding, J., Charafeddine, M., [Online]. http://www.speech.sri.com/demos/atis.html, 1996.

[103] Junqua, J. C., *"Robust speech recognition in embedded systems and PC applications"*, Ed. Kluwer Academic Publishers, 2000.

[104] Junqua, J. C., *"Robustness in automatic speech recognition : fundamentals and applications"*, Ed. Kluwer Academic Publishers, 1996.

[105] Kanal, L. N., Sastry, A. R. K., *"Models for Channels with Memory and Their Applications to Error Control,"* Proc. of the IEEE, vol. 66, pp. 724-744, Jul. 1978.

[106] Kanedera, N., Hermansky, H., Arai, T., *"Desired characteristics of modulation spectrum for robust automatic speech recognition"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp.613-616, Seattle, EE.UU., May. 1998.

[107] Kanter, T., Olrog, C., *"VoIP in Applications for Wireless Access"*, Proc. IEEE Workshop on Local and Metropolitan Area Networks (LANMAN), pp. 44-47, Nov, 1999.

[108] Kanter, T., Olrog, C., Maguire, G., *"VoIP over Wireless for Mobile Multimedia Applications"*, Proc. of the Personal Computing and Communication Workshop (Fall). Nov, 1999.

[109] Karray, L., Jelloun, A. B., Mokbel, C., *"Solutions for robust recognition over the GSM cellular network"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 261-264, 1998.

[110] Kim, H. K., Choi, S. H., Lee, H. S., *"Dynamic cepstral representations based on order-dependent windowing methods"*, IEICE Transactions on Information and Systems, vol. E81-D, no. 5, pp. 434-440, May. 1998.

[111] Kim, H. K., Choi, S. H., Lee, H. S., *"On approximating Line Spectral Frequencies to LPC Cepstral Coefficients"* IEEE Trans. On Speech and Audio Processing, vol. 8, no. 2, Mar. 2000.

[112] Kim, H. K., Cox, V., *"A bitstream-based front-end for wireless speech recognition on IS-136 communications system"*, IEEE Transactions on Speech and Audio Processing, vol. 9, no. 5, Jul. 2001.

[113] Kim, H. K., Kim, K. C., Lee, H. S., *"Enhanced distance measure for LSP-based speech recognition"*, Electronic Letters, vol 29, no. 16, pp. 1463-1465, 1993.

[114] Kleijn, W.B, Paliwal, K. K., *"Speech coding and synthesis"* Ed. Elsevier Science, 1995.

[115] Koishida, K., Tokuda, K., Kobayashi, T., Imai, S., *"Efficient encoding of mel-generalized cepstrum for CELP coders"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 1355-1358, 1997.

[116] Kondoz, A. M., *"Digital speech : coding for low bit rate communication systems"*, Ed. John Wiley & Sons, 1996.

[117] Korhonen, J., *"Introduction to 3G mobile communications"*, Ed. Artech House, 2001.

[118] Kostas, T. J., Borella, M.S, Sidhu, I., Schuster, G.M, Grabiec, J., Mahler, J., *"Real-Time Voice Over Packet-Switched Networks"*, IEEE Network, pp.18-27, Jan./Feb. 1998.

[119] Kövesi, C., Lamblin, C., Quinquis, C., Thiérion, P., Navarro, W., *"A multi-rate codec family based on GSM EFR and ITU-T G.729"* Proc. of European Conference on Speech Communication and Technology (Eurospeech), Budapest, 1999.

[120] Kroon, P., Recchione, M., *"A low complexity toll-quality variable bit rate coder for CDMA cellular systems"* Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 5-8, 1995.

[121] Kumar, A. *"Comparative performance analysis of versions of TCP in a local network with a lossy link"*, In IEEE/ACM Transactions on Networking, vol. 6, pp. 485-498, 1998.

[122] Kuthan, J., *"Internet Telephony- An overview"*, GloNe-GMD-FOKUS, 1998.

[123] LeBlanc, W., Liu, C., Viswanathan, V., *"An enhanced full rate speech coder for digital cellular applications"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Atlanta, USA, vol. 1, pp. 569-572, 1996.

[124] Li, B., Hamdi, M., Jiang, D., Cao, X-R, Hou, Y. T., *"QoS-Enabled Voice Support in the Next-Generation Internet: Issues, Existing Approaches and Challenges"*, IEEE Communications Magazine, vol. 38, no. 4, pp. 54-61, 2000.

[125] Liao, W., Liu, J.-C, *"VoIP Mobility in IP/Cellular Network Internetworking"*, IEEE Communications Magazine, vol. 38, no. 4, pp. 70-75, 2000.

[126] Lilly, B. T., Paliwal, K. K., *"Effect of Speech Coders on Speech Recognition Performance"*, Proc. International Conference on Spoken Language Processing (ICSLP) , vol. 4, pp. 2344-2347, Philadelphia, EE.UU., 1996.

[127] Lin, S., Costello, D. J., *"Error control coding : fundamentals and applications"*, Ed. Prentice-Hall, Englewood Cliffs, New Jersey, 1983.

[128] Lippman, R.P, Carlson, B.A, *"Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise"* Proc. of European Conference on Speech Communication and Technology (Eurospeech), pp. KN 37-40, Rodas, 1997.

[129] Lockwood, P., Alexandre, P. *"Root adaptive homomorphic deconvolution schemes for speech recognition in noise"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp 441-444, 1994.

[130] Maddux, M. A., *"Voice over IP. The impact of IPv6 on VoIP"* Whitepaper (Digital), May, 1998.

[131] Maes, S. H., Cohen, G., Hoory, R., Chazan, D., *"Conversational Networking: Conversational Protocols for Transport, Coding and Control"*, Proc. International Conference on Spoken Language Processing (ICSLP), Beijing, China, Oct. 2000.

[132] McCann, P. J., Hiller, T., *"An Internet Infrastructure for Cellular CDMA Networks Using Mobile IP"*, IEEE Personal Communications Magazine, vol. 7, pp. 34-41, Ag. 2000.

[133] Minoli, D., Minoli, E., *"Delivering Voice over IP Networks"* Ed. Wiley&Sons, 1998.

[134] Mokbel, C., Mauuary, L., Karray, L., Jouvet, D., Monné, J., Simonin, J., Bartkova, K., *"Towards improving ASR robustness for PSN and GSM telephone applications"*, Speech Communication vol 23. no. 1-2, pp. 141-159, 1997.

[135] Müller, J-M, Wächter, B., *"A codec candidate for the GSM half rate speech channel"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 257-260, 1994.

[136] Nadeu, C., Pachès-Leal, P., Juang, B.-H, *"Filtering the time sequences of spectral parameters for speech recognition"*, Speech Communication. vol. 22, no. 4, pp. 315-32, Sep. 1997.

[137] National Institute of Standards and Technology (NIST) (distribuidor), *"The Resource Management corpus part 1 (RM1)"*, 1992.

[138] Oliphant, M. W., *"The mobile phone meets the internet"*, IEEE Spectrum, pp. 20-28, Aug. 1999.

[139] Olivier, H., *"IP telephony : packet-based multimedia communications systems"*, Ed. Addison-Wesley , 2000.

[140] Oppenheim, A. V., Schafer, R. W., *"Discrete-time signal processing"* (2ª edición), Ed. Prentice Hall, 1999.

[141] Patel, G., Dennett, S., *"The 3GPP and 3GPP2 Movements Toward an All-IP Mobile Network"*, IEEE Personal Communications Magazine, vol. 7, pp. 62-64, Ag. 2000.

[142] Paxon, V., *"Measurements and Analysis of End-to-End Internet Dynamics"*, Tesis doctoral, University of California, Berkeley, 1997.

[143] Paxson, V., Floyd, S., *"Wide-area Traffic: The Failure of Poisson Modeling,"* IEEE/ACM Transactions on Networking, pp. 226-244, Jun. 1995.

[144] Pazos, C. M., Kotelba, M. R., Malis, A. G., *"Real-Time Multimedia Over ATM"*, IEEE Communications Magazine, vol. 38, no. 4, pp. 82-87, 2000.

[145] Peláez-Moreno, C., Gallardo-Antolín, A., Díaz-de-María, F., "*LSP-derived ASR robust feature extraction for IP environments*" IEEE Trans. On Speech and Audio Processing (propuesto para publicación).

[146] Peláez-Moreno, C., Gallardo-Antolín, A., Díaz-de-María, F., *"Recognizing Voice over IP networks: a Robust Front-End for Speech Recognition on the WWW"*, IEEE Trans. on Multimedia, vol. 3, no. 2, pp. 209-18, Jun. 2001.

[147] Peláez-Moreno, C., Gallardo-Antolín, A., Díaz-de-María, F., *"Recognizing IP over IP: towards Spoken Language Interfaces for E-business"*, Proc. eBusiness and eWork Conference, pp. 1065-1071, Madrid, 2000. o Peláez-Moreno, C., Gallardo-Antolín, A., Díaz-de-María, F., *"Recognizing IP over IP: towards Spoken Language Interfaces for E-business"* E-business: Issues, Applications and Technologies, pp 1065-1071, Smith and P.T. Kidd (Eds.) IOS Press, 2000.

[148] Peláez-Moreno, C., Zambrano-Miranda, A., Gallardo-Antolín, A., Díaz-de-María, F., *"Reconocimiento de habla en internet: una aproximación eficiente"*, Proc. Telecom I+D, Madrid, 1999.

[149] Perkins, C., Hodson, O., Hardman, V., *"A Survey of Packet-Loss Recovery Techniques for Streaming Audio"*, IEEE Network, Vol.12, no.5, pp.40-48, 1998

[150] Polyzois, C. A., Purdy, K. H., Yang, P.-F, Shrader, D., Sinnreich, H., Ménard, F., Schulzrinne, H., *"From POTS to PANS: A Commentary on the Evolution to Internet Telephony"*, IEEE Network, pp. 58, vol. 13, 1999.

[151] Rabiner, L., Juang, B.–H, *"Fundamentals of speech recognition"*, Ed. Prentice-Hall, Englewood Cliffs, 1993.

[152] Ramaswamy, G. N., Gopalakrishnan, P. S., *"Compression of acoustic features for speech recognition in network environments"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 1998.

[153] Ramjee, R., La Porta, T. F., Salgarelli, L., Thuel, S., Varadhan, K., Li, L., *"Based Access Network Infrastructure for Next-Generation Wireless Data Networks"*, IEEE Personal Communications Magazine, vol. 7, pp. 42 – 49, Aug. 2000.

[154] Rao, H., Lin, Y.-B, Cho, S.-L, *"IGSM: VoIP Service for Mobile Networks"*, IEEE Communications Magazine, vol. 38, no. 4, pp. 62-69, 2000.

[155] Rayes, A., Sage, K., *"Integrated Management Arquitecture for IP-based Networks"*, IEEE Communications Magazine, vol. 38, no. 4, pp. 48-53, 2000.

[156] Redl, S. M., Weber, M. K., Oliphant, M. W., *"GSM and personal communications handbook"*, Ed. Artech House , 1998.

[157] Rizzetto, D., Catania, C., *"Voice over IP Service Architecture for Integrated Communications"* IEEE Network, pp. 34-41, vol. 13, 1999.

[158] Rosenberg, J., Lennox, J., Schulzrinne, H., *"Programming Internet Telephony Services"*, IEEE Network, pp. 42-49, vol. 13, 1999.

[159] Rosenberg, J., Qiu, L., Schulzrinne, H., *"Integrating Packet FEC into Adaptive Voice Playout Buffer Algorithms on the Internet"* Proc. of the Conference on Computer Communications (IEEE Infocom), Mar. 2000.

[160] Ryan, J., *"Voice over IP"* The Technology Guide Series. Ed. Telogy Networks, 1998.

[161] Salami, R., Laflamme, C., Adoul, J-P, Kataoka, A., Hayashi, S., Moriya, T., Lamblin, C., Massaloux, D., Proust, S., Kroon, P., Shoham, Y., *"Design and Description of CS-ACELP: A toll quality 8 Kb/s Speech coder"* IEEE Trans. On Speech and Audio Processing, vol. 6, no. 2, pp. 116-130, 1998.

[162] Salonidis, T., Digalakis, V., *"Robust speech recognition for multiple topological scenarios of the GSM mobile phone system"*, Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. 101-104, 1998.

[163] Schulzrinne, H., Rosenberg, J., *"The IETF Internet Telephony Architecture and Protocols"* IEEE Network, pp.18-23, vol. 13, 1999.

[164] Shen, K., Bin, J., Cohen, G., *"Standardization of pitch determination, compression and transmission for DSR terminals"* Documento de trabajo STQ AURORA DSR WG, Feb. 2001.

[165] Spanias, A. S., *"Speech Coding: A tutorial review"*, Proc. IEEE vol.82, no.10, pp. 1541-1552, 1994.

[166] Sriratanaban, C., Kondoz, A., *"A full-rate GSM AMR candidate"* Proc. of European Conference on Speech Communication and Technology (Eurospeech), Budapest, 1999.

[167] Sugamura, N., Itakura, F., *"Speech analysis and synthesis methods developed at ECL in NTT –from LPC to LSP-"*, Speech Communications, vol. 5, pp. 199-215, 1986.

[168] Tebelskis, J., *"Speech Recognition using Neural Networks"*, Tesis Doctoral, Carnegie Mellon University, May. 1998.

[169] Thom, G. A., *"H.323: The multimedia communications standard for local area networks"*, IEEE Communications Magazine, pp. 52-56, 1996.

[170] Thomsen, F., Jani, Y., *"Internet telephony: going like crazy"*, IEEE Spectrum, May. 2000.

[171] TIA/EIA-829, *"Interoperatibility Specification for Tandem Free Operation"*, 2000.

[172] Tucker, R., Robinson, T., Christie, J., Seymour, C., *"Compression of acoustic features - are perceptual quality and recognition performance incompatible goals?"*, Proc. of European Conference on Speech Communication and Technology (Eurospeech), vol. 5, pp. 2155-2158, Budapest, 1999.

[173] Villette, S., Stefanovic, M., Kondoz, A., *"A multi-rate speech and channel codec: a GSM AMR half-rate candidate"* Proc. of European Conference on Speech Communication and Technology (Eurospeech), Budapest, 1999.

[174] Vizinho, A., Green, P., Cooke, M. & Josifovski, L. *"Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study"*, Proc. of European Conference on Speech Communication and Technology (Eurospeech), vol. 5, pp.2407-2410, Budapest, 1999.

[175] Watkins, C. R., Chen, J.-H, *"Improving 16 Kb/s LD-CELP coder for frame erasure channels"* Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 241-244, 1995.

[176] Yletyinen, T., *"The quality of voice over IP"*, Tesis de Máster, Helsinki University of Technology, 1998.

[177] Young, S. et al, *"HTK-Hidden Markov Model Toolkit (ver. 3.0)"*, Cambridge University, 2000.

[178] Zvonar, Z., Jung, P., Kammerlander, K. (eds.), *"GSM: evolution towards 3rd generation systems"*, Ed. Kluwer Academic Publishers, 1999.