TÉCNICAS DE ESTIMACIÓN Y RECONSTRUCCIÓN PARA LA TRANSMISIÓN ROBUSTA DE LA VOZ CODIFICADA

Tesis presentada por Domingo López Oller
para obtener el título de doctor por la Universidad de Granada en el programa de doctorado en Tecnologías de la Información y la Comunicación.

Departamento de Teoría de la Señal, Telemática y Comunicaciones.
Universidad de Granada

2017

Editor: Universidad de Granada. Tesis Doctorales
Autor: Domingo López Oller
ISBN: 978-84-9163-396-9
URI: http://hdl.handle.net/10481/47843

D. Ángel Manuel Gómez García y **D. José Luis Pérez Córdoba**, profesores de la Universidad de Granada en el Departamento de Teoría de la Señal, Telemática y Comunicaciones

CERTIFICAN:

Que la memoria titulada "**TÉCNICAS DE ESTIMACIÓN Y RECONSTRUCCIÓN PARA LA TRANSMISIÓN ROBUSTA DE LA VOZ CODIFICADA**"ha sido realizada por **Domingo López Oller** bajo nuestra dirección en el Departamento de Teoría de la Señal, Telemática y Comunicaciones de la Universidad de Granada para optar al título de Doctor por la Universidad de Granada en el programa de doctorado de Tecnologías de la Información y la Comunicación.

Granada, a 1 de junio de 2017

Directores de tesis:
Fdo. Ángel Manuel Gómez García Fdo. José Luis Pérez Córdoba

Doctorando:
Fdo. Domingo López Oller

The doctoral candidate **D. Domingo López Oller** and the thesis supervisors: **D. Ángel Manuel Gómez García** and **D. José Luis Pérez Córdoba**.

Guarantee, by signing this doctoral thesis, that this work has been done by the doctoral candidate under the direction of the thesis supervisor/s and, as far as our knowledge reaches, in the performance of the work, the rights of other authors to be cited (when their results or publications have been used) have been respected.

Place and date: Granada, june 1, 2017

Thesis supervisors:

Ángel Manuel Gómez García

José Luis Pérez Córdoba

Doctoral candidate:

Domingo López Oller

"No podemos resolver problemas usando el mismo tipo de pensamiento que usamos cuando los creamos"

Albert Einstein

"No hay más que un modo de dar una vez en el clavo, y es dar ciento en la herradura"

Miguel de Unamuno

Agradecimientos

Haciendo balance de los años que han supuesto la elaboración de esta tesis, con sus buenos y malos momentos, no quería dejar pasar esta oportunidad para agradecer en estas líneas la ayuda prestada por todos los que habéis estado a mi lado estos años.

Aunque la lista de agradecimientos sería interminable, me gustaría empezar por Antonio Miguel Peinado, catedrático de Teoría de la Señal y Comunicaciones y coordinador del grupo SIGMAT en la Universidad de Granada, por haberme prestado la oportunidad de realizar esta tesis bajo la subvención de una beca de Formación de Profesorado Investigador. Hay que reconocer que tu motivación y ayuda dentro y fuera del departamento hace que los becarios sintamos ilusión por el trabajo que realizamos.

Quiero continuar agradeciendo a mis tutores Ángel Manuel Gómez García y José Luis Pérez Córdoba, su enorme interés y dedicación para que esta tesis saliera adelante. Siempre habéis tenido una buena disposición y buenos consejos para darle forma tanto a las publicaciones en congresos y revistas como para este documento en sí mismo. Del mismo modo, también quiero expresar mi agradecimiento a Victoria Sánchez con la que he compartido las clases prácticas durante la beca y a la que agradezco su disposición para resolverme las dudas y el desarrollo de las mismas. Y como no, a los compañeros becarios Iván López Espejo y Ján Koloda con los que he compartido esta etapa de mi vida, tanto a nivel intelectual a la hora de debatir sobre algún procedimiento, como lúdico en los ICOM's o durante los fines de semana. Gracias de corazón por el apoyo que me habéis ofrecido durante estos años.

También quiero agradecer a Tim Fingscheidt, catedrático del departamento de Teoría de la Señal en la Universidad de Braunschweig, su agradable acogida en el departamento durante la estancia de tres meses en Alemania y por su interés para que me sintiera integrado dentro del mismo.

Por último, no puedo olvidar a otras personas que se han visto involucradas en

esta travesía, como son mis familiares y amigos que siempre han estado a mi lado. En especial, para mis padres y hermanos por su apoyo incondicional y motivación que me han dado desde el inicio de mis estudios de Ingeniería Informática para poder lograr mis metas, a mi mujer Aldana Sacchi por estar siempre a mi lado cuando la he necesitado y a la familia Sacchi que me ha levantado el ánimo en multitud de ocasiones. Esta tesis está dedicada para todos vosotros.

Índice general

Índice de figuras

Índice de tablas

Resumen

En los últimos años se ha producido un gran desarrollo y despliegue de las tecnologías de comunicación que posibilitan una conexión ubicua y permanente tanto a Internet como a la red de telefonía mundial. Unido a este desarrollo hay que destacar el incremento del número de aplicaciones y servicios que han modificado por completo la forma en la que nos comunicamos, transmitimos e intercambiamos ideas, sentimientos o información. Esto ha hecho que las comunicaciones actuales ya no estén diferenciadas por voz o datos (texto, video, imágenes,...) sino que cada vez son más las aplicaciones que hacen uso de la voz y los datos de manera simultánea (como, por ejemplo, el caso de las videoconferencias). Este hecho, ligado a la expansión y desarrollo que ha experimentado Internet, basada en el protocolo IP, ha dado lugar a una convergencia de las redes de comunicación de voz tradicionales y las redes de comunicación de datos hacia éstas últimas, facilitando así tanto el acceso del usuario como la escalabilidad de las aplicaciones desarrolladas.

No obstante, para proporcionar una buena calidad de servicio, es necesario que la comunicación de voz se lleve a cabo sin degradaciones y en tiempo real. Sin embargo, esta calidad está condicionada por el hecho de que las redes de comunicaciones digitales no están exentas de errores durante la transmisión, considerados éstos como alteraciones o pérdidas en los paquetes de datos enviados, debidas a las condiciones de la red y el entorno donde se realiza la transmisión.

El interés de esta tesis se centra en el estudio de la degradación producida en transmisiones sobre dos tipos de redes de diferente alcance: las redes inalámbricas de ámbito local y las redes IP. Por un lado, en las redes inalámbricas de ámbito local, la degradación produce una alteración en los paquetes recibidos causada por el efecto multitrayecto. Por otro lado, en las redes IP, la degradación conlleva la pérdida completa del paquete o paquetes enviados a consecuencia de la congestión y retardos en los nodos de la red. Para prevenir y/o mitigar esta degradación durante la transmisión, en esta tesis se desarrollarán técnicas que hacen más robusto al codec

frente a errores en el canal y mejorar así, la calidad de la voz recuperada.

Por un lado, para las transmisiones sobre redes de ámbito local, donde la comunicación de voz generalmente se realiza empleando la tecnología de telefonía digital inalámbrica o *Digital Enhanced Cordless Telephony* (DECT), se estudiará la degradación producida por el efecto multitrayecto. Como consecuencia de los obstáculos que hay entre emisor y receptor, el receptor puede recibir varias copias de la señal emitida a consecuencia de la reflexión de la onda portadora sobre los diferentes obstáculos. Este hecho provocará una serie de desvanecimientos que pueden modificar la codificación original del paquete enviado. Para mitigar esta degradación, en esta tesis se plantea el uso de la técnica de decodificación por decisiones soft o *soft-decision decoding*, con la que obtener una estimación de la componente del paquete modificado. Para ello, en la estimación se tendrá en cuenta tanto la componente recibida en el paquete como la probabilidad a posteriori obtenida a partir del comportamiento del canal.

Por otro lado, para las transmisiones sobre redes basadas en el protocolo IP, se estudiará la degradación producida a consecuencia de la pérdida de paquetes. Esta degradación se produce como consecuencia de la congestión de los nodos de la red, dando lugar a una o varias pérdidas de manera consecutiva durante la transmisión. Las técnicas empleadas en la bibliografía para reducir este tipo de degradación pueden dividirse en dos clases dependiendo de si actúan antes de enviar el paquete (basadas en el emisor) o durante la el proceso de decodificación (basadas en el receptor).

Entre las técnicas basadas en el emisor, se hará uso de códigos de corrección de error hacia delante o *Forward Error Correction* (FEC) con los que cada paquete incluye información redundante de paquetes anteriores, a una codificación inferior, y que se utilizan para recuperar uno o varios paquetes perdidos durante la transmisión. Ahora bien, aunque su aplicación conllevará una ventaja notable en la calidad perceptual obtenida, también generará un incremento en la tasa de bits final que podría no ser soportado por el ancho de banda en canales limitados. Además, este cambio en el tamaño del paquete a enviar, también conlleva una incompatibilidad para poder utilizar el codec original aunque no se produzcan pérdidas durante la transmisión. Para solventar ambos inconvenientes del uso de los códigos FEC, en esta tesis se propone el uso de una técnica esteganográfica, particularizada al codec AMR. Como resultado, el código FEC oculto en el propio paquete mantiene la compatibilidad con el codec estándar y al mismo tiempo no reducirá significativamente

la calidad perceptual en transmisiones sin pérdidas en el canal respecto al codec original.

En cuanto a las técnicas basadas en el receptor, se proporcionarán diferentes esquemas de mitigación que tratan de recuperar los paquetes perdidos en una transmisión. Aunque los codecs de voz actuales tienen algoritmos para mitigar estas pérdidas, cuando la ráfaga de pérdidas es demasiado larga, estos algoritmos aplican un proceso de apagado para evitar generar artefactos en la reproducción de la voz. Para mitigar estas pérdidas, los diferentes esquemas propuestos en esta tesis tratan de proporcionar una estimación de los parámetros de voz necesarios para su reconstrucción. Estas estimaciones se obtienen a partir de unos vectores de sustitución, previamente calculados empleando un proceso de estimación de mínimo error cuadrático medio o *Minimum Mean Square Error* (MMSE), y considerando la evolución de los parámetros previos a la pérdida. Estos parámetros de voz se obtendrán de acuerdo al modelo de predicción lineal o *Linear Prediction Coding* (LPC), y para su estimación será necesaria la obtención de los correspondientes diccionarios de cuantización. Sin embargo, uno de estos parámetros, la señal de excitación, no presenta una representación adecuada para realizar estimaciones como sí ocurre con los coeficientes LPC. Por este motivo, en esta tesis también se abordará el problema de la representación y la generación de diccionarios de cuantización para obtener estimaciones de la señal de excitación eficaces.

Por un lado, dada la dificultad para obtener un diccionario de cuantización adecuado para la señal de excitación, se plantearán diferentes métodos de cuantización que modifican los procesos de centro y celda óptimos del conocido algoritmo de cuantización vectorial *Linde-Buzo-Gray* (LBG). Sin embargo, el alto coste en recursos para obtener diccionarios representativos hace que también se presenten varios esquemas de mitigación que tratan de mejorar la estimación de la señal de excitación. Uno de estos esquemas consiste en el uso del filtro adaptativo basado en corrección recursiva o *Recursive Least Squares* (RLS) para mejorar la estimación de las primeras tramas perdidas frente a las obtenidas por el vector de sustitución. El motivo es que el error de cuantización a la hora de obtener los índices de cuantización, puede provocar que las primeras estimaciones del vector de sustitución pudieran ser peores que las propuestas por el propio algoritmo de mitigación de errores o *Packet Loss Concealment* (PLC). La nueva propuesta permitirá aprovechar mejor la correlación con las tramas precedentes a la pérdida de manera similar a un filtro LTP en los codecs basados en el paradigma CELP.

Por otro lado, se planteará una nueva representación de la señal de excitación basada en la transformada wavelet Haar. Esta transformada wavelet permite dividir una señal en componentes de menor tamaño (la mitad) que, tras ser aplicada sucesivamente, puede obtener una descomposición en árbol balanceado o no balanceado. La ventaja de esta nueva representación es que se puede mejorar la calidad de los diccionarios, ya que se reduce el error de cuantización con particiones más pequeñas. Al mismo tiempo, también permitirá realizar una minimización de error sobre cada partición de manera independiente y mejorar así la calidad de la voz recuperada. Por último, se planteará un esquema que combina las técnicas de mitigación y prevención de errores con el objetivo de proporcionar un esquema robusto frente a errores y que permita aprovechar las ventajas de ambos enfoques: La recuperación de los paquetes perdidos y evitar el error de propagación que se genera en codecs que tienen una dependencia inter-trama durante el proceso de síntesis como es el caso del codec AMR.

Para finalizar, cabe indicar que las técnicas propuestas han sido evaluadas sobre varios codecs empleados en las transmisiones de ámbito local con el estándar DECT (G.726 y G.722) y sobre redes IP (AMR e iLBC), algunos de los cuales incluyen su propio algoritmo PLC. Para el proceso de entrenamiento y test se han utilizado las bases de datos de voz TIMIT y NTT, para el desarrollo del test de calidad objetivo (PESQ), y la base de datos Albayzin, para el test subjetivo (MUSHRA). Para analizar el rendimiento de las técnicas propuestas, se ha simulado el comportamiento del efecto multitrayecto y la pérdida de paquetes con un modelo de canal donde se aplica diferente relación señal ruido (SNR), en el caso de redes de ámbito local, y diferentes tasas de pérdidas y longitudes promedio de ráfaga, en el caso de las redes basadas en el protocolo IP. En estas pruebas se ha podido observar el incremento de calidad perceptual que ofrecen las técnicas propuestas en esta tesis frente a la obtenida por el algoritmo PLC del codec utilizado en las pruebas. De este modo, las técnicas propuestas en esta tesis han demostrado ser más robustas frente a la degradación producida en el canal y mejorar así la calidad de servicio de las transmisiones de voz en tiempo real.

Abstract

In the past few years there has been a great development and deployment of communication technologies that has enabled an ubiquitous and permanent connection to the Internet and the global telephone networks. This has supposed an increase in the number of applications and services that have completely changed the way we usually communicate, transmit and exchange ideas, feelings or information. In fact, the current transmissions are not longer differentiated from speech or data (text, video, images, ...), because the current applications use speech and data simultaneously (such as the case of videoconferences). Indeed, the great expansion and development of Internet, based on the IP protocol, has supposed a convergence of traditional speech communication networks and data communication networks to the latter. As a result, this convergence eases the access of new applications to the final user and also its scalability.

However, in order to provide a good quality of service, it is necessary that the speech communication is carried out without degradations and as a real time service. However, this quality is conditioned by the fact that digital networks are not free of errors during transmission. These errors are considered as alterations or lost packets, due to the network conditions and the environment where the transmission is performed.

The concern of this thesis is focused on the study of the degradation generated during transmissions on two types of networks with different scope: the local wireless networks and the IP networks. On the one hand, in local wireless networks, the degradation causes an alteration in the sent packets which is caused by the multipath effect. On the other hand, in the IP networks, the degradation causes the complete loss of the packet or packets, which is caused by congestion and delays in the network during transmission. In order to prevent and / or mitigate this degradation during transmission, this thesis will develop several techniques in order to make the speech codec more robust against errors in the channel and to improve the quality of the

recovered speech.

On the one hand, for transmissions over local networks, where speech communications are usually performed by using the Digital Enhanced Cordless Telephony (DECT) standard, the degradation generated by the multipath effect will be studied. As a result of the obstacles between sender and receiver, the receiver can get several copies of the sent signal because of the reflection and refraction of the carrier wave on different objects. This fact will cause faddings that could modify the original encoding of the packet. In order to mitigate this degradation, it is proposed in this thesis the use of the *soft-decision decoding* technique which obtains an estimate of the modified component in the packet. To achieve this estimation, the technique will take into account both the component received in the packet and the a posteriori probability obtained from the behavior of the channel.

On the other hand, for transmissions over networks which are based on the IP protocol, the degradation produced by lost packets will be studied. This degradation occurs as a consequence of the congestion in the nodes of the network which could lead to one or consecutive lost packets, as a burst, during the transmission. In the bibliography, the techniques which are used to conceal lost packets are split into two groups depending on whether they act before sending the packet (sender-driven techniques) or during the decoding process (receiver-based tecnhiques).

From the classical sender-driven techniques, the Forward Error Correction (FEC) technique will be selected in order to retrieve losts packets during transmission. However, although its application will lead to a noticeable advantage in the perceptual speech quality, it will also generate an increment of the final bitrate that might not be supported by the available bandwidth in limited channels. In addition, the FEC code suposse a change in the packet's size and generates an incompatibility to use the original codec even if there are not losses during transmission. In order to overcome both of the drawbacks of FEC codes, this thesis proposes the use of a steganographic technique, particularized to the AMR codec. As a result, the FEC code is embedded into the packet and it keeps compatibility with the standard codec. Moreover, this technique achieves identical performance to that offered by the legacy codec in clean channel conditions.

In terms of the receiver-based techniques, different mitigation schemes will be provided to recover lost packets in a transmission. Although current speech codecs have implemmented algorithms to mitigate these losses, when the burst is too long, these algorithms apply a mutting process to avoid artifacts during speech reproduc-

tion. In order to mitigate these losses, in this thesis the different error mitigation schemes try to provide an estimate fo the speech parameters which are necessary for its reconstruction. These estimates are obtained from replacement super vectors, previously calculated by using a minimum mean squares error (MMSE) estimation process, which are obtained from the last correctly received parameters. These speech parameters will be obtained according to the linear prediction coding model (LPC) but the success of this proposal deeply relies on the quality of the corresponding quantization dictionary. However, one of these parameters, the excitation signal, does not present a suitable representation for estimation in comparison with the LPC coefficients. For this reason, this thesis will also solve the problem of the representation and generation of quantization dictionaries to obtain efficient excitation signal estimates.

On the one hand, given the difficulty to obtain a suitable dictionary which minimize the quantizaion error of the excitation signal, different quantization methods will be proposed which modify the optimum center and cell criteria of the well-known vector quantization algorithm Linde-Buzo-Gray (LBG). However, due to the high costs in resources for obtaining representative dictionaries, in this thesis several mitigation schemes are presented which improves the excitation signal estimation. In that way, the Recursive Least Squares (RLS) adaptive filter will improve the excitation signal estimation in the first lost frames instead of using the corresponding estimates from the selected replacement supervector. As a consequence that the estimations and the selected replacement super vector are affected by the quantization error, these estimates could be worst than provided by the Packet Loss Concealment (PLC) algorithm. The new proposal will use the correlation with the previous correctly received frames similar to the Long Term Prediction (LTP) filter works on speech codecs based on the CELP paradigm.

On the other hand, a new representation of the excitation signal, based on the Haar wavelet transform, will be proposed. This wavelet transform allows to split a signal into smaller components (half size) that, after being applied successively, can obtain a balanced or unbalanced tree decomposition. An advantage of this new representation is that it is possible to minimize the quantization error with smaller partitions. Moreover, this proposal allows to minimize the synthesis error independently over each partition so a better speech reconstruction is achieved. Finally, a scheme which combines mitigation and error prevention techniques will be developed with the objective of providing a robust scheme against errors and allowing to take

advantage of both approaches: Packet loss recovering and avoiding the error propagation which is generated in codecs that have an inter-frame dependency during the synthesis process such as the AMR codec.

Finally, it should be mentioned that the proposed techniques have been evaluated on several codecs over WLAN networks by using the G.726 and G.722 speech codecs which are mandatory on the DECT standard, and over IP networks by using the AMR and iLBC speech codecs. Some of these codecs implement its own Packet Loss Concealment algorithm. For the training and testing process, the TIMIT and NTT speech databases were used to performe the objective quality test (PESQ) and the Albayzin database to perform the subjective test (MUSHRA). In order to analyze the performance of the proposed techniques, the multipath effect and the packet losses are simulated with a channel model where a different signal-to-noise ratio (SNR) are considered in the case of local area networks and a different packet loss rate and average burst length are considered in the case of IP networks. In these tests it is shown the noticeable improvement of our proposals against the original codec with its own PLC algorithm. Thus, the proposed techniques are more robust against errors during the transmission in a real time service.

Capítulo 1

Introducción

Con el objetivo de presentar un contexto de desarrollo de la presente tesis doctoral, en este capítulo se presentará la motivación de la misma, así como un listado de objetivos a cubrir y cómo está desarrollada esta tesis en los diferentes capítulos que la componen.

1.1. Motivación

En los últimos años se ha producido un gran desarrollo y despliegue de las tecnologías y aplicaciones que han posibilitado una conexión ubicua y permanente a la información en cualquier lugar del mundo. Unido a este desarrollo, la sociedad ha modificado por completo la forma en que se comunica, transmite e intercambia información a través de la red. Este hecho ha provocado que las comunicaciones digitales hayan incrementado tanto su capacidad como su rendimiento ofreciendo una comunicación donde se mezclen voz, imágenes, video o texto como es el caso de las videoconferencias con Skype.

Todo este avance tecnológico tiene un punto de partida, la necesidad del ser humano para comunicarse a largas distancias. Partiendo de que la forma más natural y sencilla de comunicarse con otras personas es la voz, a lo largo de la historia se han desarrollado diversos métodos de comunicación, desde las señales de humo, pasando por el telégrafo hasta las actuales comunicaciones por medio de la telefonía móvil para lograrlo de manera eficaz. Además, dado que la sociedad actual necesita del acceso a la información en cualquier lugar del mundo y en cualquier momento, las distintas tecnologías de acceso, tanto para transmisiones de voz como de datos, han mejorado sus prestaciones y cada día aparecen nuevos servicios y aplicaciones.

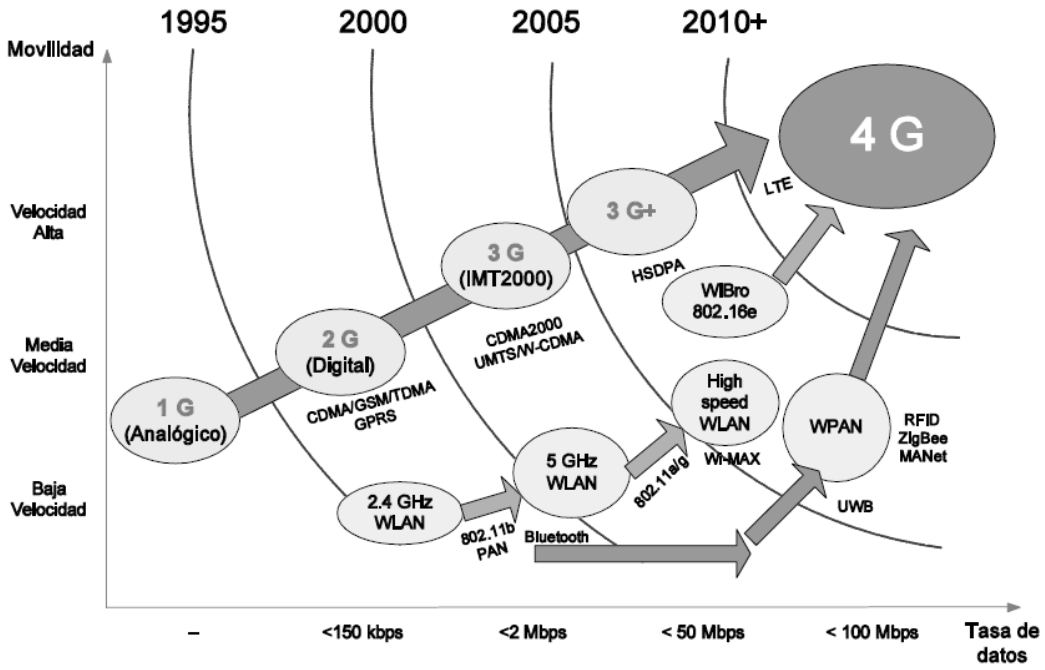

Figura 1.1: Diagrama que muestra la convergencia "todo IP"de las diferentes tecnologías hacia la cuarta generación (4G).

Hoy en día, Internet se ha convertido en una red de redes global gracias a su capacidad para conectar redes heterogéneas debido al protocolo IP o *Internet Protocol* [3]. Dada la necesidad de transmitir tanto voz como texto o video al mismo tiempo, en los últimos años se está produciendo una integración de las redes de datos y voz en las redes IP, conocida también como convergencia "todo IP"(ver figura 1.1). El resultado es que se está permitiendo que tecnologías que inicialmente eran competidoras como *WiMAX* (IEEE 802.16) o la tecnología 4G (comercialmente conocida también como *Long Term Evolution* (LTE)), puedan ser complementarias y su acceso transparente para el usuario final a la vez que se facilita la escalabilidad de las aplicaciones y servicios desarrollados. Uno de los servicios que más se ha expandido gracias a esta integración es el servicio VoIP, *Voice over IP*, ya que permite una comunicación más eficiente y económica que los sistemas de telefonía tradicionales. Prueba de ello es el incremento de suscripciones VoIP que a fecha de junio de 2015 ya se encontraba por encima de los 7.2 billones, superando así el número de personas vivas en todo el mundo [4].

Sin embargo, aunque los actuales algoritmos de codificación-decodificación de la voz, popularmente conocidos como *codecs*, permiten obtener una voz más inteligible y natural en el receptor, no hay que olvidar que durante el proceso de transmisión de

la voz, dividida en paquetes, la calidad de voz recuperada puede verse reducida a consecuencia de los errores producidos en el canal. En esta tesis, estos errores supondrán una degradación que puede ser de dos tipos: la modificación de la codificación en el paquete recibido o la pérdida del paquete enviado. El primer tipo de degradación es más característico en las redes inalámbricas de ámbito local o *Wireless Local Area Network* (WLAN), como consecuencia de la presencia de objetos cercanos entre el dispositivo emisor y receptor. Estos obstáculos provocan que la señal recibida tenga diferente fuente y se generen desvanecimientos que provocan una modificación en la codificación interna del paquete recibido. El segundo tipo de degradación se debe a que las redes IP no están desarrolladas para ser confiables, es decir, no pueden garantizar la correcta recepción del paquete para transmisiones en tiempo real [3]. Esto se debe a la congestión y retardos en la red pueden provocar que el paquete no llegue a tiempo al receptor. Además, esta congestión puede provocar que sean varios paquetes los afectados, dando lugar a una ráfaga de pérdidas consecutivas. Como consecuencia de la degradación producida, la señal de voz obtenida en el receptor es diferente a la original enviada, con lo que se produce una caída en la calidad del servicio [5].

Tradicionalmente, el esfuerzo investigador se ha centrado en el segundo tipo de degradación mediante el desarrollo de algoritmos de mitigación de paquetes perdidos o *Packet Loss Concealment* (PLC), que aprovechan las características de la voz para reducir el impacto de la degradación. Sin embargo, dado que la señal de voz no es estacionaria en el tiempo, su rendimiento decae rápidamente a medida que la longitud de ráfaga es mayor. Además, en los codecs que presenten una dependencia inter-paquete se puede producir una propagación del error hacia los paquetes recibidos correctamente tras la pérdida.

Por este motivo, esta tesis se centrará en analizar y mitigar ambos tipos de degradación proporcionando nuevas técnicas de prevención y mitigación y obtener así, codecs más robustos frente a transmisiones con errores en el canal.

1.2. Objetivos

Los objetivos a cubrir en esta tesis se pueden desglosar como sigue:

- Mitigar el efecto multitrayecto en transmisiones sobre redes inalámbricas de ámbito local.

- Proporcionar nuevos métodos de cuantización y representación para la señal de excitación que faciliten el proceso de estimación para las pérdidas de paquetes sobre redes IP.

- Presentar diferentes técnicas de mitigación de pérdidas sobre redes IP. Se centrarán en la mitigación para pérdidas de paquetes de forma consecutiva y en la mejora de la estimación de la señal de excitación.

- Mitigar el efecto de la propagación del error tras una ráfaga de paquetes perdidos con técnicas de prevención de pérdida de paquetes sobre redes IP. Además, las técnicas propuestas mantendrán la compatibilidad con el modo de funcionamiento en el codec estándar.

1.3. Estructura de la tesis

La presente tesis doctoral consta de siete capítulos. En el Capítulo 2 se presentará el proceso de generación de la voz y cómo se ha codificado ésta a lo largo de los diferentes métodos de codificación que han ido apareciendo recientemente. También se hará una breve descripción de los codecs empleados en esta tesis.

En el Capítulo 3, se describirán las características de los dos tipos de redes estudiadas en esta tesis, así como las características del canal que genera las degradaciones sobre los paquetes enviados. También se presentará el estado del arte de las técnicas que se han empleado en la bibliografía para mitigación de pérdidas de paquetes perdidos tanto desde el punto de vista del emisor como en el receptor [6]. Por último, se presentará el entorno de pruebas o *framework* utilizado para la obtención de los resultados experimentales de esta tesis.

Los Capítulos 4, 5 y 6 describirán las técnicas que se han desarrollado con el fin de minimizar el impacto de la degradación producida durante la transmisión de la voz sobre redes inalámbricas de ámbito local (WLAN) y redes IP. Así, el Capítulo 4 se centrará en mitigar el efecto de la degradación por el efecto multitrayecto en comunicaciones sobre redes WLAN. Para ello, se presentará una técnica de mitigación, basada en el receptor, donde se realizará una estimación de las componentes codificadas en cada paquete de acuerdo al conocimiento previo del comportamiento del canal, en este caso la tasa de error por bit o *Bit Error Rate* (BER) por paquete, y el paquete recibido.

Los Capítulos 5 y 6, se centrarán en mitigar la degradación producida por la

pérdida del paquete en las redes IP. Para recuperar el paquete perdido, son necesarios una serie de parámetros, entre ellos la señal de excitación. Dado que la estimación de la señal de excitación es uno de los aspectos relevantes en esta tesis, en el Capítulo 5 se presentarán los diferentes métodos de cuantización desarrollados, orientados a la estimación de la señal de excitación, y en el Capítulo 6 se presentará una nueva representación basada en la transformada wavelet Haar. Estos métodos de cuantización y la nueva representación serán empleados en las técnicas de prevención y mitigación presentados en cada capítulo.

En el Capítulo 5, se aplicarán técnicas de prevención de paquetes perdidos, basadas en el emisor, con el objetivo de recuperar el último paquete perdido en una ráfaga y al mismo tiempo reducir la propagación del error en los codecs que tienen una dependencia inter-trama. Para ello, se hará uso de los códigos de corrección de errores hacia delante o *Forward Error Correction*(FEC). Este tipo de técnica presenta el inconveniente de generar un incremento en la tasa bits, por cada paquete a enviar en el modo de funcionamiento seleccionado, que lo hace incompatible con el codec original e incluso inviable para canales con un ancho de banda limitado. Por este motivo, se utilizará una técnica esteganográfica que oculta dicho código FEC en la propia codificación del paquete a enviar. Como resultado, se mantendrá la compatibilidad con el codec original y al mismo tiempo, no se generará una pérdida significativa de calidad perceptual en las transmisiones sin pérdidas en el canal.

En el Capítulo 6, se describirán diferentes esquemas de mitigación de paquetes perdidos en el receptor. En estos esquemas se hará uso de la técnica de vectores de sustitución con los que se irá proporcionando una estimación de los parámetros de voz en los paquetes perdidos en el canal. La diferencia en los esquemas propuestos radica en cómo se mejora la estimación de la señal de excitación, ya que ésta no dispone de una representación adecuada para realizar estimaciones.

Finalmente, el Capítulo 7 se dedica a presentar las conclusiones, contribuciones realizadas y líneas futuras de investigación de esta tesis.

Capítulo 2

La voz humana

Este capítulo se centrará en describir el proceso de generación de la voz humana para a continuación presentar algunos de los modelos de representación de la voz más extendidos y conocidos en la bibliografía. Del mismo modo, se describirán las características de los algoritmos de codificación de la voz de acuerdo a su clasificación por forma de onda, paramétricos o híbridos y se presentará una breve descripción de los codecs empleados en esta tesis.

2.1. Introducción

La señal de voz es una onda sonora producida a través del aparato fonador humano y que se caracteriza por su intensidad, amplitud, timbre, duración y volumen [7]. Estas características son las que nos permiten reconocer si quien habla es un hombre, una mujer o un niño, y la emoción con la que se quiere transmitir una idea o mensaje. Sin embargo, la gran tasa de bits necesaria para transmitir la señal de voz es lo que ha propiciado la investigación para mejorar la codificación y reconocimiento de la señal de voz.

Aunque las primeras transmisiones de voz a través del teléfono no aparecieran hasta finales del siglo XIX, de la mano de Alexander Graham Bell en 1876, estas transmisiones eran analógicas a la vez que inseguras dado que cualquiera podía acceder a ellas. Este aspecto fue determinante durante la segunda Guerra Mundial (1939-1945) para que las transmisiones de voz abandonasen el mundo analógico y pasaran al mundo digital. De esta manera, la primera transmisión de voz digital se produjo utilizando la técnica de modulación por impulsos codificados o *Pulse Code Modulation* (PCM), propuesta por Alec Reeves en 1937, sobre la máquina *SIGSALY*

diseñada en los laboratorios Bell [7].

Para poder digitalizar una señal, ésta debe ser procesada mediante las operaciones de muestreo, cuantización y codificación para transformar una señal analógica en una señal discreta digital que pueda ser transmitida por un canal digital [8]. Sin embargo, en este proceso se ha de tener en cuenta la representación de la voz y cómo se va a codificar finalmente esta señal. Así, aunque PCM permite la digitalización de la voz, ésta se hace muestra a muestra con lo que la tasa de bits necesaria para transmitirla es muy alta. Además, tampoco se tiene en cuenta si hay información redundante en la señal que pueda permitir una codificación más compacta.

Con el objetivo de transmitir la voz de una forma más eficaz, Homer Dudley desarrolló en 1939 el primer codificador-decodificador de voz, que denominó como vocoder (*VOice enCODER*), en los laboratorios Bell [7]. Este vocoder extraía una serie de características de la señal de voz que permitían recuperar una aproximación bastante cercana a la señal original y que al mismo tiempo reducía la tasa de bits necesaria para su transmisión. A partir de este vocoder, ha habido numerosos trabajos que presentan nuevos algoritmos de codificación-decodificación de voz, popularmente conocidos como *codecs*, que basan su funcionamiento en modelos de producción de voz y que permiten realizar una comunicación a una baja tasa de bits y una calidad perceptual razonable.

De este modo, a continuación se presenta el proceso de generación de la voz humana a nivel biológico para posteriormente presentar algunos de los modelos matemáticos que permiten representar su funcionamiento. También se presentan los principales algoritmos de codificación de la voz de acuerdo a su clasificación por forma de onda, paramétricos o híbridos. Finalmente, se realiza una breve descripción de los codecs empleados en esta tesis.

2.2. Proceso de generación de la voz

La voz humana es una señal sonora que producimos voluntariamente y que nos permite comunicarnos con otros individuos expresando ideas o emociones. Para la generación de la voz intervienen una serie de órganos y músculos que constituyen el sistema respiratorio (pulmones, bronquios y tráquea), el aparato fonador (laringe y cuerdas vocales) y el tracto vocal (faringe y cavidades oral y nasal) [1, 7]. Como resultado se obtiene una señal caracterizada por una intensidad, frecuencia, timbre, duración y volumen. A continuación se describe el funcionamiento de cada uno de

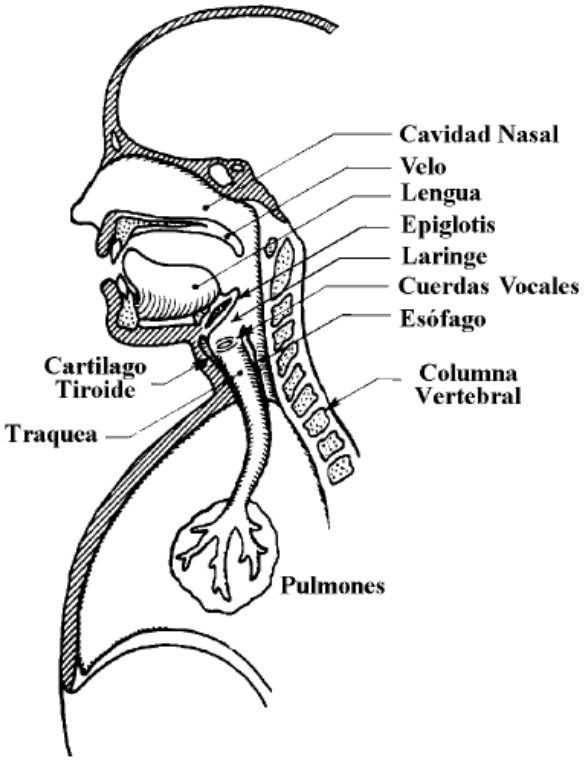

Figura 2.1: Representación e identificación de los órganos y músculos que participan en la generación de la voz humana. (Fuente:[1])

estos sistemas y que pueden verse en la figura 2.1:

- **El sistema respiratorio**: se encarga de proporcionar el aire para que en el aparato fonador se produzca una señal que pueda clasificarse como sonora o sorda. Además, el volumen de aire proporcionado al aparato fonador determinará la característica de duración de la señal de voz y que depende de la edad del individuo, la actividad física realizada (andar frente a correr) o de enfermedades respiratorias [1].

- **El aparato fonador**: es el más importante en la producción de la voz y se compone principalmente de la laringe, donde se encuentran las cuerdas vocales, y se encarga de modificar el flujo de aire procedente de los pulmones para generar un patrón periódico, en el caso de un sonido sonoro, o un patrón ruidoso, en el caso de un sonido sordo. Hay que tener en cuenta que el aparato fonador se va desarrollando con la edad del individuo y su funcionamiento se va aprendiendo a través de la relación que existe entre el proceso de generación de la voz y la audición, motivo por el cual los sordos de nacimiento no pue-

den hablar correctamente aunque no tengan ningún problema morfológico [1]. Este patrón periódico en los sonidos sonoros se genera al modificar el espacio entre los pliegues que conforman las cuerdas vocales (glotis) y cuyo periodo se denomina tono glotal o *pitch*. Esta frecuencia correspondiente al *pitch* difiere ligeramente dependiendo de si se trata de un hombre (aproximadamente 120 Hz), una mujer (aproximadamente 250 Hz) o un niño (aproximadamente 350 Hz) [1]. Por lo tanto, el aparato fonador será el responsable de la característica de la frecuencia con la que se emite la voz, que puede oscilar de 60 a 7000 Hz, y la amplitud o volumen de la voz. Esta variación en el volumen de la voz (grito frente a susurro) dependerá del flujo de aire que llega al tracto vocal según la abertura de la glotis.

- **El tracto vocal**: está compuesto por la faringe y las cavidades nasal y vocal. Su funcionalidad es la que determinará la forma de onda resultante para el sonido que se emite al exterior. Aunque la vibración de las cuerdas vocales constituye la fuente principal para generar la voz, el tracto vocal actúa como un filtro y resonador del tren de pulsos generado en la laringe y que dará lugar al resto de características de la voz. Dado que cada individuo tiene un tracto vocal diferente por su morfología de nacimiento, como consecuencia de su desarrollo o enfermedades posteriores, la voz resultante es única [1]. Este hecho hace posible identificar una persona frente a otra con sólo oír su voz, sin embargo, también complicará la obtención de un modelo determinista para la producción de la voz.

Una vez presentados los tres sistemas implicados en el proceso de generación de la señal voz, éste proceso debe modelarse de manera que permita obtener una serie de parámetros que se utilicen para realizar una transmisión y cuyo resultado final sea una señal de voz inteligible y natural en el receptor.

2.3. Representación de la voz

A la vista de la descripción indicada en la sección anterior, los sistemas para la producción de la voz presentan el problema de que son únicos para cada individuo, por tanto, no es posible desarrollar un sistema determinista que modele de manera general la voz de un hombre, una mujer o un niño. Por este motivo, es necesario obtener una adecuada representación que permita sintetizar una señal de voz lo más

cercana posible a la señal de voz original. En los siguientes apartados se describen los procesos previos al procesamiento de la voz y la representación escogida en esta tesis.

2.3.1. Conversión analógico/digital de la voz

Dada la naturaleza analógica de la señal de voz, ésta requiere de un procesamiento previo para obtener una señal de voz digital mediante la conversión analógico-digital o conversión AD. Para obtener una señal digital son necesarios tres procesos: muestreo, cuantización y codificación [8].

- **Muestreo**: Consiste en tomar valores en el tiempo de la señal analógica según un periodo T, obtenido de acuerdo al teorema de Nyquist y que garantiza que no se pierde información.
- **Cuantización**: Consiste en discretizar la amplitud de la señal muestreada, es decir, se limita el número de valores que puede tomar cada muestra de la señal dentro de un conjunto discreto, denominado como diccionario. El tamaño y valores de este diccionario es lo que determinará que el error de cuantización entre la muestra original y la cuantizada sea mayor o menor.
- **Codificación**: Consiste en asignar una códificación binaria a cada muestra cuantizada. Esta se corresponde con la codificación asignada a cada elemento del diccionario.

Este sería el proceso general para obtener una señal digital. En las siguientes secciones se particularizarán los procesos de acuerdo a las características de la señal de voz.

2.3.2. Preprocesado de la señal de voz

Partiendo de la señal de voz representada como muestras discretas, obtenidas a una frecuencia de 8 KHz, dado que la señal de voz tiene un rango de frecuencias de 250 a 4000 Hz, un paso previo a su análisis es el proceso de preénfasis. Este proceso consiste en tomar la señal de voz y aplanarla espectralmente para compensar la caída de 6 dB/década que típicamente experimenta la señal de voz [9]. Este preénfasis se realiza con un filtro de respuesta finita definido como:

$$P(z) = 1 - \mu z^{-1} \tag{2.1}$$

donde $\mu \leq 1$ es un factor real. Además, dado que este filtro aproxima la derivada de la señal, también permite eliminar la componente continua de la señal de voz.

Por otro lado, dado que la señal de voz es una señal no estacionaria, hay que dividirla en segmentos cortos o tramas de 10-40 ms que permitan realizar la suposición de estacionariedad. Tras realizar esta división en segmentos o tramas (típicamente de 20 ms), se procederá al análisis espectral en el que se extraerán los parámetros que definen la señal de voz en cada trama. Cada trama de N muestras se obtiene considerando cierto solapamiento entre tramas de acuerdo a una ventana $w(n)$ diferente a la rectangular y que reduce el fenómeno de fuga o *leakage* [8]. Para este fin, la ventana más empleada en la bibliografía ha sido la ventana de Hamming que se define como:

$$w(n) = 0{,}54 - 0{,}46\cos\left(2\pi\frac{n-1}{N}\right) \quad 0 \leq n < N \tag{2.2}$$

Como consecuencia de este enventanado, la voz puede considerarse estacionaria en cada trama permitiendo así tratar cada una de manera independiente y poder realizar el análisis de predicción lineal que se describe a continuación.

2.3.3. Análisis de predicción lineal

En la bibliografía, el modelo más conocido y extendido para el análisis y síntesis de voz es el modelo de codificación por predicción lineal o *Linear Prediction Coding* (LPC) [10] por su sencillez y por la alta calidad de la señal de voz obtenida tras el proceso de síntesis.

La idea de predicción lineal fue acuñada por Norbert Wiener, en la década de 1940, y ha sido utilizada en multitud de aplicaciones bajo diferentes formulaciones. En el caso del procesamiento de la voz, la predicción lineal fue utilizada por primera vez por Saito e Itakura en [11], pero se ha popularizó a raiz del trabajo de Atal y Schroeder en [12]. Este trabajo estableció la base del paradigma de predicción lineal por código excitado o *Coded Excited Linear Prediction* (CELP) [13], que está presente en la mayoría de los codecs actuales.

El modelo de predicción lineal (LPC) permite estimar la señal sintetizada $s(n)$ como una combinación lineal de muestras anteriores y de la señal de excitación $e(n)$ como:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + \sigma \sum_{l=0}^{q} b_l e(n-l) \tag{2.3}$$

donde p y q son ordenes de predicción, σ es la ganancia y a_k y b_l son los coeficien-

tes que representan el filtro que representa el tracto vocal (se corresponden con la cavidad oral y nasal respectivamente).

Esta señal sintetizada $s(n)$ puede expresarse en el dominio de la transformada Z como $S(z) = H(z) \cdot E(z)$ donde $H(z)$ es un filtro con polos y ceros que trata de modelar el tracto vocal en el que interviene tanto la cavidad nasal como oral (ver figura 2.1). Este filtro $H(z)$ se obtiene a partir de la expresión (2.3) en el dominio Z como:

$$H(z) = \sigma \frac{1 + \sum_{l=1}^{q} b_k z^{-l}}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{2.4}$$

Sin embargo, si se ignoran los ceros producidos por la cavidad nasal, la expresión (2.3) se puede simplificar hacia un modelo todo-polos, que es más estable, definido como:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + \sigma e(n) \tag{2.5}$$

a partir de la cual se obtiene el filtro $H(z)$ todo-polos siguiente:

$$H(z) = \frac{\sigma}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{\sigma}{1 + A(z)} \tag{2.6}$$

Partiendo de la expresión (2.5) se puede apreciar que, bajo el modelo de predicción lineal, la señal de voz, $s(n)$, depende de una señal de excitación $e(n)$, que determina si la señal va a ser sonora o sorda, y del comportamiento del tracto vocal, definido por los coeficientes a_k para un determinado orden de predicción p. Este orden de predicción determinará el número de polos que conforma el filtro $H(z)$ y se encarga de definir la forma de la envolvente espectral de la señal $s(n)$ para una trama determinada [14].

Para determinar el orden de predicción correcto se aplica la aproximación de que existe una resonancia por cada 1000 Hz [14]. Así para un ancho de banda de 4 kHz, que es típico para el muestreo de voz a 8 kHz, se tendría un orden de predicción de $p = 8$ (4 pares de polos complejos conjugados) aunque se suelen agregar algunos más. Típicamente, este orden de predicción está fijado en $p = 10$ para la mayoría de codecs.

Una vez que está determinado el orden de predicción p hay que estimar los coeficientes a_k mediante una minimización del error cuadrático. Tradicionalmente, la obtención de estos coeficientes se ha hecho siguiendo el modelo de predicción lineal basado en las muestras anteriores pero también está la versión que considera las muestras futuras [15] aunque es menos utilizada.

Siguiendo el modelo basado en las muestras anteriores, los coeficientes a_k para el filtro, conocidos también como coeficientes LPC, pueden calcularse minimizando el error ϵ definido como:

$$\epsilon = \sum_{n=0}^{N} \big(s(n) - \sum_{k=1}^{p} a_k s(n-k)\big)^2 \tag{2.7}$$

Realizando la derivada respecto a los coeficientes a_k e igualando a cero se obtiene el siguiente sistema de ecuaciones:

$$\sum_{k=1}^{p} a_k \mathbf{R}_s(|i-k|) = \mathbf{R}_s(i) \qquad 1 \leq i \leq p \tag{2.8}$$

donde $\mathbf{R}_s$ representa la autocorrelación de la señal $s(n)$.

Así, dada una señal finita, este sistema de ecuaciones puede resolverse a través de dos métodos bien conocidos basados en la autocorrelación y en la covarianza [14]. El método basado en la autocorrelación, que aplica el algoritmo Levison-Durbin [10] para resolver la matriz $\mathbf{R}_s$ de tipo Toeplitz, es el más utilizado en la bibliografía ya que permite asegurar la estabilidad del filtro todo-polos y requiere menor esfuerzo computacional [16].

Finalmente, tanto la señal de excitación como los coeficientes LPC tienen que ser codificados para ser transmitidos. En la siguiente sección se explican los métodos de codificación más relevantes en referencia a la calidad perceptual y la tasa de bits empleada.

2.4. Codificación de la voz

Aunque el modelo de predicción lineal es el más utilizado en la implementación de los codecs actuales, tanto los coeficientes LPC como la señal de excitación tienen que ser codificados de manera que se pueda obtener una señal de voz lo más aproximada posible a la señal original. Sin embargo, dado que no se dispone de un ancho de banda ilimitado para transmitir estos parámetros, es necesario codificar la señal y dependiendo de la tasa de bits empleada, así será la caída en la calidad perceptual de la señal de voz obtenida. Por lo tanto, el objetivo que se persigue en el proceso de codificación es que se pueda mantener un cierto nivel de calidad cuando se codifica una señal de voz pero que al mismo tiempo se emplee la menor tasa de bits posible.

Históricamente, los primeros codecs se basaron en codificar de manera eficiente la forma de onda de la señal de voz de entrada, haciendo uso de las particularidades de la voz y las limitaciones del oído humano para las transmisiones a través del teléfono. Por este motivo, el ancho de banda utilizado para los codificadores de voz se centraba en el rango de 300 a 3400 Hz, que es lo que se considera como banda estrecha o *narrowband*. Aunque la mayoría de los codecs utilizados sobre las arquitecturas actuales de VoIP siguen utilizando esta banda de frecuencias para mantener la compatibilidad con las redes de telefonía fija, la exigencia cada vez mayor de aportar naturalidad e inteligibilidad a la señal de voz ha hecho que los codecs cubran un espectro más amplio. Así, actualmente es posible codificar señales de audio en el rango completo de audición (20 Hz a 20 kHz). De este modo han aparecido codecs capaces de trabajar en rangos de 50 Hz a 7 kHz (banda ancha o *wideband*), de 50 Hz a 14 kHz (banda súper ancha o *superwideband*) o cubriendo todo el rango de audición (banda completa o *fullband*).

A lo largo de las últimas décadas se han desarrollado numerosos esquemas de codificación de la señal de voz que se pueden clasificar en [17]: codificadores de forma de onda, codificadores paramétricos o *vocoders* y codificadores híbridos. A continuación se detallan las principales características de los distintos esquemas de codificación, atendiendo principalmente a la tasa de bits y la calidad de voz sintetizada. Además, dentro de cada tipología se presentan los principales estándares de codificación de voz más populares que lo han utilizado.

2.4.1. Codificadores de forma de onda

Los codificadores de forma de onda se caracterizan por realizar un procesamiento muestra a muestra de la señal de voz. El objetivo es conseguir una señal de voz recuperada que se aproxime lo máximo posible a la señal original y sin emplear ningún modelo de representación de la voz. La calidad perceptual depende de la razón de transmisión por muestra, por este motivo se requiere de una alta tasa de bits para obtener una señal de voz de buena calidad. Dentro de este tipo de codificadores podemos encontrar los siguientes técnicas de codificación:

Modulación por impulsos codificados

En la introducción ya hicimos mención a esta técnica de codificación (PCM) utilizada durante la segunda Guerra Mundial. Sin embargo no sería hasta 1972

cuando sería incorporado en el estándar G.711 [18] tras finalizar los derechos de patente a finales de los años 60. Este codec emplea una cuantización no uniforme de cada muestra de la señal de acuerdo con un cuantizador logarítmico que sigue la ley μ en EE.UU. y la ley A en Europa [8]. Cada muestra se codifica con 8 bits y dado que la frecuencia de muestreo de la voz es típicamente de 8000 muestras por segundo, se tiene como resultado una tasa de bits de 64 kbps. Bajo esta codificación se obtiene una señal de voz de buena calidad que será comparada con la obtenida por otros procedimientos de codificación con diferente tasa de bits.

Modulación adaptativa por impulsos codificados

Una mejora para la codificación PCM consiste en considerar que los valores de las muestras no se encuentran repartidos de manera equiprobable en los diferentes niveles de cuantización. Además, el rango dinámico en los valores que puede tomar una muestra provoca que esta situación se agrave y que la cuantización realizada no ayude a recuperar una señal de voz con buena calidad. Para mejorar el rendimiento de la codificación PCM, se plantea el uso de un cuantizador adaptativo para obtener una codificación PCM adaptativa o *Adaptive PCM* (APCM) [19] que ayuda a cuantizar mejor las muestras de la señal de voz.

Modulación diferencial por impulsos codificados

Dado que el codificador PCM genera una tasa de bits elevada, ésta puede reducirse aplicando técnicas predictivas. La técnica de codificación diferencial de impulsos o *Differential Pulse Code Modulation* (DPCM) [20] permite reducir esta tasa de bits al cuantizar una señal compuesta por la diferencia de cada muestra con la precedente. Gracias a esta señal diferencia, se permite explotar la redundancia de la señal a corto plazo.

Modulación diferencial adaptativa por impulsos codificados

Esta alternativa, denominada también como *Adaptive Differential Pulse Code Modulation* (ADPCM), es una variante a la anterior técnica DPCM. Esta técnica fue propuesta por Nikil Jayant en 1974 [20] para mejorar la compresión de PCM utilizando un cuantizador y un predictor adaptativos con 6 coeficientes de predicción. El esquema para el codificador y el decodificador se muestra en la figura 2.2.

A partir de este esquema se puede ver cómo se trata de codificar la diferencia

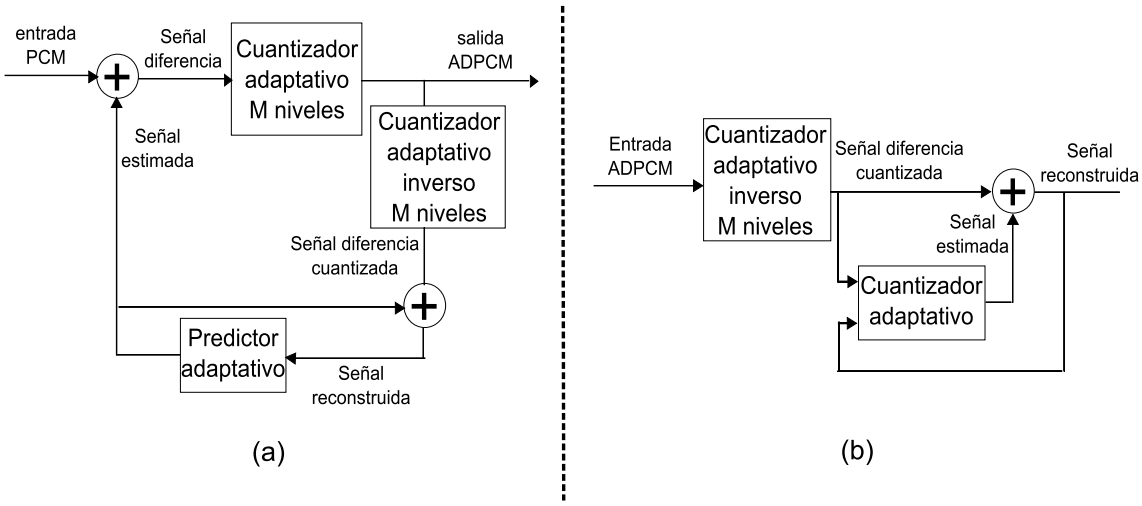

Figura 2.2: Esquema general para el codificador (a) y el decodificador (b) para la técnica de codificación ADPCM.

entre la señal original (señal de entrada PCM) y una predicción obtenida (señal estimada) a partir de la aplicación de los coeficientes de predicción sobre las muestras cuantizadas previas. Dado que el rango dinámico del error será inferior al de la codificación PCM, el resultado obtenido como señal reconstruida tiene una buena calidad de voz con una tasa de bits inferior a los 64 kbps. En este esquema de codificación están basados los estándares G.721 [21], G.722 [22], G.723 [23] o G.726 [24].

Modulación diferencial adaptativa por impulsos codificados en subbandas

Hasta ahora se ha realizado la codificación de la señal de voz sin considerar su espectro. Teniendo en cuenta que la mayor parte de la energía de la señal de voz se acumula en las frecuencias bajas (50-4000) Hz, esta técnica trata de codificar por separado la banda de frecuencias bajas y la banda de frecuencias altas con una codificación diferente por muestra pero manteniendo la misma tasa de bits final. Es decir, se busca aprovechar la codificación para la banda que tiene más información. Un ejemplo de esta codificación se puede apreciar en la figura 2.3, que se emplea en el estándar G.722 [22].

Como puede observarse, el codificador realiza la división del espectro de la señal en dos bandas que se obtienen a partir de los correspondientes filtrados paso-bajo y paso-alto utilizando un banco de filtros espejo en cuadratura o *Quadrature Mirror Filter* (QMF). Las salidas de estos filtros pasarán por un submuestreo y un

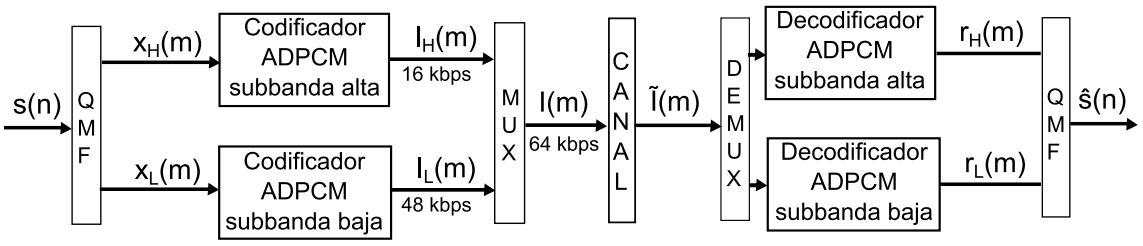

Figura 2.3: Esquema del funcionamiento general de la codificación ADPCM en subbandas en las etapas de codificación y decodificación.

codificador ADPCM con una razón de transmisión diferente a cada subbanda. Del mismo modo, la decodificación se obtiene siguiendo el proceso inverso antes comentado donde intervienen un proceso de sobremuestreo, decodificador ADPCM y el banco de filtros QMF comentado.

2.4.2. Codificadores paramétricos

A diferencia de los codificadores de forma de onda, en este caso ya no se codifica cada muestra de la señal de voz sino que se codifican una serie de parámetros para que tras realizar la síntesis, la señal reconstruida sea perceptualmente equivalente a la original. Con este fin, los codificadores paramétricos o *vocoders* determinan qué parámetros caracterizan los distintos segmentos de voz o tramas para ser enviados. Ahora bien, al no estar codificando muestras, la señal resultante de aplicar el proceso de síntesis tendrá una forma diferente a la original. No obstante, el objetivo de estos codificadores es que el oído humano la perciba como similar a la original. Es decir, perceptualmente la señal es similar a la original aunque se pierde cierta naturalidad. Por este motivo, la calidad es peor a la proporcionada por los codificadores de forma de onda.

Los parámetros necesarios para realizar esta síntesis se obtienen de acuerdo al modelo de producción de voz definido en 2.3.3 y que definen tanto la señal de excitación como el funcionamiento del tracto vocal. Dada la importancia para codificar estos parámetros, a continuación se detalla cómo se codifica cada uno de ellos bajo el modelo LPC.

Codificación de los coeficientes para el tracto vocal

Como se ha indicado anteriormente, el tracto vocal queda representado por un filtro todo-polos definido por los coeficientes a_k o coeficientes LPC. Estos coeficientes

están calculados de manera que el filtro resultante sea estable [16], sin embargo, la alta sensibilidad de los coeficientes LPC hacen que una cuantización directa sea inviable, ya que los correspondientes valores cuantizados podrían no garantizar la estabilidad del filtro. Por este motivo, se busca una representación más adecuada a la codificación de estos coeficientes.

En la bibliografía se pueden encontrar diversas representaciones para los coeficientes LPC como son: los *Reflection Coefficients* (RC) [25], los cocientes *Log Area Ratio* (LAR) [26], los *ArcSine Reflection Coefficients* (ASRC) [27], las *Line Spectral Frecuencies* (LSF) [28] o una variación de éstas últimas conocidas como los *Line Spectrum Pairs* (LSP) [29]. No obstante, la representación más utilizada para la codificación de los coeficientes LPC es la representación LSF, puesto que presentan menor distorsión tras realizar la cuantización [29, 30] y mantienen la estabilidad del filtro.

Codificación de la señal de excitación

En términos biológicos, la señal de excitación es el flujo de aire obtenido tras la glotis y que determina si el sonido resultante será sordo o sonoro. Así, si esta señal adopta un patrón de tren de pulsos (separados por un periodo o *pitch*) el sonido será sonoro o si por el contrario, presenta un patrón de ruido blanco, el sonido será sordo.

Bajo el modelo LPC, no hay una especificación sobre cómo se modela la señal de excitación sino que ésta se describe de acuerdo a una serie de parámetros para el segmento de voz en cuestión. En la mayoría de los codecs, los parámetros empleados para definir la señal de excitación son: la ganancia de la señal (G), información de si el segmento es sordo o sonoro y la estimación del periodo de *pitch* en el caso de ser sonoro. Con estos parámetros se puede obtener una señal de voz inteligible pero al no estar modelando la señal de excitación, ésta no será natural.

Otra alternativa de representar la señal de excitación sería la empleada en los codificadores sinusoidales [31]. Estos codificadores se basan en un modelo de producción de voz que considera la señal de voz como el resultado de pasar una señal de excitación generada por la glotis a través de un filtro lineal variante en el tiempo. Así, esta señal de excitación se obtiene como una suma de sinusoides, cuyas frecuencias y fases son modificadas en segmentos sucesivos para representar el carácter cambiante del espectro de la señal de voz original. Por lo tanto, los parámetros a codificar serán las fases y amplitudes utilizadas. No obstante, en la bibliografía el

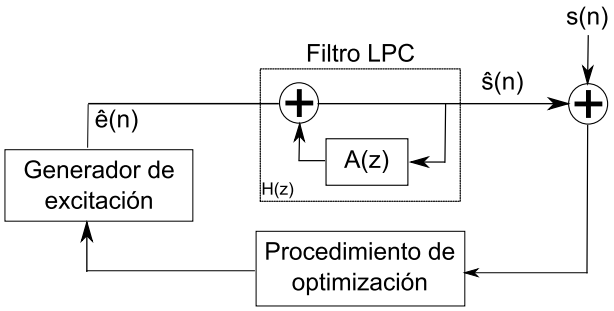

Figura 2.4: Esquema general de funcionamiento de un codificador híbrido basado en el proceso de análisis por síntesis.

modelo más utilizado es el modelo LPC.

Para finalizar, indicar que el codec más conocido entre los codificadores paramétricos es el estándar americano *FS1015* [32], también conocido como *LPC10*. Este codec divide la señal de voz en tramas de 22.5 ms para obtener una tasa de bits de 2.4 kbps e incluso 1.2 kbps para transmisiones por canales con ancho de banda muy limitado. Sin embargo, su baja tasa de bits lleva acompañada una baja calidad perceptual y esto ha hecho que su uso decaiga. Esta baja calidad perceptual es el principal motivo por el que se han ido abandonando los codificadores paramétricos y se haya centrado la investigación en los codificadores híbridos, ya que ofrecen un rendimiento mucho mayor con una baja tasa de bits.

2.4.3. Codificadores híbridos

A la vista del problema en cuanto a compromiso de calidad perceptual y tasa de bits empleada, los codificadores híbridos surgen como una unión de las ventajas de los codificadores anteriores. Por un lado, una mejora significativa en calidad perceptual muy cercana a la de los codificadores de forma de onda, y por otro lado, conseguir esta calidad con una tasa de bits reducida. Es decir, se pueden considerar como codificadores paramétricos atendiendo a que emplean un modelo paramétrico pero al mismo tiempo también se intenta preservar la forma de onda de la señal de voz recuperada.

Para ello, los codificadores híbridos emplean un procedimiento denominado análisis por síntesis, que puede verse en la figura 2.4, donde el objetivo es obtener aquellos parámetros que logran ajustar mejor la señal sintetizada $\hat{s}(n)$ respecto a la señal original $s(n)$. Esta señal sintetizada, $\hat{s}(n)$, se puede obtener de acuerdo con la expresión (2.5) y así el error a minimizar mediante un criterio de mínimo error cuadrático se

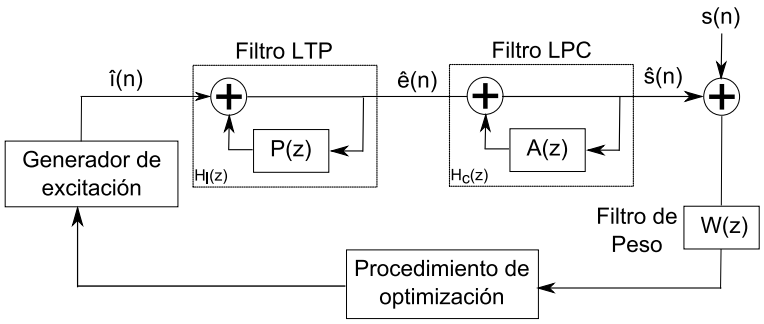

Figura 2.5: Esquema mejorado del codificador híbrido incorporando los filtros de largo retardo (LTP) y filtro de peso $W(z)$ durante el proceso de análisis por síntesis.

define como:

$$\epsilon = \sum_{n=0}^{N-1}(s(n) - \hat{s}(n))^2 = \sum_{n=0}^{N-1}\big(s(n) - \big(\sum_{k=1}^{p} a_k s(n-k) + \hat{e}(n)\big)\big)^2 \qquad (2.9)$$

donde N es el número de muestras de la trama actual y p es el orden de predicción para el filtro todo-polos que representa el tracto vocal con los coeficientes LPC (a_k).

Como se puede observar, cada muestra de la señal de excitación $\hat{e}(n)$ afecta a varias muestras en $\hat{s}(n)$, dada la recursividad presente en la expresión (2.5), por lo que la decisión sobre la mejor representación cuantizada no se realiza de forma instantánea sino que se retarda durante un intervalo. Por consiguiente, la elección de una señal de excitación se realiza midiendo su efecto sobre la señal de voz durante más de una muestra. Dado que para calcular el error en (2.9) es necesario realizar la síntesis durante el análisis, a este procedimiento se le conoce como codificación adaptable usando predicción y análisis por síntesis. Su principal característica es que el codificador incluye el propio decodificador, por tanto se conoce en cada momento la señal que se decodificará, y así mejorar la optimización mediante un proceso iterativo.

Aunque este esquema general ya supondrá una mejora notable en la calidad perceptual respecto a los *vocoders*, en la figura 2.5 se muestra un esquema de análisis por síntesis que mejora esta calidad introduciendo dos nuevos filtros en el proceso: el filtro de predicción de largo retardo y el filtro de peso.

Filtro de predicción de largo retardo El filtro de predicción de largo retardo o *long-term prediction filter* (LTP), también denominado como filtro de *pitch*, está formado por un predictor que modela las correlaciones a largo plazo debidas al *pitch* cuando la señal de voz es sonora. Su funcionamiento es similar al filtro

LPC pero en este caso se basa en observar un periodo de T muestras previas a la trama para realizar la predicción. La forma general del filtro LTP en el dominio transformado Z se define como:

$$P(z) = 1 - \sum_{k=-(q-1)/2}^{(q-1)/2} b_k z^{(T+k)} \tag{2.10}$$

donde el retardo T es un valor optimizado que toma valores en el intervalo de 2 a 20 ms (coincide con el periodo de pitch en segmentos sonoros y un valor aleatorio en segmentos sordos), b_k son los coeficientes de predicción de retardo largo y q el orden del filtro. Tanto el retardo T como los coeficientes b_k se determinan bien a partir de la señal de voz o a partir de la señal de excitación obtenida después de eliminar las correlaciones de retardo corto. De este modo, para un retardo dado T, los coeficientes b_k se calculan de forma similar a como se hace con los coeficientes LPC (a_k) [33]. Típicamente se utiliza de uno a tres coeficientes de predicción que se van adaptando a un ritmo de entre 50 y 200 veces por segundo. Ahora bien, dado que aumentar el orden de predicción incurrirá en un incremento de bits para codificar estos coeficientes, generalmente se suele utilizar un predictor de primer orden con un retardo no entero que permite una cuantización más eficiente [34]. Así, la combinación del filtro LTP y el filtro LPC determinará el comportamiento del tracto vocal final ($H(z) = H_l(z)H_c(z)$).

Filtro de peso El oído humano presenta una propiedad denominada enmascaramiento que consiste en que el sistema auditivo tiene poca capacidad de detectar ruido en las bandas donde la señal de voz tiene una alta energía [2]. Esto es importante dado que cuando se disminuye el número de bits para la codificación, también se está aumentando el ruido de codificación y lo hace más audible. La aplicación de un filtro de peso $W(z)$ antes de realizar la codificación permitirá que la energía del ruido se distribuya hacia aquellas zonas del espectro en las que el oído va a tener poca sensibilidad. Este filtro fue inicialmente investigado para los esquemas de codificación adaptable predictiva y se define como [35]:

$$W(z) = \frac{F(z/\gamma_1)}{F(z/\gamma_2)} = \frac{1 - \sum_{k=1}^{p} f_k \gamma_1^k z^{-k}}{1 - \sum_{k=1}^{p} f_k \gamma_2^k z^{-k}} \quad 0 \leq \gamma_2 \leq \gamma_1 \leq 1 \tag{2.11}$$

donde $F(z)$ es el predictor de retardo corto que en la mayoría de aplicaciones es el propio filtro LPC todo-polos, los parámetros γ_1 y γ_2 son factores con valores comprendidos entre 0 y 1 que controlan la energía del ruido en las regiones donde se

sitúan los formantes y que son establecidos heurísticamente mediante las correspondientes pruebas auditivas. Nótese que al reducir γ, se aumenta el ancho de banda de los ceros de $F(z/\gamma)$ [8].

A pesar de que este sistema permite enmascarar el ruido, este procedimiento es muy sensible a la codificación a utilizar ya que para tasas de bits pequeñas, el ruido es tan alto que aún aplicando este filtrado, sigue siendo audible. No obstante, con los codificadores híbridos se obtiene una buena calidad de la señal de voz en el rango de razones de transmisión que van de 0.5 a 2 bits/muestra, lográndose así tasas de bits comprendidas entre 4 y 16 kbps. Dada la buena relación calidad perceptual y tasa de bits utilizada, la mayoría de los codificadores actuales hacen uso de este esquema.

Dentro de los codificadores híbridos existen diferentes codificadores basados en el modelo LPC, los cuales difieren en cómo se realiza la representación de la señal de excitación. Los dos más importantes son los codificadores multipulso y los codificadores excitados por código.

Codificadores multipulso

El codificador por predicción lineal multipulso o *multipulse linear prediction coding*(MPLPC) [36], fue el primer codificador híbrido desarrollado en 1982 por B.S. Atal y J. Remde. Este codificador se caracteriza por representar la señal de excitación estimada $\hat{e}(n)$ por una serie de L pulsos, con ciertas amplitudes b_l y en determinadas posiciones n_l de acuerdo a la expresión:

$$\hat{e}(n) = \sum_{l=0}^{L-1} b_l \delta(n - n_l) \tag{2.12}$$

donde $\delta(n)$ se define como la función impulso unidad.

Tras definir los coeficientes del filtro LPC, es necesario determinar la posición y la amplitud de los L pulsos que minimizan el error en la expresión (2.9). Para poder obtenerlos habría que probar todas las combinaciones de L pulsos en N posibles posiciones y seleccionar aquella que minimice el error cuadrático de la expresión (2.9). Sin embargo, este proceso es inviable desde un punto de vista computacional, ya que existen $N!/L!(N-1)!$ posibles combinaciones. Por ello, en la práctica, este procedimiento se realiza siguiendo un algoritmo subóptimo como el propuesto por Singhal y Atal en [37] que primero determina las amplitudes óptimas de los coeficientes b_l y posteriormente sus posiciones empleando un procedimiento de mínimos

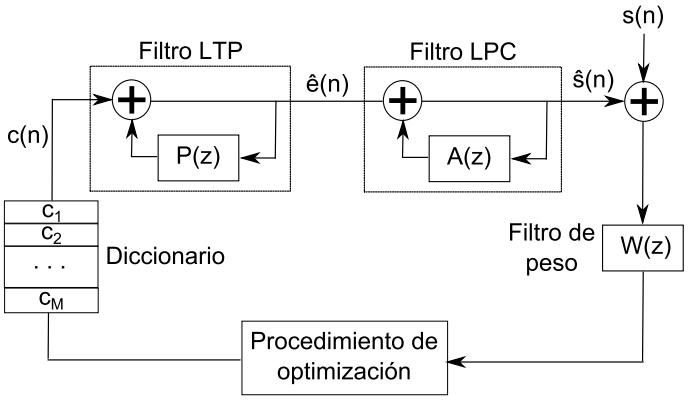

Figura 2.6: Esquema general de un codificador basado en el paradigma de predicción lineal por código excitado o *Coded Excited Linear Prediction* (CELP).

cuadrados o *Least Squares Error* (LSE).

Una particularización de esta codificación es la codificación de la señal de excitación como un tren de pulsos regulares o *Regular Pulse Excitation* (RPE-LTP) donde los pulsos se distribuyen de forma regular, por lo que no es necesario transmitir todas las posiciones sino solo la primera y las correspondientes amplitudes. Este sistema de codificación fue el implementado para el primer codec de voz digital de la segunda generación (2G) en telefonía móvil, popularmente conocido como *Global System for mobile communications* (GSM) [38]. La codificación multipulso supuso el punto de partida para desarrollar el codificador *Code Excited Linear Prediction* (CELP) que se describe a continuación y que incorporan la mayoría de los codecs actuales.

Codificadores CELP

Aunque la alternativa multipulso permite obtener una señal de excitación estimada $\hat{e}(n)$ con la que mejorar la calidad de la señal de voz sintetizada $\hat{s}(n)$, hay que tener en cuenta que requiere de la codificación de L posiciones y sus correspondientes amplitudes b_l, generando una alta tasa de bits. Para evitar esto, en 1985 se presentó la técnica de codificación CELP [13] desarrollada por B.S. Atal y M.R. Schroeder, cuyo esquema se muestra en la figura 2.6, y donde los filtros LTP y de peso comentados anteriormente se emplearán para mejorar la calidad de la señal recuperada aunque la tasa de bits sea reducida.

Esta técnica de codificación introduce una innovación que consiste en incorporar un diccionario de códigos fijos o *Fixed Codebook* (FCB), de tamaño M, donde cada

código $c(n)$ está formado por secuencias aleatorias gausianas de varianza unidad. Como resultado se tiene una señal de excitación definida como $e(n) = g_c c(n)$, donde la ganancia g_c se obtiene durante el proceso de optimización junto a la selección del código $c(n)$, de entre los M posibles, realizando una búsqueda exhaustiva para minimizar el error de cuantización. Este diccionario está disponible tanto en el codificador como en el decodificador por lo que sólo es necesario transmitir el índice de cuantización codificado con $(log_2 M)$ bits, permitiendo así tener unas tasas de bits entre 4 y 16 kbps.

La codificación CELP ha supuesto un nuevo paradigma en el que han surgido multitud de variantes. De todos ellos, cabe destacar el codec *Enhanced Full Rate* (EFR) [38], utilizado en 2G con una tasa de bits de 12.2 kbps, y el codec de tasa de bits variable o *Adaptive Multi-Rate* (AMR) [39], utilizado en 3G y 4G, hasta la aparición del nuevo codec de mejora de servicios de voz o *Enhanced Voice Service* (EVS) [40, 41], con tasas de bits comprendidas entre 4.75 y 12.2 kbps, como sus máximos referentes.

2.5. Codecs empleados en esta tesis

Para finalizar este capítulo, se ha añadido una sección donde se presenta una breve descripción de los codecs que se han empleado para la evaluación de las propuestas en esta tesis y que están basados en los codificadores explicados anteriormente.

En primer lugar, se presentan los codecs que se utilizan en las redes de ámbito local, donde se analizará el efecto de la degradación producida por el efecto multitrayecto. En esta tesis se han seleccionado los codecs del estándar para comunicaciones en teléfonos inalmbricos o *Digital Enhanced Cordless Telecommunications* (DECT). Este estándar presenta dos alternativas: el codec para transmisión en banda estrecha (codec G.726) y el codec para transmisión en banda ancha (codec G.722).

- **G.726**: El codec G.726 [24] es un estándar ITU-T desarrollado en 1982 y que utiliza la codificación ADPCM para codificar cada muestra de voz de una trama de 10 ms para transmisiones en banda estrecha. Este codec presenta varios modos de transmisión (16, 24, 32 y 40 kbps), aunque el modo utilizado tradicionalmente en las comunicaciones de voz es el de 32 kbps. De hecho, este codec fue creado para reemplazar los codecs G.721 (32 kbps) [21] y G.723 (24 y 40 kbps) [23]. Sin embargo, este codec no implementa ningún algoritmo para la mitigación de los errores producidos en el canal.

- **G.722**: El codec G.722 [22] es un estándar ITU-T desarrollado en 1983 que codifica cada muestra de voz, de una trama de 10 ms, empleando una codificación ADPCM por subbandas y cubriendo la banda ancha de 50 a 7000 Hz. El codec G.722 presenta varios modos de transmisión según los bits asignados a las componentes en la banda de frecuencias bajas, teniendo así los modos de 48, 56 y 64 kbps, aunque el modo de 64 kbps es el que se utiliza en las comunicaciones telefónicas sobre la tecnología DECT. Como consecuencia de la convergencia IP, este codec se empleará también para la tecnología de los teléfonos IP y dada la problemática con la pérdida de paquetes en este tipo de redes, en el año 2006 se le agregó un algoritmo de mitigación de errores (PLC) [42] para hacerlo robusto frente a la pérdida de paquetes en el canal.

Por otro lado, se presentan los codecs empleados para las transmisiones sobre redes IP, donde se analizará el efecto de la pérdida de paquetes. Además, para los codecs basados en el paradigma CELP, esta pérdida lleva consigo una propagación del error debido a la dependencia inter-trama producida por el uso del filtro LTP. Por este motivo, se han seleccionado dos codecs ampliamente conocidos en la bibliografía. Por un lado, el codec AMR como ejemplo de codec basado en el paradigma CELP, y por otro lado, el codec iLBC como alternativa que evita esta relación inter-trama.

- **AMR**: El codec de tasa de bits adaptativa o *Adaptive Multi Rate* (AMR) [39] es un codec que fue desarrollado en 1999 por el consorcio 3GPP. Este codec se basa en la codificación CELP, en concreto hace uso de la tecnología ACELP (*Algebraic CELP*) [39, 43], y su éxito radica en que presenta un mecanismo de adaptación en función de las degradaciones del canal móvil. De este modo, este codec dispone de 8 tasas de transferencia diferentes (4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.2, 12.2 kbps), donde cada trama tiene un tamaño de 20 ms, y a medida que la aumentan los errores en el canal se seleccionará un modo de transmisión inferior a fin de protegerse frente a pérdidas. Este codec fue el primer codec en adaptarse para funcionar tanto en redes GSM (el modo de transmisión 12.2 kbps se corresponde prácticamente con el estándar EFR) como en redes de conmutación de paquetes IP [44]. Además, aunque inicialmente este codec se desarrolló para transmisiones en banda estrecha, en el año 2001 se desarrolló el codec AMR-WB para transmisiones en banda ancha.

- **iLBC**: El codec de baja tasa de bits para Internet o *internet Low Bit-Rate Codec* (iLBC) [45] es un codec que fue desarrollado por Global IP en 2002 y

estandarizado en 2004 por la *Internet Engineering Task Force* (IETF). Así se recoge en el documento RFC3951 como una alternativa a los codecs tradicionales basados en el paradigma CELP para evitar la propagación del error. Para ello, este codec realiza la codificación de la señal de excitación de forma intratrama a partir de cierto segmento de la señal de excitación codificado mediante ADPCM. Aunque la nueva codificación aumenta su robustez frente a la pérdida de paquetes, también tiene como contrapartida, un incremento en la tasa de transmisión. Este codec presenta dos modos de transmisión para banda estrecha según al tamaño de trama escogido: 15.2 kbps para tramas de 20 ms y 13.33 kbps para tramas de 30 ms. Por último, su estandarización, así como la no necesidad de pagar de derechos, han hecho que este codec se difunda rápidamente en aplicaciones VoIP como Google Talk y Skype. De hecho, el organismo *CableLabs*, consorcio formado por los operadores de televisión por cable, ya incluye este codificador dentro de los recomendados para la codificación de voz en este tipo de redes [45].

Capítulo 3

Transmisión de voz sobre canales digitales

Este capítulo analizará el tipo de degradación que se produce en los dos tipos de redes considerados en esta tesis, las redes inalámbricas de ámbito local y las redes IP, así como los modelos empleados para simular el funcionamiento del canal de transmisión sobre estas redes.

También se presentarán las técnicas más relevantes en la bibliografía para mitigar esta degradación producida en el canal. Estas técnicas se clasificarán en dos grupos: las técnicas de prevención basadas en el emisor y las técnicas de mitigación de errores basadas en el receptor.

Por último, se describirá el entorno de trabajo o *framework* empleado en esta tesis. En esta sección se detallarán los tipos de tests objetivos y subjetivos empleados para analizar el rendimiento de las técnicas propuestas, las bases de datos utilizadas tanto para entrenamiento y test y finalmente, se indicarán las condiciones de canal simuladas para las pruebas sobre los codecs y las diferentes propuestas desarrolladas en esta tesis.

3.1. Introducción

Dado el interés de la sociedad por tener acceso a la información en cualquier momento y lugar, uno de los sectores que más ha evolucionado en los últimos años son los sistemas de telecomunicaciones, mejorando tanto las redes de comunicaciones como los dispositivos utilizados para este fin. Como consecuencia de estos avances tecnológicos, continuamente están apareciendo nuevas aplicaciones, nuevas tecno-

logías y estándares que proporcionan mayor funcionalidad y servicios al usuario final.

Sin embargo, no hay que olvidar que durante la transmisión de la señal de voz, ésta puede degradarse dependiendo de las condiciones del canal y afectar a la calidad del servicio. Por este motivo, esta tesis se centra en el estudio de dos tipos de degradación: el efecto multitrayecto en las redes de ámbito local y la pérdida de paquetes en las redes IP.

Por un lado, el efecto multitrayecto o *multipath* se produce debido a la recepción de la señal emitida como múltiples copias a consecuencia de la reflexión de la onda portadora en objetos cercanos al receptor. Este efecto provoca que la señal obtenida en el receptor presente desvanecimientos o *fadings* que pueden modificar la codificación del paquete enviado originalmente. En esta tesis centraremos el estudio de este tipo de degradación para transmisión de voz empleando la tecnología DECT, que es la que más ha expandido su uso mundial tanto en entornos domésticos como empresariales [46].

Por otro lado, la pérdida de paquetes se produce a consecuencia de la congestión en los nodos intermedios de las redes IP. Estas redes se caracterizan por emplear la conmutación de paquetes, en lugar de la conmutación de circuitos, para ser más tolerante a fallos. Sin embargo, lo que podría suponer una ventaja, también conlleva que si un nodo de la red recibe más paquetes hasta saturar la cola de entrada, el resultado será que habrá paquetes que no lleguen a su destino o en el tiempo adecuado para garantizar una comunicación en tiempo real sin cortes [3, 8]. Además, esta congestión puede afectar a más de un paquete que llegue a ese nodo saturado, dando lugar a una ráfaga de pérdidas. En esta tesis centraremos el estudio de este tipo de degradación para comunicaciones VoIP donde se evaluará el efecto de estas pérdidas sobre los codecs AMR [39] e iLBC [45].

En la bibliografía se pueden encontrar multitud de técnicas que tratan de mitigar el efecto de la degradación producida en ambos tipos de redes. Tradicionalmente, estas técnicas se clasifican como [6]:

- **Técnicas de prevención de errores basadas en el emisor**: Estas técnicas se caracterizan porque el emisor trata de prevenir el impacto de la degradación. Estas técnicas de prevención incluyen cierta información redundante, como sería el uso de códigos de corrección de errores hacia delante o *Forward Error Correction codes* (FEC), o reorganizan los paquetes enviados, mediante técnicas de entrelazado, para que, en caso de error en el canal, se pueda

compensar el efecto de la degradación.

- **Técnicas de mitigación de errores basadas en el receptor**: Estas técnicas se caracterizan porque, durante el proceso de decodificación, tratan de reducir el impacto de la degradación sobre la señal de voz recuperada. Dentro de este grupo se encuentran técnicas como la repetición, interpolación/extrapolación y métodos más sofisticados de estimación basados en un modelo estadístico de la voz.

Partiendo del estado del arte, en esta tesis se desarrollan diferentes propuestas que tratan de mitigar el efecto de la degradación y hacer al codec más robusto frente a errores en la transmisión. Para medir el rendimiento que ofrecen estas técnicas frente al resultado obtenido por el propio codec, que ya implementa un algoritmo de mitigación de pérdidas o *Packet Loss Concealment* (PLC), es necesario introducir un entorno de trabajo que permita simular las mismas condiciones de canal que genera el tipo de degradación a mitigar. Para ello, será necesario modelar el canal para medir el rendimiento del codec y nuestra propuesta en las mismas condiciones de canal. Por este motivo, en este capítulo también se presenta el entorno de trabajo empleado para la simulación de canal y obtención de resultados en esta tesis.

3.2. Transmisión de voz sobre redes locales

Las redes de comunicaciones se pueden clasificar de acuerdo con el radio de cobertura. Así, podemos encontrar, de menor a mayor cobertura, las redes de ámbito personal o *Personal Area Network* (PAN), que cubren espacios pequeños de hasta 10 metros, las redes de ámbito local o *Local Area Network* (LAN), que cubren hasta unos cientos de metros, las redes de ámbito metropolitano o *Metropolitan Area Network* (MAN), que cubren el tamaño de una ciudad, y las redes de ámbito amplio o *Wide Area Network* (WAN) como Internet [3]. En esta sección, nos centraremos en las comunicaciones realizadas sobre las redes LAN, las cuales se han popularizado en entornos empresariales para establecer comunicaciones o conexiones entre dispositivos (ordenador, terminal móvil, centralita, impresora, ...) dentro de un edificio o para dar acceso a Internet.

Además, el desarrollo de las transmisiones inalámbricas se ha impuesto en los últimos años sobre las redes LAN, ya que la reducción de cableado está permitiendo a los dispositivos mayor movilidad y accesibilidad que antes. De este modo se ha

identificado este tipo de redes como redes inalámbricas de ámbito local o *Wireless Local Area Networks* (WLAN). Multitud de estándares y servicios se han desarrollado para el ámbito de las redes WLAN y en el caso de la transmisión de voz destaca el estándar de comunicaciones para teléfonos inalámbricos *Digital Enhanced Cordless Telecommunications* (DECT) [47] que se describe a continuación.

3.2.1. El estándar de comunicaciones DECT

El estándar de comunicaciones DECT fue desarrollado durante los años 80 para entornos domésticos en Europa por la conferencia europea de administraciones de correos y telecomunicaciones, aunque finalmente sería el instituto de estándares de telecomunicaciones europeo o *European Telecommunications Standards Institute* (ETSI) el que publicara en 1992 el estándar DECT [47]. Este estándar está basado en tecnología de radio digital que permite mejorar algunos aspectos de las comunicaciones inalámbricas como son: la calidad de las comunicaciones telefónicas, la seguridad frente a escuchas y la interferencia con otros dispositivos o teléfonos cercanos. Tras la publicación del estándar en 1992, actualmente se está utilizado en más de 24 países en todo el mundo donde se utilizan diferentes frecuencias. En el caso de Europa, se emplea el rango de frecuencias de 1880 a 1980 MHz [46].

A pesar de la antigüedad de este estándar, éste ha conseguido imponerse a los dispositivos tradicionales fijos gracias al abaratamiento de los dispositivos, la fuerte implantación de las redes inalámbricas y a que permite compatibilidad e interoperatibilidad para los diferentes estándares de telefonía, ya sean sobre telefonía tradicional o la actual VoIP. De este modo, se ha convertido en la principal tecnología de comunicación inalámbrica para comunicaciones de voz en el mercado actual, representando un 73 % del mercado total en el sector [46]. Las principales características de esta tecnología son:

- **Mejora de la calidad de voz**: DECT utiliza de forma nativa una codificación ADPCM ya sea para banda estrecha, con el codec G.726 [24], o en banda ancha, con el codec G.722 [22]. Esto le permite tener una buena calidad en las comunicaciones, comparables con las realizadas a través de un teléfono cableado.

- **Capacidad de acceso múltiple**: DECT combina las técnicas de acceso múltiple por división de tiempo (TDMA) y transmisión en ambos sentidos por división de tiempo (TDD) [8] que le permite tratar posibles interferencias

y colisiones de manera similar a otras tecnologías basadas en una estructura celular.

- **Incremento de seguridad en las comunicaciones**: En DECT se incluyen algoritmos de protección de la información que se aplican durante el proceso de codificación haciendo que la escucha no deseada sea virtualmente imposible.

- **Selección dinámica de canal**: DECT implementa la tecnología de selección dinámica de canal que permite que el equipo terminal sea el que elija el canal de radio y la ventana de tiempo sobre la que realizar la comunicación en base a una monitorización periódica de las portadoras y ventanas que recibe. Esto permite dar movilidad al usuario cuando se desplaza de un área de cobertura a otra sin que tenga que reanudarse el servicio.

Por último, como consecuencia del amplio despliegue realizado por el estándar DECT, se están desarrollando diversas aplicaciones en las que DECT actúa como tecnología de acceso ya no solo en entornos domésticos sino también en entornos empresariales. Por ejemplo, el estándar DECT se está utilizando para el desarrollo de centralitas telefónicas para dar servicio a varios dispositivos móviles [46].

3.2.2. Degradación producida en redes de ámbito local

Pese a los beneficios que proporcionan las comunicaciones inalámbricas para facilitar la movilidad y acceso a los servicios en cualquier lugar, hay que tener en cuenta los numerosos desafíos que aparecen a la hora de diseñar mecanismos confiables para la transferencia de datos a altas velocidades. Esto se debe a que los canales inalámbricos son más propensos a errores.

Además de la atenuación de la potencia de la señal emitida con la distancia, hay que considerar también las interferencias con otros dispositivos que comparten el medio y los desvanecimientos de la señal provocados por el efecto multitrayecto [48]. Los dos primeros problemas se pueden solventar incorporando repetidores de señal o trabajando sobre bandas de frecuencia diferenciadas para cada tecnología. Sin embargo, los desvanecimientos no son sencillos de evitar ya que son el resultado de la recepción de múltiples copias como consecuencia de la reflexión en los objetos cercanos. Para explicar mejor esta degradación, a continuación se explicará su fundamento físico y cómo se modela para simular el canal.

Efecto multitrayecto en las comunicaciones inalámbricas

Partiendo de que las redes inalámbricas utilizan una onda electromagnética para realizar la transmisión de voz y datos desde el emisor al receptor, el efecto multitrayecto se produce cuando el receptor no recibe una única onda electromagnética sino una superposición de ondas provenientes desde diferentes direcciones con diferente amplitud y fase. Este fenómeno se produce debido a los efectos de propagación que sufre la onda (reflexión, difracción y dispersión) cuando ésta choca con un obstáculo. Así, la señal recibida en un instante de tiempo t se puede expresar como [17]:

$$x(t) = \sum_{l=1}^{L} A_l cos(2\pi f_c t + \Theta_l) \tag{3.1}$$

donde L es el número de ondas que han llegado al receptor, f_c es la frecuencia de la onda portadora y A_l y Θ_l son la amplitud y fase de cada onda recibida. Cabe destacar que el movimiento del emisor o del receptor pueden agravar los errores provocados por los desvanecimientos o *faddings* debidos al efecto Doppler [17].

3.2.3. Modelado del desvanecimiento temporal

El desvanecimiento temporal se produce cuando la señal recibida presenta una frecuencia f diferente a la frecuencia de la onda portadora f_c cuando el emisor o receptor se encuentran en movimiento. De este modo, la frecuencia f con la que el receptor recibe la señal viene definida como [17]:

$$f = \left(\frac{v_c + v_r}{v_c + v_e}\right) f_c \tag{3.2}$$

donde v_c es la velocidad de la onda tras ser emitida, v_r es la velocidad del receptor, v_e es la velocidad del emisor y f_c es la frecuencia de la onda portadora inicial. Como resultado de este cambio de frecuencia, se pueden producir fluctuaciones en la señal recibida que dan lugar a los desvanecimientos. A modo de ejemplo, si consideramos el emisor parado y el receptor se encuentra en movimiento, el incremento de la velocidad generará más desvanecimientos. Este hecho se observa en la figura 3.1 cuando el receptor se desplaza a 0,3 m/s y 3 m/s respectivamente. Así, con un desplazamiento a 3 m/s se observan muchos más desvanecimientos.

Algunos de los modelos más extendidos en la bibliografía para modelar el desvanecimiento temporal son: el desvanecimiento Rayleigh, el desvanecimiento Rician

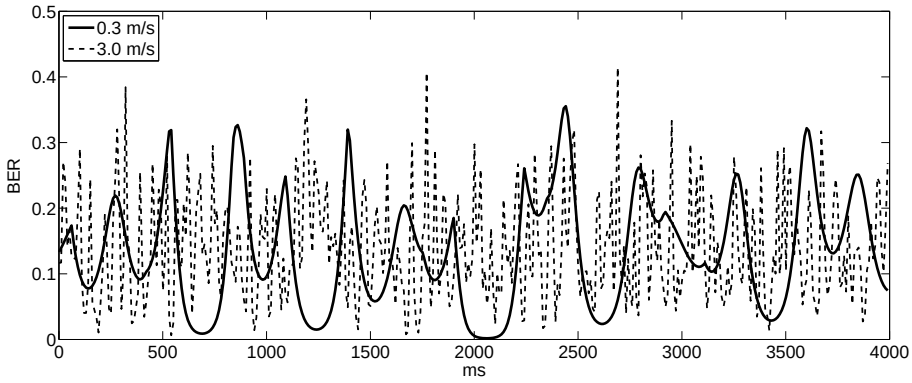

Figura 3.1: Ejemplo obtenido simulando un modelo de canal Rayleigh, para un relación señal-ruido de 0 dB, que muestra los desvanecimientos temporales producidos en las transmisiones inalámbricas cuando el emisor está parado y el receptor se desplaza a 0.3 y 3 m/s respectivamente. La tasa de error por bit o *Bit Error Rate* (BER) se empleará para modificar la codificación de la trama recibida.

y el desvanecimiento Nakagami. De entre ellos, el modelo de desvanecimiento Rayleigh es el más utilizado en la mayoría de las transmisiones y el más general al no considerar la visión directa como requisito [17].

La base de este modelo considera que la respuesta al impulso del canal se corresponde a un proceso gaussiano complejo compuesto por dos variables aleatorias gaussianas de media cero. Estas variables son independientes e idénticamente distribuidas siguiendo una distribución Rayleigh y como fase una distribución uniforme en $[0, 2\pi)$. La expresión general de una distribución Rayleigh se define como:

$$f_{Rayleigh}(r) = \frac{r}{\sigma^2} e^{\left(-\frac{r^2}{2\sigma^2}\right)} \quad r \geq 0 \tag{3.3}$$

donde σ^2 representa el valor medio de la potencia de la señal recibida y r una variable discreta temporal.

Como resultado, este modelo generará una secuencia de valores numéricos que durante la simulación del canal serán interpretados como la probabilidad de que un paquete pueda sufrir alguna modificación durante la transmisión.

3.3. Transmisión de voz y datos sobre redes IP

Como bien es sabido, las redes IP son las redes de comunicación más populares y sobre las que se desarrolla la mayoría de aplicaciones y servicios actuales. Las

redes IP tienen sus comienzos en el año 1969 en la rama científica del Departamento de Defensa de los Estados Unidos, la *Defense Advanced Research Projects Agency* (DARPA). Se inició así un ambicioso proyecto para la investigación y desarrollo de técnicas que permitan realizar una comunicación eficiente. De este modo aparecen las primeras redes de conmutación de paquetes que sucedieron a las redes de conmutación de circuitos empleadas en la telefonía tradicional.

Este fue el punto de partida de las redes IP, donde un paquete es enviado en una red no orientada a conexión, es decir, no hay un camino fijado al inicio de la comunicación. Esto supone una ventaja y es que, si un nodo se encuentra caído, la red automáticamente busca una ruta alternativa para llegar al destino. Además, al tratarse de una red no orientada a conexión, los paquetes pueden enviarse sin esperar a establecer la conexión con el equipo destino y sólo se necesita conocer la dirección de destino para realizar el encaminamiento. Es decir, no hay que reservar ni ancho de banda ni almacenamiento temporal a priori [3].

El éxito en la implantación de este tipo de redes permitió también interconectar redes de diferentes tecnologías, como podían ser las redes utilizadas para comunicaciones por satélite o redes inalámbricas, mediante el modelo *TCP/IP* [3]. Este modelo, desarrollado en 1977, se caracteriza por sus protocolos de control de transporte o *Transport Control Protocol* (TCP) [49] y el protocolo de Internet (IP) [3], y permite una comunicación transparente al usuario, de manera que éste perciba la interconexión de redes como una única red virtual. Su expansión tanto en tamaño de las redes conectadas como en funcionalidades nuevas sobre la misma es lo que dio origen a la actual Internet [3].

Uno de los servicios que más se ha popularizado es la comunicación de la voz a través de las redes IP, conocida como comunicación VoIP, que permite realizar comunicaciones más económicas y eficientes que los sistemas de telefonía tradicional. Pese a ello, la degradación que se produce en las redes IP afectará a la calidad de servicio y por este motivo, en las siguientes secciones se describirán las características de la comunicación VoIP, así como el tipo de degradación y el modelado empleado para simular canales con pérdidas sobre redes IP.

3.3.1. Las comunicaciones VoIP

Tradicionalmente, la transmisión de voz y datos se ha realizado a través de redes diferenciadas debido a que tienen requisitos de transmisión diferentes. Es decir, la red telefónica tradicional empleada para transmisión de voz utiliza conmutación de

circuitos mientras que la red IP utiliza conmutación de paquetes [3].

En la actualidad, el usuario no quiere realizar solo comunicaciones de voz sino que quiere enviar también imágenes y videos, como en el caso de una videoconferencia. Estos nuevos requisitos suponen modificaciones profundas en las redes de telefonía tradicional. Para poder minimizar los costes de mantenimiento y gestión de estos nuevos servicios, las empresas de telecomunicaciones unificaron esfuerzos para trasladar las comunicaciones de voz hacia las redes IP. Por este motivo, viendo el enorme desarrollo y prestaciones que las redes IP ofrecen, en 1995 surge la telefonía IP, empleando la voz sobre IP (VoIP), que permite realizar llamadas y videoconferencias de una forma más eficiente y menos costosa en recursos que la telefonía tradicional [17].

De este modo, VoIP trata la voz como otro tipo de dato a transmitir y, gracias a la inclusión de la voz en las redes IP, se está consumando la convergencia todo IP. Gracias a esta convergencia, se permite el uso simultáneo tanto de redes cableadas como inalámbricas y con diferentes tecnologías que, si antes eran competidoras, ahora se vuelven compatibles dentro de un único entorno como es la telefonía móvil de la cuarta generación (4G) basada en el estándar ETSI *Long Term Evolution* (LTE) [50]. De esta forma, los usuarios podrán mantener de forma ininterrumpida el uso de sus servicios incluso en el caso de estar en movimiento, atravesando diferentes redes de acceso y al mismo tiempo garantizando una calidad de servicio adecuada.

Ahora bien, al tratarse el protocolo IP de un servicio de mejor esfuerzo, no se puede garantizar que los paquetes lleguen a su destino, debido a la congestión en los nodos intermedios y los retardos variables. Esta congestión y retardos darán lugar a un tipo de degradación que es la pérdida del paquete enviado. Además del efecto de la propia pérdida para la reconstrucción de la señal de voz, en la siguiente sección se explicará qué otros efectos dañinos se producen sobre las redes IP y que provocan una baja calidad de servicio VoIP. No obstante, con la mejora en las prestaciones y calidad de servicio de las aplicaciones que hacen uso de las redes IP, se ha conseguido que las comunicaciones VoIP tengan una fuerte aceptación en la sociedad y que cada día aparezcan nuevas aplicaciones y servicios sobre ellas.

3.3.2. Degradación producida sobre redes IP

Pese al gran desarrollo de las redes de comunicaciones IP, hay que tener en cuenta que el servicio VoIP se caracteriza por tener restricciones de tiempo real, pero al mismo tiempo se quiere que el servicio VoIP sea fiable, seguro y que proporcione una

buena calidad de servicio. Sin embargo, asegurar la calidad de este servicio requiere de un tiempo adicional de acuerdo a la latencia o *jitter* de la red. Esta latencia es importante ya que si se superan los 150 ms [51], se puede estar generando el conocido efecto *walkie-talkie*, por el cual ya no se puede garantizar una comunicación fluida en ambos sentidos y en tiempo real. Es decir, este efecto provoca que haya que esperar para poder escuchar el mensaje enviado y así contestar debidamente. Como resultado, la comunicación se vuelve molesta y da una percepción de baja calidad del servicio VoIP [52].

Para mantener una comunicación en tiempo real, los codecs de voz tratan de recuperar la señal de voz a partir los paquetes que se van recibiendo. Sin embargo, la congestión en la red puede provocar que uno o varios paquetes no alcancen su destino a tiempo. En estos casos el propio codec debe de realizar la mitigación de la pérdida mediante un algoritmo *Packet Loss Concealment* (PLC). Como consecuencia, la señal de voz recuperada ya no será como la original y se produce una caída en la calidad perceptual.

Hasta ahora se ha considerado que la pérdida de calidad perceptual es debida a la pérdida de paquetes. Sin embargo, si el codec presenta una dependencia inter-trama, para la obtención de ciertos parámetros al realizar la síntesis, esta pérdida de paquetes lleva consigo errores en la obtención de los correspondientes parámetros en las tramas sucesivas aunque fueran recibidas correctamente. Este efecto, conocido como propagación del error, está presente en los codecs basados en el paradigma CELP, lo que los hace más vulnerables a las pérdidas en el canal.

A modo de ejemplo, en la figura 3.2 se muestra el efecto de la propagación del error en una simulación con pérdidas sobre el codec AMR [39], basado en el paradigma CELP. Como se aprecia, la señal recuperada se ve afectada no sólo en la zona de pérdida de paquetes sino que tras la última pérdida de la ráfaga, se observa que la señal de voz obtenida no coincide con la señal de voz original hasta pasado un tiempo. Por este motivo, es necesario el desarrollo de técnicas que minimicen este efecto o utilizar codecs como iLBC [45] que evita esta relación inter-trama o que corten la propagación del error realizando un intercalado de tramas sin relación inter-trama, como hace el codec EVS [40].

3.3.3. Modelado de la pérdida de paquetes

Con el fin de estudiar la pérdida de calidad perceptual producida en las transmisiones sobre redes IP, en 1993 Bolot et al. desarrolló un estudio exhaustivo sobre

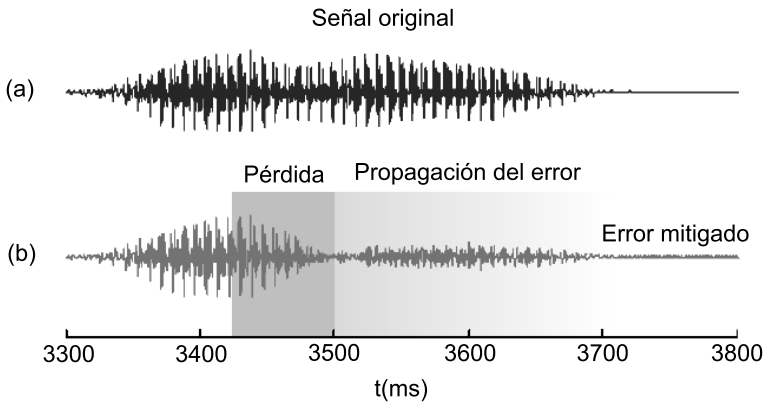

Figura 3.2: Ejemplo del impacto de una pérdida en la síntesis de voz de un codec basado en el paradigma CELP (codec AMR en el modo 12.2 kbps) a) síntesis de voz sin pérdidas. b) síntesis de voz con pérdidas aplicando el algoritmo *Packet Loss Concealment* implementado en el codec. (Fuente:[2])

el retardo y pérdida de paquetes en una conexión entre Francia y EE.UU. realizando medidas de tiempo de ida y vuelta con paquetes UDP enviados a intervalos regulares [53]. Así se observó que a partir del grado de saturación que tenían los nodos intermedios de la red, los paquetes no llegaban a intervalos regulares a su destino. Además, esta saturación provoca que las pérdidas no sean aisladas sino que se generan ráfagas de pérdidas consecutivas [54–56].

El problema es que no es posible conocer cómo de saturados se encuentran los nodos intermedios ni dónde se van a producir las pérdidas o de qué forma. Es decir, en dos transmisiones de voz, el comportamiento de la red no tiene por qué ser el mismo debido al uso del algoritmo de conmutación de paquetes. Sin embargo, para evaluar la calidad de las técnicas que se proponen en la tesis, sí es necesario comprobar, en las mismas condiciones de canal, qué técnica proporciona un rendimiento mayor. Por este motivo, es necesario desarrollar un modelo que pueda generar estas pérdidas de paquetes de manera similar a como lo haría un canal de transmisión real y así comparar el rendimiento de nuestra propuesta y la del codec en las mismas circunstancias.

Modelo de Bernoulli

El modelo de Bernoulli, también conocido como modelo de pérdidas aleatorias, genera una secuencia de 0 (paquete recibido) y 1 (paquete perdido) de acuerdo a la cantidad de pérdidas en la secuencia que se establece por una probabilidad de pérdidas incondicional o *Unconditional Loss Probability* ($r = P(X_t = 1)$). A

partir de esta probabilidad es posible determinar la probabilidad en una ráfaga de l paquetes consecutivos como:

$$P_l = r(1-r)^{l-1} \tag{3.4}$$

$$\bar{P}_l = (1-r)r^{l-1} \tag{3.5}$$

A partir de esta expresión se puede determinar la duración media de una ráfaga (L_{burst}) como:

$$L_{burst} = \sum_{l=1}^{\infty} l \cdot \bar{P}_l = (1-r)\sum_{l=1}^{\infty} l \cdot r^{l-1} = \frac{1}{1-r} \tag{3.6}$$

Aunque este modelo se haya utilizado en muchos trabajos en la bibliografía, hay que tener en cuenta que éste sólo permite ajustar la tasa media de pérdidas (r) y, puesto que no tiene en cuenta la correlación existente entre paquete enviados consecutivamente, no será posible simular casos de ráfagas de una determinada longitud promedio. De hecho, varios trabajos [54, 57, 58] coinciden en que no es un modelo apropiado para simular ráfagas de pérdidas sobre redes IP. Esto se demuestra en [59] donde se comprueba que este modelo sobreestima la aparición de pérdidas aisladas y no genera ráfagas largas.

Modelo de Gilbert

Para solventar las deficiencias del modelo anterior y acercarlo más a los canales IP reales, el modelo de Gilbert hace uso de un sencillo modelo de Markov que permite capturar la dependencia temporal entre las pérdidas [60]. El modelo más simple está representado en la figura 3.3 donde cada estado se corresponde con recepción correcta (S_0) o pérdida del paquete (S_1). Siguiendo este modelo, sólo habrá que concretar las probabilidades de las transiciones entre los estados de pérdida y no pérdida (p y q) para determinar la probabilidad de pérdida de paquetes (ulp) y la longitud de ráfaga (L_{burst}) de la siguiente forma [61]:

$$P_l = p(1-p)^{l-1}$$

$$\bar{P}_l = q(1-q)^{l-1}$$

$$L_{burst} = \sum_{l=1}^{\infty} l \cdot \bar{P}_l = \sum_{l=1}^{\infty} l \cdot q \cdot (1-q)^{l-1} = \frac{1}{q}$$

$$ulp = \frac{p}{1-q}$$

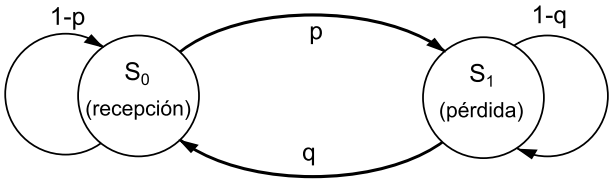

Figura 3.3: Diagrama de estados del modelo de Gilbert de dos estados.

A diferencia del modelo de Bernoulli, ahora el modelo de Gilbert dispone de un grado más de libertad que permite controlar la longitud de la ráfaga que es independiente de la probabilidad de pérdida. No obstante este modelo se puede convertir en un modelo de Bernoulli si se cumple la restricción $p + q = 1$.

Otros modelos

Aunque el modelo de Gilbert es un modelo bien aceptado para simular la pérdida de paquetes en una red IP, hay que tener en cuenta que se puede extender modificando el diagrama de la figura 3.3 para incorporar más estados y determinar mejor el comportamiento en los casos de pérdidas. Un ejemplo es el modelo de Milner y James propuesto en [62] que establece un tercer estado para simular que durante una ráfaga se puedan recibir paquetes correctos entre la ráfaga. Otro ejemplo sería extender el modelo de Gilbert con $n+1$ estados propuesto en [58] para simular mejor la longitud de la ráfaga.

Ambos ejemplos parten de una muestra previa para analizar lo que ocurre cuando se detecta una pérdida en la transmisión. Sin embargo, también sería posible obtener una representación más precisa de las correlaciones presentes en un proceso de pérdida de paquetes utilizando un modelo de Markov de orden superior para analizar así el tráfico previo al estado actual en la secuencia X_t de 0 y 1. De este modo, las probabilidades no sólo dependen de la variable previa sino de las n previas, de modo que las probabilidades de transición entre estados, que denotamos como X_t en el instante de tiempo t, responden a la forma $P(X_t|X_{t-1}, X_{t-2}, ..., X_{t-n})$. Como resultado, para desarrollar este modelo serían necesarios 2^n estados aunque en [54] se demuestra que un orden 6 es suficiente para modelar correctamente el tráfico que analizaron.

3.4. Técnicas de prevención y mitigación de errores en canales digitales

Para mitigar el efecto del error o degradación producido en el canal durante una transmisión de voz, ya sea en las redes de ámbito local o en las redes IP, es necesario el desarrollo de nuevas técnicas que ayuden a recuperar la información dañada o perdida durante la transmisión y dotando de mayor robustez al codec.

A continuación, se describirán las técnicas más relevantes de la bibliografía empleadas para mitigar el efecto de los errores producidos en el canal. Estas técnicas se clasificarán en dos grupos: las técnicas de prevención basadas en el emisor y las técnicas de mitigación de errores basadas en el receptor [6].

3.4.1. Técnicas de prevención de errores basadas en emisor

Las técnicas de prevención de errores basadas en el emisor son aquellas que requieren de la participación del emisor durante la recuperación de los paquetes modificados o perdidos durante la transmisión. Estas técnicas pueden clasificarse como activas o pasivas. Las primeras hacen referencia a la retransmisión de las tramas perdidas, mientras que las pasivas se encargan de prevenir el error o al menos, minimizar su impacto en la medida de lo posible.

Las técnicas activas se emplean frecuentemente para recuperar paquetes degradados durante la transmisión de datos, sin embargo, como se ha mencionado en la sección 3.3.2, las comunicaciones VoIP tienen unas fuertes restricciones temporales que hacen que estas técnicas sean inviables, ya que generan un incremento en el retardo que puede hacer inviable la comunicación. Por este motivo, esta sección se centra en el estudio de las técnicas pasivas que han sido mucho más empleadas en la bibliografía, destacando principalmente las técnicas de entrelazado y las técnicas de corrección hacia delante o *Forward Error Correction* (FEC).

Entrelazado

La técnica de entrelazado consiste en modificar el orden de envío de los paquetes al receptor de manera que, tras una ráfaga de paquetes degradados, su impacto se minimice a la hora de recuperar la señal de voz. Esto es, al volver los paquetes en el orden correcto, la ráfaga que ha podido afectar a varios paquetes consecutivos se distribuirán en ráfagas de menor longitud [63].

Por lo tanto, esta técnica minimiza el impacto de una ráfaga larga pero no realiza ninguna técnica de recuperación sobre los paquetes que han sufrido un error. Además, el incremento de la robustez de la transmisión de esta técnica se hace a costa de incrementar la latencia en la comunicación. Por este motivo, de acuerdo a las restricciones de cada sistema, habrá que escoger aquella técnica de entrelazado adecuada. Así, en [64] se establece un formalismo matemático que permite comparar y analizar el esquema de entrelazado respecto a su capacidad de dispersar la ráfaga de errores y los requerimientos necesarios en memoria y latencia [65]. Sin embargo, teniendo en cuenta que la comunicación debe hacerse en tiempo real y la latencia generada en el receptor para recuperarse ante ráfagas, se suele desaconsejar el uso de este tipo de técnicas.

Corrección de errores hacia delante

Las técnicas de corrección de errores hacia delante (FEC) se caracterizan por incluir información adicional en los paquetes enviados de manera que, si un paquete es degradado durante la transmisión de datos, éste pueda ser recuperado a partir de esta información adicional que se encuentra en los paquetes recibidos posteriormente.

Los códigos FEC pueden clasificarse según sean o no independientes de la información que se transmite. Así se pueden considerar códigos de corrección independientes del medio o códigos de corrección específicos del medio. Cada uno presenta unas ventajas e inconvenientes que se detallan a continuación. No obstante los códigos FEC específicos del medio son los más ampliamente estudiados en la bibliografía para codificación.

Técnicas de corrección de errores independiente del medio Los códigos FEC independientes del medio se caracterizan porque no realizan ninguna suposición acerca de la información que se transmite. De esta manera, utiliza paquetes adicionales en la transmisión que incluyen la información de reconstrucción de los paquetes previamente enviados. Es decir, cada código FEC se construye a partir de n paquetes de datos que generan k paquetes adicionales, por lo que habrá que transmitir $n + k$ paquetes en la red.

El funcionamiento de este tipo de códigos FEC puede verse en la figura 3.4. A partir de 4 paquetes a enviar, se generará un paquete adicional, tradicionalmente de paridad o empleando el código Reed-Solomon, de manera que en caso de pérdida de uno de estos 4 paquetes, se puede recuperar a partir del resto de los paquetes recibidos y este paquete FEC adicional.

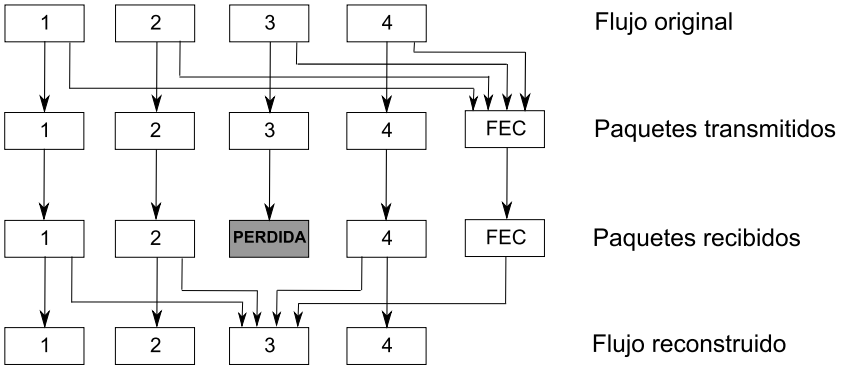

Figura 3.4: Esquema de recuperación de paquetes perdidos mediante la aplicación de un código de corrección de errores hacia delante independiente del medio. El paquete adicional (FEC) permitirá la recuperación de cualquiera de los 4 paquetes enviados.

Sin embargo, en el caso de que se produzca una ráfaga que afecte al paquete de corrección FEC, la reconstrucción será imposible, con el consecuente agravio de que ya se ha incrementado el tráfico en la red y la posible congestión al incluir más paquetes adicionales. Por este motivo, los códigos FEC más empleados en la bibliografía son los códigos FEC dependientes del medio que se detallan a continuación.

Técnicas de corrección de errores dependientes del medio Las técnicas FEC dependientes del medio se diferencian de las anteriores en que conocen el tipo de medio que se transmite y los esquemas de compresión disponibles para éste. La base de esta técnica está en replicar información de un paquete o paquetes anteriores dentro del paquete que se va a enviar, de manera que permita restaurar los paquetes perdidos previamente. Por supuesto, lo ideal sería que las réplicas fueran idénticas a los paquetes originales y así recuperar los paquetes degradados de forma exacta, sin embargo, esto multiplicaría la tasa de bits en tantas veces como se repitan los paquetes y podría ser inviable su transmisión en canales con un ancho de banda limitado.

Para reducir este incremento en la tasa de bits a transmitir por paquete, normalmente se realiza una codificación secundaria sobre cada una de las tramas a replicar y así se pueda realizar la comunicación sobre canales con un ancho de banda limitado. No obstante, el problema al realizar esta codificación secundaria radica en que el paquete recuperado tiene un error de cuantización mayor que el que se envió, por lo que la calidad perceptual de la señal recuperada también será menor. Para entender el funcionamiento de esta técnica, en la figura 3.5 se puede ver cómo cada paquete tiene una réplica del paquete anterior que es utilizado en caso de pérdida.

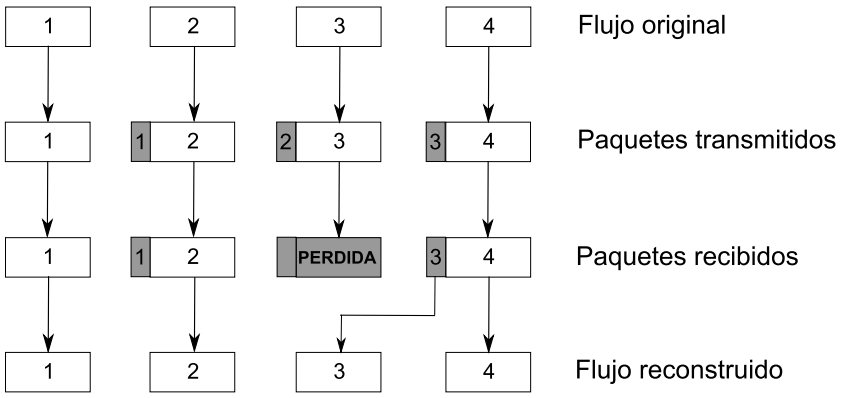

Figura 3.5: Esquema de recuperación de paquetes perdidos mediante la aplicación de un código de corrección de errores dependiente del medio. Cada paquete incluye una réplica del paquete anterior codificado con un número menor de bits que permite la recuperación del paquete modificado o perdido.

Una ventaja de esta técnica respecto a la anterior es que es más robusta frente a errores en el canal, ya que siempre habrá un código FEC en los paquetes recibidos para recuperar paquetes modificados o perdidos. Sin embargo, también está limitada su capacidad de recuperación dependiendo del número de réplicas que se incorpora a cada paquete y la longitud de ráfaga producida. Es decir, como se puede observar en la figura 3.5, si sólo se pierden paquetes aislados, es posible recuperar el paquete perdido aunque sea de forma degradada, pero si la longitud de ráfaga es superior a 2, sólo será posible recuperar el último paquete perdido, perdiendo el resto irremediablemente.

Por supuesto, cuanto más información redundante sea posible enviar, mayor será la capacidad de recuperarse frente a errores en el canal. Sin embargo, en las transmisiones siempre hay que buscar un compromiso entre la capacidad de reconstrucción y el correspondiente incremento en la tasa de bits. Además, el uso de un código FEC conlleva alterar la estructura de los paquetes enviados y como consecuencia, ya no se puede utilizar el decodificador del codec estándar aunque se realice una transmisión sin errores en el canal.

Este tipo de técnicas de corrección de errores dependiente del medio han sido muy estudiadas en la bibliografía tanto para la recuperación de paquetes perdidos [66–74] para transmisiones robustas, como para evitar el problema de la propagación del error [68, 69, 71] en los codecs basados en el paradigma CELP [13]. Así, en el Capítulo 5 se plantean esquemas que utilizan códigos FEC tanto para la recuperación de la pérdida como la mitigación de la propagación del error.

3.4.2. Técnicas de mitigación basadas en el receptor

Las técnicas de mitigación de errores basadas en el receptor se caracterizan por no requerir la participación del emisor para reconstruir los parámetros que contenía el paquete perdido o modificado. De este modo, los algoritmos de mitigación de pérdidas (PLC) que incluye la mayoría de los codecs utilizados para transmisión de voz permiten mitigar el error provocado en el canal a partir de la información recibida correctamente antes y/o después del paquete afectado.

En la bibliografía se pueden encontrar diferentes aproximaciones o esquemas para la mitigación de los paquetes modificados o perdidos. Así, en esta sección se hace una breve clasificación de estas técnicas en base a su complejidad, desde técnicas sencillas como la inserción de paquetes hasta otras técnicas más complejas basados en modelos estadísticos que mejoran la calidad perceptual de la señal de voz reconstruida.

Técnicas de inserción de paquetes

Las técnicas de inserción de paquetes son aquellas que sustituyen el paquete perdido por otro sin tener en cuenta las características de la señal [6]. De este modo, no se perturba la continuidad temporal de la señal decodificada como sí ocurriría aplicando una técnica de ensamblado [6]. Tradicionalmente las técnicas de inserción más utilizadas son la inserción de silencio, ruido o la repetición de los paquetes recibidos.

Las dos primeras alternativas se utilizan para sustituir los paquetes perdidos y mantener así la relación temporal entre los paquetes a la vez que son sencillas de implementar. Sin embargo, de entre estas dos, la inserción de ruido es más apropiada que la inserción de silencio porque en diversos estudios como en [75] se demuestra que el cerebro humano subsconscientemente recupera mejor segmentos con ruido que con silencio. No obstante, en ambos casos sólo sería efectivo el uso de estas técnicas para ráfagas cortas.

Por otro lado, el algoritmo de mitigación por inserción más utilizado en la bibliografía es la repetición de los paquetes recibidos antes y después de la ráfaga como en [76]. No obstante, la repetición en ambos sentidos presenta un problema, ya que, para poder aplicarlo, hay que esperar a obtener una trama correcta tras la ráfaga. Por lo tanto, este retardo puede hacer molesta la comunicación y dadas las restricciones en tiempo, la repetición se suele realizar únicamente hacia delante.

La técnica de repetición hacia delante ha sido la técnica más utilizada en los

codecs de voz por su simplicidad [76, 77]. En muchos trabajos presentes en la bibliografía, el parámetro que más se utiliza en las técnicas de repetición de parámetros es el *pitch*, que es replicado sobre los paquetes perdidos como en [78–83]. Sin embargo, dado que la señal de voz no presenta un comportamiento estacionario en el tiempo, la repetición de paquetes sólo será efectiva para ráfagas cortas. De lo contrario, pueden producirse discontinuidades o sonidos molestos durante la reproducción. Por este motivo, para ráfagas largas la mayoría de los algoritmos PLC, implementados en los codecs actuales, incluyen un efecto de apagado progresivo o *muting* a partir de un determinado número consecutivo de pérdidas.

Técnicas de interpolación/extrapolación de paquetes

Como alternativa a las técnicas de repetición, las técnicas de interpolación pueden proporcionar una transición más suave entre los últimos paquetes recibidos antes de la ráfaga y los primeros paquetes recibidos tras una ráfaga, que consideraremos de longitud T, y completar de manera eficaz las pérdidas ocasionadas por la ráfaga. De forma general, este proceso de interpolación para obtener el parámetro $\hat{\mathbf{x}}_t$ en el instante temporal t, a partir de los parámetros recibidos previamente y posteriormente, como [84]:

$$\hat{\mathbf{x}}_t = I(t; \mathbf{x}_{-\mathcal{M}+1}, \mathbf{x}_{-\mathcal{M}+2}, ..., \mathbf{x}_0, \mathbf{x}_{T+1}, \mathbf{x}_{T+2}, ..., \mathbf{x}_{T+\mathcal{N}},) \quad (1 \leq t \leq T) \tag{3.7}$$

donde I es la función de interpolación y $\mathcal{M}$ y $\mathcal{N}$ son el número de parámetros correctamente recibidos antes y después de la ráfaga de tamaño T. Como se puede observar en la expresión, la técnica de interpolación en ambos sentidos generará una latencia en el proceso de decodificación de $T + \mathcal{N} - 1$, para $\mathcal{N} > 0$ [85].

Partiendo de esta formulación, la técnica de repetición hacia delante puede considerarse como un caso específico al considerar $\mathcal{M} = 0$ y $\mathcal{N} = 1$. Del mismo modo, considerando únicamente el vector previo $\mathbf{x}_0$ y posterior a la ráfaga $\mathbf{x}_{T+1}$, se obtiene la interpolación lineal definida como [85]:

$$\hat{\mathbf{x}}_t = \mathbf{x}_0 + \frac{t}{T+1}(\mathbf{x}_{T+1} - \mathbf{x}_0) \tag{3.8}$$

La interpolación lineal es la técnica más ampliamente utilizada en la bibliografía para realizar una interpolación entre los parámetros x_0 y x_{T+1} como en [85, 86]. No obstante, también hay otras variantes que realizar una interpolación entre los paquetes pares o impares, como se hace en [87], o utilizando otros sistemas más complejos

basados en un modelo de estimación de la voz [88–90]. De nuevo, estos procesos de interpolación van a tener el problema del retardo que se introduce por la longitud de la ráfaga T, dado que se necesita el parámetro x_{T+1} para realizar este proceso. Una alternativa para reducir este retardo sería realizar una técnica de extrapolación que puede verse como un caso particular de las técnicas de interpolación cuando se aplican sólo hacia delante como en [91–93].

Técnicas de estimación de paquetes degradados

Las técnicas de estimación se caracterizan por ofrecer un valor para aquellos parámetros que no están disponibles o son modificados, debido a la degradación producida en el canal, mediante el uso de un modelo estadístico de la voz. A diferencia de las técnicas de interpolación anteriores, el modelo utilizado en las técnicas de estimación está entrenado con voz mientras que para los métodos de interpolación sólo se tienen en cuenta consideraciones como la forma de la trayectoria de las características de la voz para obtener el parámetro perdido.

Al disponer de un modelo de voz que caracteriza la evolución de sus parámetros en el tiempo, es posible realizar una estimación a partir de los parámetros conocidos. Estos parámetros pueden ser tanto los parámetros anteriores y posteriores a la ráfaga de pérdidas como los propios con error si se dispone de ellos (caso de las redes WLAN). Para realizar esta estimación, en esta tesis se ha considerado, un método de estimación bien conocido en la bibliografía: la estimación basada en el mínimo error cuadrático medio o *Minimimum Mean Square Error* (MMSE) [94].

Estimación por mínimo error cuadrático medio

La estimación por mínimo error cuadrático medio (MMSE) consiste en calcular el valor esperado del parámetro $\hat{\mathbf{x}}_t$, en un instante de tiempo t, a partir de los datos disponibles Δ. Así, la estimación del parámetro $\hat{\mathbf{x}}_t$ se obtiene como:

$$\hat{\mathbf{x}}_t = E[\mathbf{x}_t|\Delta] \tag{3.9}$$

Considerando una cuantización previa de los vectores de parámetros de voz, tal que $\mathbf{x}_t$ pertenece a un conjunto finito de vectores $\{\mathbf{x}^{(i)}; i = 1, ..., C\}$, la estimación

MMSE de $\hat{\mathbf{x}}_t$ puede expresarse como:

$$\hat{\mathbf{x}}_t = \sum_{i=1}^{C} \mathbf{x}^{(i)} P\Big(\mathbf{x}_t = \mathbf{x}^{(i)} | \Delta\Big) \tag{3.10}$$

donde C es el tamaño del diccionario y las probabilidades $P\Big(\mathbf{x}_t = \mathbf{x}^{(i)} | \Delta\Big)$, denominada probabilidad a posteriori, son proporcionadas por el modelo.

El conjunto de datos disponible Δ, en una ráfaga de longitud T, puede expresarse como $(\mathbf{X}^-, \mathbf{Y}_t^T, \mathbf{X}^+)$, donde $\mathbf{X}^-$ son los parámetros recibidos antes de una ráfaga, $\mathbf{Y}_t^T$ son los T parámetros degradados en la ráfaga y $\mathbf{X}^+$ son los parámetros recibidos tras la ráfaga. Cabe destacar que, dependiendo de la degradación, los parámetros $\mathbf{Y}_t^T$ pueden no estar presentes en la expresión, (como sería el caso de las redes IP).

La estimación MMSE permite combinar la información a priori disponible acerca de la evolución de la fuente y la información recibida del canal u observación $\mathbf{y}_t$ durante la ráfaga. Por un lado, la probabilidad a priori de la fuente se modela de acuerdo a un proceso discreto de Markov y que permite conocer la probabilidad de que un determinado símbolo $\mathbf{x}^{(i)}$ sea transmitido. Por otro lado, este símbolo puede ser modificado, de acuerdo a un modelo de canal, lo que permite conocer la probabilidad de observación $P(\mathbf{y}_t | \mathbf{x}^{(i)})$.

En el caso particular de una transmisión sobre una red WLAN, dado que el parámetro modificado $\mathbf{y}_t$ está disponible, la expresión (3.10) puede expresarse como:

$$\hat{\mathbf{x}}_t = \sum_{i=1}^{C} \mathbf{x}^{(i)} P\Big(\mathbf{x}_t = \mathbf{x}^{(i)} | \mathbf{y}_t\Big) = \sum_{i=1}^{C} \mathbf{x}^{(i)} \Big(\frac{P(\mathbf{y}_t | \mathbf{x}^{(i)}) P(\mathbf{x}^{(i)})}{\sum_{l=1}^{C} P(\mathbf{y}_t | \mathbf{x}^{(l)}) P(\mathbf{x}^{(l)})}\Big) \tag{3.11}$$

de este modo, la probabilidad a posteriori depende sólo de la probabilidad a priori y la probabilidad de observación.

Para el caso de las transmisiones sobre redes IP, dado que el parámetro $\mathbf{y}_t$ no está disponible, la probabilidad de observación adoptará una función de distribución uniforme al no disponer del paquete enviado con errores. Como consecuencia de no disponer del paquete enviado, el orden del modelo de Markov tiene que ser mayor o igual a 1. Bajo esta situación y de manera genérica, la probabilidad a posteriori se calculará de la siguiente forma [84, 95]:

$$P(\mathbf{x}_t = \mathbf{x}^{(i)} | \mathbf{X}^-) = \frac{P(Y^- | \mathbf{x}_t = \mathbf{x}^{(i)}) P(\mathbf{x}^{(i)})}{P(\mathbf{X}^-)} = \frac{P(\mathbf{x}^{(i)})}{C \cdot P(\mathbf{X}^-)} \tag{3.12}$$

desde donde se podrá obtener la probabilidad a posteriori de acuerdo a cuánta historia previa se considere en $\mathbf{X}^-$ [84, 95].

Finalmente, en ambos casos es posible extender las expresiones (3.11) y (3.12) para modelos de Markov de orden superior considerando también la historia previa ($\mathbf{X}^-$). No obstante, en la práctica puede resultar inviable alcanzar una estimación para un orden de predicción M. El principal inconveniente para su obtención es que será necesaria una base de datos de entrenamiento casi ilimitada para obtener la probabilidad $P\left(\mathbf{x}_t = \mathbf{x}^{(i)} | \mathbf{y}_t, \mathbf{X}^-\right)$ con todas las combinaciones posibles [84].

3.5. Metodología experimental

En esta sección se presenta el marco de trabajo experimental donde tanto el codec original como las propuestas de esta tesis van a ser evaluadas de acuerdo a unas métricas de calidad. Para ello, se describen estas métricas, así como las bases de datos empleadas para realizar entrenamiento y test y cómo se ha configurado el canal para generar diferentes situaciones de error en el canal.

3.5.1. Medidas de evaluación de la calidad perceptual

Para demostrar que una nueva técnica tiene un rendimiento mayor al funcionamiento estándar del codec original (con su propio algoritmo PLC), es necesario definir unos métodos de evaluación de calidad perceptual que permitan cuantificar esta mejora. Para obtener esta métrica se pueden emplean tests de calidad subjetivos, donde los oyentes escuchan y evalúan la calidad de la señal, o tests de calidad objetivos, donde se intenta predecir la calidad subjetiva a partir de medidas cuantitativas obtenidas a partir de las señales a evaluar. Aunque los métodos subjetivos son más realistas a la hora de valorar la calidad perceptual, también requieren de más recursos para su obtención, y dada la alta correlación que logran los métodos objetivos [96], éstos últimos son útiles para una selección previa de técnicas para realizar un test subjetivo. En esta tesis se ha querido demostrar esta correspondencia de calidad perceptual, por lo que se han utilizado ambas métricas para medir el rendimiento de las técnicas desarrolladas.

Tests de calidad subjetivos

Los tests de calidad subjetivos se caracterizan por utilizar un conjunto de oyentes que bajo unas determinadas condiciones [97, 98] califican las señales empleadas como estímulos como: (1) mala, (2) pobre, (3) razonable, (4) buena y (5) excelente. Tras obtener un determinado número significativo de tests, la calificación de valoración subjetiva media o *Mean Opinion Score* (MOS) se obtiene como la media de las valoraciones dadas por los oyentes al evaluar la calidad de la propuesta. Como se ha indicado anteriormente, este test presenta una serie de desventajas por el alto coste en tiempo y recursos humanos necesario para tener unos valores estadísticamente significativos.

Para poder reducir este coste, en la actualidad se prefiere la metodología denominada *MUltiple Stimuli with Hidden Reference and Anchor* (MUSHRA) [99]. Esta metodología se basa en una prueba doblemente ciega y multiestímulo con una referencia oculta donde se compara la señal de referencia de máxima calidad frente al resultado de la propuesta a analizar y una referencia de baja calidad o *anchor*. La prueba de audición se realiza en una o varias sesiones y está compuesta por una serie de items que contiene una señal procesada de diferentes formas (estímulos) los cuales queremos evaluar junto a los dos referencias anteriormente indicadas (referencia y *anchor*). El objetivo es que el oyente pueda detectar las degradaciones en los estímulos y pueda indicar su valoración de acuerdo con la referencia conocida en una escala que va de 0 a 100. Esta escala se corresponde con la escala MOS por cada intervalo de 20 puntos, de ahí que 0-20 se corresponde como mala y 80-100 como excelente. Además, al tratarse de valores relativos y no absolutos, el oyente puede decidir si un estímulo suena mejor que otro y cuánto mejor en relación a las referencias. De esta forma se homogenizan los resultados de los oyentes y permite reducir el número necesario para obtener resultados estadísticamente significativos en comparación con la técnica MOS. De hecho en la mayoría de las ocasiones, con 20 oyentes ya se puede obtener un resultado estadísticamente significativo [99]. Por este motivo, para las pruebas que se desarrollan en esta tesis, se buscará entre 20-25 personas para obtener resultados significativos. Del mismo modo, este grupo de personas se escogerá entre compañeros del departamento y familiares o amigos para tener un grupo diverso. Es decir, se busca que el grupo de personas no sean necesariamente expertos en procesamiento de la voz (como los compañeros del departamento) y que sean mayores de edad (hombres o mujeres con edades comprendidas entre 20 y 60 años), de manera que este grupo de personas sea representativo de la población.

Aunque existen otras metodologías complementarias en [100] que permiten evaluar pequeñas distorsiones con mayor detalle. En esta tesis se ha escogido el test MUSHRA para evaluar la calidad perceptual y en la bibliografía es una de las métricas más utilizadas para la evaluación subjetiva.

Tests de calidad objetivos

En la mayoría de las aplicaciones para el procesamiento de señales, la medida del grado de semejanza entre dos señales se obtiene calculando la relación señal ruido o *signal to noise ratio* (SNR). Sin embargo, para el caso de la voz se tiene que tener en cuenta la influencia de las características psicoacústicas, como el enmascaramiento, que hacen que este tipo de mediciones no reflejen bien la calidad percibida subjetivamente por los humanos [101].

Uno de los primeros métodos objetivos que se acercaba al modelo de percepción del oído fue desarrollado en 1998 por el organismo ITU-T bajo el nombre de medida de calidad perceptual de la voz o *Perceptual Speech Quality Measure* (PSQM) [102]. Sin embargo, esta recomendación no tiene en cuenta los efectos producidos por el filtrado, el retardo variable y las distorsiones cortas localizadas [103]. De este modo, en el año 2001 apareció el algoritmo de evaluación de la calidad perceptual de la voz o *Perceptual Evaluation of Speech Quality* (PESQ) [103] que sí considera esos efectos mediante la ecualización, la alineación en el tiempo y un nuevo bloque que promedia las distorsiones en función del tiempo. Así, como resultado de este test se obtendrá un resultado numérico entre -0.5 y 4.5, donde -0.5 representa muy mala calidad y 4.5 muy buena calidad.

En los últimos años, este algoritmo viene siendo el referente para realizar las pruebas objetivas de calidad de voz gracias a que los resultados obtenidos presentan una correlación elevada frente a una valoración subjetiva MOS que es mucho más costosa. Este hecho se puede comprobar en [103] y en la figura 3.6 donde la calificación objetiva PESQ y la calificación subjetiva MOS tienen una relación cuasi-lineal entre los intervalos de 2 y 4 en el test MOS (calidad pobre y calidad buena) con el intervalo correspondiente entre 2.5 y 4 en el test PESQ.

Por último, cabe indicar que, aunque inicialmente el algoritmo PESQ fue desarrollado para codecs de banda estrecha, éste se ha extendido para poder realizar tests objetivos sobre codecs de banda ancha. Por todas las razones expuestas, el algoritmo seleccionado para realizar los tests objetivos en esta tesis será el algoritmo PESQ.

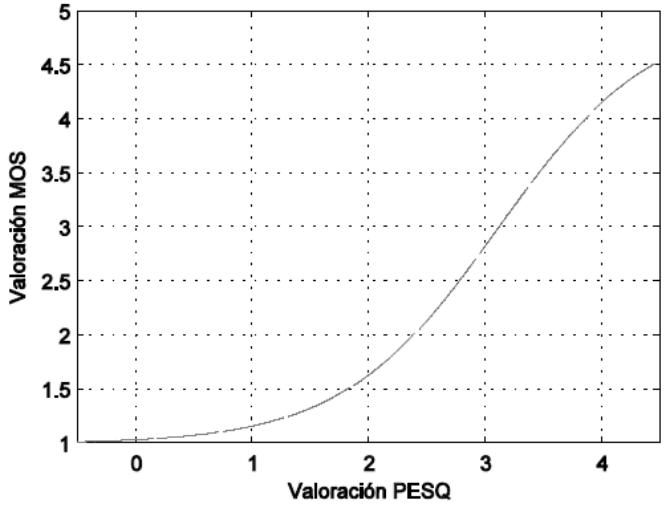

Figura 3.6: Función de correspondencia de los resultados obtenidos por el algoritmo PESQ y las valoraciones subjetivas MOS. (Fuente:[2])

3.5.2. Bases de datos

En esta tesis se han seleccionado principalmente dos bases de datos diferentes (*TIMIT* [104] y *NTT* [105]) tanto para realizar operaciones de entrenamiento (como obtener diccionarios o probabilidades), como para el desarrollo de las evaluaciones objetivas con un conjunto de test. La base de datos *TIMIT* se utilizará con las técnicas de mitigación y prevención de pérdida de paquetes sobre redes IP mientras que la base de datos *NTT* se utilizará para la mitigación del efecto multitrayecto sobre las redes de ámbito local. Aunque a lo largo de esta tesis se hará alusión a las dos bases de datos anteriores para la realización del entrenamiento y test de nuestras técnicas, para la realización de las pruebas subjetivas se ha escogido una base de datos con el idioma de los oyentes, como es la base de datos *ALBAYZIN* [106], para facilitar la realización del test MUSHRA en el idioma nativo de los oyentes. A continuación se describe cada una de las bases de datos utilizadas:

- **Base de datos TIMIT**: La base de datos TIMIT es un corpus de 6300 frases donde 630 hablantes (70 % hombres y 30 % mujeres) pronunciaron 10 frases cada uno. Estas frases se encuentran balanceadas fonéticamente y se corresponden con los 8 mayores dialectos americanos y fueron muestreadas a 16 KHz. Esta base de datos fue encargada por la *Defense Advance Research Project Agency - Information Science and Technology* (DARPA-ISTO), ahora conocido como *Softaware and Intelligent Systems Technology Office* (SISTO),

para realizar investigaciones fonético-acústicas y desarrollo y evaluación de aplicaciones de reconocimiento del habla. Para elaborar esta base de datos intervinieron varias entidades como *Texas Instruments* (TI), que grabó el corpus, y el Instituto Tecnológico de Massachussets (MIT), que realizó tareas de transcripción, que al mismo tiempo definieron el nombre actual de la base de datos.

Para poder ser utilizada en nuestras pruebas, hay que tener en cuenta que inicialmente las frases no están balanceadas por sexo y no tienen la misma longitud. Así, en primer lugar se realiza un remuestreo de cada señal de voz de acuerdo al ancho de banda necesario (4 kHz para banda estrecha), para luego concatenar las frases hasta alcanzar una duración promedio de 14 segundos. Esta duración promedio está recomendada por el estándar PESQ [103] entre 8 y 20 segundos de duración. Del conjunto de frases concatenadas se escogieron 1328 frases (50 % hombres y 50 % mujeres) de las cuales 928 serán para realizar el entrenamiento y 450 para realizar tests.

- **Base de datos NTT**: La base de datos NTT [105] fue desarrollada por *NTT Advanced Technology* entre 1989 y 1991. Es una base de datos que contiene frases de 21 idiomas incluyendo: inglés americano y británico, español, francés, tailandés, alemán, griego, chino, portugués, sueco,... entre otros, más los que se puedan incorporar en un futuro. Cada uno de estos lenguajes tiene 96 frases pronunciadas por 4 hombres y 4 mujeres nativos. La longitud de estas frases (entre 8 y 10 s) es suficiente para poder aplicar el test PESQ sin tener que realizar concatenación como se ha realizado con la base de datos TIMIT. De los lenguajes incorporados en esta base de datos, se han escogido 15 idiomas para entrenamiento y 6 para tests.

  Esta base de datos se empleará para evaluar la calidad perceptual en la propuesta realizada sobre los codecs G.726 y G.722, que no comparten el mismo ancho de banda, G.726 utiliza banda estrecha y G.722 utiliza banda ancha. Dado que las señales de voz se encuentran muestreadas a 16 kHz, un paso previo a antes de realizar el test PESQ sobre el codec G.726 será un remuestreo de las señales de voz a 8 kHz.

- **Base de datos ALBAYZIN**: La base de datos ALBAYZIN [106] fue desarrollada dentro del proyecto del mismo nombre y financiado por la Comisión Interministerial de Ciencia y Tecnología (TIC91-1488-C06). La base de datos

se finalizó en 1998 con la participación de 6 grupos de investigación de toda España. Esta base de datos consiste en un corpus en español que ha sido diseñado con el objetivo de contribuir al desarrollo y evaluación de sistemas de reconocimiento y procesado del habla. El corpus está compuesto de 15600 frases que fueron realizadas por 304 locutores, mitad hombres y mitad mujeres.

Esta base de datos ha sido considerada en esta tesis para la realización de los tests MUSHRA. De esta forma se puede comprobar el rendimiento de nuestras propuestas sobre frases que están en el lenguaje nativo de los oyentes. Además, dado el elevado coste en recursos que se requiere para el desarrollo del test MUSHRA, se seleccionarán 10 frases fonéticamente balanceadas de esta base de datos con una duración media de 6 segundos.

3.5.3. Simulación de canales de transmisión

Como ya se especificó en la secciones 3.2 y 3.3, las redes WLAN y redes IP tienen características diferentes, de modo que requieren de modelos diferentes para simular los errores en el canal. De este modo, para poder simular el desvanecimiento en las redes WLAN, se ha escogido el modelo de canal Rayleigh y para simular la pérdida de paquetes en las redes IP, se ha escogido el modelo de canal Gilbert de dos estados. A continuación se detalla la configuración empleada para implementar sendos canales:

- **Canal Rayleigh**: Para simular el canal de una red WLAN, se ha utilizado el modelo Rayleigh que genera una secuencia de números reales, cuya longitud será el número de tramas en las que se dividirá la señal de voz a transmitir. Así, cada valor representará la tasa de error por bit o *Bit Error Rate* (BER) (ver figura 3.1) que afectará a toda la trama. Este valor BER se obtiene de acuerdo a la distribución Rayleigh, determinada a partir de la velocidad del receptor definida en (3.2) con $v_r \in 0{,}3, 3m/s$, la frecuencia de la señal portadora (se ha escogido la frecuencia $f_c = 1930MHz$ dentro del rango de Europa), y la razón señal-ruido (SNR) de la densidad espectral de potencia o *signal to noise power spectral density* (E_b/N_0), que tomará valores en el intervalo $[0, 30]$ dB. Por lo tanto, asumiendo una distribución Rayleigh y una modulación BFSK, la tasa de error por bit (BER) se define como [107, 108]:

$$\mathrm{BER} = \frac{1}{2} \cdot \mathrm{erfc}\Big(\sqrt{\alpha^2 \frac{E_b}{2N_0}}\Big) \tag{3.13}$$

donde α se denomina el factor de desvanecimiento que se calcula para cada trama. De este modo, la trama resultante tras pasar por el canal presentará un BER promedio.

- **Canal Gilbert**: Para simular el canal de una red IP, se ha utilizado el modelo Gilbert de dos estados que obtiene una secuencia de 0 y 1 que determina si la trama se recibe o se pierde respectivamente. Esta secuencia se obtendrá por cada condición de canal donde se establecerá tanto la tasa de pérdida de paquetes o *Packet Loss Rate* (PER $= \{10, 20, 30, 40, 50\}$) y la longitud promedio de ráfaga o *Average Burst Length* (ABL $= \{1, 2, 4, 8, 12\}$) determinados a partir de los parámetros p y q considerados en el modelo. Para el modelo con dos estados estos parámetros se definen como:

$$q = \frac{1}{\text{ABL}}$$

$$p = \frac{\text{PER}}{\text{ABL}(1 - \text{PER})}$$

Cabe destacar que algunos valores de tasa de pérdida de paquetes (como 40 % o 50 %) y longitudes promedio de ráfaga (como 8 y 12) son condiciones improbables o inviables en las transmisiones reales. Sin embargo, se están incluyendo en la evaluación ya que para simular un número significativo de ráfagas largas, es necesaria una tasa de pérdida de paquetes alta.

Capítulo 4

Técnica de mitigación de errores sobre redes de ámbito local

En este capítulo se tratará el problema de la modificación en la codificación del paquete recibido, como consecuencia del efecto multitrayecto, en las transmisiones con errores en entornos domésticos o industriales. Para ello, se presentará una técnica de mitigación de errores basada en el receptor que considerando el paquete modificado y el conocimiento previo del comportamiento del canal, es capaz de estimar las componentes del paquete modificado.

La técnica propuesta, denominada decodificación por decisiones soft o (*soft-decision decoding*), ya fue presentada en [108] sobre el codec G.726 [24] trabajando a una tasa de bits de 32 kbps. No obstante, en esta tesis se desarrollará esta técnica para el resto de modos de funcionamiento del codec G.726 (16, 24 y 40 kbps) y se adaptará para el codec G.722 [22], ambos basados en la codificación ADPCM [20].

4.1. Introducción

Con el desarrollo de las tecnologías inalámbricas en el ámbito local o *Wireless Local Area Network* (WLAN), cada vez son más empresas y particulares los que realizan las comunicaciones de voz a través de dispositivos basados en el estándar DECT [47]. Gracias a este estándar, se ha conseguido minimizar el coste de implantación de redes en empresas para comunicaciones de voz y así facilitar la movilidad del usuario a través de zonas más amplias. Además, su tecnología presenta múltiples configuraciones de acceso e incorpora cada día más servicios y aplicaciones que mejoran tanto la seguridad como la calidad del servicio. Por ejemplo, en la nueva

generación DECT, conocida por las siglas en inglés como *NG-DECT* [109], se ha incorporado el estándar G.722 como codec de banda ancha para que la señal de voz sintetizada sea más natural. Es tal el éxito de este estándar, que supone hasta un 73 % del mercado para las comunicaciones en el ámbito doméstico o empresarial [46].

En las transmisiones de voz sobre el estándar DECT, la señal de voz se divide en paquetes o tramas de 10 ms que son codificados de acuerdo con el codec utilizado. De este modo, para transmisiones en banda estrecha, donde se utiliza el codec G.726 [24], la señal de voz es muestreada a 8 KHz, dividida en tramas de 10 ms (cada paquete tendrá 80 componentes) y codificada mediante el codificador ADPCM. Con la nueva generación *NG-DECT*, donde se utiliza el codec G.722 [22], la señal de voz es muestreada a 16 KHz, dividida en tramas de 10 ms (cada paquete tendrá 160 componentes) y codificada mediante el codificador ADPCM por subbandas. El codec G.726 dispone de varios modos de funcionamiento en banda estrecha (16, 24, 32 y 40 kbps), donde el modo utilizado tradicionalmente en las comunicaciones de voz es el de 32 kbps, aunque el resto son completamente funcionales. Del mismo modo, el codec G.722 dispone de tres modos de funcionamiento para banda ancha (48, 56 y 64 kbps), pero sólo el modo de 64 kbps está siendo utilizado para comunicaciones de voz, dejando el resto para la realización de tests experimentales [109].

Sin embargo, las comunicaciones en entornos WLAN presentan el inconveniente de que pueden producirse degradaciones en los paquetes enviados debido al efecto multitrayecto durante la transmisión. Además, dada la dependencia que se genera entre las componentes codificadas durante la codificación ADPCM, si un error modifica una componente, éste error afectará tanto a su decodificación como a la de las componentes futuras como consecuencia de la propagación del error. Por este motivo, es necesario el desarrollo de algoritmos de mitigación que ayuden a reducir el impacto de esta degradación.

4.2. Mitigación de errores en el estándar DECT

El estándar DECT dependiendo de si la transmisión es de banda estrecha o banda ancha emplea un codec diferente, los codecs G.726 y G.722 respectivamente. Ambos codecs se caracterizan por emplear una codificación ADPCM, comentado en la sección 2.4.1, y cuyo esquema de funcionamiento interno se presenta en la figura 4.1. Se puede apreciar, que si la componente de entrada al decodificador es modificada, la señal reconstruida será diferente a la señal de entrada PCM original. Además, de-

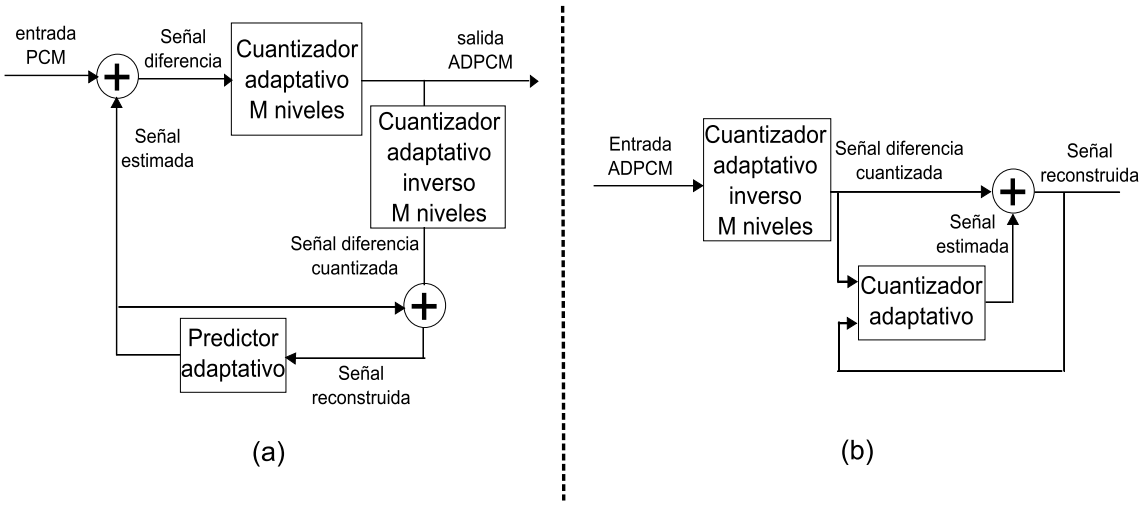

Figura 4.1: Esquema general para el codificador (a) y el decodificador (b) para la técnica de codificación ADPCM.

bido al funcionamiento interno del codificador ADPCM, esta degradación no afecta sólo a la componente modificada, ya que también provocará una desincronización entre codificador y decodificador, dando lugar a una propagación del error sobre las siguientes componentes aunque éstas no fueran modificadas.

Para mitigar esta degradación, en [110, 111] se plantea la estimación de la componente original codificada mediante el cálculo de la probabilidad a posteriori. Sin embargo, no obtienen una probabilidad a posteriori que represente correctamente el estado actual del canal sino que se basan en simplificaciones. Como alternativa, en esta tesis se plantea la técnica presentada en la figura 4.2, denominada *soft-decision decoding*, que sí tiene en cuenta la información del canal. Es decir, esta técnica considera tanto la tasa de error por bit o *bit error rate* (BER) que ha afectado a toda la trama, como la probabilidad de cada componente en la trama recibida a la hora de calcular la probabilidad a posteriori y con ella estimar los parámetros necesarios durante la decodificación ADPCM [95].

De este modo, incorporando el conocimiento previo del canal, según la tasa de error por bit (BER), la probabilidad a posteriori se puede obtener desarrollando el teorema de Bayes a partir del conocimiento de la probabilidad de observación entre la componente codificada original $\mathbf{I}^{(i)}$, de acuerdo a una entrada de diccionario i, y la componente codificada recibida $\tilde{\mathbf{I}}$, que denotaremos como $P(\tilde{\mathbf{I}}(t)|\mathbf{I}^{(i)})$, y una probabilidad a priori de cada componente codificada, que denotaremos como

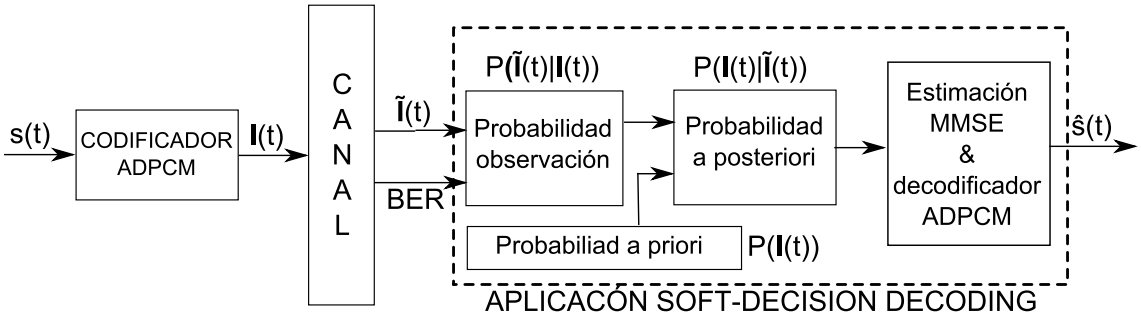

Figura 4.2: Esquema propuesto de decodificador basado en la técnica *soft-decision decoding*. Esta técnica considera la tasa de error por bit (BER) que afecta a la trama enviada para obtener la probabilidad a posteriori y estimar, mediante una estimación MMSE, los parámetros degradados en el decodificador ADPCM.

$P(\mathbf{I}(t) = \mathbf{I}^{(i)})$. Así se define la probabilidad a posteriori $P(\mathbf{I}^{(i)}|\tilde{\mathbf{I}}(t))$ como:

$$P(\mathbf{I}^{(i)}|\tilde{\mathbf{I}}(t)) = \frac{P(\tilde{\mathbf{I}}(t)|\mathbf{I}^{(i)})P(\mathbf{I}(t) = \mathbf{I}^{(i)})}{\sum_{l=1}^{C} P(\tilde{\mathbf{I}}(t)|\mathbf{I}^{(l)})P(\mathbf{I}(t) = \mathbf{I}^{(l)})} \tag{4.1}$$

donde C es el número de entradas del diccionario empleado para realizar la codificación de la componente $\mathbf{I}(t)$ en el instante de tiempo t. Queda por tanto definir la probabilidad de observación $P(\tilde{\mathbf{I}}(t)|\mathbf{I}^{(i)})$.

Para obtener la probabilidad de observación, se va a suponer que la trama enviada se verá afectada con una tasa de error por bit (BER) promedio. De esta manera, la probabilidad de cambio en la codificación de la componente $\mathbf{I}(t)$, que denotaremos como $I(t)_m$ para el bit m en la componente, a la recibida $\tilde{\mathbf{I}}(t)$, que denotaremos como $\tilde{I}(t)_m$, se obtiene como:

$$P(\tilde{I}(t)_m|I(t)_m = I_m^{(i)}) = \begin{cases} 1 - \text{BER}, \text{ si } \tilde{I}(t)_m = I_m^{(i)}, \\ \text{BER}, \text{ en otro caso} \end{cases} \tag{4.2}$$

De forma general, considerando toda la componente y no sólo bit a bit, se obtiene la siguiente probabilidad de observación:

$$P(\tilde{\mathbf{I}}(t)|\mathbf{I}^{(i)}) = \prod_{m=1}^{M} P(\tilde{I}(t)_m|I_m^{(i)}) \tag{4.3}$$

donde M es el número de bits empleado en la codificación de la componente $\mathbf{I}$, siendo $2^M = C$ el tamaño del diccionario empleado.

Una vez definido el cálculo de la probabilidad a posteriori, la estimación de la

señal decodificada $\hat{s}(t)$ vendrá dada de acuerdo a una estimación MMSE, definida en la sección 3.4.2, como:

$$\hat{s}(t) = \sum_{i=1}^{C} s^{(i)} P(\mathbf{I}^{(i)}|\tilde{\mathbf{I}}(t)) \tag{4.4}$$

donde $s^{(i)}$ es el valor decodificado de la componente $\mathbf{I}^{(i)}$ en el diccionario de tamaño C.

De esta manera se incluye en la probabilidad de observación las características del canal para realizar la estimación de la componente de voz reconstruida $\hat{s}(t)$. Así, si la trama no ha sido modificada, BER= 0, la probabilidad a posteriori será 1 y la componente recibida $\tilde{\mathbf{I}}(t)$ coincidirá con la componente enviada $\mathbf{I}(t)$. Si por el contrario, se considera el peor caso de transmisión, con BER=0.5, entonces la probabilidad a posteriori será igual a la probabilidad a priori y la estimación $\hat{s}(t)$ será un valor promedio [95].

Hay que tener en cuenta que la expresión (4.4) sólo está considerando la componente recibida $\tilde{\mathbf{I}}(t)$ para realizar la estimación. De este modo, se está desarrollando una estimación con probabilidad condicional de orden 0, aunque, como se explicó en la sección 3.4.2, la probabilidad condicional en la expresión (4.4) se puede extender para órdenes superiores sin dificultad [84]. No obstante, al estar realizando la estimación componente a componente, un estimador de orden superior acabaría por considerar las componentes ya estimadas para generar las siguientes lo que podría dar lugar a malas estimaciones [95]. Por este motivo, tanto para el codec G.726 [24] como el codec G.722 [22] se utiliza el estimador de orden 0.

4.3. Mitigación de errores sobre el codec G.726

La técnica *soft-decision decoding* ya fue aplicada con éxito sobre el codec G.726 en [108]. En este trabajo se observa el potencial que tiene esta técnica a la hora de hacer un codec robusto frente a errores en el canal. Lo que se pretende en esta sección es mostrar cómo se aplica esta técnica sobre el codec G.726, en todos los modos de funcionamiento, para posteriormente extenderlo al funcionamiento del codec G.722.

Para ello, en primer lugar se describe el proceso de codificación y decodificación para el codec G.726, basado en el codificador ADPCM, así como las expresiones utilizadas en su implementación en el estándar [24] y que son necesarias para conocer qué parámetros se ven afectados tras obtener una componente decodificada $\tilde{\mathbf{I}}$

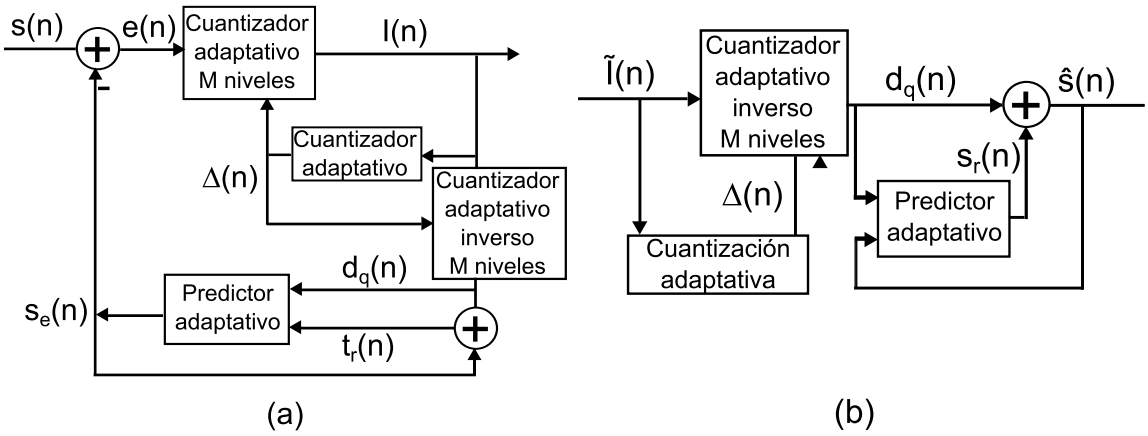

Figura 4.3: Esquemas del funcionamiento interno del codificador y decodificador estándar del codec G.726. A) Esquema del codificador. B) Esquema del decodificador.

diferente a la enviada. Posteriormente, se describe la aplicación de la técnica *soft-decision decoding* sobre el codec G.726 y cómo se estiman los parámetros internos del decodificador a fin de minimizar la degradación producida en la componente recibida $\tilde{\mathbf{I}}$. Finalmente, se presentan los resultados obtenidos de la aplicación de esta técnica sobre los distintos modos de funcionamiento del codec G.726 (16, 24, 32 y 40 kbps) en diferentes condiciones de canal.

4.3.1. Proceso de codificación y decodificación en el codec

La figura 4.3 muestra los esquemas generales de funcionamiento de los procesos de codificación y decodificación del codec G.726 [24]. Como se puede observar, a partir de la muestra $s(n)$ se extrae una señal diferencia $e(n)$. Con este valor de diferencia, el cuantizador adaptativo se encargará de generar la codificación $\mathbf{I}(n)$ de acuerdo con el número de bits indicado por el modo utilizado (16, 24, 32 y 40 kbps) para cada componente. A partir de esta representación se obtendrá un valor estimado de la muestra $s_e(n)$ a través de un predictor adaptativo, cuyos coeficientes se irán actualizando con cada componente codificada.

Si no hay errores en el canal, la componente recibida $\tilde{\mathbf{I}}(n)$ coincidirá con la codificación enviada $\mathbf{I}(n)$ y la decodificación se realizará de manera sincronizada. Sin embargo, si esto no es así, el valor de la diferencia cuantizada, $d_q(n)$, y el factor de escala, $\Delta(n)$, para la componente n no coincidirán con los que se generaron en el codificador y se producirá una desincronización. Como resultado, se producirá una propagación del error que afectará a las siguientes componentes aunque no tuvieran

errores. Por este motivo, es importante minimizar el impacto de las modificaciones sufridas en la componente recibida $\tilde{\mathbf{I}}(n)$ y que afecta directamente a la obtención de los parámetros mencionados ($d_q(n)$ y $\Delta(n)$). A continuación se detalla cómo se obtienen estos parámetros de acuerdo a las expresiones del estándar [24].

El factor de escala $\Delta(n)$ se obtiene siguiendo el principio de adaptación bimodal que establece un factor de escala rápido $y_u(n)$ y un factor de escala lento $y_l(n)$ [24]. El factor de escala rápido se utiliza para proporcionar valores de diferencia cuantizada con grandes fluctuaciones y se calcula recursivamente en el dominio logarítmico (en base dos) a partir del factor de escala logarítmico resultante, $\Delta(n)$, como:

$$y_u(n) = (1 - 2^{-5})\Delta(n) + 2^{-5}W[\tilde{\mathbf{I}}(n)] \tag{4.5}$$

donde $W[\cdot]$ es una función discreta tabulada en [24]. Como resultado, esta variable puede tomar un valor entre $1,06 \leq y_u(n) \leq 10,00$.

Por otro lado, el factor de escala de adaptación lento se encarga de proporcionar una valor de la diferencia cuantizada con fluctuaciones pequeñas y que se calcula aplicando un filtro paso bajo sobre $y_u(n)$ de la forma:

$$y_l(n) = (1 - 2^{-6})y_l(n-1) + 2^{-6}y_u(n) \tag{4.6}$$

A partir de las expresiones (4.5) y (4.6) se puede obtener el factor de escala $\Delta(n)$ como una combinación pesada por un parámetro de control $a_l \in [0, 1]$ de la siguiente forma:

$$\Delta(n) = a_l(n)y_u(n-1) + (1 - a_l(n))y_l(n-1) \tag{4.7}$$

donde $a_l(n)$, definido en el estándar G.726 [24], se actualiza conforme a los coeficientes del predictor adaptativo de la figura 4.3 y que está constituido por un filtro con 6 polos y 2 ceros.

Este factor de escala se utiliza para obtener la diferencia cuantizada $d_q(n)$ como [24]:

$$d_q(n) = \text{sign}(\tilde{\mathbf{I}}(n))2^{QM^{-1}[\tilde{\mathbf{I}}(n)]+\Delta(n)} \tag{4.8}$$

donde $\text{sign}(\cdot)$ determina el signo indicado por el primer bit de la componente recibida $\tilde{\mathbf{I}}(n)$ y $QM^{-1}[\cdot]$ es el valor correspondiente a los M bits de su codificación.

Dado que estos dos parámetros son importantes para mantener la sincronización entre codificador y decodificador, en la siguiente sección se explica cómo se ha introducido la técnica *soft-decision decoding* para minimizar el efecto multitrayecto.

4.3.2. Aplicación de la técnica *soft-decision decoding*

Para aplicar la técnica *soft-decision decoding*, con el esquema propuesto en la figura 4.2, sobre todos los modos de funcionamiento del codec G.726, se ha tomado como punto de partida el trabajo [108] que se aplica sobre el modo de funcionamiento de 32 kpbs. De este modo, se ha seguido el mismo procedimiento para estimar los correspondientes parámetros de la diferencia cuantizada $\hat{d}_q(n)$ y el factor de escala $\hat{\Delta}(n)$ a partir del conocimiento previo del error en el canal (BER) que afecta a toda la trama enviada.

Partiendo de la expresión (4.4), la estimación del parámetro de la diferencia cuantizada $\hat{d}_q(n)$ se obtiene como [108]:

$$\hat{d}_q(n) = \sum_{i=0}^{2^w-1} \Big(\text{sign}(\mathbf{I}^{(i)}) 2^{QM^{-1}[\mathbf{I}^{(i)}]+\Delta(n)}\Big) P(\mathbf{I}^{(i)}|\tilde{\mathbf{I}}(n)) \tag{4.9}$$

Aunque la aplicación de la probabilidad a posteriori en la expresión (4.9) puede reducir el error respecto al valor original $d_q(n)$, hay que tener en cuenta que en su estimación interviene el factor de escala $\Delta(n)$ y que hay que estimar previamente. Sin embargo, la estimación de este factor de escala no será como en la expresión (4.4), ya que su obtención depende de los factores de adaptación rápida $y_u(n-1)$ y adaptación lenta $y_l(n-1)$ previos a la componente actual n como se puede ver en la expresión (4.7). No obstante, asumiendo que se puede almacenar la probabilidad a posteriori de la componente anterior, la estimación del factor de escala se puede obtener como [108]:

$$\hat{\Delta}(n) = log_2\Big[\sum_{i=0}^{2^w-1} 2^{\Delta^{(i)}(n)} \cdot P(\mathbf{I}^{(i)}|\tilde{\mathbf{I}}(n-1))\Big] \tag{4.10}$$

donde $\Delta^{(i)}(n)$ se calcula como (4.7) a partir de los distintos valores de $y_u^{(i)}(n-1)$ y $y_l^{(i)}(n-1)$ obtenidos como:

$$\begin{aligned} y_u^{(i)}(n-1) &= (1-2^{-5})\hat{\Delta}(n-1) + 2^{-5}W[\mathbf{I}^{(i)}] \\ y_l^{(i)}(n-1) &= (1-2^{-6})y_l(n-2) + 2^{-6}y_u^{(i)}(n-1) \end{aligned} \tag{4.11}$$

donde $y_l(n-2)$ tiene que ser inicializado a 1.06 y posteriormente actualizarse para cada nueva componente decodificada [24, 108].

Una vez determinado el factor de escala óptimo para la componente n, éste se

utiliza posteriormente en la estimación del parámetro diferencia cuantizada $\hat{d}_q(n)$, en la expresión (4.9), y para la actualización de los factores de escala $y_u(n-1)$ e $y_l(n-1)$, en (4.11). Para ello, $y_u(n-1)$ se puede obtener a partir del factor de escala $\hat{\Delta}(n)$ actual y no del anterior despejando $y_u(n-1)$ de (4.7) y sustituyendo $y_l(n-1)$ de la expresión (4.6) para obtener $y_u(n-1)$ como [108]:

$$y_u(n-1) = \frac{\hat{\Delta}(n) - (1-a_l(n))(1-2^{-6})y_l(n-2)}{a_l(n) + (1-a_l(n))2^{-6}} \tag{4.12}$$

Por último, la formulación presentada se puede generalizar al resto de modos de funcionamiento considerando los correspondientes diccionarios, en este caso W y QM^{-1}, empleados para la obtención de los parámetros de la diferencia cuantizada y el factor de escala en el modo correspondiente.

4.3.3. Resultados experimentales

Como se ha indicado al comienzo de esta sección, la técnica *soft-decision decoding* se ha aplicado sobre todos los modos de trabajo del codec G.726, por lo que hay que realizar la estimación de los parámetros diferencia cuantizada d_q y factor de escala Δ para cada uno de ellos. Para estimar estos parámetros, hay que calcular la probabilidad a posteriori $P(\mathbf{I}^{(i)}|\tilde{\mathbf{I}}(n))$, para la cual se necesita la probabilidad a priori $P(\mathbf{I}^{(i)})$ y la probabilidad de observación $P(\tilde{\mathbf{I}}(n)|\mathbf{I}^{(i)})$.

La probabilidad a priori se ha obtenido a partir de las componentes codificadas $\mathbf{I}(n)$ de las frases de voz del conjunto de entrenamiento de la base de datos NTT [105] para cada uno de los modos de funcionamiento del codec G.726. Para el cálculo de la probabilidad de observación, la tasa de error por bit (BER) se ha obtenido simulando un canal Rayleigh, como se ha comentado en la sección 3.5.3, de acuerdo al valor de relación señal-ruido o *Signal to Noise Ratio* (SNR), que denotaremos como E_b/N_0, y la velocidad del receptor v.

Para analizar el rendimiento de la técnica *soft-decision decoding*, que denotaremos como SD, se han comparado los resultados PESQ [103] promedio obtenidos con el conjunto de test de la base de datos NTT (con los idiomas inglés americano y británico, chino, francés, alemán y español), frente a los obtenidos mediante la decodificación que realiza el propio codec o *hard-decision decoding*, que denotaremos como HD, y un algoritmo sencillo de mitigación de errores (PLC) descrito en [108]. Dado que el codec G.726 no tiene ningún algoritmo PLC implementado, este algoritmo PLC, que denotaremos como HDRep, aplica una repetición de la trama anterior

cuando se supera un BER de 0,2 %. Los diferentes valores BER promedio que afectan a las tramas se obtienen tras aplicar un canal Rayleigh con diferente relación señal-ruido ($E_b/N_0 \in [0-30]$ dB) y diferente velocidad de usuario ($v = \{0{,}3, 3\}$ m/s).

Los resultados de este test se pueden observar en las figuras 4.4 y 4.5, donde se aprecia que el algoritmo PLC (HDRep) mejora sensiblemene la calidad de la decodificación estándar HD. Del mismo modo, también se aprecia una mejora notable en los resultados obtenidos mediante la técnica SD sobre la decodificación estándar HD e incluso sobre el algoritmo HDRep.

Si se aplica el mismo procedimiento sobre el resto de modos de funcionamiento, también se puede observar en las figuras 4.6 y 4.7 un rendimiento similar de la técnica SD sobre la decodificación de estándar HD. Los valores PESQ promedio obtenidos para generar estas figuras se encuentran en la tabla 4.1 y se han obtenido con el mismo entorno de trabajo presentado anteriormente. A la vista de los resultados, se confirma el buen rendimiento que ofrece la técnica *soft-decision decoding* sobre el codec G.726. De este modo, el esquema propuesto permite hacer al codec más robusto frente a errores en el canal.

4.4. Mitigación de errores sobre el codec G.722

La técnica *soft-decision decoding* se puede aplicar también sobre el codec G.722 al tratarse de un codec que utiliza el mismo sistema de codificación, la codificación ADPCM. La única diferencia es que este codec realiza la codificación ADPCM por subbandas (ADPCM-SB), como ya se explicó en la sección 2.4.1.

De este modo, en esta sección se describe brevemente el esquema del codificador y decodificador del estándar G.722 [22], para posteriormente, presentar las expresiones que generan los correspondientes parámetros de diferencia cuantizada y factor de escala definidos en el estándar sobre cada subbanda. A continuación, se presentan las correspondientes expresiones que estiman estos parámetros, siguiendo un procedimiento similar al comentado en el codec G.726, y finalmente, se muestran los resultados obtenidos durante la aplicación de la técnica *soft-decision decoding* sobre una o ambas subbandas respecto a la decodificación *hard-decision decoding*.

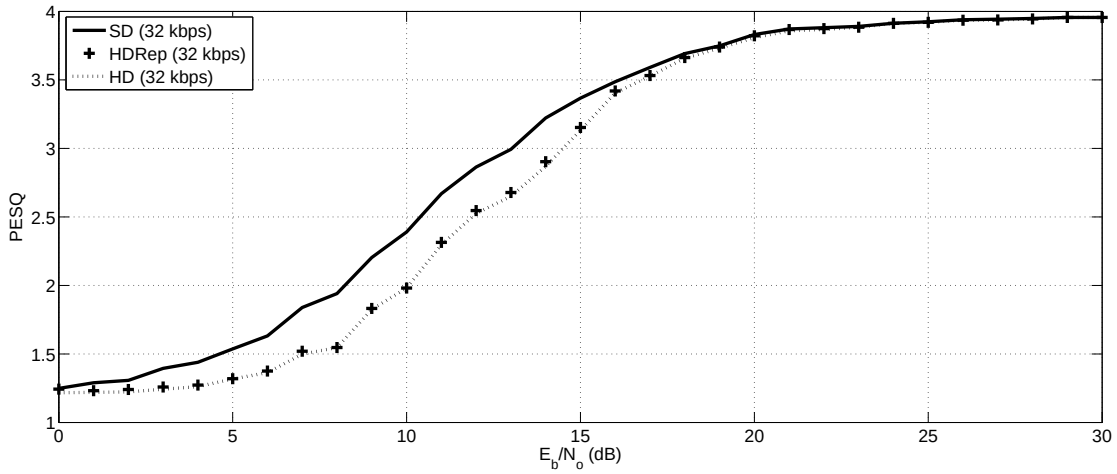

Figura 4.4: Evaluación de calidad objetiva PESQ promedio sobre diferentes valores de E_b/N_0 con una velocidad de usuario de 0.3 m/s. Los resultados obtenidos muestran el rendimiento de la técnica *soft-decision decoding* (SD) sobre la decodificación original o *hard-decision decoding* (HD) y el algoritmo PLC de repetición de tramas (HDRep) sobre el codec G.726 (en el modo 32 kbps).

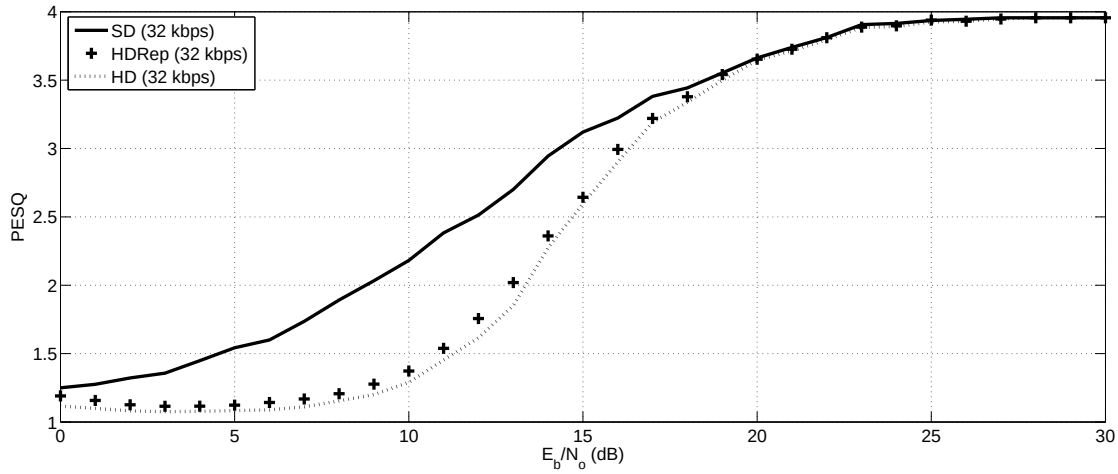

Figura 4.5: Evaluación de calidad objetiva PESQ promedio sobre diferentes valores de E_b/N_0 con una velocidad de usuario de 3 m/s. Los resultados obtenidos muestran el rendimiento de la técnica *soft-decision decoding* (SD) sobre la decodificación original o *hard-decision decoding* (HD) y el algoritmo PLC de repetición de tramas (HDRep) sobre el codec G.726 (en el modo 32 kbps).

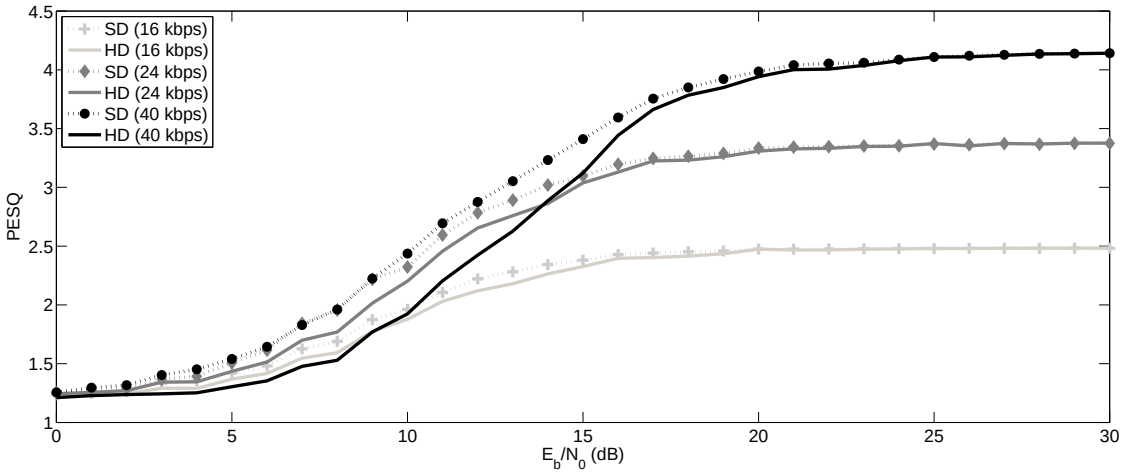

Figura 4.6: Evaluación de calidad objetiva PESQ promedio sobre diferentes valores de E_b/N_0 y velocidad de usuario de 0.3 m/s para analizar el rendimiento de la técnica *soft-decision decoding* (SD) frente a la decodificación del codec original o *hard-decision decoding* (HD) sobre el resto de modos de funcionamiento del codec G.726 (16, 24 y 40) kbps.

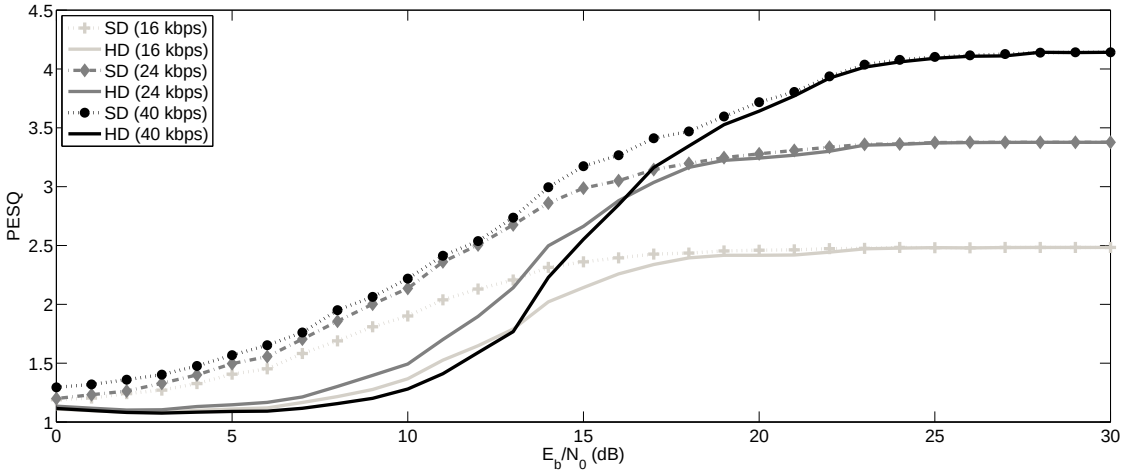

Figura 4.7: Evaluación de calidad objetiva PESQ promedio sobre diferentes valores de E_b/N_0 y velocidad de usuario de 3 m/s para analizar el rendimiento de la técnica *soft-decision decoding* (SD) frente a la decodificación del codec original o *hard-decision decoding* (HD) sobre el resto de modos de funcionamiento del codec G.726 (16, 24 y 40) kbps.

			E_b/N_0(dB)						
	tasa bits (kbps)	v	0	5	10	15	20	25	30
HD	16	0.3	1.23	1.37	1.88	2.33	2.47	2.48	2.48
		3	1.11	1.11	1.36	2.14	2.42	2.48	2.48
	24	0.3	1.24	1.43	2.20	3.04	3.31	3.37	3.38
		3	1.13	1.17	1.49	2.66	3.24	3.37	3.38
	32	0.3	1.22	1.32	1.98	3.13	3.81	3.92	3.96
		3	1.10	1.16	1.29	2.59	3.65	3.93	3.96
	40	0.3	1.21	1.30	1.92	3.12	3.94	4.11	4.14
		3	1.08	1.14	1.28	2.55	3.64	4.09	4.14
SD	16	0.3	1.24	1.42	1.96	2.38	2.47	2.48	2.48
		3	1.18	1.41	1.90	2.36	2.46	2.48	2.48
	24	0.3	1.25	1.51	2.32	3.10	3.33	3.37	3.37
		3	1.20	1.50	2.14	2.99	3.28	3.37	3.38
	32	0.3	1.25	1.54	2.39	3.37	3.83	3.92	3.96
		3	1.25	1.54	2.18	3.12	3.66	3.94	3.96
	40	0.3	1.26	1.54	2.44	3.41	3.99	4.10	4.14
		3	1.29	1.57	2.22	3.17	3.72	4.10	4.14

Tabla 4.1: Resultados PESQ obtenidos como promedio en diferentes valores de SNR (E_b/N_0) y velocidad de usuario (v) sobre el conjunto de test de la base de datos NTT al evaluar el rendimiento de la técnica *soft-decision decoding* (SD) y la decodificación *hard-decision decoding* (HD) sobre los distintos modos de funcionamiento del codec G.726 (16, 24, 32 y 40 kbps).

4.4.1. Proceso de codificación y decodificación

En la figura 4.8 se puede ver el esquema general del funcionamiento del codec G.722 [22], correspondiente a una codificación ADPCM por subbandas (ADPCM-SB). Partiendo de una trama de tamaño N, muestreada a 16 kHz, un filtro espejo en cuadratura o *Quadrature Mirror Filter* (QMF), de 23 coeficientes, divide la trama en sus componentes en alta frecuencia (4000-8000 Hz) $\mathbf{x}_H$ y sus componentes en baja frecuencia (0-4000 Hz) $\mathbf{x}_L$. Así se obtienen dos segmentos, ambos con la mitad de tamaño de la trama original ($m \in \{0, 1, ..., , \frac{N}{2} - 1\}$), a los que se aplicará una codificación ADPCM independiente.

En el caso particular del codec G.722 en el modo de 64 kbps, cada componente $\mathbf{I}_L(m)$ del segmento de frecuencias bajas se codificará con 6 bits y cada componente $\mathbf{I}_H(m)$ del segmento de frecuencias alta se codificará con 2 bits. Así, se obtiene una codificación de 8 bits por componente $\mathbf{I}(m)$ para tener una tasa de bits final de 64 kbps. La diferencia en los distintos modos de funcionamiento del codec G.722 (48, 56 y 64 kbps), radica en la codificación empleada para la componente de bajas

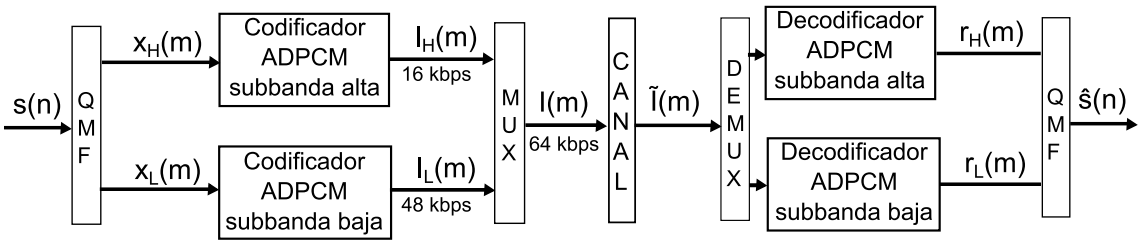

Figura 4.8: Esquema del funcionamiento general del codec G.722 en las etapas de codificación y decodificación.

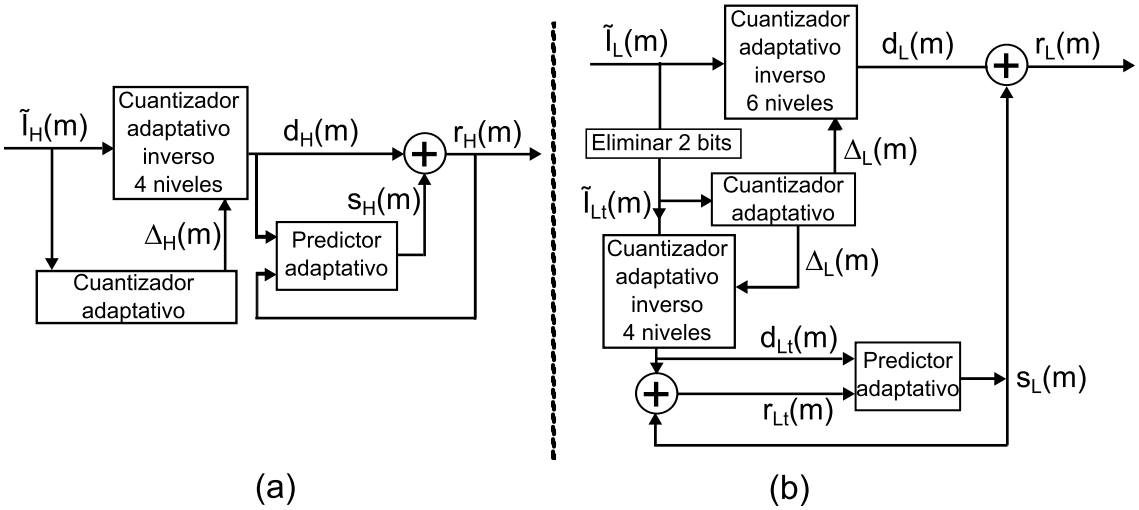

Figura 4.9: Esquema del funcionamiento interno del decodificador del codec G.722 dividido en subbandas de altas frecuencias (a) y bajas frecuencias (b).

frecuencias. No obstante, el modo de 64 kbps es el que se utiliza actualmente en las comunicaciones telefónicas sobre la tecnología DECT.

Siguiendo un procedimiento similar al aplicado sobre el codec G.726, en la figura 4.9 se puede observar que la componente recibida $\tilde{\mathbf{I}}(m)$ puede afectar a una o ambas subbandas y en consecuencia, modificar los correspondientes parámetros de diferencia cuantizada $d_H(m)$ (en la figura 4.9 a) y ($d_{Lt}(m)$, $d_L(m)$ (en la figura 4.9 b) y el correspondiente factor de escala $\Delta_H(m)$ y $\Delta_L(m)$ respectivamente. Estos parámetros se definen en el estándar G.722 [22] como:

- Diferencia cuantizada:

$$\begin{aligned} d_H(m) &= Q2^{-1}[\tilde{\mathbf{I}}_H(m)] \cdot \Delta_H(m) \cdot \mathrm{sgn}(\tilde{\mathbf{I}}_H(m)), \\ d_L(m) &= Q6^{-1}[\tilde{\mathbf{I}}_L(m)] \cdot \Delta_L(m) \cdot \mathrm{sgn}(\tilde{\mathbf{I}}_L(m)), \\ d_{Lt}(m) &= Q4^{-1}[\tilde{\mathbf{I}}_{Lt}(m)] \cdot \Delta_L(m) \cdot \mathrm{sgn}(\tilde{\mathbf{I}}_{Lt}(m)) \end{aligned} \tag{4.13}$$

- Factor de escala:

$$\begin{aligned} \Delta_H(m) &= 2^{[\bigtriangledown_H(m)]} \cdot \Delta_{\min}, \\ \Delta_L(m) &= 2^{[\bigtriangledown_L(m)+2]} \cdot \Delta_{\min}, \\ \text{con } \bigtriangledown_R(m) &= \beta \bigtriangledown_R(m-1) + W_R[\tilde{\mathbf{I}}_R(m-1)] \end{aligned} \tag{4.14}$$

donde QM^{-1} es el cuantizador inverso de 2^M entradas, $\tilde{\mathbf{I}}_{Lt}(m)$ es la codificación $\tilde{\mathbf{I}}_L(m)$ a la que se le han quitado los dos bits menos significativos, W_R es el factor de escala en el dominio logarítmico para cada subbanda ($R \in \{L, H\}$), $\text{sgn}(\tilde{\mathbf{I}}_R(m))$ es el signo de la codificación recibida $\tilde{\mathbf{I}}_R(m)$ y β y $\Delta_{\min}$ son constantes definidas en el estándar.

4.4.2. Aplicación de la técnica *soft-decision decoding*

Las expresiones definidas anteriormente serán la base para poder aplicar la técnica *soft-decision decoding* de manera análoga a como se hizo sobre el codec G.726, en este caso sobre cada subbanda. De este modo, sobre cada una de las subbandas se calculará la estimación de la diferencia cuantizada y el factor de escala siguiendo la expresión (4.4) y la probabilidad a posteriori definida en la expresión (4.1).

De nuevo, observando las expresiones anteriores, se realizará la estimación del factor de escala en primer lugar. De forma general, el cálculo del factor de escala para cada subbanda, que denotaremos como $\widehat{\Delta}_R(m)$ donde $R \in \{H, L\}$, se obtiene mediante una estimación MMSE como:

$$\begin{aligned} &\widehat{\Delta}_R(m) = \sum_{j=0}^{2^M-1} \left(\Delta_R(m)^{(j)} P(\mathbf{I}^{(j)} | \tilde{\mathbf{I}}(m-1) \right) = \\ &2^K \Delta_{\min} \sum_{j=0}^{2^M-1} \left(\left(2^{\beta \bigtriangledown_R(m-1) + W_R[\mathbf{I}^{(j)}]} \right) P(\mathbf{I}^{(j)} | \tilde{\mathbf{I}}(m-1)) \right), \end{aligned} \tag{4.15}$$

donde $K = \{0, 2\}$ se corresponde con la constante en el exponente de la correspondiente expresión en (4.14), $\mathbf{I}^{(j)}$ es la codificación de acuerdo con el índice j, $W_R[\mathbf{I}^{(j)}]$ es el factor de escala en el dominio logarítmico y $M = \{2, 4\}$ es el número de bits por subbanda en el modo de funcionamiento de 64 kbps del estándar G.722 [22]. Al igual que en la estimación en la expresión (4.10), de nuevo se hace uso de la probabilidad a posteriori calculada para la estimación anterior, $P(\mathbf{I}^{(j)} | \tilde{\mathbf{I}}(m-1))$, para obtener el factor de escala.

Finalmente, una vez realizada la estimación del factor de escala, pasamos a estimar el parámetro de diferencia cuantizada mediante una estimación MMSE de la siguiente forma:

$$\begin{aligned}
\widehat{d}_H(m) &= \left(\sum_{j=0}^{2^2-1}(Q2^{-1}[\mathbf{I}^{(j)}]\cdot \mathrm{sgn}(\mathbf{I}^{(j)})\cdot P(\mathbf{I}^{(j)}|\tilde{\mathbf{I}}(m))\right)\widehat{\Delta}_H(m);\\
\widehat{d}_L(m) &= \left(\sum_{j=0}^{2^6-1}(Q6^{-1}[\mathbf{I}^{(j)}]\cdot \mathrm{sgn}(\mathbf{I}^{(j)})\cdot P(\mathbf{I}^{(j)}|\tilde{\mathbf{I}}(m))\right)\widehat{\Delta}_L(m);\\
\widehat{d}_{Lt}(m) &= \left(\sum_{j=0}^{2^4-1}(Q4^{-1}[\mathbf{I}^{(j)}]\cdot \mathrm{sgn}(\mathbf{I}^{(j)})\cdot P(\mathbf{I}^{(j)}|\tilde{\mathbf{I}}(m))\right)\widehat{\Delta}_L(m)
\end{aligned} \tag{4.16}$$

4.4.3. Resultados experimentales

Para analizar el rendimiento de la técnica *soft-decision decoding* (SD) sobre el codec G.722 es necesario el cálculo de las probabilidades a posteriori $P(\mathbf{I}^{(i)}|\tilde{\mathbf{I}}(n))$, para las que son necesarias las probabilidades a priori $P(\mathbf{I})$ y las probabilidades de observación $P(\tilde{\mathbf{I}}(n)|\mathbf{I}^{(i)})$.

La probabilidad a priori se ha obtenido a partir de las componentes codificadas $\mathbf{I}(n)$ de las frases de voz del conjunto de entrenamiento de la base de datos NTT [105] para el modo de funcionamiento de 64 kbps del codec G.722. Para el cálculo de la probabilidad de observación, la tasa de error por bit (BER) se ha obtenido simulando un canal Rayleigh, como se ha comentado en la sección 3.5.3, de acuerdo al valor del relación señal-ruido o *Signal to Noise Ratio* (SNR), que denotaremos como E_b/N_0, y la velocidad del receptor v.

De acuerdo con el esquema de codificación ADPCM-SB mostrado en la figura 4.8, cada trama de entrada de 10 ms contiene 160 muestras, pero tras la división en subbandas, en cada una se codificarán los 80 componentes de acuerdo a los bits correspondientes para cada banda. Así, de los 8 bits que codifican cada componente en el modo de 64 kbps, se dedican 6 bits para la codificación de las bajas frecuencias y 2 a la codificación de las altas frecuencias.

Al tratarse de un esquema con dos subbandas bien diferenciadas, se ha decidido desarrollar un conjunto de pruebas donde se analice el rendimiento de la técnica (SD) respecto a la decoficicación realizada por el propio codec G.722 (HD). Por un lado, se analizará el rendimiento de aplicar la técnica SD sobre cada una de las subbandas por separado, que denotaremos como (SD_H y SD_L), y aplicando la decodificación HD en la otra subbanda. Por otro lado, se analizará el rendimiento

		E_b/N_0(dB)						
	v	0	5	10	15	20	25	30
HD	0.3	1.06	1.12	1.55	3.05	3.77	4.01	4.03
	3	1.05	1.07	1.28	2.28	3.64	4.01	4.03
HD_PLC	0.3	1.23	1.14	1.73	3.32	3.90	4.02	4.03
	3	1.10	1.18	1.68	2.74	3.75	4.02	4.03
SD_H	0.3	1.08	1.13	1.56	3.06	3.78	4.01	4.03
	3	1.05	1.08	1.28	2.29	3.65	4.01	4.03
SD_L	0.3	1.02	1.12	1.72	3.43	3.93	4.02	4.03
	3	1.02	1.09	1.74	2.86	3.84	4.02	4.03
SD_LH	0.3	1.05	1.23	1.84	3.47	3.94	4.02	4.03
	3	1.03	1.19	1.90	2.91	3.84	4.02	4.03

Tabla 4.2: Resultados de la medida de evaluación objetiva WB-PESQ promedio para diferentes valores de SNR del canal (E_b/N_0) y velocidades de usuario (v) sobre el conjunto de test de la base de datos NTT al evaluar la aplicación de la técnica *soft-decision decoding* (SD) sobre una de las subbandas (SD_L, SD_H) y sobre ambas (SD_LH), el algoritmo PLC del codec G.722 (PLC) y la decodificación estándar *hard-decision decoding* (HD) para el modo de 64 kbps.

de aplicar la técnica SD sobre ambas subbandas, que denotaremos como(SD_LH). Además, aprovechando que el codec G.722 implementa un algoritmo PLC [42, 109], también se comparará el rendimiento de las propuestas SD con los resultados obtenidos aplicando el algoritmo PLC, que denotaremos como (HD_PLC). Para ello, en estas pruebas se asume que el algoritmo PLC se aplica cuando se tiene una tasa de error por bit (BER) superior al 10 % que es un valor razonable para indicar que la trama está tan dañada como para considerarla perdida [42].

En la tabla 4.2 se pueden observar los resultados PESQ [103] promedio, en este caso la versión para banda ancha (WB-PESQ), sobre las diferentes propuestas mencionadas anteriormente. Para su obtención se ha empleado el conjunto de test de la base de datos NTT (con los idiomas inglés americano y británico, chino, francés, alemán y español) y se ha simulado un canal Rayleigh con diferente relación señal-ruido ($E_b/N_0 \in [0-30]$ dB) y diferente velocidad de usuario ($v = \{0{,}3, 3\}$ m/s).

Como puede observarse, el algoritmo PLC del codec G.722 mejora la calidad perceptual respecto a la decodificación estándar HD, incluso por encima de la propuesta SD_H, ya que sólo se está actuando sobre 2 bits de la codificación total de 8 bits por componente. Por otro lado, cuando se aplica sobre la componente de bajas frecuencias, que es donde se sitúa la mayor parte de la energía de la voz, la propuesta

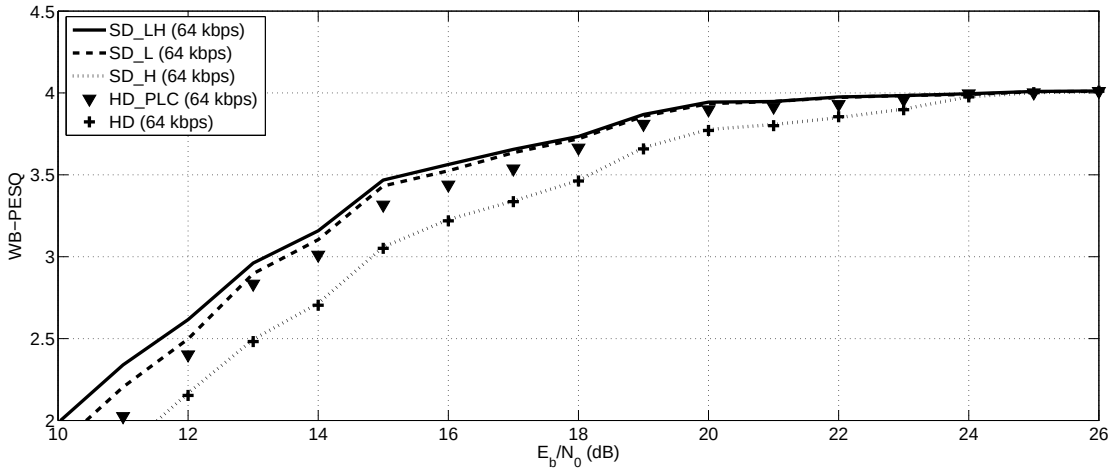

Figura 4.10: Evaluación de calidad objetiva WB-PESQ promedio en las diferentes propuestas que utilizan la técnica *soft-decision decoding* sobre una o ambas bandas (SD_H, SD_L, SD_LH) respecto a la decodificación *hard-decision decoding* (HD) y la aplicación de su propio algoritmo *Packet Loss Concealment* (PLC) con diferentes valores de E_b/N_0 y una velocidad de usuario de 0.3 m/s.

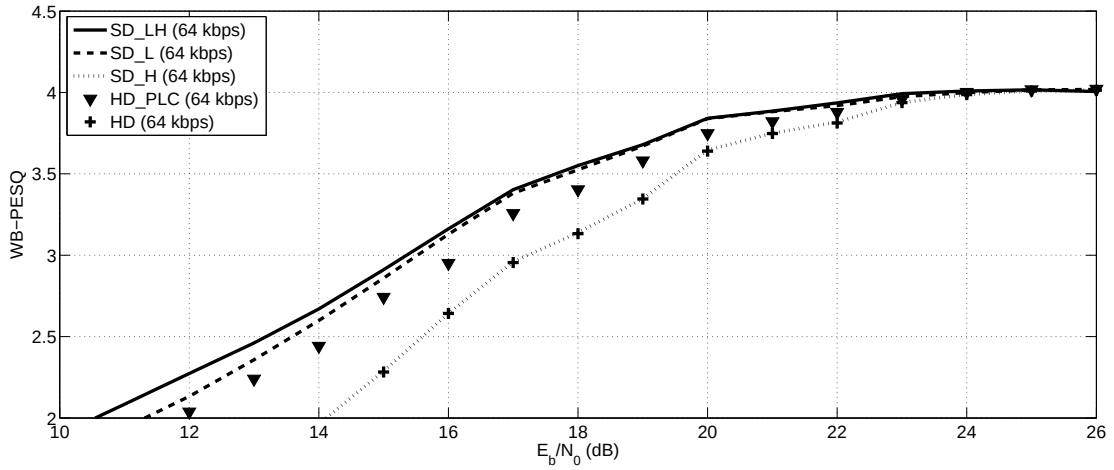

Figura 4.11: Evaluación de calidad objetiva WB-PESQ promedio en las diferentes propuestas que utilizan la técnica *soft-decision decoding* sobre una o ambas bandas (SD_H, SD_L, SD_LH) respecto a la decodificación *hard-decision decoding* (HD) y la aplicación de su propio algoritmo *Packet Loss Concealment* (PLC) con diferentes valores de E_b/N_0 y una velocidad de usuario de 3 m/s.

SD_L ofrece un rendimiento superior llegando a estar por encima de los resultados del propio algoritmo PLC (HD_PLC). Esta mejora es más significativa cuando se aplica esta técnica sobre ambas subbandas (SD_LH) y que puede comprobarse en las figuras 4.10 y 4.11.

Por último, en la tabla 4.2 se puede observar también un rendimiento inferior de las propuestas (SD_L) y (SD_LH) respecto al resultado (HD_PLC) con una SNR de 0 dB. Esto se debe a que en la técnica *soft-decision decoding* se está considerando una tasa de error por bit promedio (BER) que afecta a las componentes de la trama enviada. Por lo tanto, si una componente no fuera modificada, la decodificación (HD) podría dar mejor rendimiento que la proporcionada por la estimación aplicando la técnica *soft-decision decoding* de acuerdo al valor (BER). A pesar de ello, la propuesta (SD_LH) obtiene una notable mejora respecto a la decodificación (HD) y la aplicación de su propio algoritmo PLC (HD_PLC). Por lo tanto, la técnica *soft-decision decoding* hace que el codec G.722 sea más robusto frente a errores en el canal.

Capítulo 5

Técnicas de prevención de pérdidas sobre redes IP

Este capítulo se centrará en el uso de las técnicas de prevención de pérdida de paquetes basadas en el emisor sobre redes IP. En concreto, se emplearán los códigos de corrección de errores hacia delante (FEC) para recuperar el último paquete perdido a partir de los correspondientes parámetros de voz (coeficientes LPC y señal de excitación) codificados en el paquete recibido. Al mismo tiempo, su uso permitirá reducir la propagación del error para los codecs que presenten dependencias inter-trama.

Dado que uno de los parámetros que conforman el código FEC, la señal de excitación, no presenta una adecuada representación para su estimación, en este capítulo se presentarán diferentes algoritmos de cuantización de esta señal. Como resultado, se mejorará la obtención de los diccionarios de cuantización empleados en la minimización del error de síntesis.

Por último, el uso del código FEC incrementará la tasa de bits a transmitir por paquete, dando lugar a que no pueda utilizarse el decodificador del codec estándar, aunque no se produzcan pérdidas durante la transmisión. Para evitar esta incompatibilidad, se empleará una técnica esteganográfica que oculta el código FEC en la propia codificación del paquete.

5.1. Introducción

En la sección 3.3.2 se introdujo el segundo tipo de degradación a tratar en esta tesis y que será estudiado tanto en este capítulo como en el siguiente: la pérdida de

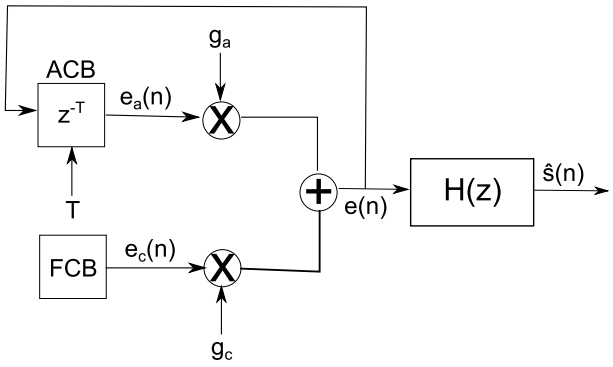

Figura 5.1: Diagrama del proceso de decodificación y generación de la síntesis de voz en los codecs basados en el paradigma CELP.

paquetes en las transmisiones de voz sobre redes IP.

Esta pérdida de paquetes se produce por la congestión en la red que puede provocar que uno o varios paquetes, generalmente de manera consecutiva, no alcancen su destino. Como consecuencia, la señal de voz sintetizada ya no será como la original y provocará una caída en la calidad de servicio. Además, los codecs que presentan dependencias inter-trama, como los codecs basados en el paradigma CELP [13], son más vulnerables ante la pérdida de paquetes ya que se generará una propagación de error hacia los paquetes recibidos correctamente tras la pérdida.

Hay que tener en cuenta que en los codecs basados en el paradigma CELP, la señal de voz sintetizada se obtiene mediante el filtrado de la señal de excitación $e(n)$ a través de un filtro de predicción lineal, $H(z) = 1/A(z)$, que representa el tracto vocal. Esta señal de excitación se obtiene en los codificadores CELP siguiendo el diagrama presentado en la figura 5.1.

En este diagrama, la señal de excitación $e(n)$ es resultado de la suma de un vector adaptativo $\mathbf{e}_a$ y un vector de código $\mathbf{e}_c$, ambos pesados con su correspondiente ganancia g_a y g_c respectivamente. Así, la señal de excitación se obtiene como:

$$e(n) = g_a e_a(n) + g_c e_c(n) \tag{5.1}$$

En el caso del vector adaptativo $\mathbf{e}_a$, éste se escoge de un diccionario adaptativo o *Adaptive Code-Book* (ACB) que tiene como objetivo modelar las correlaciones a largo plazo de la señal de excitación mediante un filtro LTP. De esta forma, la entrada $\mathbf{e}_a$ de este diccionario se construye dinámicamente a partir de las muestras

de la señal de excitación previa como:

$$e_a(n) = \sum_{k=-(q-1)/2}^{(q+1)/2} p_k e(n-(T+k)) \tag{5.2}$$

donde T es el retardo o *lag-delay*, p_k son los coeficientes de predicción y q el orden de predicción. Por otro lado, el vector de código $\mathbf{e}_c$ se obtiene a partir de un diccionario fijo o*Fixed Code-Book* (FCB), que se encarga de representar la señal residuo remanente tras eliminar las correlaciones a largo plazo, por lo que no se obtiene a partir de la señal previa.

De este modo, la señal de excitación $e(n)$ se puede calcular como:

$$e(n) = g_a \sum_{k=-(q-1)/2}^{(q+1)/2} p_k e(n-(T+k)) + g_c e_c(n) = g_a e_a(n) + g_c e_c(n), \tag{5.3}$$

donde los parámetros g_a, g_c, T y $e_a(n)$ se seleccionan siguiendo el proceso de análisis por síntesis [13].

Esta dependencia con la trama anterior para generar la señal de excitación $e_a(n)$ es la responsable de la propagación del error. Como consecuencia, la calidad perceptual de la señal de voz obtenida en los sucesivos paquetes recibidos correctamente se verá reducida.

A modo de ejemplo, en la figura 5.2 se puede observar la propagación del error en una simulación con pérdidas sobre el codec AMR [39], basado en el paradigma CELP. Así se aprecia que tras la última trama perdida, la señal de voz sintetizada tendría que ser idéntica a la original. Sin embargo, esta dependencia inter-trama genera una propagación del error que afecta a las sucesivas tramas recibidas correctamente.

En la bibliografía, la mayoría de trabajos se han centrado en recuperar la pérdida en sí, generando así diversos esquemas y algoritmos PLC que minimizan el impacto de la pérdida respecto a la señal original. Sin embargo, en los últimos años el interés científico se ha centrado en considerar la propagación del error como una importante fuente de degradación. En la bibliografía se pueden encontrar numerosas técnicas que minimizan o incluso evitan esta propagación del error como [66, 68–72, 112].

Con el objetivo de recuperar el paquete perdido y de mitigar la propagación del error, en este capítulo se utilizará una técnica de prevención de pérdidas basada en el emisor: la técnica que emplea códigos de corrección del error hacia delante o *Forward Error Correction* (FEC) dependiente del medio.

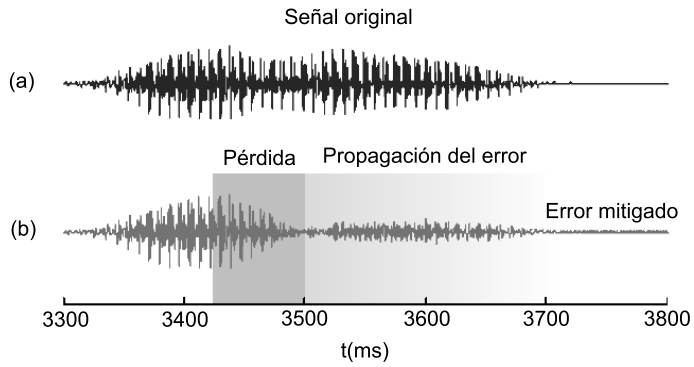

Figura 5.2: Ejemplo que muestra el error de propagación sobre la síntesis de voz empleando el codec AMR en el modo 12.2 kbps. a) Síntesis de voz sin pérdidas en el canal, b) síntesis de voz con pérdidas en el canal en el que se ha aplicado su propio algoritmo PLC. (Fuente:[2])

Como ya se comentó en la sección 3.4.1, el código FEC es una información redundante que se envía en cada paquete para recuperar los paquetes perdidos previamente durante la transmisión. Ahora bien, para generar este código FEC será necesario proporcionar una codificación eficiente para los parámetros de voz (coeficientes LPC y la señal de excitación).

Sin embargo, en la bibliografía no hay una representación adecuada para la estimación de la señal de excitación. Por este motivo, en esta tesis se ha querido abordar este problema y presentar diferentes métodos de cuantización, bajo una representación vectorial, con el objetivo de mejorar la estimación y la minimización del error de síntesis en la señal de voz obtenida.

Además, el uso de un código FEC presenta el inconveniente de incrementar la tasa de bits por paquete a enviar, provocando así que no se pueda utilizar el decodificador estándar, aunque no se produzcan pérdidas durante la transmisión. Para evitar este inconveniente, también se emplea una técnica esteganográfica que oculta este código FEC dentro de la codificación del paquete enviado. Así, no sólo se mantendrá la compatibilidad con el codec estándar, sino que además, no supondrá una pérdida significativa de calidad perceptual en la señal de voz obtenida.

5.2. Prevención de paquetes perdidos mediante el uso de códigos FEC

El código FEC del paquete recibido puede contener información redundante del paquete o paquetes previos. No obstante, hay que tener en cuenta el coste en tasa de bits que supone la codificación de los correspondientes parámetros de voz y el retardo a introducir durante el proceso de decodificación para realizar la síntesis de los paquetes perdidos. Ambos aspectos son críticos, ya que la nueva tasa de bits podría ser inviable para redes con un ancho de banda limitado y el retardo considerado podría ser excesivo para comunicaciones en tiempo real. Así, en esta tesis se plantea el uso de un código FEC dependiente del medio para recuperar la última trama perdida y/o evitar la propagación del error. Sin embargo, aunque solo se considera un retardo de una trama, todavía queda pendiente cómo se codifican los parámetros de voz (coeficientes LPC y la señal de excitación) para que no se incremente la tasa de bits de manera excesiva.

Como ya se comentó en la sección 2.4.2, hay diferentes representaciones para los coeficientes LPC. No obstante, la representación más utilizada en la bibliografía para la estimación es la representación LSF, puesto que presentan menor distorsión tras realizar la cuantización [29, 30] y mantienen la estabilidad del filtro del tracto vocal.

Sin embargo, la señal de excitación no tiene una adecuada representación para su estimación. Además, su magnitud, N muestras, y el rango dinámico en cada una de ellas no hace viable una cuantización escalar, ya que supondría una alta tasa de bits. Por este motivo, en esta tesis también se aborda el problema de la representación y la generación de diccionarios de cuantización para tanto reducir el error de síntesis con la señal de voz original como para obtener estimaciones de la señal de excitación en esquemas de mitigación. Para ello, se ha considerado una representación vectorial de la señal de excitación y a continuación se presentan diferentes métodos de cuantización vectorial, basados en algoritmo clásico *Linde-Buzo-Gray* (LBG) [113], para la obtención del diccionario.

5.2.1. Cuantización vectorial de la señal de excitación

La cuantización vectorial es una generalización de la cuantización escalar cuando se considera como unidad a cuantizar un vector de tamaño N. Cada vector estará compuesto por N componentes que pueden tomar cualquier valor continuo

y que debe ser representado por uno de las vectores N-dimensionales que tiene el diccionario de cuantización utilizado. Esta idea ya fue sugerida por Shanon para mejorar los resultados de codificación de muestras individuales en términos de eficiencia de compresión [114]. De esta manera, el vector escogido en el diccionario será aquel que minimiza la distancia euclidea con el vector de N muestras original.

Para la obtención de este diccionario de cuantización, se emplea un algoritmo de cuantización vectorial que consta de dos pasos: la búsqueda de celda óptima y la obtención del centro óptimo. Un ejemplo es el algoritmo *Linde-Buzo-Gray* (LBG) [113] en el cual se basan los métodos desarrollados en esta tesis.

El algoritmo *Linde-Buzo-Gray* (LBG) [113] aparece en 1980 como una generalización del algoritmo *Lloyd-Max* y a menudo en la bibliografía se identifica bajo la denominación de algoritmo *K-medias* o *Kmeans*. Su funcionamiento se basa en un proceso iterativo donde en un principio la población inicial se clasifica en dos celdas, de acuerdo a la menor distancia euclidea con dos centros o centroides iniciales, y posteriormente se obtiene un centro representativo, un vector promedio, de cada celda. Estos pasos se repiten hasta que se consigue la convergencia en los dos centroides obtenidos. De manera iterativa, los centroides resultantes de la iteración anterior se dividirán en otros dos (aplicando bipartición) y se repetirá el procedimiento indicado para para alcanzar la convergencia con los nuevos centroides hasta tener el número de C centroides que conformarán el diccionario de cuantización.

Sin embargo, aunque bajo una representación vectorial es posible reducir la enorme cantidad de bits necesaria para codificar la señal de excitación, hay que tener en cuenta que su cuantización mediante el algoritmo LBG presenta dos inconvenientes. Por un lado, la dimensionalidad de la señal de excitación complica la obtención de diccionarios representativos que minimicen el error de cuantización debido al coste exponencial en recursos a medida que aumenta el tamaño C del diccionario. Por otro lado, durante el paso de centro óptimo se está obteniendo un vector promedio que aunque la señal obtenida minimice la distancia euclidea con las componentes de la celda, también se puede estar obteniendo una señal que pierda características importantes como el *pitch*. Con el objetivo de mejorar tanto la obtención de celda óptima como el centro óptimo, a continuación se presentan diferentes métodos de cuantización para la señal de excitación que están basados en el algoritmo LBG.

Algoritmo modificado *Kmedoids*

Como ya se ha mencionado, en el algoritmo LBG se repiten 2 pasos en cada iteración, la búsqueda de celda o conjunto óptimo y la obtención de centro óptimo que representa a toda la celda. Durante el primero de ellos se está aplicando una distancia euclidea definida para dos vectores $\mathbf{e}_1$ y $\mathbf{e}_2$ como:

$$d(e_1, e_2) = \sqrt{\sum_{n=1}^{N} \Big(e_1(n) - e_2(n)\Big)^2} \tag{5.4}$$

donde N es la longitud de los vectores.

Aunque esta distancia permite al algoritmo LBG asignar un vector $\mathbf{e}$ a un centro $\mathbf{c}$ que minimiza la distancia $d(e, c)$, sólo clasifica de acuerdo a la magnitud que tiene sus componentes. Veámoslo con un ejemplo donde se consideran dos señales idénticas y una de ellas se multiplica por una ganancia G. Queda claro que si se aplica el algoritmo LBG, esa ganancia G puede hacer que estas señales no se clasifiquen bajo el mismo centroide debido a la distancia euclidea. Sin embargo, para obtener señales representativas, sí que interesa que señales que tengan unas características similares puedan clasificarse en la misma celda. Para ello, se ha establecido una normalización de los vectores $\mathbf{e}$ por una ganancia G, de manera que todos los vectores tengan energía unitaria ($\mathbf{u}$). De esta manera, una señal de excitación $e(n)$ puede definirse como $e(n) = Gu(n)$, donde la ganancia G vendrá definida como:

$$G = \frac{1}{\sqrt{\sum_{i=1}^{N} e(n)^2}} \tag{5.5}$$

Por otro lado, para la obtención del centro óptimo por cada celda, el algoritmo LBG obtiene un vector promedio que minimiza la distancia euclidea con todos los vectores unitarios $\mathbf{u}$ de la celda. Sin embargo, este vector promedio puede perder algunas características de la señal de excitación, como el *pitch*, como consecuencia del rango dinámico que tienen sus N componentes. Para evitar esto, durante la obtención del centro óptimo se utiliza el método *kmedoids* [115] en el cual no se escoge el vector promedio sino aquel vector de energía unitaria $\mathbf{u}$ con menor distancia euclidea a vector promedio.

Finalmente, los centroides del diccionario de cuantización resultante son las señales de excitación, $e_b(n) = G_b u_b(n)$, correspondientes a los vectores de energía unitaria $\mathbf{u}_{b,i}$, obtenidos para cada celda i-ésima. Por lo tanto, estos centroides son

vectores representativos de la base de datos de entrenamiento y que mantienen las características de la señal intactas.

Algoritmo modificado con distancia de síntesis

Hasta ahora, se ha realizado una cuantización vectorial de la señal de excitación para obtener un diccionario cuyos centroides minimicen el error de cuantización de la señal de excitación original $e(n)$. Sin embargo, no hay que olvidar que el objetivo final es la minimización del error de síntesis entre la señal de voz sintetizada $\hat{s}(n)$ y la señal de voz original $s(n)$. Por lo tanto, dado que en la obtención de la señal de voz, $s(n) = h(n) * e(n)$, intervienen tanto la señal de excitación $e(n)$ como la respuesta al impulso del filtro LPC $h(n)$, hay que considerar también el efecto de la convolución con la respuesta al impulso $h(n)$ a la hora de seleccionar la señal de excitación $\hat{e}^{(i)}(n)$, para el índice i en el diccionario de tamaño C, con menor error ϵ_i definido como:

$$\epsilon_i = \sum_{n=0}^{N-1} \left(h(n) * \hat{e}^{(i)}(n) - s(n)\right)^2, \tag{5.6}$$

donde N es el número de muestras de la trama y $s(n)$ la señal de voz original.

De este modo, se plantea una modificación del clásico algoritmo LBG considerando una nueva distancia que modifica tanto la asignación a la celda óptima como la obtención del centro o centroide óptimo $\mathbf{c}^{(i)}$. Así, para cada trama b-ésima de la base de datos de entrenamiento, se obtendrá la respuesta al impulso $h_b(n)$ y la señal de excitación $e_b(n)$ definida como $e_b(n) = G_b u_b(n)$ a partir de su ganancia G_b y la señal de excitación de energía unitaria $u_b(n)$. Esta señal $u_b(n)$ se asignará a un centroide $\mathbf{c}^{(i)}$ si se cumple que $\epsilon(\mathbf{u}_b, \mathbf{c}^{(i)}, \mathbf{h}_b, G_b) < \epsilon(\mathbf{u}_b, \mathbf{c}^{(j)}, \mathbf{h}_b, G_b)\ \forall i \neq j$. Esto es, se está considerando una nueva distancia $\epsilon(\mathbf{u}, \mathbf{c}, \mathbf{h}, G)$, que denominaremos como distancia de síntesis, y que está definida como:

$$\epsilon(\mathbf{u}, \mathbf{c}, \mathbf{h}, G) = G^2 \sum_{n=0}^{N-1} (h(n) * u(n) - h(n) * c(n))^2, \tag{5.7}$$

Una vez agrupados los vectores $\mathbf{u}_b$ de la base de datos de entrenamiento en las diferentes celdas, hay que obtener el nuevo centroide óptimo para cada celda $\mathcal{B}_i$, $\forall i \in [1, C]$. Sin embargo, el correspondiente centro $\mathbf{c}_{new}^{(i)}$ no se obtiene como un vector promedio sino que se obtiene como el centro que minimiza la distancia de síntesis entre todas las señales $u_b(n)$ clasificados en la celda $\mathcal{B}_i$ del centro i-ésimo.

Así, el centro óptimo $\mathbf{c}_{new}^{(i)}$ se obtiene como:

$$\mathbf{c}_{new}^{(i)} = \underset{\mathbf{c}}{\operatorname{argmin}} \left(\sum_{b \in \mathcal{B}_i} \epsilon(\mathbf{u}_b, \mathbf{c}, \mathbf{h}_b, G_b) \right) \tag{5.8}$$

No obstante, este proceso de minimización no es sencillo, ya que la expresión (5.7) contiene la operación de convolución. Dado que la convolución es una operación analíticamente compleja en el dominio del tiempo, es necesario tratar esta expresión en el dominio de la frecuencia. De esta manera, aplicando la transformada discreta de Fourier (DFT) se convierte la convolución de la expresión (5.7) en un producto. Ahora bien, esta conversión es válida cuando se trata de una convolución circular, por lo que hay que linealizar la convolución. Para este fin, un paso previo a la transformada de Fourier consiste en añadir ceros, o aplicar *zero-padding*, a los vectores $\mathbf{c}$ y $\mathbf{u}_b$ hasta una longitud como mínimo el doble de la longitud del vector ($K \geq 2N$). Así, la expresión (5.7) se define en el dominio de la frecuencia como:

$$\varepsilon(\mathbf{U}, \mathbf{C}, \mathbf{H}, G) = \sum_{k=0}^{K-1} (GH(k)U(k) - GH(k)C(k))^2 \tag{5.9}$$

donde $\mathbf{C}$, $\mathbf{H}$ y $\mathbf{u}$ son las respectivas transformadas de Fourier de los vectores $\mathbf{c}$, $\mathbf{h}_b$ y $\mathbf{u}_b$ y G_b es la ganancia de la señal de excitación $e_b(n)$.

De este modo, mientras que antes se comparaban las señales sintetizadas en el tiempo, ahora se comparan los espectros de las señales que han sido afectadas por un *zero-padding* por lo que se están midiendo distancias diferentes. Así, el proceso de minimización en el dominio de la frecuencia queda definido como:

$$\begin{aligned} \mathbf{C}_{new}^{(i)} &= \underset{\mathbf{C}}{\operatorname{argmin}} \left(\sum_{b \in \mathcal{B}_i} \varepsilon(\mathbf{U}_b, \mathbf{C}, \mathbf{H}_b, G_b) \right) \\ &= \underset{\mathbf{C}}{\operatorname{argmin}} \sum_{b \in \mathcal{B}_i} \sum_{k=0}^{K-1} (H_b(k)U_b(k) - H_b(k)C(k))^2 \end{aligned} \tag{5.10}$$

en cuya minimización se puede obviar la ganancia G al tratarse de una constante.

Si ahora se modifica el orden de los sumatorios, se puede ver que es posible realizar esta optimización mediante un enfoque por mínimos cuadrados o *Least Square Error* (LSE) para cada componente del vector $k \in [0, K-1]$. De modo que cada

componente en frecuencia del nuevo centroide $\mathbf{C}_{new}^{(i)}(k)$ se define como:

$$C_{new}^{(i)}(k) = \underset{C_k}{\operatorname{argmin}} \sum_{b \in \mathcal{B}_i} (H_b(k)U_b(k) - H_b(k)C_k)^2 \tag{5.11}$$

que mediante la resolución por mínimos cuadrados, se obtiene como:

$$C_{new}^{(i)}(k) = \frac{\sum_{b \in \mathcal{B}_i} H_b^*(k)U_b(k)}{\sum_{b \in \mathcal{B}_i} H_b^*(k)H_b(k)} \tag{5.12}$$

donde $H_b^*(k)$ es la transpuesta Hermítica del término $H_b(k)$.

Por último, los correspondientes centroides del diccionario se obtienen aplicando la transformada inversa de Fourier al centro $C_{new}^{(i)}$, $\forall i \in [1, C]$.

Algoritmo de división de centroides fija o dinámica

Aunque el algoritmo que considera la distancia de síntesis proporciona un centroide óptimo para toda la celda $\mathcal{B}_i$, este algoritmo puede mejorarse si se tiene en cuenta que la distribución de elementos en cada celda no suele ser homogénea. Dado que en cada iteración del algoritmo LBG todos los centroides se dividen en otros dos y posteriormente se aplica la convergencia, cabe pensar que si uno de estos centroides ya está bien entrenado al representar un conjunto pequeño, no haría falta volver a dividirlo y sería mejor entrenar las celdas más pobladas de las cuales se pueden obtener mejores centroides.

En este algoritmo, se va a considerar que sólo se divide la celda más poblada $\mathcal{B}_i$, $||\mathcal{B}_i||_0 > ||\mathcal{B}_j||_0, \forall i \neq j \in C$. Una vez dividida la celda mayor, se puede proceder de dos maneras:

- Emplear un enfoque fijo que consiste en entrenar sólo los vectores clasificados en esa celda $\mathcal{B}_i$ con los nuevos centroides obtenidos y dejar el resto de celdas con su correspondiente centroide calculado.

- Emplear un enfoque dinámico que consiste en entrenar de nuevo todos los vectores de la base de datos de entrenamiento sobre todos los centroides obtenidos de la iteración anterior y los dos nuevos obtenidos de la celda más poblada $\mathcal{B}_i$. De esta manera, se permite que vectores que estaban en una celda pasen a una nueva celda que los representa mejor.

Este proceso de división de la celda más poblada se puede realizar iterativamente hasta alcanzar C centroides que conforman el diccionario aplicando los pasos para

obtener la celda y centro óptimos explicados en la sección anterior. Como resultado de este nuevo método de cuantización, los centroides del nuevo diccionario son más representativos que los obtenidos por el propio algoritmo LBG o las alternativas anteriores, ya que se centra en entrenar mejor las celdas más pobladas.

Sin embargo, para obtener un diccionario cuyos centros sean representativos de una celda, entendiéndose éstas como celdas con un mínimo conjunto de vectores, esta división se realiza cuando la celda tiene un mínimo de vectores. Es decir, no es de gran utilidad una celda que recoge uno o muy pocos vectores cuando otra tiene un conjunto de vectores mucho mayor. Para ello se ha establecido el parámetro α como el número mínimo que tiene que tener una celda para que ésta pueda ser divida. Dependiendo de este factor se pueden obtener diccionarios de más o menos centros por lo que hay que ajustar su valor experimentalmente según la base de datos utilizada.

Por último, por el funcionamiento de las dos propuestas comentadas, se desprende que el enfoque fijo genera diccionarios subóptimos respecto al enfoque dinámico. Sin embargo, el enfoque fijo tiene la ventaja de estar clasificando celdas $\mathcal{B}_i$ con una población inferior a la de partida por cada nueva iteración. Este hecho hace que se puedan obtener diccionarios con un número de centroides muy superior al proporcionado por algoritmos como LBG pero sin que suponga un coste exponencial en recursos por cada nuevo centroide.

5.2.2. Técnica FEC basada en diccionario de señales de excitación

Con la posibilidad de minimizar el error de cuantización de la señal de excitación, en la figura 5.3 se plantea un esquema que permite recuperar la última trama perdida en una ráfaga. De este modo, el código FEC va a proporcionar tanto los coeficientes LPC, como la señal de excitación $\hat{e}(n)$ que minimiza el error de síntesis de la señal de voz de la última trama perdida.

La principal ventaja de este esquema es que bajo una codificación eficiente de los parámetros de voz, es posible no sólo minimizar el error respecto a la señal de voz original sino que también, permite mitigar la propagación del error en los codecs basados en el paradigma CELP. Esto es, al proporcionar una señal de excitación que minimiza el error de síntesis en la última trama perdida, también se está reduciendo la propagación del error cuando se genera la señal de excitación, aplicando el filtro

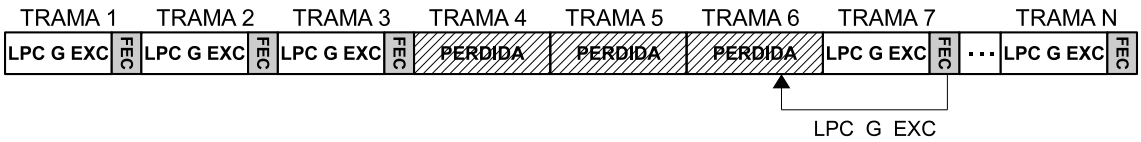

Figura 5.3: Esquema FEC que proporciona los coeficientes LPC, la gananaica G la señal de excitación de energía unitaria (EXC) para recuperar la última trama perdida y al mismo tiempo, evitar la propagación del error en los codecs basados en el paradigma CELP.

LTP, en la trama recibida correctamente respecto a cuando no se dispone de ella.

Bajo este esquema de mitigación, el código FEC está compuesto por los índices de cuantización correspondientes a la codificación de los coeficientes LPC y la señal de excitación, siendo ésta última representada como una ganancia G y su correspondiente señal de energía unitaria $u(n)$, $e(n) = Gu(n)$.

5.2.3. Aplicación de la esteganografía para ocultar el código FEC

A pesar de que la utilización de los códigos FEC dará como resultado una mejora en la calidad perceptual de la señal de voz obtenida, hay que tener en cuenta que su uso implica un incremento en la tasa de bits por paquete enviado y que éste podría ser inviable para transmisiones en canales con un ancho de banda limitado. Además, este código FEC supone una modificación en el envío de los paquetes, ya que no tienen la misma cabecera que el paquete generado por el codificador. Como resultado, el decodificador original no será capaz de procesar unos paquetes o tramas que no tienen el formato original. Es decir, la aplicación de una técnica FEC impide el uso del decodificador original aunque no se produzcan pérdidas en el canal.

Para mantener una compatibilidad con el codec original, se plantea el uso de una técnica esteganográfica. La esteganografía se caracteriza por ocultar cierta información dentro de otro formato, como puede ser una imagen, música, video o voz, de manera que si alguien obtiene el archivo no pueda detectar que hay otra información incluida en él. Este proceso no hay que confundirlo con una marca de agua, ya que ésta última está diseñada para evitar su extracción del archivo, siendo utilizada en la práctica como copyright de archivos multimedia (ya sea en su modo visible o no visible) [116].

En esta tesis se emplea la esteganografía como mecanismo para ocultar el código

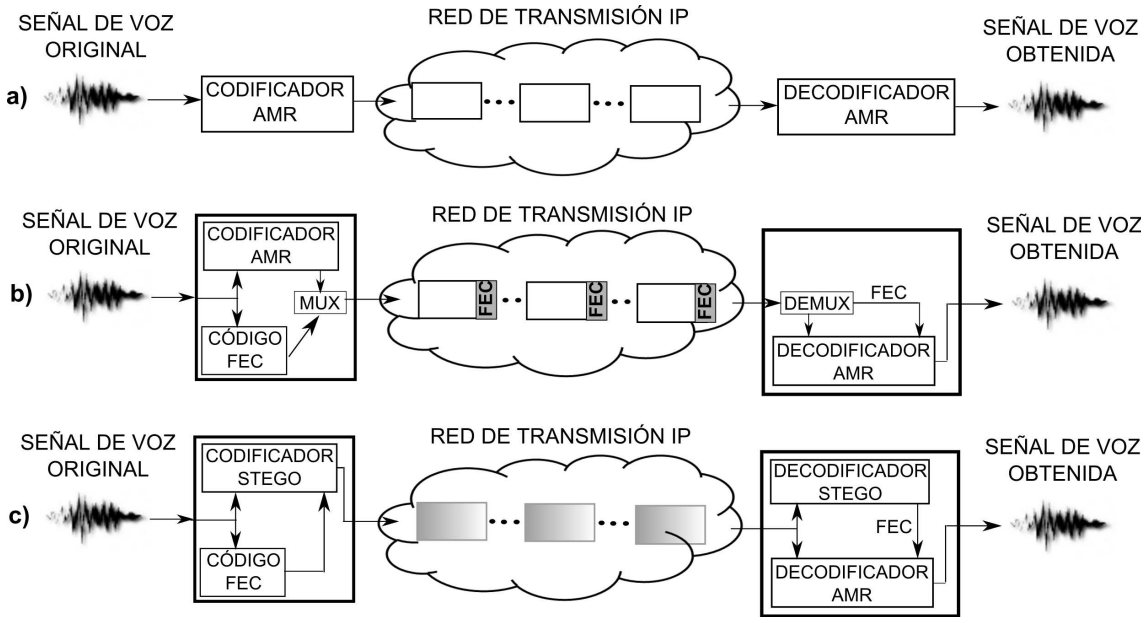

Figura 5.4: Esquemas generales de transmisión sobre un canal IP utilizando el codec AMR y sus diferentes variantes que incluyen el código FEC. En a) se presenta el esquema de transmisión en paquetes por una red IP sin código FEC utilizando el codec AMR estándar, en b) se presenta la modificación que sufriría cada paquete al incluir el código FEC introducido por cada paquete y en c) se presenta la propuesta que emplea la esteganografía para ocultar el código FEC dentro del paquete a enviar.

FEC dentro de la codificación del paquete a enviar, con el fin de evitar el incremento en la tasa de bits y la consecuente incompatibilidad con el codec original. Para este fin, se ha empleado la técnica desarrollada por Geiser et al. en [117].

En la figura 5.4 se muestra, de manera general, cómo funcionaría una transmisión con el codec AMR estándar (figura a), frente a los esquemas que utilizan un código FEC (figura b) y tras aplicar la técnica esteganográfica (figura c) sobre una red IP.

Sin embargo, esta tarea no es simple dado que la inserción del código FEC en la propia codificación del paquete genera modificaciones en los bits originales (generalmente se realiza sobre los bits menos significativos). Además, el número de modificaciones se incrementa a medida que se oculta un volumen de datos mayor, por lo que hay que determinar cuánta información redundante se puede ocultar en el paquete sin que se produzca una caída significativa en la calidad perceptual para las transmisiones sin pérdidas.

Para desarrollar esta técnica esteganografíca, Geiser et al. propone en [117] que los K bits de la información a ocultar, que denotaremos como mensaje m, se realice sobre uno de los vectores de códigos del diccionario FCB. Para ello, realiza una

división del diccionario en $M = 2^K$ subdiccionarios disjuntos $\mathcal{C}_M$ donde los códigos $\mathbf{c} \in \mathcal{C}_m$ serán los candidatos donde se encuentre el mensaje oculto m. Sin embargo, esta disminución en el número de vectores a comparar en el diccionario FCB, reducido por un factor M, conllevaría una reducción de la calidad perceptual de la señal sintetizada, al no tener en cuenta el resto de centros en el diccionario que podrían mejorar la calidad de la señal sintetizada. Para evitar esta reducción de la calidad perceptual, en [117] se plantea una modificación en el algoritmo de búsqueda, sobre la codificación ACELP [39, 43], durante los pasos de división de diccionario y expansión del espacio de búsqueda.

Por último, indicar que esta técnica esteganográfica permite ocultar un código FEC que suponga una tasa de bits de entre 200 bps a 2 kbps y que su uso está particularizado al codec AMR en el modo 12.2 kbps [117].

5.2.4. Resultados experimentales

Una vez presentada la técnica FEC y la técnica esteganográfica que permitirá ocultarlo, en esta sección se presentan los resultados de su aplicación bajo dos propuestas sobre el codec AMR en el modo 12.2 kbps.

Por un lado, siguiendo con el esquema presentado en la figura 5.3, este código FEC se utiliza para recuperar la última trama perdida en una ráfaga, cuyos parámetros minimizan el error de síntesis, y al mismo tiempo reducir la propagación del error que se produce en los codecs basados en el paradigma CELP, como es el codec AMR [39].

Para ello, el código FEC está compuesto por los índices de cuantización que proporcionan los coeficientes del filtro LPC, y la ganancia G y la señal de energía unitaria $u(n)$ que definen la señal de excitación $e(n)$. Estos índices se obtienen a partir de unos diccionarios que minimizan el error de cuantización. Así, en el caso de los coeficientes LPC se ha obtenido un diccionario mediante el algoritmo clásico LBG, bajo su representación como coeficientes LSF, para la ganancia se ha obtenido un diccionario mediante un cuantizador escalar Lloyd-Max [118], y para la señal de excitación se han empleado diferentes diccionarios de cuantización comentados en la sección 5.2.1. En concreto se han empleado dos métodos de cuantización diferentes, el método basado en la distancia de síntesis, que denotaremos como *Synthesized Linde-Buzo-Gray* (SLBG) y que ha sido comentado en la sección 5.2.1, y la modificación basada en la división de la celda más poblada y entrenando todos los datos sobre los centros resultantes, y que denotaremos como *Dynamic Synthesized Linde-Buzo-Gray*

(DSLBG), comentado en la sección 5.2.1. Todos los diccionarios obtenidos tienen un total de 1024 centroides y están obtenidos a partir del conjunto de entrenamiento de la base de datos TIMIT [104].

Con los diccionarios obtenidos, el código FEC supone un incremento en la tasa de bits de 1.65 kbps para recuperar una única trama perdida. Utilizando este código FEC, se han presentado dos propuestas de acuerdo con el método de cuantización empleado en la obtención del diccionario de la señal de excitación. Así se ha denotado como (SLBG+FEC) a la propuesta que utiliza el método SLBG y como (DSLBG+FEC) a la propuesta que utiliza el método DSLBG. Las pruebas se han realizado sobre el conjunto de test de la base de datos TIMIT [104] y las condiciones de canal se han simulado siguiendo un modelo Gilbert (comentado en la sección 3.5.3). Con estas pruebas se analizará el rendimiento de las propuestas, en términos de calidad objetiva PESQ [103], con diferentes tasas de pérdidas de paquetes o *Packet Error Rate* (PER), PER= $\{10\,\%, 20\,\%, 30\,\%, 40\,\%, 50\,\%\}$ y con diferentes longitudes promedio de ráfaga o *Average Burst Length* (ABL), ABL= $\{1, 2, 4, 8, 12\}$.

En la tabla 5.1 se presentan los resultados PESQ promedio para el codec AMR, con su propio algoritmo LPC, y las propuestas mencionadas (SLBG+FEC) y (DSLBG +FEC). A la vista de los resultados, se aprecia que, al recuperar la última trama de la ráfaga de pérdidas, se obtiene una notable mejora respecto al codec AMR, ya que no sólo se minimizará el error de síntesis en la propia trama, sino que también se reducirá la propagación del error. Además, también se puede apreciar que la propuesta que incorpora el método (DSLBG+FEC) proporciona unos resultados significativamente mejores que los de (SLBG+FEC) al poder realizar un mejor entrenamiento del diccionario.

Por otro lado, se propone aplicar la técnica esteganográfica a un esquema similar al de la figura 5.3 pero donde el código FEC se emplea sólo para evitar la propagación del error en las tramas recibidas correctamente tras la pérdida. En este caso, el código FEC está compuesto por un índice de cuantización que identifica un único pulso con el que es posible realizar la sincronización del diccionario FCB sobre el codec AMR. Este pulso se calcula empleando la técnica multipulso, descrita en [68, 69], y el índice de cuantización se obtiene de un diccionario con 1024 centroides. Así, el incremento en la tasa de bits será de 550 bps frente a los 1.65 kbps de la propuesta anterior.

Aunque ambas tasas de bits son soportadas por la técnica esteganográfica presentada en la sección 5.2.3, hay que tener en cuenta que la calidad perceptual de la señal de voz en transmisiones sin pérdidas decrece a medida que la tasa de bits a

		Tasa de pérdida de paquetes				
	ABL	10 %	20 %	30 %	40 %	50 %
AMR	1	2.819	2.273	1.829	1.348	1.009
	2	2.802	2.249	1.863	1.541	1.251
	4	2.820	2.223	1.790	1.457	1.161
	8	2.871	2.234	1.783	1.445	1.139
	12	2.925	2.303	1.831	1.450	1.183
SLBG+FEC	1	2.935	2.441	2.035	1.589	1.396
	2	2.948	2.418	2.066	1.768	1.557
	4	2.912	2.371	1.919	1.669	1.443
	8	2.958	2.385	1.899	1.649	1.422
	12	2.997	2.438	2.043	1.650	1.486
DSLBG+FEC	1	2.955	2.573	2.091	1.626	1.428
	2	2.966	2.456	2.115	1.815	1.604
	4	2.947	2.412	1.970	1.712	1.495
	8	2.989	2.434	1.953	1.682	1.476
	12	3.025	2.478	2.070	1.697	1.536

Tabla 5.1: Resultados PESQ promedio obtenidos para el codec AMR, con su propio algoritmo PLC, y las propuestas (SLBG+FEC) y (DSLBG+FEC) que utilizan el código FEC para recuperar la última trama perdida y reducir la propagación del error. Los tests se han realizado sobre diferentes condiciones de canal de acuerdo a la tasa de pérdida de paquetes y la longitud promedio de ráfaga (ABL).

ocultar del código FEC es mayor. Por este motivo, se ha decidido utilizar la técnica esteganográfica sobre una propuesta que suponga un impacto inferior y al mismo tiempo permita mostrar las ventajas de su utilización. En este caso, para un incremento de 550 bps se ha utilizado el modo de ocultación de hasta 800 bps que presenta el algoritmo esteganográfico sobre el codec AMR [117].

Tanto la propuesta que utiliza el código FEC (AMR+FEC), basado en la codificación multipulso, y la propuesta que emplea la técnica esteganográfica (STEGO), que oculta este código FEC dentro del propio paquete, se analiza su rendimiento en términos de calidad objetiva PESQ [103] sobre el conjunto de tests de la base de datos TIMIT [104]. Los resultados PESQ promedio presentados en la tabla 5.2 se han obtenido sobre diferentes condiciones de canal, descritas en [68, 69], mediante un canal aleatorio comentado en la sección 3.3.3. Así, este canal se ha simulado con diferentes tasas de pérdida de paquetes, PER= $\{4\,\%, 7\,\%, 10\,\%, 13\,\%, 16\,\%, 18\,\%, 21\,\%, 23\,\%\}$.

A la vista de los resultados, se puede observar que las propuestas (AMR+FEC) y (STEGO) proporcionan un rendimiento superior al propio codec con su propio algoritmo PLC en todas las condiciones de canal como consecuencia de la reducción de la

	Tasa de pérdida de paquetes								
	0 %	4 %	7 %	10 %	13 %	16 %	18 %	21 %	23 %
AMR	4.003	3.212	2.936	2.690	2.531	2.321	2.200	2.108	2.024
AMR+FEC	4.003	3.417	3.193	2.997	2.878	2.720	2.626	2.556	2.497
STEGO	3.991	3.398	3.170	2.974	2.852	2.693	2.597	2.527	2.466

Tabla 5.2: Resultados PESQ promedio en los tests realizados sobre el codec AMR 12.2 kbps (AMR) respecto a las propuestas que aplica el código FEC (AMR+FEC), basado en la técnica multipulso, y la que lo oculta utilizando una técnica esteganográfica (STEGO) bajo diferentes tasas de pérdidas de paquetes en un canal aleatorio.

propagación del error. También se observa que aunque la propuesta esteganográfica (STEGO) tiene un rendimiento ligeramente inferior a la propuesta (AMR+FEC), supone una diferencia de 0.026 en promedio, ésta se debe a la modificación realizada en cada paquete por la técnica esteganográfica para incluir el código FEC. No obstante, la aplicación de la técnica esteganográfica ha mantenido la compatibilidad con el codec estándar para transmisiones sin pérdidas y sin que suponga una notable caída de calidad perceptual (4.003 vs 3.991).

Capítulo 6

Técnicas de mitigación de pérdidas sobre redes IP

Este capítulo se centrará en el desarrollo de técnicas de mitigación de pérdidas basadas en el receptor sobre redes IP. Con el objetivo de recuperar las tramas perdidas, independientemente de la longitud de ráfaga, en esta tesis se propone el uso de la técnica de vectores de sustitución. Para ello, esta técnica tiene que proporcionar una estimación de los parámetros de voz (los coeficientes LPC y la señal de excitación) para cada trama perdida. Estas estimaciones se obtendrán mediante un proceso de estimación MMSE donde los diccionarios empleados para cuantizar la señal de excitación se obtendrán de acuerdo a los métodos explicados en el Capítulo 5.

Esta técnica basada en vectores de sustitución puede complementarse para obtener diferentes esquemas de mitigación de pérdidas que mejoran la calidad perceptual de la señal de voz recuperada. Por un lado, se presentará un enfoque mixto entre las técnicas de prevención basada en los códigos FEC y la técnica de mitigación basada en vectores de sustitución. Por otro lado, se presentará una propuesta que mejora la estimación de la señal de excitación al principio de una ráfaga empleando un filtro adaptativo. Finalmente, se presentará una nueva representación de la señal de excitación basada en la transformada wavelet que ayuda tanto en la generación de diccionarios sobre particiones más pequeñas, como en la obtención de mejores estimaciones.

6.1. Introducción

Cuando se producen pérdidas de paquetes en una transmisión sobre las redes IP, la mayoría de los codecs actuales aplican un algoritmo de mitigación de pérdidas o *Packet Loss Concealment* (PLC) [6] con el objetivo de reducir el impacto que ésta tiene en la señal de voz recuperada. Para ello, estos algoritmos explotan la alta correlación que existe en la señal de voz almacenada en paquetes o tramas previas, y así proporcionar unos parámetros de voz que minimicen el efecto de la pérdida en la señal de voz recuperada [6].

Sin embargo, en las redes IP no es común que se produzcan estas pérdidas de manera aislada, sino que la congestión en la red provoca que varios paquetes se pierdan de manera consecutiva, generando así una ráfaga. El problema es que los algoritmos PLC implementados en los codecs no ofrecen un buen rendimiento cuando estas ráfagas son largas, ya que la voz no es una señal estacionaria en el tiempo, y podrían generar sonidos molestos durante la reproducción. Para evitarlo, los algoritmos PLC aplican un proceso de apagado o *muting* de la señal hasta llegar a anularla si la ráfaga fuera muy larga.

Con el objetivo de minimizar el impacto de las pérdidas en ráfagas, se han desarrollado varios esquemas de mitigación de pérdidas que permiten estimar los parámetros de voz (coeficientes LPC y la señal de excitación) para realizar la síntesis de la señal voz en cualquier punto de la ráfaga. Para ello, esta estimación se realizará de acuerdo a un proceso estimación por mínimo error cuadrático medio (MMSE) [94], presentado en la sección 3.4.2.

No obstante, el éxito de los esquemas de mitigación que se proponen en esta tesis dependen de la representación escogida para la señal de excitación, ya que en la bibliografía, ésta no tiene una representación adecuada para su estimación. Por este motivo, se han empleado diferentes métodos de cuantización sobre la señal de excitación, comentados en la sección 5.2.1, para obtener diccionarios que se utilizan para generar las estimaciones que conforman los vectores de sustitución. Además, también se presenta un esquema que combina la técnica de vectores de sustitución con la técnica predictiva basada en los códigos FEC. El objetivo es presentar una técnica que además de recuperar las tramas perdidas en una transmisión, también evite la propagación del error en los codecs que presentan dependencias inter-trama como los codecs basados en el paradigma CELP [13].

Para mejorar las estimaciones de los vectores de sustitución, en particular sobre la señal de excitación, también se proponen dos variantes. Una de ellas se centra en

mejorar la estimación en las primeras pérdidas de una ráfaga mediante el uso de un filtrado adaptativo basado en corrección recursiva o *Recursive Least Squares* (RLS). La otra propuesta plantea una nueva representación de la señal de excitación basada en la transformada wavelet. Esta representación permitirá tanto la división de la señal excitación en particiones más pequeñas, con lo que se facilita su cuantización, como la reducción del error de síntesis, ya que se minimizará el error por cada una de las particiones.

6.2. Esquema de mitigación de pérdidas con vectores de sustitución

El objetivo de las técnicas de mitigación de pérdidas de paquetes es la de proveer estimaciones de los parámetros de voz (los coeficientes LPC y la señal de excitación) que minimizan el error de síntesis con la señal de voz original $s(n)$ definida como:

$$s(n) = \sum_{l=1}^{p} a_l s(n-l) + e(n) = \sum_{l=1}^{p} a_l s(n-l) + Gu(n) \tag{6.1}$$

donde p es el orden de predicción, a_l son los coeficientes LPC y la señal de excitación $e(n)$ está representada por una ganancia G y la señal de energía unitaria $u(n)$.

Para proporcionar una estimación de los parámetros de voz (los coeficientes LPC, la ganancia G y la señal de excitación de energía unitaria $u(n)$) en cada trama perdida, se emplea una técnica de estimación basada en vectores de sustitucón hacia delante [84]. La técnica propuesta se muestra en la figura 6.1, donde a partir de los índices de cuantización (i_{LPC}, i_G y i_{EXC}) de la última trama recibida correctamente, se obtienen los correspondientes vectores de sustitución ($\mathbf{V}_{\text{LPC}}$, $\mathbf{V}_G$ y $\mathbf{V}_{\text{EXC}}$). Estos vectores de sustitución serán los que proporcionen las estimaciones de los parámetros de voz durante la ráfaga.

Estos vectores de sustitución se definen como $\mathbf{V} = (\hat{\mathbf{c}}_1, \hat{\mathbf{c}}_2, ..., \hat{\mathbf{c}}_T)$, donde T es la longitud máxima para realizar reemplazos, siendo $\hat{\mathbf{c}}_\tau$ el parámetro estimado mediante una estimación MMSE [84, 94] para la $\tau-$ésima trama perdida de forma consecutiva en una ráfaga. En las siguientes secciones se explicará cómo se han obtenido los vectores de sustitución para cada parámetro de voz (los coeficientes LPC, ganancia y señal de excitación de energía unitaria) empleando un modelo de producción de voz [119, 120] y la información previa a la ráfaga.

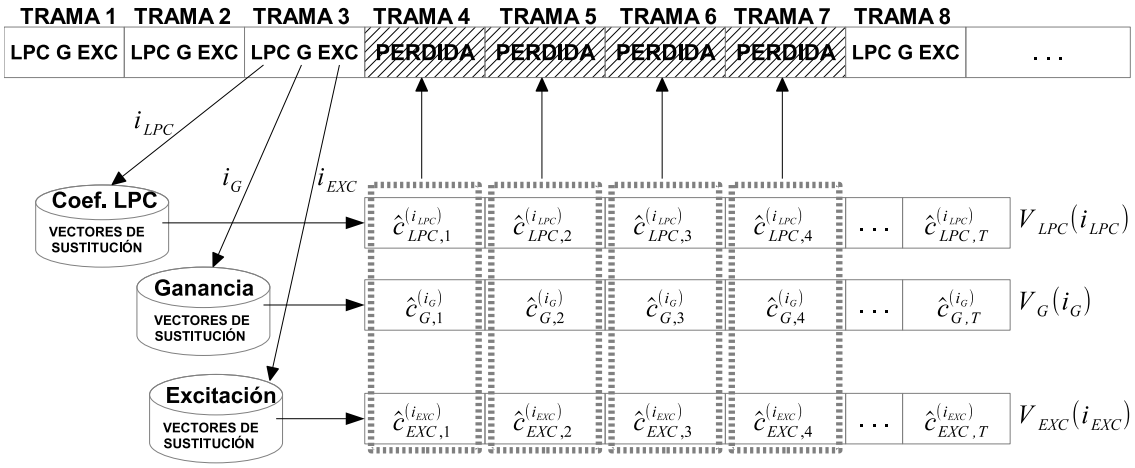

Figura 6.1: Esquema de mitigación de errores que permite recuperar tanto los coeficientes LPC como la ganancia (G) y la señal de excitación de energía unitaria (EXC) para las tramas perdidas en una transmisión. Las estimaciones se obtienen del correspondiente vector de sustitución ($\mathbf{V}_{\mathrm{LPC}}$, $\mathbf{V}_G$ y $\mathbf{V}_{\mathrm{EXC}}$) obtenido a partir de los índices cuantizados (i_{LPC}, i_G y i_{EXC}) de la última trama recibida correctamente antes de la ráfaga.

6.2.1. Estimación de los parámetros de voz

Para generar un vector de sustitución que proporcione la estimación de cada parámetro, es necesario observar la historia previa al mismo. Cada una de las estimaciones $\hat{\mathbf{c}}_\tau$ que componen el vector de sustitución, se obtienen mediante una estimación MMSE asumiendo que cada parámetro (coeficientes LPC, señal de energía unitaria o ganancia) se corresponde con el centro $\mathbf{c}^{(v)}, 1 \leq v \leq C$, siendo C el tamaño del diccionario correspondiente. De este modo, la estimación $\hat{\mathbf{c}}_\tau^{(i)}$ para la trama $\tau-$ésima, en el instante $t+\tau$, conocido el índice de cuantización i del parámetro recibido, en el instante t, se obtiene como [84]:

$$\hat{\mathbf{c}}_\tau^{(i)} = \sum_{j=0}^{C-1} \mathbf{c}^{(j)} P(i_{t+\tau} = j | i_t = i), \qquad (1 \leq \tau \leq T), \tag{6.2}$$

Esta expresión parte de la información proporcionada en la última trama recibida correctamente, de la cual se extrae el correspondiente parámetro cuantizado $\mathbf{c}^{(j)}$ en el instante de tiempo t, para generar así la estimación $\hat{\mathbf{c}}_\tau^{(i)}$ en el instante de tiempo $t+\tau$. Dado que la obtención de estos vectores de sustitución podría incrementar el retardo durante la fase de decodificación, estos vectores de sustitución se generan previamente a partir de las observaciones en la base de datos de entrenamiento [84].

Generación de los vectores de sustitución

Partiendo de que la calidad de la estimación $\hat{\mathbf{c}}_\tau^{(i)}$ dependerá del diccionario empleado para cada parámetro, a continuación se detalla la obtención del correspondiente diccionario y la obtención de las estimaciones del correspondiente vector de sustitución.

Por un lado, en la estimación de la ganancia G se utiliza el clásico algoritmo *Lloyd-Max* [118], dado que se trata un valor escalar, para obtener un diccionario de tamaño C. La correspondiente estimación $\hat{c}_\tau^{(i)}$, para cada vector de sustitución $\mathbf{V}_G^{(i)}$, $\forall i \in [1, C]$, se obtiene como un promedio de las ganancias encontradas en la base de datos de entrenamiento en el instante de tiempo $t + \tau$ cuando se ha observado la ganancia cuantizada con el centro $c^{(i)}$ en el instante de tiempo t. De este modo, aunque no se calcula explícitamente la probabilidad de transición $P(i_{t+\tau} = j | i_t = i)$ de la expresión (6.2), sí está siendo utilizada de forma implícita con el procedimiento indicado.

Para la obtención de las estimas de los coeficientes LPC, se utiliza el algoritmo clásico de cuantización vectorial LBG [113] aplicado sobre su representación como coeficientes LSF. Como ya se indicó en la sección 2.4.2, el motivo de seleccionar los coeficientes LSF es que, bajo esta representación, se puede garantizar la estabilidad del filtro todo-polos que representa el tracto vocal [29, 30]. De nuevo, cada estimación $\hat{\mathbf{c}}_\tau^{(i)}$, para cada vector de sustitución $\mathbf{V}_{\text{LPC}}^{(i)}$, $\forall i \in [1, C]$, se obtiene realizando un promedio de los coeficientes LSF que aparecen en el instante de tiempo $t + \tau$ dado que en el instante de tiempo t se ha observado los coeficientes LSF cuantizados con el centro $\mathbf{c}^{(i)}$. Finalmente, estos vectores LSF se convierten de nuevo a vectores LPC que son los aplicados para realizar la síntesis.

Hay que tener en cuenta que este proceso de estimación de los coeficientes LPC admite un amplio margen de mejora. En la bibliografía se pueden encontrar otros trabajos como [74, 121–131] donde se aplican técnicas que ofrecen mejores estimaciones que las obtenidas mediante este proceso de estimación MMSE. No obstante, en esta tesis se mantendrá esta técnica de estimación sobre los coeficientes LPC para analizar el rendimiento que ofrecen las diferentes propuestas que mejoran la estimación de la señal de excitación mientras se emplea la técnica de vectores de sustitución.

Sin embargo, obtener una estimación para la señal de excitación, siguiendo el proceso análogo al comentado sobre los coeficientes LPC, no es una tarea sencilla, ya que la señal de excitación no dispone de una representación adecuada para la

estimación. Por este motivo, la mayoría de trabajos se centran en aplicar las técnicas de repetición e interpolación/extrapolación explicadas en la sección 3.4.1. Así, con el objetivo de proporcionar estimaciones de la señal de excitación para el esquema propuesto, se emplearán los métodos de cuantización presentados en la sección 5.2.1 para obtener diccionarios de tamaño C.

Una vez obtenido el correspondiente diccionario, la estimación $\hat{\mathbf{c}}_\tau^{(i)}$, para cada vector de sustitución $\mathbf{V}_{\text{EXC}}^{(i)}$, $\forall i \in [1, C]$, se obtiene considerando todas las señales de excitación que se encuentran en el instante $t+\tau$ en la base de datos de entrenamiento, como un conjunto $\mathcal{B}$, dado que en el instante de tiempo t se ha observado la señal de excitación de energía unitaria cuantizada con el centro $\mathbf{c}^{(i)}$. Finalmente, la estimación $\hat{\mathbf{c}}_\tau^{(i)}$ es el centroide del conjunto $\mathcal{B}$ tras aplicar el criterio de centro óptimo del algoritmo de cuantización empleado en la obtención del diccionario.

Una vez obtenidas las diferentes estructuras que recogen los vectores de sustitución de cada parámetro, se aprecia que el tamaño puede ser considerable dependiendo del tamaño de diccionario C, el número de tramas a recuperar en la ráfaga T y el número de componentes que conforman la propia estimación $(p, 1, N)$ correspondientes al número de coeficientes LPC, la ganancia y el número de muestras de la señal de excitación de energía unitaria respectivamente. Así el tamaño de las estructuras queda como: $C \cdot T \cdot p$ para el caso de los coeficientes LPC, $C \cdot T \cdot 1$ para la ganancia y $C \cdot T \cdot N$ para la señal de energía unitaria.

6.2.2. Estimación de parámetros para órdenes superiores

En la sección anterior, se ha presentado el caso de considerar una única trama previa a la ráfaga como punto de partida para realizar las estimaciones. Sin embargo, la expresión (6.2) correspondiente a la predicción de primer orden puede extenderse para considerar más tramas previas como se muestra en la figura 6.2. En este ejemplo, se considera el par de índices de cuantización, $\mathbf{i} = (i_1, i_2)$, del parámetro correspondiente en las dos tramas previas a la pérdida, para seleccionar los correspondientes vectores de sustitución.

Siguiendo con el procedimiento indicado en la sección anterior, sólo habría que modificar la probabilidad de transición en la expresión (6.2), de acuerdo al orden de predicción $\mathcal{O}$. De este modo, la estimación $\hat{\mathbf{c}}_\tau^{(\mathbf{i})}$, donde $\mathbf{i}$ representa el conjunto de índices de los parámetros previos a la pérdida, se puede calcular para la trama

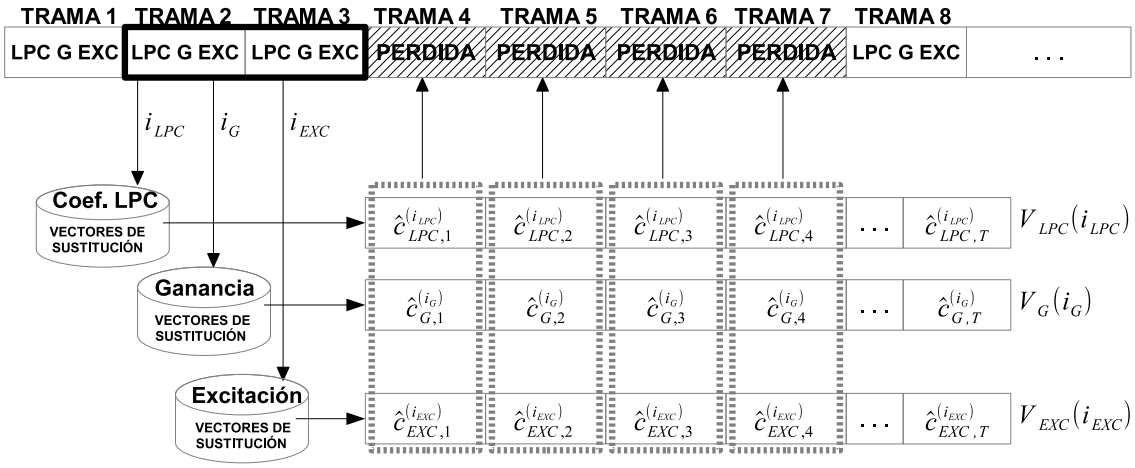

Figura 6.2: Expansión del esquema propuesto para mitigación de errores basado en vectores de sustitución donde en este caso se tienen en cuenta las dos últimas tramas recibidas correctamente para obtener los correspondientes vectores de sustitución $\mathbf{V}_{\text{LPC}}$, $\mathbf{V}_{\text{EXC}}$ y $\mathbf{V}_G$, correspondientes a los vectores de índices de cuantización obtenidos ($\mathbf{i}_{\text{LPC}}$, $\mathbf{i}_{\text{EXC}}$ y $\mathbf{i}_G$).

$\tau-$ésima como [84]:

$$\hat{\mathbf{c}}_{\tau}^{(\mathbf{i})} = \sum_{v=0}^{C-1} \mathbf{c}^{(j)} P(i_{t+\tau} = j | i_t, i_{t-1}, ..., i_{t-\mathcal{O}}), \qquad (1 \leq \tau \leq T), \tag{6.3}$$

Con este resultado y de manera análoga a lo comentado en la sección anterior, se puede obtener la estimación $\hat{\mathbf{c}}_{\tau}^{(\mathbf{i})}$ en el instante de tiempo $t+\tau$ de acuerdo al parámetro cuantizado como $\mathbf{c}^{(\mathbf{i})}$ en el instante de tiempo t para el parámetro correspondiente.

Sin embargo, hay que tener en cuenta que a medida que se consideren más tramas anteriores para obtener los vectores de sustitución, las estructuras que se necesitan almacenar van a ser cada vez mayores, del orden de $C^{\mathcal{O}}$, por lo que el tamaño final ya no podría ser almacenado en determinados dispositivos con memoria limitada. Además, hay que tener en cuenta que muchas de las probabilidades para cada uno de los posibles índices de cuantización no se pueden conocer si estas combinaciones no aparecen en toda la base de datos de entrenamiento, por lo que se estaría almacenando memoria que no va a ser útil para obtener buenas estimaciones [84].

Para aliviar este problema, en [84], se plantea considerar sólo aquellos conjuntos de índices $\mathbf{i}$ que tienen un determinado orden de aparición en la base de datos, cuyo valor se denota como μ. De esta manera, se dispondrá de un conjunto de vectores de sustitución menor y en los casos donde el conjunto de índices $\mathbf{i}$ no

se encuentre, se accederá al conjunto de vectores de sustitución de orden inferior o si esto supone un coste excesivo en memoria, considerar sólo del conjunto de primer orden comentado en la sección anterior. De este modo, será necesario para la estructura de un determinado orden $\mathcal{O}$, un vector que especifique qué índices de cuantización se han entrenado correctamente.

6.2.3. Enfoque mixto de técnicas de reconstrucción basadas en el emisor y receptor

Partiendo del esquema de mitigación de pérdidas presentado en la figura 6.1, es posible realizar una combinación de las técnicas de recuperación basadas en el emisor y el receptor, como el que se muestra en la figura 6.3, y así aprovechar sus ventajas. Por un lado, la técnica de mitigación basada en vectores de sustitución proporciona estimaciones de los parámetros en los paquetes perdidos durante la ráfaga. Por otro lado, el uso del código FEC permite reducir el error de síntesis en la última trama perdida y además, evitar la propagación del error en los codecs basados en el paradigma CELP.

Sin embargo, hay que tener en cuenta que este código FEC estaría compuesto por los coeficientes LPC, la ganancia y la señal de excitación de energía unitaria, que puede suponer un incremento en la tasa de bits, inviable para un canal con un ancho de banda limitado, y la incompatibilidad con el codec estándar en transmisiones sin pérdidas. De nuevo, para reducir el incremento en la tasa de bits y mantener la compatibilidad con el codec estándar, se ha considerado el uso de la técnica esteganográfica comentada en la sección 5.2.3. No obstante, dado que la calidad perceptual decrece a medida que la tasa de bits a ocultar se incrementa, se ha decidido que en el esquema de la figura 6.3, sólo la señal de excitación de energía unitaria sea codificada como código FEC. De esta forma, los coeficientes LPC y la ganancia correspondientes a la última trama perdida en la ráfaga son obtenidos del correspondiente vector de sustitución.

6.2.4. Resultados experimentales

Para analizar el rendimiento de los esquemas propuestos en las figuras 6.1, 6.2 y 6.3, las pruebas se han realizado sobre el conjunto de test de la base de datos TIMIT [104] y las condiciones de canal se han simulado siguiendo un modelo Gilbert (comentado en la sección 3.5.3). Así se ha medido la calidad de la señal recuperada,

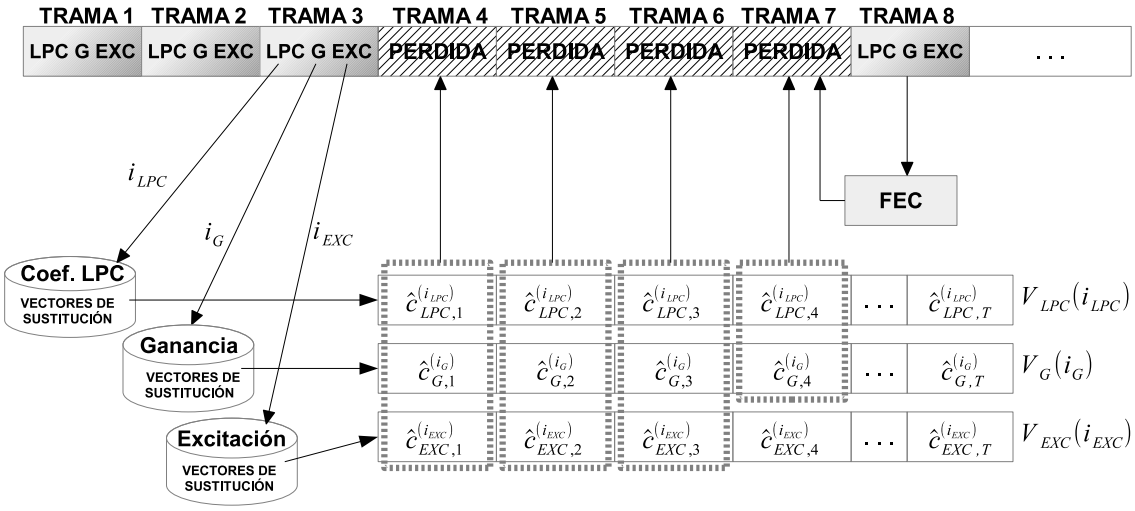

Figura 6.3: Esquema propuesto que emplea la técnica de esteganografía y un enfoque mixto entre el uso de los vectores de sustitución, como técnica de mitigación basada en el receptor, y el uso del código FEC, como técnica de prevención basada en el emisor. Los vectores de sustitución se utilizarán para estimar los parámetros de las tramas perdidas en la ráfaga mientras que el código FEC proporciona la señal de excitación (EXC) para minimizar el error en la última trama perdida y al mismo tiempo reducir la propagación del error.

en términos de calidad objetiva PESQ [103], con diferentes tasas de pérdidas de paquetes o *Packet Error Rate* (PER), PER= $\{10\,\%, 20\,\%, 30\,\%, 40\,\%, 50\,\%\}$, y con diferentes longitudes promedio de ráfaga o *Average Burst Length* (ABL), ABL= $\{1, 2, 4, 8, 12\}$. Todas las propuestas se analizaron respecto al funcionamiento del codec AMR [39] en el modo 12.2 kbps, ya que es un codec basado en el paradigma CELP, y así se verá claramente el rendimiento de los esquemas 6.1 y 6.3 en cuanto a la reducción de la propagación del error.

Los vectores de sustitución empleados en las propuestas tienen unos requisitos de memoria de acuerdo al número de centros en el diccionario C, el número de coeficientes LPC p, el número de muestras de la señal de excitación N y el valor máximo de reemplazos que se puede efectura con el vector de sustitución T. Así, en estas pruebas se ha considerado: $C = 1024$, $p = 10$, $N = 160$ y $T = 20$ respectivamente. No obstante, si una ráfaga tuviera una longitud superior a 20, se aplicaría la repetición de la última estimación.

Además, en las pruebas desarrolladas sobre cada propuesta, se ha considerado medir el rendimiento de la estimación de la señal de excitación cuando se emplean alguno de los métodos de cuantización presentados en la sección 5.2.1. Para identifi-

car cada caso, a continuación se especifican una serie de siglas de acuerdo al método de cuantización empleado:

- **LBG**: Aplicación del algoritmo clásico LBG sobre la señal de excitación (Sección 5.2.1).
- **LBGK**: Modificación de la búsqueda de centro óptimo basada en kmedoids (Sección 5.2.1).
- **SLBG**: Modificación de celda y centro óptimos basada en distancia de síntesis (Sección 5.2.1).
- **FSLBG**: Se aplica el algoritmo SLBG pero en cada iteración se aplicará sólo sobre la celda de mayor tamaño hasta alcanzar el número de centroides final (Sección 5.2.1).
- **DSLBG**: Se aplica el algoritmo SLBG pero en cada iteración se divide la celda de mayor tamaño y se vuelve a aplicar el método SLBG sobre los nuevos centroides hasta alcanzar el número de centroides final (Sección 5.2.1).

Tanto en el método FSLBG como DSLBG, se ha considerado un valor de $\alpha = 500$ componentes como el número mínimo que tiene que tener una celda para que ésta pueda ser divida y se alcancen los $C = 1024$ centroides del diccionario.

En primer lugar, en la tabla 6.1 se presentan los resultados PESQ promedio, en las distintas condiciones de canal, sobre el codec original AMR, con su propio algoritmo PLC, y las propuestas que aplican vectores de sustitución. En primer lugar, se presentan los resultados cuando se aplica sólo la técnica de los vectores de sustitución sobre los coeficientes LPC (RLPC). A continuación, se presentan los resultados cuando se aplica sobre todos los parámetros de voz (RV), representado en la figura 6.1. Por último, se presentan los resultados cuando se aplica un código FEC para evitar la propagación del error (RVFEC), representado en la figura 6.3. En estas propuestas, el método de cuantización para obtener el diccionario empleado ha sido el método FSLBG.

Como se puede apreciar, la aplicación de la técnica de los vectores de sustitución sobre todos los parámetros de voz (RV), ofrece una mejora significativa respecto a los resultados obtenidos con el codec AMR y la propuesta RLPC. De este modo, se aprecia claramente la influencia de la estimación de la señal de excitación frente a la obtenida con el algoritmo PLC implementado por el propio codec. Además,

		Longitud promedio de ráfaga				
	PER	1	2	4	8	12
AMR	10	2,82	2,80	2,82	2,87	2,93
	20	2,27	2,25	2,22	2,23	2,30
	30	1,83	1,86	1,79	1,78	1,83
	40	1,35	1,54	1,46	1,45	1,45
	50	1,01	1,25	1,16	1,14	1,18
RLPC	10	2,86	2,82	2,87	2,89	2,94
	20	2,34	2,31	2,29	2,24	2,32
	30	1,89	1,94	1,83	1,81	1,85
	40	1,41	1,60	1,50	1,46	1.47
	50	1,04	1,29	1,21	1,16	1,20
RV	10	2,99	2,99	3,01	3,06	3,08
	20	2,51	2,53	2,54	2,58	2,60
	30	2,01	2,08	2,14	2,19	2,24
	40	1,67	1,82	1,88	1,93	2.00
	50	1,32	1,54	1,60	1,64	1,69
RVFEC	10	3,15	3,18	3,20	3,23	3,26
	20	2,69	2,73	2,76	2,78	2,81
	30	2,14	2,21	2,27	2,31	2,35
	40	1,74	1.94	1,99	2,02	2,08
	50	1,39	1,60	1,65	1,70	1,74

Tabla 6.1: Resultados PESQ promedio obtenidos sobre diferentes condiciones de canal con diferentes tasas de error (PER) y longitudes promedio de ráfaga sobre el codec AMR (12.2 Kbps), la técnica de vectores de sustitución sobre los coeficientes LPC (RLPC), la técnica de vectores de sustitución sobre todos los parámetros de voz (RV) y la propuesta que incorpora un código FEC (RVFEC) para evitar la propagación del error.

la técnica de vectores de sustitución ofrece un buen rendimiento en ráfagas largas, donde el algoritmo PLC podría estar aplicando el proceso de apagado. Por último, dado que el codec AMR presenta una dependencia inter-trama para obtener la señal de excitación, también se observa que la propuesta RVFEC ofrece un rendimiento superior a la propuesta RV, ya que se está reduciendo la propagación del error. Finalmente, como ya se indicó en el Capítulo 5, la inclusión del código FEC en la propia trama no afectará significativamente a la calidad perceptual en transmisiones sin pérdidas (4.00 vs 3.99) y mantiene la compatibilidad con el codec original.

Por otro lado, en la tabla 6.2 se recogen los resultados PESQ promedio sobre el codec base AMR, utilizando su propio algoritmo PLC, y de las propuestas que hacen uso del primer y segundo orden de predicción $\mathcal{O}$, representadas en las figuras

	Tasa de pérdida de paquetes - longitud promedio ráfaga								
	10-1	**10-4**	**10-12**	**20-1**	**20-4**	**20-12**	**40-1**	**40-4**	**40-12**
AMR	2.82	2.82	2.93	2.27	2.22	2.27	1.35	1.45	1.45
$\mathcal{O} = 1$									
LBG	2.88	2.91	2.95	2.33	2.36	2.39	1.55	1.70	1.76
LBGK	2.93	2.96	3.00	2.41	2.45	2.51	1.60	1.76	1.88
SLBG	2.99	3.01	3.08	2.51	2.53	2.60	1.67	1.82	1.94
FSLBG	3.06	3.10	3.15	2.56	2.57	2.64	1.81	1.93	2.08
DSLBG	3.09	3.13	3.18	2.59	2.61	2.68	1.83	1.95	2.11
$\mathcal{O} = 2$									
LBG	2.95	2.99	3.04	2.41	2.43	2.48	1.61	1.75	1.80
LBGK	3.02	3.06	3.11	2.46	2.48	2.54	1.66	1.80	1.92
SLBG	3.04	3.07	3.13	2.56	2.58	2.66	1.73	1.86	1.98
FSLBG	3.13	3.17	3.21	2.61	2.62	2.69	1.86	1.98	2.12
DSLBG	3.17	3.20	3.25	2.65	2.67	2.72	1.88	2.00	2.16

Tabla 6.2: Resultados promedio PESQ obtenidos sobre el codec AMR original y las propuestas de vectores de sustitución para diferentes órdenes de predicción ($\mathcal{O}$) y en diferentes condiciones de canal, según el ratio de pérdidas de paquetes y la longitud promedio de ráfaga.

6.1 y 6.2 respectivamente. Estas pruebas se han realizado con el mismo entorno de trabajo presentado anteriormente y en ellas se analiza el rendimiento obtenido de las diferentes estimaciones obtenidas aplicando los distintos métodos de cuantización de la señal de excitación comentados en la sección 5.2.1.

A la vista de los resultados, se puede confirmar que, a medida que se mejora el diccionario para la señal de excitación, menor es el error de cuantización para obtener el índice de cuantización (i_{EXC}) y mejores son las estimaciones para el correspondiente vector de sustitución. Así, la propuesta que ofrece mejores resultados es la que se obtiene aplicando el método de cuantización DSLBG. No obstante, dada la pequeña diferencia y que el coste computacional del método FSLBG es significativamente menor respecto al método DSLBG, el método FSLBG puede resultar más adecuado para la obtención rápida de diccionarios.

Para terminar este análisis, cabe destacar que los resultados obtenidos para el orden $\mathcal{O} = 2$ son ligeramente superiores a los de orden $\mathcal{O} = 1$, dado que se dispone de más información previa. De esta forma, los vectores de sustitución disponen de mejores estimaciones.

6.3. Esquema de mitigación basado en vectores de sustitución y filtro RLS

A pesar de la mejora que supone la técnica de los vectores de sustitución, ya que reduce el impacto de la pérdida de paquetes producida por ráfagas en las transmisiones sobre redes IP, esta técnica tiene una fuerte dependencia con la cuantización de los parámetros previos a la pérdida. Es decir, los índices de cuantización resultantes pueden afectar tanto a la selección del correspondiente vector de sustitución como a la obtención de las correspondientes estimaciones. Como consecuencia, la estimación podría ser de peor calidad a la proporcionada por el propio algoritmo PLC en las primeras pérdidas.

Aunque este problema se resolvería obteniendo diccionarios con mayor número de centroides, ya se comentó, en la sección 5.2.1, que el coste computacional para obtener estos diccionarios, en particular sobre la señal de excitación, sería enorme dado el crecimiento exponencial en recursos que emplea el algoritmo clásico LBG [113]. Por este motivo, esta sección se centra en mejorar la calidad de la reconstrucción de la señal mejorando la estimación de la señal de excitación en las primeras pérdidas.

Para ello, se complementa el funcionamiento de la técnica de los vectores de sustitución con la aplicación de un filtro adaptativo, basado en corrección de error recursivo o *Recursive Least Squares* (RLS), para mejorar la estimación de la señal de excitación en las primeras pérdidas de una ráfaga. Como resultado de este filtrado, la señal de excitación generada puede interpretarse como resultado de un filtrado similar al filtro LTP en los codecs basados en el paradigma CELP. No obstante, el éxito de esta propuesta requiere de buscar un equilibrio entre el uso de la técnica RLS y de los vectores de sustitución. El motivo es que el filtro adaptativo requiere de una realimentación o *feedback* para que vaya actualizando sus parámetros en el tiempo, por lo que la calidad perceptual de la señal recuperada decrecerá muy rápido a medida que la ráfaga sea de longitud mayor.

6.3.1. Filtro adaptativo de corrección recursiva

El filtro adaptativo de corrección recursiva o *Recursive Least Squares* (RLS) se caracteriza por ir modificando los parámetros del filtro a medida que va cambiando la señal de entrada al sistema en el tiempo. Este es el caso de la señal de voz, ya que no es estacionaria en el tiempo, por lo que se ha decidido utilizar este tipo de

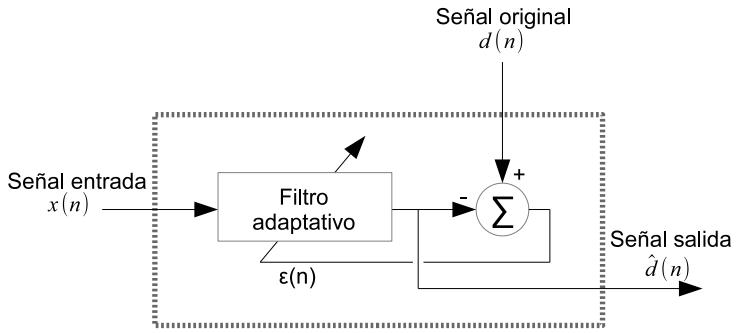

Figura 6.4: Esquema general de funcionamiento de un filtro adaptativo.

filtro para obtener una estimación de la señal de excitación a partir de su historia previa. El esquema general de funcionamiento de este tipo de filtros se puede ver en la figura 6.4, donde a partir de una señal de entrada $x(n)$ se obtiene una señal estimada $\hat{d}(n)$. Esta señal estimada genera un error $\varepsilon(n)$, respecto a la señal original $d(n)$, que es el responsable de ir modificando los parámetros del filtro adaptativo en cada instante de tiempo t.

En la bibliografía, los filtros adaptativos se basan en dos algoritmos: el algoritmo de corrección de error medio o *Least Mean Squares* (LMS) y el algoritmo de corrección de error recursivo o *Recursive Least Squares* (RLS)[132, 133]. De los dos, el algoritmo RLS es computacionalmente más complejo que el algoritmo LMS, pero tiene una convergencia más rápida [133, 134].

El rendimiento del algoritmo RLS depende de la configuración de dos parámetros: el número de muestras pasadas consideradas, que determinan la longitud del filtro, L, y el factor de olvido, que es el responsable de la convergencia y estabilidad del algoritmo RLS, λ [135]. De este modo, el vector de entrada $\mathbf{x}_n$, en el instante de tiempo $t = n$, está compuesto por las últimas L muestras de la señal de excitación previa, $\mathbf{x}_n = [e(n)\ e(n-1)\ \ldots\ e(n-L)]'$. Partiendo de este vector de entrada, se obtendrá la señal de excitación final, $(\hat{e}(n) = \hat{d}(n))$, de la trama actual m como:

$$\hat{d}(n) = \sum_{l=0}^{L} w_n(l)x(n-l) = \mathbf{w}'_n\mathbf{x}_n \tag{6.4}$$

donde $\mathbf{w}$ son los coeficientes actuales del filtro que se actualizan conforme se procesan tramas correctas y $[\cdot]'$ representa el vector traspuesto. Inicialmente, los coeficientes del filtro $\mathbf{w}$ se podrían calcular bajo un proceso de mínimos cuadrados, que minimiza el error para cada paquete o trama recibida. Sin embargo, empleando un filtro RLS

es posible estimar los coeficientes $\mathbf{w}_n$ en términos de los coeficientes previos $\mathbf{w}_{n-1}$.

Para ello, el algoritmo RLS tiene que minimizar la función de coste $J(\mathbf{w})$ a partir de los coeficientes del filtro actuales $\mathbf{w}_n$ y el error $\varepsilon(n) = (d(n) - \hat{d}(n))^2$. Esta función de coste se define como [133]:

$$J(\mathbf{w}_n) = \sum_{i=1}^{n} \lambda^{n-i} \varepsilon^2(i) \tag{6.5}$$

donde λ es el factor de olvido que tiene en cuenta las muestras anteriores con un peso menor, dando más importancia a las muestras más recientes. Este factor de olvido está limitado a valores comprendidos entre $0 < \lambda \leq 1$ y tiene que determinarse para cada aplicación donde utilicemos el algoritmo RLS. Si se reemplaza la definición del error $\varepsilon(n)$, los coeficientes $\mathbf{w}_n$ se obtendrán minimizando la función de coste tomando derivadas parciales sobre cada componente k-ésima del vector $\mathbf{w}_n$ en el intervalo $[0, L]$. De este modo se obtienen las siguientes expresiones [132, 133]:

$$\sum_{i=0}^{n} \lambda^{n-i} \Big[d(i) - \sum_{l=0}^{L} w_n(l) x(i-l) \Big] x(i-k) = 0, \quad k = 0, 1, ..., L \tag{6.6}$$

donde reagrupando términos se puede obtener la siguiente expresión equivalente:

$$\sum_{l=0}^{L} w_n(l) \Big[\sum_{i=0}^{n} \lambda^{n-i} x(i-l) x(i-k) \Big] = \sum_{i=0}^{n} \lambda^{n-i} d(i) x(i-k) \quad k = 0, 1, ..., L \tag{6.7}$$

y que puede expresarse en términos de matrices como:

$$\mathbf{R}_n \mathbf{w}_n = \mathbf{S}_n \tag{6.8}$$

donde $\mathbf{R}_n$ es la matriz de covarianza pesada del vector $\mathbf{x}_n$ y $\mathbf{S}_n$ es la matriz de correlación cruzada pesada entre los vectores $\mathbf{d}_n$ y $\mathbf{x}_n$. A partir de esta formulación matricial, los coeficientes $\mathbf{w}_n$ se pueden obtener como [132, 133]:

$$\mathbf{w}_n = \mathbf{R}_n^{-1} \mathbf{S}_n \tag{6.9}$$

Teniendo en cuenta que el método RLS obtiene los coeficientes de manera recursiva como:

$$\mathbf{w}_n = \mathbf{w}_{(n-1)} + \Delta \mathbf{w}_{(n-1)} \tag{6.10}$$

donde $\Delta\mathbf{w}_{n-1}$ es un factor de corrección en el instante de tiempo $n-1$. Igualmente, se pueden calcular también las matrices de covarianza $\mathbf{R}_n$ y covarianza cruzada $\mathbf{S}_n$ de manera recursiva como [132, 133]:

$$\begin{aligned}\mathbf{R}_n &= \sum_{i=0}^{n} \lambda^{n-i}\mathbf{x}_i\mathbf{x}_i + \lambda^0\mathbf{x}_n\mathbf{x}'_n = \lambda\mathbf{R}_{(n-1)} + \mathbf{x}_n\mathbf{x}'_n \\ \mathbf{S}_n &= \sum_{i=0}^{n} \lambda^{n-i}\mathbf{d}_i\mathbf{x}_i + \lambda^0\mathbf{d}_n\mathbf{x}'_n = \lambda\mathbf{S}_{(n-1)} + \mathbf{d}_n\mathbf{x}_n\end{aligned} \tag{6.11}$$

Este cálculo recursivo evita el cálculo de la matriz inversa $\mathbf{R}^{-1}$ en la expresión (6.8) para cada iteración, gracias al uso del lema de la inversión de matriz definido en [132]. De este modo, llamando $\mathbf{P}$ a la matriz $\mathbf{R}^{-1}$, esta matriz puede obtenerse recursivamente como:

$$\mathbf{P}_n = \lambda^{-1}\mathbf{P}_{(n-1)} + \lambda^{-1}\mathbf{g}_n\mathbf{x}_n \tag{6.12}$$

donde $\mathbf{g}_n$ es la ganancia de Kalman que se define como [132, 133]:

$$\mathbf{g}_n = \frac{\lambda^{-1}\mathbf{P}_{(n-1)}\mathbf{x}_n}{1 + \lambda^{-1}\mathbf{x}'_n\mathbf{P}_{(n-1)}\mathbf{x}_n} \tag{6.13}$$

Tras aplicar el lema de la inversión de matriz, finalmente se obtienen los coeficientes del filtro w_n de la forma:

$$\mathbf{w}_n = \mathbf{w}_{(n-1)} + \mathbf{g}_n\big(\mathbf{d}_n - \mathbf{x}'\mathbf{w}_{(n-1)}\big) \tag{6.14}$$

donde $\mathbf{g}_n\big(\mathbf{d}_n - \mathbf{x}'\mathbf{w}_{(n-1)}\big)$ es el factor de corrección $\Delta\mathbf{w}_{n-1}$, conocido también como el error a priori.

6.3.2. Esquema de mitigación que combina filtrado adaptativo y vectores de sustitución

El esquema que se propone para combinar tanto el uso del filtro adaptativo RLS como la técnica de los vectores de sustitución puede verse en la figura 6.5. Este esquema es muy similar al presentado en la sección 6.2, salvo que en las primeras estimaciones de la señal de excitación se emplea la técnica de filtrado adaptativo RLS, cuyos parámetros L y λ se han obtenido de manera experimental. Para ello, se ha ido modificando L de 1 a 160 y λ de 0.1 a 1 y se ha buscado, sobre la base de datos de entrenamiento, el par (L, λ) que obtiene mejores prestaciones en promedio

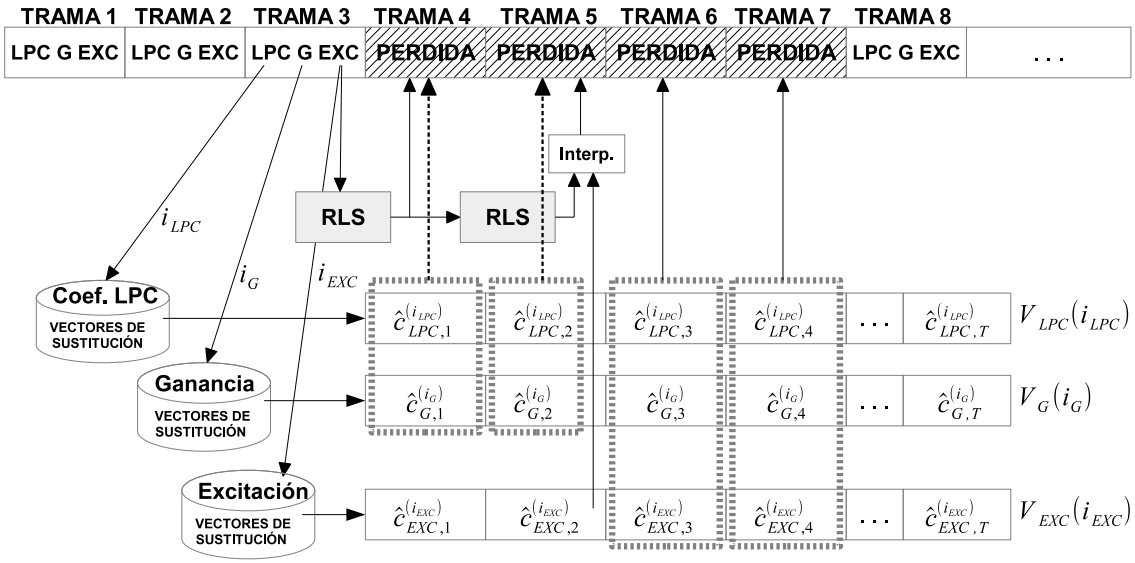

Figura 6.5: Esquema de mitigación de errores que incorpora la técnica RLS junto con la técnica de los vectores de sustitución para mejorar la estimación de la señal de excitación (EXC). En las primeras tramas perdidas de una ráfaga se emplea el filtro adaptativo RLS y posteriormente se utiliza el correspondiente vector de sustitución hasta el final de la ráfaga. Los parámetros de ganancia y coeficientes LPC obtienen la estimación desde el correspondiente vector de sustitución desde el inicio de la ráfaga.

sobre diferentes condiciones de canal, simulado empleando un modelo Gilbert como se indicó en la sección 3.5.3. En la obtención de este par, se ha priorizando los valores L menores para que las matrices $\mathbf{R}_n$ y $\mathbf{S}_n$ no sean demasiado grandes. Así se pudo obtener unos valores de $L = 13$ y $\lambda = 0{,}985$.

Una vez configurado el filtro adaptativo RLS, hay que tener en cuenta su bajo rendimiento cuando hay ráfagas largas. Por este motivo, es necesario aplicar un enfoque mixto entre las técnicas RLS y los vectores de sustitución para aprovechar la ventaja de ambas técnicas. Por un lado, obtener mejores estimaciones para la señal de excitación en las primeras pérdidas. Por otro lado, obtener las estimaciones del correspondiente vector de sustitución para evitar la gran caída en calidad perceptual que produciría la técnica RLS. El objetivo, por tanto, será encontrar la trama $\mathcal{R}$-ésima de la ráfaga donde ambas técnicas se complementen y se obtengan los mejores resultados.

6.3.3. Resultados experimentales

Para conseguir que ambas técnicas funcionen de manera complementaria siguiendo la propuesta de la figura 6.5, en primer lugar, hay que determinar la trama $\mathcal{R}$ donde se obtienen los mejores resultados. Para ello, se ha utilizado el conjunto de entrenamiento de la base de datos TIMIT [104], con la que se obtendrán los resultados PESQ [103] promedio en diferentes condiciones de canal. Condiciones que serán simuladas mediante un modelo de canal Gilbert, como se indicó en la sección 3.5.3, con diferentes tasas de pérdida de paquetes o *Packet Error Rate* (PER), PER= $\{10\,\%, 20\,\%, 30\,\%, 40\,\%, 50\,\%\}$, y longitud promedio de ráfagas o *Average Burst Length* (ABL), ABL= $\{1, 2, 4, 8, 12\}$. Por último, para generar de los vectores de sustitución se han obtenido los diccionarios de cuantización, con 1024 centroides, siguiendo el método de cuantización vectorial LBG para los coeficientes LPC, en su representación como coeficientes LSF, el método de cuantización *Lloyd-Max* para la ganancia y el método de cuantización basado en distancia de síntesis y con división fija (FSLBG), comentado en la sección 5.2.1, para la señal de excitación de energía unitaria.

En la tabla 6.3 se presentan los resultados de este entrenamiento sobre el codec iLBC [45] empleando el nuevo esquema de mitigación sobre la señal de excitación, que denominaremos RLSRV. Así, en primer lugar se aplica la técnica RLS hasta la trama $\mathcal{R}-1$ perdida de manera consecutiva en la ráfaga, posteriormente, sobre la trama $\mathcal{R}$ se aplica una interpolación lineal entre las estimaciones RLS y del correspondiente vector de sustitución (RV) y finalmente, se aplica la técnica (RV) hasta el final de la ráfaga. Así, se han obtenido los resultados promediando los resultados PESQ por la longitud promedio de ráfaga (derecha) y la tasa de pérdida de paquetes (izquierda) respectivamente. Como se puede apreciar, los mejores resultados se obtienen cuando la interpolación se produce en la segunda trama perdida en la ráfaga.

Partiendo de este contexto, en la tabla 6.4 se puede apreciar el rendimiento de la técnica propuesta RLSRV (empleando diccionarios de diferente tamaño para la señal de excitación) en comparación con los resultados que se obtienen para el codec estándar iLBC y las técnicas RV y RLS por separado. Estas pruebas se han realizado sobre el conjunto de test de la base de datos TIMIT [104] y empleando las mismas condiciones de canal presentadas anteriormente. Los valores obtenidos son un promediando del resultado PESQ por la longitud promedio de ráfaga (derecha) y la tasa de pérdida de paquetes (izquierda) respectivamente.

		Tasa de pérdida de paquetes					Longitud de ráfaga promedio				
	$\mathcal{R}$	10 %	20 %	30 %	40 %	50 %	1	2	4	8	12
RLSRV	1	3,11	2,75	2,48	2,26	2,10	2,67	2,43	2,49	2,51	2,57
	2	**3,12**	**2,76**	**2,49**	**2,28**	**2,11**	**2,67**	**2,45**	**2,51**	**2,54**	**2,58**
	3	3,11	2,75	2,48	2,27	2,10	2,67	2,44	2,50	2,53	2,57
	4	3,11	2,73	2,46	2,25	2,09	2,67	2,44	2,49	2,51	2,54
	6	3,10	2,72	2,45	2,24	2,08	2,67	2,44	2,47	2,49	2,52

Tabla 6.3: Resultado PESQ promedio sobre diferente tasa de pérdida de paquetes (izquierda) y diferente longitud promedio de ráfaga (derecha) para encontrar el mejor rendimiento en la propuesta (RLSRV), que combina las técnicas RLS y los vectores de sustitución RV, cuando se aplica la interpolación sobre un paquete diferente ($\mathcal{R}$) en la ráfaga.

	Tasa de pérdida de paquetes					Longitud de ráfaga promedio				
	10 %	20 %	30 %	40 %	50 %	1	2	4	8	12
iLBC	3,03	2,56	2,22	1,96	1,75	2,62	2,34	2,20	2,18	2,18
RV	3,06	2,69	2,42	2,21	2,01	2,57	2,37	2,43	2,48	2,53
RLS	3,07	2,62	2,29	2,03	1,83	2,67	2,40	2,26	2,24	2,24
RLSRV(1024)	3,12	2,76	2,49	2,27	2,12	2,67	2,46	2,52	2,54	2,59
RLSRV(2048)	3.19	2,83	2,58	2,36	2,19	2,75	2,53	2,60	2,62	2,66

Tabla 6.4: Resultado PESQ promedio con diferentes tasas de pérdida de paquetes (izquierda) y longitud de ráfaga promedio (derecha) sobre el codec iLBC, las técnicas de vectores de sustitución (RV) y filtro adaptativo (RLS) y la propuesta que une ambas en la segunda trama consecutiva perdida y empleando un diccionario de 1024 (RLSRV 1024) y 2048 (RLSRV 2048) centroides respectivamente.

A la vista de los resultados, las técnicas RLS y RV mejoran significativamente la calidad perceptual de la señal de voz obtenida respecto al codec iLBC. No obstante, se observa un rendimiento ligeramente inferior de la técnica RV con ráfagas cortas, así como en la técnica RLS a medida que la ráfaga es de mayor longitud. Del mismo modo, se observa que la técnica RLSRV (con 1024 centroides) obtiene los mejores resultados al aprovechar las características de ambas técnicas. También se puede comprobar que la propuesta que emplea 2048 centroides mejora los resultados, ya que obtiene mejores estimaciones y se reduce el error de cuantización para seleccionar el vector de sustitución correspondiente.

Para finalizar esta sección, comentar que la mejora significativa observada con el test objetivo PESQ se puede corroborar con el test subjetivo MUSHRA [99] en la figura 6.6. En este test se comparó la calidad de la propuesta RLSRV con 1024 centros con el propio codec iLBC y las técnicas RV y RLS siguiendo el proceso

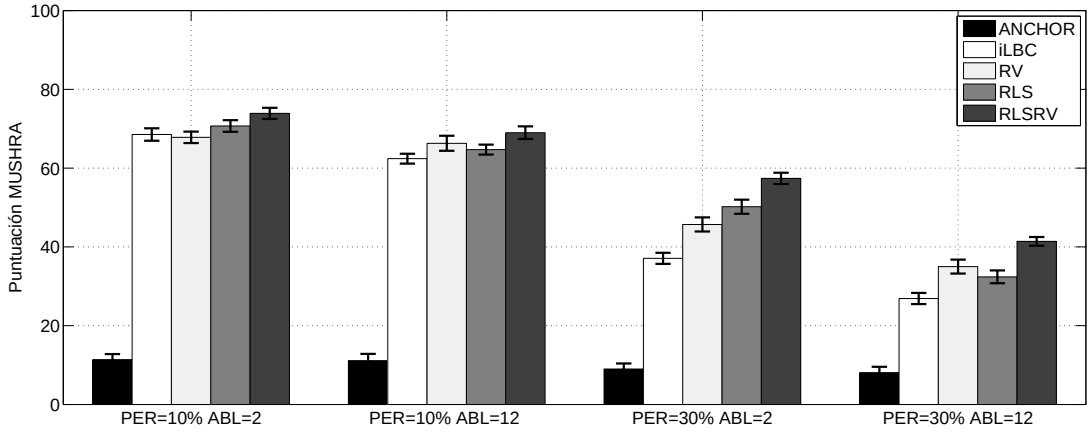

Figura 6.6: Resultados del test MUSHRA sobre la propuesta RLSRV con 1024 centros en comparación con los resultados alcanzados por las técnicas RLS, RV, el propio codec iLBC con su propio algoritmo PLC, todos ellos comparados con la peor situación (anchor), en diferentes tasas de pérdida de paquetes (PER) y longitud promedio de ráfaga (ABL).

indicado en la sección 3.5.1. De este modo se han realizado las pruebas con 20 oyentes y se han empleando 10 frases de voz de la base de datos Albayzin [106]. En la prueba se escogieron las señales de voz obtenidas en las siguientes condiciones de canal: tasas de pérdidas de 10 % y 30 % y longitud de ráfagas promedio de 2 y 12 tramas. A la vista de los resultados, se puede comprobar que la propuesta RLSRV ofrece un rendimiento superior, confirmando los resultados PESQ obtenidos en la tabla 6.4.

6.4. Esquema de mitigación basado en transformada Wavelet

Hasta ahora, en los esquemas de mitigación de pérdidas propuestos, la señal de excitación se ha considerado como una unidad indivisible a la hora de generar el diccionario de cuantización y posteriormente obtener las estimaciones para los vectores de sustitución. Sin embargo, dado el tamaño de este vector (generalmente $N = 160$ muestras) no es sencillo reducir el error de cuantización y por tanto obtener mejores estimaciones.

Aunque dividir un vector en trozos de menor tamaño podría facilitar su cuantización, hay que tener en cuenta que la señal sintetizada, a partir de los trozos

obtenidos mediante la técnica de vectores de sustitución, tiene que ser continua y que minimice el error de síntesis con la señal de voz original. Con este objetivo, en esta tesis se presenta una nueva representación, basada en la transformada wavelet, que permite dividir la señal de excitación en particiones de menor tamaño y también minimizar el error de síntesis sobre cada una de ellas de manera independiente.

6.4.1. Representación de la señal de excitación basada en wavelet

Uno de los objetivos destacados de esta tesis es el desarrollo de una nueva representación de la señal de excitación orientada a estimación. Con este propósito se plantea el uso de la transformada wavelet. A continuación, se presenta una breve descripción teórica de la transformada wavelet, su particularización en la transformada wavelet Haar y finalmente cómo se ha aplicado en el contexto de la transmisión de voz.

Introducción a la transformada wavelet

La transformada wavelet surgió con el objetivo de cubrir el problema que presenta la transformada de Fourier para poder analizar una señal tanto en el tiempo como en la frecuencia. Sin embargo, la transformada de Fourier presenta el problema de no poder identificar cambios bruscos en la señal al perder información temporal tras realizar la transformada en el dominio de la frecuencia. Por este motivo, era necesario introducir una variación de esta transformada que permitiera representar las características de la función y al mismo tiempo analizar la señal en tiempo y frecuencia. Ya en 1946, Dennis Gabor plantea una modificación de la transformada de Fourier permitiendo este análisis mediante el proceso de enventanado, conocida como *Short Time Fourier Transform* (STFT), con el que poder analizar las componentes de altas frecuencias usando ventanas temporales cortas o las componentes de bajas frecuencias usando ventanas más anchas. Sin embargo, no se puede realizar un análisis empleando ambos procesos de enventanado, mientras que ésta es la principal ventaja que nos proporcionar la transformada wavelet.

La primera alusión en la bibliografía de la transformada Wavelet se produce en 1909 de la mano de Alfred Haar que en [136] propone la modificación del cambio del análisis en frecuencia por un análisis en escala a partir de una onda compacta que se va ajustando para que se aproxime a la función $f(t)$. Esta onda compacta, también

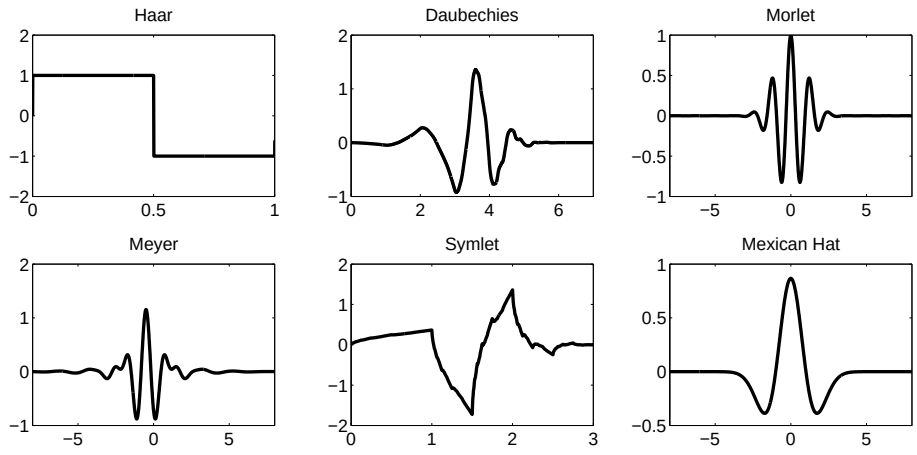

Figura 6.7: Algunas de las funciones wavelet madre más populares que se han utilizado en el procesamiento de señales.

denominada wavelet madre Haar, puede verse en la figura 6.7 (arriba-izquierda). A pesar de lo innovadora que fue esta idea, esta onda no es derivable por lo que sus aplicaciones eran limitadas inicialmente. Habría que esperar a finales de los años 70 para que físicos e ingenieros como Levy, Marr, Grossmann, Morlet, Goupillaud, Weis, Mallat, Daubechies o Coifman entre otros, definieran los conceptos básicos de la teoría wavelet y se popularizara su uso en el procesamiento de señales [137]. Algunas de estas wavelet madre se pueden ver en la figura 6.7.

De hecho, no sería hasta el trabajo desarrollado por Jean Morlet [138] donde se defina formalmente una formulación matemática para la transformada wavelet. Morlet planteó un desarrollo similar a Fourier pero considerando, en lugar de funciones seno y coseno, funciones madre generadas por dilatación o contracción para representar la señal $s(t)$. De este modo, la transformada wavelet continua de la señal $s(b, a, t)$, en función de un factor de escala a y traslación b, se define como [138]:

$$s(b, a, t) = \int_{-\infty}^{\infty} \frac{s(t)}{\sqrt{a}} \psi^* \Big(\frac{t-b}{a} \Big) dt \tag{6.15}$$

donde $a = f/f_0$ es la escala considerando la frecuencia de muestreo f_0 de la señal, b es la traslación en el tiempo y ψ es la wavelet madre. Posteriormente, en 1984 Morlet y Grossmann publicarán en [139] el fundamento teórico de la transformada wavelet continua y de su inversa. Siendo, la wavelet Haar una de las wavelets ortonormales más simples.

En el análisis de señales discretas, estas familias wavelets cumplen la propiedad

de ser ortonormales, ortogonales y con energía normalizada [140], cuya expresión general se define de la siguiente forma:

$$\psi_{ab}(n) = 2^{\frac{-a}{2}}\psi(2^{-a}n - b) \tag{6.16}$$

donde a y b son enteros que escalan y dilatan la wavelet madre ψ. Sobre esta función madre modificada se aplicarán los coeficientes que implementa cada transformada wavelet particular para dividir la señal en su componente en baja frecuencia (aproximación) y su componente en alta frecuencia (detalle). Mallat introduciría también en [140] el concepto de multiresolución que consiste en aplicar la transformada wavelet sobre los resultados que se van obteniendo en cada iteración, con lo que se obtendrá una descomposición en árbol.

Gracias a las características y propiedades que ofrece la transformada wavelet con cada una de sus wavelets madre, se ha podido utilizar sobre diferentes temáticas como astrofísica, geofísica, óptica, mecánica, compresión y codificación de datos o en el procesamiento de señales y análisis multifractal. En esta tesis se emplea la transformada wavelet, en concreto la wavelet Haar, para desarrollar una representación alternativa de la señal de excitación orientada a la estimación.

La transformada wavelet de Haar

La transformada wavelet Haar además de ser la más básica de las transformadas wavelet, tiene la propiedad de que, tras aplicarla sobre una señal, el resultado son dos señales cuyo tamaño es la mitad de la señal original. De este modo, una señal de excitación de N muestras, podrá ser dividida en unidades de menor tamaño (la mitad), con lo que se facilita el proceso de cuantización. Además, como propuso Mallat en [140], esta transformada puede volver a ser aplicada sobre cada una de estas particiones para tener una descomposición en árbol balanceado (todos los nodos tienen el mismo número de componentes) o desbalanceado (no hay mismo número de componentes en los nodos del árbol) según el criterio escogido para la aplicación [137].

Para obtener una descomposición de un vector de excitación $\mathbf{e}$, de tamaño N par, en 2 particiones, que en la bibliografía se denominan como aproximación ($\mathbf{e}_0$) y detalle ($\mathbf{e}_1$), la transformada wavelet Haar aplica las siguientes operaciones [137]:

$$\left.\begin{aligned} e_0(m) &= (e(2m) + e(2m+1))/\sqrt{2} \\ e_1(m) &= (e(2m) - e(2m+1))/\sqrt{2} \end{aligned}\right\} \; m = 0, ..., N/2 - 1. \tag{6.17}$$

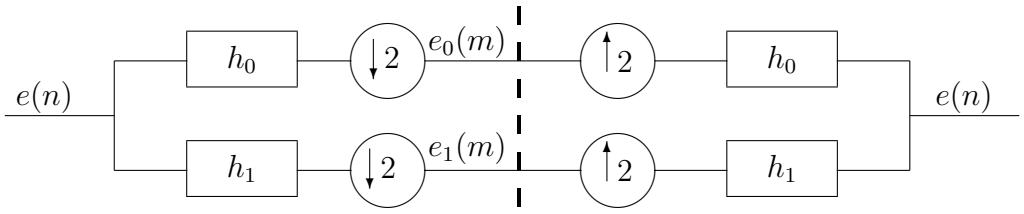

Figura 6.8: Esquema que presenta el proceso de división y reconstrucción empleando la transformada wavelet Haar desde un punto de vista de procesamiento de señales con filtros y operaciones de sobremuestreo y submuestreo.

A partir de estos vectores es posible recuperar el vector original de excitación $\mathbf{e}$ como:

$$\mathbf{e} = \frac{1}{\sqrt{2}}[e_0(1) + e_1(1), e_0(1) - e_1(1), \ldots, e_0(n/2) + e_1(n/2), e_0(n/2) - e_1(n/2)]'. \tag{6.18}$$

donde el operador $[\cdot]'$ representa la operación de trasposición.

Estas mismas operaciones pueden expresarse como operaciones de filtrado, submuestreo y sobremuestreo como se puede ver en la figura 6.8. Así, sobre este esquema se puede observar que la señal de aproximación $e_0(m)$ se puede obtener tras la aplicación de un filtro de análisis (filtro paso bajo) $h_0(n)$, para obtener las componentes en bajas frecuencias, y la señal de detalle $e_1(m)$ obtenida tras la aplicación del correspondiente filtro de análisis (filtro paso alto) $h_1(n)$, para obtener las componentes en altas frecuencias. De esta manera la formulación para la transformada wavelet de Haar de la excitación $\mathbf{e}$ puede representarse como:

$$\begin{aligned} e_0(m) &= [h_0(n) * e(n)]_{\downarrow 2} \\ e_1(m) &= [h_1(n) * e(n)]_{\downarrow 2} \end{aligned} \tag{6.19}$$

donde el operador $[\cdot]_{\downarrow 2}$ representa la operación de submuestreo de un factor 2 y los filtros de análisis $h_0(n)$ y $h_1(n)$ tienen por respuesta al impulso $h_0(n) = (\delta(n) + \delta(n-1))/\sqrt{2}$ y $h_1(n) = (\delta(n) - \delta(n-1))/\sqrt{2}$, respectivamente.

De nuevo, el vector de excitación $\mathbf{e}$ se puede obtener aplicando los correspondientes filtros de síntesis, que en el caso particular de la transformada wavelet Haar, coincidirán con los filtros de análisis $h_0(n)$ y $h_1(n)$ respectivamente, garantizando la reconstrucción perfecta [141]. De este modo, siguiendo con esta formulación, la señal de excitación $e(n)$ se puede obtener a partir de las componentes $e_0(m)$ y $e_1(m)$

como:

$$e(n) = h_0(n) * [e_0(m)]_{\uparrow 2} + h_1(n) * [e_1(m)]_{\uparrow 2} \tag{6.20}$$

donde el operador $[\cdot]_{\uparrow 2}$ representa la operación de sobremuestreo por un factor 2.

Así pues, con la transformada wavelet Haar se puede dividir una señal de excitación $e(n)$ en particiones más pequeñas y que asumiremos independientes a la hora de realizar el proceso de cuantización y posterior recuperación de la señal sintetizada. Así, la señal original $s(n)$ se puede obtener a partir de la descomposición de la señal de excitación en dos componentes $e_0(m)$ y $e_1(m)$ como:

$$s(n) = h(n) * h_0(n) * [e_0(m)]_{\uparrow 2} + h(n) * h_1(n) * [e_1(m)]_{\uparrow 2} \tag{6.21}$$

donde $h(n)$ es la respuesta al impulso del filtro LPC.

En las siguientes subsecciones se presenta cómo se obtienen los diccionarios de cuantización considerando una descomposición de la señal de excitación. Así se explicará tanto para el caso de una descomposición de primer nivel hasta un caso general donde se puede obtener un árbol balanceado y no balanceado con más de un nivel de profundidad cuando ésta transformada es aplicada de manera iterativa.

6.4.2. Obtención del diccionario en el primer nivel de descomposición Haar

La descomposición de primer nivel Haar se obtiene cuando sólo se aplica la transformada wavelet una vez, dando como resultado la descomposición en el árbol balanceado de la figura 6.9 (a).

De esta manera, la expresión (6.21) puede expresarse como $s(n) = s_0(n) + s_1(n)$. Es decir, la señal sintetizada $s(n)$ puede verse como la aportación de ambas particiones $\hat{e}_0(m)$ y $\hat{e}_1(m)$, con los correspondientes filtros de síntesis $h_0(n)$ y $h_1(n)$ respectivamente. Además, bajo esta representación, el error de síntesis se puede definir como:

$$\begin{aligned}\epsilon = \sum_{n=0}^{N-1} \Big(\Big(h(n) * h_0(n) * [\hat{e}_0(m)]_{\uparrow 2} - s_0(n) \Big) + \\ \Big(h(n) * h_1(n) * [\hat{e}_1(m)]_{\uparrow 2} - s_1(n) \Big) \Big)^2 .\end{aligned} \tag{6.22}$$

Alternativamente, se puede definir un error para cada partición ϵ_u como $\epsilon = \epsilon_0 + \epsilon_1$. Sin embargo, si se desarrolla esta expresión se obtiene un término cruzado que no es nulo. Asumiendo que la transformada wavelet Haar proporciona particiones

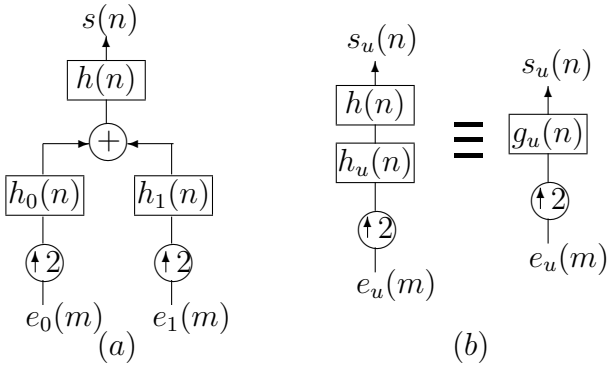

Figura 6.9: Representación en árbol balanceado de la reconstrucción de la señal a partir de (a) una descomposición de primer nivel de la señal de excitación mediante la transformada wavelet Haar y (b) la simplificación aplicable en la reconstrucción de cada rama $u = 0, 1$.

independientes, el término cruzado puede considerarse nulo para cada partición $\mathbf{e}_u$, con $u = 0, 1$. Así, este error ϵ_u se puede definir de modo general como:

$$\epsilon_u = \sum_{n=0}^{N-1} (g_u(n) * [\hat{e}_u(m)]_{\uparrow 2} - s_u(n))^2, \tag{6.23}$$

donde $g_u(n) = h(n) * h_u(n)$ y $s_u(n)$ la correspondiente señal sintetizada obtenida con la partición $e_u(m)$ sin cuantizar. La ventaja de realizar esta separación en errores parciales es que ahora se puede minimizar el error de manera independiente sobre cada uno de ellos como se muestra en la figura 6.9(b).

Para generar diccionarios eficientes siguiendo esta representación, se utilizará el método de cuantización basado en distancia de síntesis, explicado en la sección 5.2.1, para obtener el centro óptimo de cada partición u-ésima. Para ello hay que definir una nueva distancia de síntesis donde la correspondiente partición Haar (u) de la excitación $\mathbf{e}_b(n)$ en una base de datos de entrenamiento, definida como $\mathbf{e}_{u,b}$, se asignará al centroide $\mathbf{c}_u^{(i)}$ si se cumple $\epsilon_u(\mathbf{e}_{u,b}, \mathbf{c}_u^{(i)}, \mathbf{g}_{u,b}) < \epsilon_i(\mathbf{e}_{u,b}, \mathbf{c}_i^{j)}, \mathbf{g}_{u,b})) \; \forall i \neq j$, con la distancia de síntesis $\epsilon_u(\mathbf{e}_u, \mathbf{c}_u, \mathbf{g}_u)$ definida como:

$$\epsilon_u(\mathbf{e}_u, \mathbf{c}_u, \mathbf{g}_u) = \sum_{n=0}^{N-1} (g_u(n) * [e_u(m)]_{\uparrow 2} - g_u(n) * [c_u(m)]_{\uparrow 2})^2. \tag{6.24}$$

donde la respuesta al impulso $h(n)$ se encuentra convolucionada con el filtro Haar correspondiente en $g_u(n)$ y el centroide correspondiente $\mathbf{c}$ tiene un tamaño de acuerdo con el de la partición $e_u(m)$, $m = 0, \ldots, N/2 - 1$ en este caso.

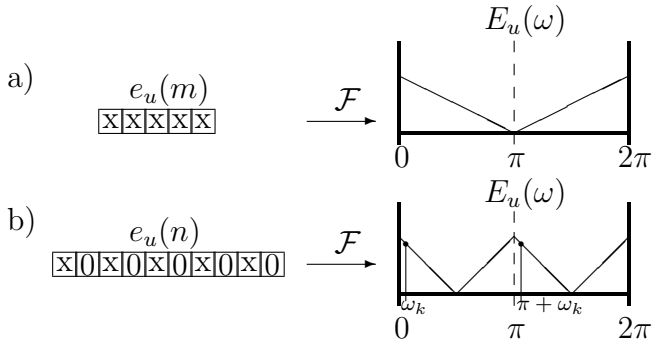

Figura 6.10: Ejemplo gráfico que muestra el proceso de compresión y réplica tras aplicar la transformada de Fourier a una señal $e_u(m)$ (a) y tras realizar un sobremuestreo de factor 2 sobre la señal $e_u(n)$ (b).

A la vista de la expresión (6.24), ésta es análoga a la distancia de síntesis $\epsilon(\mathbf{e}, \mathbf{c}, \mathbf{h})$ presentada en la sección 5.2.1. Así, se puede seguir el mismo desarrollo matemático para calcular el centro óptimo $\mathbf{c}_{new}^{(i)}$. Sin embargo, hay que tener en cuenta que, debido a la operación de sobremuestreo o *upsampling* a la partición $e_u(m)$, el resultado en el dominio de la frecuencia es un efecto de compresión y réplica como se muestra en la figura 6.10. Como consecuencia de esta réplica en el espectro, el número de muestras a calcular se reduce a la mitad, considerando las muestras k y $K/2 + k$ (correspondientes a las frecuencias ω_k y $\omega_k + \pi$) al mismo tiempo. De este modo, la expresión (6.24) en el dominio de la frecuencia para la partición $\mathbf{e}_u$ puede definirse como:

$$\begin{gathered}\varepsilon_u(\mathbf{E}_u, \mathbf{C}_u, \mathbf{G}_u) = \sum_{k=0}^{K/2-1} \left(\mathcal{A}_u(k)E_u(k) - \mathcal{A}_u(k)C(k)\right)^2 \\ \text{con } \mathcal{A}_u(k) = G_u(k) + G_u(K/2 + k)\end{gathered} \tag{6.25}$$

donde $\mathbf{G}_u$, $\mathbf{E}_u$ y $\mathbf{C}_u$ son las transformadas DFT de la extensión con ceros de la señales $g_u(n)$, $e_u(m)$ y $c_u(m)$, donde el número de muestras en frecuencia es $K \geq 2N - 1$ para una señal de N muestras en el tiempo.

Bajo esta definición de error, las $K/2$ componentes del espectro del centro óptimo $\mathbf{C}_{u,new}^{(i)}$ para la partición u se calculan, sobre cada celda $\mathcal{B}_i$, como:

$$C_{u,new}^{(i)}(k) = \frac{\sum_{b \in \mathcal{B}_i} \mathcal{A}_{u,b}^*(k) E_{u,b}(k)}{\sum_{b \in \mathcal{B}_i} \mathcal{A}_{u,b}^*(k) \mathcal{A}_{u,b}(k)} \quad 0 < k < K/2 - 1; \tag{6.26}$$

A partir del espectro de $\mathbf{C}_{u,new}^{(i)}$, se obtiene el correspondiente centro $\mathbf{c}_{u,new}^{(i)}$ mediante la transformada inversa de Fourier. El procedimiento para obtener el diccionario consiste en aplicar de nuevo el proceso de minimización anterior sobre el conjunto $\mathcal{B}_{u,\tau|i}$ formado por las particiones u encontradas en el instante $t + \tau$ cada

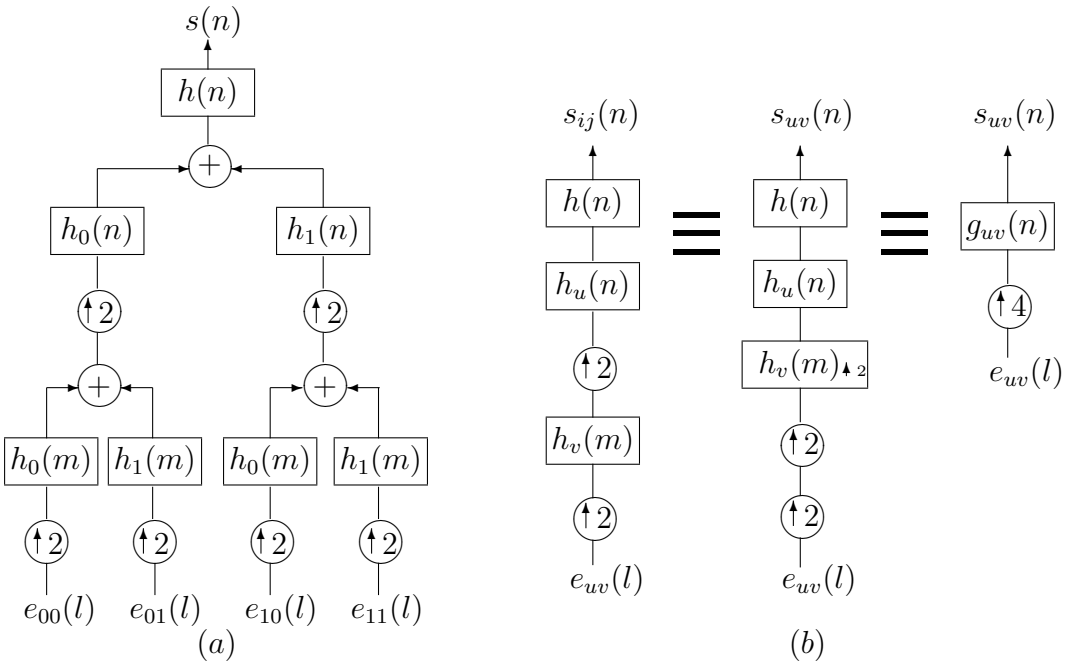

Figura 6.11: Representación en árbol balanceado de una reconstrucción de una señal a partir de una descomposición de segundo nivel de la señal de excitación mediante la transformada wavelet Haar (a) y la simplificación en la reconstrucción de una de sus ramas uv (b).

vez que es encontrado el centro $c^{(i)}_{u,new}(m)$ en el instante t en la base de datos de entrenamiento. De este modo, hay un diccionario para cada partición u.

6.4.3. Obtención del diccionario en una descomposición Haar multinivel

Partiendo de la descomposición de la figura 6.9 es posible obtener una descomposición en árbol balanceado o no balanceado a partir de las particiones $\mathbf{e}_0$ y $\mathbf{e}_1$ aplicando de nuevo la transformada wavelet Haar. Este proceso podría llevarse hasta el extremo donde sólo se obtendría un escalar en la partición obtenida. A continuación, se detalla el modo en el que se pueden generar los diccionarios de acuerdo con una descomposición balanceada en dos niveles como la que se muestra en la figura 6.11 (a). Después, se presentará la expresión general que se podría aplicar tanto a descomposiciones en árbol balanceado o no balanceado.

Como se puede observar, bajo esta descomposición de segundo nivel se obtienen cuatro particiones, $e_{00}(l), e_{01}(l), e_{10}(l)$ y $e_{11}(l)$ con $l = 0, \ldots, N/4 - 1$, donde la longitud de cada una de ellas es $N/4$. Siguiendo el mismo procedimiento indicado

anteriormente, en cada rama de este árbol se obtiene una reconstrucción parcial de la señal sintetizada de manera que $s(n)$ se pueda expresar como $s(n) = s_{00}(n) + s_{01}(n) + s_{10}(n) + s_{11}(n)$ tal que cada una de estas componentes pueden expresarse como:

$$s_{uv}(n) = h(n) * h_u(n) * \Big[h_v(n) * [e_{uv}(l)]_{\uparrow 2}\Big]_{\uparrow 2} = g_{uv}(n) * [e_{uv}(l)]_{\uparrow 4}\,, \qquad u, v \in [0, 1] \tag{6.27}$$

donde $g_{uv}(n)$ se define como $g_{uv}(n) = h(n) * h_u(n) * [h_v(m)]_{\uparrow 2}$, tal y como se aprecia en la figura 6.11 (b).

Aunque en la figura 6.11 se ha representado un árbol balanceado de profundidad 2, este mismo procedimiento puede aplicarse sobre un árbol no balanceado. La diferencia es que, mientras que en un árbol balanceado todos los nodos van a tener la misma cantidad de componentes, en un árbol no balanceado cada nodo tiene un número diferente de componentes de acuerdo a la profundidad. Así por ejemplo, el correspondiente árbol no balanceado resultante estaría formado por las particiones $e_{00}(l)$ y $e_{01}(l)$, que tendrían $N/4$ componentes, y la partición $e_1(m)$, que tendría $N/2$ componentes.

Para el ejemplo presentado de árbol balanceado, si se asume independencia entre cada partición uv, se puede definir el error de síntesis por cada rama, ϵ_{uv}, como:

$$\epsilon_{uv} = \sum_{n=0}^{N-1} \Big(g_{uv}(n) * [\hat{e}_{uv}(l)]_{\uparrow 4} - s_{uv}(n)\Big)^2, \tag{6.28}$$

donde $\hat{e}_{uv}(l)$ es la correspondiente partición cuantizada en el árbol de la señal de excitación.

Con la definición de error indicada, se puede definir la distancia de síntesis por cada partición uv como:

$$\epsilon_{uv}(\mathbf{e}_{uv}, \mathbf{c}_{uv}, \mathbf{g}_{uv}) = \sum_{n=0}^{N-1} (g_{uv}(n) * [e_{uv}(l)]_{\uparrow 4} - g_{uv}(n) * [c_{uv}(l)]_{\uparrow 4})^2 \tag{6.29}$$

donde $e_{uv}(l)$y $c_{uv}(l)$ son el correspondiente vector de excitación y el centro de cuantización para la rama uv-ésima. Como se aprecia, la nueva distancia es muy similar a la expresada en (6.23) con la diferencia de un nuevo factor de sobremuestreo por tener una profundidad 2.

Al igual que se consideró en la subsección anterior, al tener ahora dos procesos

de sobremuestreo, la transformada de Fourier genera 4 réplicas del espectro. Por lo tanto, de manera similar a como se hizo en la expresión (6.25), el error a minimizar queda definido como:

$$\varepsilon_{uv}(\mathbf{E}_{uv}, \mathbf{C}_{uv}, \mathbf{C}_{uv}) = \sum_{k=0}^{K/4-1} (\mathcal{A}_{uv}(K)E_{uv}(K) - \mathcal{A}_{uv}(k)C(K))^2$$
$$\text{con } \mathcal{A}_{uv}(k) = (G_{uv}(k) + G_{uv}(K/4+k) + G_{uv}(K/2+k) + G_{uv}(3K/4+k) \tag{6.30}$$

donde $\mathbf{G}_{uv}$, $\mathbf{E}_{uv}$ y $\mathbf{C}_{uv}$ son la transformada de Fourier de las señales extendidas con ceros de $g_{uv}(n)$, $e_{uv}(m)$ y $c_{uv}(m)$ respectivamente. Con esta nueva expresión, las $K/4$ componentes replicadas del centro óptimo $\mathbf{C}^{(i)}_{uv,new}$ se puede obtener de nuevo aplicando mínimos cuadrados, sobre cada celda $\mathcal{B}_i$, como:

$$C^{(i)}_{uv,new}(k) = \frac{\sum_{b\in\mathcal{B}_i} \mathcal{A}^*_{uv,b}(k)E_{uv,b}(k)}{\sum_{b\in\mathcal{B}_i} \mathcal{A}^*_{uv,b}(k)\mathcal{A}_{uv,b}(k)} \qquad 0 < k < K/4 - 1; \tag{6.31}$$

De manera general, estas expresiones se pueden extender sin problemas para árboles de mayor profundidad y bajo cualquier configuración (balanceado o no balanceado) sólo con tener en cuenta los procesos de sobremuestreo, y las convoluciones con los filtros $h_0(n)$ y $h_1(n)$ correspondientes para generar el filtro equivalente $g(n)$ de cada rama.

Finalmente, a partir del espectro de $\mathbf{C}^{(i)}_{uv,new}$, se obtiene el correspondiente centro $\mathbf{c}^{(i)}_{uv,new}$ mediante la transformada inversa de Fourier. El procedimiento para obtener el diccionario consiste en aplicar de nuevo el proceso de minimización anterior sobre el conjunto $\mathcal{B}_{uv,\tau|i}$ formado por las particiones uv encontradas en el instante $t+\tau$ cada vez que es encontrado el centro $c^{(i)}_{uv,new}(m)$ en el instante t en la base de datos de entrenamiento. De este modo, hay un diccionario para cada partición uv.

6.4.4. Obtención del diccionario en el último nivel de descomposición Haar. Un enfoque matricial

Siguiendo el procedimiento de la sección anterior es posible alcanzar el último nivel del árbol balanceado y obtener así los K escalares que definen el comportamiento de la señal de excitación. No obstante, hay un procedimiento más rápido para poder alcanzar estos mismos K escalares mediante una formulación matricial de la transformada wavelet Haar sobre el vector de excitación $\mathbf{e}$.

La principal ventaja que aporta este procedimiento es que se puede recuperar el

vector de excitación $\mathbf{e}$ a partir de los escalares obtenidos en el último nivel de descomposición, coeficientes que conforman el vector $\mathbf{w}$, mediante operaciones matriciales, reduciendo así, la carga computacional que supone el procedimiento anterior con K escalares (por las operaciones de filtrado y sobremuestreo necesarias). De este modo, si el vector de excitación se define como $\mathbf{e} = W\mathbf{w}$, donde W es la matriz de transformación wavelet Haar [137], la señal de voz $s(n)$ se puede expresar de la forma:

$$s(n) = h(n) * \left(\sum_{k=0}^{K-1} W_k(n) w_k \right) = \sum_{k=0}^{K-1} \mathcal{H}_k(n) w_k \tag{6.32}$$

donde $W_k(n)$ es el k-ésimo vector columna de la matriz de transformación wavelet y $\mathcal{H}_k(n)$ representa el k-ésimo vector columna de la matriz W después de su convolución con la respuesta al impulso $h(n)$, y $\mathbf{w} = [w_0, w_1, \ldots, w_{K-1}]'$ son los escalares finales resultantes de la descomposición Haar.

Bajo esta nueva representación matricial, la distancia de síntesis se obtiene de la forma:

$$\epsilon(\mathbf{w}, \mathbf{c}, \mathcal{H}) = (\mathcal{H}\mathbf{c} - \mathcal{H}\mathbf{w})'(\mathcal{H}\mathbf{c} - \mathcal{H}\mathbf{w}). \tag{6.33}$$

Tras la clasificación de los vectores $\mathbf{w}$ en C centros, el centro óptimo ($\mathbf{c}_{new}^{(i)}$) para cada celda $\mathcal{B}_i, \forall i \in [1, \mathcal{C}]$, se puede obtener realizando una minimización por mínimos cuadrados sobre la expresión (6.33) como:

$$\mathbf{c}_{new}^{(i)} = (\mathcal{H}'_{\mathcal{B}_i} \mathcal{H}_{\mathcal{B}_i})^{-1} \mathcal{H}'_{\mathcal{B}_i} \mathbf{s}_{\mathcal{B}_i} \tag{6.34}$$

donde $\mathcal{H}_{\mathcal{B}_i}$ y $\mathbf{s}_{\mathcal{B}_i}$ se obtienen apilando sucesivamente las matrices de transformación $\mathcal{H}$ y las señales de voz $s(n)$ de los elementos clasificadas en el conjunto $\mathcal{B}_i$. Es decir, éstas se definen como $\mathcal{H}_{\mathcal{B}_i} = [\mathcal{H}_{b_1}, \mathcal{H}_{b_2}, \ldots]'$ y $\mathbf{s}_{\mathcal{B}_i} = [\mathbf{s}_{b_1}, \mathbf{s}_{b_2}, \ldots]'$, con $b_1, b_2, \cdots \in \mathcal{B}_i$.

De este modo, mientras que con el procedimiento general serían necesarios K diccionarios para realizar la optimización, siguiendo este procedimiento sólo hace falta calcular un diccionario empleando la expresión (6.34) para cada índice i y donde cada centroide es un vector $\mathbf{w}$ con K escalares.

6.4.5. Resultados experimentales

El esquema de mitigación que emplea la nueva representación de la señal de excitación, mediante la transformada wavelet Haar, se muestra en la figura 6.12. Este esquema es similar al de la figura 6.1 pero en lugar de tener un único índice de

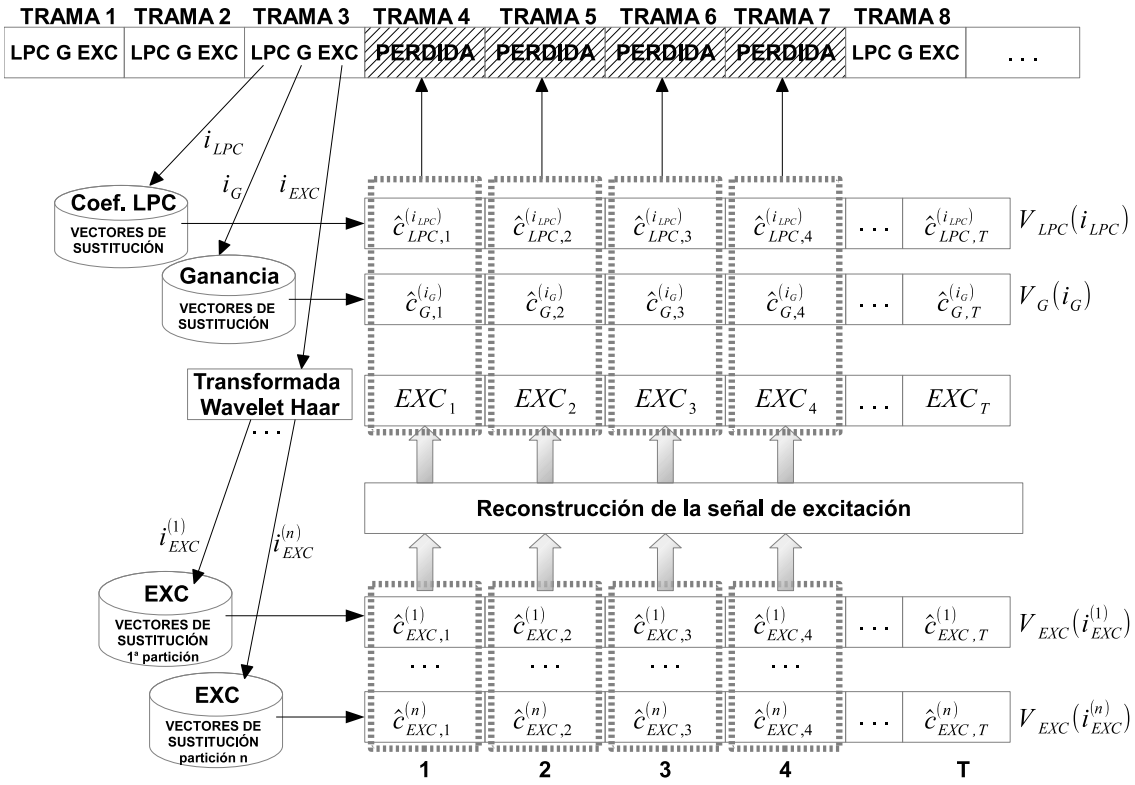

Figura 6.12: Esquema de mitigación de errores que aplica la nueva representación de la señal de excitación basada en la transformada de wavelet Haar para mejorar las estimaciones en los vectores de sustitución.

cuantización para la señal de excitación de energía unitaria, se tienen tantos como particiones obtenidas en la descomposición en árbol balanceado o no balanceado de la última trama recibida. Los vectores de sustitución obtenidos para cada uno proveerán de las estimaciones empleadas para generar la señal de excitación y posterior síntesis de voz tal y como se ha indicado anteriormente.

Para comprobar el rendimiento de este esquema de mitigación, las pruebas se han realizado sobre el conjunto de test de la base de datos TIMIT [104] y las condiciones de canal se han simulado siguiendo un modelo Gilbert (comentado en la sección 3.5.3). Con estas pruebas se analizará el rendimiento de diferentes propuestas, en términos de calidad objetiva PESQ [103], con diferentes tasas de pérdidas de paquetes o *Packet Error Rate* (PER), PER= $\{10\,\%, 20\,\%, 30\,\%, 40\,\%, 50\,\%\}$, y con diferentes longitudes promedio de ráfaga o *Average Burst Length* (ABL), ABL= $\{1, 2, 4, 8, 12\}$, sobre los codecs AMR [39] e iLBC [45].

En las tablas 6.5 y 6.6 se muestran los resultados PESQ promedio sobre la longitud de ráfaga (izquierda) y sobre la tasa de pérdidas (derecha) obtenidos sobre

los codecs AMR (modo 12.2 kbps) e iLBC (modo 15.2 kbps) respectivamente. En esta tabla, se incluyen los resultados de la técnica de vectores de sustitución (RV), presentada en la sección 6.2, las diferentes propuestas basadas en la representación wavelet según se obtenga una descomposición en árbol balanceado (BT$\mathcal{W}$) o no balanceado (UT$\mathcal{W}$), con $\mathcal{W}$ particiones, la aplicación de la descomposición completa que aplica un enfoque matricial (BTM) y la aplicación de la técnica que emplea un procesamiento por subtramas (SRV$\mathcal{W}$). Ésta última, se ha querido incorporar para analizar el rendimiento obtenido por la nueva representación de la señal de excitación, frente al caso de realizar una estimación sobre particiones realizadas sobre la propia señal de excitación y tratadas de manera independiente.

Por último, para generar de los vectores de sustitución, se han obtenido los diccionarios de cuantización, con 1024 centroides, siguiendo el método de cuantización vectorial LBG para los coeficientes LPC, el método de cuantización *Lloyd-Max* para la ganancia y el método de cuantización basado en distancia de síntesis y con división fija (FSLBG), comentado en la sección 5.2.1, sobre la señal de excitación en la técnica RV y las particiones obtenidas. Así, se puede comprobar el rendimiento que supone la nueva representación sobre el funcionamiento del esquema de mitigación propuesto en la sección 6.2.

A la vista de los resultados, se puede apreciar que en el codec AMR son ligeramente inferiores a los obtenidos por el codec iLBC, aunque en transmisiones sin pérdidas, la calidad perceptual del códec AMR es superior a la del codec iLBC (4.02 vs 3.05) en métrica PESQ. Esto se debe a que el codec AMR es más vulnerable frente a la pérdida de paquetes como consecuencia de la propagación del error, dado que el codec AMR está basado en un paradigma CELP. Del mismo modo, la nueva representación wavelet, tanto balanceado como no balanceado, supone una mejora significativa frente al esquema de los vectores de sustitución RV, presentados en la sección 6.2, en todas las condiciones de canal.

Por otro lado, se puede comprobar que el rendimiento ofrecido por la nueva propuesta, con una descomposición balanceada o no balanceada obtenida mediante la transformada wavelet Haar, es superior a la propuesta que realiza un procesamiento por subtramas (SRV) a igualdad en número de particiones ($\mathcal{W}$). Es decir, aunque en ambos casos se parte de particiones de menor tamaño que facilitan el proceso de cuantización, hay que tener en cuenta que la estimación que conforma la señal de excitación final y la posterior síntesis $\hat{s}(n)$, tiene que ser una señal continua y que minimice el error de síntesis con la señal original $s(n)$. Sin embargo, en una señal

		Tasa de pérdida de paquetes					Longitud promedio de ráfaga				
	$\mathcal{W}$	10 %	20 %	30 %	40 %	50 %	1	2	4	8	12
AMR	-	2,85	2,26	1,82	1,45	1,15	1,86	1,94	1,89	1,90	1,94
RV	-	3,03	2,55	2,13	1,85	1,56	2,10	2,20	2,23	2,28	2,31
SRV	4	2,97	2,43	2,06	1,77	1,49	2,04	2,10	2,12	2,15	2,18
	8	3,01	2,46	2,09	1,78	1,51	2,07	2,12	2,15	2,18	2,19
UT	4	3,07	2,58	2,21	1,93	1,66	2,17	2,25	2,29	2,35	2,39
	8	3,09	2,60	2,25	1,98	1,71	2,20	2,28	2,32	2,38	2,42
BT	2	3,04	2,56	2,15	1,87	1,58	2,12	2,21	2,24	2,30	2,33
	4	3,08	2,59	2,23	1,96	1,69	2,19	2,27	2,30	2,35	2,41
	8	3,10	2,64	2,27	2,02	1,74	2,23	2,31	2,34	2,40	2,45
	16	3,09	2,62	2,25	2,01	1,73	2,22	2,30	2,32	2,39	2,44
	32	3,08	2,60	2,22	2,00	1,72	2,20	2,28	2,31	2,38	2,42
	128	3,06	2,57	2,19	1,98	1,70	2,19	2,27	2,29	2,36	2,40
BTM	256	3,09	2,63	2,26	2,01	1,73	2,22	2,30	2,32	2,39	2,43

Tabla 6.5: Resultados PESQ promedio sobre longitud promedio de ráfaga (izquierda) y sobre tasa de error en paquetes (derecha) para el codec AMR y la aplicación de varias propuestas: la técnica de vectores de sustitución (RV), el procesamiento por subtramas (SRV), la representación wavelet en árbol balanceado (BT) y no balanceado (UT) con el correspondiente número de particiones $\mathcal{W}$ y la versión matricial (BTM).

de excitación, las muestras de voz están correladas y, al tratar cada partición de manera independiente, se va a introducir un efecto de corte o discontinuidad de una partición con otra.

Este hecho puede observarse en la figura 6.13, donde se muestra la señal de síntesis obtenida (arriba) a partir de la señal de excitación (abajo) original (a), la señal de excitación obtenida mediante el particionado en 4 subtramas (SVR4) (b) y la señal de excitación obtenida con la representación wavelet Haar en una descomposición balanceada con 4 particiones (BT4) (c). Dado que el procedimiento seguido por la transformada wavelet Haar genera la señal de síntesis aplicando filtrado y sobremuestreo sobre cada rama, cada una va a tener una parte de la señal sintetizada completa y se va a preservar tanto la continuidad como sus propiedades (como el *pitch*) mejor que siguiendo un procesamiento por subtramas. Así se muestra en las señales sintetizadas (arriba), donde la propuesta basada en wavelet permite conservar mejor el *pitch*, mientras que la propuesta basada en subtramas puede perder la estructura del *pitch* e incluso generar artefactos.

		Tasa de pérdida de paquetes					Longitud promedio de ráfaga				
	$\mathcal{W}$	10 %	20 %	30 %	40 %	50 %	1	2	4	8	12
iLBC	-	3,02	2,56	2,22	1,96	1,75	2,62	2,34	2,20	2,18	2,18
RV	-	3,06	2,65	2,38	2,16	1,97	2,53	2,33	2,39	2,44	2,48
SRV	4	3,03	2,60	2,33	2,09	1,93	2,58	2,27	2,33	2,38	2,41
	8	3,05	2,61	2,34	2,10	1,94	2,60	2,29	2,35	2,39	2,42
UT	4	3,11	2,75	2,48	2,26	2,09	2,66	2,38	2,48	2,54	2,61
	8	3,13	2,77	2,56	2,34	2,16	2,72	2,44	2,54	2,59	2,66
BT	2	3,09	2,69	2,43	2,18	2,04	2,63	2,35	2,43	2,48	2,55
	4	3,12	2,76	2,55	2,33	2,15	2,71	2,44	2,53	2,58	2,65
	8	3,15	2,81	2,59	2,37	2,20	2,76	2,48	2,57	2,62	2,69
	16	3,14	2,79	2,57	2,36	2,19	2,74	2,46	2,55	2,61	2,68
	32	3,12	2,77	2,55	2,35	2,18	2,73	2,45	2,54	2,59	2,66
	128	3,10	2,75	2,54	2,34	2,17	2,70	2,43	2,52	2,58	2,65
BTM	256	3,14	2,79	2,58	2,37	2,19	2,74	2,45	2,56	2,61	2,68

Tabla 6.6: Resultados PESQ promedio sobre longitud promedio de ráfaga (izquierda) y sobre tasa de pérdida en paquetes (derecha) para el codec iLBC y la aplicación de varias propuestas: la técnica de vectores de sustitución (RV), el procesamiento por subtramas (SRV), la representación wavelet en árbol balanceado (BT) y no balanceado (UT) con el correspondiente número de particiones $\mathcal{W}$ y la versión matricial (BTM)

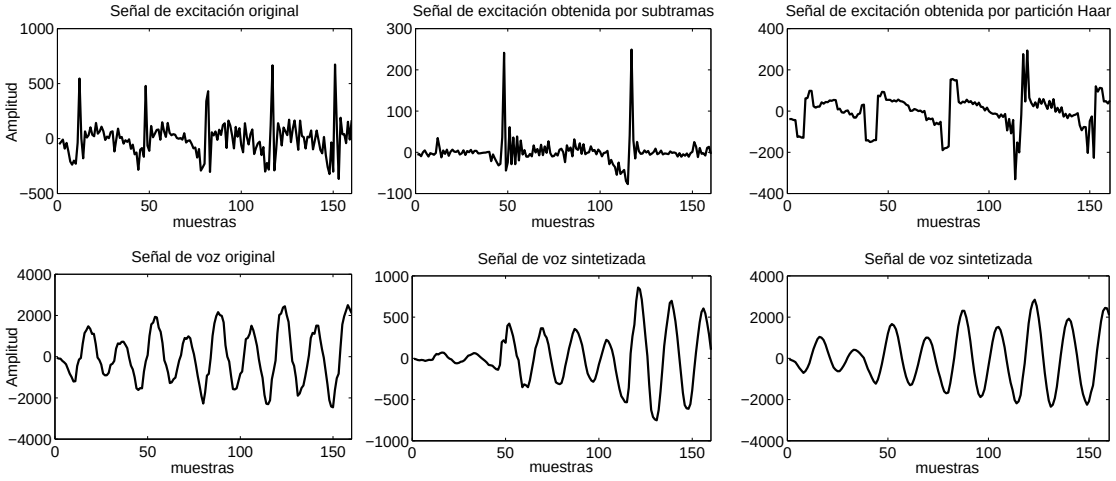

Figura 6.13: Comparación en la síntesis de voz (arriba) para diferentes señales de excitación (abajo): señal de excitación original (a), señal de excitación obtenida por procesamiento en 4 subtramas (b) y señal de excitación obtenida mediante descomposición wavelet Haar en 4 componentes (c).

Del mismo modo, de entre las propuestas basadas en la transformada wavelet Haar (BT$\mathcal{W}$ y UT$\mathcal{W}$), se puede observar que la propuesta con árboles balanceados (BT) ofrece un rendimiento superior al de los no balanceados (UT). El motivo de esto puede radicar en el tamaño que tienen las distintas particiones en cada descomposición. Mientras que un árbol balanceado todas las particiones tienen el mismo tamaño, en un árbol no balanceado cada partición tendrá un número diferente de muestras. Por ejemplo, para un árbol balanceado con 8 particiones de una señal de 512 componentes, cada partición tendrá 128 muestras y alcanzará una profundidad 2. Sin embargo, en un árbol no balanceado estas mismas 8 particiones se alcanzan en el nivel de profundidad 7 (si se expande sólo la componente de aproximación) generando particiones desde 4 a 256 muestras. Como consecuencia de ello, los diccionarios para particiones de mayor número de muestras se van a entrenar peor y esto afecta a la calidad perceptual final de la señal de voz recuperada.

No obstante, siguiendo una descomposición en árbol balanceado hasta el final, se puede pensar que a medida que se consideren particiones más pequeñas, mejores resultados se obtendrán al reducir el error de cuantización y minimizar el error de síntesis. Sin embargo, como puede verse de manera gráfica en la figura 6.14, el resultado PESQ promedio decrece ligeramente sobre ambos codecs a partir de una descomposición balanceada con 8 particiones (BT8). El motivo de que esto ocurra, se debe a que a medida que aumenta el número de particiones, también se están perdiendo las relaciones intra-trama, haciendo que las particiones estén más aisladas entre sí, como ocurre con el procesamiento por subtramas. Este es el motivo por el que el procedimiento matricial BTM consigue un rendimiento similar a BT16, ya que en este procedimiento sí se tiene en cuenta la influencia de unos coeficientes sobre otros durante el proceso de síntesis matricial.

Para finalizar esta sección, comentar que la mejora significativa observada con el test objetivo PESQ [103] se puede corroborar con el test subjetivo MUSHRA [99] en la figura 6.15. En este test se comparó la calidad de la propuesta BT8 con respecto a los codecs analizados (AMR e iLBC) siguiendo el procedimiento indicado en la sección 3.5.1. De este modo se han realizado las pruebas con 20 oyentes y se han empleado 10 frases de voz de la base de datos Albayzin [106]. En la prueba se escogieron las señales de voz obtenidas en las siguientes condiciones de canal: tasas de pérdidas de 10 % y 30 % y longitud de ráfagas promedio de 2 y 12 tramas. A la vista de los resultados, se puede comprobar que la propuesta BT8 ofrece un

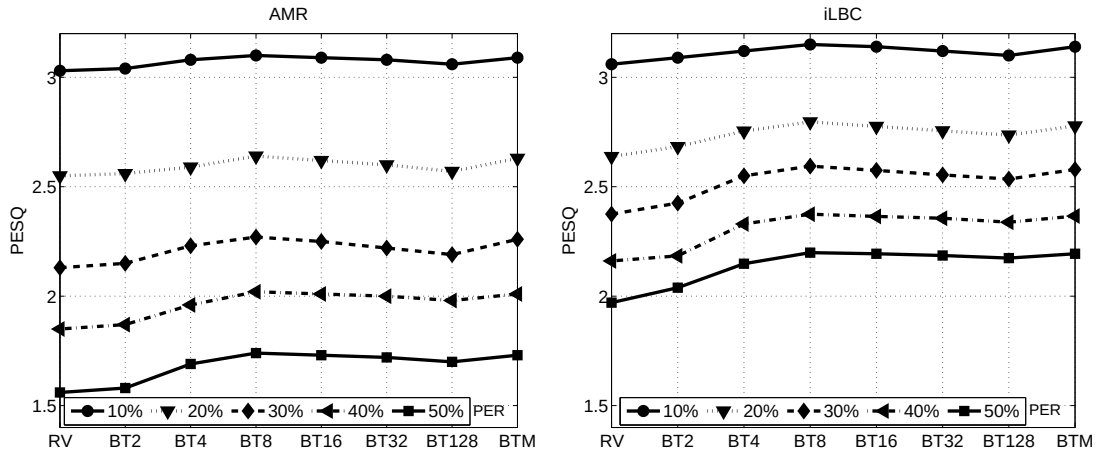

Figura 6.14: Rendimiento PESQ promedio de diferentes descomposiciones en árbol balanceado BT(de 2 a 128 particiones) y comparación con el caso límite con un enfoque matricial (BTM) sobre los codecs AMR e iLBC para diferentes valores de tasa de pérdidas (PER).

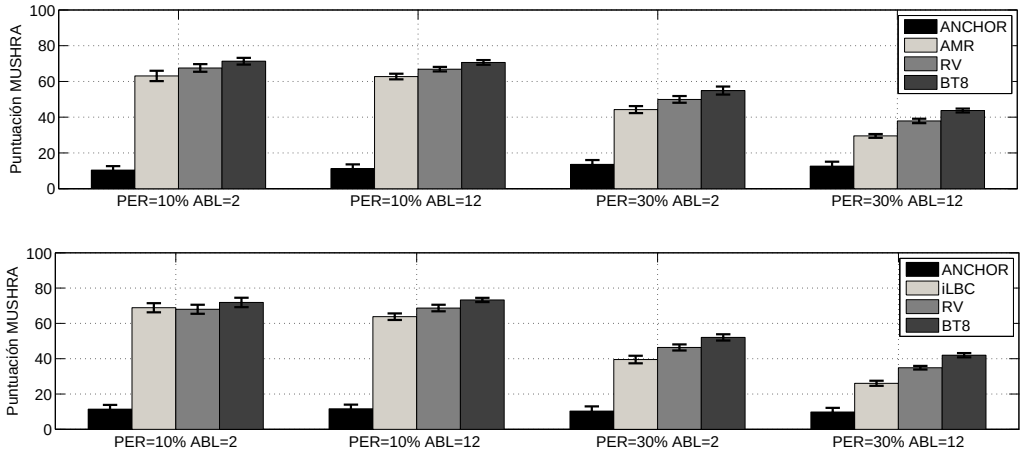

Figura 6.15: Resultados del test MUSHRA obtenidos comparando la propuesta basada en wavelet con 8 particiones (BT8) respecto a la técnica de vectores de sustitución (RV) y el codec estándar (AMR e iLBC) con su propio algoritmo PLC en diferentes tasas de pérdidas de paquetes y longitud promedio de ráfaga.

rendimiento superior, confirmando los resultados PESQ obtenidos en las tablas 6.5 y 6.6.

Capítulo 7

Conclusiones

Este capítulo resume las principales conclusiones alcanzadas durante el desarrollo de esta tesis. Además se describen las principales contribuciones de este trabajo y algunas sugerencias de trabajo futuro.

7.1. Conclusiones

Las conclusiones de esta tesis se pueden desglosar en:

- En esta tesis se ha estudiado la degradación provocada por el efecto multitrayecto y la pérdida de paquetes, ya que son característicos en las transmisiones de voz en redes WLAN y redes IP respectivamente.
- Para mitigar el efecto multitrayecto, se ha aplicado la técnica *soft-decision decoding* que permite estimar las componentes del paquete dañado mediante una estimación MMSE. En esta estimación se tiene en cuenta el *bit error rate* que afecta al todo el paquete y la componente recibida en el paquete recibido. De esta manera, se ha podido mejorar el rendimiento de los codecs G.726 y G.722 que se utilizan en el estándar DECT.
- Para la recuperación de los paquetes perdidos, en esta tesis se han presentado nuevos métodos de cuantización, basados en el algoritmo LBG, que ha permitido generar diccionarios adecuados para la cuantización y estimación de la señal de excitación de una manera eficaz.
- Se ha planteado una técnica de prevención de pérdida basada en el emisor que emplea códigos FEC compactos. Considerando los nuevos métodos de

cuantización de la señal de excitación, se ha podido no sólo minimizar el error de síntesis en la última trama perdida en la ráfaga, sino también, reducir la propagación del error que se origina en los codec con dependencia inter-trama, como los codecs basados en el paradigma CELP.

- Para que el código FEC no suponga un incremento en la tasa de bits por paquete ni la incompatibilidad con el codec original en transmisiones sin pérdidas, se ha aplicado una técnica esteganográfica que lo oculta en el propio paquete y no supone una pérdida de calidad perceptual significativa en transmisiones sin pérdidas.

- También se han presentado diferentes esquemas de mitigación, basados en la técnica de vectores de sustitución, que ha permitido mejorar el rendimiento del codec al ofrecer estimaciones tanto de los coeficientes LPC como de la señal de excitación durante la ráfaga de pérdidas. Se ha podido comprobar que ofrece un alto rendimiento especialmente en ráfagas largas donde los algoritmos PLC no ofrecen buenas estimaciones.

- Del mismo modo, la propuesta de un enfoque mixto entre técnicas de prevención y mitigación de pérdidas ha permitido aprovechar las ventajas de ambos enfoques. Esto es, recuperar la pérdida en una ráfaga y por otro lado, evitar la aparición de la propagación del error en los codecs con dependencia inter-trama.

- Las diferentes propuestas de esquemas de mitigación han permitido mejorar la estimación de la señal de excitación y mejorar así la calidad de la señal sintetizada.

- La aplicación de un filtro adaptativo RLS como complemento a la técnica de vectores de sustitución ha permitido mejorar la estimación de la señal de excitación en las primeras pérdidas de una ráfaga. De esta manera se evita el posible error producido con las primeras estimaciones del correspondiente vector de sustitución que están afectadas por el error de cuantización.

- Finalmente, la nueva representación basada en transformada wavelet Haar ha mejorado no sólo la posibilidad de cuantizar la señal de excitación a partir de una descomposición en particiones de menor tamaño, sino también, minimizar el error de síntesis en la señal recuperada y sin que este proceso genere discontinuidades.

7.2. Contribuciones

Las principales aportaciones de esta tesis se pueden resumir en:

- Se ha aplicado la técnica *soft-decision decoding* para reducir el impacto de los paquetes degradados a consecuencia del efecto multitrayecto sobre comunicaciones basadas en el estándar ETSI DECT [142].
- Se han desarrollado diferentes métodos de cuantización vectorial para reducir el error de cuantización de la señal de excitación [143].
- Se ha utilizado una técnica esteganográfica para mantener la compatibilidad del codec AMR en canales sin pérdidas [144].
- Se ha propuesto una técnica de mitigación de pérdidas novedosa basada en vectores de sustitución [145].
- Se ha desarrollado una representación de la señal de excitación orientada a mitigación que está basada en la transformada wavelet Haar [146].
- Se ha mejorado la estimación de la señal de excitación en las primeras pérdidas mediante el uso de un filtro adaptativo RLS [147].
- Se ha propuesto un esquema mixto entre las técnicas de prevención de errores basados en el emisor y las técnicas de mitigación de errores basados en el receptor con el fin de aprovechar las ventajas de ambos enfoques sobre el codec AMR[148].

7.3. Trabajo futuro

La presente tesis se ha centrado en el desarrollo de técnicas que minimizan el impacto de la degradación producida en las transmisiones sobre redes WLAN e IP. A pesar de que el resultado es un codec más robusto frente a errores en el canal, la calidad perceptual de la señal recuperada se podría mejorar aplicando otras técnicas que ayuden a hacerla más natural e inteligible. Un ejemplo sería la aplicación de técnicas de realce empleadas para el reconocimiento remoto.

Para las transmisiones WLAN se ha utilizado la técnica *soft-decision decoding* que realiza una estimación de las tramas dañadas en una transmisión considerando un *bit error rate* promedio para toda la trama recibida. Una alternativa de mejora

de este enfoque sería considerar el *bit error rate* promedio a nivel de componente de la trama, con lo que las estimaciones serían mucho más precisas.

En esta tesis se ha presentado una nueva representación para la señal de excitación, basada en la transformada wavelet Haar, que hace tratable el problema de la dimensionalidad de la señal de excitación para realizar una cuantización eficiente y orientada a la mitigación de pérdidas. No obstante, se podrían explorar otras representaciones basadas en otras transformadas, como la transformada coseno o la transformada de *Karhounen-Loeve*, con las que desarrollar nuevos métodos de cuantización que eviten el alto coste computacional que supone obtener los diccionarios de cuantización.

La propuesta de utilizar un código FEC compacto permite restaurar la última trama perdida con el menor error de síntesis. Esto permitirá reducir también el efecto del error de propagación pero no lo evita. De este modo, una posible mejora de este esquema sería que el código FEC proporcionara la señal de excitación que no sólo minimiza el error de síntesis con la señal original sino también aquella que evita el error de propagación.

Por último, con vista a emplear la nueva representación de la señal de excitación como código FEC, durante el proceso de cuantización de las particiones obtenidas, todas ellas se han cuantizado con 10 bits. Esta codificación podría suponer una alta tasa de bits por lo que sería conveniente realizar una cuantización diferente por partición, de acuerdo a su entropía, o profundizando en la reconstrucción de la señal mediante el enfoque matricial.

Capítulo 8

Conclusions

This chapter summarizes the main conclusions reached during the development of this thesis. It also describes the main contributions of this paper and some suggestions for future work.

8.1. Conclusions

The conclusions of this thesis can be broken down into:

- In this thesis, the degradations caused by the multipath effect and packet losses, which are characteristics of the speech transmissions over WLAN and IP networks respectively, have been addressed.

- In order to mitigate the multipath effect, the soft-decision decoding technique has been applied to estimate the components of the damaged packet by using an MMSE estimation process. This estimation takes into account the bit error rate that affects the whole packet and the received component in the received packet. In this way, it is possible to improve the performance of the G.726 and G.722 codecs which are mandatory in the DECT standard.

- A sender-driven technique which apply a compact FEC code has been proposed. Considering the new quantization methods of the excitation signal, it is possible not only to minimize the synthesis error in the last lost frame in the burst, but also to reduce the error propagation which apper in codecs with inter-frame dependence, as codecs based on the CELP paradigm.

- Since the use of FEC codes implies an increase in the bit rate per packet and the incompatibility with the original codec in lossless transmissions. A steganographic technique has been applied that embeds it into the packet and it does not imply a significant loss of perceptual speech quality in lossless transmissions.

- In addition, the proposal for a mixed approach between prevention and mitigation techniques has allowed us to take the advantages of both approaches. That is, recover the loss in a burst and avoid the error propagation in codecs with inter-frame dependency.

- Different error mitigation schemes, based on the replacement vector technique, have been presented. These replacement supervectors provide estimates of both speech parameters (the LPC coefficients and the excitation signal) during the burst of packet losses. The different proposals have shown a noticeable improvement in long bursts where the PLC algorithms can not provide estimates.

- The different proposals of error mitigation schemes have improved the estimation of the excitation signal and as a consecuence, the speech quality of the synthesized signal.

- The application of an adaptive RLS filter as a complement to the replacement vector technique has allowed the improvement of the estimation of the excitation signal in the first lost packet in a burst. Thus, the possible error produced by the first estimates of the corresponding replacement vector, which could be affected by error quantization, are avoided.

- Finally, the new representation based on Haar wavelet transform has improved not only the possibility of quantizing the excitation signal from a decomposition in smaller partitions, but also, the minimization of the synthesis error in the recovering process of the speech signal and without generating discontinuities.

8.2. Contributions

The main contributions of this thesis can be summarized as follows:

- Apply the soft-decision decoding technique to reduce the impact of modified packets as a consequence of the multipath effect on communications based on the ETSI DECT standard [142].
- Develop different vector quantization methods which minimize the quantization error of the excitation signal [143].
- Apply a steganographic technique to keep compatibility of the AMR codec in lossless channels [144].
- Develop a loss mitigation technique based on replacement vectors [145].
- Develop a new representation of the excitation signal based on the Haar wavelet transform [146].
- Improve the estimation of the excitation signal in the first losses by using the RLS adaptive filter [147].
- Propose a mixed scheme between the sender-driven and receiver-based techniques in order to take advantage of both approaches on the AMR speech codec [148].

8.3. Future work

The present thesis is focused on the development of techniques that minimize the impact of the degradation produced during the transmissions over WLAN and IP networks. Although the proposals achieve a more robust codec against errors in the channel, the perceptual quality of the recovered signal could be improved by applying other techniques that could make it more natural and intelligible. An example could be the application of enhancement techniques used for speech recognition.

For WLAN transmissions, the *soft-decision decoding* technique has been used in order to estimates the damaged componentes in a frame by considering an average *bit error rate* over the entire frame. An alternative to improve this approach would be the consideration of the average *bit error rate* at the component level, so the estimates would be much better.

In this thesis a new representation has been introduced for the excitation signal, based on the Haar wavelet transform, which makes the excitation signal dimensionality problem treatable for efficient quantization and estimation. Nonetheless,

we could investigate another type of wavelet transform that besides splitting the signal into partitions, these are independent of each other. In addition, other representations based on other transformations, such as the cosine transform or the Karhounen-Loeve transform, could be explored in order to develop new quantization methods that avoid the high computational cost in the generation of quantization dictionaries.

The proposal for using a compact FEC code allows the restoration the last lost frame with the least synthesis error. This will also reduce the effect of the error propagation but it does not avoid it. Thus, a possible improvement of this scheme would be that the FEC code would provide the excitation signal which not only minimizes the synthesis error with the original signal but also avoids the error propagation.

Finally, in order to use the new representation of the excitation signal as a FEC code, during the quantization process of the obtained partitions, all of them have been quantized with 10 bits. This coding could imply a high bitrate so it would be advisable to perform a different quantization process for each partition, according to its entropy, or improving the synthesized signal by using the matrix approach.

Bibliografía

[1] J. Joskowicz, "Codificación de voz y video," *PhD Thesis. Instituto de Ingeniera Elćtrica. Universidad de la República de Uruguay*, March 2013.

[2] J.L. Carmona Maqueda, "Reconocimiento de voz codificada sobre redes IP," *PhD. Thesis*, 2010.

[3] W. Stallings, "Comunicaciones y Redes de Computadores," *Prentice Hall*, 2000.

[4] "Ericsson mobility report "http://www.ericsson.com/mobility-report"[Disponible online]," .

[5] A. Arjona, C. Westphal, A. Ylä-Jääski, and M. Kristensson, "Towards High Quality VoIP in 3G Networks: An Empirical Study," *In proceedings of the 2008 Fourth Advanced International Conference on Telecommunications. IEEE Computer Society*, pp. 143–150, 2008.

[6] Orion Hodson, Colin Perkins, and Vicky Hardman, "A survey of packet loss recovery techniques for streaming audio," *IEEE Network*, vol. 12, no. 5, pp. 40–48, Sept. 1998.

[7] F. Le Huche and A. Allali, "La Voz. Anatomía y Fisiología de los Órganos de la Voz y el Habla," 1996.

[8] A.V. Oppenheim, R.W. Schafer, and J.R. Buck, "Discrete-time Signal Processing," *Prentice Hall*, 1975.

[9] J. Markel and A. Gray, "Linear Prediction of Speech," *Springer-Verlag*, , no. 2.3.1, 1976.

[10] P.P. Vaidynathan, "The Theory of Linear Prediction," *Morgan & Claypool publishers*, 2008.

[11] S. Saito and F. Itakura, "The Theoretical Consideration of Statistically Optimum Methods for Speech Spectral Density," *Electrical Communication Laboratory N.T.T Tokyo*, vol. 43, no. 3107, 1966.

[12] B.S. Atal and M.R. Schroeder, "Adaptive Predictive Coding of Speech Signals," *Bell System Technical Journal*, vol. 49, no. 8, pp. 1973–1986, Oct.

[13] M.R. Schroeder and B.S. Atal, "Code-excited linear prediction (CELP): high-quality speech at very low bit rates," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1985)*, vol. 10, pp. 937–940, Apr. 1985.

[14] A.S. Spanias, "Speech coding: a tutorial review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994.

[15] J.P. Burg, "Maximum Entropy Spectral Analysis," *Proceeding 37th meeting of Society of Exploration Geophysicists*, 1967.

[16] W.B. Kleijn and K.K.Paliwal, "Speech coding and synthesis," *Elsevier*, 1995.

[17] T.S. Rappaport, "Wireless communications. Principles & Practice," *Prentize Hall PTR*, 1996.

[18] ITU.T Recommendation G.711, "Pulse Code Modulation (PCM) of voice frequencies," p. 3.4, Oct. 1988.

[19] J. Gibson, "Adaptive prediction in speech diferential encoding systems," *Proceedings of IEEE*, vol. 68, pp. 488–525, 1982.

[20] N.S. Jayant, "Digital coding of speech waveforms: PCM, DPCM and DM quantizers," *Proceedings of IEEE*, vol. 62, pp. 611–632, 1974.

[21] ITU-T Recommendation G.721, "32 kb/s Adaptive Differential Pulse Code Modulation (ADPCM)," Oct. 1988.

[22] ITU-T Recommendation G.722, "7 kHz Audio-Coding Within 64 kbit/s," Nov. 1988.

[23] ITU-T Recommendation G.723, "Extensions of Recommendation G.721 ADPCM to 24 and 40 kbits/s for DCME applications," Oct. 198.

[24] ITU-T Recommendation G.726, "40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM)," 1990.

[25] J.D. Markel and A.H. Gray, "Linear Prediction of Speech," *Berlin: Springer Verlag*, 1976.

[26] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictivesystems," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1975)*, vol. ASSP-23, pp. 309–325, 1975.

[27] J.D. Markel and A.H. Gray, "Quantization and bit allocation in speech processing," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1976)*, vol. ASSP-24, pp. 459–473, 1976.

[28] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Am.*, vol. 57, Apr. 1975.

[29] F.K. Soong and B.H. Juang, "Line Spectrum Pair (LSP) and Speech Data Compression," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1984)*, vol. 9, pp. 37–40, Mar. 1984.

[30] N. Sugamura and N. Farvardin, "Quantizer design in LSP speech analysis-synthesis," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 432–440, 1988.

[31] R.J. McAuley and T.F. Quatieri, "Speech Analysis-Synthesis Based on a Sinusoidal Representation," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1986)*, vol. ASSP-34, pp. 744–754, 1986.

[32] T. Tremain, "The goverment standard linear predictive coding algorithm: LPC-10," *Speech Technology*, pp. 40–49, April 1982.

[33] R.P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1989)*, pp. 467–478, 1989.

[34] P. Kroon and B.S. Atal, "On the use of pitch predictors with high temporal resolution," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1991)*, vol. 3, no. 39, pp. 733–735, 1991.

[35] B.S. Atal and M.R. Schoroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1979)*, pp. 247–254, June 1979.

[36] B.S. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit-rates," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1982)*, vol. 7, pp. 614–617, 1982.

[37] S. Sinhal and B.S. Atal, "Amplitude optimization and pitch prediction in multipulse coders," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1989)*, vol. 37, pp. 317–327, 1989.

[38] ETSI EN 300 726 v8.0.1, "Enhanced Full Rate (EFR) speech transcoding," 2000.

[39] 3GPP TS 26.090, "Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec," .

[40] S. Bruhn et al., "Standardization of the new EVS Codec," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2015)*, Apr. 2015.

[41] 3GPP TS.126.445, "Codec for Enhanced Voice Services (EVS); Detailed algorithmic description," .

[42] C. Lamblin and H. Taddei, "Terms of Reference (ToR) and Time Schedule for the G.722 Packet Loss Concealment (G.722 PLC) standardisation," *ITU-T WP3/16 Doc. AC-06-14*, Jun. 2006.

[43] C. Laflamme, J. Adoul, H. Su, and S. Morisette, "On reducing computational complexity of codebook search in CELP coders through the use of algebraic codes," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1990)*, pp. 177–180, 1990.

[44] 3GPP TS.26.235, "Packet switched conversational multimedia applications; default codecs," .

[45] IETF RFC 3951, "Internet Low Bit Rate Codec (iLBC)," 2004.

[46] "DECT Technologies: http://www.etsi.org/technologies-clusters/technologies/dect," *[Disponible online]*.

[47] ETSI EN 300 175, "Digital Enhanced Cordless Telecommunications Common Interface," 1992.

[48] S. Sain, "Modelling and Characterization of Wireless Channels in Harsh Environments," *PhD Thesis 1154. Vasteras*, June 2011.

[49] J. Postel, "Transmission Control Protocol," *RFC 793*, 1981.

[50] ETSI TS 36.300, "LTE:E-ULTRA & E-ULTRAN," 2009.

[51] "Understanding Delay in Packet Voice Networks "http://www.cisco.com/c/en/us/support/docs/voice/voice-quality/5125-delay-details.html"[Disponible online]," .

[52] Recommendation ITU-T G.114, "General Recommendations on the transmission quality for an entire international telephone connection," 2003.

[53] J. Bolot, "End-to-end packet delay and loss behavior in the Internet," *ACM Sig-comm*, pp. 289–298, 1993.

[54] M. Yajnik and S. Moon and J. Kurose and D. Towsley, "Measurement and modelling of the temporal dependence in packet loss," *IEEE INFOCOM 99*, vol. 1, pp. 345–352, 1999.

[55] M. Borella and D. Swider and S. Uludag, "Internet packet loss: Measurement and implications for end-to-end QoS," *Proceedings of the 1998 ICPP Workshop*, pp. 3–12, Aug. 1998.

[56] N.F. Maxemchuk and S. Lo, "Measurement and interpretation of voice traffic on the Internet," *Proceedings of IEEE ICC*, vol. 1, pp. 500–507, 1997.

[57] J. Bolot, "Adaptive FEC-based error control for Internet telephony," *Proceedings of IEEE INFOCOM 99*, vol. 3, pp. 1453–1460, 1999.

[58] H. Sanneck and G. Carle, "A framework model for packet loss metrics based on loss runlengths," *in proceedings of IEEE Global Internet*, pp. 554–557, 1996.

[59] W. Jiang and H. Schulzrinne, "Modeling of packet loss and delay and their effect on Real-Time multiemdia service quality," *in proceedings of NOSSDAV*, 2000.

[60] K. Salamatian and S. Vaton, "Hidden Markov modelling for network communication channels," *in proceedings of ACM SIGMETRICS*, pp. 92–101, June 2001.

[61] A.M. Gomez, A.M. Peinado, V. Sánchez, B.P. Milner, and A.J. Rubio, "Statistical-based reconstruction methods for speech recognition in IP networks," *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, 2004.

[62] B. Milner and A. James, "An analyisis of packet loss models for distributed speech recognition," *in proceedings of INTERSPEECH-ICSLP*, 2004.

[63] B. Delaney, "Increased robustness against bit errors for distributed speech recognition in wireless enviroments," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP 2005)*, vol. 1, pp. 313–316, 2005.

[64] K. Andrews, C. Heegard, and D. Kozen, "A theory of interleavers," *Computer Science Department in Cornell University (Tech. Rep. 97-1634)*, 1997.

[65] J. Ramsey, "Realization of optimum interleavers," *IEEE Transactions on Information Theory*, vol. 6, pp. 338–345, 1970.

[66] M. Serizawa and H. Ito, "A packet loss recovery method using packet arrived behind the playout time for celp decoding," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, pp. 169–172, 2002.

[67] T. Vaillancourt, M. Jelinek, R. Salami, , and R. Lefebvre, "Efficient frame erasure concealment in predictive speech codecs using glottal pulse resynchronisation," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2007)*, vol. 4, pp. 1113–1116, apr 2007.

[68] A.M. Gomez, J.L. Carmona, J.A. González, and V. Sánchez, "One-pulse FEC coding for robust CELP-coded speech transmission over erasure channels," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 894–904, 2011.

[69] A.M. Gomez, J.L. Carmona, A.M. Peinado, and V. Sánchez, "A multipulse-based forward error correction technique for robust CELP-coded speech transmission over erasure channels," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1258–1268, Aug. 2010.

[70] F. Lahouti, A.K. Khandani, and A. Saleh, "Robust Transmission of Multistage Vector ChannelsApplications to MELP Speech Codec," *IEEE Trans. on Vehicular Technology*, vol. 55, no. 6, pp. 1805–1811, Nov. 2006.

[71] M. Chibani, R. Lefebvre, and P. Gournay, "Fast recovery for a CELP-like speech codec after a frame erasure," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2485–2495, 2007.

[72] A.M. Gomez, J.L. Pérez-Córdoba, and B. Geiser, "Backwards-Compatible Error Propagation recovery for the AMR codec over erasure channels," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2013)*, pp. 8174–8178, May 2013.

[73] F. Merazka, "The Use of FEC method for Packet Loss Concealment for CELP Based Coders in Packet Networks," *2012 Wireless Advanced*, pp. 138–142, June 2012.

[74] C. Feldbauer and W.B. Kleijn, "An Adaptive, Scalable Packet Loss Recovery Method," *IEEE Acoustics, Speech and Signal Processing*, vol. 4, pp. 1117–1120, 2007.

[75] R.M Warren, "Auditory perception," *Pergamon Press Inc.*, 1982.

[76] W.T. Liao, J.C. Chen, and M.S. Chen, "Adaptive Recovery Techniques for Real-Time Audio Streams," *IEEE INFOCOM*, vol. 2, pp. 815–823, 2001.

[77] E. Gunduzhan and K. Momtahan, "Linear prediction based packet loss concealment algorithm for PCM coded speech," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 8, pp. 778–785, 2001.

[78] M. Serizawa and Y. Nozawa, "A packet loss concealment method using pitch waveform repetition and internal state update on the decoded speech for the sub-band ADPCM wideband speech codec," *IEEE Workshop Proceedings on Speech Coding*, vol. 11, pp. 68–70, Aug. 2002.

[79] M. Toyoshima and T. Shinamura, "Packet loss concealment for VoIP based on Pitch Waveform Replication and Linear Predictive Coding," *APCCAS*, pp. 89–92, 2014.

[80] N. Aoki, "Voip packet loss concealment based on two-side pitch waveform replication technique using steganography," *TENCON*, vol. 3, pp. 52–55, Nov 2014.

[81] D.J. Goodman, G.B. Lockhart, O.J. Wasem, and W.C. Wong, "Waveform substitution techniques for recovering missing speech segments in packet voice communications," *IEEE Trans. Acoustics, Speech Signal Processing*, vol. 34, pp. 1440–1448, Dec. 1986.

[82] O.J. Wasem, D.J. Goodman, C.A. Dvorak, and H.G. Page, "The effect of waveform substitution on the quality of pcm packet communications," *IEEE Trans. Acoustics, Speech Signal Processing*, vol. 36, pp. 342–348, March 1988.

[83] O.J. Wasem, D.J. Goodman, C.A. Dvorak, and H.G. Page, "Implementation Aspects of a Novel Speech Packet Loss Concealment Method," *International Symposium on Circuits and Systems*, vol. 3, pp. 2867–2870, 2005.

[84] A.M. Gómez García, "Tratamiento de la Degradación debida al Canal en Sistemas de Reconocimiento Remoto," *PhD. Thesis*, 2006.

[85] E. Zavarehei and S. Vaseghi, "Interpolation of lost speech segments using LP-HNM model with codebook-mapping post-processing," *IEEE Workshop on Automatic Speech Recognition and Understanding*, vol. 2, pp. 14–18, 2007.

[86] M. Yuito and N. Matsuo, "A new sample-interpolation method for recovering missing speech samples in packet voice communications," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1989)*, pp. 381–384, 1989.

[87] F. Merazka, "Packet loss concealment by interpolation for speech over IP network services.," *CIWSP*, pp. 1–4, 2013.

[88] M. Erdol, C. Castelluccia, and A. Zilouchian, "Recovery of missing speech packets using the short-time energy and zero-crossing measurements," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 295–303, July 1993.

[89] E. Zavarehei and S. Vaseghi, "Interpolation of Lost Speech Segments Using LP-HNM Model with Codebook Post-Processing," *IEEE Trans. on Multimedia*, vol. 10, no. 3, pp. 493–502, April 2008.

[90] S.R. Miralavi, S. Ghorshi, M. Mortazavi, and J. Choupan, "Packet loss replacement in VoIP using a recursive low-order autoregressive model-based speech," *8th International Multi-Conference on Systems, Signals and Devices*, pp. 1–4, 2011.

[91] J. Lindbrom and P. Hedelin, "Packet los concealment based on sinusoidal extrapolation," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP)*, vol. 1, pp. 173–176, 2002.

[92] C.A. Rodbro, M.G. Christensen, S.V. Andersen, and S.H. Jensen, "Compressed domain packet loss concealment of sinusoidally coded speech," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP 2003)*, vol. 1, pp. 104–107, April 2003.

[93] J. Chen, "Packet loss concealment based on extrapolation of speech waveform," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP 2009)*, pp. 4129–4132, 2009.

[94] A.M. Peinado, V. Sánchez, J. Pérez-Córdoba, and A. Torre, "HMM-based channel error mitigation and its application to distributed speech recognition," *Speech Communication*, , no. 41, pp. 549–563, 2003.

[95] T. Fingscheidt and P. Vary, "Softbit Speech Decoding: A New Approach to Error Concealment," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 240–251, 2001.

[96] Recommendation ITU-P.862.1, "Mapping function for transforming P.862 raw result scores to MOS-LQO," 2001.

[97] Recommendation ITU P.800, "Methods for subjective determination of transmission quality," 1996.

[98] Recommendation ITU P.830, "Subjective performance assesment of telephone-band and wideband digital codecs," 1996.

[99] Recommendation ITU-R BS.1534-1, "Method for the subjective assessment of intermediate quality level of coding systems," 2001.

[100] Recommendation ITU-R BS.1116-1, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," 1997.

[101] P. Vary and R. Martin, "Digital Speech Transmission: Enhancement, Coding and Error Concealment," *Wiley*, 2006.

[102] Recommendation ITU-P.862, "Objective quality measurement of telephone-band (300-3400 Hz) speech codecs ," 1998.

[103] Recommendation ITU-T P.862, "Perceptual evaluation of speech quality (PESQ)," 2001.

[104] J.S. Garofolo, L. Lamel, W. Fisher, J. Fiscus adn D. Pellett, N. Dehlegren, and V. Zue, "The structure and format of the DARPA TIMIT CD-ROM Prototype," *in Proceedings of NIST*, pp. 1–9, 1988.

[105] NTT, "Multi-Lingual Speech Database for Telephonometry," 1994.

[106] A. Moreno and D. Poch, "An analyisis of packet loss models for distributed speech recognition," *in proceedings of INTERSPEECH-ICSLP*, 2004.

[107] J.G. Proakis, "Digital Communications," *McGraw-Hill, New York*, 1989.

[108] T. Fingscheidt, "Graceful Degradation in ADPCM Speech Transmission," *Proceedings of DAGA*, pp. 748–749, March 2003.

[109] ETSI TS 102 527, "Digital Enhanced Cordless Telecommunications New Generation DECT," 2007.

[110] S. Feldes, "Enhancing Robustness of Coded LPC-Spectra to Channel Errors by Use of Residual Redundancy," *in Proceedings of EUROSPEECH*, pp. 1147–1150, 1993.

[111] C.G. Gerlach, "A probabilistic Framework for Optimum Speech Extrapolation in Digital Mobile Radio," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 1993)*, vol. 2, pp. 419–422, 1993.

[112] V. Eksler and M. Jelinek, "Transition mode coding for source controlled CELP codecs," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP 2008)*, pp. 4001–4004, March 2008.

[113] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. on Commun.*, vol. 28, no. 1, pp. 84–95, 1980.

[114] C.E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, 1948.

[115] L. Kaufman and P. Rousseuw, "Finding Groups in Data: An Introduction to Cluster Analysis," *Willey*, 1990.

[116] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, "Digital watermaking and steganography," *Morgan Kaufmann Publishers Inc.*, 2008.

[117] B. Geiser and P. Vary, "High Rate Data Hiding in ACELP Speech Codecs," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP 2008)*, pp. 4005–4008, March 2008.

[118] S.P. Lloyd, "Leasts squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. 28, no. 2, pp. 129–137, March 1982.

[119] M.A. Kohler and R.K. Yarlagadda, "Markov chain prediction for missing speech frame compensation," *IEEE Workshop on Speech Coding*, pp. 75–77, Sep. 2000.

[120] C.A. Rodbro, M.N. Murthi, S.V. Andersen, and S.H. Jensen, "Hidden Markov Model-Based Packet Loss Concealment for Voice Over IP," *IEEE Transactions on audio, speech and language processing*, vol. 14, no. 5, pp. 1609–1622, Sep. 2006.

[121] R. Martin, C. Hoelper, and I. Wittke, "Estimation of missing LSF parameters using gaussian mixture models," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP 2001)*, vol. 2, pp. 729–732, May 2001.

[122] G. Zhang and W.B. Kleijn, "Autoregressive model-based speech packet-loss concealment," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP)*, vol. 1, pp. 4797–4800, 2008.

[123] Z. Ma, R. Martin, J. Guo, and H. Zhang, "Nonlinear estimation of missing LSF parameters by a mixture of Dirichlet distributions," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP 2014)*, pp. 6929–6933, May 2014.

[124] Y.Agiomyrgiannakis and Y.Stylianou, "Coding with side information techniques for LSF reconstruction in voice over IP," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP 2005)*, vol. 1, pp. 141–144, 2005.

[125] C. Boubakir and D. Berkani, "The estimation of line spectral frequencies trajectories based on unscented kalman filtering," *International Multi-Conference on Systems, Signals and Devices*, pp. 1–6, 2009.

[126] C. Feldbauer and W.B. Kleijn, "Scalable coding with side information for packet loss recovery," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP 2009)*, vol. 57, no. 8, pp. 2309–2319, Aug. 2009.

[127] F. Merazka, "Differential quantization of spectral parameters for CELP based coders in packet networks," *IECON*, pp. 1495–1498, Oct. 2012.

[128] F. Lahouti and A.K. Khandani, "Approximating and exploiting the residual redundancies-applications to efficient reconstruction of speech over noisy channels ," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP 2001)*, vol. 2, pp. 721–724, 2001.

[129] Y. Agiomyrgiannakis and Y. Stylianou, "Coding with Side Information Techniques for LSF Reconstruction in Voice over IP," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP 2005)*, vol. 1, pp. 141–144, 2005.

[130] D. Chazan, R. Hoory, G. Cohen, and M. Zibulski, "Speech reconstruction from MEL frequency cepstral coefficients and pitch frequency," *IEEE International Conference on Acoustics, Speech and Signal Proccessing (ICASSP 2000)*, vol. 3, pp. 1299–1302, 2000.

[131] S. Subasingha, M.N. Murhi, and S.V. Andersen, "A Kalman filtering approach to GMM predictive coding of LSFs for packet loss conditions," *16th International Conference on Digital Signal Processing*, pp. 1–6, July 2009.

[132] S.O. Haykin, "Adaptive filter theory," *Prentice-Hall*, 1991.

[133] S.C. Douglas, "Introduction to adaptive filters," *Digital signal processing handbook (Ch.18)*, 1999.

[134] C. Papaodysseus, G. Roussopoulos, and A. Panagopoulos, "Using a fast RLS adaptive algorithm for efficient speech processing," *Mathematics and Computers in Simulation*, vol. 68, no. 2, pp. 105–113, April 2005.

[135] C. Paleologu, J. Benesty, and S. Ciochină, "A Robust Variable Forgetting Factor Recursive Least-Squares Algorithm for System Identification," *IEEE Signal Processing*, vol. 15, pp. 597–600, 2008.

[136] A. Haar, "Zur Theorie der Orthogonalen Funktionensysteme," *Mathematische Annalen*, vol. 69, no. 3, pp. 331–371, 1910.

[137] J.S. Walker, "A primer on wavelets and their scientific applications," *Chapman and Hall-CRC*, 1999.

[138] S.S. Iyengar, E.C. Cho, and V.V. Phoha, "Foundations of Wavelet Networks and Applications," *Chapman and Hall*, 2002.

[139] A. Grossmann and J. Morlet, "Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape," *Siam Jorunal On Mathematical Anlysis*, vol. 15, pp. 723–736, 1984.

[140] S. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Trans. on Patt. Anal. and Machine Intell.*, vol. 11, no. 7, pp. 674–693, 1989.

[141] J. Kovacevic and M. Vetterli, "Non separable multidimensional perfect reconstruction filter banks and wavelet bases for R^n," *IEEE Trans. on Information Theory*, vol. 38, pp. 533–555, 1992.

[142] D. López-Oller, S. Han, A.M. Gomez, J.L. Pérez-Córdoba, and T. Fingscheidt, "System-compatible robustness improvement for New Generation DECT decoders by G.722 soft-decision decoding," *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP 2016)*, pp. 5945–5949, 2016.

[143] D. López-Oller, A.M. Gomez, and J.L. Pérez-Córdoba, "Residual VQ-quantization for speech frame loss concealment," *IberSpeech 2014*, pp. 91–100, Nov. 2014.

[144] D. López-Oller, A.M. Gomez, J.L. Pérez-Córdoba, B. Geiser, and P. Vary, "Steganographic Pulse-Based Recovery for Robust ACELP Transmission over Erasure Channels," *IberSpeech 2012*, pp. 257–266, Nov. 2012.

[145] D. López-Oller, A.M. Gomez, and J.L. Pérez-Córdoba, "Source-based error mitigation for speech transmissions over erasure channels," *Proceeding of EUSIPCO 2014*, pp. 1242–1246, Sept 2014.

[146] D. López-Oller, A.M. Gomez, J.L. Pérez-Córdoba, and V. Sánchez, "An error mitigation technique for erasure channels based on a Wavelet representation of the speech excitation signal," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1245–1256, July 2016.

[147] D. López-Oller, N. Benamirouche, A.M. Gomez, and J.L. Pérez-Córdoba, "Novel excitation signal estimation method to regenerate lost speech frames for transmissions over erasure channels," *[In revision] Speech Communication*, 2017.

[148] D. López-Oller, A.M. Gomez, and J.L. Pérez-Córdoba, "A novel error mitigation scheme based on replacements vectors and FEC codes for speech recovery in loss-prone channels," *IberSpeech 2016*, pp. 44–53, Nov. 2016.