Universidad Zaragoza
1542

Proyecto Fin de Carrera

Optimization of an automotive high-definition voice hands-free system

Autor:

Ester Morte Júdez

Director:

Marc-André Jung

Ponente:

Luis Vicente Borruel

Escuela de Ingeniería y Arquitectura

2015

Resumen

Los algoritmos encargados de compensar el eco acústico (ing: acoustic echo cancelation, AEC) reducen las componentes de dicho eco que aparecen frecuentemente en sistemas de comunicación manos libres. Sin embargo, y especialmente en aquellos sistemas de aplicación en automóviles implementados para soportar voz en banda ancha (HD Voice), los requisitos necesarios establecidos por recomendaciones internacionales no se pueden alcanzar únicamente mediante algoritmos de cancelación de eco; a veces es necesaria la implementación de otros bloques adicionales que ayuden a cumplir con dichos requisitos. De esta manera, el objetivo de este proyecto es el de perfeccionar un ya existente sistema de manos libres y ampliarlo con un algoritmo de control de ganancia. Este algoritmo permite reducir el eco durante una conversación telefónica a la vez que mantiene la mejor inteligibilidad posible para ambos usuarios.

Así, este proyecto consiste en la implementación en MATLAB de un algoritmo robusto para reducir el efecto de las señales de eco presentes en la comunicación. Se ha comprobado la idoneidad de los algoritmos para el control de ganancia descritos en la literatura y el elegido ha debido adaptarse a las necesidades del sistema manos libres existente. Además, ha debido mantenerse la compatibilidad de este algoritmo para su futura implementación en un procesador digital de la señal. Los procedimientos matemáticos y otros específicos del algoritmo descritos en la literatura se han recogido y han sido críticamente revisados y probados. Adicionalmente, se ha desarrollado y utilizado un entorno de simulación realista capaz de proporcionar todos los datos de entrada relevantes al algoritmo a desarrollar y además se ha realizado una evaluación de las señales intermedias y de salida. Para ello, se han realizado las siguientes medidas y han sido posteriormente discutidas: cancelación de eco, calidad de la señal de salida y calidad del sistema en casos de habla simultánea. Además se ha proporcionado información y datos sobre los parámetros utilizados y los distintos tipos de errores cometidos por el algoritmo. De la misma manera, se ha examinado críticamente la compatibilidad con el algoritmo de cancelación de eco ya existente incluso en condiciones adversas de ruido.

El algoritmo finalmente implementado se ha probado y documentado de manera completa. Se han mostrado las posibles limitaciones tanto del propio algoritmo como de sus posibles aplicaciones, así como las diferentes recomendaciones para su mejora. Ha sido también necesario llevar a cabo una evaluación detallada y sistemática basada en recomendaciones y métodos de medida que son aceptados internacionalmente.

Palabras clave sistema manos libres, cancelación de eco acústico, detección de actividad de voz, control de ganancia

Índice de contenidos

Índice de tablas

Índice de figuras

Lista de abreviaturas

– # –	
3GPP	3rd Generation Partnership Project
– A –	
AEC	Acoustic Echo Canceler
AP	Affine Projection
– C –	
CSS	Composite Source Signal
– D –	
DFT	Discrete Fourier Transformation
DSP	Digital Signal Processor
– E –	
ETSI	European Telecommunications Standards Institute
– F –	
FAP	Fast Affine Projection
FBE	Filter Bank Equalizer
FDAF	Frequency Domain Adaptive Filter
FFT	Fast Fourier Transformation
FIR	Finite Impulse Response
– G –	
GLC	Gain Loss Control
– H –	
HD	High-Definition
– I –	
IDFT	Inverse Discrete Fourier Tranformation
IIR	Infinite Impulse Response
ITU	International Telecommunication Union
ITU-T	Standardization Sector of the International Telecommunication Union

– **L** –

LDF	Low Delay Filter
LEM	Loudspeaker Enclosure Microphone
LMS	Least Mean Square
LTSD	Long-Term Spectral Divergence
LTSE	Long-Term Spectral Envelope

– **M** –

MIPS	Million Instructions Per Second
MOS	Mean Opinion Score
MRP	Mouth Reference Point
MSE	Mean Square Error

– **N** –

NB	Narrowband
NLMS	Normalized Least Mean Square

– **P** –

PESQ	Perceptual Evaluation of Speech Quality
PF	Postfilter
PN	Pseudo-Noise
PSD	Power Spectral Density

– **Q** –

QMF	Quadratic Mirror Filter

– **R** –

RLMS	Recursive Least Mean Square
RLS	Recursive Least Square

– **S** –

SER	Signal to Echo Ratio
SNR	Signal to Noise Ratio

– **V** –

VAD	Voice Activity Detector

– **W** –

WB	Wideband
WGN	White Gaussian Noise

– **Z** –

ZCR	Zero-Crossing Rate

1 Introducción

En los últimos años, el uso de sistemas de comunicación manos libres ha aumentado de tal manera que dicha tecnología ha resultado ser muy útil e incluso indispensable para algunas situaciones. Por ejemplo, se utiliza habitualmente para llevar a cabo teleconferencias, puede resultar de gran ayuda a personas con discapacidad y la mayoría de los automóviles lleva instalado uno de estos sistemas. Un sistema de manos libres es un sistema *front-end* [LDC04], o interfaz, que se utiliza en situaciones que requieren un cierto grado de atención. Este interfaz se encuentra normalmente situado en una habitación o espacio cerrado y está formado por uno o más micrófonos que recogen el audio presente en la habitación y uno o más altavoces que se encargan de reproducir la señal proveniente del usuario o usuarios remotos que participan en la conversación.

En la figura 1.1 se representa un esquema de este tipo de comunicación en la que participan un usuario local y un usuario remoto. El dispositivo que permite la comunicación manos libres se sitúa en el llamado extremo cercano, o en inglés *near-end side*, que se corresponde con el extremo en el que se encuentra usuario local. En la parte opuesta, se ve el extremo lejano, o en inglés *far-end side*, correspondiente al usuario remoto. Dos canales o enlaces independientes permiten la comunicación entre estos usuarios. Así, el usuario local recibe la señal proveniente del usuario remoto a través del enlace *downlink*, o de bajada, que se reproduce por medio del altavoz; y de la misma manera, el micrófono recoge las señales presentes en el extremo cercano y se transmiten hacia el otro extremo a través del enlace *uplink*, o de subida.

Cuando se usa este tipo de dispositivos, el usuario se encuentra normalmente posicionado a una cierta distancia del micrófono y del altavoz. Por esta razón, la calidad de la señal de voz que recoge el micrófono es inferior a la de aquella que pueda tenerse en una comunicación telefónica al uso. De hecho, en una comunicación manos libres, el micrófono capta todas las señales de audio presentes en la habitación o habitáculo en el que se encuentra situado. Así, y como puede verse en la figura 1.1, la señal captada por el micrófono estará finalmente compuesta por:

- Voz del usuario local *(near-end speech).*
- Ruido de fondo presente en la habitación (voz de otras personas, ruido ambiente, ruido propio del automóvil, etc.)
- Señal de eco proveniente del usuario remoto.

Como puede verse, la señal del micrófono está formada por distintas señales, de las cuales sólo nos interesa la voz del usuario local, que es la que se desea transmitir al extremo remoto. Sin embargo, la presencia tanto de ruido de fondo como del eco degrada la señal de voz y hacen que ésta no sea fácilmente inteligible, produciéndose un efecto molesto para el usuario a la vez que se dificulta la conversación y disminuye su calidad. Con el objetivo

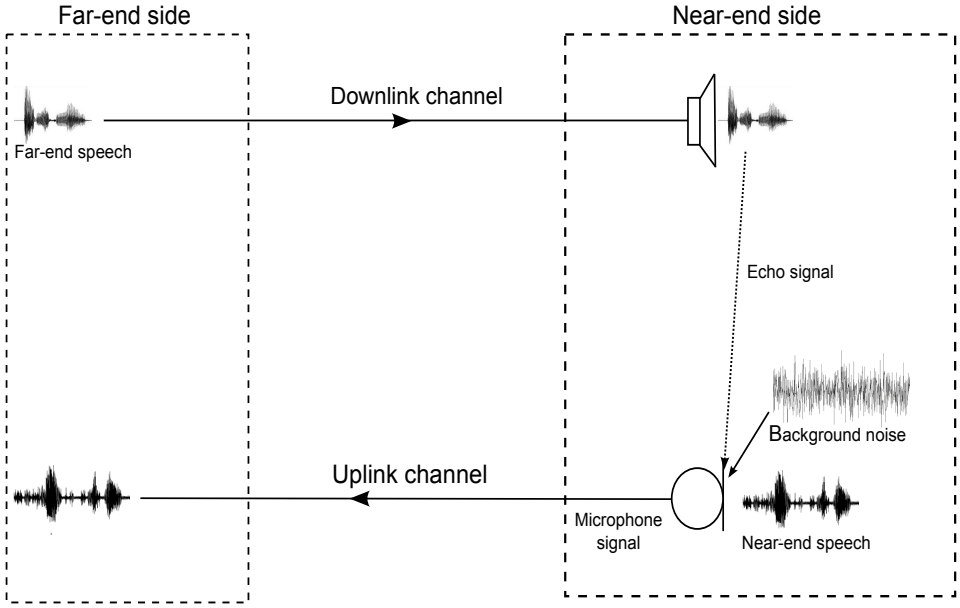


Figura 1.1: Esquema de una comunicación manos libres.

de reducir estos efectos negativos que producen dichas señales indeseadas, se implementan diferentes bloques de procesamiento en los sistemas manos libres. Éstos son el cancelador de eco acústico – *Acoustic Echo Cancelers* (AECs) – y post-filtro – *Postfilters* (PFs) – que se encargan de suprimir la señal de eco y reducir el ruido respectivamente. Después de que la señal del micrófono sea procesada por estos dos bloques, el resultado es una 'señal de voz aproximada' a la del usuario local donde tanto el efecto del eco como del ruido se han visto altamente reducidos, mejorando de esta manera la calidad de la señal y de la conversación. El proyecto que aquí se documenta se ha desarrollado con el objetivo de mejora de un sistema manos libres para automóviles desarrollado por el *'Institut für Nachrichtentechnik'* de la 'TU Braunschweig', en Alemania.

1.1 Motivación y objetivos

Uno de los puntos principales de investigación del departamento de tratamiento de señal del mencionado *'Institut für Nachrichtentechnik'* es el procesamiento de señales de voz y audio. A su vez, dentro de este campo, se encargan de la investigación y desarrollo de métodos de reducción de ruido, cancelación de eco acústico y de métodos de medición de calidad en el marco de la estandarización internacional ITU-T, presentando gran atención a este último punto.

A este respecto, en el sistema manos libres desarrollado en dicho departamento, se detectó una carencia en cuanto al cumplimiento de las recomendaciones exigidas. La recomendación ITU-T G.167 [ITU93] trata los controladores de eco acústico utilizados en

aplicaciones de teleconferencia, teléfonos de altavoz (manos libres), terminales telefónicos y aplicaciones móviles y personales. Además, se especifican las características y valores de calidad de funcionamiento que deben cumplir los controladores; siendo estas características las que se espera que el sistema manos libres tenga. Así, en la sección 5.4 de dicha recomendación se establecen distintas especificaciones con respecto a la atenuación que debe sufrir la señal de eco en distintos caso, y que se recoge de la siguiente manera:

- 'Atenuación ponderada por acoplamiento del terminal – monoloquia': atenuación que debe sufrir la señal de eco cuando únicamente haya un interlocutor activo, situación a la que nos referimos en este documento como *single talk*. Dicha atenuación deberá ser de, por lo menos, 45 dB.
- 'Atenuación ponderada por acoplamiento del terminal – habla simultánea': atenuación que debe sufrir la señal de eco cuando ambos interlocutores estén activos, situación que se conoce como *double talk*. El valor de esta atenuación debe ser de, por lo menos, 30 dB.

Como ya se ha comentado, se detectó que el sistema no alcanzaba estos valores de atenuación exigidos y fue por ello que se decidió realizar este proyecto con el objetivo de solucionar esta carencia a la vez que mantener o incluso mejorar la calidad ya proporcionada por el sistema ya existente.

1.2 Estructura del documento

Después de esta breve introducción, el documento está formado por otros 4 capítulos en los que se recogen los distintos aspectos de este proyecto.
En el capítulo 2 se realiza una breve introducción sobre los sistemas manos libres haciendo hincapié en aquellos que ofrecen servicios de voz HD. Además se hace una pequeña diferenciación entre los distintos bloques que los componen, realizando un breve análisis de los distintos algoritmos que los pueden implementar. Con ello, se llega a la conclusión de cuáles son los mejores algoritmos para implementar un sistema manos libres de aplicación en automóviles, lo que lleva a una descripción más detallada del algoritmo implementado en el '*Institut für Nachrichtentechnik*'. Finalmente se explica cuál ha sido el problema a tratar, objetivo de este proyecto.
A continuación, el capítulo 3 trata concretamente la solución adoptada, describiendo detalladamente el algoritmo que se ha decidido implementar, distinguiendo entre las distintas partes que lo componen con el objetivo de facilitar su comprensión. Igualmente, se abordan problemas encontrados durante la implementación y las modificaciones que han sido necesarias realizar tanto para solucionar dichos problemas como para mejorar los resultados obtenidos.
En el capítulo número 4 se analizan en profundidad los experimentos realizados para la evaluación del sistema llevado a cabo en este proyecto, mostrando los resultados obtenidos más relevantes. Para ello, se explican los métodos utilizados para cada experimento, se presentan los resultados obtenidos de esta metodología y finalmente se analizan dichos resultados discutiendo su idoneidad.
Por último, en el capítulo 5 se recogen y resumen las conclusiones que se han obtenido durante la realización del proyecto y al realizar la evaluación. Se presentan también posibles líneas futuras de desarrollo y mejora, proponiendo mejoras y ampliaciones al sistema.

2 Sistema manos libres de aplicación en el automóvil con calidad de voz HD

Los sistemas manos libres que soportan servicios de voz en alta definición o banda ancha – ancho de banda desde los 50 Hz hasta los 7000 Hz – o 'High-Definition (HD) Voice' permiten llevar a cabo comunicaciones de elevada calidad e inteligibilidad, mejorando así los sistemas de comunicación desarrollados para voz en banda estrecha. Cada vez más, los sistemas de comunicaciones para oficinas, viviendas o automóviles requieren las prestaciones que proporcionan los sistemas de manos libres de alta calidad, e incluso, en el caso de los automóviles, son de obligado cumplimiento en muchos países.

2.1 Introducción a los sistemas manos libres

En muchas ocasiones, los sistemas manos libres deben ser diseñados para hacer frente a degradaciones de la señal de voz que vienen dadas por el entorno acústico. El eco acústico y el ruido de fondo son normalmente la causa de estas degradaciones, llevando así a una reducción tanto de la inteligibilidad como de la calidad de las señales de voz. Estos efectos son especialmente importantes cuando hablamos de los sistemas instalados en automóviles donde pueden darse, por ejemplo, señales de ruido de elevada amplitud. Como ya se ha dicho anteriormente en el capítulo 1, AECs [Shy92, WS85, Say03, EM83] y PFs [YIEB10] se implementan con el objetivo de paliar o eliminar estos efectos negativos. Los algoritmos propuestos para ello, se han desarrollado normalmente para voz en banda estrecha, trabajando así con una frecuencia de muestreo de $f_s = 8\,\text{kHz}$; sin embargo, a los nuevos sistemas que soportan voz en banda anda, les corresponde una frecuencia de muestreo de $f_s = 16\,\text{kHz}$. Es por ello que los antiguos algoritmos para banda estrecha deben adaptarse para banda ancha, lo que lleva a un aumento de la complejidad del algoritmo e incluso pueden producirse efectos negativos e indeseados [BSV06].

Algunos de los algoritmos en dominio temporal típicamente implementados para los canceladores de eco son: '*Normalized Leas Mean Square*' (NLMS) [SSS04], '*Affine Projection*' (AP) [SSS04], [GT95], '*Reccursive Least Squares*' (RLS) [Cio84] o Kalman [Kal60]. Estos algoritmos normalmente suelen presentar una estructura sencilla basada en un procesamiento de la señal muestra a muestra, además de introducir un retardo muy pequeño. Este filtro debe aproximar el camino del eco y en el caso de que dicho filtro deba adaptarse para cada muestra de la señal para modelar de respuestas impulsionales largas, puede tener como consecuencia una elevada complejidad computacional. Además, los AECs implementados en dominio temporal no obtienen buenos resultados en los casos de *double talk* ya que la presencia de actividad en el extremo local puede llevar a una adaptación incorrecta del filtro, y con ello a una falsa estimación de la respuesta impulsional.

Aunque puedan introducir otro tipo de problemas, los algoritmos desarrollados para señales divididas en sub-bandas pueden salvar muchas de las deficiencias que presentan los algoritmos en dominio temporal. Al separar la señal de banda completa en sub-bandas, se puede reducir el coste computacional de la adaptación del filtro en el caso de que sea necesario modelar respuestas impulsionales largas. Además, estos algoritmos presentan también una mayor velocidad de convergencia. No obstante, ha de considerarse el mal funcionamiento en los casos de *double talk* y que el retardo introducido por este tipo de algoritmos es mayor que en el caso de los de dominio temporal.

Como alternativa a estos últimos, se presentan los filtros adaptativos en dominio frecuencial, *Frequency Domain Adaptive Filter* (FDAF) [SG65]. En ellos, tanto la adaptación de la respuesta impulsional como la estimación de la señal de eco se realizan en el dominio frecuencial. Esto permite calcular parámetros que sean dependientes de la frecuencia que lleven a obtener resultados más óptimos. Otra gran ventaja que presentan algunos algoritmos FDAF, es su buen rendimiento en los casos de *double talk*. Además, son capaces de preservar la calidad de la componente de voz en el canal ascendente. Desafortunadamente, para computar la transformación discreta de Fourier – *Discrete Fourier Transform* (DFT) – es necesario almacenar bloques de muestras, lo cual introduce un retardo elevado en el canal ascendente.

Normalmente, ya que el bloque de AEC no es capaz de suprimir la cantidad de eco requerida o deseada, es necesaria la implementación de un postfiltro que ayude al cumplimiento de esta tarea. Este tipo de filtros también son capaces de suprimir las componentes no lineales de eco y pueden reducir la componente de ruido presente en la señal. Normalmente, tanto AEC como postfiltro se implementan en el mismo dominio. Así, los postfiltros en dominio temporal son computacionalmente eficientes pero presentan mal rendimiento en los casos de *double talk*, mientras que los postfiltros en dominio frecuencial o en sub-bandas presentan un mejor rendimiento con el inconveniente de añadir retardo en la señal.

El sistema manos libres desarrollado por el '*Institut für Nachrichtentechnik*' está pensado para proporcionar servicios de voz HD en aplicaciones para automóviles con respuestas impulsionales relativamente cortas. En estas aplicaciones, es necesario obtener buenos resultados en las situaciones de *double talk* con el objetivo de que el conductor no se distraiga. Ya que presenta un excelente rendimiento en las situaciones de *double talk* e introduce un retardo aceptable, se decidió implementar un algoritmo adaptativo en dominio frecuencial basado en filtrado de Kalman [Enz06, EV06]. En la siguiente sección, se presenta el sistema manos libres implementado junto con una breve explicación de su funcionamiento.

2.2 Sistema basado en algoritmo FDAF

En esta sección, se va a presentar el sistema descrito por *Jung* y *Fingscheidt* en [JF14] y que ha sido la base sobre la que se ha trabajado en este proyecto. Como ya se ha comentado anteriormente, los sistemas manos libres implementados en el dominio frecuencial son una buena elección si se desea que se cumplan las siguientes características: baja complejidad computacional en el caso de respuestas impulsionales largas, buen rendimiento en situaciones de *double talk* y poca degradación de la señal de voz local. El filtro adapta-

tivo encargado de suprimir la señal de eco se sitúa en paralelo al camino acústico del eco, también conocido como sistema *loudspeaker-enclosure-microphone* (LEM), intentando así estimar una réplica de la señal de eco. En el caso del algoritmo FDAF que aquí se ha implementado, la adaptación de los coeficientes del filtro y la estimación de la señal de eco se realizan en dominio frecuencial.

Tal y como se puede ver en la figura 2.1, la señal de eco $d(n)$ es el resultado de la convolución entre la señal de voz remota $x(n)$ y la respuesta impulsional LEM.
La señal del micrófono viene dada por:

$$\mathbf{y} = [y\,(n - R + 1)\,, ..., y\,(n)]^T \tag{2.1}$$

donde R es la longitud de la trama e $y(n) = s(n) + d(n) + r(n)$, siendo $s(n)$ la señal de voz local y $r(n)$ el ruido de fondo.
La señal $x(n)$ se transforma al dominio frecuencial obteniendo:

$$\begin{aligned}\mathbf{X} &= DFT\left\{[x(n-K+1), ..., x(n-R), x(n-R+1), ..., x(n)]^T\right\} \\ &= [X(l,0), ..., X(l,k), ..., X(l,K-1)]^T\end{aligned} \tag{2.2}$$

siendo l el índice de la trama l-ésima y k el índice frecuencial. A través de la aproximación FDAF basada en la teoría de filtrado de Kalman, se estiman los coeficientes $\hat{W}_1(l,k)$ del filtro adaptativo [EV03]. Con esto, se puede estimar la señal de eco en el dominio frecuencial de la siguiente manera:

$$\hat{D}(l,k) = \hat{W}_1(l,k)X^*(l,k) \tag{2.3}$$

para $k = 0, ..., K-1$. La transformada discreta de Fourier inversa (IDFT) lleva a obtener $\hat{d}(n)$, a partir de la cual se puede calcular la señal de error:

$$e(n) = y(n) - \hat{d}(n). \tag{2.4}$$

La señal de eco residual $d_r(n) = d(n) - \hat{d}(n)$ está presente en la señal de error, ya que esta se puede escribir como:

$$e(n) = r(n) + s(n) + d_r(n) \tag{2.5}$$

Calculando la señal de error en dominio frecuencial:

$$\mathbf{E}_l = DFT\left\{[\mathbf{0}^T_{K-R-O}, (\mathbf{e}^-_{l-1})^T, \mathbf{e}^T_l]\right\} \tag{2.6}$$

donde $\mathbf{0}^T_{K-R-O}$ es un vector nulo (K-R-O)-dimensional,
$\mathbf{e}^-_{l-1} = [e(n-R-O+1), ..., e(n-R)]^T$ y $\mathbf{e}_l = [e(n-R+1), ..., e(n)]^T$ y O el número de muestras que sufren 'overlap'.

En este punto, la señal llega al postfiltro donde tanto el eco residual $R(l,k)$ como la componente de ruido $N(l,k)$ son suprimidas utilizando un filtro de Wiener en dominio frecuencial tal que:

$$\hat{S}(l,k) = \hat{W}^c_2(l,k)E^*(l,k) \tag{2.7}$$

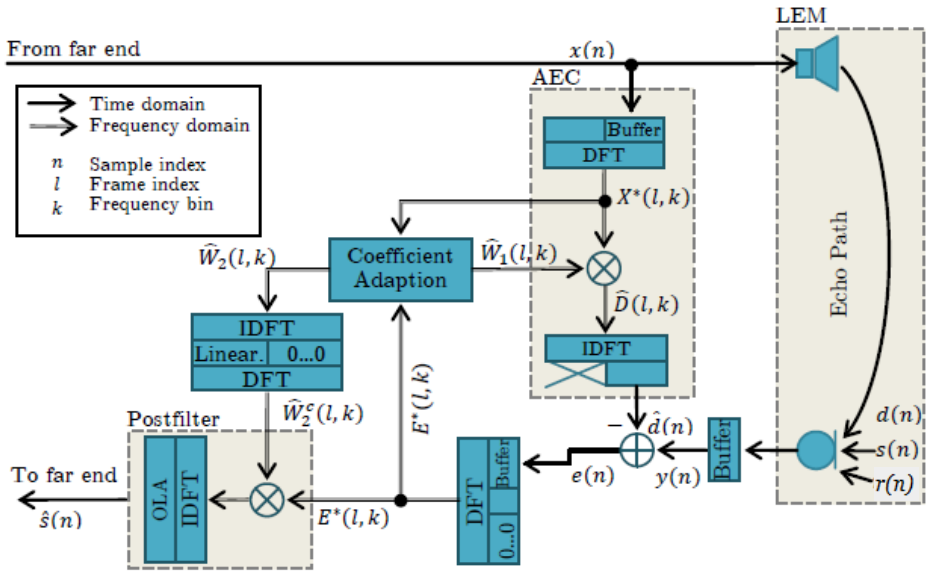


Figura 2.1: Estado del arte del sistema manos libres basado en algoritmo FDAF [JF14].

, obteniendo $\hat{W}_2^c(l,k)$ por medio de $\hat{\mathbf{w}}_{2,l} = IDFT\left\{\hat{\mathbf{W}}_{2,l}\right\}$ de la siguiente manera:

$$\begin{aligned}\hat{w}_{2,l}^c =&[\hat{w}_{2,l}(n=K-N_p/2),...,\hat{w}_{2,l}(n=K-1),\\ &\hat{w}_{2,l}(n=0),...,\hat{w}_{2,l}(n=N_p/2-1),\mathbf{0}_{K-N_p}^T]^T\end{aligned} \tag{2.8}$$

que contiene la respuesta impulsional de fase lineal del postfiltro de longitud $N_p \leq K-R-O$. Tal y como se muestra en el diagrama de bloques de la figura 2.1, los coeficientes en dominio frecuencial tanto del AEC $\hat{W}_1(l,k)$ como del postfiltro $\hat{W}_2(l,k)$ se estiman de manera síncrona. Después del postfiltro se calcula la transformada inversa de la señal resultante, para finalmente obtener la señal de voz estimada $\hat{s}(n)$ y transmitirla al interlocutor situado en el extremo remoto.

2.3 Problema a tratar

Se ha comentado anteriormente en el capítulo 1 que el sistema existente descrito en la sección anterior, no cumple con los requisitos exigidos por la recomendación ITU-T G.167 [ITU93] por escasamente unos 4 dBs. Es por ello, que para solucionar dicho problema se ha decidido implementar un control de ganancia y añadirlo al sistema ya existente. El control de ganancia, que en sus orígenes funcionaba como un conmutador entre canales cancelando el eco pero dando lugar a una comunicación *half-duplex*, en este caso, permite distinguir entre los distintos estados en los que se puede encontrar la conversación e introducir la ganancia deseada para cada uno de ellos. Estos estados son: *single talk* remoto, *single talk* local, *double talk* y silencio; en los dos primeros casos se atenuará el canal no activo, mientras que en los dos últimos la atenuación se repartirá en ambos canales. Los valores de atenuación a insertar, deberán de ser lo suficientemente elevados como para cumplir con

la recomendación ITU-T G.167 [ITU93], pero no más de lo necesario ya que, como esta atenuación se aplica a la señal total captada por el micrófono también se atenúa la señal de voz local, y si esta atenuación fuera muy grande, la voz que se desea transmitir se haría inaudible y se degradaría la calidad de la comunicación.

Así pues, el objetivo de este proyecto consiste en implementar un bloque de control de ganancia que complemente y funcione conjuntamente con el AEC y postfiltro del sistema ya existente, como puede verse en la figura 2.2 y que permita además que el sistema cumpla con los requisitos ya descritos en el capítulo 1.

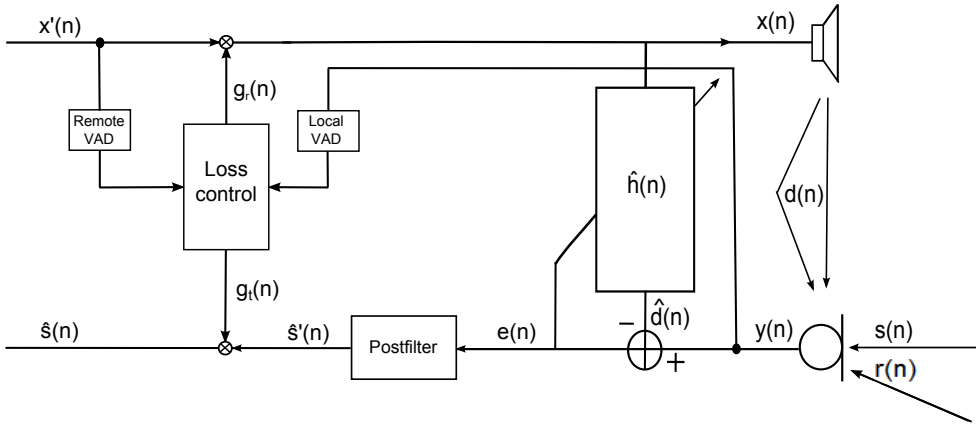


Figura 2.2: Sistema manos libres formado por un AEC, postfiltro y un bloque de control de ganancia.

.

Es por ello que con los valores que el sistema manos libres ya alcanzaba, se decide que la atenuación a insertar por el control de ganancia sea la siguiente:

- 6dB en el canal inactivo para los casos de *single talk*.
- 3dB en cada canal para los casos de *double talk*.

Ya que el control de ganancia decidido presenta la ventaja de poder implementarse en cualquier sistema manos libres y con el objetivo de conocer y comprender mejor el tema, se ha recurrido a la literatura para investigar sobre los distintos tipos de algoritmos de AECs y PFs, viendo las ventajas y desventajas que presenta su implementación. Como resultado de esta investigación, se han recogido en el anexo A los aspectos principales de algunos de los algoritmos ya comentados.

3 Control de ganancia

En el capítulo anterior ya se ha descrito someramente el objetivo del proyecto y se ha introducido brevemente el concepto de control de ganancia. En este capítulo vamos a presentar de manera concreta el algoritmo que ha decidido implementarse y las modificaciones que han sido necesarias realizar para tener un funcionamiento óptimo. Además, se incluyen las fórmulas necesarias para su correcto entendimiento.

3.1 Algoritmo de control de ganancia

El algoritmo para el control de ganancia que se ha implementado en la realización de este proyecto es el propuesto por *Hänsler y Schmidt* en [HS06]. Se ha llevado a cabo una investigación en la literatura concluyendo que éste es el algoritmo más actual y con el que se obtienen buenos resultados. Este bloque de control de ganancia se debe situar, según [BSV06], al final de la línea de procesamiento teniendo así que tratar solamente con las componentes residuales de eco y de ruido de fondo. Las señales de entrada a este bloque son las presentes en el inicio de los canales de comunicación, las cuales se corresponden con la señal que proviene del usuario remoto $x(n)$ y la señal capturada por el micrófono $y(n)$, tal y como puede verse en la figura 2.2. A continuación se describe el algoritmo propuesto por los autores; para cualquier información más detallada, el lector puede consultar [HS06].

3.1.1 Detección de actividad de voz

El algoritmo de control de ganancia a implementar, puede dividirse a su vez en distintas partes. La primera de ellas, consiste en sendos bloques de detección de actividad de voz, o *Voice Activity Detection* (VAD), tanto para la señal de voz remota como para la señal del micrófono local. Estos detectores determinarán qué partes de la señal presentan actividad vocal y cuáles no, discriminando así la voz tanto del ruido de fondo como de la señal de eco. Para realizar esta detección y de acuerdo con [HS06], este algoritmo calcula la magnitud *short-term* de las señales correspondientes, la cual puede estimarse por medio de un filtro de suavizado de respuesta impulsional infinita, filtro *Infinite Impulse Response* (IIR). A fin de mejorar la detección, la constante de tiempo del filtro va a tener diferentes valores para detectar el principio y final de las palabras; esto se hace así porque la detección del comienzo de una palabra, o frase, debe realizarse más rápidamente que la detección del final.
Las magnitudes y constantes de tiempo se calculan de la siguiente manera.
Para la señal que proviene del usuario remoto:

$$\overline{|x(n)|} = (1 - \gamma_x(n))|x(n)| + \gamma_x(n)\overline{|x(n-1)|} \tag{3.1}$$

$$\gamma_x(n) = \begin{cases} \gamma_r, & \text{si } |x(n)| > \overline{|x(n-1)|}, \\ \gamma_f, & \text{en otro caso.} \end{cases} \tag{3.2}$$

Y de la misma manera para la señal del micrófono:

$$\overline{|y(n)|} = (1 - \gamma_y(n))|y(n)| + \gamma_y(n)\overline{|y(n-1)|} \tag{3.3}$$

$$\gamma_y(n) = \begin{cases} \gamma_r, & \text{si } |y(n)| > \overline{|y(n-1)|}, \\ \gamma_f, & \text{en otro caso.} \end{cases} \tag{3.4}$$

donde γ_r y γ_f son las constantes de tiempo de subida y bajada respectivamente y las magnitudes $\overline{|x(n)|}$ e $\overline{|y(n)|}$ se conocen como magnitudes *short-term*.

Teniendo ya estas magnitudes, puede ejecutarse el algoritmo propio de VAD, el cual se basa en un mecanismo de conteo. Cuando se detecta actividad de voz, se establece un contador a su valor máximo, el cual se mantiene durante el tiempo correspondiente a N muestras, periodo que se establece para evitar falsas detecciones en cortos periodos de habla y durante los que se realizaría una detección de silencio. En el momento en el que ya no se detecta actividad de voz, el valor del contador se decrementa de manera lineal.

$$t_{act,far}(n) = \begin{cases} N_{far}, & \text{si } \overline{|x(n)|} > K_{far}, \\ max\left\{t_{act,far}(n-1) - 1, 0\right\}, & \text{en otro caso.} \end{cases} \tag{3.5}$$

Como ya se ha comentado, esta detección debe realizarse para las dos señales de entrada, siendo más sencilla en el caso de la señal que proviene del usuario remoto que en la del micrófono, debido a la presencia del eco en esta última. Así pues, para la señal remota, la detección de voz será positiva siempre que la magnitud $\overline{|x(n)|}$ tenga un valor superior al del umbral K_{far}, el cual debe ser unos decibelios mayor que el nivel de ruido de fondo remoto.

No obstante, debido a la señal de eco que se crea por el acople existente entre el altavoz y el micrófono, otros parámetros deben de tenerse en cuenta a la hora de detectar actividad del usuario local. Con el objetivo de reducir las falsas detecciones que la componente de eco puede causar, el nivel del eco debe ser estimado. Por ello, es necesario introducir en la detección de actividad local el parámetro K_c, o factor de acople, el cual debe elegirse unos decibelios por debajo del nivel de eco. Por lo demás, el algoritmo de detección sigue los mismos principios que para el caso remoto. Se detecta actividad de voz cuando $\overline{|y(n)|}$ presenta un valor mayor al del ruido de fondo local K_{loc} y/o el eco.

$$t_{act,loc}(n) = \begin{cases} N_{loc}, & \text{si } \overline{|y(n)|} > max\left\{K_{loc}, K_c\overline{|x(n)|}g_r(n-1)\right\} \\ max\left\{t_{act,loc}(n-1) - 1, 0\right\}, & \text{en otro caso,} \end{cases} \tag{3.6}$$

donde $g_r(n-1)$ es la ganancia previa aplicada en el canal *uplink*.
Por ser en este caso la detección más complicada, el tiempo durante el cual se mantiene la actividad de voz detectada se aumenta, siendo $N_{loc} > N_{far}$. De esta manera también nos aseguramos de no cometer un gran número falsas detecciones durante los periodos de actividad vocal local y de que la ganancia aplicada no presente muchas variaciones en cortos periodos de tiempo.

3.1.2 Decisión de estados e inserción de ganancia

Una vez que ya realizada la detección de voz en ambas señales se puede distinguir, a partir del valor de los contadores, en qué estado de la conversación nos encontramos en cada momento, *single talk* remoto, *single talk* local, *double talk* o silencio, y dependiendo de esto, calcular el valor de ganancia que debe insertarse en cada canal. Tanto la ganancia en recepción como en transmisión se calculan de la misma manera, únicamente deben sustituirse los parámetros del extremo remoto por los del local y viceversa. Por ello, aquí se muestran solamente los cálculos necesarios para obtener el valor de la ganancia en recepción. Las condiciones que describen los posibles estados de la comunicación son las siguientes:

$$\textit{Single talk} \text{ local} \leftarrow (t_{act,far}(n) \leq 0) \wedge (t_{act,loc}(n) > 0), \tag{3.7}$$

$$\text{Silencio} \begin{cases} (t_{act,far}(n) \leq 0) \wedge (t_{act,loc}(n) \leq 0) \wedge (g_r(n) > \sqrt{g_{min}}) \\ (t_{act,far}(n) \leq 0) \wedge (t_{act,loc}(n) \leq 0) \wedge (g_r(n) \leq \sqrt{g_{min}}) \end{cases} \tag{3.8}$$

$$\textit{Single talk} \text{ remoto} \leftarrow (t_{act,far}(n) > 0) \wedge (t_{act,loc}(n) \leq 0), \tag{3.9}$$

$$\textit{Double talk} \begin{cases} (t_{act,far}(n) > 0) \wedge (t_{act,loc}(n) > 0) \wedge (g_r(n) > \sqrt{g_{dt}}) \\ (t_{act,far}(n) > 0) \wedge (t_{act,loc}(n) > 0) \wedge (g_r(n) \leq \sqrt{g_{dt}}) \end{cases} \tag{3.10}$$

donde, cada una de las ecuaciones anteriores describe las condiciones que deben darse para cada uno de los posibles estados. A partir de estas condiciones, se pueden calcular los valores de ganancia que han de insertarse en cada canal de comunicación y se hará de la siguiente manera:

$$g_r(n) = \begin{cases} max\left\{g_r(n-1)g_{dec}, g_{min}\right\}, & \text{si 3.7} \\ max\left\{g_r(n-1)g_{dec}, \sqrt{g_{min}}\right\}, & \text{si 3.8/1} \\ min\left\{g_r(n-1)g_{inc}, \sqrt{g_{min}}\right\}, & \text{si 3.8/2} \\ min\left\{g_r(n-1)g_{inc}, g_{max}\right\}, & \text{si 3.9} \\ max\left\{g_r(n-1)g_{dec}, \sqrt{g_{dt}}\right\}, & \text{si 3.10/1} \\ min\left\{g_r(n-1)g_{inc}, \sqrt{g_{dt}}\right\}, & \text{si 3.10/2} \end{cases} \tag{3.11}$$

Así queda completamente descrita la forma de calcular la ganancia que debe insertarse en el canal de recepción. Si nos encontramos en la situación de *single talk* local, debería minimizarse la influencia del canal en recepción para que, de esta manera, no haya una completa interrupción hacia el interlocutor local. Esta es la razón por la que $g_r(n)$ – ganancia insertada en el canal de recepción – debe decrementarse hasta que alcanza un valor mínimo, g_{min}, que se ha decidido previamente para los casos de *single talk*. Para las situaciones de *single talk* remoto, el canal de recepción debe de estar completamente abierto, para que la actividad del interlocutor remoto se reciba sin problemas. Es por eso que $g_r(n)$ debe incrementarse hasta que alcance su valor máximo $g_{max} = 1$. En las situaciones de silencio y *double talk*, la correspondiente ganancia g_{min} y g_{dt} debe repartirse respectivamente para cada caso entre ambos canales. Dependiendo del valor actual de $g_r(n)$, éste debe aumentarse o decrementarse hasta que se alcance el valor final deseado. El caso del canal de transmisión se desarrolla de la misma manera, únicamente cambiando desde la ecuación

3.7 hasta la 3.11 los parámetros de ganancia en recepción por transmisión. Los valores de los que se ha hablado, necesarios para la implementación del algoritmo, son los siguientes:

- $g_{max} = \quad 0\text{dB} \; = 1$
- $g_{min} = -6\text{dB} \; = 0.5012$
- $g_{dt} \quad = -3\text{dB} \; = 0.7079$

Durante el desarrollo del proyecto, la correcta estimación de los niveles de ruido de fondo y de la señal de eco ha resultado ser un problema complicado y en el que su resolución ha tenido mucho que ver con los resultados experimentales. En un inicio se intentó dar un valor constante a los parámetros K_{far}, K_{loc} y K_c (de los que se ha hablado previamente en la sección 3.1) para luego realizar un ajuste manual y ver si existe algún patrón o relación entre ellos y otros parámetros del sistema. Esta tarea no tuvo resultados positivos ya que, cualquier pequeño cambio realizado en las características del experimento conllevaba la necesidad de realizar un ajuste muy fino. Es por ello que se decidió buscar un algoritmo que estimara los niveles del ruido de fondo y del eco.

3.1.3 Ruido de fondo

Algunos métodos como el filtrado IIR durante las pausas de la voz y algunos otros basados en *minimum statistics* proporcionan unos buenos resultados a la hora de estimar el nivel de ruido de fondo. De la misma manera que le método propuesto en [Mar93], no es necesario distinguir entre periodos de actividad e inactividad vocal para calcular el nivel de ruido; si no que la potencia del ruido en cada instante de tiempo es el mínimo de la potencia estimada de la señal dentro de una ventana de longitud específica. Sin embargo, ya que nuestro algoritmo trabaja muestra a muestra, no nos permite la implementación de este método, el cual necesita tomar información de la señal dentro de la longitud de dicha ventana. Por ello, se implementa un algoritmo propuesto igualmente por *Hänsler y Schmidt* en [HS06] basado en un filtrado IIR y que se implementa de la siguiente manera:

$$\overline{y^2(n)} = (1 - \gamma_y(n))y^2(n) + \gamma_y(n)\overline{y^2(n-1)} \tag{3.12}$$

$$\gamma_y(n) = \begin{cases} \gamma_r, & \text{si } y^2(n) > \overline{y^2(n-1)}, \\ \gamma_f, & \text{en otro caso.} \end{cases} \tag{3.13}$$

La magnitud $\overline{y^2(n)}$ se conoce como *short-term power* y debe calcularse tanto para la señal $y(n)$ como para $x(n)$ y así estimar el nivel de ruido de fondo en ambos extremos de la comunicación. A las constantes de tiempo γ_r and γ_f se les darán los mismos valores que los calculados en la sección 3.1.1.
Con esto, el nivel de ruido de fondo ya puede estimarse de la siguiente manera:

$$\overline{\hat{b}^2(n)} = min\left\{\overline{y^2(n)}, \overline{\hat{b}^2(n-1)}\right\}(1+\epsilon) \tag{3.14}$$

donde ϵ es una pequeña constante determinada experimentalmente que evita que el valor de la estimación se estanque en un mínimo.

3.1.4 Factor de acoplamiento

Además del nivel de ruido de fondo, también debe calcularse el factor de acoplamiento, que permite estimar el nivel de la señal de eco, para así tenerla en cuenta en la detección de actividad local y descartarla como voz del usuario local. Aquí se ha recurrido también a otro método propuesto por los autores *Hänsler y Schmidt* en [HS06] ya que de esta manera, se parte ya de un alto grado de compatibilidad y robustez entre estos métodos con el algoritmo de VAD. El método estima el factor de acoplamiento durante periodos de actividad remota, ya que es cuando se produce un acople entre esta señal transmitida por el altavoz y la capturada por el micrófono, y se formula de la siguiente manera:

$$K_c(n) = \begin{cases} \gamma \mathrm{K}_c(n-1) + (1-\gamma)\frac{\overline{y^2(n)}}{\overline{x^2(n)}}, & \text{si se detecta actividad remota,} \\ K_c(n-1), & \text{en otro caso.} \end{cases} \tag{3.15}$$

donde γ es una constante de suavizado, con un valor cercano a 1.
Tanto este algoritmo como los explicados en las secciones previas se han programado, implementado y optimizado hasta obtener los mejores resultados posibles.

3.2 Modificaciones del algoritmo

Con el objetivo de que el algoritmo de control de ganancia funcione de la mejor manera posible, hay que realizar una serie de modificaciones que se recogen a continuación.

3.2.1 Compatibilización con la estimación del ruido de fondo y factor de acoplamiento

La primero que hay que realizar para que el algoritmo de control de ganancia funcione sin problemas es hacerlo completamente compatible con los algoritmos que estiman el nivel de ruido de fondo y de factor de acoplamiento. Para ello habrá que realizar modificaciones en algunas fórmulas de aquel. En [HS06], para las ecuaciones 3.12 y 3.15 se recomienda el cálculo de la *short-term power* para realizar las estimaciones. Sin embargo, también se cita que el mismo procedimiento puede realizarse mediante la *short-term magnitude*, como en [Hei97]. Es importante tener en cuenta qué método ha sido el utilizado pues, a la hora de comparar señales que se han calculado con diferente método, habría de aplicarse un factor que las relaciona. Como puede verse en las ecuaciones 3.5 y 3.6, se realiza una comparación entre la *short-term magnitude* de la señal remota y de la señal del micrófono con el nivel de ruido de fondo. Así pues, en este caso, ambos términos a comparar deberían calcularse utilizando el mismo método. Se ha realizado una comprobación experimental sobre cuál de ellos es el que permite obtener mejores resultados, y éstos se consiguen mediante el cálculo de *short-term power*. Es por ello que debe realizarse un cambio en las fórmulas del algoritmo de detección de actividad vocal previamente explicado, donde ha de tomarse el valor de la señal al cuadrado en lugar de su valor absoluto. Este cambio afecta a todas las ecuaciones recogidas en 3.1.1.

3.2.2 Compatibilización con el sistema manos libres

Como también se ha comentado anteriormente, el algoritmo dedicado al control de ganancia debe implementarse en un sistema manos libres ya existente y funcionar en consonancia con éste. Debido a la configuración de dicho sistema, tanto la generación de la señal captada por el micrófono $y(n)$ como los cálculos necesarios de los valores de ganancia deben realizarse al principio del canal de transmisión, antes del bloque correspondiente al AEC. Es por ello que es necesario programar una función que llamamos **ini-loss-control.m** y que combina la inicialización del sistema (parámetros del sistema, generación de señales con concretos niveles de SER y SNR, etc.) con el algoritmo de control de ganancia encargado de calcular los valores necesarios.

En la inicialización del algoritmo original, la señal del micrófono podría generarse directamente teniendo en cuenta únicamente el valor de SER deseado. Sin embargo, al añadir el control de ganancia, esta señal debe generarse muestra a muestra, porque cada una de ellas depende de la muestra de la señal de eco, que a su vez depende de la señal $x(n)$ que en este caso se ve afectada por los distintos valores de ganancia introducidos en el canal de recepción en cada momento. Esta situación puede observarse más fácilmente en la figura 2.2. Una vez que a cada muestra de la señal $x(n)$ se le aplica el valor de ganancia correspondiente, deberá convolucionarse con la respuesta impulsional que modela la cabina del coche, obteniendo así la componente de eco. Con esto, se puede generar cada muestra de la señal $y(n)$ y utilizarse tanto para el VAD local como para el AEC. En la figura 3.1 se puede ver un ejemplo en el que se representan las diferentes señales en cada punto del canal de recepción así como la señal de eco generada con el valor de SER deseado.

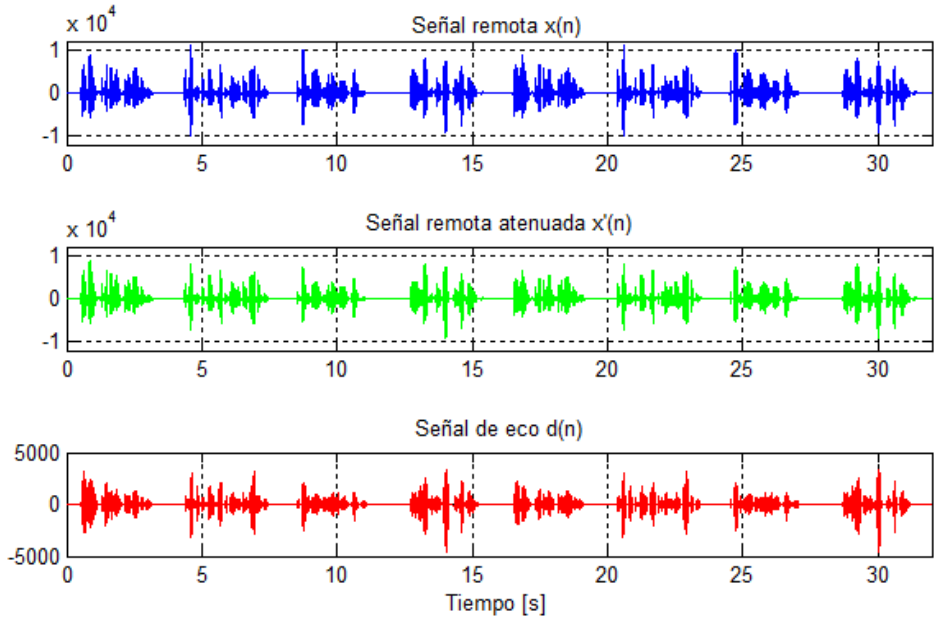


Figura 3.1: Representación en dominio temporal de las señales presentes en el canal de recepción (original y atenuada) $x'(n)$, $x(n)$ así como la señal de eco $d(n)$. Para un valor de $SER = 10\,\mathrm{dB}$

De la misma manera, en la figura 3.2 puede verse la señal del micrófono formada como suma de la señal correspondiente a la voz local y de la señal de eco. En este ejemplo, no se ha añadido ruido de fondo para facilitar la comprensión del algoritmo.

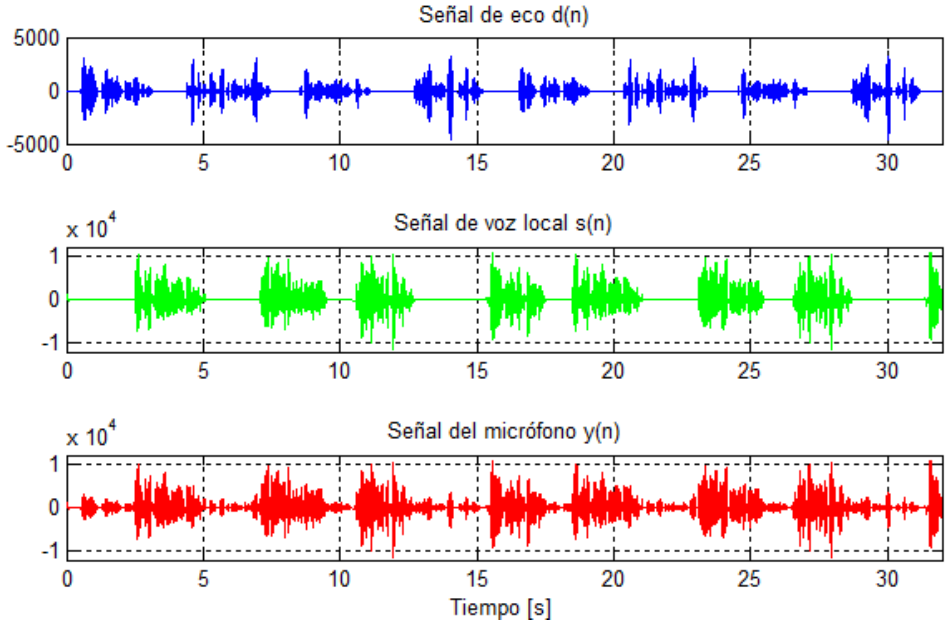


Figura 3.2: Representación en el dominio temporal de la señal de voz $s(n)$, la señal de eco $d(n)$ y la señal total capturada por el micrófono $y(n)$. Para valores de $SER = 10\,\text{dB}$ y $SNR = \infty\,\text{dB}$.

Otro aspecto importante a tener en cuenta es que este algoritmo que trabaja muestra a muestra debe poder implementarse en un procesador digital de la señal o *Digital Signal Processor* (DSP), por lo que debe garantizarse la compatibilidad entre algoritmo y procesador. La frecuencia de muestreo de este DSP para procesamiento muestra a muestra es de $f_s = 48\,\text{kHz}$. Por ello, esta frecuencia de muestreo es la que se habrá de utilizar a la hora de programar el algoritmo de inicialización/control de ganancia muestra a muestra en *Matlab*, como por ejemplo, para calcular los valores de las constantes de tiempo de los filtros. En [HS06] el algoritmo se ha desarrollado y optimizado para una frecuencia de muestreo de $f_s = 8\,\text{kHz}$ y con este valor se calculan las constantes de tiempo. Así pues, dichas constantes deben transformarse para que cumplan con la nueva frecuencia de muestreo de $f_s' = 48\,\text{kHz}$ por medio de las siguientes fórmulas:

$$\tau = \frac{-1}{f_s ln\,(\gamma)} \tag{3.16}$$

$$\gamma' = e^{-\frac{1}{f_s' \tau}} \tag{3.17}$$

donde γ se corresponde con la constante de tiempo para $f_s = 8\,\text{kHz}$, τ con la constante medida en unidades de tiempo absoluto y γ' se corresponde con la constante de tiempo para $f_s' = 48\,\text{kHz}$.

Por otra parte, hay que tener en cuenta que este sistema está diseñado para trabajar con señales de voz en banda ancha y así proveer servicios HD en los que se tiene una frecuencia de muestro de $f_s = 16\,\text{kHz}$. Es por esto, que el sistema original trabaja con este valor existiendo una compatibilidad con el DSP ya que se corresponde con su valor de frecuencia de muestro para algoritmos que procesan la señal en tramas, como es el caso del AEC. Sin embargo, no puede olvidarse que al realizarse muestra a muestra, la frecuencia de muestro en el algoritmo programado en **ini-loss-control.m** debe ser de $f_s^{'} = 48\,\text{kHz}$ garantizando así su posible implementación en el DSP. Es por ello que dichas señales, antes de que sean tomadas como entradas por el AEC, deben diezmarse por un factor 3. Esta función de diezmado se valdrá de un filtro FIR de orden 90 que garantice la compatibilidad entre todos los bloques del sistema manos libres y con su futura implementación en un DSP.

3.2.3 Seguimiento de las variaciones en la señal de voz

Teniendo en cuenta todas estas consideraciones, se pasa a programar el algoritmo. Para los experimentos iniciales, se utilizan señales de voz proporcionadas por la recomendación [ITU12b] las cuales se han modificado con ayuda del programa *Adobe Audition* para crear secuencias de voz y así simular conversaciones en las que estén presentes todas las situaciones posibles. Estos experimentos no han sido útiles a la hora de evaluar el algoritmo, pero han permitido analizar diversas opciones de mejora. El valor de las constantes de tiempo del filtro IIR será la primera modificación a realizar. El objetivo es que el VAD siga las variaciones de la señal de voz de forma tan precisa como sea posible; sin embargo, con las constantes definidas por *Hänsler y Schmidt* el inicio de la actividad de voz no se realiza tan rápidamente como se desea. Las constantes definidas en el libro tienen un valor mayor a 10 ms, siendo éste la duración de las tramas en las que el algoritmo de AEC divide la señal. Con la intención de realizar una detección más precisa, el valor de las constantes de tiempo debe ser menor para que así el VAD trabaje más rápido. Después de analizar varias opciones posibles se ha decidido que el valor de la constante para las pendientes de subida 1 ms, mientras que la de bajada sea de 2 ms. Con esto tendremos que:

$$\gamma_r = e^{-\frac{1}{48\,\text{kHz}1\,\text{ms}}} = 0.9793$$
$$\gamma_f = e^{-\frac{1}{48\,\text{kHz}2\,\text{ms}}} = 0.9896$$

Otra modificación que se ha observado que puede mejorar los resultados obtenidos, es modificando la manera en que se decrementan los contadores $t_{act,far}$ y $t_{act,far}$. Según el algoritmo, estos deben decrementarse linealmente en una unidad por muestra y en el momento en que este valor llegue a cero, se determina que la actividad vocal ha finalizado. Este decremento lineal es muy lento, llevando a una detección de voz incorrecta, ya que la actividad de voz concluye mucho antes que el contador alcance su valor mínimo. Es por ello que prefiere implementarse una caída exponencial, ya que ésta se asemeja más a las características que presentan las señales de voz. Además, se ha incrementado el valor umbral del contador, el cual se encarga de determinar cuál es el momento en el que finaliza la actividad de voz. Los valores tanto de la constante de decaimiento como del umbral se han determinado experimentalmente hasta que se ha observado la mayor similitud posible con el decaimiento de las señales de voz. En la figura 3.3 se representa un ejemplo de la detección de actividad de voz.

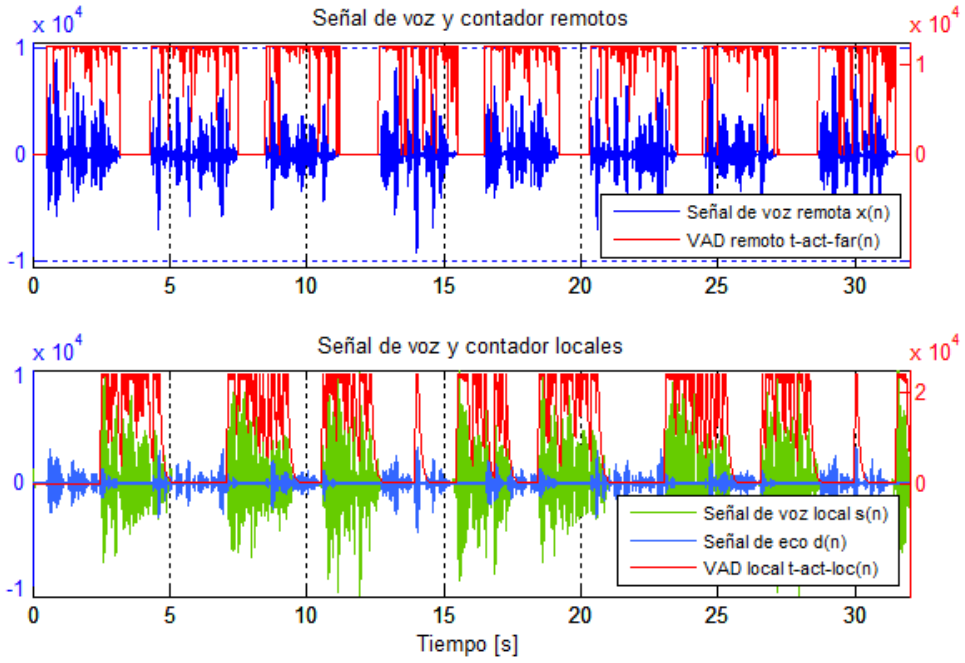


Figura 3.3: Ejemplo de la detección de actividad de voz tanto local como remota con los correspondientes contadores encargados de realizar esta detección. Para un valor de $SER = 10\,\mathrm{dB}$.

Como puede apreciarse en la figura 3.4 se producen falsas detecciones de actividad de voz del usuario local en torno a los instantes equivalentes con 14 y 30 segundos.

En estos instantes, se aprecian picos de elevada amplitud en la señal proveniente del usuario remoto, presentando la señal de eco valores de amplitud tan grandes como los que se alcanzan en la señal local. Por esta razón, cabe esperarse que el sistema cometa estas falsas detecciones.

Además se ha generado y representado en dicha figura una señal de verificación dependiendo del estado de la conversación la cual describe en cuál de ellos se encuentra la conversación en cada instante de tiempo. De esta manera, se puede visualizar de manera sencilla si la detección se realiza correctamente. Para generar esta secuencia se ha asignado un número a cada estado, teniendo 2 para silencio, 1 para *single talk* remoto, 0 para *double talk* y -1 para *single talk* local. El ejemplo representado en esta figura se corresponde con la detección de actividad para las señales de la figura 3.3.

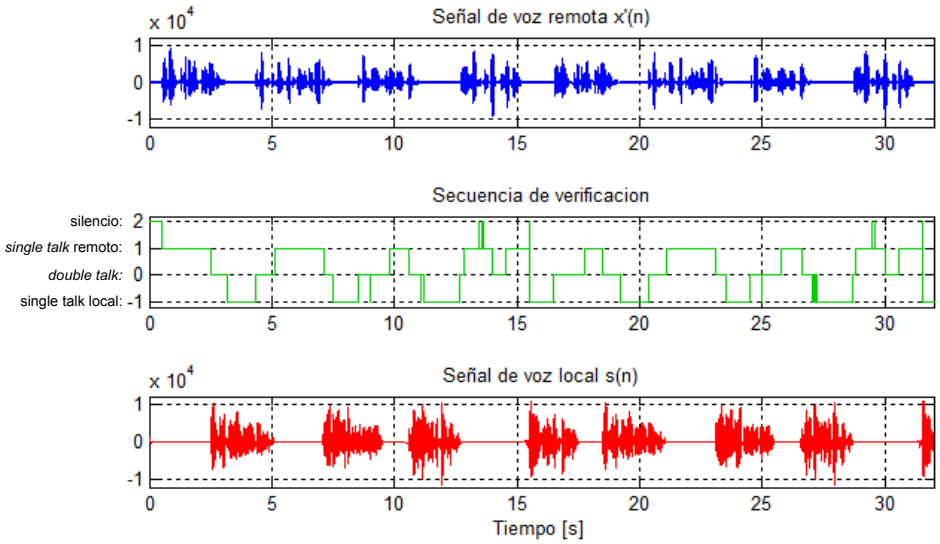


Figura 3.4: Representación de las señales $x(n)$ e $y(n)$ así como de una secuencia de verificación.

4 Experimentos y evaluación

En este capítulo comentan todos aquellos experimentos realizados que han sido relevantes así como sus resultados. Estos experimentos se dividen principalmente en: evaluación de los errores en detección cometidos por el algoritmo de control de ganancia, análisis del comportamiento del sistema de manos libres para situaciones de *double talk* y atenuación sufrida por la señal de eco.

4.1 Evaluación del VAD

Primeramente, se ha querido realizar una evaluación del funcionamiento del detector de actividad de voz presente en el control de ganancia. Para ello, se ha realizado una distinción entre los distintos errores que este bloque puede cometer al realizar una falsa detección, se han evaluado para distintos escenarios y se han analizado los resultados obtenidos.

4.1.1 Experimento con ruido blanco

Con el objetivo de evaluar el algoritmo de control de ganancia y particularmente, de la detección de voz que este realiza, se ha llevado a cabo en primer lugar un experimento con secuencias de ruido blanco. Para ello se generan dos señales de 41 segundos cada una, que representan la actividad remota y local y que están formadas por secuencias de ruido blanco de un segundo de duración, con cuatro valores de amplitud diferentes. Estas dos señales intentan representar todos los diferentes estados de comunicación y aunque no se asimilen a señales de voz reales, su forma permite realizar un seguimiento preciso de los resultados del algoritmo. Estas secuencias presentan una menor variación en amplitud, es por ello que la evolución de algunos factores, parámetros o magnitudes puede apreciarse con mayor claridad, por ejemplo la detección de actividad/inactividad, el valor de ganancia insertado en cada momento, etc. Estas características permiten encontrar errores en la programación y mejorar el algoritmo de una manera más sencilla posibles.

Para realizar el experimento, se crean dos secuencias de ruido blanco, una para cada interlocutor. Las señales están formadas por segmentos de ruido blanco – de distribución normal estándar – de un segundo de duración, donde el nivel de cada uno de ellos puede tomar un valor entre cuatro diferentes. Estos valores son: -16 dBov, -26 dBov, -36 dBov y $-\infty$ dBov[1]. Los diferentes valores representan distintos niveles de volumen de una 'señal de voz' y combinándolos adecuadamente, se pueden generar unas señales que representen todos los estados posibles presentes en una comunicación de voz normal, como pueden ser:

[1]la unidad dBov se refiere a los dB existentes con relación al punto de carga del sistema, más concretamente, es el nivel relativo entre la amplitud de la señal y la máxima amplitud que éste puede soportar sin que se produzcan recortes en la señal. El nivel que corresponde con habla activa, *active speech*, es de -26 dBovsegún ITU-T P.56 2011[ITU11]

silencio, *single talk* remoto, *single talk* local, *double talk*, interrupciones de los usuarios, etc. En la figura 4.1 se representan ambas señales generadas como combinación de segmentos de ruido blanco de diferentes valores de amplitud. En estas secuencias, los segmentos con amplitud nula se corresponden con -∞ dBoV, los que tienen menor amplitud con -36 dBoV, media amplitud con -26 dBoV y los de amplitud mayor con -16 dBoV. Si se analiza la combinación de las dos señales, pueden ver que están representados todos los estados que pueden ocurrir en una conversación: *double talk* con interrupción del interlocutor local al rededor del segundo 15, completa superposición de actividad de voz en el segundo 25, etc. A cada uno de los segmentos de un segundo de duración en la conversación se les ha 'etiquetado' dependiendo de los niveles que ambas señales presentan durante ese tiempo. En la tabla 4.1 se puede ver la clasificación de estos estados y a su vez, en la figura a continuación, se representa también la sucesión de estados correspondiente a las señales generadas, donde se puede apreciar que todos ellos están incluidos.

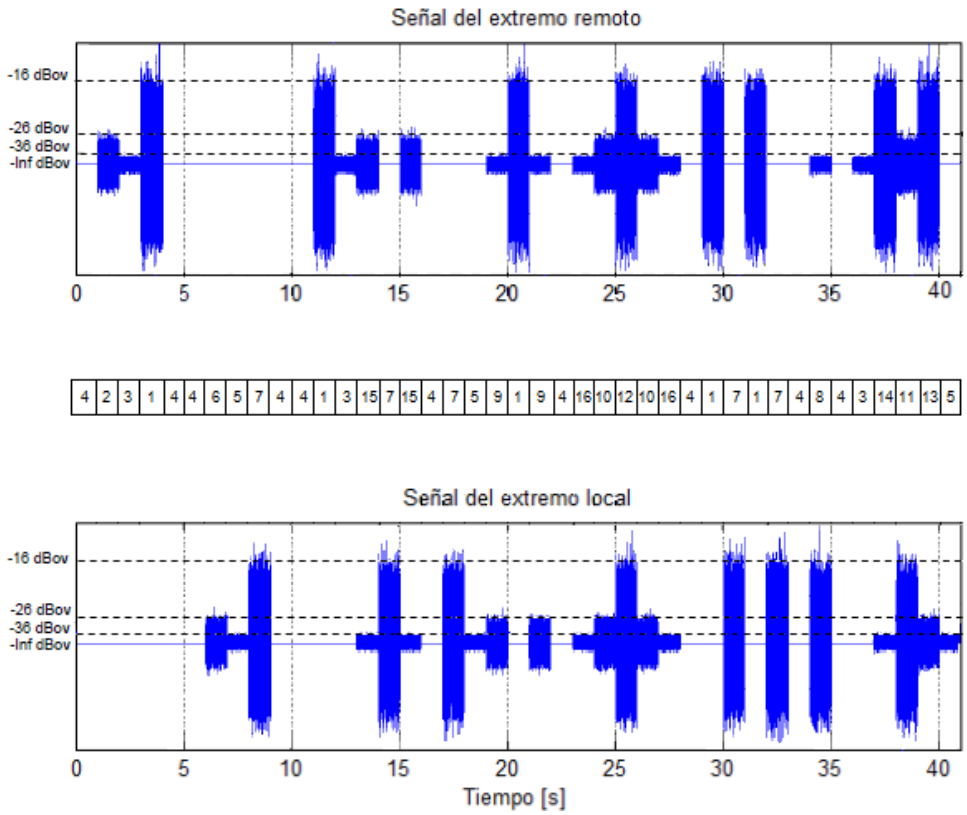


Figura 4.1: Secuencias de ruido blanco generadas para representar la actividad de voz de los interlocutores remoto y local junto con la correspondiente secuencia de estados.

4.1.2 Detección de errores

Según la clasificación realizada que puede verse en la tabla 4.1 se va a proceder a dos etiquetados de los estados de la conversación; uno de ellos es manual y corresponde con los

Tabla 4.1: Estados de la conversación dependiendo del nivel de las señales local y remota.

Nivel de la señal remota	Nivel de la señal local	Etiqueta del estado	Estado de la conversación
−16 dBov	−∞ dBov	1	
−26 dBov	−∞ dBov	2	*single talk* remoto
−36 dBov	−∞ dBov	3	
−∞ dBov	−∞ dBov	4	silencio
−∞ dBov	−36 dBov	5	
−∞ dBov	−26 dBov	6	*single talk* local
−∞ dBov	−16 dBov	7	
−36 dBov	−16 dBov	8	
−36 dBov	−26 dBov	9	
−26 dBov	−36 dBov	10	
−26 dBov	−16 dBov	11	
−16 dBov	−16 dBov	12	*double talk*
−16 dBov	−26 dBov	13	
−16 dBov	−36 dBov	14	
−26 dBov	−36 dBov	15	
−36 dBov	−36 dBov	16	

estados reales y que representaremos en un vector llamado *ground truth*, el otro se realiza automáticamente según la clasificación del algoritmo VAD y se almacena en un vector de verificación que llamamos *check vector* y que se representa en la figura 3.4.

Tras realizar la comparación entre los estados presentes en el vector de *ground truth* y la decisión que el VAD ha realizado, pueden obtenerse los errores cometidos y calcular tanto su ratio en relación a todas las decisiones realizadas como el ratio de cada tipo de error en relación todos los errores cometidos. De esta manera puede comprobarse qué tipos de errores son los que se cometen con mayor frecuencia para diferentes *setups* del algoritmo, permitiéndonos conocer cómo se comporta el VAD en distintas situaciones. Por supuesto, cuanto menos errores se cometan, mejor es el resultado, pero en el caso de que estos errores ocurran, es más favorable si éstos no son los más críticos. La idea a la hora de realizar este experimento con el ruido blanco es el de hacer de manera sencilla un primer análisis del bloque de control de ganancia y ver cómo de buenos son los resultados proporcionados por el VAD. Por esta razón sólo va a simularse este bloque, desactivando el AEC y el postfiltro para que se pueda analizar el funcionamiento del control de ganancia sin la presencia de cualquier otro bloque que pueda influenciar en sus resultados. Por otro lado, en esta situación en la que tenemos secuencias de ruido blanco, no tiene sentido añadir ruido de fondo, es por ello que se evalúa sólo para 'voz limpia' y la caída de los contadores que representan la actividad vocal se realiza de forma lineal como en las ecuaciones 3.6, 3.5, ya que aquí no debe aproximarse a la forma de caída de la voz real.

Clasificación de errores

Asumiendo que el bloque VAD no va a proporcionar resultados perfectos, se definen distintos grados de relevancia entre los distintos tipos de errores que se obtienen al realizar falsas detecciones, como se explica en [GBSA13], y así poder clasificar cuáles de ellos afectan en mayor o menor grado al sistema. Así pues, se compara el vector de *ground truth* con el resultado decidido por el VAD representado en *check vector*, y de esta manera se crea una matriz de ponderación en la que se representa el peso que se le asigna a cada tipo de fallo en la clasificación realizado por el VAD. Esta matriz viene representada en la tabla 4.2. Los pesos de mayor valor se asignan a los errores más críticos, mientras que los de menor valor se asignan a los errores en los que el fallo en la clasificación no es tan desmesurado. De esta manera, los errores se dividen en tres grupos que se explican a continuación, asignándoles un peso diferente según su gravedad.

Tabla 4.2: Matriz de ponderación para el error cometido entre el estado real y la clasificación realizada.

		Estado real			
		Single talk remoto	Silencio	*Single talk* local	*Double talk*
Clasificación	*Single talk* remoto	–	3	6	1
	Silencio	3	–	3	1
	Single talk local	6	3	–	1
	Double talk	1	1	1	–

- Error crítico: en casos de *single talk*, decisión errónea del interlocutor que se encuentra activo; en otras palabras, el VAD clasifica *single talk* remoto como *single talk* local o viceversa. Este error en la clasificación es el peor que puede cometerse, ya que toda la ganancia que debe insertarse en los estados de *single talk* – en nuestro caso -6 dB – se destina al canal incorrecto, atenuando al interlocutor activo es ese momento. A este error se le asigna el peso de mayor valor de 6.
- Error medio: clasificación errónea de silencio como *single talk* o viceversa. El caso en el que se detecte *single talk* en lugar de silencio puede ser irrelevante, pero no es así, ya que este silencio puede ser ruidoso, y en este caso se detecta ruido como si fuera voz. En el caso contrario en que se detecta silencio en lugar de *single talk*, se introduce parte de la atenuación en cada canal aunque debería atenuarse con el valor máximo sólo el canal inactivo. Sin embargo, este error no es tan grave como el anterior en el que se realiza la acción contraria a la deseada, y por ello se le asigna el peso 3.

- Error leve: clasificación errónea de los casos de *single talk* o silencio como *double talk* y viceversa. Esta situación no se considera tan crítica como las anteriores pues se detecta al menos un usuario activo correctamente. Es por ello que se le asigna el peso con el mínimo valor 1.

Resultados obtenidos

A continuación se presentan, de distinta manera, tanto en la tabla 4.3 como en la figura 4.2 los ratios de error para diferentes *setups* del experimento.

Tabla 4.3: Ratios de error para el experimento con secuencias de ruido blanco como señales remota y local.

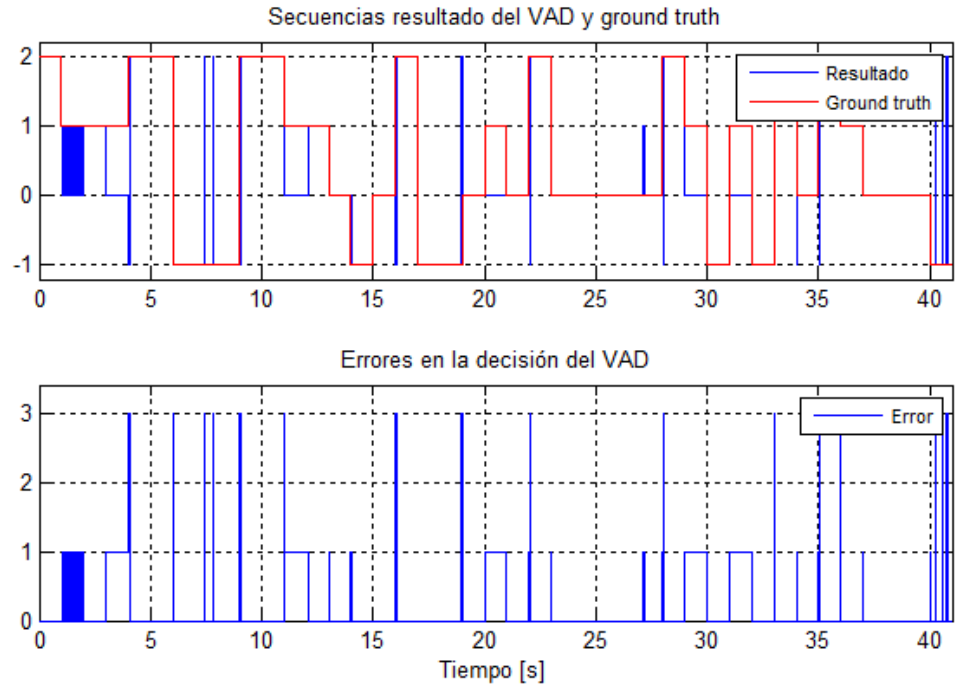
$SER[dB]$	%errores	%error$_1$[a]	%error$_3$[b]	%error$_6$[c]
0 dB	19.11%	96.01%	3.99%	0%
5 dB	16.21%	95.63%	4.37%	0%
10 dB	14.73%	95.42%	4.58%	0%
20 dB	6.05%	90.52%	9.48%	0%
30 dB	1.58%	62.57%	37.42%	0%
∞ dB	1.58%	62.57%	37.42%	0%

[a] Error leve
[b] Error medio
[c] Error crítico

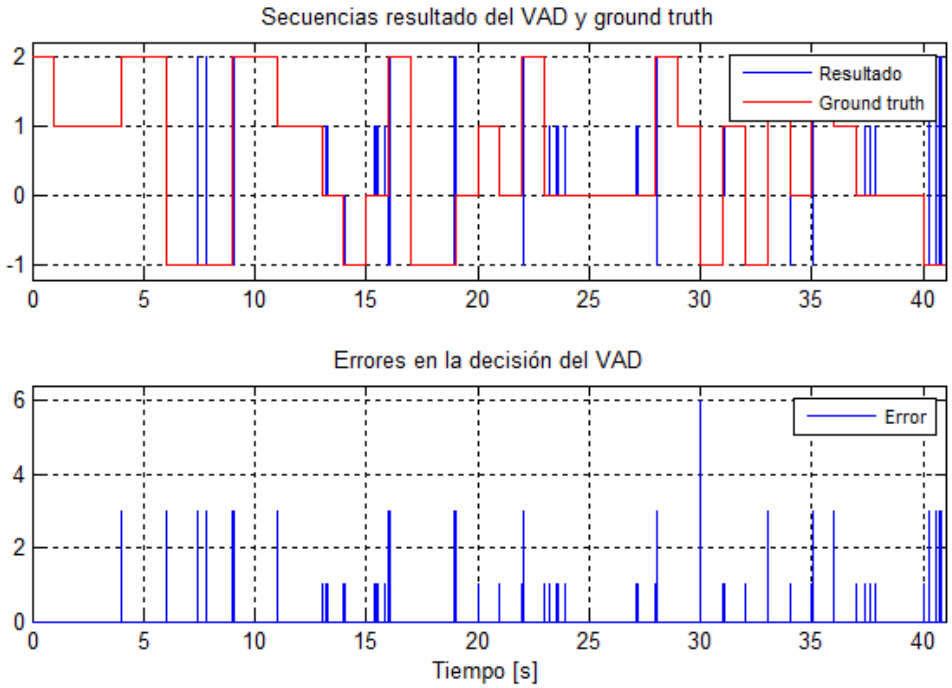
En la figura 4.2 se representan los vectores de *ground truth* y decisión del VAD, correspondientes a los estados reales y calculados para cada momento de la conversación[2]; así como el vector de errores con los pesos explicados anteriormente.

Como es de esperar, los resultados experimentales muestran que para valores más elevados de SER – donde la señal de eco presenta niveles más pequeños comparados con el habla del interlocutor local – el ratio de errores es menor, pues la componente de eco no afecta de manera tan negativa a la comunicación y puede distinguirse con más precisión de la actividad local, llevando así a realizar una mejor detección. Si nos fijamos en la figura 4.2 en el número absoluto de errores de cada tipo, vemos que con el aumento del SER, los $errores_3$ se mantienen, mientras que el número de $error_1$ disminuye. El $error_3$ es un error típico que se comete debido al retardo a la hora de seguir la actividad de voz en el principio y fin de palabras o frases. Por otro lado, en situaciones de *single talk* remoto, para un mayor valor de SER, el nivel de eco presente en la señal del micrófono es menor; por ello es más difícil que éste se confunda con voz real y que estas situaciones de *single talk* se clasifiquen incorrectamente como *double talk*, disminuyendo consecuentemente el número de $errores_1$. Además, la presencia de errores críticos es imperceptible, siendo un resultado positivo para el VAD.

[2] 2: silencio, 1: *single talk* remoto, 0: *double talk*, -1: *single talk* local



a)



b)

Figura 4.2: Para dos diferentes *setups* del experimento, secuencias que representan el *ground truth* de los estados de la conversación y la decisión de dichos estados realizada por el VAD, así como el consecuente error cometido. En la figura a) valor de $SER = 10\,\text{dB}$ y en b) valor de $SER = 30\,\text{dB}$.

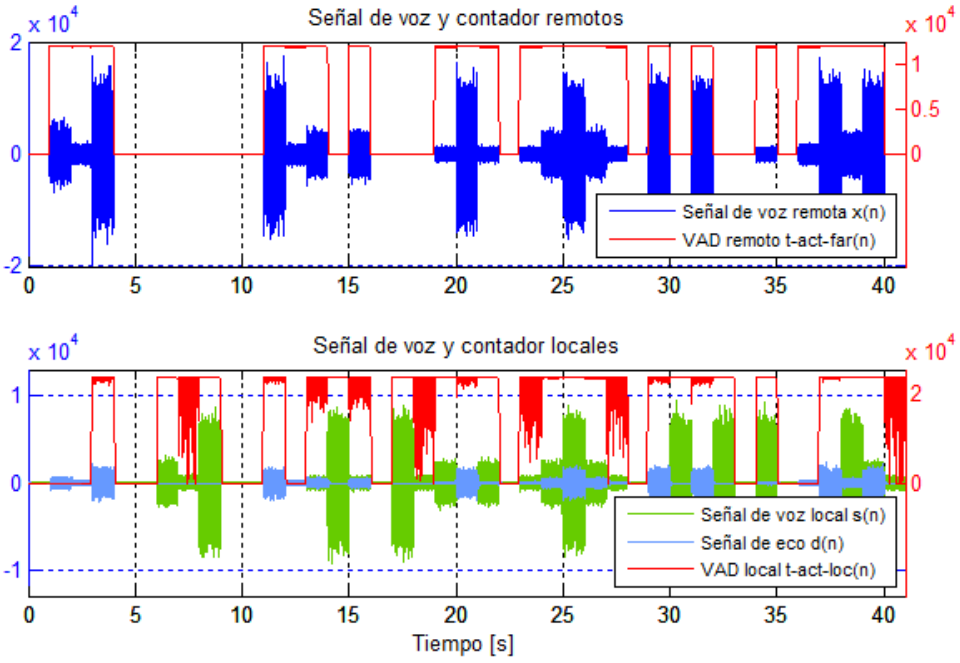


Figura 4.3: Detección de actividad representada por los contadores remoto y local para las secuencias generadas de ruido blanco y con un valor de $SER = 10\,\mathrm{dB}$.

En general, observando la tabla 4.3 se puede afirmar que el ratio de errores no es exageradamente elevado; sin embargo, la figura 4.3 representa el resultado de la detección de actividad de voz para el *setup* de $SER = 10\,\mathrm{dB}$ y en ella se puede observar que hay varios momentos en los que se produce una falsa detección aunque el valor de SER no es extremadamente bajo. Estas falsas detecciones ocurren en los casos en los que el nivel del eco – por ejemplo en el segundo 20 – es al menos tan elevado como el nivel más bajo de la señal de voz local, es por eso de esperar que en estas situaciones, el VAD cometa fallos.

4.2 Evaluación de *double talk*

Tal y como se describe en la recomendación ITU-T P.110 [ITU09], un sistema de manos libres puede clasificarse de acuerdo con su comportamiento con respecto a las situaciones de *double talk*. Las diferentes clases se definen en la recomendación ITU-T P.340 [ITU00a] y se resumen en la figura 4.4 junto con las características que el sistema debe cumplir para pertenecer a cada una de ellas.

A la vez que alcanzar los 45 dB de atenuación de la señal de eco en el caso de *single talk* y 30 dB en el caso de *single talk* que la recomendación ITU-T G.167 [ITU93] establece, se presenta también como objetivo de este proyecto que el sistema manos libres alcance una capacidad de comunicación *full duplex*. De esta manera, el sistema pasaría a pertenecer a la mejor categoría tal y como se especifica en 4.4. Como ahí se recoge, para alcanzar esta meta, el sistema debe una atenuación total igual o menor a 3 dB durante los periodos

Category (according to ITU-T P.340)	1	2a	2b	2c	3
	Full duplex capability	Partial duplex capability			No duplex capability
att_dt [dB]	≤3	≤6	≤9	≤12	>12

Figura 4.4: Clasificación de sistemas de manos libres de acuerdo con su comportamiento y atenuación introducida durante periodos de *double talk*.

de *double talk*. Esta cantidad de decibelios también permitiría en principio alcanzar los requisitos ya comentados en ITU-T G-167 [ITU93]. Para poder clasificar el sistema según la atenuación insertada en *double talk*, se ha implementado un método de evaluación el cual analiza su comportamiento en los periodos de *double talk*. Esta evaluación se realiza principalmente con dos tipos de señales de entrada diferentes:

- Evaluación de *double talk* para 'señal de fuente compuesta' – *Composite Source Signals* (CSS) –, tal y como se describe en la recomendación ITU-T P.501 [ITU12b] utilizando el análisis estandarizado descrito en la recomendación ITU-T P.502 [ITU00b].
- Evaluación de *double talk* para señales de voz real tal y como se describe en [3GP12].

Ambos métodos se presentan con más detalle en las siguientes secciones, donde se realiza una descripción de cada uno de ellos así como la presentación de los resultados obtenidos tras su aplicación. [3GP12]

4.2.1 Algoritmo para el análisis de *double talk*

El procedimiento implementado para analizar el comportamiento del sistema manos libres en las situaciones de *double talk* se realiza según el Apéndice II de la recomendación ITU-T P.502 [ITU00b] y se utilizará tanto para el análisis con *Composite Source Signals* como con voz real, siendo necesarias para ésta última unas pequeñas modificaciones que se recogen en [3GP12]. Dicho procedimiento permite obtener resultados que son independientes tanto de la percepción de la persona que realiza la evaluación como de las condiciones del laboratorio, que en muchas situaciones pueden influenciar el resultado. Mediante el método propuesto se pretende reducir la posible ambigüedad presente en las interpretaciones finales.

En este algoritmo de evaluación, el sistema de manos libres debe ejecutarse dos veces. En la primera sólo hay actividad presente en el extremo local, *single talk*, y la salida del sistema – aproximación de la señal de voz local, $\hat{s}(n)$ – se guarda como señal de referencia. Para la segunda ejecución, se añade también actividad en el extremo remoto, lo que llevará a tener situaciones de *double talk* y de la misma manera que en el caso anterior, se guarda la salida del sistema. Con estas dos señales que llamaremos de referencia y de *double talk*, han de realizarse los siguientes pasos:

- Tanto para la señal de referencia, como la de *double talk*, cálculo del nivel de potencia respecto al tiempo. Esta tarea se realiza a través del filtrado IIR de ambas señales,

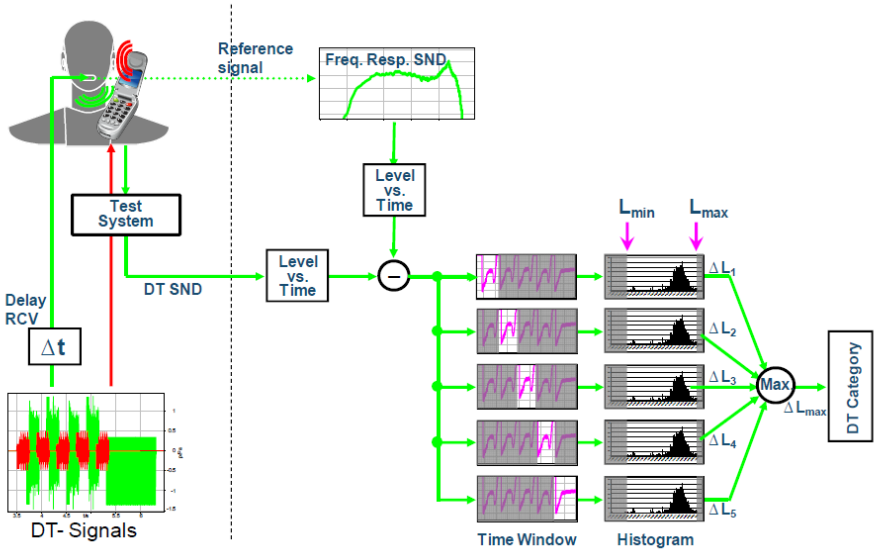


Figura 4.5: Esquema del algoritmo encargado de evaluar el comportamiento del sistema manos libres en situaciones de *double talk*.

teniendo el filtro una constante de tiempo adecuada, que en este caso se corresponde con 5 ms.

- Cálculo de la diferencia entre ambos niveles. Se obtendrá así una señal con valor nulo en los periodos de *single talk* – pues existe la misma información presente en ambas señales guardadas – y con el valor de atenuación introducida en los periodos de *double talk*.

- Creación del histograma de la atenuación introducida para cada periodo de *double talk*. Los límites del histograma se corresponden con los valores máximo y mínimo de atenuación presentes en cada periodo. El intervalo entre este máximo y mínimo debe dividirse en 100 valores equiespaciados siendo estos 100 valores resultantes aquellos para los que se crea el histograma. Así se obtiene cuantas veces se aplica cada uno de estos valores de atenuación.

- 'Suavizado' del histograma mediante la eliminación del 20% de los valores superiores y el 15% de los inferiores. Con esto, se suprimen los picos de atenuación de elevado valor, pero irrelevantes por su corta duración que pueden estar presentes.

- La atenuación introducida en cada periodo se calcula como aquella que presenta el valor más elevado de apariciones en el histograma correspondiente.

- El valor máximo de todas estas atenuaciones individuales será el valor de la atenuación global para *double talk*.

Conociendo la atenuación global, se puede clasificar directamente el sistema manos libres en la categoría correspondiente de acuerdo con la tabla 4.4. Una representación gráfica del algoritmo propuesto, puede verse en la figura 4.5.

4.2.2 Evaluación del comportamiento en *double talk* para *Composite Source Signals*

Primeramente, se explican los fundamentos de las señales CS para posteriormente presentar y discutir los resultados que se han obtenido al aplicar el algoritmo de evaluación explicado en la sección 4.2.1.

Composite Source Signals

Composite Source Signals presentan características similares a las de voz real y se utilizan como señales de test estándares en algunas aplicaciones, como en AECs, tal y como se describe en la recomendación ITU-T G.168 G.168 [ITU12a]. Estas señales se dividen en las siguientes componentes:

- Sonido de voz: para simular las propiedades de la voz y que se genera mediante la señal de voz artificial de la recomendación ITU-T P.50 [ITU99]. Normalmente, los sistemas que se utilizan para la transmisión del habla reaccionan positiva y rápidamente en la presencia de sonidos de voz, es por eso que este componente tiene como finalidad activar los posibles detectores de habla presentes en el sistema.

- Señal determinística o señal de pseudo-ruido: presenta características de una señal de ruido, con magnitud constante en el dominio de la frecuencia, mientras que su fase cambia. La señal suele generarse, produciendo primero un espectro complejo en el dominio de la frecuencia, para ser transformado al dominio del tiempo mediante la transformada inversa de Fourier, tal y como se describe en ITU-T P.501 [ITU12b].

- Pausa: la pausa puede tener dos objetivos, el primero es poner el sistema en un estado inicial definido y el segundo consiste en poner el sistema en un estado constantemente activado (habla constante).

La secuencia de *single talk* está completamente definida en ITU-T P.501 para frecuencias hasta 20 kHz, caso de voz en banda ancha, y presenta las siguientes duraciones: 48.62 ms para el sonido de voz, 200 ms para la señal de pseudo-ruido, y finalmente 101.38 ms para la pausa. Esta señal se representa en la figura 4.6.

Tal y como se expone en ITU-T P.501, la secuencia necesaria para generar situaciones de *double talk* debe crearse de la misma manera que la de *single talk* y únicamente se diferenciarán en las duraciones de cada componente, que se modifican para obtener situaciones más realistas de *double talk*. Sin embargo, también se afirma que para cierto tipo de aplicaciones, puede ser recomendable generar una secuencia de *double talk* con las mimas duraciones que en la señal de *single talk*, siendo éstas en este caso: 48.62 ms para el sonido de voz, la longitud del pseudo-ruido ha de mantenerse en 200 ms y la pausa para ambas señales CS ha de prolongarse hasta tener 151.38 ms. Para el experimento que aquí nos concierne, han de generarse dos señales, una para *single talk* y otra para tener condiciones de *double talk*. Éstas se han de generar tal y como se representan en la recomendación ITU-T P.502 [ITU00b] y en [ITU12c], donde ambas señales presentan las mismas características en relación a la duración de cada una de sus componentes. Es por ello que las señales van a generarse siguiendo estas recomendaciones utilizando las voces artificiales proporcionadas por la recomendación ITU-T P.50 [ITU99] y adaptándolas con el programa *Adobe Audition*,

para al final obtener las señales que cumplen con los requisitos de la recomendación, tanto de duración, como de potencia. Estas señales pueden verse representadas en la figura 4.7.

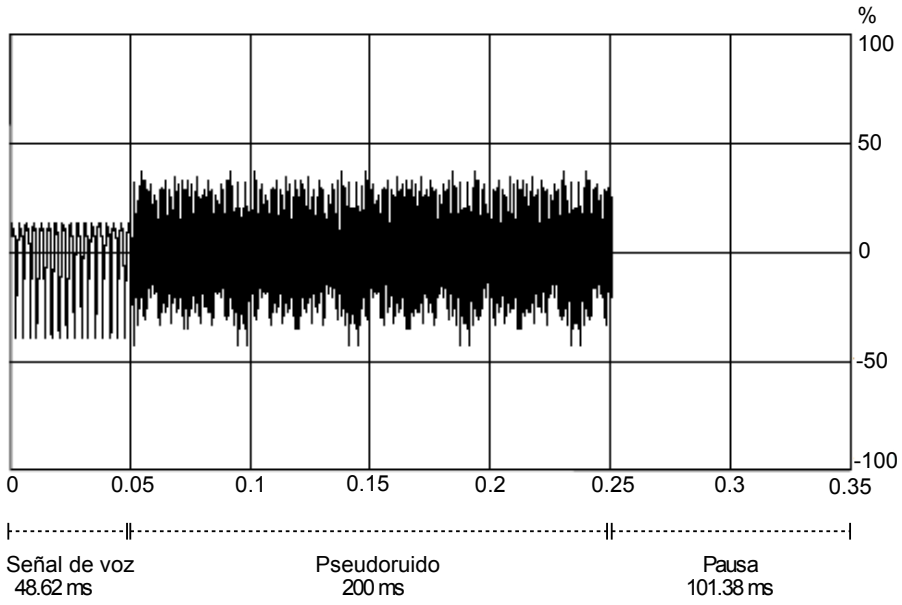


Figura 4.6: Composite Source Signal.

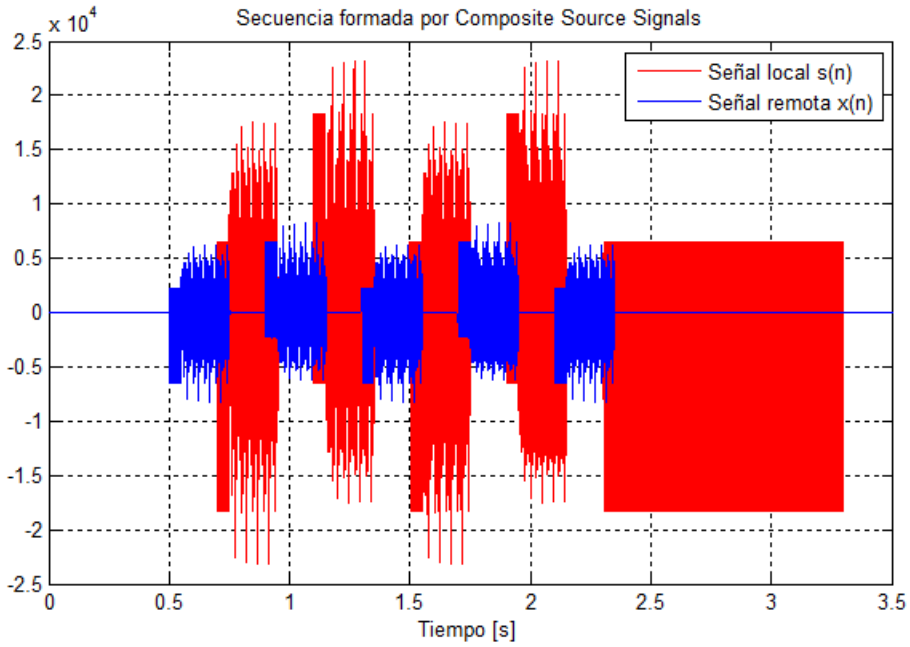


Figura 4.7: *Composite Source Signals* generadas respectivamente para simular la actividad del usuario remoto y local.

Resultados de la evaluación de *double talk* con CSS

En esta sección, se presentan los resultados de la evaluación de *double talk* obtenidos para diferentes configuraciones de los parámetros del sistema manos libres. Estas evaluaciones se han llevado a cabo mediante el método de análisis descrito en la sección 4.2.1. A su vez, se ha evaluado la efectividad del VAD, utilizando el método de clasificación de errores explicado en la sección 4.1.2.

En la tabla 4.4 que se muestra a continuación se han recogido los resultados para diferentes setups evaluados. Dicha tabla se puede subdividir en dos partes, que se diferencian en la atenuación que el control de ganancia inserta en los casos de *double talk*; en la parte superior, ésta es de $att_{dt} = 3\,\text{dB}$ y en la inferior de $att_{dt} = 0\,\text{dB}$. La razón de realizar estos experimentos son las de comprobar si el bloque de control de ganancia inserta la atenuación deseada durante los periodos de *double talk*.

Analizando los resultados experimentales en la tabla, la primera apreciación que debe realizarse es que para todos los setups del algoritmo evaluados, la atenuación insertada durante los periodos de *double talk* es $att < 3\,\text{dB}$ cumpliendo así el requisito necesario que le permite pertenecer a la mejor categoría de acuerdo con ITU-T P.340 [ITU00a]. Este resultado puede parecer obvio, ya que para *double talk* el control de ganancia inserta 3 dB de atenuación; sin embargo, si el VAD no fuera suficientemente robusto y realizara muchas falsas detecciones, el valor de atenuación global que se obtiene podría ser más elevado, resultando así en una degradación de la calidad del control de ganancia. Así pues, el resultado final depende en gran medida de los bloques de detección de voz local y remoto.

Con respecto a las tasas de error se pueden realizar las mismas afirmaciones que las ya comentadas en la sección 4.1.1 y además, como es de esperar, para valores más pequeños de SNR aumenta el porcentaje de detecciones falsas, ya que el ruido tiene suficientemente potencia como para confundirse con voz real. Aun así, se puede afirmar que para ninguno de los experimentos llevados a cabo esta tasa de errores es especialmente elevada y que los errores que se cometen mayoritariamente son los leves y en ningún caso los graves.

Observando ahora con más detenimiento la parte inferior de la tabla – donde el control de ganancia no inserta atenuación en las situaciones de *double talk* –, vemos que se obtienen los resultados esperados, pues aunque la atenuación global no alcanza los 0 dB, se tienen valores muy cercanos. Estos dBs de más se deben a las detecciones falsas de voz en los casos de *double talk* como *single talk*, ya que para este último si que se introduce atenuación en el canal. Entre estos valores y los de la parte superior de la tabla existe una diferencia de entre 1 y 2 dB, lo que nos ayuda a afirmar que el control de ganancia funciona correctamente. Por su parte, las tasas de error son las mismas o muy similares a las de la parte superior, pues la atenuación insertada por el bloque de control de ganancia no influye en la detección de los estados de la conversación, y menos aún para unos valores tan pequeños como con los que se está trabajando en este proyecto.

Así pues, ya que no existe mucha diferencia entre los valores de atenuación y que el sistema pertenece a la mejor categoría, podemos decir que para el experimento con *Composite Source Signals*, tenemos un sistema robusto que cumple con los objetivos deseados.

Tabla 4.4: Resultados del análisis *double talk* y tasas de error obtenidas para diferentes setups con CSS como entrada del sistema.

att_{dt} [a]	att_{st} [b]	SNR	SER	att [c]	%error	$\%error_1$	$\%error_3$	$\%error_6$
3 dB	6 dB	∞ dB	∞ dB	1.41 dB	2.08%	86.90%	13.13%	0%
3 dB	6 dB	∞ dB	10 dB	1.41 dB	2.25%	87.93%	12.07%	0%
3 dB	6 dB	∞ dB	5 dB	1.40 dB	5.28%	94.85%	5.15%	0%
3 dB	6 dB	∞ dB	0 dB	1.41 dB	9.41%	97.11%	2.89%	0%
3 dB	6 dB	10 dB	∞ dB	2.24 dB	2.10%	87.08%	12.92%	0%
3 dB	6 dB	10 dB	10 dB	2.25 dB	2.18%	87.52%	12.48%	0%
3 dB	6 dB	10 dB	5 dB	2.38 dB	5.42%	94.99%	5.01%	0%
3 dB	6 dB	10 dB	0 dB	1.45 dB	9.80%	97.23%	2.77%	0%
3 dB	6 dB	5 dB	∞ dB	2.51 dB	3.70%	52.80%	47.20%	0%
3 dB	6 dB	5 dB	10 dB	2.17 dB	5.27%	66.84%	33.16%	0%
3 dB	6 dB	5 dB	5 dB	2.43 dB	7.88%	77.82%	22.18%	0%
3 dB	6 dB	5 dB	0 dB	2.38 dB	11.33%	84.58%	15.42%	0%
3 dB	6 dB	0 dB	∞ dB	2.67 dB	15.32%	26.55%	73.45%	0%
3 dB	6 dB	0 dB	10 dB	2.64 dB	16.45%	31.57%	68.43%	0%
3 dB	6 dB	0 dB	5 dB	2.57 dB	18.29%	38.47%	61.53%	0%
3 dB	6 dB	0 dB	0 dB	2.55 dB	21.09%	46.64%	53.36%	0%
0 dB	6 dB	∞ dB	∞ dB	0.53 dB	2.08%	86.91%	13.09%	0%
0 dB	6 dB	∞ dB	10 dB	0.52 dB	2.26%	87.94%	12.06%	0%
0 dB	6 dB	∞ dB	5 dB	0.52 dB	5.28%	94.85%	5.15%	0%
0 dB	6 dB	∞ dB	0 dB	0.72 dB	9.83%	97.23%	2.77%	0%
0 dB	6 dB	10 dB	∞ dB	0.37 dB	2.10%	87.09%	12.91%	0%
0 dB	6 dB	10 dB	10 dB	0.35 dB	2.24%	87.86%	12.14%	0%
0 dB	6 dB	10 dB	5 dB	0.56 dB	6.84%	96.03%	3.97%	0%
0 dB	6 dB	10 dB	0 dB	0.59 dB	9.90%	97.26%	2.74%	0%
0 dB	6 dB	5 dB	∞ dB	0.31 dB	3.70%	52.81%	47.19%	0%
0 dB	6 dB	5 dB	10 dB	0.21 dB	5.74%	69.58%	30.42%	0%
0 dB	6 dB	5 dB	5 dB	0.51 dB	8.82%	80.20%	19.80%	0%
0 dB	6 dB	5 dB	0 dB	0.50 dB	11.37%	84.64%	15.36%	0%
0 dB	6 dB	0 dB	∞ dB	0.30 dB	15.32%	26.56%	73.44%	0%
0 dB	6 dB	0 dB	10 dB	0.32 dB	16.46%	31.64%	68.36%	0%
0 dB	6 dB	0 dB	5 dB	0.51 dB	20.04%	43.84%	56.16%	0%
0 dB	6 dB	0 dB	0 dB	0.54 dB	21.11%	46.70%	53.30%	0%

[a] att-dt: atenuación máxima insertada por el control de ganancia en los casos de *double talk*

[b] att-st: atenuación máxima insertada por el control de ganancia en los casos de *single talk*

[c] att: atenuación resultante obtenida del análisis *double talk* descrito en la sección 4.2.1

4.2.3 Evaluación del comportamiento en *double talk* para señales de voz real

En esta evaluación se va a abordar un problema más realista, ya que se realiza con el objetivo de analizar el comportamiento del sistema en situaciones de *double talk* para señales de voz real como entrada al sistema. Las señales necesarias para llevar a cabo este experimento se obtienen de la recomendación ITU-T P.501 [ITU12b]. Las señales remota y local utilizadas están representadas en la figura 4.8. Para realizar simulaciones, ITU-T recomienda el uso de voz real en lugar de CSS, ya que los sistemas son suficientemente inteligentes como para poder adaptarse a estas últimas y así proporcionar mejores resultados. El método de evaluación sigue el mismo principio que el explicado en la sección 4.2.1, sin embargo, ya que ahora se usan señales de voz es necesario realizar algunos cambios a la hora de generar los histogramas. Estos cambios son los siguientes:

- Modificación de la constante de tiempo a 30 ms.
- La atenuación global introducida en periodos de *double talk* se calcula como la mediana de todas las atenuaciones individuales.

El resto del algoritmo se mantiene igual que para el caso de las CSS, teniéndose así que ejecutar el sistema dos veces.

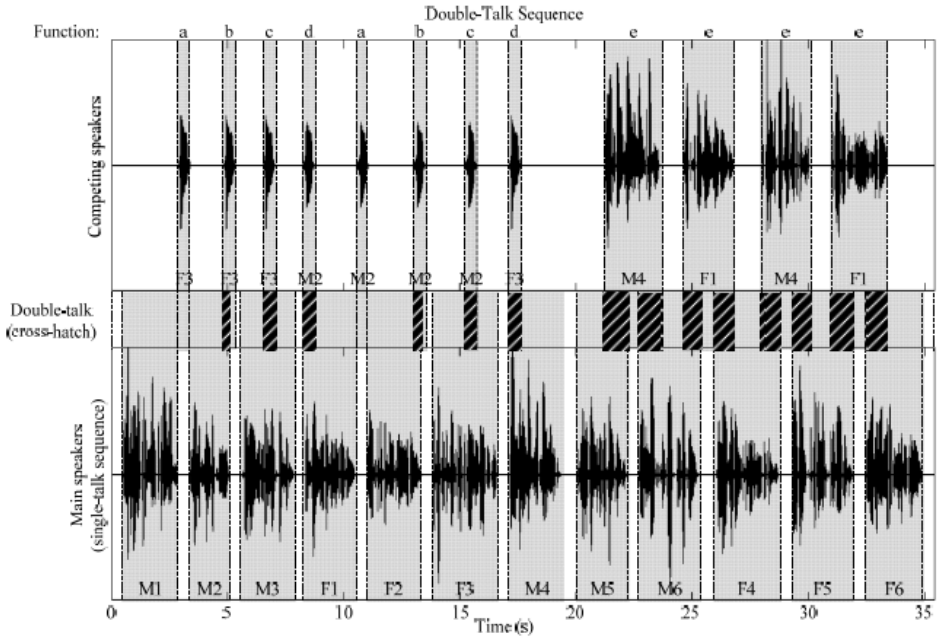


Figura 4.8: Secuencias para la evaluación de *double talk* con voz real femenina 'F' y masculina 'M', mostrando diferentes situaciones de double talk etiquetadas de la 'a' a la 'e' [ITU12b].

Resultados de la evaluación de *double talk* con real speech

Los resultados de esta evaluación se recogen en la tabla 4.5, resultados que son similares a los obtenidos para la evaluación con *Composite Source Signals*. Como puede apreciarse, se ha realizado para los mismos setups que en el caso de señales CS para así poder determinar que el algoritmo funciona como se espera. Ya que los resultados obtenidos son similares a los del caso anterior, presentando además las mismas variaciones, no van a analizarse en profundidad ya que los comentarios son los mismos que los realizados en la sección 4.2.2.

La diferencia más remarcable de este experimento con el realizado para *Composite Source Signals* es que el detector de actividad de voz realiza más falsas detecciones en el caso de señales de voz real; esto se puede apreciar claramente en las tasas de error. Dicho resultado se debe a que, aunque las CSS aproximan la voz real, ésta última presenta muchas más variaciones de amplitud a lo largo del tiempo, incluso pueden presentar pequeños silencios en la mitad de una palabra que pueden hacer que se realice una mayor cantidad de falsas detecciones. Aun así, para el caso más desfavorable – que se supone improbable – con $SNR = 0\,\text{dB}$ y $SER = 0\,\text{dB}$, el ratio de errores apenas alcanza en 20%. El resto de apreciaciones son las mismas que las ya realizadas en la sección 4.2.2 y es por eso que no nos vamos a detener en ellas. En este caso, el sistema se clasifica igualmente en la mejor categoría de acuerdo con ITU-T P.501 [ITU12b] con respecto a su comportamiento para *double talk* según los resultados obtenidos.
Ya que estas señales de habla real simulan una conversación realista, con ellas se va a realizar un análisis más exhaustivo llevando a cabo diferentes experimentos, teniendo en cuenta otros bloques y señales presentes en el sistema manos libres. Sin embargo, dichos experimentos no se van a realizan para tantas configuraciones del sistema como se ha hecho anteriormente, ya que el análisis en profundidad del control de ganancia ya ha sido realizado.

Influencia de la secuencia de ruido Hasta ahora se ha analizado el comportamiento que presenta el bloque de control de ganancia en situaciones de *double talk*. En este apartado se continúa con el mismo experimento pero con una señal diferente de ruido de fondo, para de esta manera, comprobar si el sistema es robusto bajo otras condiciones. Se eliminan los evaluaciones para valores extremos de SER, pero no para el ruido, pues es el componente que en este caso se trata de analizar. Los valores de atenuación son los establecidos de 6 dB en casos de *single talk* y 3 dB en casos de *double talk*, para los que se obtienen unos resultados que se recogen en la tabla 4.6.

Aquí han de compararse estos resultados con los correspondientes de la tabla 4.5 y, como puede apreciarse, la nueva secuencia de ruido no afecta a dichos resultados. Esto es positivo ya que significa que el comportamiento y resultados del sistema manos libres no se ven afectados por la señal de ruido de fondo presente en el extremo local. Como en los casos anteriores, el sistema vuelve a clasificarse en la mejor categoría. Por otro lado, los errores realizados en la detección de actividad de voz se ven únicamente incrementados en no más de un 2% por lo que también podemos afirmar la robustez del VAD respecto a diferentes secuencias de ruido.

Tabla 4.5: Resultados del análisis *double talk* y tasas de error obtenidas para diferentes setups con señales de voz real como entrada del sistema.

att_{dt} [a]	att_{st} [b]	SNR	SER	att [c]	%error	$\%error_1$	$\%error_3$	$\%error_6$
3 dB	6 dB	∞ dB	∞ dB	2.67 dB	10.72%	42.45%	55.33%	2.22%
3 dB	6 dB	∞ dB	10 dB	2.92 dB	10.61%	48.21%	49.88%	1.91%
3 dB	6 dB	∞ dB	5 dB	2.79 dB	14.26%	50.86%	47.77%	1.70%
3 dB	6 dB	∞ dB	0 dB	2.47 dB	16.71%	57.89%	41.02%	1.09%
3 dB	6 dB	10 dB	∞ dB	2.67 dB	11.02%	42.23%	56.50%	1.20%
3 dB	6 dB	10 dB	10 dB	2.53 dB	11.22%	42.93%	55.81%	1.26%
3 dB	6 dB	10 dB	5 dB	1.90 dB	15.21%	55.43%	43.64%	0.93%
3 dB	6 dB	10 dB	0 dB	1.92 dB	17.60%	58.71%	40.87%	0.42%
3 dB	6 dB	5 dB	∞ dB	2.39 dB	11.55%	50.17%	48.61%	1.22%
3 dB	6 dB	5 dB	10 dB	2.07 dB	12.54%	54.06%	44.81%	1.13%
3 dB	6 dB	5 dB	5 dB	1.76 dB	16.90%	63.48%	36.06%	0.44%
3 dB	6 dB	5 dB	0 dB	1.37 dB	17.79%	60.67%	38.91%	0.42%
3 dB	6 dB	0 dB	∞ dB	2.61 dB	20.05%	54.88%	44.74%	0.30%
3 dB	6 dB	0 dB	10 dB	2.32 dB	20.27%	55.37%	44.26%	0.37%
3 dB	6 dB	0 dB	5 dB	2.32 dB	20.27%	55.37%	44.26%	0.37%
3 dB	6 dB	0 dB	0 dB	2.37 dB	20.27%	55.37%	44.26%	0.37%
0 dB	6 dB	∞ dB	∞ dB	0.12 dB	10.16%	41.92%	55.84%	2.24%
0 dB	6 dB	∞ dB	10 dB	0.15 dB	13.69%	48.53%	49.58%	1.90%
0 dB	6 dB	∞ dB	5 dB	0.37 dB	16.68%	55.22%	43.63%	1.15%
0 dB	6 dB	∞ dB	0 dB	0.37 dB	17.81%	57.00%	41.92%	1.08%
0 dB	6 dB	10 dB	∞ dB	0.76 dB	10.98%	41.67%	57.04%	1.29%
0 dB	6 dB	10 dB	10 dB	0.65 dB	12.36%	48.19%	50.67%	1.14%
0 dB	6 dB	10 dB	5 dB	1.24 dB	16.47%	56.76%	42.79%	0.45%
0 dB	6 dB	10 dB	0 dB	1.17 dB	17.40%	58.52%	41.07%	0.41%
0 dB	6 dB	5 dB	∞ dB	0.29 dB	11.19%	48.52%	50.21%	1.26%
0 dB	6 dB	5 dB	10 dB	0.45 dB	15.83%	61.03%	38.08%	0.89%
0 dB	6 dB	5 dB	5 dB	0.50 dB	17.43%	61.32%	38.26%	0.42%
0 dB	6 dB	5 dB	0 dB	0.66 dB	17.61%	60.70%	38.88%	0.42%
0 dB	6 dB	0 dB	∞ dB	0.05 dB	19.79%	54.29%	45.33%	0.38%
0 dB	6 dB	0 dB	10 dB	0.15 dB	20.16%	55.13%	44.49%	0.37%
0 dB	6 dB	0 dB	5 dB	0.15 dB	20.16%	55.13%	44.49%	0.37%
0 dB	6 dB	0 dB	0 dB	$[0.15]dB$	20.16%	55.13%	44.49%	0.37%

[a] att-dt: atenuación máxima insertada por el control de ganancia en los casos de *double talk*

[b] att-st: atenuación máxima insertada por el control de ganancia en los casos de *single talk*

[c] att: atenuación resultante obtenida del análisis *double talk* descrito en la sección 4.2.1

Tabla 4.6: Resultados del análisis *double talk* y tasas de error obtenidas para diferente señal de ruido. Voz real como entrada del sistema y atenuación de 3dB para periodos de double talk.

att_{dt} [a]	att_{st} [b]	SNR	SER	att [c]	%error	$\%\text{error}_1$	$\%\text{error}_3$	$\%\text{error}_6$
3 dB	6 dB	∞ dB	10 dB	2.92 dB	13.61%	48.21%	49.88%	1.91%
3 dB	6 dB	∞ dB	5 dB	2.76 dB	15.26%	50.86%	47.77%	1.70%
3 dB	6 dB	10 dB	10 dB	2.26 dB	11.92%	41.18%	57.63%	1.19%
3 dB	6 dB	10 dB	5 dB	1.67 dB	16.15%	54.78%	44.76%	0.46%
3 dB	6 dB	5 dB	10 dB	2.04 dB	15.08%	54.72%	44.46%	0.82%
3 dB	6 dB	5 dB	5 dB	1.76 dB	18.21%	57.21%	42.38%	0.41%
3 dB	6 dB	0 dB	10 dB	2.61 dB	19.35%	59.00%	40.61%	0.39%
3 dB	6 dB	0 dB	5 dB	2.59 dB	19.48%	57.61%	42.01%	0.39%

[a] att-dt: atenuación máxima insertada por el control de ganancia en los casos de *double talk*

[b] att-st: atenuación máxima insertada por el control de ganancia en los casos de *single talk*

[c] att: atenuación resultante obtenida del análisis *double talk* descrito en la sección 4.2.1

4.3 Evaluación de la atenuación de la señal de eco

Como ya se ha recogido en la sección 1.1, el objetivo principal de este proyecto consiste en cumplir con los requisitos exigidos por la recomendación ITU-T G.167 [ITU93] para cancelación de eco y que de esta manera, se consiga que el sistema manos libres desarrollado en el 'Institut für Nachrichtentechnik' de la 'TU Braunschweig' llegue a cumplir todos los estándares internacionales. Estos requisitos son los siguientes:

- Atenuación de 45 dB para la señal de eco cuando la comunicación presente situaciones de *single talk*.
- Atenuación de 30 dB para la señal de eco cuando la comunicación presente situaciones de *double talk*.

Debido a que el sistema ya existente presenta unas atenuaciones cercanas a las que aquí se refieren – especialmente para *double talk* –, la atenuación insertada por el control de ganancia debería ser suficiente para alcanzar los valores exigidos por la recomendación. Para comprobarlo, vamos a comparar para *single talk* y *double talk* la atenuación que sufre la señal de eco tanto sin haber implementado el control de ganancia como ya implementado. En estas simulaciones, el interlocutor remoto está continuamente activo tanto para *single talk* como para *double talk*, mientras que el interlocutor local sólo se encuentra activo intermitentemente en las simulaciones de *double talk*. Este experimento se ha realizado para diversas señales, recogiendo aquí a modo comparativo los resultados obtenidos con las señales de voz real utilizadas en experimentos comentados anteriormente. Además, las configuraciones del sistema son las mismas que para los demás experimentos exceptuando $SER = \infty$ dB ya que en esta situación la señal de eco tiene una amplitud nula.

Con objetivo de valorar la capacidad del sistema para suprimir la señal de eco se calcula la magnitud conocida como 'Echo Return Loss Enhancement' (ERLE), magnitud que sirve para determinar la mejora percibida de la perturbación que produce la señal de eco. El ERLE se define de manera recursiva y se estima de la siguiente manera:

$$\begin{aligned} ERLE(n) &= \frac{E\left\{d^2(n)\right\}}{E\left\{\left(d(n)-\hat{d}(n)\right)^2\right\}} \\ &\approx \frac{(1-\beta)d^2(n)+\beta\hat{d}^2(n-1)}{(1-\beta)\left(d(n)-\hat{d}(n)\right)^2+\beta\left(d(n-1)-\hat{d}(n-1)\right)^2} \end{aligned} \tag{4.1}$$

con un factor de suavizado de $\beta = 0.9996$; donde $d(n)$ es la señal de eco y $\hat{d}(n)$ la señal de eco después de ser procesada por todo el sistema. Como se ha dicho, esta magnitud se va a calcular con y sin la presencia del control de ganancia obteniéndose los resultados que se recogen en la siguiente tabla 4.7.

Tabla 4.7: Resultados de la evaluación para la atenuación de la señal de eco.

		Single talk		*Double talk*	
SNR	SER	sin c.g. [a]	con c.g. [b]	sin c.g.	con c.g.
∞ dB	10 dB	43.78 dB	48.40 dB	29.25 dB	31.69 dB
∞ dB	5 dB	43.69 dB	47.94 dB	28.29 dB	31.14 dB
∞ dB	0 dB	41.28 dB	45.45 dB	29.05 dB	31.14 dB
10 dB	10 dB	41.02 dB	45.26 dB	29.25 dB	31.91 dB
10 dB	5 dB	43.01 dB	48.51 dB	28.02 dB	30.11 dB
10 dB	0 dB	40.59 dB	45.54 dB	28.14 dB	31.35 dB
5 dB	10 dB	42.63 dB	47.03 dB	29.26 dB	31.83 dB
5 dB	5 dB	42.50 dB	47.66 dB	27.57 dB	30.14 dB
5 dB	0 dB	39.67 dB	44.56 dB	27.65 dB	30.30 dB
0 dB	10 dB	42.92 dB	47.50 dB	28.00 dB	30.54 dB
0 dB	5 dB	43.39 dB	48.94 dB	28.37 dB	30.72 dB
0 dB	0 dB	40.28 dB	45.21 dB	29.86 dB	31.96 dB

[a] sistema con el control de ganancia desactivado
[b] sistema con el control de ganancia activado

En la tabla anterior se recogen los valores de atenuación del eco obtenidos para los casos de *single talk* y *double talk* tanto con el bloque de control de ganancia activado como desactivado. Con esto se pretende evaluar si dicho bloque cumple con su principal objetivo y así se llegan a alcanzar 45 dB de atenuación en situaciones en las que un sólo interlocutor está activo y 30 dB cuando ambos lo están. Por los valores que se recogen en las columnas tercera y quinta – control de ganancia desactivado – se puede ver que éstos son cercanos al objetivo; sin embargo se desea superar esta pequeña diferencia para hacer

que el sistema cumpla con las recomendaciones. Es por ello que no ha sido necesario que el control de ganancia introduzca unos valores de atenuación elevados y aun así se cumplan los valores objetivo. Dichos valores vienen reflejados en las columnas cuarta y sexta en las que se recogen los resultados para el control de ganancia activado. Como se puede apreciar, los valores deseados se alcanzan y superan levemente. Si se quisieran superar con mayor amplitud, se podría hacer que el control de ganancia introdujera más atenuación en el canal *uplink*, pero esto podría perjudicar el comportamiento del sistema global en situaciones de *double talk*. Al introducir valores de atenuación más elevados, la señal de eco se vería más atenuada e igualmente la atenuación global durante los periodos de *double talk* sería mayor, afectando a los resultados que se recogen en las secciones 4.2.2 y 4.2.3. De esta manera, el parámetro att_{dt} sería más elevado y probablemente, en muchas ocasiones, mayor a 3 dB. En el caso de que esto ocurriese, el sistema ya no presentaría capacidad total *full duplex* y se clasificaría en otras categorías según la tabla 4.4. Así pues, se ha conseguido satisfacer con éxito el compromiso entre estos dos requisitos: la capacidad *full duplex* del sistema y los valores exigidos para la atenuación de la señal de eco. Es por esto que se puede concluir que la implementación del control de ganancia escogido ha sido satisfactoria.

4.4 Evaluación de *Perceptual Evaluation of Speech Quality*

Perceptual Evaluation of Speech Quality (PESQ) es un método matemático estandarizado utilizado para evaluar la calidad de las señales de voz. En este caso, PESQ compara la señal de voz limpia $s(n)$, con su versión aproximada $\hat{s}(n)$ obtenida de la señal de micrófono $y(n)$ al final de ser procesada por todo el canal de comunicación. Este método proporciona una estimación de la calidad percibida por un grupo de oyentes en un test de escucha que, por supuesto, es subjetivo. Para obtener dicha estimación deben calcularse dos parámetros cuya combinación resulta en un tercero llamado *Mean Opinion Scores*(MOS) con un valor comprendido 1.0 y 5.0, donde dichos valores representan la peor y mejor calidad respectivamente.

Debido a factores psíquicos, el mejor valor de PESQ se sitúa alrededor de 4.7 y un resultado cercano o mayor a 3.0 ya es indicador de buena calidad. El objetivo de este experimento es conocer si la implementación del control de ganancia en el sistema mejora la calidad de la señal transmitida al interlocutor remoto, $\hat{s}(n)$. Para ello, se calcula el PESQ tanto para la señal antes del control de ganancia como la de después y así poder analizar si dicho bloque mejora la calidad de la señal. Los dos valores obtenidos se comparan, y el resultado será positivo siempre y que obtengamos un mayor valor de PESQ para la señal procesada por el control de ganancia, pues esto indicaría que el bloque ha mejorado la calidad de la señal de voz.

Observando los resultados que se presentan en la tabla 4.8, puede afirmarse que el control de ganancia sí mejora la calidad de la señal de voz. Sin embargo, esta mejora es tan pequeña, que la voz se percibe de manera similar con o sin el control de ganancia, aun así, esto es positivo, pues la calidad no se degrada lo que se vería como un resultado peor. La mayoría de los valores obtenidos – excepto en los peores casos, con SNR bajo – son

inferiores pero cercanos a 3, por lo que se alcanza una calidad aceptable.

Tabla 4.8: Resultados de *Perceptual Evaluation of Speech Quality* (PESQ) evaluados para las señales antes y después de ser procesadas por el bloque de control de ganancia.

att_{dt} [a]	att_{st} [b]	SNR	SER	PESQ ŝ(n) [c]	PESQ ŝ'(n) [d]
3 dB	6 dB	∞ dB	∞ dB	3.1115	3.1028
3 dB	6 dB	∞ dB	10 dB	3.0935	3.0850
3 dB	6 dB	∞ dB	5 dB	3.0701	3.0637
3 dB	6 dB	∞ dB	0 dB	3.0202	3.0157
3 dB	6 dB	10 dB	∞ dB	2.7390	2.7361
3 dB	6 dB	10 dB	10 dB	2.7299	2.7282
3 dB	6 dB	10 dB	5 dB	2.7255	2.7248
3 dB	6 dB	10 dB	0 dB	2.6909	2.6905
3 dB	6 dB	5 dB	∞ dB	2.5031	2.4997
3 dB	6 dB	5 dB	10 dB	2.5008	2.4977
3 dB	6 dB	5 dB	5 dB	2.4953	2.4916
3 dB	6 dB	5 dB	0 dB	2.4779	2.4759
3 dB	6 dB	0 dB	∞ dB	2.2701	2.2692
3 dB	6 dB	0 dB	10 dB	2.2691	2.2665
3 dB	6 dB	0 dB	5 dB	2.2667	2.2630
3 dB	6 dB	0 dB	0 dB	2.2563	2.2513

[a] att-dt: atenuación máxima insertada por el control de ganancia en los casos de *double talk*

[b] att-st: atenuación máxima insertada por el control de ganancia en los casos de *single talk*

[c] ŝ(n): señal después del control de ganancia

[c] ŝ'(n): señal antes del control de ganancia

5 Conclusiones

El objetivo de este proyecto ha sido el de optimizar un sistema de manos libres ya existente de manera que así pueda cumplir los requisitos exigidos por las recomendaciones internacionales a la vez que mantenga una buena calidad durante la comunicación.

Detección de actividad de voz La primera evaluación que se ha realizado ha sido la del bloque detector de actividad de voz incluido dentro del control de ganancia. Aquí, se evalúa la precisión de este detector y se contabilizan las falsas detecciones que éste realiza. Se estudian tanto el porcentaje de errores totales, como los porcentajes individuales dependiendo de la gravedad del error cometido. Esta evaluación se ha llevado a cabo para distintos tipos de señales de entrada al sistema: ruido blanco, señales CS y voz real. Los resultados obtenidos de esta evaluación son positivos, puesto que el porcentaje total de errores no es elevado y en el caso más desfavorable, con mayor nivel de ruido y eco, apenas alcanza el 20%. Además, como ya se ha dicho, los errores se han dividido en distintas clases según su importancia. El ratio de cada uno de estos errores individuales se ha calculado con respecto al número total de errores cometidos y asumiendo que el sistema no va a funcionar perfectamente, el resultado será tanto mejor cuanto menos errores graves se produzcan. La valoración obtenida tras realizar este análisis también ha sido en cualquier caso positiva, ya que el porcentaje correspondiente al error se ha obtenido siempre cercano a cero. La situación ideal será aquella en la que no se produzca ningún fallo en la detección, pero ya que este hecho es prácticamente imposible de conseguir, es mejor que no se cometan errores graves. Por todo esto, tras realizar esta primera evaluación al VAD, se puede decir que este bloque funciona según las expectativas y que lleva a cabo una buena detección de voz, discriminando ésta tanto de componentes de eco como de ruido de fondo.

Comportamiento frente a *double talk* Esta evaluación ya se ha realizado únicamente para las situaciones más realistas en las que las señales de entrada al sistema son o bien señales CS o bien voz real. En este se ha llevado a cabo mediante un método internacionalmente estandarizado descrito en la recomendación ITU-T P.502 [ITU00b]. Dicho método calcula la atenuación global media que el sistema introduce en la señal del micrófono en situaciones de *double talk*. Dependiendo del valor obtenido se puede clasificar el sistema con respecto a su capacidad *full duplex* en estos casos de habla simultánea. A la mejor categoría pertenecen aquellos sistemas que introduzcan una atenuación en situaciones de *double talk* menor o igual a 3 dB. Es por ello que ya se decidió que el control de ganancia no aplicara valores elevados de atenuación.
Tras realizar varias evaluaciones para parámetros del algoritmo con diferentes valores, se ha observado que este primer requisito se cumple satisfactoriamente; y es por ello que el sistema manos libres se clasifica en la mejor categoría con respecto a su comportamiento en situaciones de *double talk* permitiendo en todo caso una comunicación *full duplex*. Además de esto, se ha comprobado también que elevados valores tanto de eco como de ruido de

fondo no afectan en gran medida al resultado y tampoco así a la calidad del sistema.

Cancelación de eco La atenuación que sufre la señal de eco después de procesarse ha sido, por supuesto, evaluada ya que este ha sido el principal objetivo a alcanzar durante este proyecto. A lo largo de este documento, ya se ha comentado la importancia de cumplir con los requerimientos que establecen las recomendaciones internacionales y uno de ellos es que los sistemas manos libres deben atenuar la señal de eco al menos 45 dB en casos de *single talk* y 30 dB en casos de *double talk*. Ya que el sistema original no los cumplía, es por eso que se decidió incluir un bloque de control de ganancia. Estos valores se han medido y comprobado calculando las pérdidas de las señal de eco en retorno, mediante la magnitud *ERLE*. Debido a que los valores originales ya eran cercanos a los valores objetivos – y para garantizar la capacidad *full duplex* –, se decidió que la atenuación introducida por el control de ganancia no fuera muy elevada.
Con ello se han alcanzado sin problemas los valores objetivo, cumpliendo además el compromiso que se presentaba entre atenuar suficientemente la señal de eco y garantizar una comunicación *full duplex*.

Calidad del habla Finalmente, se ha evaluado la calidad de la señal que es transmitida al usuario remoto mediante la magnitud *PESQ*, que ya ha sido previamente explicada. Esta evaluación simula un escenario de escucha objetivo dando al final un valor numérico que se corresponde con lo buena o mala que es la señal de voz. Este método se aplica para la señal antes y después de ser procesada por el control de ganancia, de esta manera vemos si este bloque influye en la calidad de la señal y, en el caso ideal, si la mejora.
Después de llevar a cabo este análisis, se aprecia en una pequeña mejora de la calidad de la señal, aunque es prácticamente imperceptible. Esto es porque la atenuación adicional que introduce el control de ganancia en la señal de eco es realmente muy pequeña, aunque suficiente para alcanzar los objetivos propios de este proyecto.

5.1 Líneas futuras

Si bien, como se ha comentado, los objetivos sobre los que partía este proyecto se han cumplido con el algoritmo implementado, existen varios procedimientos y mejoras para garantizar su perfecto funcionamiento.

El primero de ellos sería el de implementar el algoritmo en un procesador digital de la señal para así poder evaluarlo directamente en un coche real en un laboratorio y de esta manera poder realizar una evaluación mucho más realista. Para ello, la programación realizada ya se ha llevado a cabo teniendo en cuenta esta tarea futura respetando los valores que deben darse a la frecuencia de muestreo.

Por otro lado, se decidió este algoritmo por la facilidad de ser incorporado tanto en este como en otros sistemas manos libres, ya que no depende de señales concretas del sistema, como por ejemplo las proporcionadas por el AEC. Es posible la implementación de otros algoritmos, tanto en tiempo como en dominio frecuencial, prestando especial atención a estos últimos ya que en ellos, se puede extraer información del espectro de frecuencia de la

señal y realizar una detección de voz más precisa.

Por último, uno de los mayores problemas encontrados durante la realización de este proyecto ha sido la estimación de los niveles de ruido de fondo y acoplamiento cuyos valores son decisivos para llevar a cabo correctamente la detección de voz. Es por ello que esto puede ser un punto importante en una mejora futura, realizando una estimación más robusta de estos parámetros o incluso midiéndolos directamente en la cabina del vehículo.

A ANEXO I: Cancelación de eco acústico

En las comunicaciones que se realizan a través de un sistema manos libres, el camino del eco viene dado por un sistema *Loudspeaker-Enclosure-Microphone* (LEM). Esto significa que el micrófono está situado en el campo sonoro del altavoz y por ello están conectados acústicamente, tanto por un camino directo como por diversos caminos secundarios debidos a las reflexiones del sonido en los extremos del recinto. Este efecto ocurre en el interior de los vehículos, por lo que este acoplamiento afecta al correcto funcionamiento de los sistemas manos libres que en ellos se instalan. Cuando se da una situación en la que el interlocutor remoto está activo, la señal que se recibe y escucha a través del altavoz, $x(n)$, convoluciona con la respuesta impulsional de la habitación, $h(n)$. Como resultado de esta convolución, se genera la señal de eco, $d(n)$, que es capturada por el micrófono. De esta manera es como el interlocutor remoto, escucha un eco de su propia voz, lo que afecta negativamente a la comunicación. Este efecto negativo empeora cuando hay actividad del usuario local, haciendo que la calidad de la comunicación se reduzca, llegando ésta a ser molesta. En este caso, la señal capturada por el micrófono que es enviada al interlocutor remoto se expresa de la siguiente manera:

$$y(n) = x(n) * h(n) + s(n) = d(n) + s(n). \tag{A.1}$$

donde $(*)$ representa el operador de la convolución. Este proceso y las señales involucradas pueden verse en la figura A.1.

Debido a los efectos negativos del eco en las comunicaciones manos libres, éste debe ser eliminado o al menos reducido de la señal enviada al un remoto. Una solución sencilla para ello, sería permitir únicamente una comunicación half-duplex, en la que sólo un interlocutor puede hablar mientras que el otro canal de comunicación es bloqueado. Pero una comunicación half-duplex se encuentra muy lejos de la solución deseada, que es una comunicación full-duplex en la que ambos interlocutores pueden intervenir simultáneamente. Es por esta razón por la que se necesita la implementación de un cancelador de eco acústico (AEC), para cancelar la realimentación acústica entre el micrófono y el altavoz y mejorar la inteligibilidad de la conversación. Para alcanzar este objetivo, el AEC realiza una aproximación del camino del eco $h(n)$, obteniendo $\hat{h}(n)$. Con esta aproximación, puede estimarse una réplica de sa señal de eco, $\hat{d}(n)$.

A.1 Filtrado adaptativo

El interior de un vehículo no es un espacio estático, si no que en el se producen cambios debido a, por ejemplo, el movimiento de las personas presentes en él. Por estos cambios, su respuesta impulsional $h(n)$ cambia a lo largo del tiempo, no es estacionaria, y la estimada $\hat{h}(n)$ debería adaptarse a estos cambios, aproximándose en mayor medida a $h(n)$. Por tanto, para alcanzar este objetivo, en el que existe una dependencia del entorno y del estado del sistema, es necesario implementar un filtrado adaptativo [FB13]. Los coeficientes de

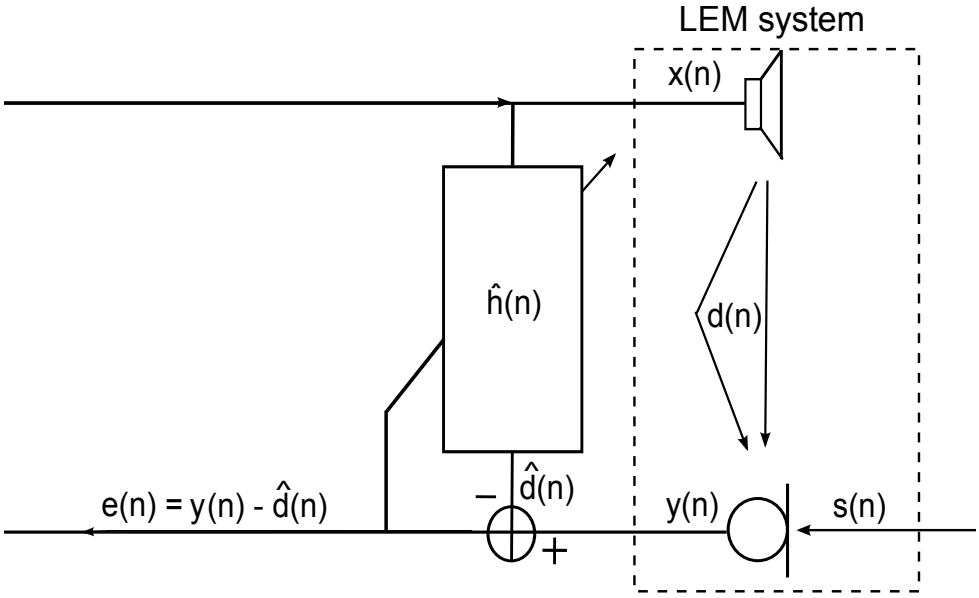


Figura A.1: Estructura de un sistema manos libres con AEC de filtrado adaptativo.

estos filtros no son valores constantes, si no adaptativos. Para calcular dichos valores, se implementa un algoritmo adaptativo, el cual se ejecuta hasta su convergencia; o lo que es lo mismo, la función error objetivo se minimiza. Esta función de error $e(n)$ permite determinar la forma en la que los coeficientes del filtro han de modificarse.

Como puede verse en la figura Figure A.1, la señal $x(n)$ proveniente del usuario remoto, se utiliza como entrada del AEC para obtener la predicción de la señal de eco $\hat{d}(n)$. La resta de esta señal de la captada por el micrófono, $y(n)$, resulta en la señal de error $e(n)$

$$e(n) = y(n) - x(n) * \hat{h}(n) = d(n) + s(n) - \hat{d}(n), \tag{A.2}$$

donde, en el resultado óptimo, es igual a la señal del interlocutor local $s(n)$. Sin embargo, normalmente, queda presente una señal de eco residual, $r(n)$, que se define:

$$r(n) = d(n) - \hat{d}(n). \tag{A.3}$$

Tras esta breve introducción, se presentan algún algoritmos de filtrado adaptativo. Debido a que este proyecto no está enfocado hacia el AEC, hay aspectos, como el desarrollo de algunas ecuaciones, que no van a explicarse en detalle.

A.1.1 Algoritmo 'Normalized Least Mean Square'

Los algoritmos *Least Mean Square* (LMS) son de los más utilizados para el filtrado adaptativo. Este algoritmo ajusta los coeficientes de acuerdo con la función de *Mean-Square Error* (MSE); los coeficientes definitivos serán aquellos que minimicen el error cuadrático medio. La velocidad de convergencia de este algoritmo depende del factor de convergencia o parámetro de paso, μ. La condición de convergencia de este algoritmo en error cuadrático

medio es:

$$0 < \mu < \frac{2}{\lambda_{max}} \tag{A.4}$$

donde λ_{max} es el mayor autovalor de la matriz de autocorrelación $\boldsymbol{R}$ de la señal de entrada. La velocidad de convergencia de este método depende pues de la dispersión de estos autovalores. Así, para valores muy dispersos, la velocidad será menor. De esta manera, el algoritmo presenta un buen comportamiento para ruido blando, mientras que se ralentiza para señales de entrada coloreadas, como señales de voz, las cuales presentan una alta correlación [SSS04].

Para resolver este problema de convergencia y acelerar el algoritmo LMS sin usar estimaciones de la matriz de autocorrelación de la señal de entrada, una posible solución es utilizar un factor de convergencia variable [Din13]. Al introducir esta modificación, se tendrá el llamado algoritmo LMS normalizado o *Normalized Least Mean Square* (NLMS). Este algoritmo converge normalmente más rápido que el LMS, ya que usa el factor de convergencia variable con el objetivo de minimizar el error instantáneo. Así, se puede utilizar la fórmula de actualización del algoritmo LMS de la manera siguiente:

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) + 2\mu_n e(n)\mathbf{x}(n) \tag{A.5}$$

donde μ_n debe elegirse adecuadamente para alcanzar una convergencia más rápida. Una posible estrategia es la de reducir el error cuadrático instantáneo lo máximo posible debido a que éste es un buen y simple estimador del MSE. Debido a que la explicación de los todos los cálculos no procede ser explicada en este proyecto, se le remite al lector que esté interesado en ellos a la sección 4.4 de [Din13]. Como conclusión, aquí se extrae que el valor del factor de convergencia que minimiza este error cuadrático se corresponde con:

$$\mu_n = \frac{1}{2\mathbf{x}^T(n)\mathbf{x}(n)} \tag{A.6}$$

Sustituyendo esta expresión en (A.5) e introduciendo un parámetro μ_k para controlar posibles desajustes y otro γ para evitar valores muy altos del tamaño de paso, se obtiene la ecuación de actualización de coeficientes

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) + \frac{\mu_k}{\gamma + \mathbf{x}^T(n)\mathbf{x}(n)} e(n)\mathbf{x}(n) \tag{A.7}$$

donde γ es una constante de valor muy pequeño y μ_k se elige en el rango $0 < \mu_k \leq 1$.

A.1.2 Algoritmo 'Affine Projection'

Hay situaciones en las que es posible reutilizar valores anteriores de la señal de entrada para mejorar la convergencia de los algoritmos de filtrado adaptativo. Estos algoritmos en los que se reutilizan datos son una alternativa para aumentar la velocidad de convergencia en situaciones donde la señal de entrada es altamente correlada. Por el contrario, esta reutilización de los datos implica un alto desajuste del algoritmo; la compensación entre el desajuste final y la velocidad de convergencia se consigue introduciendo un factor de convergencia.
Asumimos que conservamos los últimos $L+1$ vectores de entrada en una matriz:

$$\mathbf{X}_{ap}(n) = [\mathbf{x}(n), \mathbf{x}(n-1), ..., \mathbf{x}(n-L)], \tag{A.8}$$

Además de la señal de entrada, también se pueden definir los vectores que representan el resto de las señales involucradas en el proceso que son la salida del filtro adaptativo $\hat{\mathbf{d}}_{ap}(n)$, la señal del micrófono, $\mathbf{y}_{ap}(n)$ y el vector de error, $\mathbf{e}_{ap}(n)$. Este algoritmo intenta que el siguiente vector de coeficientes $\hat{\mathbf{h}}(n+1)$ sea lo más similar posible al actual $\hat{\mathbf{h}}(n)$ a la vez que fuerza al error a posteriori a ser cero.[1]
Como en el caso anterior, no se va a explicar el procedimiento completo de cálculo de las ecuaciones, que puede verse en detalle en la sección 4.6 de [Din13]. La ecuación que actualiza los coeficientes del filtro es la siguiente:

$$\hat{\mathbf{h}}(n+1) = \hat{\mathbf{h}}(n) + \mu \mathbf{X}_{ap}(n)[\mathbf{X}_{ap}^T(n)\mathbf{X}_{ap}(n)]^{-1}\mathbf{e}_{ap}(n). \tag{A.9}$$

donde μ es el factor de convergencia cuyo valor debe encontrarse dentro del rango $0 < \mu \leq 1$ y que realiza la compensación entre el desajuste final del algoritmo y la velocidad de convergencia, como se ha explicado antes.
La complejidad computacional de este algoritmo se debe al número de vectores de datos que se reutilizan, los que determinan el orden de la matriz que debe ser invertida. Para reducir este coste computacional se utilizan estrategias de selección de datos que también pueden verse en el Capítulo 6 de [Din13].

A.1.3 Algoritmo 'Recursive Least Squares'

A diferencia de los algoritmos previamente explicados, el objetivo de los algoritmos *least-squares* es el de minimizar la suma del error cuadrático. Si para cada iteración se reciben nuevas muestras de la señal de entrada, la solución puede calcularse de forma recursiva teniendo así algoritmos *Recursive Least Squares* (RLS). Estos algoritmos buscan alcanzar una velocidad de convergencia rápida incluso cuando la dispersión de los autovalores de la señal de entrada es muy grande. Estos algoritmos obtienen buenos resultados cuando se implementan en entornos variantes en el tiempo. Sus principales desventajas son el elevado coste computacional y problemas en la estabilidad del algoritmo.
Es necesario disponer de las muestras de la señal de entrada hasta el presente instante de tiempo para la minimización de la suma del error cuadrático. Este algoritmo suele implementarse con un filtro FIR de orden N. El vector que contiene la información de la señal de entrada para un instante n viene dado por:

$$x(n) = [x(n)x(n-1)...x(n-N)]^T. \tag{A.10}$$

La función a minimizar en el caso de los algoritmos *least-square* es de la forma:

$$\xi(n) = \sum_{i=0}^{n} \lambda^{n-1}\epsilon^2(i) = \sum_{i=0}^{n} \lambda^{n-1}[y(i) - \mathbf{x}^T(i)\hat{\mathbf{h}}(n)]^2 \tag{A.11}$$

donde $\hat{\mathbf{h}}(n) = [\hat{h}_0(n)\hat{h}_1(n)...\hat{h}_N(n)]^T$ es el vector de los coeficientes del filtro adaptativo y $\epsilon(i)$ es el error a posteriori en el instante i. El parámetro λ es un factor de ponderación que debe elegirse en el rango $0 << \lambda \leq 1$. Este parámetro hace que el efecto que la información de instantes de tiempo lejanos tiene sobre la actualización de coeficientes sea

[1] El error a posteriori es el error calculado con los datos disponibles hasta el momento actual (instante n) usando el vector de coeficientes ya actualizado $\hat{\mathbf{h}}(n+1)$.

cada vez menos influyente. Minimizando la función del error a posteriori, se obtiene que la expresión para los coeficientes óptimos del filtro adaptativo es:

$$\hat{\mathbf{h}}(n) = \mathbf{R}_D^{-1}(n)\mathbf{p}_D(n) \tag{A.12}$$

donde $\mathbf{R}_D(n)$ y $\mathbf{p}_D(n)$ son, respectivamente, la matriz de correlación de la señal de entrada y el vector de correlación cruzada entre la señal de entrada y la deseada, en nuestro caso, la señal $y(n)$ capturada por el micrófono. Se asume que la matriz $\mathbf{R}_D(n)$ es no singular, en caso de que no lo fuera, se utilizará una generalizada inversa. En el algoritmo RLS convencional se evitará realizar la operación de invertir la matriz de autocorrelación, utilizando el lema de inversión de matrices y reduciendo así el coste computacional. También hay una manera alternativa de reescribir el algoritmo RLS, siendo entonces la ecuación de actualización de pesos de la siguiente manera:

$$\hat{\mathbf{h}}(n) = \hat{\mathbf{h}}(n-1) + e(n)\mathbf{R}_D^{-1}(n)\mathbf{x}(n) \tag{A.13}$$

donde $e(n) = y(n) - \mathbf{x}^T(n)\hat{\mathbf{h}}(n-1)$. Debido a la gran cantidad de operaciones que involucra, el lector interesado puede observar tanto esta inversión matricial como el algoritmo desarrollado completamente en la sección 5.2 de [Din13]. Posteriormente también se discuten algunas propiedades correspondientes a este algoritmo.

A.1.4 Filtrado adaptativo en sub-bandas

Existen muchos casos en las que el orden requerido para los filtros adaptativos es alto; esto implica un alto coste computacional. Una de las soluciones a este problema es la implementación del filtrado adaptativo en sub-bandas. En este tipo de filtrado, tanto la señal de entrada $x(n)$, como la señal deseada $y(n)$ se dividen en sub-bandas utilizando bancos de filtro. Asumiendo que esta descomposición es efectiva, cada sub-banda puede ser diezmada y aplicar un filtrado adaptativo a cada una de ellas. El filtro para cada sub-banda tiene, normalmente, una respuesta impulsional más corta que su homólogo en banda completa. Además, en algunos algoritmos, el factor de convergencia puede adaptarse individualmente para cada sub-banda, lo que resulta en una velocidad de convergencia mayor que en los casos de filtrado en banda completa. Cada sub-banda puede diezmarse por un factor de diezmado r, menor o igual al número de sub-bandas, M, sin que afecte negativamente a la información de la señal original. A la hora de diezmar las sub-bandas, se pueden provocar efectos de aliasing, principalmente para casos de sub-muestreo crítico en los que $r = M$. Para los casos en los que $r > M$ se pierde información debido a los efectos de aliasing, lo que impide recuperar la señal original. La solución más obvia a estos problemas consiste en introducir huecos entre las diferentes sub-bandas, lo que sin embargo empeorará la calidad señal en banda completa.
Se han propuesto varias estructuras de filtrado adaptativo en las que se utiliza un submuestro crítico, $r = M$, como aquellas que usan bancos de filtros *Quadratic-mirror Filter* (QMF) o pseudo-QMF. Los problemas de aliasing pueden solucionarse utilizando términos cruzados que aumentarán tanto el coste computacional como la velocidad de convergencia. Una solución alternativa es utilizar sobremuestreo, esto es $r < M$. Sin embargo, en este caso, la complejidad computacional es más elevada de lo necesario, pues después de diezmar la señal, el número de muestras es mayor que aquellas a la entrada del banco de filtros.
En los filtros adaptativos en sub-bandas, es común que la señal de error se evalúe localmente

para cada sub-banda y que una función objetivo global se minimice teniendo en cueanta todos estos errores locales. Este tipo de estructura se conoce como *open-loop*, o de lazo abierto, en la que se intenta minimizar la energía de la señal de error en cada sub-banda. Para la estructura *open-loop* la función objetivo puede verse como una combinación lineal de cada error local, de la siguiente manera:

$$\xi = \sum_{i=0}^{M-1} E[|e_i(m)|^2] \tag{A.14}$$

donde $e_i(m)$ es la señal de error para cada sub-banda. Asumiendo que la matriz del filtro adaptativo es diagonal y que las señales de cada subbanda son complejas, la ecuación que actualiza los coeficientes, puede basarse en el algoritmo NLMS y viene dada por:

$$e_i(m) = \tilde{y}_i(m) - \hat{\mathbf{h}}_i^T(m)\mathbf{u}_i(m) \tag{A.15}$$

$$\hat{\mathbf{h}}_i(m+1) = \hat{\mathbf{h}}_i(m) + \frac{\mu}{\gamma + N_s\sigma_i^2(m)} e_i(m)\mathbf{u}_i^*(m) \tag{A.16}$$

donde N_s es la longitud de cada filtro adaptativo para la sub-banda i (se considera igual para todas las sub-bandas por simplicidad de notación). Por otro lado,

$$\sigma_i^2(m) = (1-\alpha)\sigma_i^2(m-1) + \alpha|u_i(m)|^2 \tag{A.17}$$

siendo α una constante cuyo valor debe encontrarse entre el intervalo $0 < \alpha \leq 0.1$ y γ otra constante pequeña para evitar que el paso sea muy grande. La señal $e_i(m)$ es la señal de error en la i-ésima sub-banda, y $\mathbf{u}_i(m)$ el vector de entrada al i-ésimo filtro adaptativo. Por último, el rango de valores del factor de convergencia es típicamente:

$$0 < \mu \leq 1 \tag{A.18}$$

Para conocer más en detalle el filtrado adaptativo en sub-bandas, se remite al lector a la sección 12.4 de [Din13]. A pesar de las ventajas que este sistema, el delay introducido durante todo el proceso es significativo; esto es debido a las estructuras de bancos de filtros tanto de análisis como de síntesis que deben ser implementadas. Por esta razón, se han desarrollado algoritmos de filtrado adaptativo sin retardo como puede verse en [SSBF08].

A.1.5 Frequency Domain Adaptive Filter

El filtrado en dominio frecuencial o *Frequency Domain Adaptive Filter* (FDAF), es otro de los métodos frecuentemente usados para la cancelación de eco. En este caso, tanto la adaptación de los coeficientes del filtro como la estimación de la señal de eco se realizan en domino frecuencial. FDAF soluciona dos problemas que presentan aquellos algoritmos implementados en dominio temporal: elevado coste computacional en caso de respuestas impulsionales largas y su mal funcionamiento frente a señales de entrada de elevada autocorrelación.

El primer problema se soluciona debido a que este tipo de filtrado, como el de subbandas, implementa un procesamiento en bloques de manera que al procesar un bloque de entrada se obtiene un bloque de datos de salida. Este procesamiento no requiere un

elevado coste computacional ya que se utiliza la multiplicación en lugar de convolución y otros algoritmos eficientes como la transformación rápida de Fourier, *Fast Fourier Transform* (FFT). De esta manera, los coeficientes se actualizarán sólo una vez por bloque de datos.
Las propiedades de decorrelación que presenta la transformada de Fourier solucionan el segundo problema. El factor de convergencia puede calcularse separadamente para cada slot frecuencial de manera que se asegura una convergencia uniforme de los coeficientes del filtro aunque la señal de entrada sea altamente correlada.

En este tipo de algoritmos se separa la señal de entrada en tramas a las que después se aplica una transformación discreta de Fourier, *Discrete Fourier Transformation* (DFT). Así se obtiene $X(l,k)$ a partir de $x(n)$ donde k se corresponde con el número de trama y l con el índice frecuencial. Los coeficientes $\hat{W}(l,k)$ dados por el algoritmo de adaptación se multiplicarán con dicha señal para obtener la estimación de la señal de eco en dominio frecuencial $\hat{D}(l,k)$.

$$\hat{D}(l,k) = \hat{W}(l,k)X^*(l,k), \tag{A.19}$$

para $k = 0, ..., K-1$, con K el número de puntos de la DFT. La transformada inversa de Fourier, *Inverse Fourier Transformation* (IDFT), permite calcular la señal de error qeu se utilizará en el bloque de adaptación

$$e(n) = y(n) - \hat{d}(n). \tag{A.20}$$

Diferentes algoritmos para la actualización de los coeficientes del filtro pueden encontrarse, por ejemplo, en [GMND10], [SG65].

A.2 Postfiltro

Como ya se ha comentado en el capítulo 2, es posible que después del bloque AEC todavía quede presente información de eco residual. Esto puede ser debido al limitado orden del filtro, cambios en el camino del eco que el algoritmo no es capaz de seguir, etc.

$$r(n) = d(n) - \hat{d}(n) \tag{A.21}$$

donde $r(n)$ es la señal de eco residual, $d(n)$ la señal de eco capturada por el micrófono y $\hat{d}(n)$ la señal de eco estimada por el AEC. La implementación de un postfiltro tiene normalmente como objetivo la eliminación o reducción de este eco residual, obteniendo a su salida la señal mejorada del micrófono, $\hat{s}(n)$ que debe ser lo más similar posible a la voz del interlocutor local, $s(n)$. El esquema de un sistema manos libres en el que se han implementado un AEC y un filtro para el eco residual puede verse en A.2.

Una forma simple y ampliamente utilizada para eliminar el eco residual es la implementación de un *gain loss control* (GLC). Estos algoritmos aplican atenuación en el enlace de subida, uplink. Aunque este valor de ganancia se calcula en función de la potencia de la señal que transmite el altavoz, también influye en los periodos de *double talk* pues se aplica independientemente de que haya actividad local o no. Si se realiza un postfiltrado en sub-bandas en lugar de implementar un GLC se soluciona el problema que este último

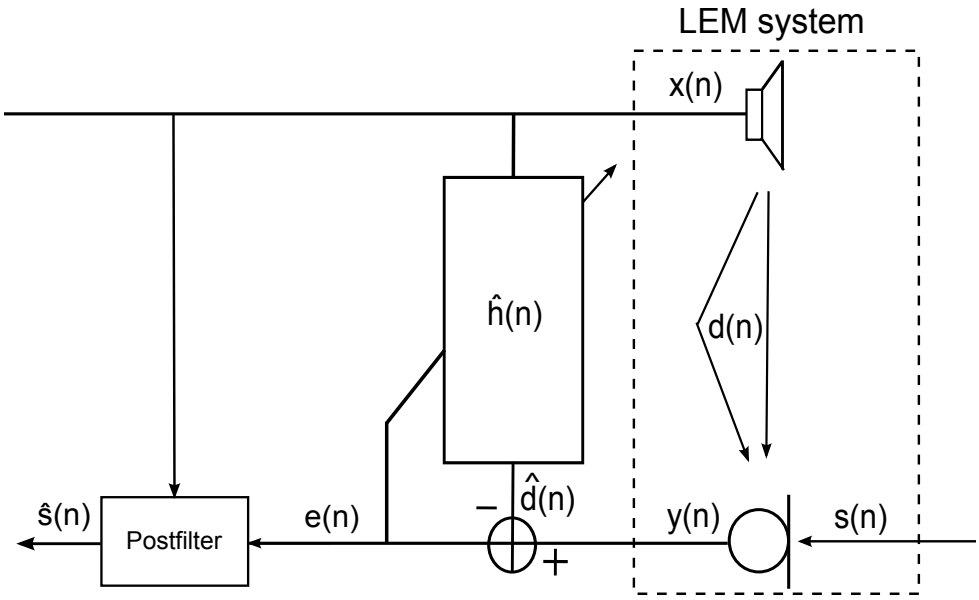


Figura A.2: Esquema de un sistema de manos libres formado por un AEC y un postfiltro.

tiene en los casos de *double talk*. Ya que este filtrado en sub-bandas se utiliza con frecuencia, se explica a continuación en mayor profundidad. Éstos se prefieren a los GLC ya que introducen diferentes valores de ganancia en cada sub-banda y de esta manera puede atenuar frecuencias en las que el eco residual es audible.

Tal y como puede verse en la figura A.3, tanto la señal de error $e(n)$ como la señal del altavoz $x(n)$, se dividen en sub-bandas, teniendo las señales $e_i(n)$ y $x_i(n)$ respectivamente, donde i denota el índice de la sub-banda, tomando valores entre 0 y $M-1$, siendo$M$ el número de sub-bandas en las que se divide la señal. En este caso, el análisis y síntesis de las señales se realiza a través de un banco de filtros modulado DFT. Una propiedad de este tipo de bancos de filtros es que cada filtro paso banda es un duplicado de un filtro paso bajo, $h_{pb}(n)$, desplazado en frecuencia. A este filtro paso bajo $h_{pb}(n)$ se le llama filtro prototipo.

Posteriormente, el filtrado del eco residual se puede hacer tanto mediante un filtro en sub-bandas como con un filtro en dominio temporal. El filtrado en sub-bandas tiene la ventaja de que posee una complejidad computacional baja, sin embargo introduce un delay significativo en la señal de salida. Dicho delay puede reducirse realizando el filtrado en el dominio temporal. En este caso, la ganancia de cada sub-banda se utiliza para determinar la respuesta impulsional de un filtro *Finite Impulse Response* (FIR). Las diferencias entre el filtrado en sub-bandas y en dominio temporal con un filtro FIR pueden observarse en la figura A.3.

Para cada sub-banda i, la ganancia del postfiltro se actualiza siguiendo el filtrado de

Wiener[YIEB10]:

$$G_i(n) = \frac{\xi_i(n)}{1 + \xi_i(n)}, \tag{A.22}$$

donde $G_i(n)$ es la ganancia que debe insertarse en cada sub-banda y $\xi_i(n)$ es el *Signal to Echo Ratio* (SER) para la voz del interlocutor local. En la práctica, el valor del SER es desconocido y consecuentemente, ha de realizarse una estimación. En este caso, en [YIEB10] esta estimación se realiza a través de la aproximación de Ephiram y Malah, en la que:

$$\xi_i(n) = \beta \frac{\hat{s}_i^2(n-1)}{\hat{\gamma}_i^{e_r e_r}(n-1)} + (1-\beta) max(\xi_i^{post}(n), 0), \tag{A.23}$$

donde β es una constante de suavizado cuyo valor se encuentra en el intervalo $0 < \beta < 1$, $\hat{s}_i(n)$ corresponde a la i-ésima sub-banda estimada de la voz local, $\hat{\gamma}_i^{e_r e_r}(n)$ es la densidad espectral correspondiente al eco residual y, por último, $\xi_i^{post}(n)$ es el SER a posteriori. La densidad espectral del eco residual, $\hat{\gamma}_i^{e_r e_r}(n)$, se debe estimar con la siguiente fórmula:

$$\hat{\gamma}_i^{e_r e_r}(n) = \frac{\gamma_i^{xe}(n)}{\gamma_i^{xx}(n)}, \tag{A.24}$$

donde $\gamma_i^{xe}(n)$ es la densidad espectral cruzada entre la señal remota, $x(n)$, y la señal de error, $e(n)$, y $\gamma_i^{xx}(n)$ la densidad espectral de potencia – *Power Spectral Density* (PSD) – de $x(n)$. De igual manera, el valor del SER a posteriori se calculará:

$$\xi_i^{post}(n) = \frac{e_i^2(n)}{\hat{\gamma}_i^{e_r e_r}(n)} - 1. \tag{A.25}$$

En todos los casos, las densidades espectrales $\gamma_i^{xx}(n)$ y $\gamma_i^{xe}(n)$ se estiman mediante un proceso autorregresivo.
Como ya se ha dicho, el eco residual puede filtrarse tanto en sub-bandas como en dominio temporal. Estos dos métodos se describen a continuación con mayor profundidad.

Filtrado en sub-bandas En el esquema a) de la figura A.3 se representa este filtrado en sub-bandas. Puede verse que tanto la señal de voz remota como la del error se dividen en sub-bandas a través de bancos de filtros de análisis. Posteriormente se aplica la ganancia $G_i(n)$ a la i-ésima sub-banda correspondiente de la señal de error $e_i(n)$ como un factor multiplicativo

$$\hat{s}_i(n) = G_i(n) e_i(n) \tag{A.26}$$

Mediante la implementación de un banco de filtros de síntesis que procesa las sub-bandas $\hat{s}_i(n)$, es posible recuperar la señal de voz mejorada en banda completa $\hat{s}(n)$. Todo el sistema completo introduce un retardo en la señal de $N-1$ muestras, siendo N la longitud del filtro prototipo paso bajo.

Filtrado en dominio temporal En la figura A.3(b) se representa el diagrama para un postfiltro en sub-bandas con filtrado FIR. Las señales en sub-bandas $x_i(n)$ y $e_i(n)$ se utilizan para calcular las ganancias $G_i(n)$ como ya se ha explicado anteriormente, éstas permiten determinar la respuesta impulsional del filtro FIR de acuerdo con distintos métodos. En este caso, el eco residual se elimina a través de la convolución de la señal $e(n)$ con

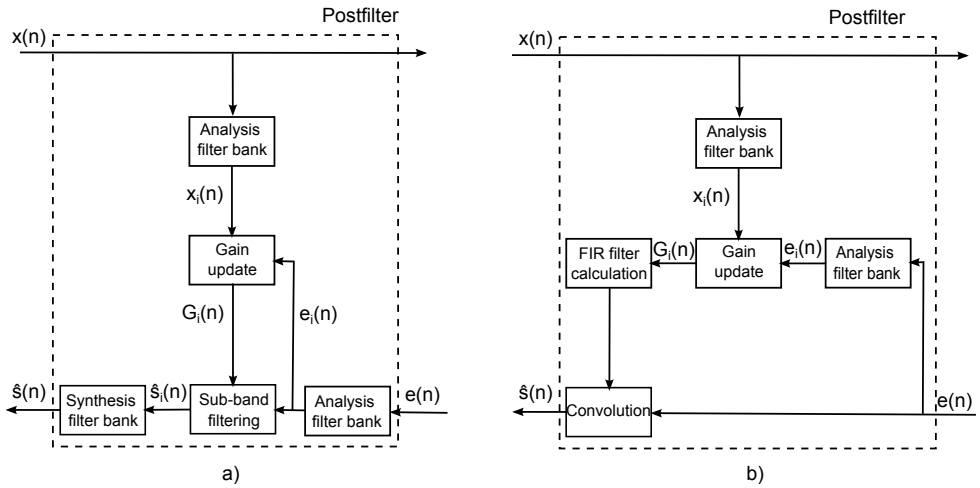


Figura A.3: Diagramas para: (a) filtro de reducción de eco en sub-bandas (b) filtro de reducción de eco en el dominio temporal mediante un filtro FIR.

el filtro FIR. Para evitar distorsiones de fase y asegurar un retardo constante en la señal, se deben implementar filtros de fase lineal. A continuación se presentan dos tipos de filtros FIR.

El *Filter Bank Equalizer* (FBE) fue desarrollado por Löllman and Vary in [LV07] y es el equivalente al filtrado en sub-bandas. El filtro FBE se calcula de la siguiente manera:

$$g_{fbe}(n) = h_{pb}(n)\tilde{g}(n) \tag{A.27}$$

donde $h_{pb}(n)$ es el filtro prototipo paso bajo del banco de análisis y $\tilde{g}(n)$ es la transformada inversa de Fourier – *Inverse Discrete Fourier Transformation* (IDFT) – de las ganancias en dominio frecuencial $G_i(n)$. Ya que éstas tienen un valor positivo – fase cero – el filtro $g_{fbe}(n)$ tendrá fase lineal si $h_{pb}(n)$ también la tiene. Por ejemplo, en el caso de que $h_{pb}(n)$ es un filtro simétrico.
Este proceso introduce únicamente un retardo en la señal de $\frac{N-1}{2}$. La mitad que en el caso de filtrado del eco en sub-bandas.

Aunque el FBE ya disminuye considerablemente el retardo introducido, éste todavía puede reducirse más aproximando el FBE por un filtro de menor grado [LV07].
El *Low Delay Filter* (LDF) se obtiene truncando el FBE con una ventana de longitud P tal que $P < N$. Esta ventana puede elegirse arbitrariamente o de manera que pueda mantener la propiedad de fase lineal. El retardo que introduce este sistema es de $\frac{P-1}{2}$ muestras.

Bibliografía

[3GP12] 3GPP. Technical Specification TS 126 132 - Universal Mobile Telecommunications System (UMTS); LTE; Speech and video telephony terminal acoustic test specification. 2012.

[BSV06] Cristophe Beaugeant, Martin Schönle, and Imre Varga. Challenges of 16 khz in acoustic pre and post-processing for terminals. 2006.

[Cio84] T. Cioffi, J. andKailath. *'Fast, Recursive-Least-Squares transversal filters for Adaptive Filtering' IEEE Transactions on acoustics, Speech and Signal Processing.* 1984.

[Din13] Paulo S. R. Diniz. *Adaptive Filtering.* Springer US, Boston, MA, 2013.

[EM83] Y. Ephraim and D. Malah. *'Speech Enhancement Using Optimal Non-Linear Spectral Amplitude Estimation' Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP).* 1983.

[Enz06] G. W. Enzner. *'A Model-Based Optimum Filtering Approach to Acoustic Echo Control: Theory and Practice.' PhD thesis, Rheinisch-Westfälische Technische Hochschule Aachen.* 2006.

[EV03] Gerald Enzner and Peter Vary. *Robust and Elegant, Purely Statistical Adaptation of Acoustic Echo Canceler and Postfilter: in International Workshop on Acoustic Echo and Noise Control.* 2003.

[EV06] G. W. Enzner and P. Vary. *'Frequency-Domain Adaptive Kalman Filter for Acoustic Echo Control in Hands-Free Telephones' Signal Processing, Elsevier.* 2006.

[FB13] B. Farhang-Boroujeny. *Adaptive filters: Theory and applications.* Second edition edition, 2013.

[GBSA13] Virginie Gilg, Christophe Beaugeant, Martin Schönle, and Bernt Andrassy. *Methodology for the Design of a Robust voice Activity Detector for Speech Enhancement: International Workshop on Acoustic Echo and Noise Control.* 2013.

[GMND10] K. G. Gunale, S. N. Motade, S. L. Nalbalwar, and S. B. Deosarkar. Frequency domain adaptive filter using fft algorithm for acoustic echo cancellation. pages 582–587. 2010.

[GT95] Steven L. Gay and Sanjeev Tavathia. The fast affine projection agorithm: - acoustics, speech, and signal processing, 1995. icassp-95., 1995 international conference on. 1995.

[Hei97] P. Heitkamper. An adaptation control for acoustic echo cancellers. *IEEE Signal Processing Letters*, 4(6):170–172, 1997.

[HS06] E. Hänsler and Gerhard Schmidt. *Acoustic Echo And Noise Control*. Adaptive and learning systems for signal processing, communications, and control. Wiley-Interscience, Hoboken, N.J., 2006.

[ITU93] ITU. ITU-T Recommendation G.167 (03/93) Acoustic echo cancelers. 1993.

[ITU99] ITU. ITU-T Recommendation P.50 (09/99) Artificial voices. 1999.

[ITU00a] ITU-T Study Group 12. Recommendation P.340 (05/2000) Transmission characteristics and speech quality parameters of hands-free terminals2. 2000.

[ITU00b] ITU-T Study Group 12. Recommendation. P.502 (05/2000) Objective test methods for speech communication systems using complex test signals. 2000.

[ITU09] ITU. Recommendation P.1110 (12/2009) Wideband hands-free communication in motor vehicles. 2009.

[ITU11] ITU. ITU-T Recommendation P.56 (12/11) Objective measurement of active speech level. 2011.

[ITU12a] ITU. Recommendation G.168(02/2012) Digital Network Echo Cancelers. 2012.

[ITU12b] ITU. Recommendation P.501 (01/2012) Test signals for use in telephonometry. 2012.

[ITU12c] ITU-T Study Group 12 - Contribution 188. *Double Talk Testing Using Different Double Talk testing Metholdologies*. 2012.

[JF14] Marc-André Jung and Tim Fingscheidt. A wideband automotive hands-free system for mobile hd voice services. In Gerhard Schmidt, Huseyin Abut, Kazuya Takeda, and John H.L Hansen, editors, *Smart Mobile In-Vehicle Systems*, pages 81–96. Springer New York, New York, NY, 2014.

[Kal60] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35, 1960.

[LDC04] Fredric Lindtrom, Mattias Dahl, and Ingver Claesson. *On Audio Hands-Free System Design*. 2004.

[LV07] Heinrich W. Löllmann and Peter Vary. Uniform and warped low delay filterbanks for speech enhancement. *Speech Communication*, 49(7-8):574–587, 2007.

[Mar93] Rainer Martin. An efficient algorithm to estomate the instantaneous snr of speech signals: 3rd european conference on speech communication and technology eurospeech'93. 1993.

[Say03] A.H. Sayed. *'Fundamentals of Adaptive Filtering' Wiley-IEEE Press, 1 ed.* 2003.

[SG65] John J. Shynk and Richard P. Gooch. Frequency-domain adaptive pole-zero filtering. 1965.

[Shy92] J.J. Shynk. *'Frequency-Domain and Multirate Adaptive Filtering' IEEE Signal Processing Magazine*. 1992.

[SSBF08] Kai Steinert, Martin Schönle, Christophe Beaugeant, and Tim Fingscheidt. Hands-free system with low-delay subband acoustic echo control and noise reduction. 2008.

[SSS04] H.-C. Shin, A. H. Sayed, and W.-J. Song. Variable step-size nlms and affine projection algorithms. *IEEE Signal Processing Letters*, 11(2):132–135, 2004.

[WS85] B. Widrow and P.N. Stearns. *'Adaptive Signal Processing' Prentice Hall 1 ed.* 1985.

[YIEB10] Christelle Yemdji, Moctar M. Idrissa, Nicholas W.D. Evans, and Christophe Beaugeant. Efficient low delay filtering for residual echo suppression: 18th european signal processing conference. 2010.