UNIVERSIDAD DE CHILE

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS

DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

ROBUSTEZ A EFECTOS DE CANAL EN VERIFICACIÓN DE LOCUTOR

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELECTRICISTA

MATÍAS JOSÉ TORRES RISSO

Profesor Guía:
Néstor Becerra Yoma

Miembros de la Comisión:
Claudio Garretón Vender
Fernando Huenupán Quinán

Santiago de Chile
Abril 2009

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL ELECTRICISTA
POR: MATÍAS TORRES RISSO
FECHA: 3/08/2009
PROF. GUÍA: SR. NÉSTOR BECERRA

"Robustez a efectos de canal en verificación de locutor"

Se denomina verificación de locutor a la tarea de aceptar o rechazar la identidad que un determinado usuario declara tener mediante la información extraída de su voz. Este tipo de aplicación toma especial interés cuando es operado sobre la red telefónica pues otorga una interfaz hombre-máquina de gran naturalidad para las personas. Sin embargo, para que un sistema de este tipo sea comercialmente atractivo, debe exhibir un rendimiento acorde a las exigencias de seguridad de la aplicación a ser implementada. Finalmente, la plataforma debe ser robusta a efectos indeseados como son el ruido y en especial el *mismatch* de canal.

El *mismatch* de canal es la condición a la cual es sometido un motor verificador de locutor donde las etapas de entrenamiento y de verificación son realizadas mediante canales de transmisión distintos, considerando el canal transmisor como la suma del aparato telefónico más el medio de comunicación. Dicha condición es uno de los factores que más degradan el desempeño de un sistema verificador de locutor, más aun si este es operado sobre la red telefónica.

Para otorgar robustez al motor verificador frente a la distorsión de canal, se realizaron experimentos con tres técnicas implementadas durante este proyecto. El primer método propuesto es una transformación de parámetros que actúa en el dominio del espectro de log-energía del banco de filtros Mel, capaz de reducir la tasa de error del sistema hasta en un 9 % al ser aplicado solo y en un 41,5 % al ser combinada con un procedimiento clásico tal como CMN (*Cepstral Mean Normalization*). La segunda propuesta consiste en un filtro pasa bandas aplicado en el espacio del espectro de las trayectorias temporales de log-energía del banco de filtros Mel, con el cual se logran reducciones en la tasa de error de 10,4 % y 5,5 % considerando y sin considerar *mismatch* de canal, respectivamente. Finalmente se plantea un método de normalización y compensación de *scores* basado en la selección automática del canal, procedimiento con el cual se logran mejoras del orden del 40 % en el error del motor verificador.

Agradecimientos

Este trabajo no habría sido posible sin el apoyo del profesor Néstor Becerra y todo el equipo que conforman el laboratorio de procesamiento y transmisión de voz. Para ellos mis más sinceros agradecimientos.

Mención especial merece mi hermano Tomás, quien colaboró de forma activa en los diagramas utilizados en este trabajo y Natalia, quien soportó con paciencia el tiempo invertido en esta memoria.

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

En el siglo XX y lo que corre del siglo XXI, la tecnología ha avanzado más rápido que en cualquier otro periodo conocido por el hombre. Gran responsable de aquello es la aparición del procesamiento masivo de datos y la computación, lo cual otorga rapidez y eficiencia en cualquier desarrollo que el ser humano quiera realizar. Un deseo de las personas es lograr imitar y mejorar las tareas realizadas por seres humanos, logrando con ello facilitar nuestras vidas. Es por ello el creciente desarrollo en áreas como robótica, control automático, etc. Un área de estudio que tomó mucha fuerza a principios del siglo pasado es el procesamiento de la voz, producto del creciente mercado de las telecomunicaciones, donde se intentó con respetable éxito sintetizar la voz humana, codificar y comprimir las señales de audio con el fin aprovechar mejor el ancho de banda escaso en ese entonces, entre muchos otros desarrollos. Junto con ello, aparece la biometría, disciplina que se encarga de reconocer a los usuarios de una determinada aplicación mediante una interacción biológica. Parte de esto es, por ejemplo, el reconocimiento mediante el rostro, el iris, las huellas digitales y, de particular interés para este trabajo, la voz.

El estudio de la voz, como disciplina enmarcada dentro de la biometría, toma mucha relevancia frente al creciente y cada vez más presente mercado de las comunicaciones móviles. Es en este contexto donde el reconocimiento de locutor aparece con especial fuerza. Este pretende determinar la identidad de un usuario mediante los atributos de su voz, la cual se asume como una característica única en cada ser humano. Dentro del reconocimiento de voz, existen dos familias de aplicaciones, por un lado están los llamados identificadores de locutor, los cuales consisten en un sistema capaz de identificar a un locutor dentro de un conjunto de usuarios afiliados a dicha aplicación, es decir, un clasificador de N clases.

Mientras que por otro están los verificadores de locutor, clasificadores binarios capaces de aceptar o rechazar la identidad que un determinado locutor dice tener. Esta última aplicación, inmersa dentro de lo que son las redes de telefonía, es lo que se ha desarrollado con mucha fuerza durante las últimas décadas, pues otorga a los usuarios comunes nuevas formas de interactuar con su aparato telefónico, entregándoles al mismo tiempo mayor naturalidad a un servicio que hasta hace poco no era más que presionar teclas. Así por ejemplo, un verificador de locutor operado sobre la red de telefonía, permite al usuario el acceso a información restringida o privada de forma segura y remota. Aplicaciones típicas de este tipo de tecnología son por ejemplo el acceso a operaciones remotas de un banco, acceso a bases de datos restringidas y muy importante también aplicaciones forenses

Sin embargo, y a pesar de los grandes avances logrados en esta área, aplicaciones de este tipo requieren por parte del usuario, una robustez altísima, puesto que malas decisiones por parte del sistema pueden desencadenar fraudes u otras consecuencias quizás menos graves pero igualmente indeseadas. Un factor que implica bajas en la robustez de este tipo de sistemas es el ruido convolucional o distorsión de canal. Esta es producto de la constante degradación a la que es expuesta una señal al ser transmitida por un canal de transmisión, a lo cual se suma el hecho de que en las distintas etapas que conforman un verificador de locutor pueden haber inconsistencias o *mismatch* de canal, por lo que es de suma importancia el lograr atenuar o cancelar dichos efectos negativos.

1.1. Motivación

El *mismatch* de canal es uno de los factores que más degradan el desempeño de un sistema verificador de locutor y más aun si este es operado sobre la red telefónica, pues constantemente se verá enfrentado a cambios en el canal transmisor. Este problema se puede solucionar enseñándole al motor verificador los distintos canales a los que se ve expuesto, lo cual se traduce en horas y horas de entrenamiento. Dicha solución es infactible por dos motivos. La primera razón es que constantemente salen al mercado nuevos aparatos telefónicos por lo que mantener al día el sistema sería necesario tener un constante entrenamiento, operación que claramente no se puede realizar por asuntos de tiempo y económicos. La segunda y más importante razón es que el locutor, usuario al que se pretende vender este sistema, no desea perder su tiempo pasando largas horas repitiendo frases y entrenando un sistema, que para Él es una caja negra, razón por la cual los datos de entrenamiento son escasos. Así, la principal motivación de este trabajo es aportar una

serie de técnicas capaces de darle robustez al sistema frente a los efectos degradantes del canal, de forma eficiente y con la limitada cantidad de datos que se dispone.

1.2. Objetivos

En este trabajo se plantea como principal objetivo el darle robustez a un sistema verificador de locutor texto dependiente operado sobre la red telefónica frente a los efectos adversos del canal. Para ello se desea plantear una serie de técnicas capaces de cancelar dicho efecto de distintas formas. En una primera instancia se plantea una transformación de los parámetros ingresados al sistema, tal que estos sean más robustos a los efectos del canal. De aquí se desprenden los siguientes objetivos específicos:

- Proponer un espacio capaz de representar al canal y la información de locutor de una forma tal que puedan ser separadas.
- Construir una transformada que le de robustez a los parámetros frente a la distorsión de canal.
- Establecer una metodología capaz de calcular el tipo de transformación más adecuada al sistema.
- Probar y comparar la transformada propuesta con técnicas convencionales utilizada en la literatura.

Luego de lo anterior, se plantea el desafío de complementar la técnica propuesta con el filtrado de trayectorias temporales. Para ello se establecen los siguientes objetivos:

- Determinar un filtro óptimo para darle sistema robustez al efecto del canal.
- Complementar el filtro calculado con la transformación antes propuesta.
- Probar y concluir respecto a los resultados obtenidos.

Finalmente se desea establecer una técnica capaz de normalizar y compensar los *scores* entregados por el motor verificador basado todo en una selección de canal mediante un clasificador adecuado. Para ello se definen los siguientes objetivos:

- Construir un selector de canal.
- Proponer una técnica de normalización y compensación de *scores*.
- Probar y concluir respecto de la técnica propuesta.

1.3. Estructura de la Memoria

En este documento se presentará de forma ordenada el trabajo realizado con el fin de atenuar los efectos degradantes que aporta la distorsión de canal a las señales ingresadas al sistema. Para ello se comienza en el capítulo dos dando un marco teórico capaz de introducir al lector en los temas relacionados con el ruido de canal y cómo esto afecta el desempeño de la verificación de locutor. En él además se explica qué es y cómo funciona la verificación de locutor, poniendo especial énfasis en el algoritmo de decodificación de Viterbi, técnica esencial en la modelación estocástica del proceso de verificación vía modelos ocultos de Markov. Luego se desarrolla la problemática del canal y cómo este problema es crítico en los sistemas abordados en este trabajo, presentando técnicas clásicas en la literatura que abordan dicha problemática.

En el capítulo tres se muestra una técnica denominada transformación de parámetros, la cual es aplicada en un espacio propuesto en este trabajo llamado espectro de la log-energía del banco de filtros Mel. Además se propone una nueva forma de realizar el análisis de importancia relativa (AIR), procedimiento utilizado con el fin de estimar una transformación óptima para el sistema. Con lo anterior se llevan a cabo una serie de experimentos sobre una versión telefónica de la base de datos YOHO, probando la técnica propuesta y mezclándola con el clásico procedimiento denominado *cepstral mean normalization* (CMN), para luego discutir y concluir respecto de los resultados obtenidos.

Posteriormente, se continúa en el capítulo cuatro mostrando el filtrado de trayectorias temporales, procedimiento diseñado con el fin de atenuar los efectos negativos del canal a través de los frames. Para ello se presenta el espacio llamado espectro de la trayectoria temporal del espectro de la log-energía del banco de filtros Mel, dominio en el cual actúa dicha propuesta. Luego se revisa el desempeño de lo mostrado y la mezcla con la técnica descrita en el capítulo anterior. Se concluye este capítulo con la normalización y compensación no supervisada de canal, técnica que combina las bondades de normalizaciones descritas en la literatura con la clasificación o detección de canal mediante mezcla de modelos Gaussianos.

Se finaliza este trabajo realizando en el capítulo cinco conclusiones obtenidas a partir de las propuestas hechas en los capítulos anteriores, además de plantear algunas tareas futuras que pueden ser llevadas a cabo a partir de los desarrollos hechos en este documento.

Capítulo 2

Verificación de Locutor y Robustez a Canal

2.1. Reseña Histórica

Mucho de lo que hoy se conoce en tecnologías basadas en el procesamiento de la voz comenzaron a desarrollarse a principios del siglo veinte. Alexander Graham Bell, conocido por patentar el teléfono, fue uno de los primeros investigadores que se planteó el desafío de lograr transcribir voz a texto sin que esto llegara a buen término, pero abriendo dicha área de investigación. Los grandes aportes comienzan a aparecer en la década de los sesenta bajo el alero de los laboratorios BELL, teniendo un reconocedor de dígitos capaz de realizar su tarea con un 2 % de error en condiciones dependientes de locutor y en ausencia de ruido.

En la década del ochenta se produce un cambio de paradigma en la clasificación de patrones utilizada en sistemas operados por voz. Se comienza a utilizar un modelamiento estadístico del proceso y se modela el habla como un fenómeno markoviano, reemplazando técnicas como las redes neuronales y el muy extendido hasta entonces *Dynamic Time Warping* (DTW) (L. Rabiner y Schmidt, 1980). La modelación asumiendo la voz como proceso de Markov es la estructura que se sigue utilizando en casi todos los sistemas actuales basados en el uso de la voz.

Dado el avance de los computadores, las capacidades de cómputo permiten tratar otros problemas no abordados con anterioridad como lo es el ruido presente en ambientes no controlados, dando a los motores de voz mayor robustez frente a condiciones reales. Es sin

duda este uno de los mayores desafíos que enfrentan este tipo de tecnologías.

En este capítulo se presenta un marco teórico en el cual se envuelve el tema de verificación de locutor texto dependiente. Se hace una modelación estadística del problema, se explican los parámetros a utilizar y las métricas usadas para evaluar el desempeño del sistema. Además se explica el problema del ruido y las diversas técnicas que abordan dicho tema en la actualidad.

2.2. Modelo de la Voz

Diversas características humanas son utilizadas en biometría, como ya se ha expuesto en partes anteriores, el iris, las huellas dactilares y el rostro son algunas de las posibles opciones que pueden ser utilizadas con el fin de lograr determinar la identidad de un usuario. La voz, el modo de comunicación más natural en los seres humanos, también puede ser utilizado para estos fines y gracias al crecimiento explosivo en el mercado de telefonía móvil, toma especial interés el trabajo en esta área. De esta forma, el estudio de el habla como interfaz hombre-máquina y su aplicación a sistemas biométricos es el principal centro de interés de este trabajo.

La producción de la voz está dominada fuertemente por órganos como el diafragma, cuerdas vocales, faringe, lengua, labios y las cavidades vocal y nasal. Para emitir palabras el proceso comienza con la contracción del diafragma produciendo de esta forma un flujo de aire por la garganta. Dicho flujo es el que transita a través de las cuerdas vocales, que producto de este tránsito comenzarán a vibrar a una frecuencia definida que se denomina f_0 o frecuencia fundamental. Gracias a estos dos procesos, por la faringe transita una excitación sonora que se aproxima muy bien a la forma de un tren de impulsos o a un ruido blanco, según el sonido a efectuar sea sonoro o sordo. Independiente del tipo de excitación, esta comienza a resonar y a producir armónicos a medida que pasa por el tracto vocal y nasal. Así, las características de un sonido u otro están fuertemente determinadas por la morfología que tenga en ese momento toda la cavidad resonadora (que incluirá faringe y cavidades nasal y vocal) y con ello los distintos armónicos que se le indujeron a la excitación. Por último, los labios también permiten caracterizar sonidos otorgándoles comportamientos explosivos, fricativos entre otros. En la figura 2.1 se muestran los distintos órganos que participan en la producción del habla junto con un esquema del transito de aire y la posterior producción de armónicos. Una segunda opción es la que ocurre cuando la excitación corresponde a un ruido blanco, la cual produce toda la gama de sonidos

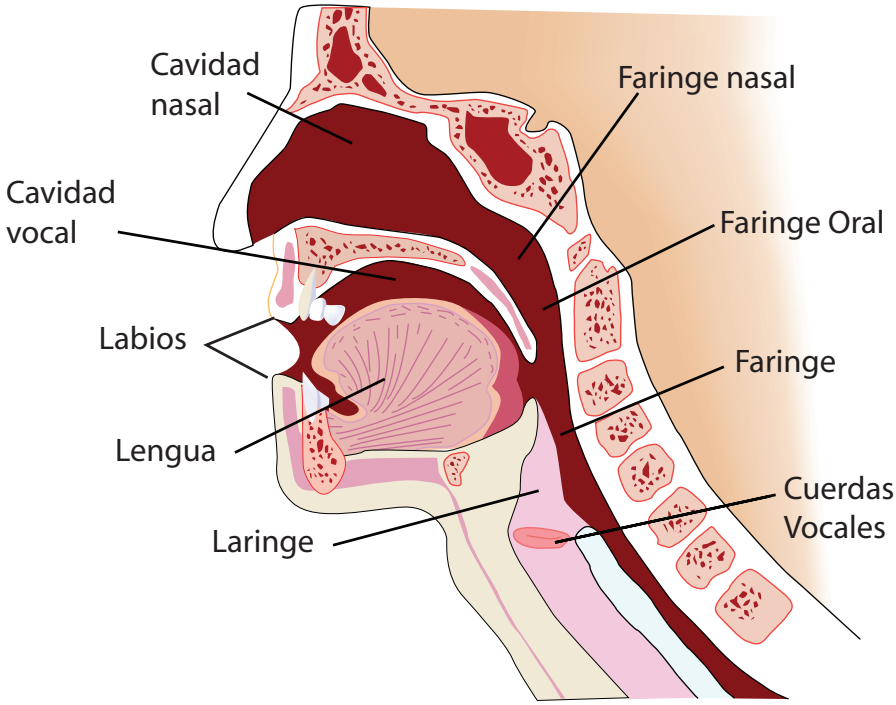


Figura 2.1: Esquema de producción de habla mediante los distintos órganos que participan en el proceso.

sordos tal y como ocurre con las letras 's', 'f', entre muchas otras

Una forma de modelar la producción de voz es observar dicho proceso como la interacción de una fuente excitadora y un filtro predictor del cual finalmente saldrá una señal representativa del habla. Para ello es necesario definir dos posibles entradas o excitaciones al sistema. Por un lado está la excitación mediante trenes de impulsos a frecuencia f_0 que caracterizarán a los sonidos sonoros o *voiced* mientras que otra opción es excitación mediante ruido blanco que caracterizará a los sonidos sordos o *unvoiced*. Independiente de la excitación que se tenga, está será modulada en amplitud por una ganancia, responsable principalmente de dar cuenta de la energía propia de la voz. Finalmente, la señal modulada es ingresada al filtro predictor, del cual saldrá una señal con armónicos denominados *formantes*, caracteristicas capaces distinguir a un sonido de otro. El asumir que la acción de las cavidades vocal y nasal además de otros órganos se resume en un filtro predictor es bastante cercano a la realidad, tanto así que en lo años treinta diversos desarrolladores construyeron maquinas mecánicas capaces de producir algunos sonidos de voz mediante la acción de flujos de aire y filtros predictores compuestos por cilindros móviles. En la figura 2.2 se muestra un diagrama de bloques que esquematiza el modelo de producción de habla.

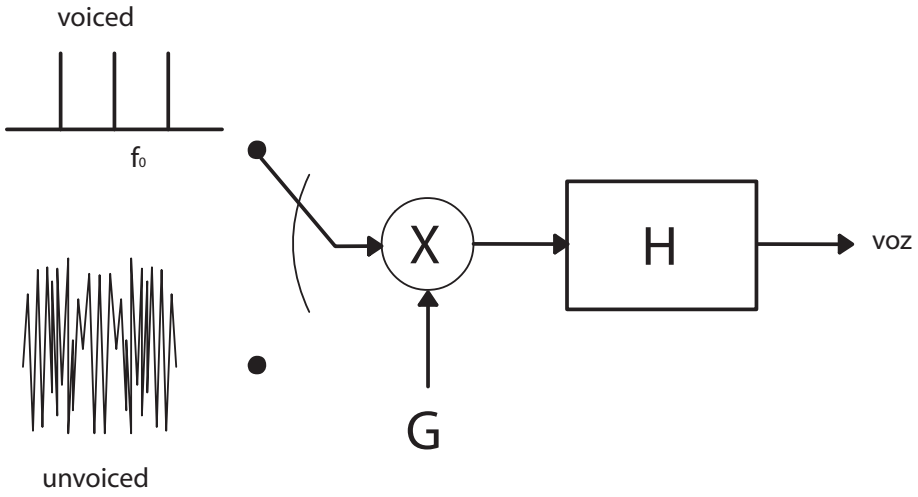
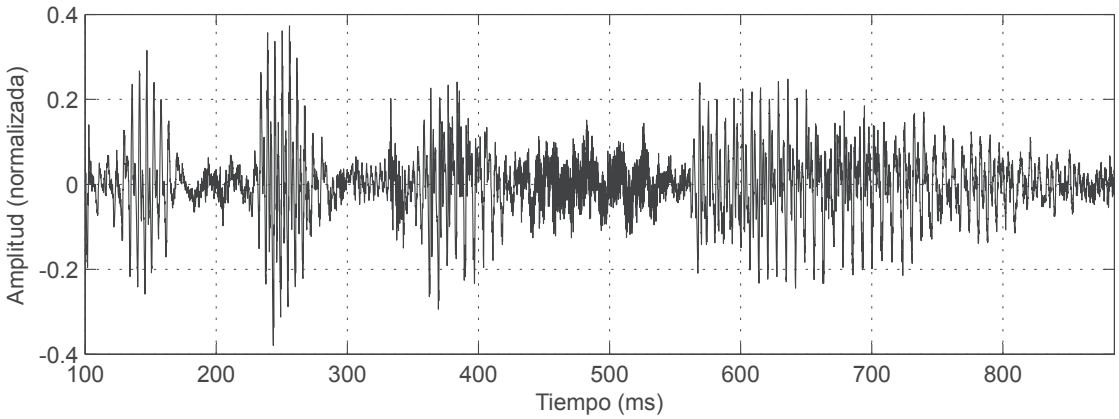


Figura 2.2: Diagrma de bloques del modelo de producción de voz.

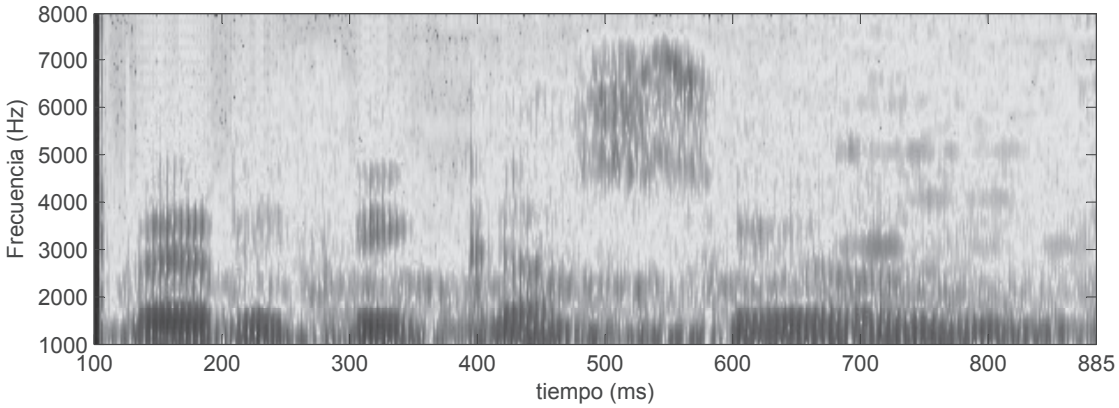
Una representación clásica de una señal sonora cualquiera es aquella en que se muestra la variación de la amplitud de la señal a través del tiempo. Este tipo de visualización, como la mostrada en la figura 2.3a, permite entre otras cosas el cálculo directo de la energía y la estimación de la frecuencia fundamental de la señal. Sin embargo, este tipo de representación no es útil en la mayoría de las tareas relacionadas con procesamiento de voz. Una forma más clara de observar las señales es mediante el llamado espectrograma, el cual muestra la variación del espectro a través del tiempo mediante una visualización en tres dimensiones. Para ello se grafica el espectro de intervalos de la señal representando mediante colores la presencia o ausencia de bandas de este. En la figura 2.3b se muestra el espectrograma para una determinada señal, donde las zonas de color más oscuro muestran bandas del espectro con mayor energía y por ende los formantes de la señal. A pesar de que las *formantes* (armónicos principales en una señal de voz) distinguen a un sonido de otro, para un mismo sonido en dos locutores distintos dichas formantes serán distintas y por ende permitirán distinguir a ambos locutores mediante su voz. Es este el principio básico de funcionamiento de la verificación de locutor, tema central de este trabajo.

2.3. Sistemas de Verificación de Locutor

Diversas son las aplicaciones que pueden desarrollarse mediante el análisis y procesamiento del habla y producto del creciente mercado de telefonía móvil y todo lo que esto conlleva, el sacar provecho a este modo de interacción hombre-máquina se ha vuelto un



(a) Representación Temporal



(b) Representación Espectral

Figura 2.3: Representación temporal y espectral de una señal voz. En el cuadro (a) se representa la voz como una onda mecánica a través del tiempo. En el cuadro (b) se muestra la variación del espectro a través del tiempo, donde los sectores más oscuros son de mayor energía y por ende muestran la presencia de formantes.

desafío importante de abordar. Como se muestra en la figura 2.4, el procesamiento de voz puede ser clasificado en varias áreas (Campbell, 1997). En este trabajo se da énfasis a la rama de reconocimiento y en particular al de locutor, tarea que consiste en lograr establecer la identidad de un sujeto mediante su voz. Para llevar a cabo dicha tarea existen dos posibilidades, en la primera de ellas se puede identificar a un sujeto dentro un conjunto acotado de usuarios abonados al sistema identificador mientras que una segunda posibilidad es aceptar o rechazar una hipótesis de identidad, tarea denominada verificación de locutor.

Bajo la mirada de clasificación de patrones, un verificador de locutor es un clasificador de dos estados, que bajo cierta hipótesis, acepta o rechaza dicha presunción, esto es,

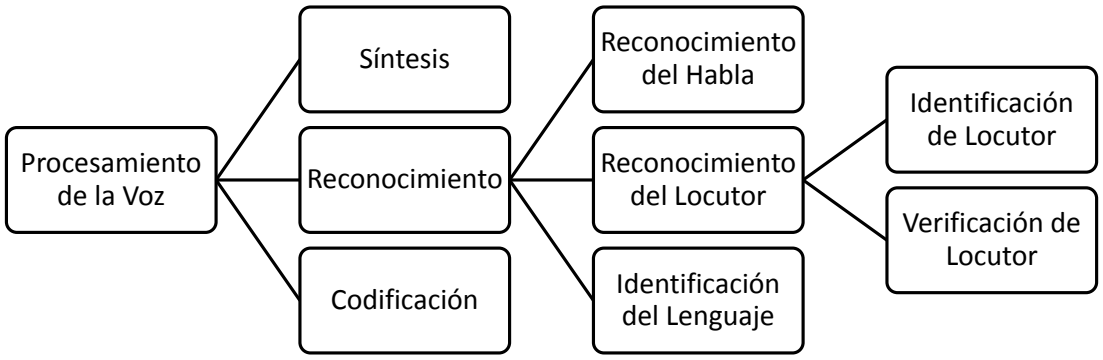


Figura 2.4: Tecnologías de Voz.

suponiendo una identidad definida por el locutor en cuestión, se acepta (el usuario es un *CLIENTE*) o se rechaza (el locutor es un *IMPOSTOR*) según el puntaje o *score* esté sobre o bajo un umbral de decisión.

Todo verificador de locutor basa su funcionamiento en dos etapas perfectamente definidas que son la etapa de entrenamiento y la etapa de verificación o *test*. En la etapa de entrenamiento se crean el o los modelos que dan cuenta de forma única la identidad de un locutor afiliado al sistema. Para ello es necesario que el locutor emita un cierto número de palabras de forma tal de extraer características adecuadas de sus voz y con ello entrenar modelos matemáticos capaces de distinguir a un usuario de otro. En secciones posterior se detalla que tipo de parametrización se utiliza así como también los modelos matemáticos a utilizar. En la etapa de verificación, se parametriza de igual forma el habla del sujeto para luego evaluar dichas características en el modelo de locutor que se clama tener. Con esto se obtiene un puntaje o *score*, el cual indicará si la identidad clamada es verdadera (*score* sobre el umbral de decisión) o es falsa (*score* bajo el umbral de decisión). En la figura 2.5 se muestra de forma esquemática las etapas de entrenamiento y verificación.

Dentro de los sistemas de verificación de locutor existen dos grandes familia clasificadas según el vocabulario que se utiliza. En primer lugar se tienen los motores verificadores texto dependientes, los cuales requieren un vocabulario fijo a pronunciar. En estos sistemas tanto en etapa de entrenamiento como verificación, se pronuncian palabras definidas por el sistema por lo que los modelos de locutor solo dan cuenta de un conjunto acotado y pequeño de unidades fonéticas y por ende no requieren un número masivo de elocuciones para entrenar el modelo del locutor. Este tipo de sistema basa su funcionamiento en los llamados *modelos ocultos de Markov*, tema que se explica en secciones posteriores. Sistemas de este tipo presentan hoy en día resultados muy buenos que pueden llegar

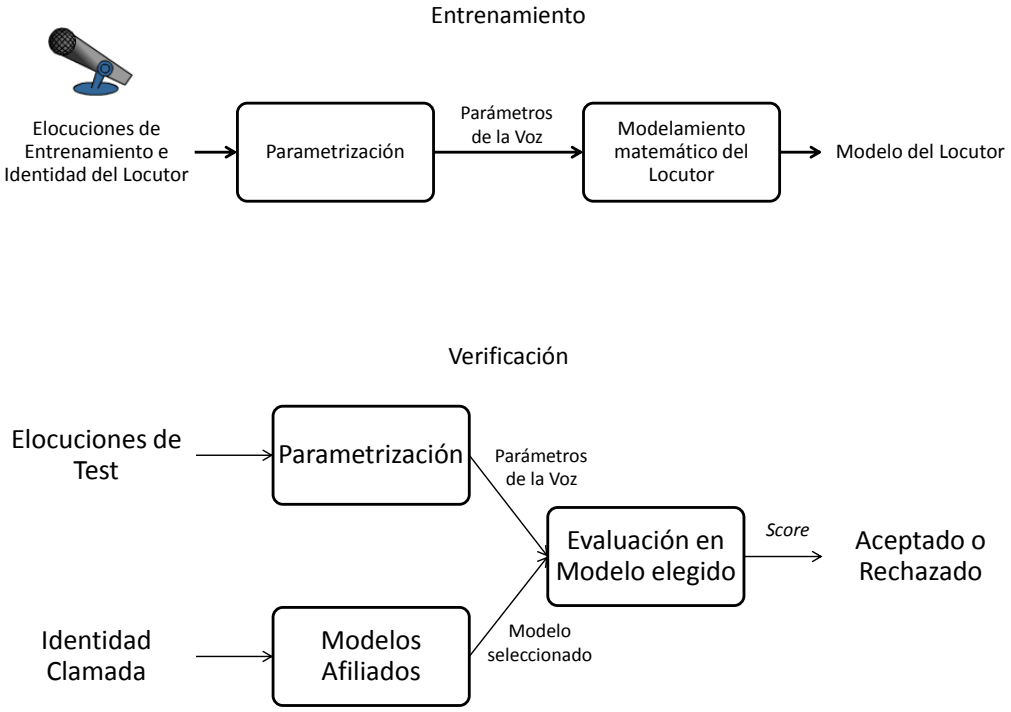


Figura 2.5: Esquema de las etapa de entrenamiento y verificación.

incluso bajo el 1 % de error bajo condiciones *matched*. Por otro lado existen también los motores verificadores texto independientes, los cuales relajan la restricciones de pronunciar determinadas palabras dejando a elección del usuario que elocuciones decir. Para lograr que un sistema de este tipo funciones se requieren varios minutos de habla para lograr así un entrenamiento capaz de reflejar en los modelos suficientes unidades fonéticas para tener un desempeño satisfactorio.Esto es claramente una dificultad puesto que se sabe el entrenamiento es un bien escaso y un usuario común no desea perder largo tiempo en entrenar un sistema. El funcionamiento, a diferencia de los sistemas texto dependientes esta basado en una técnica llamada GMM (D. Reynolds y Rose, 1995) (del inglés *Gaussian Mixtures Models* o mezcla de modelos gaussianos) con la cual se modela la voz de un usuario sin la restricción del orden de las palabras pronunciadas. Técnicas como SVM (del inglés *Support Vector Machine* o máquinas de soporte vectorial) son utilizadas también con un éxito comparable a GMM.

En el caso particular de este trabajo, se utiliza un sistema verificador de locutor texto dependiente, en donde el vocabulario a pronunciar en etapa de entrenamiento son los números del cero al nueve repetidos tres veces, mientras que en etapa de verificación se

pronuncian una serie de 4 números subconjunto de los dígitos utilizados en entrenamiento repetidos 2 veces.

2.4. Medidas de Desempeño

Con el fin de evaluar y comparar el desempeño de un motor verificador de locutor y de las diversas técnicas que le dan robustez frente al ruido, se definen a continuación una serie de métricas e indices utilizados ampliamente en la literatura tanto en verificación de locutor como en problemas más generales de clasificación de patrones. Para este trabajo en particular, es decir un verificador de locutor texto dependiente, existen 4 posibles salidas en el sistema. Estas son:

- Aceptar al locutor cuando realmente es quien dice ser
- Rechazar al locutor cuando no es quien dice ser
- Aceptar al locutor cuando no es quien dice ser (Falsa Aceptación)
- Rechazar al locutor cuando realmente es quien dice ser (Falso Rechazo)

Los dos primeros casos corresponden al comportamiento deseado del sistema mientras que los dos últimos son salidas indeseadas y por ende quienes definen el desempeño en cuanto a error del sistema. Es natural el querer minimizar la aparición de casos correspondientes a las últimas dos salidas por lo que se puede definir o medir el desempeño del sistema mediante un conteo de la aparición de este tipo de casos. De esta forma se define el EER (del inglés *Equal Error Rate* o tasa de igual error) que corresponde al error que se tiene cuando la tasa de falsa aceptación y falso rechazo se igualan. Una forma de calcular dicha medida es mover el umbral de decisión de modo de obtener una curva de falso rechazo y falsa aceptación según dicho umbral. Luego, la intersección de ambas curvas entrega la medida del EER ademas de otras métricas que se definirán a continuación. Esta explicación queda clara observando la figura 2.6. En ella se grafican las curvas antes mencionadas.

Se define además del EER el umbral de igual tasa de error o TEER (del inglés *Threshold Equal Error Rate*) que es el umbral que minimiza el el error y por ende con en el que se obtiene el EER. Ambos valores (EER y TEER), pueden ser definidas tanto para el sistema global como también para cada uno de los usuarios que estén enrolados, es decir, se puede tener un umbral de decisión general con el cual se clasificará a todos los usuarios o también tener uno para cada usuario en particular, el cual se aplicará según la identidad que el

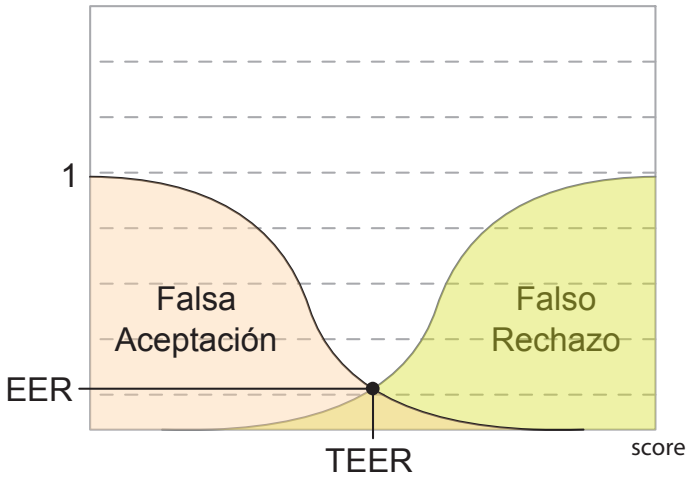


Figura 2.6: EER y TEER en las curvas de falsa aceptación y falso rechazo.

usuario clama tener. Esto es bastante intuitivo pues las distribuciones de probabilidad de cada usuario no tienen porque ser siquiera parecidas. De este modo, definido el TEER por persona, se tendrá también un EER particular, con los cuales se puede calcular un desempeño global del sistema mediante un promedio simple.

Con las curvas de falsa aceptación y falso rechazo se pueden definir otras medidas o formas de ilustrar el desempeño de un sistema. Este es el caso de las curvas ROC (del inglés *Reciever Operating Curve*) y DET (del inglés *Detection Error Tradeoff*. En ambas gráficas se muestra la tasa de falso rechazo versus la de falsa aceptación de forma lineal en la curva ROC y de forma logarítmica en la curva DET. En dichas curvas también es posible apreciar el EER el cual corresponde a la intersección de la diagonal con la curva, sea esta ROC o DET (ver figura 2.7).

2.5. Parametrización de la Voz

Crítico en la tarea de caracterizar de forma única a un usuario es la obtención y elección de las características propias de la señal de interés. En sistemas modernos y desde hace ya bastantes años, los parámetros utilizados ampliamente en la tarea de verificación de locutor son los coeficientes mel-cepstrales o MFCC
shortcitecepstralfurui,ss-wiener (del inglés *Mel Frequency Cepstral Coefficients*) los cuales pueden ser obtenidos siguiendo los pasos esquematizados en la figura 2.8. Estos coeficientes, se pueden interpretar como una combinación lineal del espectro logarítmico, por lo que

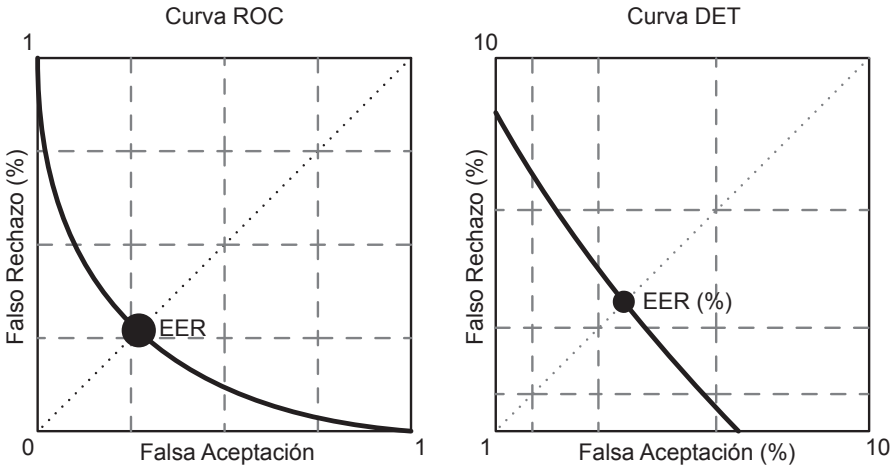


Figura 2.7: Curva ROC y DET.

están profundamente ligados al espectro de la señal.

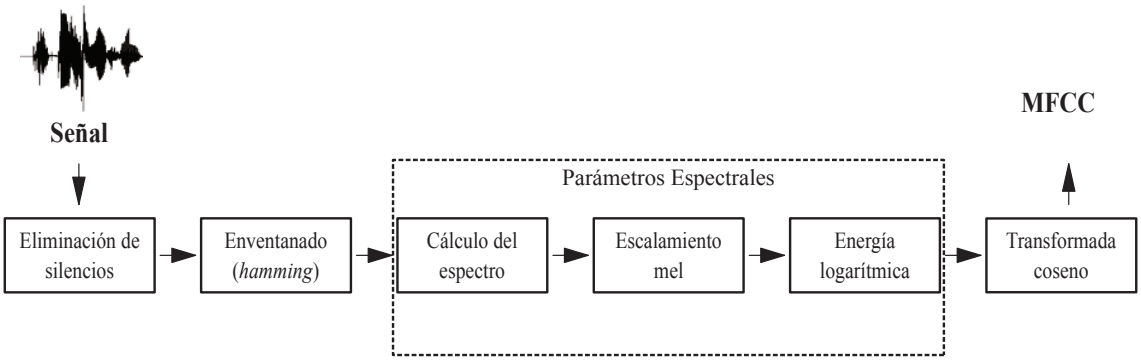


Figura 2.8: Diagrama de bloques de la extracción de los MFCC.

2.5.1. Eliminación de Silencios

Una buena forma de ahorrar tiempo computacional es eliminar información inútil en etapas tempranas del tratamiento de la señal. De esta forma, lograr eliminar silencios en el dominio temporal antes de hacer cálculos de espectro e incluso entrenar o verificar, permite tener un sistema más eficiente y óptimo en cuanto a tiempo de procesamiento se refiere. Además de los anterior, el excluir trozos de silencio de la etapa de entrenamiento permite dar cuenta de mejor forma de las características del locutor a parametrizar. Dicha tarea se realiza mediante el calculo de energía por frame y con ello se logra limitar la duración del comienzo y final además de silencios prolongados en secciones intermedias de la señal.

2.5.2. Enventanado

Una hipótesis en el manejo de señales de voz es el tratarlas como información estacionaria por tramos de tiempo definido (Hermansky, 1994; L. R. Rabiner y Schafer, 2007), es decir, para intervalos no mayores a cierta duración (20 a 50 ms) una señal de voz tiene la propiedad de ser estacionaria. Dicha suposición no siempre es del todo cierta pues existen factores como efectos de borde en las transiciones de un alófono a otro además de cierto grado de aleatoriedad en toda la señal. Sin embargo lo anterior, el modelar el habla como una señal periódica por pedazos agregando algunas consideraciones presenta los mejores resultados en sistemas modernos.

Una forma de evitar el efecto de transición de alófonos es utilizar traslape en el enventanado de esta. En el caso de este trabajo, se considera un traslape del 50 % por lo que cada intervalo o frame de análisis comparte la mitad de información con el intervalo siguiente. Agragado a lo anterior, se desea dar énfasis a la información presente en el centro de la ventana para lo cual se recurre a una ponderación de *hamming*. En la ecuación 2.1 se muestra la definición formal de la ventana de *hamming*, donde N en el número de muestras deseadas en el intervalo.

$$v(n) = 0,53836 - 0,46164 \cos\left(\frac{2\pi n}{N-1}\right) \tag{2.1}$$

De forma sencilla se pondera punto a punto la ventana $v(n)$ con el frame de interés de lo cual se obtiene una ventana a caracterizar con información concentrada en el medio (ver figura 2.9).

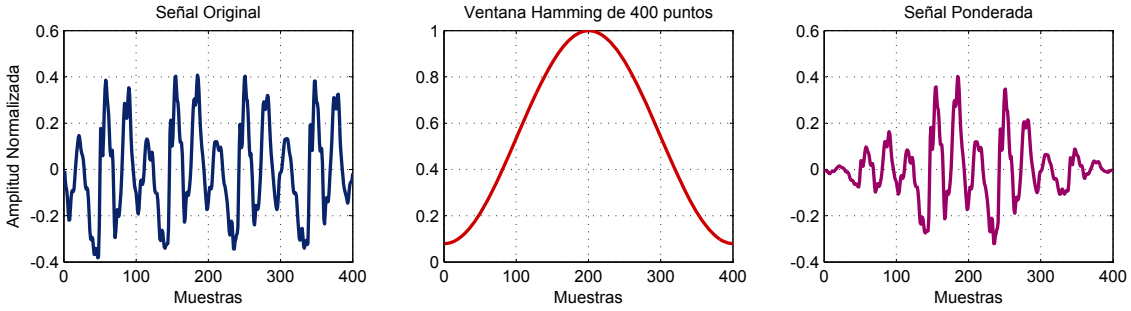


Figura 2.9: Acción de la ventana de hamming sobre un intervalo de interés.

2.5.3. Log-Energía del Banco de Filtros Mel

Un método para el cálculo espectral, análogo al cálculo de la FFT (del inglés *Fast Fourier Transform* o transformada rápida de Fourier) es la estimación de la densidad espectral mediante bancos de filtros de banda angosta. Esto es, determinada cierta resolución deseada para el espectro, se obtienen las energías de dichos filtros y con ello el espectro. Este procedimiento permite además, quitar énfasis en excesivos armónicos, dando especial atención a la envolvente del espectro, que en definitiva es la información relevante en la señal, pues es esta envolvente quien muestra la estructura de los formantes, identidades que determinan el sonido de un alófono u otro.

Existen tres parámetros a fijar en los filtros de banda angosta que son su posición en la banda de frecuencia, su ancho de banda y su forma. Para este trabajo se utilizan formas triangulares y anchos tal que exista traslape de 50 % entre filtro y filtro. Para el espaciemiento de los filtros se toma muy en cuenta lo que es la percepción humana, la cual es incapaz de distinguir variaciones de frecuencia a lo largo del espectro de forma lineal, es decir, el notar la diferencia entre 100 Hz y 200 Hz no es equivalente a distinguir entre 5,1 kHz y 5,2 kHz, y de manera empirica se ha mostrado que existe una relación logaritmica en la percepción. Dicha variación en la percepción es la que trata de dar cuenta escalas como la de Bark y Mel, la cuales transforman el espacio de la frecuencia a un espacio en que las diferencias de percepción sean lineales a lo largo del espectro. En este trabajo en particular, se utiliza la escala Mel (ver ecuación 2.2) como transformación de percepción y es acorde a esta escala que se poscicionan de forma equiespaciada los filtros triangulares.

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.2}$$

De esta forma, la posición de los 14 filtros triangulares está determinada por la ubicación de las frecuencias centrales de los filtros, posicionados en forma equiespaciada en escala Mel (ver figura 2.10), los cuales, por teorema del muestreo de Nyquist, se distribuyen a los largo de la porción de espectro comprendida entre las frecuencias 0 Hz y 8000 Hz dado el muestreo a 16 kHz utilizado en este trabajo.

Teniendo las energías correspondientes a los 14 filtros de banda angosta, se aplica una compresión logarítmica compensando de esta forma la percepción no lineal del volumen de un sonido, además de separar de forma aditiva la excitación y el filtro impuesto por el tracto vocal. Si se asume que una señal s es producto de la convolución en el tiempo de la

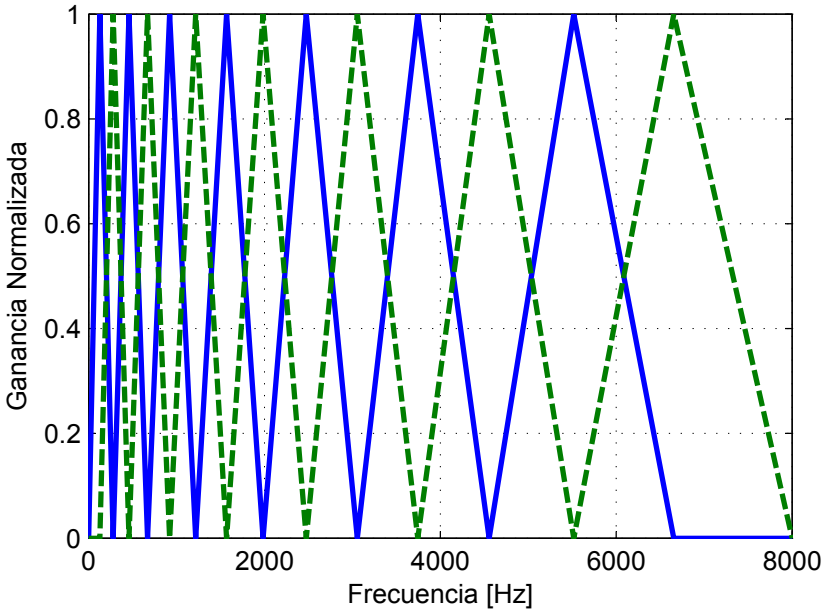


Figura 2.10: Banco de 14 filtros triangulares entre 0 y 8 kHz.

excitación h impuesta por las cuerdas vocales y el filtro g característico del tracto vocal, la aplicación del logaritmo al espectro de dicha señal puede verse en la ecuación 2.3.

$$\begin{aligned} s(t) &= h(t) * g(t) \\ &\Downarrow \\ S(f) &= H(f)\, G(f) \\ &\Downarrow \\ \log(S(f)) &= \log(H(f)) + \log(G(f)) \end{aligned} \tag{2.3}$$

2.5.4. Transformada Coseno

La transformada coseno o DCT (del inglés *Discrete Cosine Transform* o transformada coseno discreta) es equivalente a calcular la DFT (del inglés *discrete fourier transform*) y utilizar solo la parte real. La definición matemática formal de la DCT se muestra en la ecuación 2.4

$$
\begin{aligned}
X_k &= a_0 \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right] \\
k &= 0, \ldots, N-1 \\
a_0 &= \begin{cases} \frac{1}{\sqrt{2}} & \text{si } k = 0 \\ 1 & \text{resto} \end{cases}
\end{aligned} \tag{2.4}
$$

Aplicando esta transformación al logaritmo del espectro (cálculo del bloque anterior) se obtienen 14 coeficientes cepstrales de los cuales se conservan solo 10.

2.5.5. Coeficientes Delta y Delta-Delta Cepstrales

Frecuentemente en problemas de clasificación de patrones se introducen como parámetros adicionales la variaciones dinámicas de las características ya utilizadas. Esto introduce al sistema información adicional a la hora de clasificar. Se definen de esta forma los coeficientes dinámicos de primer y segundo orden, los cuales tienen la propiedad de ser relativamente robustos a variaciones en el canal transmisor. Los coeficientes delta cepstrales o coeficientes dinámicos de primero orden se definen según la ecuación 2.5 donde $c_{t,n}$ corresponde al coeficiente cepstral n del frame t de un total de T frames.

$$
\delta c_{t,n} = \begin{cases} \frac{c_{t+1,n} - c_{t,n}}{2} & t = 0 \\ \frac{c_{t+1,n} - c_{t-1,n}}{2} & 1 \leq t \leq T-1 \\ \frac{c_{t,n} - c_{t-1,n}}{2} & t = T-1 \end{cases} \tag{2.5}
$$

El cálculo de los coeficientes dinámicos de segundo orden se realiza de forma análoga reemplazando los coeficientes estáticos por los dinámicos de primer orden. Se agrega como parámetro a los coeficientes estáticos la energía del frame y de igual forma se calcula su variación temporal (primera y segunda derivada) cono lo cual se tiene un total de treinta y tres características a utilizar en el sistema verificador de locutor.

2.6. Verificación de Locutor Texto Dependiente y Modelos Ocultos de Markov

En un problema de clasificación de patrones existen en la literatura diversos enfoques según la naturaleza de la tarea a realizar. Una enfoque ámpliamente usado es la clasificación bayesiana basada en puntajes obtenidos según la probabilidad de ocurrencia de las clases. En verificación de locutor texto dependiente, mirado desde dicho enfoque, se tienen señales a evaluar de largo T *frames* y en donde de cada uno se extraen N parámetros o características. De esta forma se define como Observación a el conjunto de vectores $O\,(t) = [o_1(t)\, o_2(t)\, \cdots o_N(t)]$ donde $t \in [1, T]$. Además dado un modelo de locutor λ_i, el problema de verificación de locutor texto dependiente se traduce en que la probabilidad de que el locutor S_j corresponda a la hipótesis S_i dado un conjunto de observaciones O y un modelo de locutor λ_i sea mayor a cierto umbral de decición. Esto se muestra formalmente en la ecuación 2.6.

$$P\,(S_j = S_i\,|O, \lambda_i\,) > \delta \tag{2.6}$$

Utilizando el teorema de Bayes para probabilidades condicionales se puede llegar a que la condición de aceptación de un locutor, mostrada en la ecuación 2.6, se reescriba como sigue.

$$\frac{P\,(O\,|S_j = S_i, \lambda_i\,)\, P\,(S_j = S_i)}{P\,(O)} > \delta \tag{2.7}$$

En verificación de locutor solo algunos de los terminos de dicha condición son importantes, así la probabilidad de observar O y la de que el locutor S_j sea S_i se consideran constantes. Por ende la ecuación 2.7 se puede simplificar y llegar a la siguiente expresión.

$$P\,(O\,|S_j = S_i, \lambda_i\,) > \delta \tag{2.8}$$

Dicha probabilidad es imposible de determinar de forma directa y por lo tanto se recurren a técnicas capaces de estimarla de forma adecuada. Una de esas técnicas consiste en modelar el proceso del habla como un *Modelo oculto de Markov*, técnica que se explica en las siguientes secciones.

2.6.1. Modelos Ocultos de Markov

Una de las técnicas más ámpliamente usadas en verificación de locutor es el modelamiento mediante *Modelos ocultos de Markov* (HMM, del inglés *Hidden Markov Models*)(L. R. Rabiner, 1989). Un HMM se puede entender como una máquina de estado o autómata en donde cada estado, transición de estado y observaciones posibles son representadas mediante distribuciones de probabilidad, de esta forma un modelo tiene como parámetros a definir el número de estados, topología y probabilidades de los estados y transiciones. Un ejemplo de HMM es el mostrado en la figura 2.11, en la cual se esquematiza un *Modelo oculto de Markov* de 5 estados, topología de izquierda a derecha y sin salto de estados.

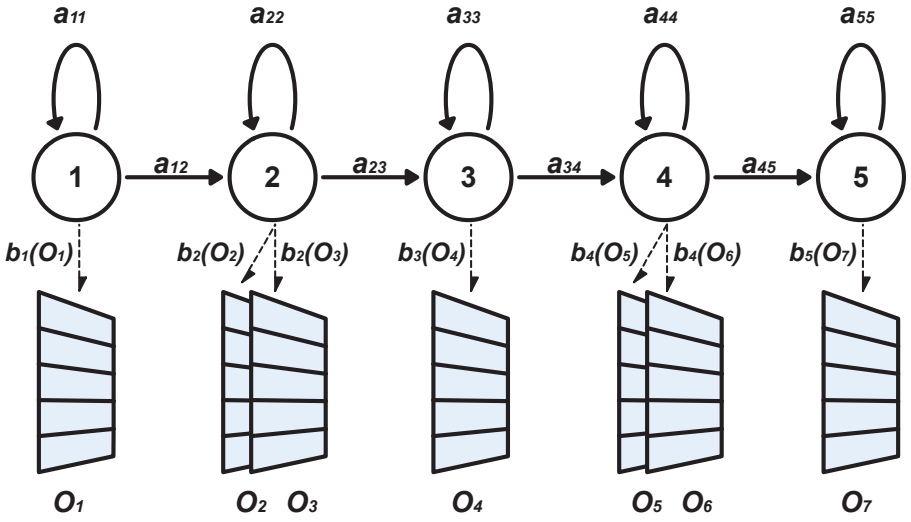


Figura 2.11: HMM de 5 estados y topología Izquierda-Derecha.

Asumiendo que el modelo corresponde a la secuencia de estados que se da para las observaciones $O = [O_1 \cdots O_7]$, dichos estados quedan definidos como $X = \{1, 2, 2, 3, 4, 4, 5\}$ de donde se desprende que la probabilidad conjunta de observación O dado un modelo de locutor clamado λ_i moviendose a través de la secuencia de estados X es la indicada en la ecuación 2.9.

$$P(O, X \mid \lambda_i) = b_1(O_1)a_{12}b_2(O_2)a_{22}b_2(O_3)a_{23}b_3(O_4)\ldots \tag{2.9}$$

Esta probabilidad no es factible de ser calculada pues la secuencia de estados X no es conocida y se denomina oculta (es por eso el nombre de *Modelos oculto de Markov*) y

la información a la que se tiene acceso solo proviene de la secuencia de observaciones O. De esta forma, la única forma de obtener la verosimilitud deseada es calculando todas las secuencias de estados posibles (figura 2.10) o una aproximación de aquello, la secuencia más probable (figura 2.11).

$$P(O, X|\lambda_i) = \sum_X a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(O_t) a_{x(t)x(t+1)} \tag{2.10}$$

$$P(O, X|\lambda_i) \cong Max\left\{ a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(O_t) a_{x(t)x(t+1)} \right\} \tag{2.11}$$

Estimar la ecuación 2.11 no es factible en forma directa, por lo que se aprovecha el algoritmo de Viterbi, originalmente ideado para decodificación de señales, en la tarea de calcular en forma rápida y eficiente dicha verosimilitud, obteniendo además la secuencia de estados más probable. Este algoritmo se explicará en secciones posteriores.

2.6.2. Probabilidad de Observación

En diversas aplicaciones, los HMM son utilizados con observaciones dentro de un universo finito de posibilidades y por ende las probabilidades asociadas a estas también lo son. Sin embargo, en aplicaciones de voz, dichas observaciones pertenecen a un conjunto infinito, en particular un vector de observación en un tiempo t pertenece a $\Re^N$, donde N es la cantidad de parámetros extraídos de un *frame* de cierta señal. De esta forma, las probabilidades de observar cierto suceso en algún estado tienen que responder a una función densidad de probabilidad continua. En verificación de locutor de dice que las probabilidades de observación corresponden a Gaussianas multivariables o sumas ponderadas de estas. En la ecuación 2.12 se muestra una estructura genérica para cada una de las probabilidades de observación, donde N_e representa el número de estados, G el número de gaussianas y c_{jg} la ponderación de cada una de ellas.

$$
\begin{aligned}
b_j(O_t) &= \sum_{g=1}^{G} c_{jg} \aleph(O_t; \mu_{jg}, \Sigma_{jg}) \ , 1 \leq j \leq N_e \\
\sum_{g=1}^{G} c_{jg} &= 1 \\
c_{jg} &\geq 0 \\
1 \leq & j \leq N_e \\
1 \leq & g \leq G
\end{aligned}
\tag{2.12}
$$

Una gaussiana multivariable o distribución normal multivariable, se define en forma individual según la ecuación 2.13, en la cual N es la dimensión de las observaciones, μ es el vector de medias y Σ es la matriz de covarianzas.

$$
\aleph(O_t; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} e^{-\frac{1}{2}(O_t - \mu)^T \Sigma^{-1} (O_t - \mu)} \tag{2.13}
$$

2.6.3. Algoritmo de Búsqueda de Viterbi

Recordando secciones anteriores, el que un locutor tenga o no la identidad que clama tener, esta determinado si la probabilidad conjunta de observación O dado un modelo de locutor clamado λ_i moviéndose a través de la secuencia de estados X es mayor a un umbral establecido δ. Dicha probabilidad no es factible de calcular de forma directa y por ende debe ser estimada. Una aproximación a aquello es la secuencia de estados más probable con la cual puede ser calculada la verosimilitud requerida. Para estimar dicha secuencia de mayor probabilidad es que se utiliza el algoritmo de de codificación de Viterbi, procedimiento iterativo y de programación dinámica ideado originalmente para la decodificación de señales aplicadas principalmente a transmisión inalámbrica como lo es la telefonía móvil. Esta técnica es utilizada en verificación de locutor texto dependiente para lograr un alineamiento óptimo entre los frames que componen una señal y la secuencia de estados correspondiente o más probable. Dicho alineamiento puede ser visualizado como la solución a un problema de maximización de caminos dentro de una malla, es decir, el camino más probable. En la figura 2.12 se muestran los caminos posibles en la decodificación, donde en el eje vertical se muestran los estados posibles, en el vertical los *frames* de la señal a verificar, los arcos que unen la malla son las probabilidades de transición y cada vértice en la malla es la probabilidad de observación.

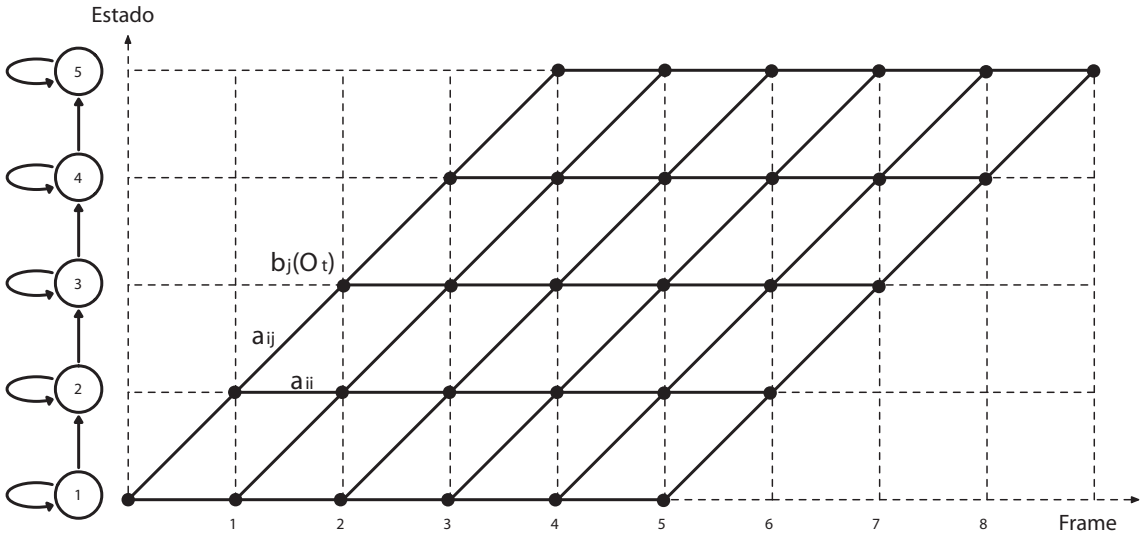


Figura 2.12: Malla correspondiente a la decodificación de Viterbi.

El recorrer la malla obteniendo la mayor probabilidad, es decir, encontrar la secuencia de estados $Q = \{q_1\, q_2\, \cdots\, q_t\}$ óptima, dada la secuencia de observaciones a lo largo del tiempo $O = \{O_1\, O_2\, \cdots\, O_t\}$, se resuelve definiendo la cantidad $\delta_t\,(i)$ tal y como se muestra en la ecuación 2.14.

$$\delta_t\,(i) = \max_{q_1, q_2, \cdots, q_{t-1}} P\,(q_1\, q_2\, \cdots\, q_t = i, O_1\, O_2\, \cdots\, O_t\,|\lambda) \tag{2.14}$$

Dicha probabilidad representa la mayor verosimilitud a lo largo del camino óptimo hasta el tiempo t, que incluye las primeras t observaciones terminando en el estado S_i. Luego, por inducción se consigue lo expresadoen la ecuación 2.15.

$$\delta_{t+1}\,(j) = \max_i \{\delta_t\,(i)\, a_{ij}\}\, b_j\,(O_{t+1}) \tag{2.15}$$

De esta forma se define el siguiente algoritmo.

- Inicialización

$$\begin{aligned} \delta_1\,(i) &= \pi_i b_i\,(O_1)\,,\ 1 \leq i \leq N \\ \psi_1\,(i) &= 0 \end{aligned} \tag{2.16}$$

  Donde ψ es un arreglo que guarda el camino óptimo seguido seguido y N el número de estados. Además se define la cantidad π_i que refleja la probabilidad que es estado

i sea el primero.

- Recursión

$$\begin{aligned} \delta_t(j) &= \max_{1\leq j\leq N} \{\delta_{t-1}(i)\, a_{ij}\}\, b_j(O_t) \\ \psi_t(j) &= \arg\max_{1\leq j\leq N} \{\delta_{t-1}(i)\, a_{ij}\} \end{aligned} \tag{2.17}$$

Para $2 \leq t \leq T$ y $1 \leq j \leq N$

- Finalización

$$\begin{aligned} P^* &= \max_{1\leq j\leq N} \{\delta_T(i)\} \\ q_T^* &= \arg\max_{1\leq j\leq N} \{\delta_T(i)\} \end{aligned} \tag{2.18}$$

- Recuperación del Camino

$$q_t^* = \psi_{t+1}\left(q_{t+1}^*\right),\ t = T-1,\ T-2,\ \cdots,1 \tag{2.19}$$

Para simplificar los cálculos y aliviar la carga computacional del sistema, se aplica logaritmo a todas las probabilidades de tal forma de transformar las multiplicaciones en suma. De esta forma se tiene el alineamiento óptimo y la respectiva verosimilitud.

2.6.4. Normalización de Verosimilitud

En una primera versión de un sistema verificador de locutor, tal y como se plantea en secciones anteriores, se aceptará o rechazará un determinado locutor según su puntaje (verosimilitud) sea mayor o menor que un determinado umbral. Esto se traduce en clasificar según la distancia al modelo del locutor clamado. Sin embargo esta distancia no tiene porque ser pareja para todos los locutores puesto que algunos de los usuarios pueden presentar bajas verosimilitudes sin que esto corresponda a un caso de impostor. Una forma de enfrentar este problema es crear un modelo *speaker independent* (SI) para el caso de verificación de locutor texto dependiente o UBM (del inglés *Universal Background Model*) para el caso texto independiente, proceso que consiste en realizar dos evaluaciones de modelo. Una en el modelo de locutor clamado y otra en un modelo de universo, construido o entrenado a partir de una gran cantidad de señales de locutores fuera del conjunto de clientes registrados en el sistema. Luego, para aceptar a un locutor, se debe cumplir la ecuación 2.20, donde Ω es el modelo universal SI o UBM según sea el caso.

$$P\left(O\left|\lambda_j\right.\right) > P\left(O\left|\Omega\right.\right) \Leftrightarrow \frac{P\left(O\left|\lambda_j\right.\right)}{P\left(O\left|\Omega\right.\right)} > 1 \tag{2.20}$$

En una aplicación práctica, el umbral de decisión no tiene porque ser 1 y podría variar un poco de este valor.

Una variación de lo anterior y lo que actualmente se utiliza en la mayoría de los verificadores de locutor texto dependientes es el uso de varios modelos de impostor denominados *cohorts*. Para ello se define una cantidad C de verosimilitudes logarítmicas extraídas de modelos de locutores no pertenecientes al conjunto de clientes registrados en el sistema. Luego, se le llama *llr* (del inglés *Log-Likelihood Ratio* o radio de verosimilitud logarítmica) a la resta del *score* (verosimilitud logarítmica) del locutor en el modelo clamado y el promedio de los M mayores *scores* del locutor en los modelos de *cohorts*.

$$llr_j = \log\left(P\left(O\left|\lambda_j\right.\right)\right) - \log\left(\frac{1}{M}\sum_{k=1}^{M} P\left(O\left|\lambda_{cohort_k}\right.\right)\right) \tag{2.21}$$

En la ecuación 2.21 se define el llr, en donde λ_{cohort} son los modelos de *cohort*. De esta forma el aceptar o rechazar a un locutor j está determinado si el llr_j es mayor o menor que un umbral δ. Este tipo de normalización es el que entrega los mejores resultados en verificadores texto dependientes actuales.

2.7. Robustez a la Distorsión de Canal en Verificación de Locutor

En el trabajo de esta memoria se presta especial atención al canal de transmisión y como este es responsable de parte del ruido introducido en las señales, cosa que se traduce en una baja en el desempeño del sistema verificador de locutor. En lo que sigue se explica en mayor profundidad la influencia del canal de transmisión, un modelamiento de este y formas de abordar esta problemática.

2.7.1. Influencia del Canal de Comunicación

Como se ha mencionado en secciones anteriores, la verificación de locutor consta de dos etapas distintas que son el entrenamiento y la verificación. En una aplicación sobre la red telefónica, dado que un usuario no desea perder mucho tiempo en afiliarse al sistema, la cantidad de datos que se dispone para entrenar modelos para dicho usuario es limitada. Producto de esta limitante, el modelo del locutor solo reflejará las condiciones que existían en el momento en que se entrenó el modelo. Por otro lado, al momento de verificación, el locutor podría no estar bajo las condiciones en que se encontraba cuando hizo el entrenamiento y por lo tanto el sistema tendrá una baja en su desempeño. Estas inconsistencias entre el momento de entrenamiento y verificación o *test* incluyen por ejemplo: cambios en el estado de ánimo del locutor; ruido ambiente; canales de transmisión distintos; variabilidad natural de la voz humana; etc. Luego es un desafío importante el lograr atenuar los efectos adversos que introducen estos factores en el motor verificador.

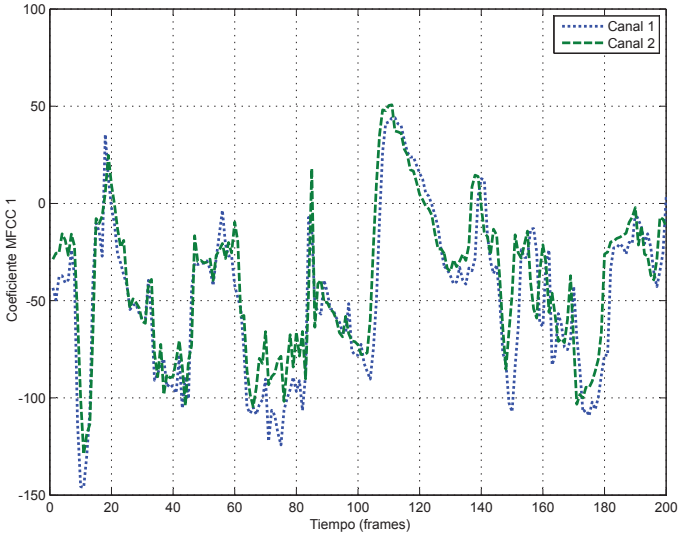


Figura 2.13: Efecto de la acción del canal sobre los coeficientes MFCC. En la linea punteada se muestra la trayectoria temporal del coeficiente uno para un canal mientras que en la linea a rayas se muestra la trayectoria temporal del coeficiente uno para una señal idéntica transmitida por un canal distinto.

Uno de los grandes problemas que se tiene al parametrizar la voz mediante coeficientes MFCC, es la gran dependencia que tienen estos sobre el canal de transmisión (D. A. Reynolds *et al.*, 1995; D. A. Reynolds, 1996) tal y como se muestra en la figura 2.13, por lo que inconsistencias entre el canal de comunicación utilizado en entrenamiento y el usado en la verificación hacen que la parametrización varíe y por ende los modelos no respondan

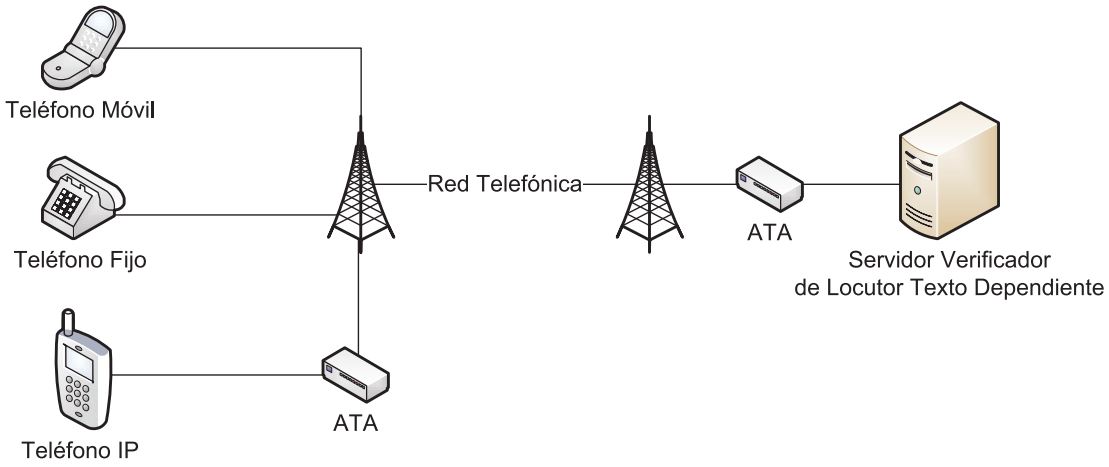


Figura 2.14: Esquema de la red sobre la que opera un verificador de locutor.

de la forma que se espera, traduciéndose esto en una baja en el desempeño del sistema. A esta condición de inconsistencia es a la que se denomina *mismatch* de canal, objeto de estudio central en este trabajo.

2.7.2. Modelo del Canal de Comunicación

Un verificador de locutor texto dependiente operado sobre la red telefónica presenta de forma critica problemas de *mismatch* de canal, esto pues al ser remoto el proceso de enrolamiento (acción de entrenar un modelo para un locutor) es común que exista disparidad de canal entre la fase de entrenamiento y *test*. En la figura 2.14 se esquematiza la estructura de red del sistema verificador de locutor aplicado a telefonía.

Se define *handset* como el aparato encargado de transmitir una señal de voz o sonido en general a través de la red telefónica. Así por ejemplo un teléfono celular es un *handset* distinto a un teléfono fijo producto de la diferencia entre los micrófonos (transductor) y circuitería en general. Si al *handset* le agregamos el canal transmisor (red telefónica en este caso) se tendrá el denominado *canal*. De esta forma, para que exista un *mismatch* de canal solo es necesario que haya disparidad en una de sus partes, ya sea *handset* o canal transmisor.

Una forma de modelar la acción del canal sobre las señal de voz es representar dicho canal como un filtro invariante en el tiempo e independiente de la señal con lo que el problema de canal se transforma en un problema de ruido convolucional. Es necesario

mencionar que los supuesto de invariancia e independencia del filtro no son del todo ciertas, sin embargo una buena aproximación de la realidad es aceptar dichos supuestos.

Para formalizar matemáticamente se supondrá que la señal de voz pura esta representada por la función en el tiempo $s(t)$ mientras que el ruido aditivo, producto del medio ambiente, será $n(t)$. En un verificador de locutor texto dependiente operado sobre la red telefónica, son ambas fuentes sumadas las que ingresan al canal, en donde son convolucionadas con el filtro característico h de dicho canal (recordar que el canal lo compone el handset y el canal de transmisión). Esta operación es la que se muestra en la ecuación 2.22, donde $y(t)$ es la señal que finalmente ingresa al sistema verificador.

$$y(t) = (s(t) + n(t)) * h \tag{2.22}$$

Si a la ecuación anterior se le calcula el espectro mediante transformada de Fourier y luego se le aplica logaritmo, se llega a una separación aditiva entre la acción del canal y la señal de voz, tal y como se ve en la ecuación 2.23. Sin embargo, el ruido aditivo queda inserto en la señal limpia y por ende debe ser tratado y eliminado en una fase anterior con algunas de las técnicas ideadas para aquello, como *spectral substraction* (Hardt y Fellbaum, 1997), algoritmo ponderado de Viterbi (Becerra Yoma *et al.*, 1998), etc.

$$\log(Y) = \log(S + N) + \log(H) \tag{2.23}$$

Notar que la operación realizada, es decir, cálculo de espectro y logaritmo, es muy similar a la extracción de los coeficientes MFCC, por lo que compensar el canal puede entenderse como lograr estimar el *bias* o promedio no nulo de los parámetros y con ello limpiar la señal del ruido convolucional.

2.7.3. Técnicas de Cancelación de Canal

Para cancelar el ruido convolucional o de canal, existen técnicas agrupadas en tres familias según el nivel de acción. El primer grupo reúne a los procedimientos que transforman los parámetros tal que estos sean robustos a los efectos de canal (Burget *et al.*, 2007; Zhonghua *et al.*, 2004). El segundo trabaja a nivel de modelos, adaptándolos según el canal al que se enfrente el sistema (Surendran *et al.*, 1999). Por último, el tercer grupo opera

sobre los *scores* que entrega el sistema. Este tipo de técnica se denomina normalización de *scores*. En las siguientes secciones se explican en más detalles cada una de estas tres familias y se dan ejemplos de técnicas clásicas en la literatura.

2.7.4. Transformación de Parámetros

Este tipo de técnica opera sobre un nivel de abstracción bajo puesto que el funcionamiento de esta familia de métodos trabaja prácticamente sobre la señal, aplicando transformaciones a los parámetros tal que estos sean robustos al accionar del canal. Una estructura general de este tipo de procedimientos se muestra en la ecuación 2.24, donde O son los parámetros originales, A una matriz de trasformación, b un vector *bias* y $\hat{O}$ los parámetros transformados.

$$\hat{O} = AO + b \tag{2.24}$$

Ejemplos clásicos de este tipo de técnicas son CMN, RASTA y en general cualquier tipo de filtro aplicado a las trayectorias temporales de los parámetros, resaltando filtros Wienner (Bai *et al.*, 2004; Xie *et al.*, 2006), mapeo probabilístico (Neumeyer y Weintraub, 1994), entre tantos otros. La gran ventaja que ofrece este tipo de procedimientos es que no requieren complejos modelos estocásticos ni algoritmos iterativos de demasiada carga y por ende el tiempo computacional utilizado en los cálculos implicados suele ser bajo.

2.7.4.1. Normalización de la Media Cepstral (CMN)

Como se vio en la sección 2.7.2, la acción del canal puede verse como un bias en el espectro logarítmico de la señal. Luego, calcular los coeficientes MFCC es simplemente realizar una combinación lineal de dicho espectro y por ende en cada coeficiente se mantendrá la presencia de un promedio no nulo. Basado en el supuesto de que el promedio de los coeficientes MFCC a lo largo del tiempo es nulo, se desarrolla la técnica denominado CMN (del inglés *Cepstral Mean Normalization*), la cual simplemente elimina el *bias* presente en los coeficientes MFCC dejándolos con promedio nulo. Aquel procedimiento es el que se muestra en la ecuación 2.25, donde N es el número total de frames que componen la señal, O_i el parámetro transformado i a través del tiempo y c_i el coeficiente MFCC i a través del tiempo.

$$O_i(t) = c_i(t) - \frac{1}{N}\sum_{k=1}^{N} c_i(k) \tag{2.25}$$

Existen algunos problemas con esta técnica pues el supuesto de que los coeficientes MFCC tienen promedio nulo a través del tiempo deja de cumplirse para señales muy cortas (caso predominante en verificación texto dependiente). Luego, la extracción del promedio elimina información propia del locutor. Otra desventaja que muchas veces no es considerada dado el largo de las señales utilizadas en verificadores texto dependiente, es que esta técnica requiere el cálculo del promedio y por ende no puede ser obtenido hasta que se tenga la totalidad de la señal. Así, esta técnica no es en tiempo real y requiere un tiempo adicional al de obtención de la señal para procesar los cálculos requeridos. Una solución a este problema es el cálculo de un promedio mediante una ventana móvil con lo cual se pierde un poco de precisión a cambio de poder procesar datos mientras se recibe la señal. Autores proponen también considerar no solo el promedio sino también la varianza de las muestras, llevando a los parámetros a una distribución de probabilidad deseada que típicamente responde a la forma de una normal (Skosan y Mashao, 2006). Otras formas de optimizar el proceso o evitar el cálculo directo del promedio, es reemplazar dicho cálculo por un filtro pasa altos de muy baja frecuencia de corte, tal y como se ha hecho en varias técnicas propuestas en la literatura (Naik, 1995; Lo *et al.*, 1999; Zilovic *et al.*, 1998).

Una variante interesante de CMN es la que propone Tukekci en su trabajo (Tukekci, 2007), en el cual sugiere normalizar el espectro extraído del banco de filtro Mel.

Sistemas basados en otras parametrizaciones como coeficientes LPC y todo tipo de parámetros extraídos a partir de filtros predictivos también cuentan con una gama de técnicas equivalentes a CMN aplicadas en otros dominios (Yu y Hsiao-ChuanWang, 2003). El procedimiento común es la normalización de los parámetros cualquiera que estos sean o el remplazo de dicha normalización por un filtro pasa altos capaz de extraer el promedio.

2.7.4.2. Filtrado RASTA

RASTA (del inglés *Relative Spectra*)(Hermansky y Morgan, 1994), es una técnica de filtrado de trayectorias temporales de los parámetros (Hermansky, 1995) desarrollada a partir del supuesto de que el rango de variación del tracto vocal es limitado y acotado. Luego, presencia de variaciones fuera de dicho rango son solo producto de la acción del canal y por ende pueden ser eliminadas. El accionar de esta técnica es el siguiente:

- Cálculo del Espectro
- Aplicación de alguna compresión no lineal
- Filtrado Pasa Banda (en el dominio Z)

$$H(z) = 0,1z^4 \left(\frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0{,}98z^{-1}} \right) \tag{2.26}$$

- Expansión correspondiente a la compresión aplicada

Existen dos versiones de RASTA denominadas Log-RASTA y J-RASTA. La primera esta diseñada para eliminar la acción del canal mientras que la segunda fue diseñada para eliminar ruido aditivo. Ambas difieren en la compresión no lineal utilizada. Mientras que en Log-RASTA se aplica simplemente logaritmo (Logartimo del espectro), en J-RASTA se aplica una compresión que responde a la ecuación 2.27, donde J es una constante dependiente de la señal. Notar que para $J >> 1$ la transformación se vuelve logarítmica mientras que para $J << 1$ se vuelve lineal

$$y = \ln(1 + Jx) \tag{2.27}$$

2.7.5. Cancelación de Máxima Verosimilitud de la Componente de Canal (ML-SBR)

Tal como se ve en la sección 2.7.2, la parametrización de señales mediante coeficientes MFCC conduce a que la distorsión de canal, asumiendo invariancia de esta en el tiempo, se modela como una componente aditiva y por ende la estimación de dicha componente llevaría a solucionar los problemas de *mismatch* de canal, puesto que conociendo la componente aditiva H se podría tener el vector de parámetros limpios O restando el canal H de los parámetros distorsionados O^d, tal como se explica en la ecuación 2.28.

$$O = O^d - H \tag{2.28}$$

Es obvio que en un sistema real, el canal H no puede ser estimado de forma exacta, sin embargo si es posible tener una aproximación de ello mediante el criterio de máxima verosimilitud o ML (Juang y Rahim, 1996; Sankar y Lee, 1996). Para ello se utiliza una metodología de *codebooks* compuestos con *codewords* representantes de las unidades fonéticas presentes en el sistema. El procedimiento consiste en construir un modelo (*codeword*)

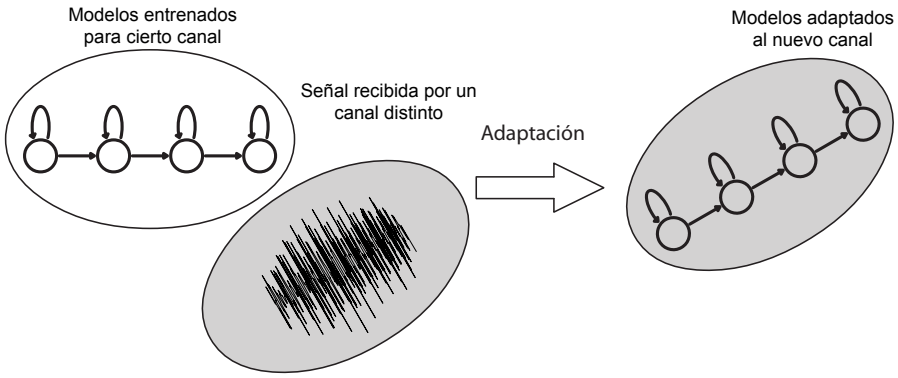


Figura 2.15: Vista esquemática de como opera la adaptación de modelos.

de tipo gaussiano representativo de señales limpias o grabadas sobre un canal referencial. El criterio ML, expresado en la ecuación 2.29, debe ser resuelto mediante el algoritmo de *Expectation Maximization* (EM).

$$\hat{H} = \arg\max_{H} P\left(O^d \,|\lambda, H\right) \tag{2.29}$$

En la ecuación anterior, se muestra que bajo el criterio ML se puede lograr una estimación del canal $\hat{H}$ teniendo el modelo de canal limpio o referencial λ y entregando señales distorsionadas por el canal O^d, dado que bajo los supuestos antes mencionados, los modelos de canal limpio y distorsionado solo difieren en las medias μ, las se encuentran separadas por un *bias* H representante del canal.

2.7.6. Adaptación de Modelos

Esta técnica, al igual RASTA, fue ideada originalmente para ser aplicada en reconocimiento de voz, sin embargo su uso se ha expandido a otras ramas del procesamiento de voz llegando incluso a utilizarse en sistemas texto a voz o TTS. A grandes rasgos su objetivo es adaptar los modelos de locutor construidos para cierto tipo canal a un canal diferente mediante el uso de una limitada cantidad de datos. Dicho procedimiento es el que se esquematiza en la figura 2.15.

Existen dos familias principales que agrupan a este tipo de técnicas que son las de adaptación directa y las de adaptación indirecta. En las siguientes secciones se explican dos procedimientos representativos de cada una de estas familias y que son ampliamente

utilizadas en la literatura.

2.7.6.1. MAP

MAP (Mak *et al.*, 2006) (del inglés *Maximum a Posteriori*) es una técnica del tipo directo puesto que su objetivo es reestimar los parámetros del modelo de locutor a verificar. Para ello se debe maximizar una función densidad de probabilidad g y f tal como se muestra en la ecuación 2.30, donde λ es el modelo original, λ_{MAP} es el modelo adaptado y X es la señal proveniente del locutor.

$$\lambda_{MAP} = \arg\max_{\lambda} f\left(X\,|\lambda\right) g\left(\lambda\right) \tag{2.30}$$

Clave para el funcionamiento de MAP es el disponer buenas estimaciones de las distribuciones de probabilidad f y g tal que la adaptación sea correcta. Para ello se han desarrollado diversos métodos en la literatura.

2.7.6.2. MLLR

A diferencia de MAP, MLLR (Yiu *et al.*, 2007) (del inglés *Maximum Likelihood Linear Regression*) no adapta de forma directa los parámetros sino que mediante la maximización de la verosimilitud se calculan transformadas con el propósito de adaptar con ello los parámetros que componen un modelo de locutor. Un caso sencillo es la adaptación de las medias del modelo y no de las varianzas, con lo cual el procedimiento a realizar es encontrar una matriz cuadrada de transformación W para ser aplicada en el vector de medias original μ. Se añade ademas un *bias* a dicha transformación obteniendo lo mostrado por la ecuación 2.31.

$$\hat{\mu} = W\mu + b \tag{2.31}$$

Para obtener el vector de medias transformado $\hat{\mu}$, es necesario resolver un problema de optimización utilizando el criterio de máxima verosimilitud mostrado en la ecuación 2.32, donde el procedimiento estándar de resolución es el algoritmo de *Expectation Maximization* o EM.

$$\hat{W}, b = \arg\max_{W,b} P\left(O \,|\lambda, W, b\right) \tag{2.32}$$

Con O las observaciones o información necesaria para estimar las transformaciones y λ el modelo a adaptar.

2.7.7. Normalización de Scores

Ideada originalmente para ser aplicada en verificadores de locutor texto independientes (Yin *et al.*, 2008; Barrus y Gauvain, 2003), la normalización de *scores* intenta simplificar en cierta medida la difícil elección de umbral. Tiene como objetivo disminuir la varianza presente tanto en *scores* de cliente (hipótesis de locutor correcta) como también de impostor (hipótesis de locutor falsa), además de centrar dichas distribuciones. Todo tipo de normalización sigue la estructura mostrada en la ecuación 2.33 (Mariéthoz y Bengio, 2005), donde llr_i es el *score* normalizado para el locutor que clama tener la identidad i, ll_i la verosimilitud que tiene el mismo locutor, μ una media normalizadora y σ una desviación estándar de igual propósito.

$$llr_i = \frac{ll_i - \mu}{\sigma} \tag{2.33}$$

En las siguientes secciones se explican dos tipos de normalizaciones muy utilizadas en la literatura además de ser parte esencial de parte de este trabajo.

2.7.7.1. H-Norm

H-NORM (del inglés *Handset Normalization*) es una técnica no ciega o supervisada cuyo objetivo es compensar la variabilidad que presentan los *scores* frente a diversos canales. Para su funcionamiento es necesario conocer el tipo de handset con el que se está intentado verificar algún determinado locutor. Una opción para aquello es utilizar el ANI (número identificador del teléfono), sin embargo solo es posible distinguir entre telefonía móvil y fija mediante este sistema. Para normalizar el *score* se calcula una media y desviación estándar acorde al *handset* indicado, tal como lo muestra la ecuación 2.34, donde llr_i es el *ratio* o *score* de desición normalizado para el locutor que clama tener la identidad i, ll es la verosimilitud logarítmica del mismo usuario y mu_i^H y σ_i^H términos normalizadores

propios de la identidad clamada i.

$$llr_i = \frac{ll_i - \mu_i^H}{\sigma_i^H} \tag{2.34}$$

El cálculo de mu_i^H y σ_i^H se realiza evaluando distintas señales de usuarios no afiliados al sistema sobre el modelo del locutor que se clama ser i. De dichos *scores* se seleccionan los M mayores y se calcula una media mu_i^H y una desviación estándar σ_i^H. Una ventaja de esto es que ambos términos pueden ser calculados en etapa de entrenamiento, ahorrando así tiempo computacional.

2.7.7.2. T-Norm

Como se ha mencionado antes, la voz humana y sus características pueden variar en el tiempo según el estado de animo del locutor, edad, estado de salud y un sin número de factores influyentes. Dicho fenómeno es el que se denomina variabilidad intra-locutor y es uno de los problemas a solucionar en verificación de locutor.

Con el fin de compensar el efecto degenerativo propio de la variabilidad intra-locutor es que se desarrollo el método de normalización T-Norm (del inglés *Test Normalization*). Su aplicación es similar a la realizada en H-NORM, cambiando solamente los términos normalizadores, tal como se muestra en la ecuación 2.35, donde llr_i es el *ratio* o *score* de decisión normalizado para el locutor que clama tener la identidad i, ll es la verosimilitud logarítmica del mismo usuario y mu^T y σ^T términos normalizadores propios de la identidad clamada i.

$$llr_i = \frac{ll_i - \mu^T}{\sigma^T} \tag{2.35}$$

La obtención de μ^T y σ^T se realiza evaluando la señal proveniente del locutor a verificar en varios modelos de usuarios no afiliados al sistema (*cohorts*), utilizando solo los M mayores *scores* y calculando con ello la media mu^T y la desviación estándar σ^T. Como el cálculo de los términos normalizadores se realiza con la señal proveniente del usuario a verificar, los *scores* de *cohorts* deben ser determinados en ese momento, por lo que esta técnica implica una cantidad considerable de tiempo computacional extra.

2.8. Conclusiones

En el presente capítulo se presentaron las bases teóricas necesarias para abordar los temas relacionados con la verificación de locutor texto dependiente operada sobre la red telefónica y como dicha tarea se ve afectada por las condiciones adversas que son introducidas al operar bajo condiciones reales, es decir, introduciendo al sistema ruido aditivo y distorsión de canal. Se explicó de esta forma la manera en que la señal de voz es parametrizada y como opera el motor verificador de locutor mediante la modelación estocástica de los HMM. Posteriormente se explica que es la distorsión de canal y como es modelado dicho efecto, para luego presentar al lector una serie de técnicas presentes en la literatura capaces de tratar el problema planteado.

Capítulo 3

Transformación de Parámetros Espectrales para Robustez a Canal

3.1. Introducción

La inconsistencia de canal entre las etapas de entrenamiento (enrolamiento) y de verificación (*test*) conocida también como *mismatch* de canal es el enfoque central de este trabajo. Con el fin de compensar el efecto producido por el canal es que en el siguiente capítulo se desarrolla una transformación de parámetros basada en el filtrado de estos en el espacio del espectro de la log-energía del banco de filtros Mel, logrando mejoras cercanas al 9 %. Para ello se comienza explicando el espacio de acción de la transformada denominado espectro de la MFBLE (del inglés *Mel Filter Bank Log Energy*). Luego se muestra la estructura de una transformación en dicho espacio de forma matemática siguiendo a continuación el análisis de importancia relativa (Vuuren y Hermansky, 1998), procedimiento ideado con el fin estimar una transformación adecuada. Finalmente se presenta la transformación desarrollada para luego da pasos a los experimentos realizados, sus resultados y la correspondiente discución y conclusiones.

3.2. El Espectro de la Log-Energía del Banco de Filtros Mel (MFBLE)

Es común y ampliamente utilizado en la literatura, una parametrización en el dominio cepstral. Es ahí donde se aplican técnicas ya mencionadas como CMN y otro tipo de filtrado de trayectorias temporales. En este trabajo se propone un espacio distinto para la transformación de parámetros denominado *Espectro de la Log-Energía del Banco de Filtros Mel* (MFBLE). Dicho espacio es capaz discriminar mejor la información propia del habla (identidad del locutor por ejemplo) y la información no relevante o basura (ruido).

Recordando secciones anteriores, el banco de filtros Mel espaciados es una estimación del espectro de una señal. Luego, el cálculo o aplicación de logaritmo a las energías obtenidas de dicho banco logra separar en forma aditiva la señal del canal. Si suponemos que el ruido aditivo ya ha sido extraído, a la salida de este procedimiento se tendrá lo mostrado en la ecuación 3.1, donde $\log(Y_t)$ es el vector de log-energías del banco de filtros Mel para el frame t de la señal corrompida por la distorsión de canal, $\log(S_t)$ es el vector de log-energías del banco de filtros Mel para la señal limpia S en el frame t y $\log(H_t)$ la distorsión de canal en el frame t en el espacio MFBLE. Sin embargo, la suposición de que el filtro H, representante del canal, es invariante en el tiempo no tiene porque ser cierta, y de hecho en gran medida no lo es. Así, el supuesto *bias* representante del canal en realidad no es tal y por ende se deben buscar formas alternativas de compensar el efecto.

$$\log(Y_t) = \log(S_t) + \log(H_t) \tag{3.1}$$

Se define como espectro de la MFBLE a la aplicación de la transformada discreta de Fourier (DFT del inglés *Discrete Fourier Transform*) a Log-Energía del banco de filtro Mel, espacio en el cual canal y señal presentan una mayor discriminabilidad. Con esto la señal quedaría representado el espectro de la MFBLE tal como se muestra en la ecuación 3.2, donde el superindice l indica la aplicación de logaritmo, es decir $\log(Y_t) = Y_t^l$.

$$DFT\left\{Y_t^l\right\} = DFT\left\{S_t^l\right\} + DFT\left\{H_t^l\right\} \tag{3.2}$$

De esta forma, el procedimiento para obtener dicho espacio es:

- Estimación el espectro de cada *frame* de la señal mediante un banco de filtros Mel-

espaciados
- Aplicación de logaritmo
- Cálculo de la DFT a cada *frame*

3.3. Transformación en el Espacio MFBLE

Tal y como se procede en RASTA, la transformación en el espacio MFBLE responde a la forma de un filtro pasa banda. Sin embargo difiere de esta última técnica en que no es aplicado en el dominio de la transformada Z sino en el espacio definido anteriormente. El transformar en este espacio esta definido de la siguiente forma.

Sea G el filtro pasa banda de orden K con el cual se transforman los parámetros y sea Z_t el vector de parámetros de orden K en el espacio del espectro de la MFBLE para el frame t, donde Z_t es igual a $DFT\left\{Y_t^l\right\}$ visto en la ecuación 3.2. Aplicar la transformada es simplemente ponderar punto a punto al filtro con los parámetros tal como se muestra en la ecuación 3.3, donde $\hat{Z}_t$ representan a los parámetros transformados en el espectro de la log-energía del banco de filtros Mel en el frame t.

$$\hat{Z}_t[k] = Z_t[k]\, G[k], \quad 1 \leq k \leq K \tag{3.3}$$

Luego, con los parámetros ya transformados $\hat{Z}_t$, es solo cosa de aplicar DFT inversa y luego estimar los coeficientes MFCC. Dicho procedimiento puede ser simplificado y dado que la trasformación es estimada con anterioridad y no al momento de la verificación, se puede reescribir el procedimiento como se indica en la ecuación 3.4 aplicando convolución circular, donde g es la transformada discreta de Fourier inversa de orden K para el filtro G y $\hat{Y}_t^l$ los parámetros transformados en la log-energía del banco de filtros Mel para el frame t.

$$\begin{aligned} \hat{Y}_t^l &= g \circledast Y_t^l \\ g &= DFT^{-1}\{G\} \end{aligned} \tag{3.4}$$

3.4. Definición de la Transformación

El algoritmo de robustez propuesto en este trabajo, como se menciona en secciones anteriores, será aplicado en los parámetros del espectro de la log-energía del banco de filtro Mel *frame* a *frame* y por tal razón, la dimensión de la transformación debe ser la misma que posee la parametrización de cada uno de dichos *frames*. Esta transformación se define como un filtro pasa bandas aplicado en el espectro discreto de la log-energía de la MFBLE, de esta forma, el filtro DFT G, estará completamente determinado si se conoce su frecuencia de corte inferior BF, la frecuencia de corte superior AF y dos ganancias denominadas BFG y AFG aplicadas en las secciones pasa bajos y pasa altos respectivamente. En la figura 3.1 se muestra la típica estructura de un filtro pasa banda propio de esta aplicación. Respetando el teorema del muestreo de Nyquist, en dicho filtro discreto debe existir simetría en su estructura, luego, si en el dominio MFBLE se tiene un vector de parámetros de dimensión K, entonces en el espectro de dicho dominio, el filtro DFT definido queda totalmente determinado con solo $K/2+1$ puntos, ya que los restantes corresponden a un reflejo horizontal de los otros puntos ya definidos. Así, las frecuencias de corte BF y AF cumplen que $1 \leq BF \leq K/2$ y $1 \leq AF \leq K/2$ y el filtro $G[k]$ en el dominio espectral de la MFBLE queda definido formalmente según la ecuación 3.5.

$$G[k] = \begin{cases} BFG & \text{si } k < BF \\ 1 & \text{si } BF \leq k \leq AF \\ AFG & \text{si } AF < k \end{cases} \tag{3.5}$$

3.5. Análisis de Importancia Relativa

Para calcular un filtro G que sea útil en la transformación definida en la ecuación 3.4 se utiliza en este trabajo el llamado análisis de importancia relativa. Este técnica diseñada originalmente para aplicarse a reconocimiento de voz consiste en realizar un estudio de que partes de la información son más sensibles al canal y cuales otras preservan en su mayoría características del habla. En este caso la información corresponde al espectro de la MFBLE y las partes de dicha información vienen a ser porciones de dicho espectro, por lo que se desea determinar mediante este análisis es que partes del espectro de la MFBLE son más sensibles a los efectos del canal variante en el tiempo y cuales otras no lo son y por ende reflejan en su mayoría información del habla. Es su versión original, el análisis

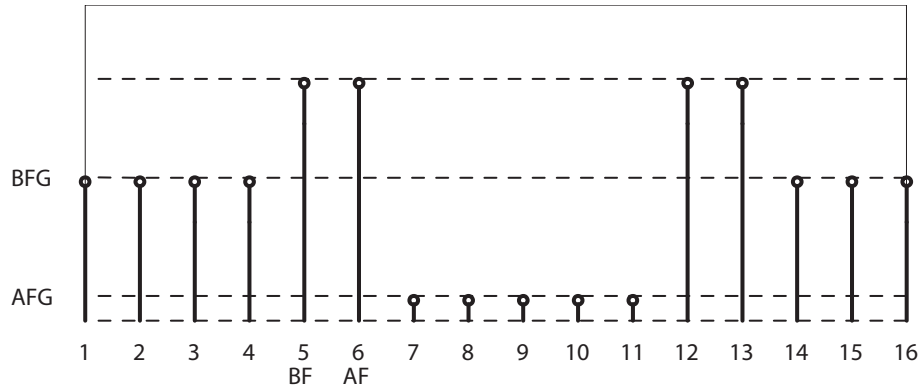


Figura 3.1: Estructura general para un filtro de orden 16 en el dominio del espectro de la log-energía del banco de filtros Mel, donde BFG es la ganancia para k menor a la frecuencia de corte inferior BF y AFG es la ganancia para k mayor a la frecuencia de corte superior AF.

de importancia relativa utilizaba el EER como medida discriminadora. En este trabajo se propone una función discriminante J basada en la variabilidad intra e iter-locutor en lugar del EER del sistema. Para calcularla se utiliza una base de datos de 40 usuarios grabados cada uno en tres canales distintos de donde se genera un modelo GMM para cada uno de los distintos locutores en cada uno de los canales. De esta forma la variabilidad intra-locutor se relaciona con la variabilidad que posee el locutor bajo tres canales distintos mientras que la inter-locutor lo hace con la variabilidad que presente un locutor frente a otros locutores.

$$J = \frac{1}{P}\sum_{p=1}^{P}\frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{s=1}^{S}\sigma_{s,p}^{2}[n]}{\sum_{s=1}^{S}\left(\mu_{s,p}[n]-U_{p}[n]\right)^{2}} \tag{3.6}$$

En la ecuación 3.6 se muestra la forma de la función discriminante J, en la cual $\sum_{s=1}^{S}\sigma_{s,p}^{2}[n]$ representa la variabilidad intra-locutor y $\sum_{s=1}^{S}\left(\mu_{s,p}[n]-U_{p}[n]\right)^{2}$ la inter locutor, $\mu_{s,p}[n]$ y $\sigma_{s,p}^{2}[n]$ son la media y la varianza[1] de la característica n en la gaussiana p del locutor s, P es el número de gaussianas, S el número de locutores y $U_p[n]$ el promedio de una característica n en una gaussiana p para todos los usuarios. Esta definición puede verse como un *ratio* entre la variabilidad intra-clase y la inter-clase muy semejante al análisis de discriminante de Fisher.

Para determinar la importancia relativa de las distintas bandas del espectro de la

[1]Se utiliza una covarianza del tipo matriz diagonal

MFBLE, se construye un set de filtros $G_{BF,AF}[k]$ tal como se definen en la ecuación 3.5 en los cuales se tendrá que BFG y AFG serán cero, $0 \leq BF \leq K/2$ y $0 \leq AF \leq K/2$, teniendo en cuenta que los casos en que $AF < BF$ no son interesantes y por ende no se calculan dichos filtros. Estimado el set de filtros se construye el set de filtros para el espacio MFBLE mediante la aplicación de la trasformada discreta inversa de Fourier explicada en la ecuación 3.4. Realizado esto, se aplican las transformaciones y para cada una de ellas se estima la función discriminante J, que en este caso pasa a tener una dependencia de las frecuencias de corte BF y AF, es decir, ahora la función discriminante puede escribirse como $J(BF, AF)$. De esta forma es claro porque los filtros en que AF $<$ BF no son relevantes.

Con el set de funciones discriminante $J(BF, AF)$ calculado, se define la medida importancia relativa $R(k)$ de la componente de frecuencia k en el espectro de la MFBLE según la ecuación 3.7, donde $0 \leq k \leq K/2$.

$$R(k) = \frac{2}{K}\left(\sum_{BF=0}^{k-1}\left(J(BF,k) - J(BF,k-1)\right) + \sum_{AF=k+1}^{K/2}\left(J(k,AF) - J(k+1,AF)\right)\right) \tag{3.7}$$

Esta cantidad $R(k)$ es quien muestra que tan importante es cada banda del espectro de la MFBLE para que el *ratio* definido en J de cuenta de una buena separabilidad entre clases y poca variabilidad en las clases.

3.6. Experimentos

Para probar el desempeño del sistema frente a la aplicación de la transformada propuesta, se utiliza en este trabajo el sistema verificador de locutor texto dependiente desarrollado en el laboratorio de procesamiento y transmisión de voz de la Universidad de Chile (LPTV), el cual se basa en modelos del tipo HMM donde cada unidad fonética es modelada con tres estados de izquierda a derecha sin saltos entre ellos. Cada estado es modelado mediante una gaussiana multivariable donde la matriz de covarianza se considera diagonal. En la etapa de verificación se calculan scores normalizados para cada una de las elocuciones según lo mostrado en la sección 2.6.4.

La base datos utilizada para esta prueba es una versión telefónica de YOHO, en la

cual se utiliza un subconjunto de setenta usuarios dividida a su vez en cuarenta utilizados para entrenar modelos de impostor y treinta para la verificación, tal como se ve resumido en la tabla 3.1

	Número de Impostores	Usuarios para Verificar
Hombres	20	20
Mujeres	20	10
Total	40	30

Tabla 3.1: Estructura general de la base telefónica YOHO

Cada usuario posee 24 señales destinadas a entrenamiento y otras dieciséis para verificación de donde se seleccionan conjuntos de cuatro con tal de obtener cuatro sesiones de *test* para cada usuario.

Para obtener la variabilidad de canal se graba la base antes mencionada a través de siete distintos canales denominados hset1, hset2, etc. Cada canal difiere del otro solo en el *handset*, manteniendo constante el canal transmisor. La grabación se hace a una tasa de muestreo de ocho kHz y 16 bits por muestra. Luego, la base compuesta por siete tipos de canal, es subdividida en tres. El primer subconjunto, compuesto por los canales hset2, hset3 y hset4, agrupa las elocuciones de entrenamiento de los cuarenta usuarios utilizados para los modelos de impostor y es utilizado para la estimación de la función discriminante $J\left(BF, AF\right)$ y la importancia relativa $R\left(k\right)$. El segundo conjunto solo utiliza el canal hset1, del cual se utilizan las elocuciones de entrenamiento para los modelos de impostor utilizados en la normalización de verosimilitud del sistema. Finalmente, el conjunto tres agrupa los canales hset1, hset5, hset6 y hset7, usando en este caso las elocuciones de entrenamiento y *test* de los treinta usuarios a verificar. Para entrenar los modelos de locutor se utilizan las elocuciones de entrenamiento de hset1, mientras que para la verificación las elocuciones de *test* de hset1, hset2, hset3 y hset4.

El cálculo del desempeño del sistema se hace computando el EER, para lo cual es necesario estimar las curvas de falsa aceptación y falso rechazo. El FR es computado con 1920 elocuciones provenientes de los cuatro canales mencionados, 30 locutores por cada canal y 16 elocuciones por locutor. La FA se calcula mediante 20880 provenientes de cuatro canales, 30 locutores, 29 impostores y seis elocuciones por impostor.

Con esto, el sistema entrega un *baseline* de 2,71 % bajo condiciones *matched* (entrenamiento y *test* con hset1) y 3,99 % cuando se utiliza toda la base de datos (hset1, hset5, hset6 y hset7).

3.7. Resultados

Para evaluar la respuesta del sistema verificador de locutor desde el enfoque de EER, se realizan seis experimentos distintos, de los cuales se pueden extraer las mejoras obtenidas en el desempeño gracias a la técnica propuesta y su posterior mezcla con métodos clásicos de la literatura. Los seis escenarios son:

- *Baseline*
  Se evalúa el desempeño sin aplicar transformación alguna
- Transformada
  Se evalúa el desempeño aplicando la transformada propuesta
- CMN
  Se evalúa el desempeño aplicando CMN a los parámetros
- Transformada y CMN
  Se evalúa el desempeño aplicando la transformada y posteriormente CMN a los parámetros.
- RASTA
  Se evalúa el desempeño aplicando RASTA a los parámetros.
- Transformada y RASTA Se evalúa el desempeño aplicando la transformada y posteriormente RASTA.

Utilizando el primer subconjunto de canales explicado en la sección anterior, es decir, hset2, hset3 y hset4, se calcula la función discriminante J para casos en que se aplica a los parámetros CMN y RASTA además de un caso para parámetros sin aplicación de técnica alguna. Este resultado puede visualizarse como un manto tridimensional tal como se muestra en la figura 3.2, en el cual se puede notar la existencia de puntos (bandas del espectro) en que la discriminabilidad es mayor que para el caso base (sin aplicar transformación o para BF $= 0$ y AF $= 8$) cosa que sugiere que bandas del espectro de la MFBLE pueden ser eliminadas y con ello aumentar el desempeño del sistema.

Con la función J calculada, se estima la importancia relativa R, de la cual se desprende un estudio completo de la importancia de las distintas bandas del espectro de la MFBLE. Dicha función es ilustrada en la figura 3.3, en la cual se muestra la importancia relativa de las bandas para los parámetros transformados aplicando CMN, RASTA y sin aplicación de técnica alguna. Es claro que secciones del espectro de la MFBLE son menores a otras e incluso negativas, dejando de manifiesto su bajo aporte de información relevante. De

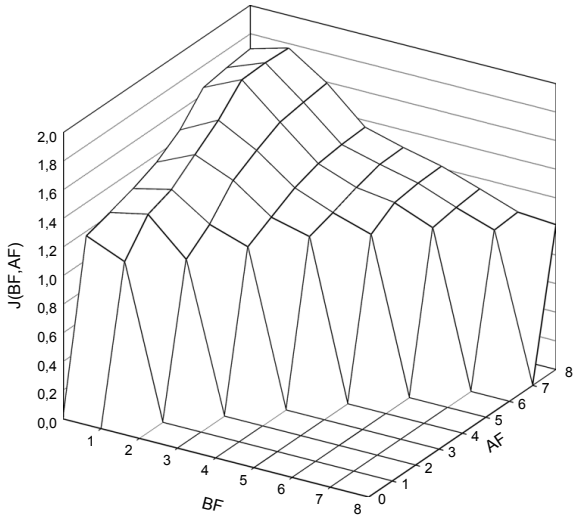


Figura 3.2: Visualización de la función $J\,(BF, AF)$.

aquí se desprende que las frecuencias de corte BF y AF sean $k = 1$ y $k = 6$ aplicando o no CMN y $k = 0$ y $k = 6$ aplicando RASTA.

Con todo lo anterior se define la transformación G según la tabla 3.2 para los casos mencionados anteriormente (según la aplicación o no de CMN o RASTA). Se debe notar como en el último caso (transformación y RASTA), la frecuencia de corte para la sección pasa alta no tiene sentido pues la frecuencia de corte BF es cero y por ende la transformación toma la forma de un filtro pasa bajos.

	Frecuencia BF	Frecuencia AF	Ganancia BFG	Ganancia AFG
Transformada sin CMN	1	6	0,4	0
Transformada con CMN	1	6	0,8	0
Transformada con RASTA	0	6	-	0

Tabla 3.2: Definición del transformación G para el sistema de prueba

De esta forma y recordando los seis escenarios a los que se somete el sistema verificador de locutor, los resultados del desempeño medido mediante el EER son los mostrados en la tabla 3.3, destacando el hecho de que la aplicación de la transformada propuesta introduce mejoras en el sistema para el caso en que es aplicada de forma aislada como también cuando es mezclada con CMN y RASTA.

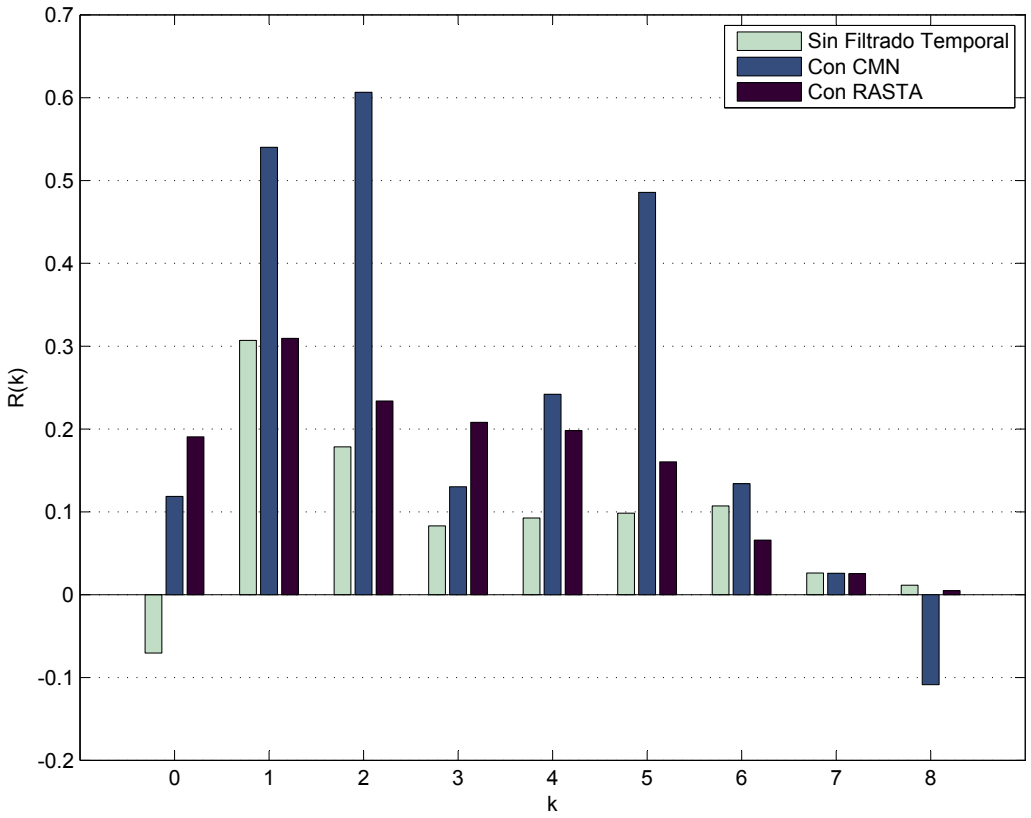


Figura 3.3: Importancia relativa $R(k)$ del espectro de la MFBLE. Las barras negras muestran el resultado sin CMN y las barras blancas con CMN.

	Baseline	Transf.	CMN	Transf. y CMN	RASTA	Transf. y RASTA
EER (%)	4,02	3,65	2,6	2,35	3,27	3,13

Tabla 3.3: Resultados del Sistema en sus 6 escenarios posibles

3.8. Discusión

Como se ve en el análisis hecho mediante el cálculo de la función discriminante J, el eliminar ciertas bandas de frecuencia permite al sistema, enfrentado a un escenario de *mismatch* de canal, aumentar su capacidad de separar entre clases (cliente-impostor) basado en el hecho de que la función J toma valores superiores al obtenido en el caso *baseline*. Esto se ve de forma muy clara en la figura 3.2. Del mismo modo, al observar los resultados obtenidos en el cálculo de la importancia relativa R, se puede notar como las bandas mas bajas y altas carecen de importancia frente a frecuencias centrales, reafirmando así el hecho de que ciertas bandas del espectro de la MFBLE reflejan en su mayoría información de la variabilidad de canal mientras que otras presentan información de la voz y del locutor en su mayoría, con lo cual se puede establecer un filtro adecuado para los casos de interés. Es interesante notar por ejemplo que al aplicar CMN, la banda baja del espectro de la MFBLE aumenta su importancia relativa debido a que esta técnica disminuye el efecto de canal presente en esta banda en el espectro del dominio expuesto. Por último, en una mirada al desempeño del sistema frente a la aplicación de la transformación expuesta, es notable como se logra una mejora de 9,2 % al aplicar solo la transformación respecto al caso *baseline*, un 9,61 % al aplicar la transformación de parametros junto a CMN respecto a la aplicación de solamente CMN y un 41,5 % al aplicar ambas técnicas respecto al *baseline*. Las mejoras al combinar la transformación con el filtrado temporal RASTA son menores y solo alcanzan el 4,28 % al compararlo con la aplicación de RASTA de forma aislada. Estos resultados pueden verse gráficamente en la figúra 3.4, en donde se muestran las curvas DET para los seis escenarios antes explicados.

3.9. Conclusiones

Se presentó en las secciones anteriores todo un marco de procedimientos con el fin de atenuar el efecto degenerativo que tiene el canal sobre la verificación de locutor. Para ello se propuso un nuevo espacio de trabajo denominado espectro de la log-energía del

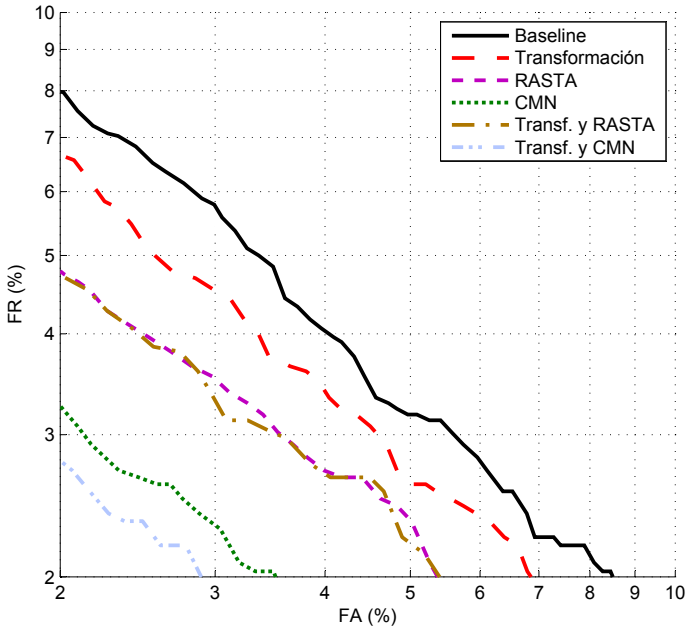


Figura 3.4: Curva DET para los seis escenarios posibles.

banco de filtros Mel, sobre el cual se llevo a cabo una transformación de parámetros de la forma de un filtro pasa bandas. Los resultados son alentadores, resaltando el 9,61 % de mejoras a nivel de EER logrado al combinar la técnica propuesta con CMN. Además es importante destacar la generalización del análisis de importancia relativa desarrollado aquí, el cual logra demostrar que la información presentada en el espacio propuesto se concentra principalmente en bandas centrales del espectro, dejando en los bordes casi en su totalidad la acción del canal, razón por la cual pueden ser eliminadas y con ello mejorar el desemepeño del sistema.

Capítulo 4

Comparación y Combinación de Técnicas de Transformación de Parámetros y Normalización de Scores

4.1. Introducción

En este capítulo, se utiliza la transformación de parámetros propuesta en el capítulo anterior en una base de datos distinta. Esta base, a diferencia de la mostrada anteriormente (versión telefónica de YOHO), presenta *mismatch* de canal tanto en el *handset* como en el canal transmisor por lo que la distorsión de canal es más severa y por ende es más difícil de cancelar. Además, se compara y combina la transformada propuesta con un filtrado de trayectorias temporales desarrollado como parte de este trabajo. Por último, se propone además una normalización y compensación de *scores* no supervisada, la cual es puesta a prueba en una base de datos con tres tipos de canal distintos. Con ello se pretende discutir y concluir frente al tema central de este desarrollo que es el efecto degenerativo del canal.

Para estructurar este capítulo se comienza explicando sobre la selección de género, técnica complementaria desarrollada para el ahorro de tiempo computacional, luego se explica el tema del filtrado de trayectorias temporales y como esta técnica es complementaria a la transformación de parámetros propuesta en el capítulo anterior, logrando

mejoras significativas en el sistema. Con esto se presentan y discuten los resultados obtenidos con las técnicas propuestas. Finalmente se plantea un tipo de normalización similar a H-NORM y T-NORM implementando una técnica de selección de canal sin supervisión, complementando esto con una compensación de *scores* basado en la selección antes explicada, presentando enseguida los resultados pertinentes y las conclusiones que se extraen de este capítulo.

4.2. Selección de Género Basado en HMM y Parámetros MFCC

Hoy en día, una de las formas de normalizar la verosimilitud de un sistema verificador de locutor es mediante el cálculo de *scores* en modelos de impostor. Este modo de operar, tal como se menciona en capítulos anteriores, presenta uno de los mejores resultados existentes en la literatura. Sin embargo acarrea un gran problema de retardo pues agrega como tarea al sistema la evaluación en numerosos modelos adicionales al modelo de locutor clamado. Por esta razón, escoger de forma inteligente que impostores utilizar con el fin de normalizar la verosimilitud es una tarea importante a realizar con el objetivo de ser más eficiente en procesamiento de información. Con este fin es que se implementa una clusterización supervisada basada en una selección de género.

Tal como se ha hecho en trabajos anteriores, se implementa un clasificador de género basado en la utilización de HMM (Parris y Carey, 1996), a diferencia de otros clasificadores bayesianos basados en el *pitch* (frecuencia fundamental del habla) del locutor (Ting *et al.*, 2006),característica muy poco robusta dada la facilidad con que puede ser modificada a gusto por el locutor, en el clasificador propuesto se utilizan como parámetros los coeficientes MFCC. Esto implica que el clasificador estará tomando decisiones según características que dan cuenta del tracto vocal y no de la excitación, dándole robustez a dicho clasificador ante eventuales intentos de engaño mediante la modificación del *pitch* del locutor. Además de lo anterior, la estructura de tipo HMM, permite aprovechar características propias de las elocuciones como son las unidades fonéticas implicadas. La toma de decisión en el clasificador está estrictamente basado en la evaluación de los parámetros del locutor en los modelos de género y con ello las verosimilitudes correspondientes. El clasificar como hombre o mujer a un locutor simplemente se realiza viendo cual verosimilitud es mayor. Este proceder es el que se formaliza en la ecuación 4.1, donde λ_{genero} corresponde al modelo de género, O los parámetros del locutor, ll_{hombre} la verosimilitud logarítmica relacionada

a la clase hombre y ll_{mujer} la verosimilitud logarítmica relacionada a la clase mujer

$$\begin{aligned} ll_{hombre} &= \log\left(P\left(O\,|\lambda_{hombre}\right)\right) \\ ll_{mujer} &= \log\left(P\left(O\,|\lambda_{mujer}\right)\right) \\ \text{Si } ll_{hombre} > ll_{mujer} &\Rightarrow \text{ el locutor es hombre} \\ \text{Si } ll_{hombre} < ll_{mujer} &\Rightarrow \text{ el locutor es mujer} \end{aligned} \tag{4.1}$$

La resolución de este problema es exactamente igual al que se utiliza para calcular los *scores* de locutor relacionados con su identidad, es decir, mediante el alineamiento de Viterbi se encuentra la secuencia de estados más probable y con ella se calculan las verosimilitudes relacionadas al género.

Se sabe que el número de modelos de impostor a evaluar es C, de los cuales, en una configuración común, $C/2$ serán de cada género, por lo cual la reducción de tiempo de procesamiento debería estar cercana al 50 %. En el caso de este trabajo dicha disminución ronda el 60 %.

Otras formas de realizar la selección de género es mediante clasificadores basados en mezclas de modelos Gaussianos (Zeng *et al.*, 2006), máquinas de soporte vectorial (Bocklet *et al.*, 2008), entre otras técnicas de clasificación de patrones.

4.3. Filtrado Temporal de la DFT de los Parámetros Espectrales

El filtrado de trayectorias temporales no es un tema novedoso, numerosas técnicas en la literatura realizan dicho procedimiento con relativo éxito según la aplicación y condiciones en que son utilizadas. En este trabajo se propone un nuevo tipo de filtrado de trayectorias temporales realizado en el espectro de la trayectoria temporal del espectro de la log-energía del banco de filtro Mel y aplicado a condiciones bastante adversas en cuanto a distorsión de canal se refiere.

Tal y como se define en el capítulo anterior, sea $Y_k^l\,[t]$ la log-energía k obtenida a partir del banco de filtros Mel para un *frame* t, es decir, el parámetro k para un tiempo t en la MFBLE de determinado locutor. Se define como el espectro de la trayectoria temporal del espectro de la log-energía del banco de filtros Mel al cálculo de la transformada discreta

de Fourier a través de los *frames* al espectro de la MFBLE. Dicho procedimiento es el que se formaliza a través de la ecuación 4.2.

$$\begin{aligned} Z_k^{DFT}[t] &= DFT\{Z_k[t]\}, \quad 1 \leq t \leq T \\ Z[t] &= DFT\left\{Y^l[t]\right\} \end{aligned} \tag{4.2}$$

A diferencia de la transformación antes propuesta, es decir en el espectro de la MFBLE, en donde se transformaban los parámetros a través de un filtro pasa bandas aplicado *frame* a *frame*, el filtrado de trayectorias temporales es realizado en un parámetro fijo k a través de los *frames*. Con ello se define la transformación propuesta H como un filtro pasa bandas (Jung y Lee, 2000) aplicado a la variación temporal de un determinado parámetro en el espacio del espectro de la trayectoria temporal del espectro de la MFBLE. En la ecuación 4.3 se explica como la transformación H es aplicada, donde $\hat{Z}^{DFT}[t]$ son vectores de parámetros transformados a través del tiempo.

$$\hat{Z}_k^{DFT}[t] = Z_k^{DFT}[t]\, H_t, \quad 1 \leq t \leq T \tag{4.3}$$

Al igual que en procedimientos anteriores, la ecuación 4.3 puede ser reescrita mediante convolución circular de forma de evitar el cálculo de la DFT tal como se nota en la ecuación 4.4.

$$\begin{aligned} \hat{Z}_k &= Z_k \circledast h \\ h &= DFT^{-1}\{H\} \\ \hat{Z}_k[t] &= DFT^{-1}\left\{\hat{Z}_k^{DFT}[t]\right\} \end{aligned} \tag{4.4}$$

Luego, es factible regresar al espacio MFBLE mediante la aplicación de la DFT inversa a cada uno de los frames de la señal con lo cual finalmente es posible estimar los coeficientes MFCC que son los parámetros con los que funciona el sistema verificador de locutor utilizado en este trabajo. En la figura 4.1 se esquematiza el procedimiento para aplicar el filtrado de trayectorias temporales h, teniendo en cuenta que los parámetros para cada uno de los frames corresponden a vectores columna y que su evolución temporal se representa en un eje horizontal. De esta manera, el filtrado de trayectorias temporales propuesto se resume como la aplicación de la transformada discreta de Fourier a cada uno de los vectores columna de la matriz mostrada y posteriormente el filtrado (convolución circular) de los

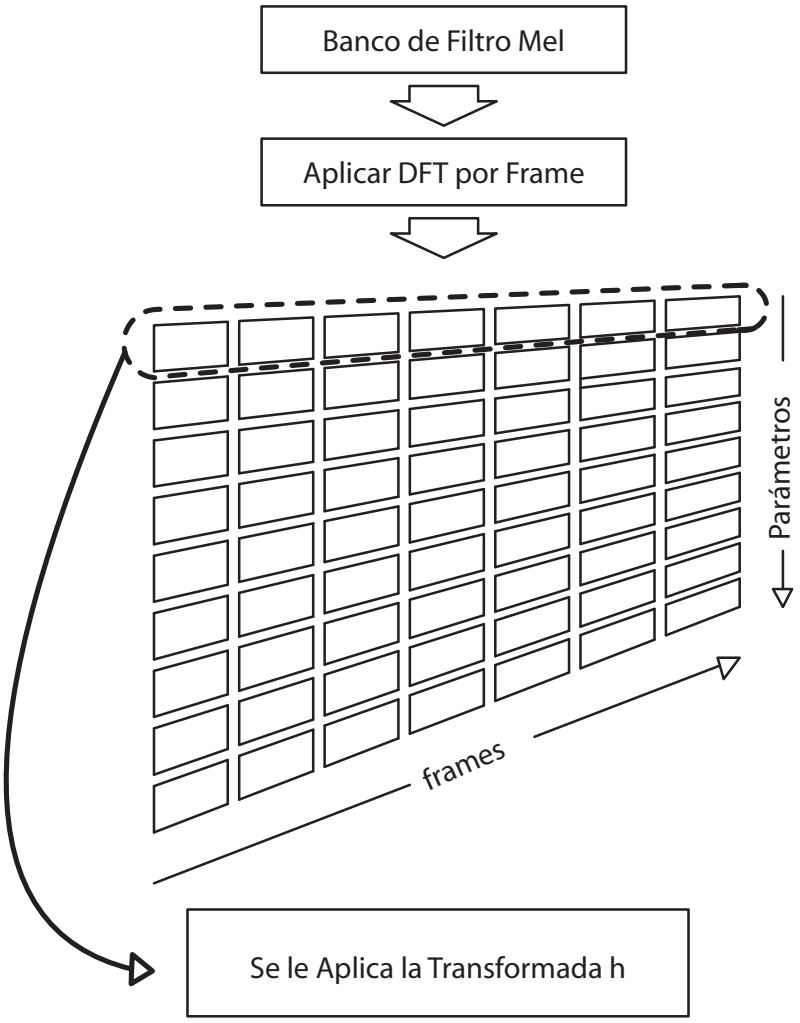


Figura 4.1: Esquema en bloques de la aplicación del filtro temporal h a los parámetros en el espacio de las trayectorias temporales del espectro de la MFBLE.

vectores fila.

Tal como se ha anticipado, e inspirado en el trabajo realizado en el filtro RASTA, el filtro de trayectorias temporales propuesto en este trabajo H responde a la forma de un filtro pasa bandas con frecuencia de corte inferior BF y superior AF, donde las bandas inferior y superior presentan ganancias BFG y AFG respectivamente. De esta forma, son estos los parámetros a estimar con el fin de definir el filtro H. En la ecuación 4.5 se formaliza el filtro H, donde la cantidad T es la duración en *frames* de la señal.

$$H_t = \begin{cases} BFG & \text{si } 1 \leq t < BF \\ 1 & \text{si } BF \leq t \leq AF \\ AFG & \text{si } AF < t \leq T \end{cases} \tag{4.5}$$

Dada la naturaleza temporal del filtro H, es una dificultad definir las frecuencias de cortes para un t discreto fijo, pues la duración T de las señales, varia entre cada una de ellas. Es por ello que BF y AF son definidos en frecuencia normalizada (entre 0 y π) llevando a la duración T a la potencia de dos más cercana a la duración original con el fin de realizar el cálculo de la DFT mediante el algoritmo óptimo FFT (del inglés Fast Fourier Transform).

4.4. Experimentos

Los experimentos para la técnica propuesta en este capítulos son desarrollados sobre una base de datos grabada en el LPTV de la Universidad de Chile la cual será denominada ASTERISK LPTV. Dicha base esta compuesta por cuatro canales distintos que a diferencia de la versión telefónica de YOHO, presentan *mismatch* de canal tanto en el *handset* como en el canal transmisor. La composición de ella se detalla en la tabla 4.1.

Canal	Handset	Locutores	Impostores	Elocuciones de Entrenamiento	Elocuciones de Test
Fijo	Laboratorio	40	60	3	9
	Particular	40	-	3	9
Móvil	Samsung	20	-	3	10
	Particular	20	-	3	10
	Nokia	-	60	3	-

Tabla 4.1: Base de Datos ASTERISK LPTV

En la base de datos antes expuesta, el canal fijo se refiere al uso de telefonía analógica convencional, mientras que el canal móvil se refiere a la telefonía celular vía auricular. En ambos canales, el *handset* particular se refiere a que cada locutor utiliza su teléfono propio, luego, las condiciones de *mismatch* ocurren cuando el enrolamiento se realiza en un teléfono común a todos los locutores y la verificación en otro teléfono distinto al de entrenamiento y particular para cada locutor. El caso *matched* se produce cuando tanto el entrenamiento como la verificación es realizada con un teléfono común para todos los

locutores. Existe un tercer caso que se da cuando el entrenamiento y verificación se lleva a cabo con un teléfono particular en cada usuario pero distinto entre ellos, el que se denomina *matched-casa*.

En el canal móvil, el *handset samsung* corresponde a un teléfono celular marca samsung común a todos los locutores, el *handset laboratorio* del canal fijo corresponde a un teléfono convencional utilizado en el LPTV común también para todos los locutores (caso *matched* o de paridad de canal). El *handset* Nokia solo es utilizado para los modelos de impostor aplicados en el canal móvil. Cada señal de ASTERISK LPTV es grabada a una tasa de muestreo de 8 kHz y a 16 bits por muestra.

La composición por genero de la base expuesta es equitativa, teniendo un 50 % de locutores e impostores por género en cada uno de los canales. Así por ejemplo, el *handset laboratorio* para el canal fijo, tiene veinte locutores hombres y la misma cantidad de mujeres, además de 30 impostores hombres y la misma cantidad de impostores mujer.

El sistema verificador de locutor texto dependiente utilizado para estas pruebas es idéntico al utilizado en el capítulo anterior salvo el modelo *speaker independent* base con el cual se entrenan los modelos de locutor, el cual esta vez está adaptado al idioma español y por ende las unidades fonéticas correspondientes a dicho idioma. La estructura HMM utilizada sigue siendo la misma utilizada en el capítulo anterior (modelos de 8 estados sin saltos y con Gaussianas multivariables) al igual que la parametrización de las señales. Para calcular el desempeño del sistema, se ocupa el EER como indicador y con ello las curvas de falsa aceptación (FA) y falso rechazo (FR).

El cálculo de la curva FR se realiza utilizando dos sesiones de *test* por locutor y promediando los *scores* obtenidos. Para ello se realizan todas las combinaciones de pares posibles descartando los casos en que la elocución es la misma para cada sesión, evaluando a cada locutor en el modelo correspondiente (CLIENTE). Misma lógica es la que se utiliza en el cálculo de la curva FA, teniendo en cuanta eso sí, el hecho de que la evaluación de los locutores se hace en los modelos de otros (IMPOSTORES). En la tabla 4.2 se muestra la dimensión que tienen los experimentos de CLIENTE e IMPOSTOR en los canales fijo y móvil.

Tal como en la sección anterior, el filtrado de trayectorias temporales propuesto es contrastado con CMN y RASTA. Junto con esto, la transformación mostrada en el capítulo anterior es combinada con cada una de las técnicas de filtrado temporal mencionadas, evaluando el desempeño para cada uno de los tres escenarios explicados en la tabla 4.3.

Canal	Experimento	Locutores	Combinaciones	Total
Fijo	CLIENTE	2x20	27	1080
	IMPOSTOR	2x20x19	27	20520
Móvil	CLIENTE	2x10	40	800
	IMPOSTOR	2x10x9	40	7200

Tabla 4.2: Dimensión experimentos de CLIENTE e IMPOSTOR

Canal	Caso	*Handset* de entrenamiento	*Handset* de test
Fijo	*Matched*	Laboratorio	Laboratorio
	Unmatched	Laboratorio	Particular
	Matched-casa	Particular	Particular
Móvil	*Matched*	Samsung	Samsung
	Unmatched	Samsung	Particular
	Matched-casa	Particular	Particular

Tabla 4.3: Casos probados en los experimentos

4.5. Resultados

Como se menciona en el comienzo del presente capítulo, el objetivo central de esta parte del trabajo es lograr combinar la transformación de parámetros propuesta en el capítulo anterior con un filtrado de trayectorias temporales. Para ello se prueban ambas técnicas por separado y luego combinadas se muestra el desempeño que esta presenta. Para ello se aplica en todos los casos la selección de género evaluando así solo 30 modelos de impostor de los cuales se eligen los diez mayores.

Basado en el desempeño del sistema mediante la medición del EER, se determina que la transformación de parámetros óptima a aplicar en esta base de datos toma la forma de un filtro pasa bandas con frecuencias de corte BF= 0 (filtro pasa bajos) y AF= 7 con una ganancia de la sección alta de AFG= $0,8$, siendo esta la transformación es la que se combina con cada una de las técnicas de filtrado temporal probados en este trabajo. Por otra parte la estructura que toma el filtro de trayectorias temporales se determina de igual forma y teniendo en cuenta su forma de filtro pasa bandas, se estima que las frecuencias de cortes normalizadas son BF= 0,04 y AF= 3 con ganancias BFG=AFG=0.

Es importante recordar que el filtro de trayectorias temporales es expresado mediante frecuencias normalizadas. Esto pues la duración de las señales en *frames* T no es constante. En el caso de la base de datos utilizada a continuación la duración T no supera los mil

frames ni es inferior a los trecientos. Dado que el volver a los coeficientes MFCC implica un calculo de DFT inversa, se utiliza con el fin de optimizar dicho proceso el algoritmo FFT inverso. Para ello se debe tener un número de *frames* potencia de 2, por lo que se manejan en esta aplicación señales de largo 512 y 1024, rellenando con ceros los *frames* adicionales requeridos. De esta forma, solo es necesario almacenar dos filtros distintos que se adecuen al número de *frames* que le corresponda a la señal, respondiendo ambos de igual forma en frecuencia.

4.5.1. Resultados Canal Fijo

Para en canal fijo, se realizan las pruebas con las técnicas antes explicadas, cuyos resultados son expresados mediante el EER de la tabla 4.4, en la cual se expresa además el resultado para cada uno de los escenarios definidos.

	Matched	Unmatched	Matched-Casa
Baseline	4,74	18,81	1,55
Filtrado de Trayectorias Temporales	4,26	17,84	1,3
Transformación de Parámetros	4,73	18,26	1,57
CMN	5,07	16,92	3,24
RASTA	5,96	16,61	4,54
Transformación de Parámetros y Filtrado de Trayectorias	4,21	17,78	1,39
Transformación de Parámetros y CMN	4,91	17,11	3,24
Transformación de Parámetros y RASTA	6,37	16,58	4,66

Tabla 4.4: Resultados para el canal fijo

Dichos resultados pueden ser vistos en forma gráfica mediante las curvas DET de cada una de las técnicas propuestas tal y como se muestra en la figura 4.2.

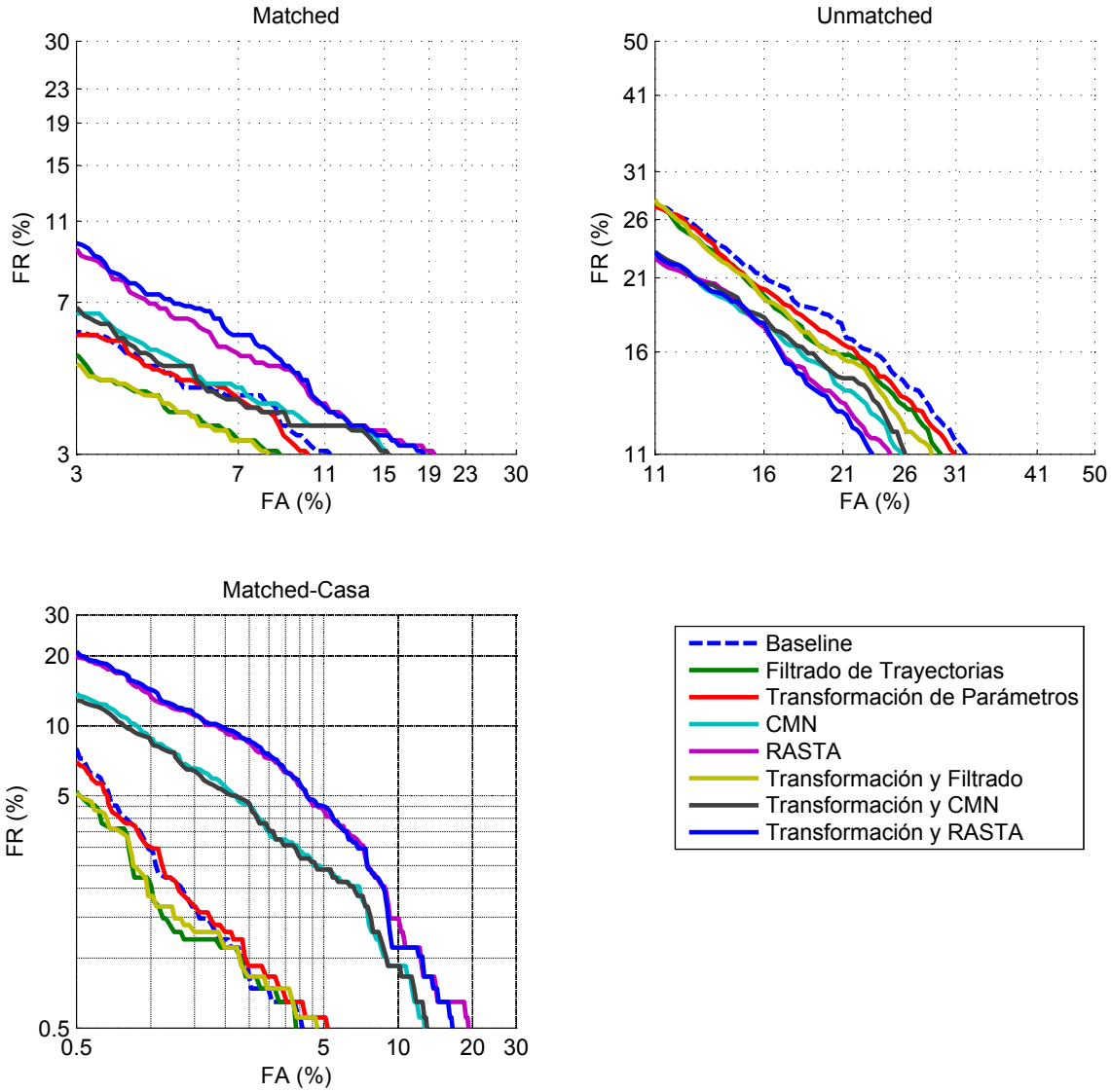


Figura 4.2: Curva DET para canal fijo.

4.5.2. Resultados Canal Móvil

Siguiendo la misma lógica aplicada para canal fijo, se calculan los desempeños del sistema frente a la aplicación de las diversas técnicas mostradas y descritas anteriormente. Así, los niveles de EER son presentados en la tabla 4.5, la cual presenta los resultados para los escenarios *Matched*, *Unmatched* y *Matched-Casa*.

	Matched	Unmatched	Matched-Casa
Baseline	2,55	10,44	2,75
Filtrado de Trayectorias Temporales	3	8,99	2,92
Transformación de Parámetros	2,69	10,53	2,75
CMN	6,22	8	4,12
RASTA	8,67	17,09	8,13
Transformación de Parámetros y Filtrado de Trayectorias	2,9	9,09	2,58
Transformación de Parámetros y CMN	6,18	7,39	4,05
Transformación de Parámetros y RASTA	7	15,63	8,75

Tabla 4.5: Resultados para canal móvil.

De igual forma a lo presentado anteriormente, los resultados pueden visualizarse mediante la estimación de las distintas curvas DET para cada uno de las técnicas propuestas. Dicho gráfico es el que se muestra en la figura 4.3

4.6. Discusión

Frente a los resultados obtenidos, no se debe perder de vista la motivación principal de este trabajo que es lograr darle robustez a la distorsión de canal al sistema verificador de locutor operado sobre una red telefónica, puesto que dado esa motivación, los escenarios que más interesan frente a una aplicación comercial son sin duda alguna los estados *Matched-Casa* y *Unmatched*, pues son estos los casos que se dan de forma natural. De esta forma es importante obtener mejoras transversales en esos escenarios, condición que

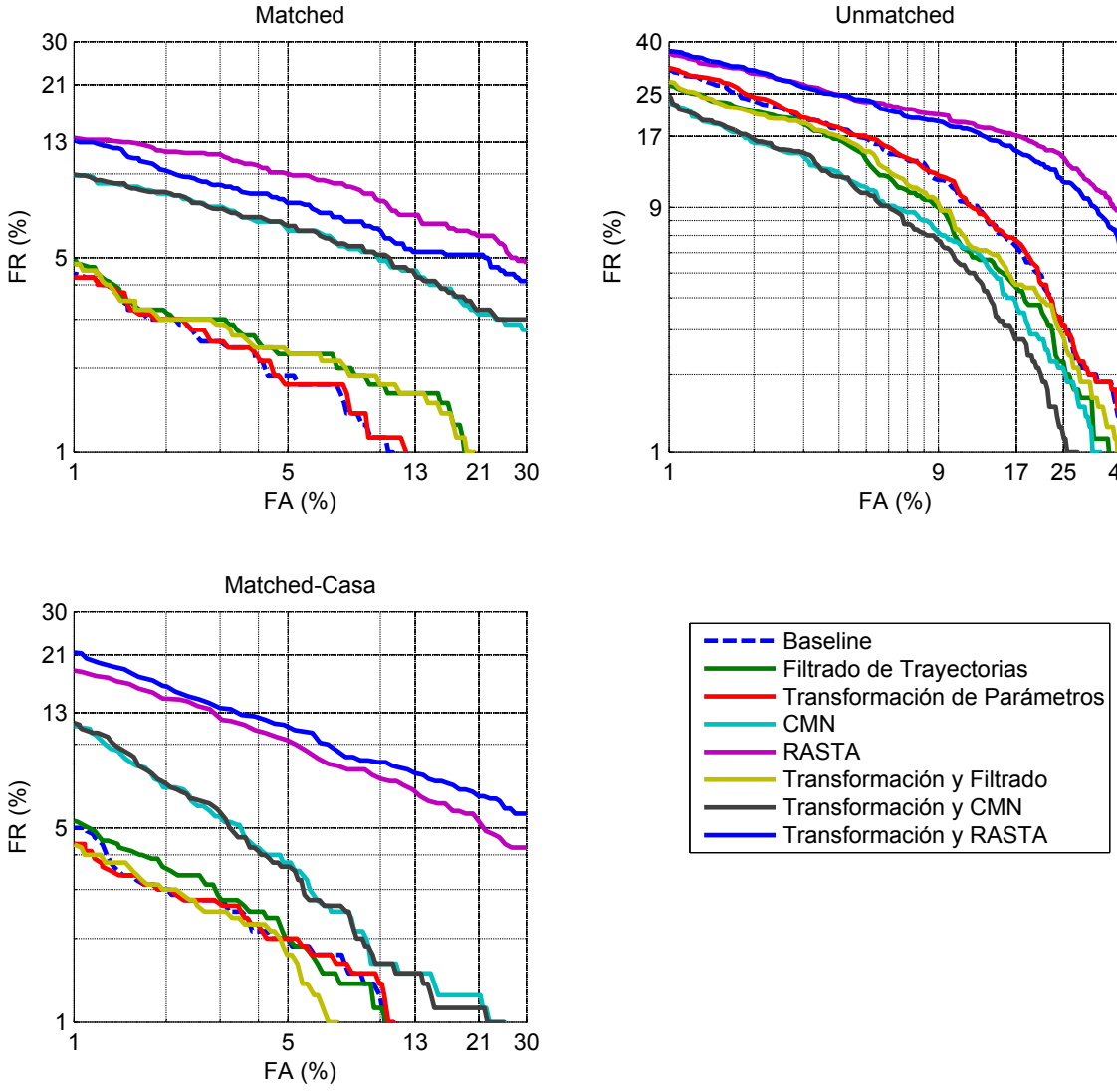


Figura 4.3: Curva DET para canal móvil.

solo se logra mediante la aplicación del filtrado de trayectorias temporales propuesto y la combinación de dicho filtrado con la transformación de parámetros ya explicada, ambos para canal fijo. En canal móvil, las mejoras deseadas solo se logran con la combinación de las técnicas ya mencionadas.

4.6.1. Discusión Canal Fijo

Las mejoras respecto del *baseline* para la aplicación del filtrado de trayectorias temporales son de un 16,1 % y 5,2 % para los casos *Unmatched* y *Matched-Casa* respectivamente, mientras que para las combinación del filtrado de trayectorias con la transformación de parámetros las mejoras son de 10,4 % y 5,5 % para los mismos casos.

Independiente de lo anterior, es importante analizar también los sucedido en el escenario *Matched* ya que de forma clara se ve como la aplicación de algunas técnicas disminuye el desempeño del sistema. Tal es el caso de CMN y RASTA, procedimientos que eliminan distorsión de canal junto con información de locutor (uno en el espacio de la MFBLE y el otro en los coeficientes MFCC), la cual frente a condición de paridad de canal (*Matched*) toma especial importancia y por ende su eliminación deriva en bajas en el rendimiento que van desde el 7 % al 25,7 % para CMN y RASTA respectivamente. Misma situación ocurre para el caso *Matched-Casa*, donde la importancia de la información de locutor eliminada por CMN y RASTA también es importante para el desempeño del sistema.

Por último, la combinación de las diversas técnicas de filtrado temporal con la transformación de parámetros en el espectro de la MFBLE puede ayudar a mejorar en el caso *Matched* y *Matched-Casa* tal y como ocurre con CMN, donde el EER sube de 5,07 % a 4,91 %. Sin embargo, dicha mejora no es suficientemente importante como para mejorar el *baseline* y por ende no son relevante dichas combinaciones.

4.6.2. Discusión Canal Móvil

Para el canal móvil, solo la combinación de la transformación de parámetros en el espectro de la MFBLE con el filtrado de trayectorias temporales entrega resultados positivos de forma transversal para los casos denominados de importancia para un sistema aplicado sobre la red telefónica. Las mejoras obtenidas son 6,2 % y 12,9 % para los casos *Matched-Casa* y *Unmatched*. Para el caso *Matched*, las técnicas de filtrado temporal introducen en el sistema una degradación en el desempeño, el cual se explica por las mismas razones

dadas para el canal fijo.

Con estos resultados y dado el deseo de operar el sistema para telefonía tanto fija como móvil, es claro que la combinación de la transformación de parámetros en el espacio del espectro de la MFBLE con el filtrado de trayectorias temporales otorga un marco de funcionamiento capaz de atenuar los efectos degenerativos del canal.

4.7. Normalización de Canal no Supervisada

Numerosas son la técnicas de normalización disponibles en la literatura, de las cuales ya se han explicado en el presente trabajo dos de ellas. Se propone a continuación una versión no supervisada de la técnica conocida como HT-Norm (Bimbot *et al.*, 2004), técnica muy similar a T-Norm pero que aprovecha el conocimiento del canal para seleccionar los modelos adecuados con los cuales se calculan los *scores* normalizadores (cohort adecuado). En ese sentido, este trabajo propone un clasificador de canal con el cual seleccionar los modelos adecuados para la normalización de *score*.

Como se ha visto es secciones anteriores, la distorsión de canal es un problema crítico en la verificación de locutor, esto pues se pueden presentar inconsistencias entre el modelo entrenado y los parámetros de *test* obtenidos para la verificación. Sin embargo, además de lo anterior, existe una dificultad extra que es la variabilidad de los *scores* producto de los distintos canales y por ende el umbral de decisión no es capaz de discriminar de forma adecuada. Así por ejemplo, *scores* provenientes de telefonía celular y otros provenientes de telefonía fija presentarán promedios muy distintos. En la figura 4.4 se muestra como los *scores* de CLIENTES poseen distinto promedio y varianza de acuerdo al canal. También se muestra una distribución correspondiente a la mezcla de todos los canales, la cual de forma esperada se ubica al centro de todas las demás.

Se plantea entonces como necesidad el que hayan condiciones *matched* no solo para las fases de entrenamiento y verificación, sino muy importante es también que los modelos de impostor sean consistentes con el canal en que se hace la verificación. Con este fin se define la técnica HT-Norm como la normalización de la verosimilitud mediante el calculo de una media normalizadora μ_{HT}. Es factible utilizar también una desviación estándar σ_{HT}, sin embargo, en este trabajo se ha probado que dicha aplicación disminuye el desempeño del sistema y por ende se descarta. En la ecuación 4.6 se muestra el procedimiento a realizar, donde llr_j es la verosimilitud logarítmica normalizada del usuario con identidad clamada

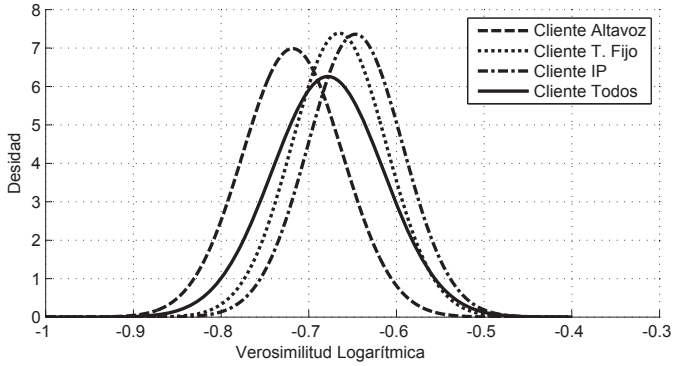


Figura 4.4: Distribución de los *scores* de CLIENTE para canales de telefonía fija, IP y móvil mediante altavoz.

j y ll_j es la verosimilitud logarítmica del usuario con la misma identidad clamada.

$$llr_j = ll_j - \mu_{HT} \tag{4.6}$$

El cálculo del término μ_{HT} se realiza evaluando la señal de *test* en un número C de modelos de impostor, donde dichos modelos provienen de un *set* de canales seleccionables. El canal que se selecciona se toma de acuerdo a una información otorgada en la llamada, ya sea mediante la numeración (ANI) u otorgada por el locutor, siendo en ambos casos una elección supervisada. El problema de aquello es que por ejemplo el tipo de numeración detectable mediante el ANI es solo de teléfonos fijos y móviles, luego la variedad de *handsets* para un canal es imposible de distinguir mediante esta técnica. Es por eso que el aporte de este trabajo es realizar la elección de canal mediante un clasificador no supervisado (Mak, 2002)(Mak *et al.*, 2004) basado en mezclas de modelos Gaussianos.

Un modelo GMM queda definido de forma muy sencilla como la suma ponderada de un grupo M de Gaussianas de dimensión N igual a la dimensión del vector de parámetros tal como se muestra en la ecuación 4.7, donde x es el vector de parámetros, λ el modelo y w_i las ponderaciones de cada una de las gaussianas p_i.

$$
\begin{aligned}
p\left(x\,|\lambda\right) &= \sum_{i=1}^{M} w_i p_i\left(x\right) \\
\sum_{i=1}^{M} w_i &= 1 \\
p_i\left(x\right) &= \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} e^{-\frac{1}{2}(O_t-\mu)^T \Sigma^{-1}(O_t-\mu)}
\end{aligned}
\tag{4.7}
$$

En este trabajo se inicializan las Gaussianas y por ende sus parámetros mediante la aplicación de *k-means*, para luego entrenar un modelo de todos los canales utilizando el algoritmo iterativo de *expectation maximization* (EM). Siguiente a aquello y utilizando el modelo multicanal como condición inicial, se entrenan los modelos de los diferentes canales mediante ejemplos adecuados aplicando EM nuevamente. En etapa de *test*, la verosimilitud logarítmica del modelo es calculada de forma muy sencilla mediante la evaluación de los parámetros de test en los modelos de canal, donde la mayor verosimilitud logarítmica indicará a que canal pertenece la señal de interés.

Aprovechando el canal seleccionado es que se define una compensación en el *score* tal que el umbral general sea capaz de separa de forma correcta los casos de CLIENTE e IMPOSTOR. Dichas compensaciones son estimadas con una base de datos aparte, de la cual se extrae el promedio de las verosimilitudes logarítmicas sin normalizar de cada uno de los canales involucrados. Dichos promedios son restados a un promedio de *scores* general, definiendo de esta forma un factor de compensación de *score* aplicado según la selección de canal.

4.8. Experimentos

Para probar la técnica propuesta se utiliza una base de datos grabada en el LPTV de la Universidad de Chile compuesta por tres canales distintos, canal celular vía altavoz (*altavoz*), canal telefónico fijo con *handset* de telefonía analógica (*fijo*) y canal telefónico IP (*IP*). La composición de dicha base se muestra en la tabla 4.6.

Mediante un subconjunto de 30 usuarios de la base explicada (10 por canal), se estiman tres compensaciones de canal. Para ello se calculó el promedio de los *scores* de CLIENTE (sin normalización de verosimilitud) de cada canal por separado y se compararon dichos valores con el *score* de CLIENTE promedio de todos los canales (ver figura 4.5). Estas

Canal	Locutores	Señales de entrenamiento	Señales de test
Altavoz	20	3	10
IP	20	3	10
Fijo	20	3	9

Tabla 4.6: Estructura de la base de datos usada

compensaciones son aplicadas según la decisión de canal que es tomada por el clasificador GMM. Luego de ello se calculan los *scores* de modelos de impostor según el canal seleccionado por el clasificador antes mencionado. Con ello se estima el EER de forma convencional mediante las curvas de FA y FR. Dichos resultados se contrastan con una elección manual de *cohort* con y sin aplicación de la compensación definida, es decir, se utilizarán *cohorts* IP, ALTAVOZ y FIJO de forma común a todos los canales de forma tal de visualizar el desempeño de la normalización.

Para el clasificador GMM se utilizó una cantidad de 512 mezclas de gaussianas entrenadas por 100 iteraciones mediante el algoritmo EM tanto en el cálculo del modelo general como en el de los canales específicos.

Los experimentos realizados con el fin de mostrar el desempeño de la normalización y compensación de *scores* combinados y separados fueron la evaluación del sistema utilizando normalización de verosimilitud con *cohort* automático, IP, Fijo y Altavoz. Luego se lleva a cabo el mismo procedimiento incluyendo esta ves la compensación explicada.

4.9. Resultados

En la tabla 4.7 se muestran los resultados de la normalización de *scores* no supervisada en contraste con el uso de *cohorts* comunes para todos los canales. Estos escenarios son probados tanto con y sin utilización de compensación de *scores* según la selección de canal realizada por el sistema.

La compensación de *score* se calcula estimando la verosimilitud logarítmica no normalizada de una serie de CLIENTES presentes en una base de datos compuesta de los mismo tres canales en análisis, pero utilizando usuarios distintos. De allí se calcula el promedio las verosimilitudes para cada canal y se compara con el promedio general de los tres canales, definiendo así los valores con que se compensa el score calculado en la aplicación.

	Con compensación			
Cohort	*Automático*	*Altavoz*	*IP*	*Fijo*
EER	4,4227	10,7009	1,9979	2,6095
	Sin compensación			
Cohort	*Automático*	*Altavoz*	*IP*	*Fijo*
EER	5,0571	10,873	3,5706	4,3948

Tabla 4.7: Resultado EER aplicando normalización y compensación de scores

Los valores obtenidos se muestran en la tabla 4.8 y de forma gráfica en la figura 4.5.

Canal	*IP*	*ALTAVOZ*	*FIJO*	*TODOS*
Promedio Verosimilitud cliente	-0,6814	-0,7086	-0,6839	-0,6915
Compensación	-0,0101	0,0171	-0,0076	

Tabla 4.8: Cálculo de las compensaciones aplicadas

4.10. Discusión

De los resultados obtenidos en la normalización de *scores*, es claro que el utilizar un *cohort* de acuerdo al canal no presenta los mejores resultados, de hecho la utilización de un *cohort* fijo proveniente del canal IP es quien muestra el mejor desempeño. Esto sugiere que los modelos de normalización de verosimilitud entregarán mejores resultados mientras mejor sea su calidad de grabación (caso que se cumple en canal IP). Este importante resultado se puede visualizar de forma clara mediante las curvas DET mostradas en la figura 4.6.

De esta forma, utilizando la mejor configuración, es decir, *cohort* de canal IP y compensación de *score*, se logra una mejora del 44 % respecto del mismo caso pero si compensación, lo cual muestra que la variabilidad del score producto del *canal* puede ser tratada mediante la selección del canal y posteriormente la aplicación de un factor compensador, teniendo los resultados ya expuestos.

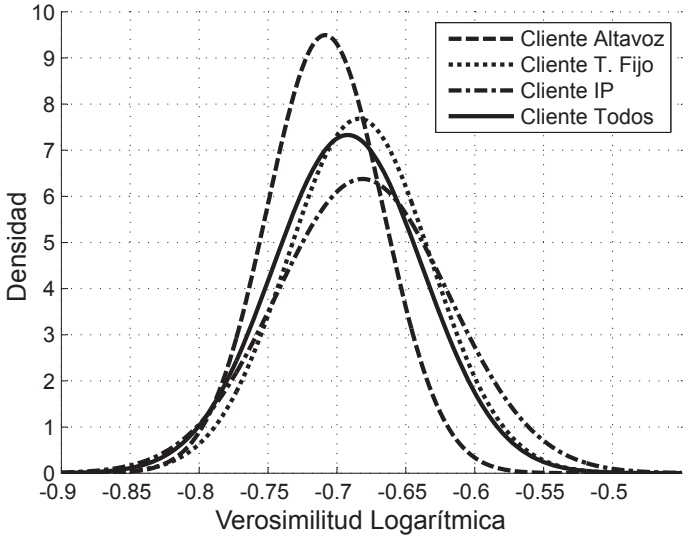


Figura 4.5: Densidades de las verosimilitudes logarítmicas de CLIENTE.

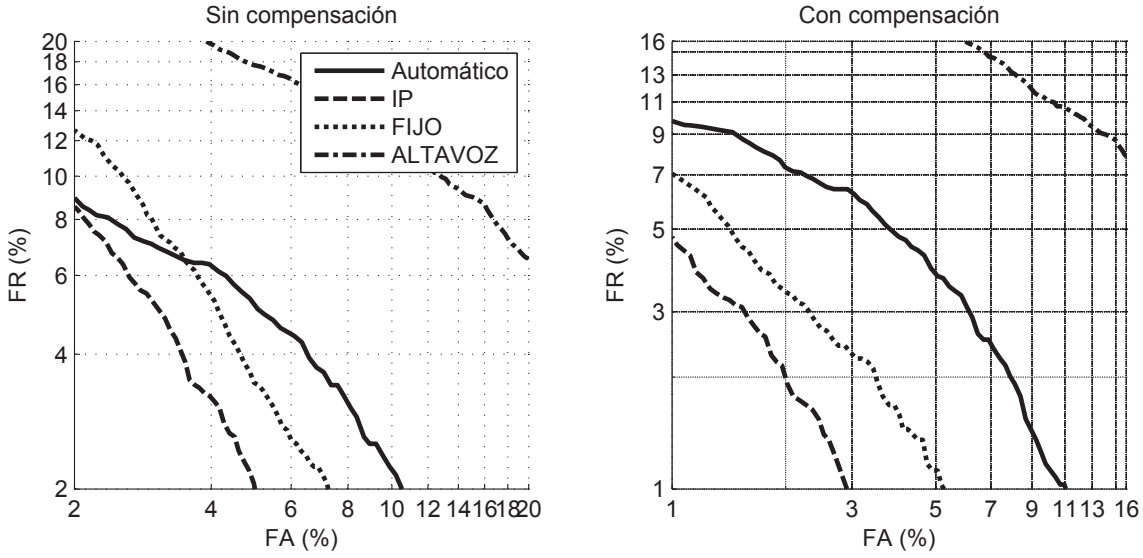


Figura 4.6: Curva DET de la normalización y compensación. Automático indica selección de *cohort* según canal. IP indica utilización de *cohort* IP. ALTAVOZ indica utilización de *cohort* ALTAVOZ. FIJO indica utilización de *cohort* FIJO.

4.11. Conclusiones

Es claro que la aplicación de las técnicas propuestas tanto en este capítulo como en el anterior sobre la base de datos ASTERISK LPTV presentan mejoras menores a las obtenidas en YOHO, teniendo incluso casos en que el desempeño del sistema se ve degradado tal como ocurre en condiciones *matched* para el canal móvil. Por otro lado, para condiciones de *mismatch* de canal, la combinación de técnicas presentadas mostró los mejores resultados, lo cual lleva a plantear el hecho de que la información presente en las bandas canceladas por los dos procedimientos propuestos poseen principalmente información no útil o del canal frente a canales *unmatched*. Del mismo modo, bajo canales *matched*, la poca información de locutor presente en las bandas canceladas, toma relevancia producto del alto desempeño del sistema. Esta degradación introducida por estas técnicas ya ha sido analizada en otros trabajos (Mak *et al.*, 2007), resaltando el hecho de que técnicas de filtrado como RASTA y CMN producen bajas en el desempeño bajo la condición señalada.

Finalmente se propuso una normalización y compensación de *scores* basado en la selección automática de canal, la cual terminó sugiriendo que el mejor modelo de *cohort* que se puede utilizar es aquel que presente mayor integridad en su grabación, razón por la cual el *cohort* IP mostró los mejores resultados. Así, se descarta la normalización propuesta y se aprovecha la compensación mostrada, la cual presentó mejoras de 44 % en el error EER del sistema.

Capítulo 5

Conclusiones

En este trabajo se mostró la manera en que la acción del canal sobre la verificación de locutor incide directamente en el desempeño que tiene el sistema, ocurriendo por ejemplo que de un caso *matched* con un nivel de error EER del 5 % se pase a un EER de 20 % en condiciones *unmatched*. Esto plantea un importante desafío con el fin de compensar dicho efecto. Por esta razón se presentaron tres formas distintas y complementarias de compensar la distorsión de canal. La primera técnica expuesta corresponde a una transformación de parámetros aplicada sobre el espectro de la MFBLE, obteniendo una mejora de 9,61 % al combinar dicha técnica con CMN respecto del caso en que se aplica CMN por si solo. Además de ello y quizás más importante es el desarrollo realizado en torno al análisis de importancia relativa, el cual mostró que el espacio propuesto (espectro de la MFBLE) presenta grandes beneficios en cuanto a factibilidad de separar la información útil y la no útil correspondiente a la distorsión de canal. En dicho estudio se logró visualizar que el espacio del espectro de la MFBLE concentra la distorsión de canal en las bandas extremas, definiendo con esto un filtro óptimo para el sistema y con el transformar los parámetros tal que ellos sean robustoz a los efectos negativos del canal. Junto con lo anterior, quedó claro que CMN actúa sobre una región del espectro fuertemente correlacionada con las bajas frecuencias del espectro de la MFBLE y por ende, la transformada a utilizar en combinación con esta técnica concentra su atención principalmente en la zona de bandas altas. Mismas conclusiones son las que se extraen al analizar la combinación de la transformación propuesta con el filtrado temporal RASTA. En dicho caso, las mejoras son del orden 4 % respecto de la aplicación de RASTA solamente. También se extrae del análisis de importancia relativa de este caso, que la banda baja del espectro de la MFBLE esta aun más correlacionada con la acción de RASTA y por ende no debe ser cancelado,

con lo cual se concluyó que para este caso la transformación ideal tomaba la forma de un filtro pasa bajos.

Seguido de esto se planteó comparar y combinar el procedimiento propuesto con un filtrado de trayectorias temporales aplicado todo sobre una base de datos de mayor dificultad. En ella se estableció que la combinación de ambas técnicas presenta los mejores resultados para los casos de *mismatch* de canal y *matched-casa*. En contraste con esto se utilizó CMN y RASTA con sus respectivas combinaciones con la transformación de parámetros en el espectro de la MFBLE, con lo cual los resultados obtenidos tomaron especial relevancia pues se demostró que solo la combinación de la transformación presentada con el filtrado de trayectoria temporales propuesto presentaron mejoras en los casos de interés para los canales fijo y móvil. Los resultados sobre el canal telefónico fijo mejoraron en un 5,5 % para *mismatch* de canal y 10,3 % para *matched-casa*, mientras que sobre el canal telefónico móvil los resultados mejoraron un 13 % y 6,2 % para los mismos casos. Sin embargo, se hizo notoria una degradación en el desempeño del sistema para canal *matched* sobre el canal telefónico móvil. Producto de esto se plantea que la información presente en las bandas canceladas por la mezcla de las técnicas propuestas refleja en parte información respecto del locutor y, dada la baja tasa de error en este caso (bordeando el 3 %), dicha información es importante y necesaria para verificar de forma adecuada. Afortunadamente, el caso de canal *matched* es poco usual en un sistema verificador de locutor texto dependiente operado sobre la red telefónica, y por ende los resultados obtenidos en las otras condiciones (casos muy comunes en este tipo de sistemas) se consideran satisfactorios.

Finalmente, se propuso un forma de normalización inspirada en HT-Norm y se incorporó a dicha técnica una selección de canal automática basada en mezclas de modelos Gaussianos (GMM). Este método intentaba probar que los modelos de *cohorts* que mejor resultado entregan son los se encuentran sobre el canal de la señal de interés (canal *matched*), sin embargo, esto no aconteció y por el contrario se logró ver de forma clara cómo los mejores resultados se obtienen con modelos de *cohorts* sobre un solo canal siendo el mejor de ellos el canal IP. Esto sugiere que la mejor forma de normalizar la verosimilitud es mediante modelos de *cohorts* grabados en un canal limpio y de la mejor calidad posible. Adicional a lo anterior, se aplicó una compensación de *score* aprovechando la selección de canal y basado en el hecho de que el *mismatch* de canal introduce una fuerte variabilidad en los scores de decisión. Esto resulto en una mejora del 44 % utilizando modelos de cohorts sobre canal IP repecto del caso en que no se compensan los *scores* y se utiliza el mismo *cohort*, lo cual deja al sistema operando a un nivel de error bajísimo (del orden del 2 %).

5.1. Trabajo Futuro

Como se mencionó en el capítulo dedicado al marco teórico, hoy en día los verificadores de locutor texto independiente son un área de desarrollo fuertemente trabajado. Es por ello que se propone como trabajo a futuro el aplicar los métodos aquí propuestos sobre un sistema verificador de este tipo. En un primer caso, la aplicación de la técnica de transformación de parámetros puede ser aplicada de forma rápida y sin modificaciones sobre las señales ingresadas al sistema, sin embargo, la técnica de filtrado de trayectorias temporales, dado que se aplica sobre una evolución en el tiempo, debe ser modificada en un sistema con requerimientos de tiempo real y de características texto independiente que posee señales de análisis bastante más largas que las obtenidas en un sistema texto dependiente, por lo que el esperar a terminar de grabar una señal y luego de eso procesarla es inviable y debe ser remediado. Una forma de abordar aquello es analizar la señal sobre ventanas móviles o aplicar filtros IIR de bajo orden y con ello aprovechar el tiempo de grabación en el procesamiento propio del filtrado de parámetros.

Inspirado también en los resultados obtenidos utilizando modelos de impostor sobre el canal IP, se propone llevar las suposiciones hechas más allá y probar el sistema utilizando *cohorts* grabados mediante micrófonos de alta calidad y en ausencia de ruido aditivo (condiciones controladas), con el fin de bajar aun más el error EER presentado.

Referencias

Bai, J., Zheng, R., Xu, B., y Zhang, S. (2004, Dec.). Robust speaker recognition integrating pitch and wiener filter. En *Chinese spoken language processing, 2004 international symposium on* (p. 69-72).

Barrus, C., y Gauvain, J.-L. (2003). Feature and score normalization for speaker verification of cellular data. En *In: Proc. icassp 03.*

Becerra Yoma, N., McInnes, F., y Jack, M. (1998, May). Weighted viterbi algorithm and state duration modelling for speech recognition in noise. *Acoustics, Speech and Signal Processing, 1998. ICASSP-98., 1998 IEEE International Conference on*, *2*, 709-712.

Bimbot, F., Bonastre, J., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., et al. (2004). A tutotrial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, *4*, 430-451.

Bocklet, T., Maier, A., Bauer, J., Burkhardt, F., y Noth, E. (2008, 31 2008-April 4). Age and gender recognition for telephone applications based on gmm supervectors and support vector machines. En *Acoustics, speech and signal processing, 2008. icassp 2008. ieee international conference on* (p. 1605-1608).

Burget, L., Matejka, P., Schwarz, P., Glembek, O., y Cernocky, J. (2007). Analysis of feature extraction and channel compensation in a gmm speaker recognition system. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(7), 1979-1986.

Campbell, J. P. (1997, Sep). Speaker recognition: a tutorial. *Proceedings of the IEEE*, *85*(9), 1437-1462.

Hardt, D., y Fellbaum, K. (1997). Spectral subtraction and rasta-filtering in text-dependent hmm-based speaker verification. En *In: Proc. icassp 97.*

Hermansky, H. (1994). Speech beyond 10 milliseconds (temporal filtering in feature domain). En *International workshop on human interface technology.*

Hermansky, H. (1995). Exploring temporal domain for robustness in speech recognition.

En *Proc. of 15th international congress on acoustics* (p. 61-64).

Hermansky, H., y Morgan, N. (1994). Rasta processing of speech. *IEEE Transactions on Speech and Audio Processing*, *2*, 578-589.

Juang, B.-H., y Rahim, M. (1996, Jan). Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. *IEEE Transactions on Speech and Audio Processing*, *4*(1), 19-30.

Jung, H.-Y., y Lee, S.-Y. (2000). On the temporal decorrelation of feature parameters for noise-robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, *8*, 407-416.

Lo, T., Mak, M., y Yiu, K. (1999). A new cepstrum-based channel compensation method for speaker verification. En *Eurospeech '99* (pp. 775–778).

Mak, M.-W. (2002). Combining stochastic feature transformation and handset identification for telephone-based speaker verification. En *In: Proc. icassp 02* (pp. 701–704).

Mak, M.-W., Hsiao, R., y Mak, B. (2006, May). A comparison of various adaptation methods for speaker verification with limited enrollment data. *Acoustics, Speech and Signal Processing, 2006. ICASSP-2006., 2006 IEEE International Conference on*, *1*, 929-932.

Mak, M.-W., Tsang, C. leung, y Kung, S. yuan. (2004). Stochastic feature transformation with divergence-based out-of-handset rejection for robust speaker verification. *EURASIP Journal on Applied Signal Processing*, *4*, 452-465.

Mak, M.-W., Yiu, K.-K., y Kung, S.-Y. (2007). Probabilistic feature-based transformation for speaker verification over telephone networks. *Neurocomput.*, *71*(1-3), 137-146.

Mariéthoz, J., y Bengio, S. (2005). A unified framework for score normalization techniques applied to text independent speaker verification. *IEEE Signal Processing Letters*, *12*, 532-535.

Naik, D. (1995, May). Pole-filtered cepstral mean subtraction. *Acoustics, Speech and Signal Processing, 1995. ICASSP-95., 1995 IEEE International Conference on*, *1*, 157-160.

Neumeyer, L., y Weintraub, M. (1994, Apr). Probabilistic optimum filtering for robust speech recognition. *Acoustics, Speech and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, *1*, 417-420.

Parris, E. S., y Carey, M. J. (1996). Language independent gender identification. En *Icassp '96: Proceedings of the acoustics, speech, and signal processing, 1996. on conference proceedings., 1996 ieee international conference* (p. 685-688). Washington, DC, USA: IEEE Computer Society.

Rabiner, L., y Schmidt, C. (1980, Aug). Application of dynamic time warping to connected

digit recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *28*(4), 377-388.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. En *Proceedings of the ieee* (p. 257-286).

Rabiner, L. R., y Schafer, R. W. (2007). Introduction to digital speech processing. *Found. Trends Signal Process.*, *1*(1), 1-194.

Reynolds, D., y Rose, R. (1995, Jan). Robust text-independent speaker identification using gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on*, *3*(1), 72-83.

Reynolds, D. A. (1996). The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus. En *Icassp '96: Proceedings of the acoustics, speech, and signal processing, 1996. on conference proceedings., 1996 ieee international conference* (p. 113-116). Washington, DC, USA: IEEE Computer Society.

Reynolds, D. A., Zissman, M. A., Quatieri, T. F., O'Leary, G. C., y Carlson, B. A. (1995). The effect of telephone transmission degradations on speaker recognition performance. En *In: Proc. icassp 95.*

Sankar, A., y Lee, C.-H. (1996). A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, *4*, 190-202.

Skosan, M., y Mashao, D. (2006). Modified segmental histogram equalization for robust speaker verification. *Pattern Recognition Letters*, *27*(5), 479-486.

Surendran, A., Lee, C.-H., y Rahim, M. (1999, Nov). Nonlinear compensation for stochastic matching. *Speech and Audio Processing, IEEE Transactions on*, *7*(6), 643-655.

Ting, H., Yingchun, Y., y Zhaohui, W. (2006, 16-20). Combining mfcc and pitch to enhance the performance of the gender recognition. En *Signal processing, proceedings. icsp '06. 2006 8th international conference on* (Vol. 1, p.-).

Tukekci, Z. (2007). Convolutional bias removal based on normalizing the filterbank spectral magnitude. *IEEE Signal Processing Letters*, *14*, 485-488.

Vuuren, S. V., y Hermansky, H. (1998). On the importance of components of the modulation spectrum for speaker verification. En *In proc. icslp '98* (p. 3205-3208).

Xie, Y., Liu, M., Yao, Z., y Dai, B. (2006). Improved two-stage wiener filter for robust speaker identification. En *Icpr '06: Proceedings of the 18th international conference on pattern recognition* (pp. 310–313). Washington, DC, USA: IEEE Computer Society.

Yin, S.-C., Rose, R., y Kenny, P. (2008, 31 2008-April 4). Adaptive score normalization for

progressive model adaptation in text independent speaker verification. En *Acoustics, speech and signal processing, 2008. icassp 2008. ieee international conference on* (p. 4857-4860).

Yiu, K. K., Mak, M. W., y Kung, S. Y. (2007). Environment adaptation for robust speaker verification by cascading maximum likelihood linear regression and reinforced learning. *Comput. Speech Lang.*, *21*(2), 231-246.

Yu, A.-T., y Hsiao-ChuanWang. (2003). Channel effect compensation in lsf domain. *EURASIP Journal on Applied Signal Processing*, *2003*, 922-929.

Zeng, Y.-M., Wu, Z.-Y., Falk, T., y Chan, W.-Y. (2006, Aug.). Robust gmm based gender classification using pitch and rasta-plp parameters of speech. En *Machine learning and cybernetics, 2005. proceedings of 2005 international conference on* (p. 3376-3379).

Zhonghua, F., Lei, X., y Rongchun, Z. (2004, Aug.-4 Sept.). Channel robust speaker verification via extended feature mapping. En *Signal processing, 2004. proceedings. icsp '04. 2004 7th international conference on* (Vol. 3, p. 2417-2420).

Zilovic, M. S., Ramachandran, R. P., y Mammone, R. J. (1998). Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions. *IEEE Transactions on Speech and Audio Processing*, *6*, 260-267.