UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

ROBUSTEZ A VARIABILIDAD DE CANAL EN RECONOCIMIENTO DE PATRONES ACÚSTICOS CON APLICACIONES EN ENSEÑANZA DE IDIOMAS Y BIOMETRÍA

TESIS PARA OPTAR AL GRADO DE DOCTOR EN INGENIERÍA ELÉCTRICA

CLAUDIO ANDRÉS GARRETÓN VENDER

PROFESOR GUÍA:
NÉSTOR BECERRA YOMA

MIEMBROS DE LA COMISIÓN:
JORGE SILVA SÁNCHEZ
CARLOS BUSSO RECABARREN
ISMAEL SOTO GÓMEZ

SANTIAGO DE CHILE
MAYO 2011

ROBUSTEZ A VARIABILIDAD DE CANAL EN RECONOCIMIENTO DE PATRONES ACÚSTICOS CON APLICACIONES EN ENSEÑANZA DE IDIOMAS Y BIOMETRÍA

Resumen de Tesis para optar al grado de Doctor en Ingeniería Eléctrica

Claudio Andrés Garretón Vender

Profesor guía: Néstor Becerra Yoma

Santiago de Chile, Mayo 2011.

La robustez a la variabilidad en el canal de comunicaciones entre condiciones de entrenamiento y evaluación es uno de los más graves problemas que enfrentan los sistemas de procesamiento de patrones acústicos en aplicaciones reales. Además, por motivos de usabilidad, la cantidad y la duración de las elocuciones con las que debe operar una aplicación es limitada. Estas restricciones llevan a un escenario desfavorable: modelos con un bajo nivel de entrenamiento y elocuciones cortas en la etapa de prueba implican una reducción en la exactitud del sistema, situación que puede empeorar si existen diferencias de canal entre los procesos de entrenamiento y evaluación. De aquí deriva la importancia de generar sistemas de procesamiento de patrones acústicos robustos al canal de comunicaciones.

Dentro de esta tesis se presentan dos modelos para los efectos de la distorsión de canal que derivan en técnicas de cancelación. Se pone especial énfasis en aplicaciones que funcionan con datos limitados o señales de corta duración y así generar propuestas aplicables a situaciones reales. La estrategia que sigue el primer modelo desarrollado es obviar la hipótesis de invariabilidad en el tiempo de la distorsión de canal. El segundo esquema propuesto considera la interdependencia de las componentes espectrales al modelar la distorsión. Para evaluar las técnicas presentadas en esta tesis se utilizan dos plataformas de reconocimiento de patrones acústicos: a) un sistema biométrico basado en verificación de locutor texto-dependiente (TD-SV, Text Dependent - Speaker Verification); y b) un sistema de evaluación automática de pronunciación (CAPT, Computer Aided Pronunciation Training) para enseñanza de segundo idioma basado en tecnología de reconocimiento de voz.

La primera técnica propuesta es una transformación de características frame-por-frame para TD-SV con datos limitados. La transformación es aplicada como un filtro pasa-banda a lo largo del vector de características que representa la envolvente espectral. El objeto de este filtrado es reducir el efecto variable en el tiempo de la componente de distorsión de canal en el dominio cepstral, el que es generado por la dependencia de la respuesta del canal en la señal de voz. La transformación se define empleando análisis de importancia relativa en combinación con una función discriminativa basada en la razón de dispersión intra-locutor/inter-locutor.

A continuación, se presenta una nueva estrategia de compensación de la distorsión de canal en el dominio de las características basada en una aproximación polinomial aplicada a TD-SV y CAPT con datos limitados. El método modela la distorsión empleando una función polinomial en el dominio del logaritmo de las energías del banco de filtros Mel. La técnica modela la continuidad de la respuesta en frecuencia del canal y reduce el número de variables requeridas en la estimación de la distorsión al usar un modelo paramétrico. El método usa esquemas de búsqueda de vecino más cercano en la etapa de estimación, lo que mantiene controlada carga computacional.

Las técnicas presentadas consiguen sustanciales mejoras en los sistemas de TD-SV y CAPT al ser aplicadas de forma aislada y en combinación con técnicas convencionales para robustez a canal como CMN (Cepstral Mean Normalization) y RASTA (Relative Spectral). Cabe destacar que los métodos propuestos en esta tesis operan en el dominio de los parámetros acústicos de la señal de voz, por lo que son eventualmente aplicables a cualquier tarea de procesamiento de patrones acústicos.

Agradecimientos

En primer lugar quiero agradecer a mi familia por el apoyo, motivación y comprensión que me han brindado durante estos años. Nada de esto sería posible sin ellos. Después de todo, han pasado 29 años aún soy un estudiante (ups).

Deseo entregar mi especial agradecimiento a mi profesor guía, Dr. Néstor Becerra Yoma, por haberme entregado de su tiempo y experiencia, y haberme orientado en mi formación. Gracias también a mis compañeros investigadores y estudiantes del Laboratorio de Procesamiento y Transmisión de Voz de la Universidad de Chile por su soporte y compañía.

Agradezco además al Dr. Horacio Franco por haber compartido su experiencia y conocimientos en mi pasantía en el Speech Technology and Research Laboratory de SRI International en Menlo Park, California, EE.UU. También agradezco especialmente al Dr. Martín Graciarena por la orientación y el constante apoyo que me brindó durante la pasantía.

Finalmente cabe mencionar que el desarrollo de esta tesis contó con el apoyo de los siguientes programas de beca de CONICYT Chile: beca para estudios de Doctorado año 2007; y, beca de apoyo a la realización de tesis doctoral 2da convocatoria 2009 (24091017). También se contó con el apoyo otorgado por el proyecto MECESUP FSM0601 para realizar una pasantía en SRI International, EE.UU. Además, parte del trabajo mostrado en la presente tesis se realizó en el marco de los proyectos Fondef D05I10243 y Fondecyt 1070382/1100195.

Índice

Lista de figuras

Lista de Tablas

Glosario

Alineamiento: Proceso para asociar a cada vector de parámetros acústico un estado de los modelos que describen el ASR (HMMs).

ASR: *Automatic Speech Recognition* - Reconocimiento Automático de Voz.

Baseline: Resultado de evaluar el sistema *ASR,* entrenado con modelos acústicos (*HMM*) sin compresión, con señales sin distorsión.

CDHMM: *Continuous Density Hidden Markov Model.*

CMN: *Cepstral Mean Normalization.*

Coeficientes Cepstrales: Parámetros acústicos que caracterizan a una señal de voz. Se basan en análisis en frecuencia de la señal.

Conjunto de Entrenamiento: Señales acústicas que se utilizan para determinar los parámetros de los modelos que describen el *ASR.*

Conjunto de Test: Señales acústicas que evalúan el reconocedor y que no fueron utilizadas para el entrenamiento de los modelos que describen el *ASR.*

DCT: *Discrete Cosine Transform.*

DET: *Detection Error Tradeoff.*

DFT: *Discrete Fourier Transform.*

EER: *Equal Error Rate*

EM: *Expectation Maximization.*

Estado: Etapa de un *HMM* que representa un período estacionario de una señal acústica. Su valor es escalar.

FA: Falsa Aceptación.

FR: Falso Rechazo.

Frame: Segmentación de la señal acústica.

GMM:	*Gaussian Mixture Model*
HMM:	*Hidden Markov Model.*
LPTV:	Laboratorio de Procesamiento y Transmisión de Voz.
MAP:	*Maximum a Posteriori.*
MFCC:	*Mel Frequency Cepstral Coefficient.*
Mismatch:	Situación que ocurre cuando las condiciones de evaluación y entrenamiento difieren. Pueden ser diferencias en el ambiente, locutor, ruido, etc.
ML:	*Maximum Likelihood.*
MLLR:	*Maximum Likelihood Linear Regression.*
N-best:	Lista de las N mejores hipótesis que se derivan del ASR.
p.d.f.:	*Probability density function.*
ROC:	*Receiver Operating Characteristic.*
SD:	*Speaker-dependent.*
SI:	*Speaker-independent.*
SV:	*Speaker Verification.*
TD-SV:	*Text-dependent Speaker Verification.*
TEER:	*Threshold of Equal Error Rate.*
TI-SV:	*Text-independent Speaker Verification.*
WER:	*Word Error Rate.*

Capítulo 1

Introducción

1.1 Motivación

La identificación o reconocimiento de figuras, sonidos, aromas, texturas y semántica en el lenguaje son capacidades habituales en el ser humano las que manifiesta sin ni siquiera pensar al respecto. Sin embargo, a pesar de la capacidad de realizar largos y complejos cálculos numéricos que tienen los sistemas computacionales, el diseño de algoritmos para el desarrollo de estas tareas no es para nada trivial (Gales y Young, 2008; Jelinek, 1997). Es por esto que la investigación en este tipo de algoritmos es uno de los mayores desafíos en el área de la ingeniería para el siglo que recién comienza.

A partir de las últimas décadas del siglo pasado la industria de las telecomunicaciones y sistemas multimediales ha penetrado enormemente a nivel

mundial. De la mano de este vertiginoso crecimiento también han evolucionado las plataformas computacionales de procesamiento de datos, las que cada vez son más alcanzan mayores capacidades a menores costos (Zheng-Hua Tan y Lindberg, 2010; Neustein, 2010). Amalgamando los avances y conocimiento generado en ambas áreas, los investigadores del área de procesamiento de la voz han concentrado sus esfuerzos en la generación de métodos cada vez más complejos y eficientes. Desde comienzos de los años 80 ya se han venido desarrollando diversos sistemas y se han estandarizado diversas técnicas, especialmente para las tareas de reconocimiento de voz y locutor, las que han alcanzado tasas de desempeño muy aceptables en condiciones controladas o de laboratorio. Este tipo de condiciones se puede describir como ambientes acústicos limpios para la grabación de la voz, la que es generalmente realizada con dispositivos de captura de buena calidad. No obstante, cuando las condiciones acústicas ambientales o de grabación en las etapas de entrenamiento y/o evaluación se acercan a las de aplicación real, el rendimiento de los sistemas de reconocimiento de voz y locutor decaen significativamente (Furui, 1997; Neustein, 2010).

A pesar de esta vulnerabilidad a las condiciones de operación que tienen las aplicaciones masivas basadas en tecnologías de voz, estas se han masificado cada vez más, sobretodo en la última década (Zheng-Hua Tan y Lindberg, 2010; Neustein, 2010). Es por esto que ya es una situación común ver un sistema de procesamiento de voz enfrentado a situaciones prácticas reales. Aplicaciones como control por comandos, máquinas de dictado, plataformas de autentificación biométrica o sistemas educativos,

entre otras, son las que más ha desarrollado esta incipiente industria. De esta manera el desarrollo de sistemas de reconocimiento de voz y locutor robustos, es decir, que ofrezcan un desempeño aceptable sin depender de las condiciones ambientales en que estos operen es una necesidad real, y a la vez, un tremendo desafío de ingeniería, el que aún sigue abierto (Zheng-Hua Tan y Lindberg, 2010; Neustein, 2010).

Básicamente existen tres factores que degradan el desempeño o rendimiento de sistemas de reconocimiento de voz y locutor: la variabilidad en el habla de cada locutor, el ruido ambiental y la distorsión generada por el canal de comunicaciones (micrófono, codificación y canal telefónico). En las últimas décadas han aparecido un sinnúmero de métodos para eliminar o atenuar los diversos factores de degradación a afectan los sistemas de reconocimiento de patrones acústicos, de forma individual o en conjunto. A pesar de ser este un tema que ha sido ampliamente abordado en la literatura especializada reciente, la comunidad está lejos de encontrar una solución definitiva a estos problemas. Si el objetivo final es que ciertas aplicaciones de procesamiento de la voz puedan lograr una alta penetración, deben ser capaces de operar correctamente bajo las condiciones adversas descritas.

Se puede denominar como desacople o *mismatch* a la diferencia por variabilidad acústica entre las condiciones de grabación de las señales de entrenamiento y evaluación. En particular, el conjunto de distorsiones a nivel de parámetros acústicos que se generan por las variaciones en las condiciones en que se captura y transmite la señal de voz, tales como: diferentes medios de transmisión (micrófono, teléfono, celular, red de datos, etc.),

tipos de micrófono (auricular o *handset* telefónico, micrófonos de computador o de alta calidad, etc.) se denomina "variabilidad de canal de comunicaciones" (Sorell, 2009). A su vez, el *mismatch* que se genera por esta variabilidad se denomina "*mismatch* de canal" y este será el principal problema a abordar en el desarrollo de esta tesis. Para trabajar con el problema del *mismatch* de canal es necesario utilizar técnicas para cancelar los efectos que produce la distorsión de canal en la señal, es decir, generar aplicaciones robustas a canal (Neustein, 2010; Sorell, 2009). Así, la presente tesis propone técnicas para enfrentar el problema de la distorsión en el canal de comunicaciones, las que operan en el espacio de las características acústicas que se extraen de las señales de voz, en los módulos de parametrización acústica y clasificación acústico-fonética, respectivamente. Los métodos propuestos son probados en dos plataformas de reconocimiento de patrones acústicos: un sistema de autentificación biométrica y un sistema educativo en evaluación de pronunciación

1.2 Definición del problema a abordar

La presente tesis se enmarca en el desarrollo de aplicaciones de procesamiento de patrones acústicos robustas a la variabilidad en el canal de comunicación. En particular, se pretende abordar el problema del *mismatch* en el canal de comunicación entre condiciones de entrenamiento y evaluación en sistemas de procesamiento de voz. Dentro de esta tesis se proponen modelos para los efectos de la distorsión de canal que derivan

en técnicas de cancelación de canal para señales de voz. Se pondrá especial énfasis en aplicaciones que funcionan con datos limitados y/o señales de corta duración en sus etapas de entrenamiento y prueba. La justificación de enfocar el trabajo de investigación en sistemas que operan con datos limitados es la de generar métodos aplicables a situaciones prácticas reales. Por motivos de usabilidad, en estos escenarios las elocuciones que la aplicación solicita al potencial usuario no pueden tener una larga duración. Entonces, en el caso de la etapa de entrenamiento, una cantidad reducida de grabaciones de entrenamiento implicará modelos con un bajo nivel de entrenamiento, lo que lleva a una reducción en la exactitud del sistema. Por otro lado, el rendimiento del sistema se verá también afectado por la restricción en el tiempo y cantidad de señales en la etapa de prueba. De aquí deriva la importancia de generar robustez al canal de comunicaciones en este tipo de sistemas.

La estrategia que siguen los modelos desarrollados en esta tesis para generar robustez a canal en señales cortas es la de considerar la hipótesis de dependencia de la distorsión de canal en la señal de voz. Cada esquema de compensación propuesto está respaldado por un marco teórico apropiado y es comparado con técnicas convencionales de supresión de los efectos del canal. Cabe destacar que las técnicas propuestas en esta tesis operan en el dominio de las características acústicas que se extraen de la señal de voz por lo que son, eventualmente, aplicables a cualquier tarea de procesamiento de patrones acústicos.

Para evaluar los esquemas de cancelación de canal de comunicaciones que se proponen en esta tesis se utilizan dos plataformas de reconocimiento de patrones acústicos: a) un sistema biométrico basado en verificación de locutor texto dependiente (TD-SV, *Text Dependent - Speaker Verification*); y b) un sistema de evaluación automática de pronunciación (CAPT, *Computer Aided Pronunciation Training*) basado en tecnologías de reconocimiento de voz (ASR, *Automatic Speech Recognition*) para enseñanza de segundo idioma (CALL, *Computer Aided Language Learning*).

Finalmente, cabe destacar el impacto tecnológico y social que ciertas aplicaciones basadas en tecnología de procesamiento de patrones acústicos pueden llegar a alcanzar, lo que hace al problema de la variabilidad de canal un desafío sumamente interesante. La robustez de un sistema de evaluación automático de pronunciación con micrófonos de bajo costo abre las puertas para crear una novedosa herramienta aplicada a la enseñanza de idiomas. Además, una plataforma de verificación de locutor robusta a variabilidad de canal puede derivar una amplia gama de atractivas aplicaciones tales como el *password* o huella de voz para control de acceso en diálogos telefónicos y sistemas de identificación forense. Los métodos de cancelación de canal que se investigan en esta tesis ayudarán al desarrollo de sistemas de procesamiento de voz robustos que permitan su operación con dispositivos de captura de bajo costo (micrófonos, teléfonos, etc.) y que además permitan el cambio de estos durante su operación. De esta forma, se facilita el acceso a este tipo de aplicaciones y por ende su masificación. Vale la pena destacar que el desafío

de la cancelación de la distorsión de canal es un tema de interés actual en la comunidad y que las técnicas propuestas en la presente tesis son originales y no han sido publicadas previamente en la literatura. Prueba de ello son las dos publicaciones[1,2] en revistas ISI logradas del trabajo de esta tesis (ver anexo).

1.3 Objetivos generales y específicos

1.3.1 Objetivo general

Mejorar la robustez a *mismatch* de canal en sistemas de procesamiento de patrones acústicos mediante la generación de técnicas de cancelación de la distorsión de canal de comunicaciones.

1.3.2 Objetivos específicos

- Generar un modelo para la distorsión por efectos de canal de comunicación en el dominio espectral y/o cepstral. Incluir la dependencia de la respuesta del canal con

[1] Claudio Garretón y Néstor Becerra Yoma, "Telephone channel compensation in speaker verification using a polynomial approximation in the log-filter-bank energy domain," Aceptado para su publicación en IEEE Transactions on Audio, Speech and Language Processing. 2011.

[2] Claudio Garretón, Néstor Becerra Yoma y Matías Torres. "Channel robust feature transformation based on filter-bank energy filtering," IEEE Transactions on Audio, Speech and Language Processing, Vol. 18, No. 5, pp. 1082 - 1086. 2010.

la señal de entrada. Se dará un especial énfasis a modelar la distorsión por canal telefónico y micrófonos de baja calidad en señales cortas.

- Desarrollar un método de parametrización robusto a efectos de canal aplicado en el dominio espectral y/o cepstral.
- Definir una estrategia de compensación de *mismatch* por distorsión de canal a nivel de parámetros acústicos. Esta debe generar una carga computacional acorde a una aplicación on-line.
- Combinar las técnicas a desarrollar con métodos convencionales aplicados en distintos dominios de operación (i.e. transformaciones para parametrización robusta, compensación de parámetros, etc.).

1.4 Estructura de la tesis

La presente tesis se ha estructurado de modo de introducir gradualmente al lector al problema abordado, comenzando con una visión macro de la temática estudiada para llegar posteriormente a los detalles del los métodos propuestos. Así, se comienza con un marco introductorio sobre sistemas de reconocimiento de voz, verificación de locutor y evaluación de pronunciación para luego dar paso a la descripción técnica del problema del *mismatch* por distorsión de canal y como se ha abordado este tema en la literatura especializada. A continuación, se describe en detalle cada método propuesto, se muestran resultados y comparaciones con esquemas convencionales hallados en la

literatura especializada. De esta forma se tendrá un soporte conceptual adecuado para seguir el desarrollo de las técnicas propuestas y los experimentos realizados. La tesis se compone de 5 capítulos, cada uno trata temas relevantes relacionados con el trabajo de documentación, investigación y resultados experimentales. A continuación se describe brevemente la estructura de cada uno de ellos.

El capítulo 2 tiene como objetivo especifico introducir al lector de forma genérica en la problemática del *mismatch* por distorsión de canal. En éste se busca entregar una base teórica suficiente para adentrarse en las técnicas y análisis propuestos en esta tesis. Se presentan las técnicas en el estado-del-arte más utilizadas. En éste capítulo, se realiza además una descripción de las tecnologías de los sistemas de reconocimiento de voz, verificación de locutor y evaluación de pronunciación comenzando con el procesamiento de las señales de voz, metodologías de evaluación, técnicas de clasificación. Especial énfasis se hace a los sistemas desarrollados en el Laboratorio de Procesamiento y Transmisión de Voz (LPTV) de la Universidad de Chile.

En el capítulo 3, se propone una novedosa transformación de características *frame*-por-*frame* para robustez a la variabilidad de canal de comunicaciones en verificación de locutor texto-dependiente con datos limitados. La transformación es aplicada como un filtro pasa-banda a lo largo del vector de características que representa la envolvente espectral. El objeto de este filtrado es reducir el efecto variable en el tiempo de la componente de distorsión de canal en el dominio cepstral, el que es

generado por la dependencia de la respuesta del canal de comunicaciones en la señal de voz de entrada. La transformación presentada en este capítulo consigue una compensación de canal variable en el tiempo. La transformación es definida empleando análisis de importancia relativa en combinación con una función discriminativa basada en la razón de dispersión intra-locutor/inter-locutor. Los resultados presentados en este capítulo muestran que el espectro del vector de envolvente espectral provee de una representación concisa del efecto de la distorsión de canal. Más aún, por operar en el bloque de parametrización, el esquema propuesto puede ser aplicado a cualquier tarea de reconocimiento de patrones de voz.

El capítulo 4 presenta una nueva estrategia de compensación de la distorsión de canal en el dominio de las características basada en una aproximación polinomial aplicada a verificación de locutor texto-dependiente y evaluación de pronunciación basada en reconocimiento de voz con datos limitados. El método modela la distorsión empleando una función polinomial en el dominio del logaritmo de las energías del banco de filtros Mel. Además, la técnica incluye el modelamiento la continuidad de la respuesta en frecuencia del canal. Al usar un modelo paramétrico, el esquema mostrado reduce el número de variables requeridas en la estimación de la distorsión. A pesar de ser un método estadístico, al usar métodos de búsqueda de vecino más cercano en la etapa de estimación, se mantiene controlada carga computacional.

Finalmente, el capítulo 5 presenta las conclusiones y análisis de las técnicas propuestas en esta tesis además del trabajo propuesto a futuro.

1.5 Contribuciones de la tesis

Como se menciona en la sección anterior, en esta tesis se proponen dos esquemas para robustez a la variabilidad de canal: un método de parametrización robusto al canal de comunicaciones y una técnica de cancelación de los efectos del canal en el espacio de las características. De la primera propuesta se desprenden las siguientes contribuciones: un esquema de transformación de características *frame*-por-*frame* para compensar los efectos del canal de comunicaciones con datos limitados, el que reduce la componente variable en el tiempo de la distorsión aplicando un filtrado apropiado a lo largo del vector de características espectrales; un esquema de estimación de parámetros de filtrado basado en análisis de importancia relativa en combinación con una función discriminativa que usa la razón de dispersión intra-locutor/inter-locutor; un nuevo espacio de paramétrico, que entrega una representación certera del efecto de la distorsión de canal variable en el tiempo: el espectro del vector características espectrales.

En el marco de la segunda propuesta de esta tesis, las contribuciones son: un método de compensación de distorsión de canal con datos limitados que usa una

aproximación polinomial en el dominio del logaritmo de las energías del banco de filtros Mel; un modelo paramétrico para la distorsión que considera la continuidad de la respuesta en frecuencia del canal; y, un esquema eficiente para la estimación de la componente aditiva de la distorsión de canal en el dominio cepstral.

Capítulo 2

Marco teórico

2.1 Reconocimiento de voz

El proceso de reconocimiento automático de la voz tiene como función obtener la secuencia de palabras asociada a una elocución en lenguaje natural de entrada. La tarea de reconocer la voz se enmarca dentro del amplio campo de reconocimiento de patrones, es decir, la función de un sistema ASR consiste básicamente en convertir una señal acústica en un conjunto de clases conocidas. La solución estocástica al problema de ASR emplea el teorema de Bayes para modelar el proceso de reconocimiento de voz como un problema de maximización de probabilidad *a posteriori*, el que puede descomponerse de la siguiente manera (Gales y Young, 2008; Jelinek, 1997):

$$P(W,O) = P(W \mid O) \cdot P(O) \tag{1}$$

donde $W = \left[w_1, w_2, ..., w_J \right]$ representa a la secuencia de palabras y $O = \left[O_1, O_2, \ . \ . \ . , O_I \right]$ corresponde a la secuencia de vectores de parámetros generados de la señal acústica.

Si se asume que la ocurrencia de una secuencia de vectores de parámetros O, es igualmente probable, es decir, $P(O)$ es igual para todo O (Jelinek, 1997), el problema de maximización se transforma en encontrar la mejor secuencia de palabras W que maximice la probabilidad $P(W|O)$:

$$\hat{w} = \arg\max_{w} P(W \mid O) \tag{2}$$

al aplicar el teorema Bayes se obtiene:

$$\hat{w} = \arg\max_{w} P(W \mid O) = \arg\max_{w} \left(\frac{P(O \mid W) \cdot P(W)}{P(O)} \right). \tag{3}$$

Al considerar nuevamente la probabilidad de producir una secuencia de parámetros igualmente posible (Gales y Young, 2008; Jelinek, 1997), el problema de maximizar la probabilidad conjunta de la clase y los parámetros de la señal se reduce a:

$$\hat{w} = \arg\max_{w} P(W \mid O) = \arg\max_{w} P(O \mid W) \cdot P(W). \tag{4}$$

Esto quiere decir que se debe encontrar aquella clase que maximice la verosimilitud de la señal dado un modelo de clases (probabilidad *a posteriori*) ponderada por la probabilidad de la clase (probabilidad *a priori*). Finalmente, la decisión en el *ASR*

será para aquel conjunto de palabras que maximice (2) (Rabiner et. al., 1996; Jelinek, 1997). La primera expresión del argumento de la maximización en (4), $P(O \mid W)$, entrega la probabilidad de que dada una secuencia de palabras, denotadas por W, haya generado una secuencia de vectores de parámetros O. Esta probabilidad se conoce como el modelo acústico del reconocedor de voz. El segundo término representa a la ocurrencia *a priori* de las clases (palabras), lo que es conocido como el modelo de lenguaje del sistema.

Para obtener los parámetros de los modelos estadísticos que se representan en (4) existe una etapa de entrenamiento del sistema, la que entrega como resultado un modelo acústico-fonético y un modelo de lenguaje. Estos, en general, se determinan aplicando algoritmos numéricos que satisfacen criterios estadísticos como el de máxima verosimilitud (*ML, maximum likelihood*). La descripción sobre la forma en cómo se modelan cada uno de estos componentes se detalla en las secciones siguientes.

2.1.1 Medidas de desempeño usadas en ASR

El término que se usa en la literatura para evaluar el rendimiento de un *ASR* es la tasa de palabras erradas o mal clasificadas o *WER* (*Word Error Rate*) definido como:

$$WER = \frac{S + I + D}{N} \tag{5}$$

donde N corresponde al número total de palabras de *test*, es decir las palabras que efectivamente fueron pronunciadas, y S, I y D son el número de palabras sustituidas, insertadas y eliminadas, respectivamente. Para medir la carga computacional en el reconocedor se usa el término *times-real-time*. Esto se define como la razón entre el tiempo que demora el proceso de reconocimiento y la duración de la señal. Esto es ampliamente utilizado para las comparaciones en tiempo real del algoritmo.

2.2 Evaluación automática de pronunciación basada en ASR

Mejoras en la habilidad del habla generan mejoras en la habilidad de escuchar y viceversa. Reforzar ambas habilidades en conjunto mediante la comunicación interactiva es el propósito principal de los sistemas CALL. Es este contexto, la tarea de un sistema de evaluación automática de pronunciación es la de obtener una medida de similitud, para una cierta palabra o frase, entre la pronunciación del usuario y la correcta o de referencia. En general esto se realiza contrastando la elocución de prueba con la secuencia de unidades fonéticas que representa la pronunciación correcta y con secuencias que representan errores comunes de pronunciación. La tecnología CAPT en plataformas de CALL debe entregar una calificación de pronunciación confiable por cada frase que se evalúe en al sistema. Este *score* debe estar altamente correlacionado con aquel que entregaría un experto en el idioma a evaluar. De esta forma, al tener una

correcta evaluación, el sistema será capaz de informarle al estudiante los errores cometidos y como mejorarlos.

La forma más básica de entregar una evaluación cuantitativa de la pronunciación de una elocución es mediante la verosimilitud entregada por el algoritmo de Viterbi descrito en la sección 2.4.4 (Franco et al., 1997). Es por esto que el método más usado para medir calidad de pronunciación es el basado en el alineamiento forzado de Viterbi, empleando como referencia la secuencia de estados de la palabra correctamente pronunciada. Una forma de mejorar la precisión de este método consiste en normalizar la verosimilitud por la duración de cada estado (Franco et al., 1997). Los métodos de verosimilitud de características espectrales segmentadas y de duración de estados presentan un buen desempeño al medir su correlaciones con evaluaciones subjetivas (realizadas por expertos humanos) (Neumeyer et. al., 1996). El mejor resultado, dentro de la literatura investigada, se obtiene mediante el método de la probabilidad posterior (Neumeyer et. al., 1996; Franco et al., 1997; Cucchiarini et al., 1998; Sevenster et al., 1998). En este método se realiza una evaluación posterior donde se compara para cada *frame* la probabilidad de la unidad fonética asociada al alineamiento con todas las demás unidades fonéticas del modelo. De este modo, se obtiene una medida de distancia entre el fonema seleccionado y el resto. Además, en la literatura se puede encontrar medidas de distancias entre la pronunciación no-nativa y la pronunciación objetivo

basadas en cuantización vectorial y en alineamiento temporal dinámico (Hamada et al., 1993).

Una variante al método de CAPT mediante el algoritmo de Viterbi forzado es sencillamente no restringir la búsqueda a sólo una palabra, sino que realizarla sobre palabras competidoras. Un ejemplo de esto es utilizar tecnología de ASR y ampliar la búsqueda a un conjunto finito de errores predefinidos (búsqueda 1:N). El uso explícito de la tecnología ASR usando búsqueda 1:N en CAPT ha sido recientemente mencionado en algunos trabajos como (Hamid y Rashwan, 2004; Abdou *et al.*, 2006; Moustroufas y Digalakis, 2007). En (Hamid y Rashwan, 2004; Abdou *et al.*, 2006) el modelo de lenguaje es generado por intermedio de reglas que permiten tomar en cuenta hipótesis de errores, supresiones y substituciones de posibles faltas en la pronunciación. En este caso la medida de confiabilidad para medir la calidad de la pronunciación está basada en un análisis de la duración de los fonemas en el alineamiento. En (Moustroufas y Digalakis, 2007) con el objetivo de utilizar modelos nativos y no nativos se usaron dos reconocedores en paralelo. Cabe notar que el vocabulario usado en ASR para aplicarlo en CAPT ha estado principalmente basado en reglas empíricas. Un ejemplo de esto se presenta en (Bonaventura *et al.*, 2000) donde por medio de reglas y entrenamiento de datos que contienen los errores típicos de una lengua determinada se genera el vocabulario competitivo. El problema que se presenta de inmediato es que el conocimiento acerca de los errores de pronunciación típicos es altamente dependiente del

dominio del sistema y de las palabras objetivos o las palabras con las cuales se está practicando el segundo idioma. Este tipo de vocabularios fuerza a una competición simultánea de pronunciaciones correctas e incorrectas lo que es crucial para que las tecnologías de ASR tengan éxito siendo aplicadas en CAPT.

El sistema empleado en esta tesis está basado en la generación automática del vocabulario competitivo de una palabra objetivo (Molina et al, 2009). Esto permite que el ASR pueda contrastar en forma simultánea el léxico competitivo y la correcta pronunciación de una palabra objetivo dada. Además, se incorpora el uso modelos acústicos entrenados con fonemas no-nativos para la generación de palabras competidoras (en el caso de esta tesis el idioma nativo es el inglés y el no-nativo es el español). Con esto es posible contrastar la pronunciación a evaluar con modelos de fonemas nativos y no-nativos. Es importante recalcar que el método usado para generar el set de palabras competidoras no requiere de ningún tipo de modelamiento *a priori* o información acerca de reglas que representen los errores comunes en pronunciación dada una lengua y además la metodología enseñada para dicha generación no depende de la palabra objetivo. Como resultado, el sistema de ASR queda intrínsecamente habilitado para la integración de nuevo material de manera más eficiente y directa. Debido al hecho de que el ASR puede comparar una señal grabada con un vocabulario competitivo y una correcta pronunciación de manera más eficiente que una comparación 1:1, el ASR está más capacitado para la extracción simultánea de diversas métricas. En efecto, un

simple análisis sobre el resultado de las N-mejores hipótesis que se acostumbra entregar en un reconocedor de voz puede entregar variadas mediciones objetivas o métricas acerca de una señal en particular. Finalmente, el sistema hace un mapeo de las salidas del ASR a una medida *score* de la calidad de la pronunciación.

2.2.1 Medidas de desempeño usadas en CAPT

En el caso de CAPT, la medida de desempeño a usar es la correlación entre los *scores* que entrega del sistema de evaluación y el score *s*ubjetivo, que se obtiene a través de la evaluación de la pronunciación realizada por un grupo de expertos. Al evaluar un conjunto de señales, el coeficiente de correlación se calcula como (Molina et al., 2009):

$$Correlación = \frac{Cov\Big(\text{score objetivo, score subjetivo}\Big)}{Var\Big(\text{score objetivo}\Big) \cdot Var\Big(\text{score subjetivo}\Big)}. \tag{6}$$

Mientras más alto es el coeficiente de correlación, mejor es el desempeño del sistema de evaluación de pronunciación.

2.3 Verificación de locutor

Dentro de los sistemas de reconocimiento de identidad basados en información biométrica se destacan, entre otros, aquellos basados en voz, iris y huellas dactilares. En particular los métodos biométricos basados en información de la voz humana se

denominan técnicas de reconocimiento de locutor. Estas se dividen en dos grandes áreas: identificación de locutor y verificación de locutor (VL). Un sistema de identificación de locutor asociará a un usuario la identidad de alguno de los individuos registrados en el sistema, es decir, la salida del sistema será la identidad del que mejor se aproxime a las características de la señal de voz. Por otra parte, un sistema de verificación de locutor debe decidir si un usuario que declara una cierta identidad es o no quien dice ser (Subramanian et al., 2010; Furui, 1994). La señal de voz de un locutor cualquiera es comparada con el modelo del individuo cuya identidad fue declarada. Así, si el modelo de locutor y la pronunciación coinciden dentro de los límites permitidos (umbral de decisión), la identidad será aceptada y en caso contrario será rechazada.

Existen diversos tipos de sistemas de verificación de locutor. Entre ellos se pueden distinguir los sistemas texto-dependientes (TD-SV) y texto-independientes (TI-SV). Los primeros requieren que el usuario pronuncie una palabra o frase determinada por el sistema. El segundo tipo de plataforma está diseñada para realizar el proceso de verificación cualquiera sea la palabra o frase pronunciada. Se pueden distinguir dentro de cada uno de estos tipos de sistema, aquellos de pronunciación continua o los de palabra aislada. En estos últimos las palabras deberán estar separadas entre sí por pequeños instantes de silencio. Todo sistema de verificación de locutor cuenta con una base de datos de usuarios registrados, denominados "clientes". Esta base de datos está compuesta por modelos que representan las características del habla de cada uno de los

clientes. Estos modelos se consiguen mediante el procesamiento datos capturados en sesiones de entrenamiento en las cuales el usuario del sistema pronunciará varias frases.

Como se mencionó, un sistema de SV puede clasificarse como dependientes o independientes del texto (Subramanian et al., 2010; Furui, 1994). Ambos tipos de sistemas tienen diferentes aplicaciones en el marco de una plataforma biométrica. Los sistemas TD-SV están más enfocados en aplicaciones de interés comercial tales como el *password* o huella de voz para control de acceso en diálogos telefónicos. Por otro lado, las aplicaciones de interés para las tecnologías de TI-SV son las relacionadas con sistemas de identificación forense. En la presente tesis se trabajará con un sistema de verificación de locutor texto dependiente que opera en un diálogo telefónico con datos limitados de entrenamiento y elocuciones cortas en la etapa de prueba.

2.3.1 Verificación de locutor texto-dependiente basada en HMM

Si dada una señal o elocución se considera que el *frame* en el instante i es representado por un vector de parámetros espectrales $O_i = \left[O_{i,1},...,O_{i,n},... ,O_{i,N} \right]$, donde N es el número total de parámetros y $O_{i,n}$ es el n-ésimo parámetro en el *frame* i, entonces una elocución estará representada por una secuencia de vectores O:

$$O = \left[O_1, O_2, \ . \ . \ . , O_I \right] \tag{7}$$

donde I es la duración en *frames* de la señal.

La tarea de clasificación de patrones acústicos para un sistema de verificación de locutor consiste en medir la verosimilitud entre el modelo del locutor j y la secuencia de vectores de observación O del locutor k. La verosimilitud obtenida es comparada con un umbral de decisión. De esta forma se decide aceptar o rechazar la afirmación de identidad recibida del usuario.

En un sistema basado en modelos ocultos de Markov (HMM), la medida usada para evaluar una secuencia de observación O corresponde a la probabilidad de que esta haya sido pronunciada por el cliente j (S_j) cuya identidad dice tener el usuario que se está verificando, dado el modelo de referencia del cliente j (λ_j). Los términos O y λ_j son generados a partir de la señal de entrada y las elocuciones de entrenamiento, respectivamente. Utilizando el teorema de Bayes esta medida de probabilidad, $\Pr\left(S_j \middle| O, \lambda_j\right)$, puede escribirse como:

$$\Pr\left(S_j \middle| O, \lambda_j\right) = \frac{\Pr\left(O \middle| S_j, \lambda_j\right) \cdot \Pr\left(S_j \middle| \lambda_j\right)}{\Pr\left(O \middle| \lambda_j\right)}. \tag{8}$$

Debido a que $\Pr\left(S_j\right)$ y $\Pr\left(O\right)$ pueden ser considerados constantes e independientes del locutor, el término relevante para estimar la probabilidad

$\Pr\left(S_j \middle| O, \lambda_j\right)$ corresponde al valor de la verosimilitud definida por $\Pr\left(O \middle| S_j, \lambda_j\right)$, o simplemente, $\Pr\left(O \middle| \lambda_j\right)$.

El estado del arte en tecnologías de verificación de locutor texto-dependientes operando en canal telefónico muestra sistemas que alcanzan un EER entre 0,3% a 1% usando señales de 3 a 5 segundos de duración y con niveles bajos de ruido. En general estos sistemas utilizan para los procesos de enrolamiento y verificación elocuciones de 10 a 30 y de 2 a 10 segundos de duración, respectivamente (Becerra Yoma y Villar, 2002; Sivakumaran et al. 2003; Mahadeva Prasanna et al. 2004; Nealand et al., 2005).

2.3.2 Normalización de la verosimilitud

En un sistema de verificación de locutor las decisiones son tomadas calculando la verosimilitud de la elocución de verificación con respecto al modelo HMM de la identidad que un locutor afirma. En el caso de un sistema de verificación de locutor texto dependiente, en el cálculo del valor de verosimilitud también se considera la información lingüística de la señal de testeo. De esta forma, el valor de la verosimilitud deseada presentará una fuerte dependencia de la variabilidad natural del locutor, por lo que un umbral de decisión estándar es difícil de fijar. Una forma de enfrentar el problema de la variación del umbral de decisión es aplicar normalización de la verosimilitud (Furui, 1997; Bimbot, et al., 2004; Matsui y Furui, 1993; Higgins, et al.,

1991). Esta normalización puede mostrar mejoras significativas en el desempeño del sistema, al generar una decisión por contraste, y se aplica evaluando la relación entre las verosimilitudes de la elocución de *test* con respecto al HMM de referencia del usuario objetivo, el que representa la hipótesis de cliente y con respecto a un HMM que representa la hipótesis de impostor (Subramanian et al., 2010; Furui, 1994):

$$L\left(O\right) = \frac{\Pr\left(O \mid \lambda_S\right)}{\Pr\left(O \middle| \lambda_{\bar{S}}\right)} \tag{9}$$

donde λ_S y $\lambda_{\bar{S}}$ son los modelos que representan las hipótesis de cliente (SD, *speaker-dependent*) y de impostor, respectivamente. La estimación de la verosimilitud normalizada se realiza en el dominio logarítmico:

$$\log L\left(O\right) = \log\left[L\left(O\right)\right] \tag{10}$$

donde el término $\log L\left(O\right)$ se denomina verosimilitud logarítmica (*loglikelihood*) normalizada.

La probabilidad que la secuencia de vectores de observación O corresponda al modelo de referencia del locutor $\Pr\left(O \mid \lambda_S\right)$ se calcula estimando la verosimilitud de O en el modelo SD. Por su parte la probabilidad $\Pr\left(O \mid \lambda_{\bar{S}}\right)$, denominada término normalizador, corresponde a la verosimilitud calculada con respecto a un modelo general de impostores o modelo *speaker-independent* (SI). Este modelo, se entrena idealmente

con elocuciones pertenecientes a una gran cantidad usuarios que no se encuentren registrados en el sistema. Una alternativa a usar un único modelo SI es entrenar un conjunto de modelos, cada uno entrenado con uno de los locutores incluido en el modelo SI. De esta forma, se puede calcular $\Pr\left(O \mid \lambda_{\bar{S}}\right)$ como el promedio de las verosimilitudes de la señal de prueba en el grupo de locutores más probable (o *cohort* de locutores). Para independizar el cálculo de $\log L\left(O\right)$ de la duración de las señales de voz, se divide el resultado por el número total de *frames* total de la señal de voz, *I*:

$$\log L\left(O\right)' = \frac{\log L\left(O\right)}{I}. \tag{11}$$

El uso de normalización de la verosimilitud ha demostrado una reducción significativa del error provocado por la presencia de ruido convolucional al usar distintos tipos de micrófono. Existen variadas formas adicionales de aplicar una normalización a la verosimilitud o *score* de una elocución de verificación. Cada una de estas ha sido diseñada con algún objetivo en particular (eliminar dependencia al locutor, compensación de *mismatch* de canal, etc.).

2.3.3 Medidas de desempeño usadas en SV

En un sistema de verificación de locutor sólo existen dos respuestas posibles: aceptar o rechazar al usuario testeado. Lo que lleva a cuatro casos posibles, dos correctos y dos

errados: aceptar un cliente, rechazar un impostor, aceptar un impostor y rechazar un cliente. Los dos primeros casos corresponden a respuestas correctas por parte del sistema de verificación de locutor, mientras que las dos últimas opciones son erradas. Estos errores corresponden a los denominados errores de "falsa aceptación" (FA) y "falso rechazo" (FR), respectivamente. El valor en el que el ajuste del sistema iguala estos niveles de error es denominado *Equal Error Rate* (EER), este valor es comúnmente utilizado para medir el desempeño en sistemas de verificación de locutor y otros sistemas biométricos. El nivel de umbral de decisión en el que el sistema opera bajo una tasa de error igual al EER se denomina TEER (*Threshold Of Equal Error Rate*). El desempeño del sistema se puede representar gráficamente generando curvas de FA y FR en función del umbral de decisión, como se puede ver en la Figura 1.

Otra herramienta utilizada para medir el desempeño de un sistema de verificación de locutor es la curva DET (*Detection Error Tradeoff*) (Martin et al., 1997). Esta curva se genera computando el error de FR y FA en un rango amplio de valores del umbral de decisión. En los ejes vertical y horizontal se ubican las tasas de error de FA y FR, respectivamente (NIST, 2006). Dados los niveles de error manejados por este tipo de sistemas, estas curvas generalmente se grafican en escala logarítmica (gráfico *log-log*). La Figura 2 muestra un ejemplo de este tipo de curva. En el dominio lineal, la curva que representa la relación FA vs. FR es comúnmente denominada curva ROC (*Receiver Operating Characteristic*), el valor del área bajo la curva ROC sirve como indicador la

habilidad discriminativa del sistema bajo el rango completo de valores de umbral de decisión en el que este es probado. Mientras mejor desempeño presentado por el sistema, menor será el área bajo la curva ROC.

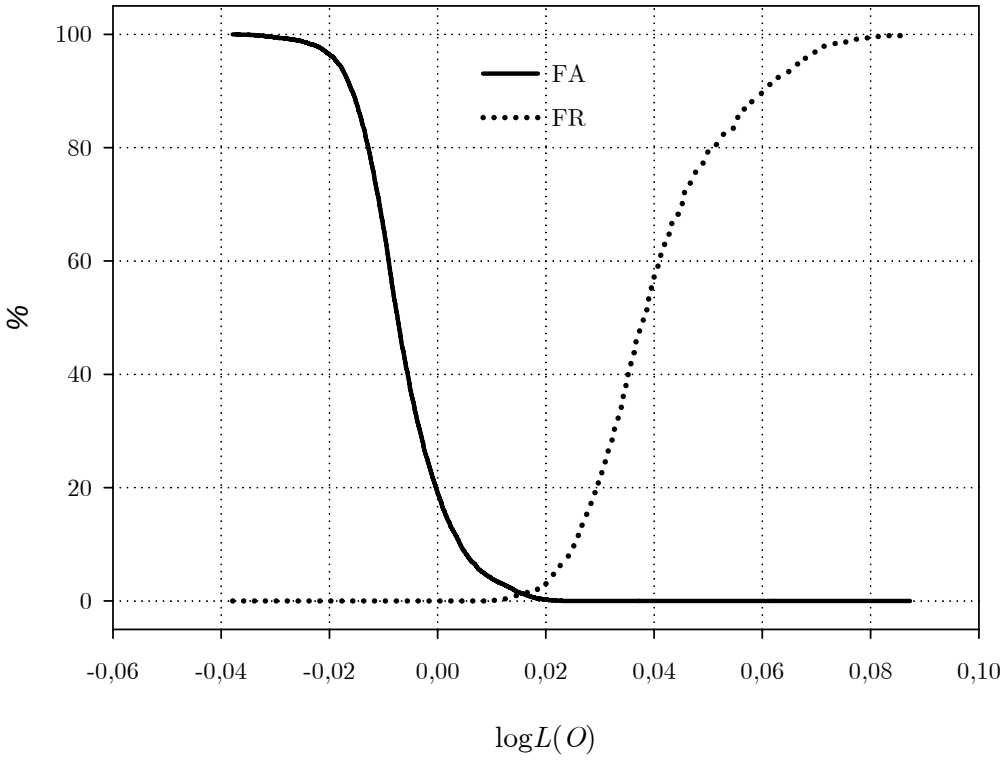


Figura 1. Curvas de falsa-aceptación y falso-rechazo en función del umbral de decisión.

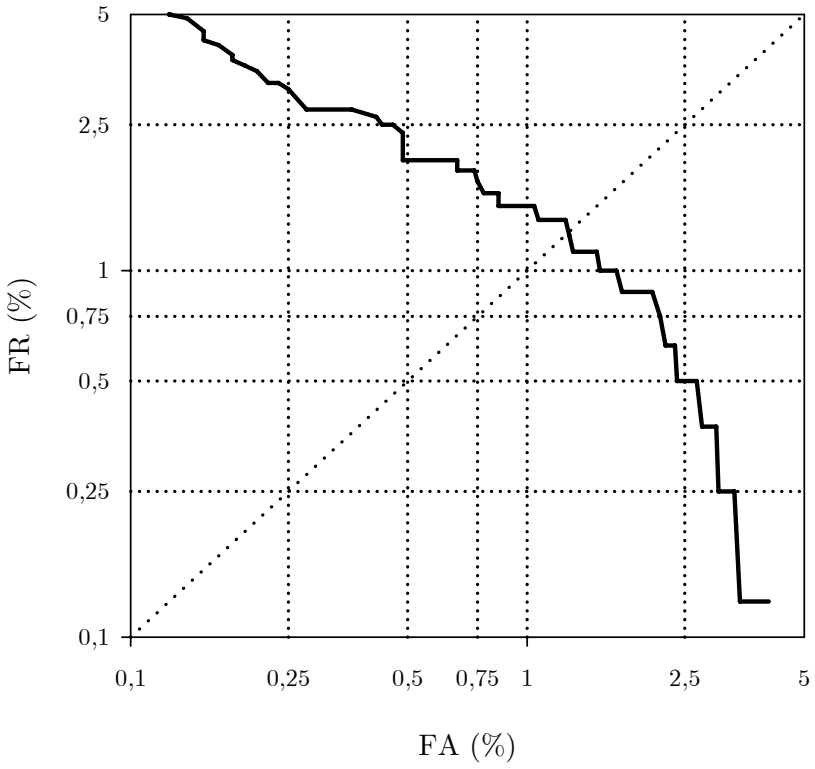


Figura 2. Curva DET: FR en función de FA.

2.4 Técnicas usadas en reconocimiento de patrones acústicos

2.4.1 Parametrización acústica

Para realizar la tarea de caracterizar una señal de voz, se deben tener en cuenta dos importantes factores: a) la señal de voz es un proceso estocástico no-estacionario; y b) las variaciones temporales entre señales que contienen la misma información fonética.

La variabilidad temporal en las señales de voz puede deberse a factores relacionados con el locutor, el entorno y la fuente o medio de captura de la voz. El primer factor es denominado variabilidad intra-locutor (Yang et al., 1996) y se describe como la variación entre elocuciones de un mismo individuo de la información acústico fonética que se extrae de la señal voz. De forma análoga se desprende el concepto de variabilidad inter-locutor, relacionada con las variaciones entre elocuciones pertenecientes a un grupo amplio (o universo) de locutores. El siguiente factor que puede introducir una componente de variabilidad no deseada al momento de parametrizar una señal de voz, es la cantidad de ruido ambiental y la variabilidad de este en el tiempo. Finalmente se tiene el efecto del medio de captura de la voz o canal de comunicación. Este factor que puede generar fuertes distorsiones en elocuciones con idéntica información fonética de un mismo usuario

En la Figura 3 se puede apreciar el proceso completo de extracción de características acústicas. El método usado para la parametrización de señales de voz se basa en el cálculo de coeficientes *cepstrales*. Analizar una señal de voz en el dominio cepstral o *cepstrum* contribuye a realzar las componentes asociadas a los formantes del tracto vocal, incluso en señales con ruido. Los parámetros basados en el *cepstrum* se han convertido en uno de los métodos más usados en clasificación de patrones acústicos y ya se ha transformado en un estándar dentro del área de procesamiento de voz (Forsyth, 1995). Antes de efectuar la extracción de parámetros generalmente se le da un

tratamiento de pre-procesamiento a la señal de voz. Esta etapa tiene por objeto realzar la información de voz por sobre otro tipo de información que pueda contener la señal. De esta forma dejar todas las señales a analizar en condiciones similares para su caracterización. Esto se puede lograr mediante las siguientes tareas: detección del inicio y fin de la información de voz; supresión de segmentos de silencio; y, compensación de ruido aditivo y/o convolucional.

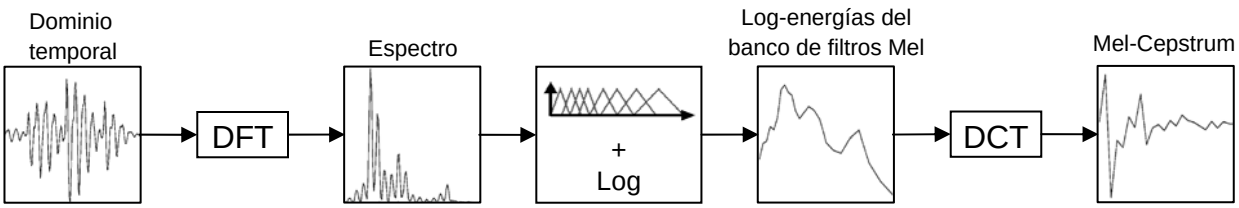


Figura 3. Diagrama de bloques que describe el proceso de parametrización cepstral del *frame* de una señal de voz.

La primera parte del pre-procesamiento es la conversión análogo-digital de la señal de voz. Esta tarea es realizada por el hardware de captura o por interfaces telefónicas. Luego la señal es procesada por un filtro inicio-fin el que elimina la información irrelevante que esta antes y después del primer y último pulso de voz detectados (Lamel et al., 1981; Savoji, 1989). El siguiente paso es dividir la señal en segmentos que pueden ser considerados estadísticamente estacionarios, los que se denominan ventanas o *frames*. Con esto se busca lograr una caracterización de la señal ventana a ventana.

Para esta segmentación generalmente se toman intervalos de 10 a 30 [mseg], los que pueden tener un traslape de hasta 50% entre ventanas consecutivas. Para evitar las distorsiones en el análisis espectral que pueden generar las discontinuidades en los límites de cada ventana, se utiliza la técnica de enventanado de *Hamming* (Picone, 1993). Este es el último paso de la etapa de pre-procesamiento.

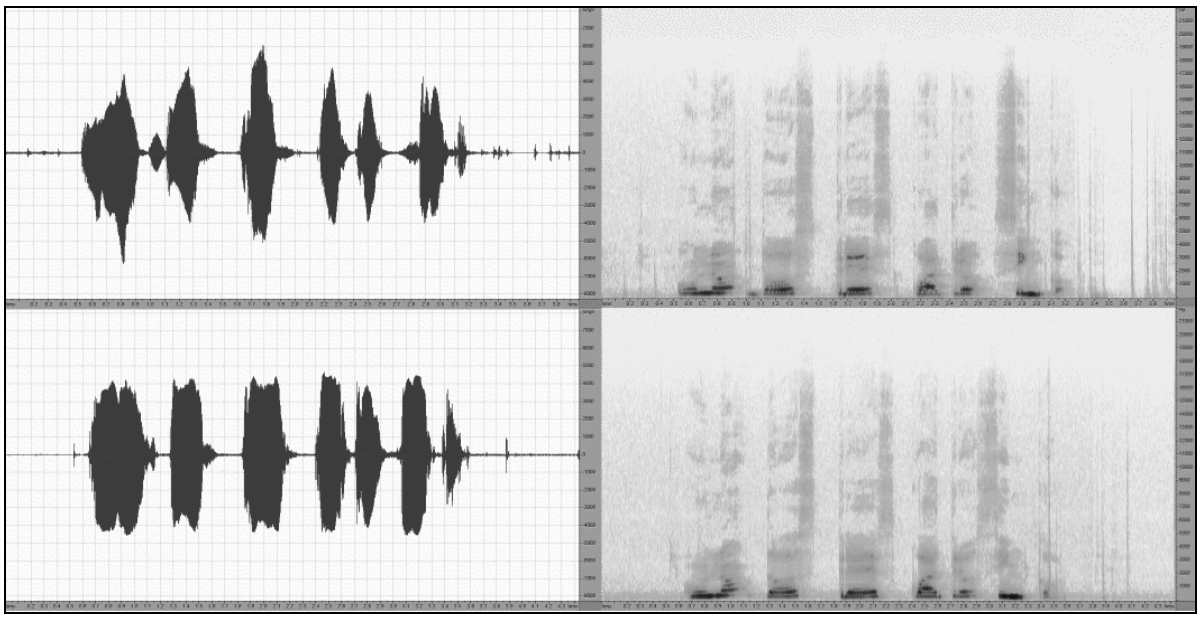


Figura 4. Paralelo en el dominio temporal (izquierda) y espectral (derecha) de dos señales de un mismo locutor pronunciando la secuencia de dígitos "1-2-3-4-5", las señales fueron muestreadas a 8KHz. El eje horizontal representa el tiempo (muestras). En los espectrogramas el eje vertical representa frecuencia (en Hertz), el nivel de energía asociado a la frecuencia se representa por colores (blanco a azul, menor a mayor energía).

La etapa de parametrización comienza con un análisis espectral por cada *frame,* el que consta de un análisis por transformada discreta de Fourier (DFT, *Discrete Fourier Transform*) y de la aplicación de bancos de filtros por bandas. La utilización de estos filtros se debe a que la percepción auditiva humana no es capaz de distinguir frecuencias individuales, sino que capta franjas de frecuencias. Además la respuesta del sistema auditivo humano en el espectro de frecuencias no es lineal, lo que lleva a utilizar una escala en que la concentración de las frecuencias producto del filtrado simule la capacidad discriminativa del oído humano (en un rango de frecuencias aproximado de entre 300 y 3400 [Hz]). Una de las escalas más utilizada para estos efectos es la escala Mel. En (12) se describe la transformación asociada a esta escala, para un valor de frecuencia *f*:

$$Mel(f) = 2595 \cdot \log_{10}(1 + \frac{f}{700}) \quad ; \quad f \text{ en Hertz}\,. \tag{12}$$

El banco de filtros se compone de un conjunto de funciones triangulares y simétricas de ganancia unitaria para la frecuencia central, con superposición de 50% y un ancho de banda constante en escala Mel. Para cada filtro se calcula el logaritmo de energía, con esto se obtienen las características LFBE (Log *Filter-Bank Energy*). A continuación, se realiza el cálculo de coeficientes cepstrales en escala Mel (MFCC, *Mel Frequency Cepstral Coefficient*), para esto se aplica la transformada coseno discreta (DCT, *Discrete Cosine Transform*) sobre los parámetros LFBE. En procesamiento de voz, se

obtiene un vector de parámetros MFCC para cada *frame* a analizar, es decir, una señal de voz es caracterizada como una secuencia de vectores de observación en el dominio MFCC.

Como se puede ver en (Bimbot et al., 2004, Furui, 2005) el uso de las características basadas en MFCC es predominante en las áreas de reconocimiento de voz/locutor y se mantiene relativamente invariante. Sin embargo, la tarea de generar parámetros robustos a diversas condiciones adversas es aún un problema abierto y ampliamente abordado. En la literatura especializada reciente es posible encontrar técnicas de parametrización cepstral robusta a factores como: el ruido ambiental (Damper y Higgins, 2003; Skosan y Mashao, 2006); la reverberancia (Thomas et al., 2008; Wolfel, 2009); y la distorsión de canal (Heck et al., 2000; Tufekci, 2007). La Figura 4 muestra ejemplos de variabilidad a nivel espectral - temporal en dos señales de voz de un mismo locutor, capturadas con distintos micrófonos y bajo distintas condiciones de ruido ambiental. Los métodos de parametrización robusta a la influencia del canal, relacionados con los objetivos de este trabajo de investigación, serán descritos con mayor detalle en el capítulo 3.

2.4.2 Modelamiento acústico con modelos ocultos de Markov

Las cadenas de Markov consisten en una secuencia de estados conectados por probabilidades de transición en las cuales se van generando las observaciones. Cada

estado tiene una función de distribución de probabilidad la cual entrega la verosimilitud de que una observación haya sido generada por él (Rabiner, 1989). Los modelos ocultos de Markov o HMM (*"Hidden Markov Models"*), han sido ampliamente utilizados en los sistemas de reconocimiento de voz y locutor. En particular, los modelos más usados son los de de primer orden, donde el estado actual de una señal depende solamente del anterior (Rabiner, 1989; Jelinek, 1997). La salida de una secuencia de estados permanece oculta en el proceso y sólo se conoce el conjunto de parámetros acústicos de la señal producidos por la secuencia de estados. La topología usada en HMM aplicado *a ASR* se denomina *"left-to-right"*, es decir, permiten sólo transiciones al siguiente o al mismo estado. Con esto se limitan los saltos o retrocesos.

Un HMM queda definido por: las probabilidades de transición de estados, la función de distribución de probabilidad y las probabilidades iniciales (Rabiner, 1989). Las probabilidades de transición para un HMM con M estados debe cumplir con la siguiente restricción:

$$\sum_{k=1}^{M} A(j,k) = 1 \qquad \forall j = 1,...,M \tag{13}$$

donde $A(j,k)$ corresponde a la probabilidad de estar en el estado k dado que el anterior estado fue j. La función de distribución de probabilidad (f.d.p.) de que una observación haya sido generada por el estado k se representa en (14). Cabe mencionar que en la tarea de reconocimiento de voz es usual utilizar poblaciones para modelar las f.d.p. de

estados. En este caso suponemos una población de G distribuciones normales independientes, cada una con un peso de probabilidad asignado, y restringido por (15):

$$b_k(O_i) = \sum_{g=1}^{G} \left\{ p_g \cdot \prod_{n=1}^{N} \left[\left(2 \cdot \pi\right)^{-0.5} \cdot \left(Var_{k,g,n}\right)^{-0.5} \cdot e^{-\frac{1}{2}\frac{\left(O_{i,n}^{o} - E_{k,g,n}\right)}{Var_{k,g,n}}} \right] \right\} \tag{14}$$

$$\sum_{g=1}^{G} p_g = 1 \tag{15}$$

donde *k, g, n* son los índices para el estado, la componente Gaussiana y el coeficiente del vector de observación, respectivamente; p_g corresponde al peso de probabilidad de la población *g-esima*; $O_i = \left[O_{i,1}, O_{i,2}, ..., O_{i,N} \right]$ es el vector de observación de la señal acústica de dimensión N en el instante i; $E_{j,g,n}$ y $Var_{j,g,n}$ son la media y varianza para un determinado modelo en el estado k, componente Gaussiana g y coeficiente cepstral n. Cabe mencionar que la matriz de covarianza de las Gaussianas es supuesta diagonal, es por esta razón que se hace mención a la varianza.

Un HMM representa una unidad fonética. En este caso se utiliza los denominados tri-fonemas, estos se componen de una unidad fonética (formantes) central más dos segmentos de fonemas que preceden y suceden a la unidad central (Schwartz, 1985). Una palabra está formada por una secuencia de tri-fonemas, lo que a su vez lleva a que cada palabra está formada por una secuencia de HMM, que si lo generalizamos aún más, se llega a que una frase también lo es. Así, la probabilidad de que un vector de

parámetros acústicos O haya sido generado por el HMM de la secuencia de palabras W queda dada por (Gales y Young, 2008; Rabiner, 1989; Jelinek, 1997):

$$P(O|W) = \sum_{S \in \Lambda} P(O, S|W) = \sum_{S \in \Lambda} P(S|W) \cdot P(O|S) \tag{16}$$

donde $S = \left[s_1, s_2, ..., s_I \right]$ representa cualquier secuencia de estado dentro del conjunto Λ; el conjunto Λ son todas las posibles secuencias de estados que son capaces de generar la secuencia de vectores de parámetros acústicos O. Si ahora al descomponer (16) según la descripción de un HMM y se reemplazar en (4), se obtiene:

$$\begin{aligned} \hat{W} &= \arg \max_{w,S} \left\{ P(W) \cdot P(O|W) \right\} \\ &= \arg \max_{w,S} \left\{ P(W) \cdot \left(A(0, S_1) \cdot \prod_i A(S_i, S_{i+1}) \right) \cdot \left(\prod_i b_{S_i}^W (O_i) \right) \right\} \end{aligned} \tag{17}$$

Un ejemplo de arquitectura HMM se muestra en la Figura 5.

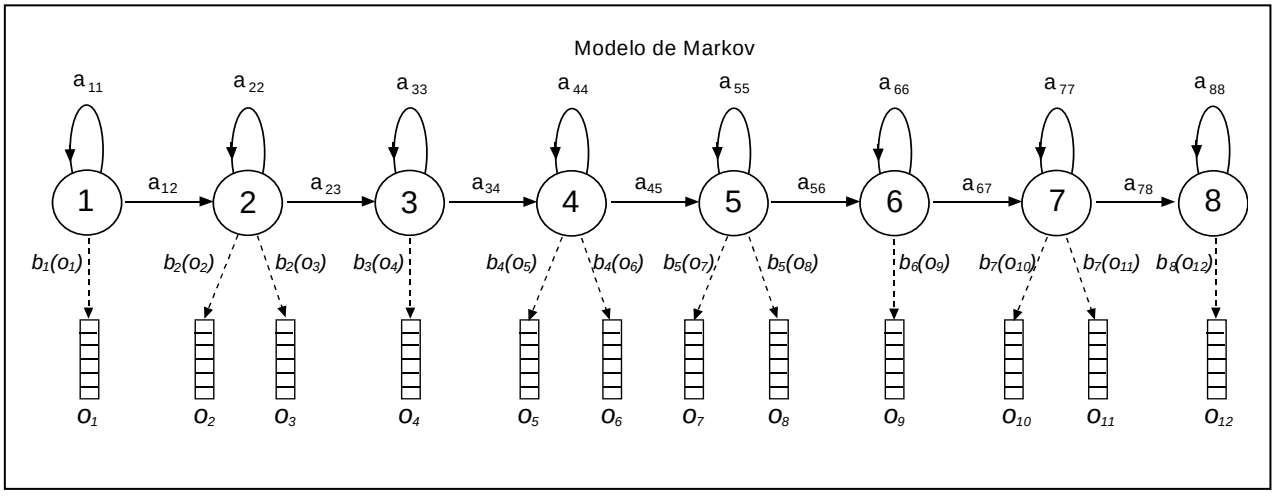


Figura 5. Ejemplo de topología izquierda derecha sin salto de estado de un HMM.

2.4.3 Modelo de lenguaje

El modelo de lenguaje entrega información a priori en la tarea de reconocimiento de la voz, $P(W)$ en (4). Los métodos para estimarlo pueden variar desde ser un algoritmo de reglas gramaticales, hasta ser netamente una representación estadística del lenguaje utilizado. Los más usados son los modelos estocásticos de tipo *M*-grama. Esto considera que la ocurrencia de una palabra dentro de una sucesión de ellas está condicionada a la probabilidad de las *M*-1 palabras anteriores. Un modelo M-grama se representa como:

$$P(w_1, w_2, w_3, \ldots, w_N) = \prod_{i=1}^{N} P\left(w_m \left| w_{m-M+1}, \ldots, w_{m-2}, w_{m-1}\right.\right). \tag{13}$$

El criterio para la estimación de los parámetros que determinan el modelo de lenguaje es el estimador de máxima verosimilitud. En él se maximiza la probabilidad de observar las secuencias de algún conjunto de entrenamiento. Uno de los problemas de los modelos estocásticos es que no considera probabilidad para las secuencias de palabras que no se encuentran en el conjunto de entrenamiento. Según la definición, estas probabilidades quedan en cero para aquellos casos en que no existe ocurrencia. El problema de generalización del modelo de lenguaje es tratado con diversas técnicas. Por ejemplo, existe el modelo de lenguaje a nivel de clases, depuración de parámetros o modelos de lenguaje por palabras (Gales y Young, 2008; Laurila et. al., 1998; Becchetti y Prina, 1999).

2.4.4 El algoritmo de Viterbi

Una secuencia de estados (S) determina inmediatamente una secuencia de HMMs, estos a su vez determinan la secuencia de palabras reconocidas (W). Se debe resolver la tarea de encontrar la secuencia de estados óptima que genera un vector de parámetros acústicos. Para esto, se necesita evaluar todas las posibles secuencias de estado para cada instante de tiempo en la señal de voz. Como es de suponer, implica una excesiva carga computacional. Para minimizar la carga existe el algoritmo de Viterbi. El método consiste en ir optimizando a nivel local las secuencias de estado. Con ello, en forma inductiva, se resuelve el problema de optimización global (Gales y Young, 2008; Jelinek, 1997). El algoritmo de Viterbi al optimizar a nivel local va descartando secuencias. Con ello logra reducir el campo de búsqueda y generar un algoritmo más viable desde el punto de vista computacional. Sea $\delta_i(j)$ la probabilidad de observar la secuencia de parámetros O hasta el tiempo i junto con la secuencia de estados más verosímil hasta i y que además, el estado s en i sea j:

$$\delta_i(j) = \max_{s_1,s_2,\ldots,s_{i-1}} P\left(s_1,s_2,\ldots,s_{i-1},s_i = j,o_1,o_2,\ldots,o_i \mid \lambda_w\right). \tag{18}$$

Suponiendo recursividad se obtiene:

$$\begin{aligned}\delta_i(j) &= \max_{s_1,s_2,\ldots,s_{i-1}} P\left(o_i,s_i = j \mid s_1,\ldots,s_{i-1},o_1,\ldots,o_{i-1},\lambda_w\right)\cdot P\left(s_1,\ldots,s_{i-1},o_1,\ldots,o_{i-1} \mid \lambda_w\right)\\ &= b_j(o_i)\cdot \max_{s_1,s_2,\ldots,s_{i-1}} P\left(s_1,s_2,\ldots,s_{i-1},o_1,o_2,\ldots,o_{i-1} \mid \lambda_w\right)\\ &= b_j(o_i)\cdot \max_{s\in\Gamma}\left(a(s,j)\cdot\delta_{i-1}(s)\right)\end{aligned} \tag{19}$$

Para llegar al cálculo de $\delta_i(j)$ se debe evaluar todos los posibles caminos para llegar a $s_i = j$. Estos posibles caminos están agrupados en el espacio Γ, por lo tanto Γ es un conjunto de secuencias de t estados, es decir $\Gamma \in \Re^t$. λ_W es el modelo de la secuencia de palabra W hasta el instante i. el término $a(s, j)$ determina la probabilidad de transición del último estado en la secuencia S al estado dado en i que es j.

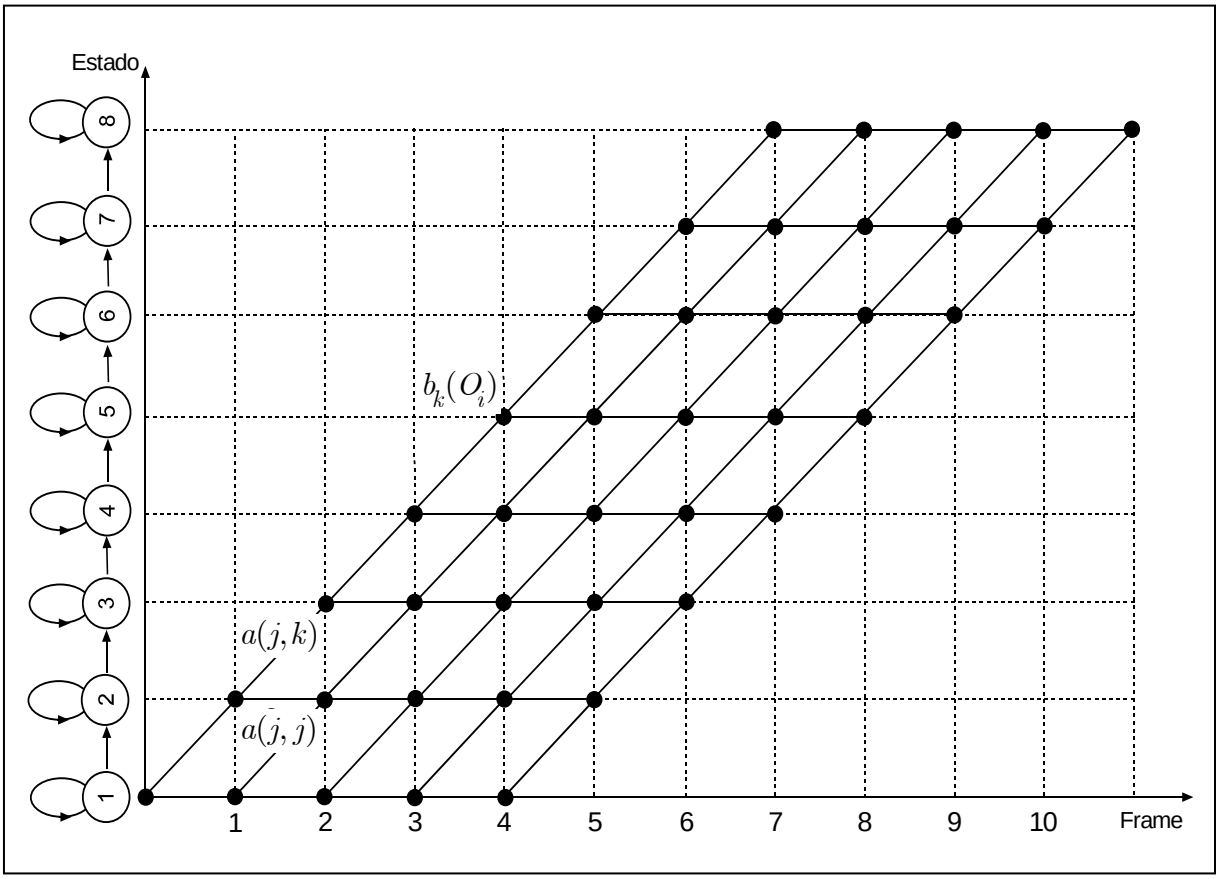


Figura 6. Representación gráfica del algoritmo de Viterbi.

Asumiendo la recursividad del algoritmo, si buscamos la secuencia de estados más verosímil para llegar a s_i, la secuencia anterior debe ser $\delta_{i-1}(s)$ donde s pertenece al conjunto Γ. Con esto, en forma recursiva, se llega a que $\delta_i(j)$ es la secuencia más probables de estados para llegar al tiempo i con el estado j (Gales y Young, 2008; Jelinek, 1997). Luego, para obtener la información del estado en el cual se está en el tiempo i se define la función $\psi_i(j)$, que a medida que se avanza en el algoritmo guardará la información del estado óptimo. Finalmente, el algoritmo se define según la secuencia que se describe a continuación.

1) Inicialización:

$$\begin{aligned} &\delta_1(j) = \pi_j \cdot b_1(o_1) && j \in \Gamma \\ &\psi_1(j) = 0 \end{aligned} \tag{20}$$

2) Recursión:

$$\begin{aligned} &\delta_i(j) = b_j(o_i) \cdot \max_{s\in\Gamma}\left(a(s,j)\cdot\delta_{i-1}(s)\right) && j \in \Gamma \quad 2 \le i \le k \\ &\psi_i(j) = \arg\max_{s\in\Gamma}\left(a(s,j)\cdot\delta_{i-1}(s)\right) && j \in \Gamma \quad 2 \le i \le k \end{aligned} \tag{21}$$

3) Finalización: se determina la probabilidad de la secuencia de estados más verosímil y el último estado de dicha secuencia:

$$P_{\max} = \max_{j\in\Gamma}\left(\delta_k(j)\right) \quad ; \quad \hat{s} = \arg\max_{j\in\Gamma}\left(\delta_k(j)\right) \tag{22}$$

4) Alineamiento: se reconstruye la secuencia de estados más verosímil.

$$s_i = \psi_{i+1}(s_{i+1}) \qquad t = 1,\ldots,k-1 \tag{23}$$

La variante del algoritmo de Viterbi usada en sistemas de verificación texto-dependiente, se denomina "Algoritmo Forzado de Viterbi". Esta consiste en limitar la búsqueda de Viterbi a sólo una frase y que el algoritmo sólo se encargue de alinear los estados dentro de esa transcripción. El objetivo de esta variante del algoritmo es estimar la probabilidad de que un vector de parámetros acústicos haya generado cierta secuencia "forzada" de estados.

2.5 Robustez a la variabilidad en el canal de comunicaciones

Si se analiza el proceso de la señal de voz captada por un transductor, este puede ser resumido como: el paso de un flujo de aire sobre el tracto vocal del locutor; la radiación de dicho sonido al exterior; su propagación acústica hasta el transductor; y el paso de la señal de voz a través de sistemas electrónicos con sus respectivas respuestas en frecuencia y fase (amplificadores, filtros, canales telefónicos, etc.). Es posible observar que cada uno de los elementos que compone esta cadena introduce su propio efecto, los que claramente son perjudiciales para cualquier sistema de clasificación de patrones. El concepto de "canal de comunicación" considera todas aquellas etapas que componen el proceso de captura y transmisión de la voz, las que potencialmente modificarán la información que contiene la señal. Algunos elementos que pueden componer un canal de comunicación son: el dispositivo de captura de la voz (micrófono o *handset*); el medio

físico de transmisión; el procesamiento en las centrales telefónicas; conversiones análogo/digitales y digital/análogas; y procesos de codificación y decodificación.

2.5.1 Modelo del canal de comunicación

El modelo más sencillo de un canal de comunicaciones más simple, se compone de un filtro lineal invariante de respuesta impulsiva $h[t]$ y una señal de ruido aditivo $n[t]$. Estos componentes se suman y distorsionan de forma lineal la señal $x[t]$, respectivamente (Wolfel, 2009; Oppenheim et al., 1997). La Figura 7 describe el modelo.

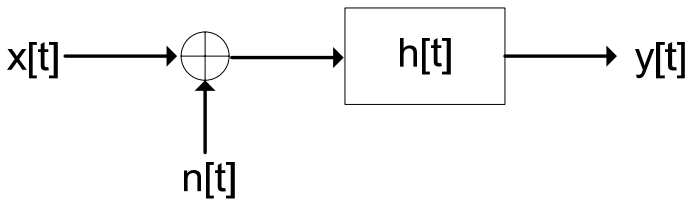


Figura 7. Modelo de canal de transmisión de la señal de voz.

La señal $y[t]$ a la salida del modelo de canal se puede estimar como

$$y[t] = \Big(x[t] + n[t]\Big) * h[t] \tag{24}$$

al aplicar transformada de Fourier en (24) se obtiene:

$$Y[\omega] = \Big(X[\omega] + N[\omega]\Big) \cdot H[\omega] = X[\omega] \cdot H[\omega] + N[\omega] \cdot H[\omega] \tag{25}$$

donde $N[\omega]$ representa el ruido aditivo y $H[\omega]$ el ruido convolucional. Tomando el modulo y aplicando logaritmo en (25) se tiene:

$$\log\Big(Y[\omega]\Big) = \log\Big(X[\omega] + N[\omega]\Big) + \log\Big(H[\omega]\Big) \tag{26}$$

Se puede ver en (26) que el término para la distorsión lineal $H[\omega]$ puede ser expresado como una componente aditiva en el dominio logarítmico. El modelo generado en (26) resulta de gran interés pues muestra que en el dominio del logaritmo del modulo del espectro los efectos del ruido aditivo y convolucional se distorsionan la señal de entrada de forma aditiva. De esta forma, si el logaritmo del espectro de la señal es conocido, es posible eliminar el ruido aditivo y convolucional en la medida que se conozca el logaritmo de su respuesta espectral. En la literatura especializada las hipótesis más usadas para el modelamiento de la distorsión de canal son:

H1) La respuesta del canal es independiente de la señal de entrada

H2) El canal puede ser modelado como un filtro lineal

H3) La distorsión lineal del canal es constante o varía muy lentamente en el tiempo en los dominios log-espectral o cepstral.

2.5.2 Influencia del canal de comunicación

El ingreso de datos en forma masiva de un sistema de verificación de locutor operando en condiciones reales (ambiente no controlado) implicará una serie de inconvenientes. El trabajar en ambientes ruidosos y poco predecibles genera grandes dificultades al momento de modelar y compensar el ruido. Como se mencionó en la sección 2.4.2, los parámetros que definen los modelos acústico-fonéticos, son estimados mediante la maximización de la verosimilitud de un conjunto elocuciones de entrenamiento. Si una aplicación es evaluada con un conjunto de elocuciones de evaluación grabadas en un ambiente con características distintas a las que presentó el ambiente de entrenamiento. Se generará una componente de distorsión a nivel de parámetros acústico-fonéticos (denominada *mismatch*), lo que deteriorará considerablemente el desempeño del sistema. Esta es una de las principales causas de error en las aplicaciones reales (Neustein, 2010; Openshaw, 1993). En resumen, los elementos que generan *mismatch* en un sistema en condiciones reales son: variabilidad intra-locutor; condiciones del entorno (ruido aditivo); y el canal de comunicación (compuesto por distorsiones provocadas por el medio de captura usado y efectos del canal de transmisión). La Figura 8 enumera algunos ejemplos de estos factores.

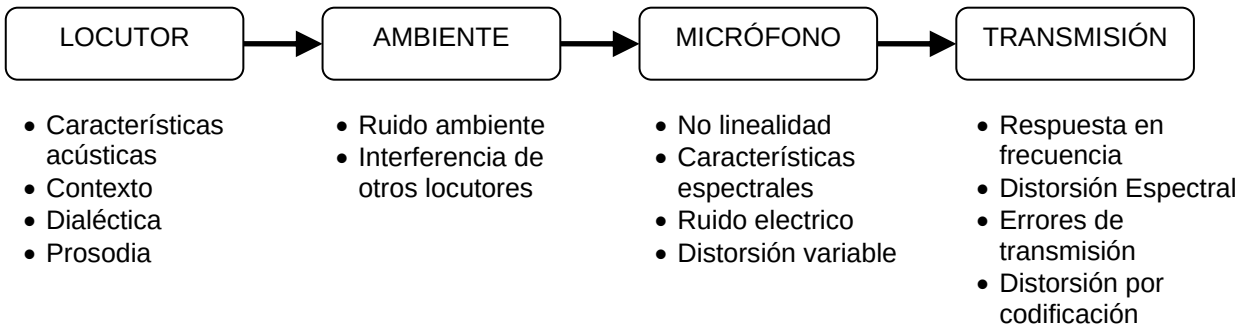


Figura 8. Factores que generan *mismatch* en procesamiento de voz.

Cuando una aplicación opera en un diálogo telefónico, se deben tener en cuenta peculiaridades asociadas a esta plataforma de comunicaciones, las que agregan dificultades adicionales a la tarea. Por ejemplo, el hecho de trabajar en sistemas telefónicos implica una disponibilidad limitada del tiempo de captura de la información de voz, ya que un servicio ofrecido en una plataforma telefónica debe garantizar un dialogo natural, fluido y sin largas esperas. Teniendo en cuenta la explicado anteriormente, resulta de vital importancia en una plataforma de procesamiento de patrones de voz el realizar una compensación del canal de comunicaciones y ajustar las condiciones de canal entre el entrenamiento y la evaluación de manera de disminuir la degradación en el funcionamiento del sistema. Otro factor a considerar es que los métodos a emplear deben ser esquemas sencillos y de carga computacional ligera de forma de no alterar la usabilidad del sistema.

En la Figura 9 se puede apreciar la distorsión que sufren los parámetros de un modelo de locutor (coeficientes cepstrales), al ser estimados con elocuciones grabadas bajo condiciones de canal diferentes. Cada punto de los gráficos mostrados representa el valor del parámetro cepstral de un mismo *frame* grabado en dos condiciones de canal (grabación estéreo). Como se explica en la sección 2.5.1, el modelo más aceptado en la literatura para la distorsión de canal corresponde a un componente constante en el dominio log-espectral o cepstral que asume las hipótesis H1, H2 y H3. Sin embargo, en (Reynolds et al., 1995) se muestra que la gran mayoría de los dispositivos de captura de baja calidad (incluidas cápsulas telefónicas) tienen una respuesta en frecuencia dependiente de la energía de la señal de entrada. El resultado que muestra la Figura 9 comprueba empíricamente esa afirmación. De esta forma, asumir que el canal es lineal e independiente de la señal de entrada puede ser una pobre aproximación (Wolfel, 2009; Mak et al., 2004). De lo anterior se puede concluir que es necesario desarrollar una estrategia de compensación de canal que modele a la componente de distorsión como un efecto variable en el tiempo y dependiente de la señal de entrada. En consecuencia, las hipótesis H1 y H2 no se deben considerar.

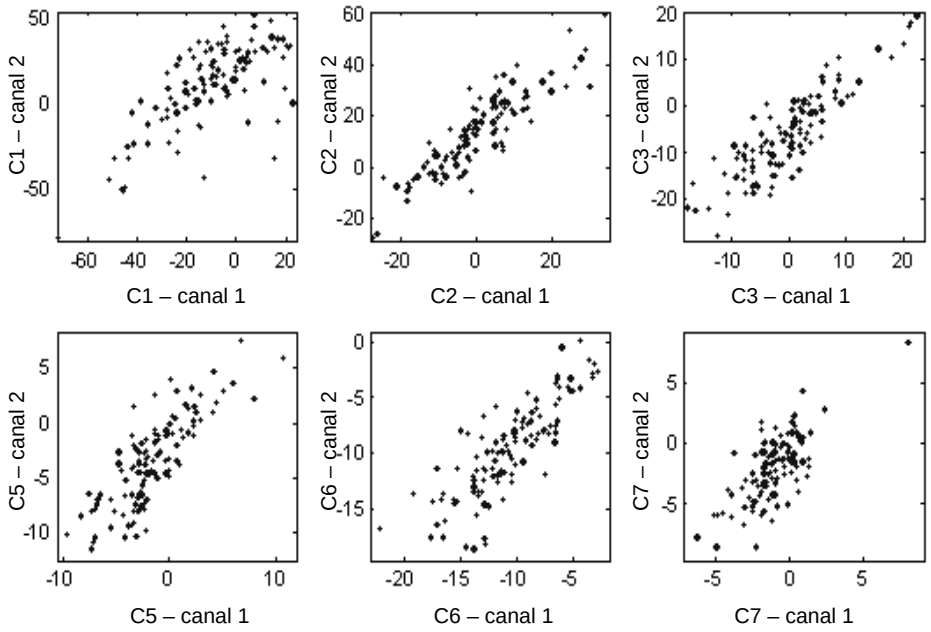


Figura 9. Distorsión que sufren algunos parámetros de un *frame* de voz (coeficientes cepstrales estáticos 1, 2, 3, 5, 6 y 7). Los ejes horizontal y vertical representan el valor del parámetro calculado con señales de voz de un grupo de locutores, grabados bajo dos condiciones de canal de distintas características.

Según lo afirmado en (Reynolds et al., 1995) y mostrado en la Figura 9, la distorsión de canal de comunicación es dependiente del *frame*. Sin embargo, la cantidad de información contenida en un *frame* aislado no es suficiente para estimar la distorsión de canal asociada a este. Es por esto que el uso de un esquema de transformación de parámetros *frame*-a-*frame* generará inexactitud en la estimación de la distorsión.

2.5.3 Técnicas de cancelación de canal

El objetivo de las técnicas de cancelación o compensación de la distorsión producida por el canal de comunicaciones es lograr que un sistema de reconocimiento de voz / locutor muestre de desempeño similar al que logra en ausencia de *mismatch* de canal. Los métodos que abordan el problema de *mismatch* de canal pueden dividirse en dos grupos: adaptación de modelos y transformación de características (Mak et al., 2004). La estrategia de adaptación de modelos consiste en agregar información o reentrenar los modelos acústico-fonéticos para que estos contengan información de todas las condiciones de canal posibles y así lograr una mejor representatividad de las condiciones de evaluación. Así, es factible que una aplicación funcione con modelos entrenados solo con señales de voz limpias. Por otro lado, las técnicas basadas en transformación de características tienen por objetivo aplicar una modificación a los parámetros observados distorsionados de modo de que estos se ajusten de mejor forma a los modelos acústico-fonéticos, entrenados en condiciones limpias o diferentes a la señal de evaluación. Los esquemas de transformación de características pueden clasificarse en dos conjuntos que siguen distintas filosofías de compensación de los efectos de canal: las técnicas basadas en parametrización robusta (basadas en elocución o *utterance-based*) y las basadas en compensación en el espacio de las características usando un modelo de referencia (basadas en modelo o *model-based*). En la presente tesis se trabajará con ambos tipos de metodologías para transformación de características, de las que a continuación se entrega una revisión bibliográfica.

2.5.3.1 Técnicas de parametrización robusta a efectos de canal (basadas en elocución)

La técnicas de caracterización robusta basadas en la elocución se aplican sobre el modulo de parametrización del sistema, es decir, se usan indistintamente en las etapas de entrenamiento y testeo. Tienen por objetivo minimizar el efecto del canal de comunicaciones sobre los parámetros acústico-fonéticos, sin hacer uso de un modelo canal de referencia o canal limpio. Esto es, atenuando o realzando ciertas características de la elocución observada bajo cierto criterio observado previamente. Las técnicas más representativas de este grupo son CMN (Furui, 1981) y RASTA (Hermansky y Morgan, 1994).

a. Normalización de la media cepstral (CMN)

La técnica de normalización de la media cepstral o CMN (*Cepstral Mean Normalization*) (Furui, 1981). Toma como punto de partida el modelo de canal explicado en la sección 2.5.1, sin considerar la componente del ruido aditivo $N[\omega]$. Al calcular las energías del banco de filtros Mel de la señal de voz en (24), y asumiendo que la respuesta del canal $H[\omega]$ es aproximadamente constante dentro de la banda de interés de cada filtro, se obtiene:

$$\begin{aligned} Y_m^S &= \sum_{\omega} F_m[\omega] \cdot Y[\omega] \\ &= \sum_{\omega} F_m[\omega]\Big[X[\omega] \cdot H[\omega]\Big] \\ &\approx H_m^S \cdot \sum_{\omega} F_m[\omega] \cdot X[\omega] \\ &= H_m^S \cdot X_m^S \end{aligned} \tag{27}$$

donde X_m^S y Y_m^S son las energías asociadas al filtro m del banco de filtros Mel (MFBLE) para la señal limpia y distorsionada, respectivamente; H_m^S es la respuesta del canal en el filtro m y $F_m[\omega]$ es el vector de ganancias del filtro m en el dominio DFT. Aplicando logaritmo y calculando la DCT sobre Y_m^S, se obtiene el vector de parámetros cepstrales de la señal distorsionada O_n^d:

$$\begin{aligned} O_n^d &= DCT\left\{\log\left(Y_M^S\right)\right\} \\ &= DCT\left\{\log\left(H_M^S \cdot X_M^S\right)\right\} \\ &= DCT\left\{\log\left(H_M^S\right)\right\} + DCT\left\{\log\left(X_M^S\right)\right\} \\ &= H_n + O_n^o \end{aligned} \tag{28}$$

donde O_n^o y H_n representan a la señal limpia y a la respuesta del canal en el dominio cepstral, respectivamente. La expresión en (28) muestra que en el dominio MFCC la componente de distorsión de canal puede ser modelada como una constante que se suma a la señal limpia. Este un resultado muy interesante y práctico, ya que sugiere que si se elimina la componente constante en cada dimensión n del vector de observación O, se suprimirá todo efecto de distorsión de canal sobre la señal limpia. Esta es la hipótesis que sostiene la técnica de CMN. De esta forma, la ecuación de cancelación de canal es:

$$\tilde{O}_{i,n} = O_{i,n} - \frac{1}{I}\sum_{i=1}^{I} O_{i,n} \tag{29}$$

donde $O_{i,n}$ y $\tilde{O}_{i,n}$ representan el parámetro cepstral n en el *frame* i, antes después de aplicar la normalización cepstral, I es el número total de *frames* de la señal. Si se aplica CMN sobre los datos de entrenamiento y evaluación en un sistema de procesamiento de voz, el sistema funcionara en un espacio sin la influencia del canal de comunicación.

En la literatura especializada es posible encontrar diversas variantes de CMN, todas ellas basadas en el modelo de canal constante en el dominio MFCC, algunas son (Acero, 1993):

- Global CMN (GCMN): Se basa en calcular el promedio cepstral para un conjunto de señales (ya sea en etapa de entrenamiento o evaluación), en vez una media para cada señal de entrenamiento o evaluación. Esta técnica se aplica forzando la media del conjunto de señales sobre la señal a normalizar.
- SNR-dependent cepstral normalization (SDCN): Las componentes de corrección cepstrales se estiman para diferentes rangos de SNR (razón señal a ruido) aplicando el criterio máxima verosimilitud en una base de datos estéreo.

b. Filtrado RASTA

La técnica de filtrado RASTA (Hermansky y Morgan, 1994) tiene su origen en el modelo auditivo humano. Este modelo muestra que el oído presenta la máxima sensibilidad para captar información lingüística alrededor de los 4 [Hz], y que ésta disminuye a mayores y menores frecuencias. Además, se comprueba que la percepción de sonidos similares a los

de la voz depende del sonido precedente. Tomando en cuenta que la información que se quiere transmitir al oído humano (lenguaje) es enviado dentro de la banda de frecuencias de mayor sensibilidad. Se sugiere la idea de suprimir la información acústica que no se encuentre en esta banda, ya que no aportar información relevante. Así, se disminuye la variabilidad a nivel de parámetros acústicos.

En la sección 2.5.2 se explica que el efecto distorsionador del canal de comunicaciones tiene una variación lenta, es decir, se sitúa en la zona de bajas frecuencias. A diferencia del ruido aditivo, el que frecuentemente se asocia a la zona de frecuencias altas. Basándose en estos hechos nace la técnica de filtrado RASTA (*RelAtive SpecTrAl*). Básicamente, este método aplica un análisis espectral-temporal aplicando un filtrado pasa-banda a la trayectoria temporal de los componentes del espectro (o el *cepstrum*) de la señal de voz. Originalmente esta técnica fue propuesta en el marco de un esquema de parametrización denominado RASTA-PLP (Hermansky, 1992). En este esquema se propone el uso de parámetros denominados RASTA-PLP, diferentes de los parámetros MFCC estudiados en la sección 2.2.1, utilizados en esta tesis y ampliamente usados en la literatura. Sin embargo, el filtrado RASTA puede realizarse sobre la trayectoria temporal de cualquier tipo de parámetros. Dado que los parámetros MFCC se relacionan con el desarrollo en serie del logaritmo del espectro, el filtrado RASTA se puede aplicar directamente a éstos, salvo al coeficiente cepstrum cero

o energía total, no se filtra. La función de transferencia para la técnica de filtrado RASTA es:

$$H[z] = 0.1z^{-4} \cdot \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - Kz^{-1}} \tag{30}$$

Como se puede ver en (30) el denominador contiene una K asociada con la frecuencia de corte inferior. Cuando el valor de esta constante es 0,98 le corresponde un grado de integración de unos 500[ms], que es el valor teórico para el cual se filtraría de manera óptima. Sin embargo, se ha demostrado que con un valor de 0,94 (correspondiente a 160[ms]), se observa el mejor desempeño del filtrado (30). El numerador contiene una regresión de orden dos. Además, se puede observar en (30) que la ganancia del filtro es cero a la frecuencia de 0[Hz], en otras palabras, se suprime componente continua. Esto muestra que la técnica de RASTA y CMN están relacionadas: ambos esquemas cancelas efectos lineales de distorsión de canal, ya que los filtros lineales se pueden interpretar como una componente constante en los dominios MFBLE y MFCC. Sin embargo, estas técnicas no contemplan el efecto del ruido aditivo en su modelado del canal de comunicaciones, por lo que no son capaces de cancelar este problema. De todas formas, es posible aplicar una técnica de compensación del ruido aditivo en combinación con RASTA o CMN según sea apropiado.

c. Avances recientes

Entre las contribuciones recientes de técnicas de transformación de parámetros es posible encontrar un primer grupo de técnicas, inspiradas básicamente en CMN y RASTA, que proponen metodologías de compensación de canal que hacen uso de las hipótesis H1, H2 y H3. Tomando elementos de CMN y/o RASTA y agregando cambios o mejoras. Dentro de estos trabajos se puede destacar la técnica MSN (*Mean Spectral Normalization*), propuesta en (Tufekci, 2007), en la que la normalización por promedio temporal se hace en el dominio MFBLE en vez de usar el MFCC. Otra interesante propuesta se encuentra en (Hung y Lee, 2006), donde se proponen una serie de métodos de optimización de filtrado de las trayectorias temporales de las características cepstrales. En este trabajo los filtros temporales siguen las hipótesis de RASTA, pero la estimación de parámetros de estos considera técnicas convencionales usadas en reconocimiento de patrones, tales como LDA, PCA y MCE. Siguiendo también las hipótesis H1, H2 y H3 se destaca el trabajo mostrado en (Chen y Bilmes, 2007), denominado MVA (*Mean Variance Autoregressive*), en el que se utiliza un filtro autoregresivo de media móvil para estimar la distorsión de canal *frame-a-frame*.

Existe otro grupo de técnicas en las que se va más allá del modelo de canal como una componente aditiva en el dominio MFCC y plantea la normalización de estadísticos de mayor orden para la robustez a efectos de canal. La primera técnica que se observa de este grupo es CVN (*Cespstral Variante Normalization*), usada en (Cook et al., 1997),

en este método se normaliza la varianza de cada coeficiente cepstral tomando como referencia un valor constante. Esfuerzos posteriores que continúan esta línea de investigación son los basados en la normalización de histogramas de las características cepstrales. Por ejemplo, en (De la Torre et al., 2005) cada señal de voz es modificada con el objeto de forzar una función de distribución de probabilidad de referencia para cada componente del vector de características MFCC. Además, en (Skosan y Mashao, 2006) se propone un método de normalización del histograma utilizando una serie de transformaciones que varían por segmentos de voz.

Es posible encontrar algunas técnicas en las que las hipótesis H1 y H3 no son consideradas, es decir, la distorsión de canal se considera dependiente de la señal de entrada y variable en el tiempo. Por ejemplo, en (Souilmi et al, 2002) se propone un método que hace uso de la correlación de los vectores de parámetros de cada *frame* para estimar una compensación de canal variable en el tiempo.

2.5.3.2 Compensación de efectos de canal en el espacio de las características (basadas en modelo)

Las técnicas basadas en compensación de parámetros tienen por objetivo modificar los parámetros distorsionados de modo de que estos se ajusten de mejor forma a los modelos acústico-fonéticos, entrenados en condiciones limpias o diferentes a la señal de evaluación. Ejemplos de estos métodos se detallan a continuación.

a. Cancelación de máxima verosimilitud de la componente de canal (ML-SBR)

La metodología de ML-SBR (*Maximum Likelihood - Signal Bias Removal*) (Rahim y Huang, 1996) se basa en la estimación de un *codebook* independiente del locutor, entrenado con señales limpias, o en su defecto, con señales capturadas en un canal de referencia (generalmente el canal usado para el enrolamiento). Este *codebook* se compone de *codewords*, cada uno de estos modela una unidad fonética.

Si se denota por O^d la secuencia de datos observados (señal de voz distorsionada por efectos de canal), H la componente aditiva de canal y $\tilde{H}$ la estimación de máxima verosimilitud de H estimada a partir de O^d. Además, si la secuencia de estados a evaluar se denota por S, entonces $\tilde{O}$, que representa la estimación de la secuencia de observación limpia O, se puede expresar como:

$$\tilde{O} = O^d - \tilde{H} \tag{31}$$

donde el valor de la estimación del canal, $\tilde{H}$, se obtiene a partir de la siguiente expresión basada en el principio de máxima verosimilitud (ML):

$$\tilde{H} = \arg\max_{H} \left\{ \Pr\left(O \middle| \lambda, H \right) \right\} \tag{32}$$

La derivación de la expresión de ML para la estimación de ruido convolucional es explicada en detalle en (Afify et al., 1998). Básicamente este método funciona de forma iterativa, basándose en el algoritmo *Expectation Maximization* (EM). En cada iteración se re-estima el *bias* de canal H, utilizando información del vector de observación O, la

estimación de H de la iteración anterior, y los parámetros de cada *codeword*. En general, los métodos basados en el algoritmo EM se caracterizan por generar una importante carga computacional. Sin embargo, existen alternativas que utilizan menos recursos computacionales, sin comprometer la exactitud del sistema. Una de estas alternativas al algoritmo descrito es el uso del un alineamiento, el que puede ser estimado mediante el algoritmo de Viterbi forzado, descrito en la sección 2.4.4. Este alineamiento asocia a cada *frame* de la señal de voz un único estado del modelo acústico-fonético. De esta forma, al contar con una relación *frame/*estado, es posible deducir una expresión analítica para estimar la distorsión de canal. Si se considera una distribución de probabilidad Gaussiana multivariable, es posible calcular una solución analítica para esta expresión:

$$\tilde{H} = \frac{\sum_{i=1}^{I}\left(\frac{O_i - \mu_\lambda(i)}{\sigma_\lambda^2(i)}\right)}{\sum_{i=1}^{T}\left(\frac{1}{\sigma_\lambda^2(i)}\right)} \tag{33}$$

b. RATZ

En el método de *RATZ (Multivariate-Gaussian-Based Cepstral Normalization*) (Moreno et al., 1995) la compensación cepstral es modelada en un esquema "*codeword-dependent*", es decir, se genera un *codebook* a partir de un conjunto de observaciones,

para cada *codeword* se estima la componente constante asociada al ruido convolucional. La técnica sigue básicamente las siguientes etapas:

1. Estimación estadística de la señal libre de distorsión.
2. Estimación estadística de la señal ruidosa.
3. Compensación de señal ruidosa.

La solución para estimar la distorsión se hace mediante el algoritmo de *EM* de forma no-supervisada, sin bases datos estéreo (es por esto que esta variante de la técnica se denomina *blind* RATZ). Para calcular la nueva señal se utiliza la siguiente expresión:

$$\hat{r}_k^{l+1} = \frac{\sum_{i=0}^{N-1} z_i \hat{p}^l\left(\frac{k}{z_i}\right)}{\sum_{i=0}^{N-1} \hat{p}^l\left(\frac{k}{z_i}\right)} - \mu_{x,k} \tag{34}$$

$$\hat{x} \cong y - \sum_{k=1}^{K} r_k \cdot P\Big(k \mid y\Big) \tag{35}$$

donde r_k es la distorsión del nivel de cuantización k y $P\Big(k \mid y\Big)$ es la probabilidad de esa celda dado la señal observada y. En este caso la celda corresponde a un segmento del universo que representan los coeficientes cepstrales. El conjunto de ellas genera un mapa representativo de las señales. La ecuación (35) representa a la señal libre de distorsión que es estimada a partir del algoritmo de *EM*. Según los resultados mostrados en

(Moreno et. al., 1995) se demuestra la dependencia del rendimiento de la adaptación con la tasa de señal a ruido (*SNR, Signal Noise Rate*). A medida que el *SNR* aumenta la compensación deja de tener efectos en la señal. Además, el número de señales con las cuales se estima la distorsión es fundamental.

c. Avances recientes

En la literatura reciente es posible encontrar técnicas en las que el canal es modelado como una transformación lineal en el dominio cepstral, es decir, una rotación y traslación del vector de parámetros. Este procedimiento también es conocido como *feature mapping*. Entre estas técnicas destacan la propuesta en (Reynolds, 2003), donde los parámetros del mapeo lineal son entrenados de un conjunto de modelos dependientes de canal utilizando adaptación MAP (*máximum a posteriori*). Otro ejemplo se observa en (Mak et al., 2007). En este trabajo la transformación de parámetros considera información de todas las distribuciones de probabilidad asociadas a las clases existentes, no solo a la perteneciente a la clase seleccionada, siguiendo un esquema de estimación tipo EM.

En los trabajos mencionados, es posible ver que las técnicas suelen superar el desempeño logrado con CMN, RASTA o ML-SBR, según corresponda la comparación. Sin embargo, a pesar de mostrar mejoras relevantes no solucionan completamente el problema de la distorsión de canal. Bajo ambientes de distorsión severa como

grabaciones con canales telefónicos o micrófonos de baja calidad, las mejoras que se observan por sobre lo que se obtiene con CMN, RASTA o ML-SBR pueden llegar a ser irrelevantes. Otro problema importante que presentan estas técnicas es que no funcionan adecuadamente con señales de voz muy cortas. Esto se debe a que se necesita una cantidad suficiente de *frames* de voz para poder estimar de forma confiable la componente de canal. Además, algunos métodos cuando son aplicados en condiciones limpias (sin distorsión de canal) pueden disminuir el desempeño del sistema. Por ejemplo, las técnicas de filtrado temporal inspiradas en RASTA pueden eliminar información asociada a una clase acústico-fonética o información del locutor.

Como se menciona en la sección 2.5.2 y según lo afirmado en (Reynolds et al., 1995) y mostrado en la Figura 9, el efecto distorsionador del canal de comunicación depende del *frame*. Por lo tanto, la componente de canal el en dominio MFCC no es constante en el tiempo y se debe estimar *frame*-a-*frame*. Esta afirmación resta validez a las hipótesis H1, H2 y H3 y lleva a la idea de que se debe estimar la distorsión del canal de comunicación como la suma de una componente constante y una componente variable en el tiempo. Dado que la cantidad de información contenida en un *frame* aislado no es suficiente para estimar la componente de distorsión en este, según se menciona en la sección 2.5.2, una estimación de canal *frame*-a-*frame* sería inexacta. Una solución a este problema es visualizar el efecto de la distorsión de canal sobre los parámetros de un *frame* empleando un dominio apropiado (p.e. espectro discreto de los

parámetros). Esta tesis considera la hipótesis de que con esta información es posible definir una parametrización robusta que atenúe el efecto del canal variable en el tiempo. Además, esta parametrización robusta podría ser combinada con técnicas convencionales de filtrado temporal, generando un efecto complementario.

Desde el punto de vista de la usabilidad, las técnicas basadas en elocución (en especial las de filtrado temporal de características) presentan una carga computacional relativamente baja y no deberían degradar mayormente los tiempos de procesamiento de la aplicación en la que fueran implementados. Sin embargo, las técnicas transformación de parámetros basadas en el criterio de máxima verosimilitud descritas en esta sección (como ML-SBR y RATZ), a pesar de que han mostrado mayores reducciones que los métodos basados en filtrado temporal en las tasas de error de sistemas SV y ASR (Mak et al., 2004), generan una alta carga computacional, lo que hace que la implementación de estos métodos en una aplicación real (on-line) no sea factible. Esto se debe a que estas metodologías tienen una base estadística y hacen uso de algoritmos numéricos como EM.

Dado este escenario, es necesario desarrollar métodos de compensación de parámetros haciendo uso del criterio de máxima verosimilitud de forma rápida. Para lograr esto, por ejemplo, es posible reducir el menor número de operaciones empleando un criterio de asociación *frame*/estado de tipo "duro" (p.e. algoritmo de Viterbi) en vez de uno "suave" (p.e. algoritmo EM).

Capítulo 3

Transformación de características robusta a canal basada en el filtrado de las energías del banco de filtros Mel

3.1 Introducción

Como se ha mencionado en la presente tesis, la robustez al *mismatch* en el canal de comunicaciones entre las condiciones de entrenamiento y prueba es uno de los problemas más severos enfrentados por los sistemas de verificación de locutor, reconocimiento de voz, y aplicaciones de procesamiento de voz en general. En particular, los sistemas que operan integrados a plataforma telefónicas se ven enfrentados a restricciones de usabilidad, ya que los diálogos del usuario con el sistema deben ser rápidos y eficientes. En particular, desde el punto de vista recién descrito, en un sistema de verificación de locutor los procesos de enrolamiento y verificación deben cumplir con estas condiciones. Esto implica que el sistema debe entrenar los modelos para cada usuario con una

cantidad limitada de elocuciones, las que no pueden tener una larga duración. Una cantidad reducida de grabaciones de entrenamiento implica modelos con un bajo nivel de entrenamiento, lo que lleva a una reducción en la exactitud del sistema. Además, el rendimiento del sistema se verá también afectado por la restricción en el tiempo y cantidad de señales en la etapa de test. Esto implica severas restricciones en la cantidad de datos en caso de aplicar alguna técnica para remover o reducir el ruido convolucional y/o aditivo que puede presentar la señal de test.

Cómo se menciona en la sección 2.5.3, la motivación de las técnicas de cancelación de la distorsión del canal de comunicaciones es la de poder alcanzar el desempeño observado en condiciones de canal *matched* (i.e., sin *mismatch*) minimizando los requerimientos de datos adicionales a la elocución de test. Los métodos para enfrentar este problema pueden dividirse en dos grandes grupos (Mak et al., 2004): compensación en el espacio de las características (*feature-space compensation*) (Tufekci, 2007; Rahim y Juang, 1996; Furui, 1981); y adaptación de modelos (Leggetter y Woodland, 1995; Gauvain y Lee, 1994). En ambos grupos de técnicas, el modelos de más amplio uso para la distorsión de canal corresponde a un *bias* o componente constante en el espacio cepstral o log-espectral. Este modelo simple e históricamente muy efectivo viene como resultado de las siguientes hipótesis, las que son ampliamente consideradas en la literatura especializada: H1, la respuesta del canal es independiente de la señal de entrada; y H2, el canal de comunicaciones puede ser modelado como un filtro lineal.

El objetivo de las técnicas actuales de compensación de características, como se menciona en la sección 2.5.3, es el de estimar la señal limpia o sin distorsión mediante la remoción de una componente constante o de baja frecuencia de la trayectoria temporal de las de cada dimensión del vector de características en el dominio cepstral o log-espectral (Tufekci, 2007; Rahim y Juang, 1996; Furui, 1981). Usualmente estas técnicas pueden reducir dramáticamente las tasas de error en presencia de *mismatch* de canal. Sin embargo, su efectividad se reduce notoriamente cuando se cuenta con datos limitados de enrolamiento y/o testeo, sobretodo cuando sólo hay disponibles elocuciones de corta duración. Por ejemplo, CMN remueve la componente constante a lo largo de la secuencia temporal de cada característica, pero su efectividad se ve reducida cuando el tiempo de duración de cada elocución disminuye (Wang et al., 2007). Existen técnicas que emplean métodos estadísticos técnicamente más sofisticados que CMN, MVA o RASTA para estimar la componente constante cepstral o log-espectral asociada a la distorsión de canal. Estas técnicas usualmente se basan en el algoritmo EM y pueden entregar mejoras mayores a las conseguidas por las técnicas convencionales de filtrado temporal de características (Rahim y Juang, 1996). Sin embargo, el algoritmo de EM es también muy sensible a la duración de las elocuciones y generalmente requiere una carga computacional significativamente mayor a las técnicas convencionales de filtrado temporal de características, y aún así, no remueve el efecto de la distorsión de canal en un 100%. Este resultado puede deberse al hecho de que las componentes constantes o de baja frecuencia en el dominio cepstral o log-espectral sólo consideran una parte del real

efecto del la distorsión generada por el canal de comunicaciones en el vector de características. Consecuentemente, las hipótesis H1 y H2 pierden validez. De hecho, como se puede ver en los resultados mostrados en esta tesis, al realizar experimentos de verificación de locutor texto-dependiente se observan aumentos en la tasa de error sólo por el hecho de operar sobre una plataforma telefónica (en comparación a la misma prueba realizada con micrófonos de alta calidad) a pesar de que los auriculares o *handsets* empleados en las etapas de entrenamiento y test sean los mismo, es decir, no existe *mismatch* de canal. Esto, junto con la información encontrada den la literatura especializada (Reynolds et al., 1995) genera una potente evidencia empírica que muestra que la distorsión de canal es altamente dependiente de la señal de entrada. Como resultado, la dependencia de la distorsión de canal en la señal de entrada generará una componente de ruido convolucional -variable en el tiempo- en la trayectoria temporal de los vectores de características. Este hecho restringe aún más la validez de las hipótesis H1 y H2 y el modelo de eliminación de componente constante en el dominio cepstral o log-espectral.

La estimación de la componente variable en el tiempo de la distorsión convolucional es una complicada tarea que ha sido vagamente abordada en la literatura especializada. Esta componente de distorsión es dependiente del *frame*, pero no puede ser estimada usando la información de un único *frame*, ya que esta no es suficiente. Algunas técnicas que abordan el problema de la distorsión de canal variable en el

tiempo sin considerar las hipótesis H1 y H2 ya han sido propuestas. Sin embargo, no toman en consideración el problema de los datos limitados en entrenamiento y/o test y generalmente no ayudan a eliminar el problema cuando la cantidad o duración de las elocuciones es baja. Por ejemplo CMVN (Kajarekar et al., 2009; Preti et al., 2008; Zheng et al, 2005) ha demostrado ser una eficiente herramienta para recobrar las características de la señal en sistemas de verificación de locutor texto-independiente en ambientes ruidosos. En estas tareas la duración de las elocuciones es notoriamente mayor a las disponibles en un sistema de verificación de locutor texto-dependiente. En los primeros es posible trabajar con grabaciones de duración del orden de los cientos de segundos (por ejemplo, en las evaluaciones de reconocimiento de locutor de NIST) (NIST, 2006), en cambio en los últimos, la cantidad de audio disponible por elocución es, como máximo, 5 segundos (por ejemplo, en la evaluación de YOHO) (Campbell y Higgins, 1994). Esto trae como consecuencia para los sistemas de verificación de locutor texto-dependiente la restricción de operar con cantidades limitadas de audio, y por ende con una reducida variabilidad fonética. Así, dificulta la correcta estimación de la componente de canal a nivel de trayectoria temporal de características tanto en la etapa de entrenamiento como en la de verificación. Como se muestra en éste capítulo, cuando CMVN es aplicado a una tarea que considera elocuciones cortas, la varianza cepstral es estimada de forma poco confiable, lo que lleva a una inexacta cancelación de la componente de canal y por ende a un eventual aumento en la tasa de error del sistema.

Los esquemas de filtrado en frecuencia aplican transformaciones a los largo del vector de características espectrales a diferencia de los ya convencionales métodos de filtrado temporal de características. La diferencia entre los dominios de operación entre ambos tipos de técnicas se grafica en la Figura 10.

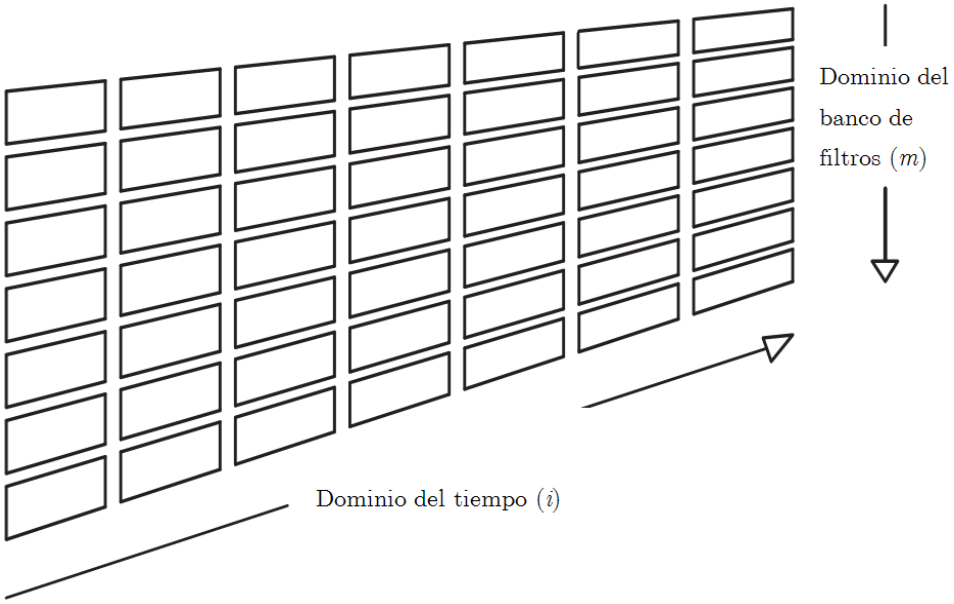


Figura 10. Representación gráfica de los dominios del tiempo y del logaritmo de las energías del banco de filtros Mel.

Las técnicas de filtrado en frecuencia ya han sido propuestas con el objetivo de aumentar la habilidad discriminativa de sistemas de reconocimiento de voz basados en HMM (Jung, 2004; Nadeu et al., 2001). Como se muestra en (Jung, 2004; Nadeu et al., 2001), hacer un filtrado a lo largo del vector de características LFBE puede producir principalmente dos efectos: la decorrelación de la secuencia temporal de características y

una ponderación de los coeficientes cepstrales. A pesar del hecho de que estas técnicas pueden mejorar las tasas de reconocimiento de los sistemas al ser comparadas con métodos de parametrización convencionales (e.g. MFCC), no han sido empleadas para abordar el problema de la robustez al *mismatch* por condiciones de distorsión en el canal de comunicaciones.

La transformada propuesta en este capítulo intenta reducir el efecto de la componente variable en el tiempo de la distorsión de canal telefónico al aplicar *frame*-por-*frame* un filtro pasa-bandas a lo largo cada vector de características LFBE sin tomar en cuenta las hipótesis H1 y H2. Este procedimiento es equivalente a desenfatizar algunas componentes y realzar otras en el dominio del espectro del vector de características LFBE. Como resultado, el *mismatch* de canal de comunicaciones entre modelos y señales de verificación puede ser significativamente reducido. Observar que el esquema propuesto en este capítulo no reemplaza las técnicas convencionales de filtrado temporal de características tales como CMN y RASTA. Al contrario, la técnica propuesta puede ser altamente complementaria y aumentar aún más las mejoras conseguidas con éstas. Más aún, en contraste a las técnicas existentes de filtrado de frecuencias, el esquema propuesto en este capítulo considera la aplicación de un análisis de importancia relativa o RIA (*Relative Importance Analysis*) que hace uso de una base de datos de evaluación, grabada bajo variadas condiciones de canal de comunicaciones, para estimar los parámetros de la transformación. Este capítulo además propones el uso

de una función discriminativa para aplicar RIA en el dominio del espectro de las características LFBE.

3.2 Transformación frame-por-frame en el dominio del espectro LFBE

El objetivo del esquema de filtrado propuesto en esta tesis es incrementar la habilidad de discriminación del sistema en que se aplica por medio de la reducción de la componente variable en el tiempo de la distorsión de canal. Esto implica que no se considerarán las simplificaciones impuestas por la hipótesis H1 y H2. La idea principal es reducir las componentes en el espectro del vector de características LFBE que son más sensibles a la distorsión de canal y conservar aquellas componentes que contienen principalmente información específica de la voz. Considerando que $Y_i[m]$ denota el *m-ésimo* valor de log-energía del banco de filtros Mel en el *frame i*. El espectro discreto de Y_i, Z_i, puede ser obtenido empleando la transformada discreta de Fourier de dimensión K, $DFT\{\}$:

$$Z_i = DFT\left\{Y_i\right\}. \tag{36}$$

La técnica de filtrado presentada en esta tesis puede ser aplicada en el espectro del vector LFBE, Z_i, como una función de ponderación $G[k]$, con $0 \le k < K$. Consecuentemente, el *k-ésimo* componente del espectro LFBE filtrado $\hat{Z}_i[k]$ es:

$$\hat{Z}_i[k] = G[k] \cdot Z_i[k]. \tag{37}$$

El vector filtrado de características LFBE en tiempo i, $\hat{Y}_i$, es obtenido aplicando la transformada DFT inversa en (37). Como resultado, $\hat{Y}_i$ es igual a la convolución circular de Y_i con la respuesta impulsiva de G, g:

$$\hat{Y}_i = g \circledast Y_i . \tag{38}$$

Observe que (38) permite aplicar el filtrado a lo largo del vector de características en el dominio LFBE. De esta forma el esquema de filtrado propuesto en esta tesis no es equivalente a los métodos convencionales basados en el filtrado de las trayectorias temporales de las características. En los experimentos realizados en el marco de esta tesis, se emplearon características de tipo cepstral (ver sección 2.4.1). Consecuentemente para incluir el filtrado propuesto en el vector de características definitivo se debe aplicar en una primera etapa la convolución entre Y_i y g, mostrada en (38), y luego estimar los parámetros cepstrales aplicando la DCT al vector filtrado $\hat{Y}_i$.

3.3 Uso de análisis de importancia relativa para la definición de la función G

Esta tesis propone el uso de RIA (*Relative Importance Analysis*) (Kanedera et al., 1999; van Vuuren y Hermansky, 1998) para definir la función G en (37). La función del análisis RIA es estimar la contribución de las componentes espectrales en el desempeño del sistema en el que es aplicado. Al hacer esto, es posible identificar aquellas componentes que son más sensibles a la distorsión de canal, y aquellas que contienen

principalmente información específica de voz. Si las componentes sensibles a la distorsión son atenuadas y las componentes asociadas a la voz son realzadas, el efecto de la distorsión de canal variable en el tiempo debería ser disminuido. En particular, en este trabajo RIA es empleado para evaluar la robustez al ruido convolucional de la componente espectral k en el domino Z_i. Originalmente RIA fue propuesto para ser estimado haciendo un uso directo de la medida de desempeño del sistema en el que se aplica (i.e. EER en SV y WER en ASR, etc.) (Kanedera et al., 1999; van Vuuren y Hermansky, 1998). Con el objeto de generalizar este procedimiento a cualquier problema de reconocimiento de patrones, esta tesis propone la aplicación de una medida de importancia relativa basada en una función discriminante en vez de la medida de desempeño del sistema. Para estimar la importancia relativa de cada componente en el domino del espectro del vector LFBE se utilizó una base de datos auxiliar compuesta por 40 locutores, quienes fueron grabados con tres diferentes tipos de micrófonos telefónicos (ver sección 3.4). Como se mencionó anteriormente, es sistema de TD-SV usado en esta tesis hace uso de los coeficientes cepstrales como medio de parametrización principal (ver sección 3.4). Consecuentemente, la función discriminativa de la que RIA hace uso también debe ser estimada en el dominio cepstral. Notar que el esquema de filtrado propuesto en esta tesis es aplicado en el dominio LFBE como se indica en (38) en la etapa previa a la estimación de los coeficientes cepstrales. Para estimar la función discriminativa, definida en esta tesis como J, se genera un modelo GMM *speaker-dependent* de cada locutor de los 40 que componen la base de datos

auxiliar, estos modelos son generados utilizando adaptación MAP (Gauvain y Lee, 1994). Al seguir ese procedimiento se logra que la correspondencia entre las componentes Gaussianas dentro de cada GMM *speaker-dependent* se preserve (Bimbot et al., 2004; Reynolds et al., 2000). La función discriminativa J se define como:

$$J = \frac{1}{P}\sum_{p=1}^{P}\frac{1}{N}\sum_{n=1}^{N}\frac{\text{Variabilidad Inter-locutor en característica } n}{\text{Variabilidad Intra-locutor en característica } n} = \frac{1}{P}\sum_{p=1}^{P}\frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{s=1}^{S}\left(\mu_{s,p}[n]-U_p[n]\right)^2}{\sum_{s=1}^{S}\sigma_{s,p}^2[n]} \tag{39}$$

donde N denota el número total de características cepstrales $\mu_{s,p}[n]$ y $\sigma_{s,p}^2[n]$ son la media y varianza en la dimensión n de la Gaussiana p del modelo GMM del locutor s. S es el número total de locutores y P es el número de componentes Gaussianas en cada GMM. En esta tesis se trabajó con P igual a 128. Además, $U_p[n]$ es el valor promedio de la característica n asociada a la componente Gaussiana p en la base de datos auxiliar. Vale la pena enfatizar que la función discriminativa J definida en (39) es claramente una razón de dispersión intra-clase/inter-clase, la que se puede interpretar como una medida de la separación entre clases considerando el criterio de minimización de la distancia cuadrático medio, similar a la función objetivo utilizada en el análisis discriminante lineal de Fisher (Duda y Hart, 1973). Note que la razón de variabilidad intra-locutor/inter-locutor adoptada en esta tesis no depende de la manera en que las clases acústicas son agrupadas o definidas en el espacio de las características. Como resultado,

es razonable suponer que la optimización de la función discriminativa J es independiente de la tarea de reconocimiento de patrones en la que se utiliza.

Considere que la función G en (37) es definida por sus frecuencias de corte, alta y baja, k_L y k_H, respectivamente. Si K denota la dimensión en que se realiza en análisis DFT propuesto en (36), el dominio de G puede definirse usando únicamente las primeras $K/2+1$ componentes de la DFT. Así, se puede definir $G[k]$ como:

$$G[k] = \begin{cases} w_L & k < k_L \\ 1 & k_L \le k \le k_H \\ w_H & k_H < k \end{cases} \tag{40}$$

donde $0 \le k_L \le K/2$, $0 \le k_H \le K/2$ y K=16. Además, w_L y w_H son las ganancias en la bandas de corte baja y alta, respectivamente. La medida de importancia relativa asociada a la componente k, $R(k)$, con $0 \le k \le K/2$ se estima como:

$$R(k) = \frac{1}{K/2}\left\{\sum_{k_1=0}^{k-1}\left[J(k_1,k) - J(k_1,k-1)\right] + \sum_{k_2=k+1}^{K/2}\left[J(k,k_2) - J(k+1,k_2)\right]\right\} \tag{41}$$

donde $J(k_1,k_2)$ denota la función discriminativa objetivo definida en (39) al ser estimada con el esquema de filtrado propuesto de acuerdo a (38) utilizando $k_L = k_1$, $k_H = k_2$ y $w_L = w_H = 0.0$, definidos en (40). De acuerdo a (Kanedera et al., 1999; van Vuuren y Hermansky, 1998), $R(k)$ se obtiene derivando la superficie de $J(k_1,k_2)$ con respecto a una componente espectral k, la que puede ser considerada como punto de corte superior

o inferior. Así, $R(k)$ en (41) corresponde al promedio de los diferenciales considerando todos los puntos de corte superiores e inferiores dependiendo si k es un punto de corte inferior o superior en (40), respectivamente. Este análisis puede ser interpretado como una medida de la mejora promedio que se obtiene con la inclusión de una cierta componente espectral k, lo que explica la sumatoria en (41). Como se muestra en la Figura 10, la función discriminativa $J(k_1, k_2)$ puede tener valores mayores que los obtenidos con el sistema *baseline* (es decir, sin aplicar ninguna transformación $G[k]$, o con $k_1 = 0$ y $k_2 = K / 2$) cuando ciertas componentes en el dominio del espectro LFBE son suprimidas.

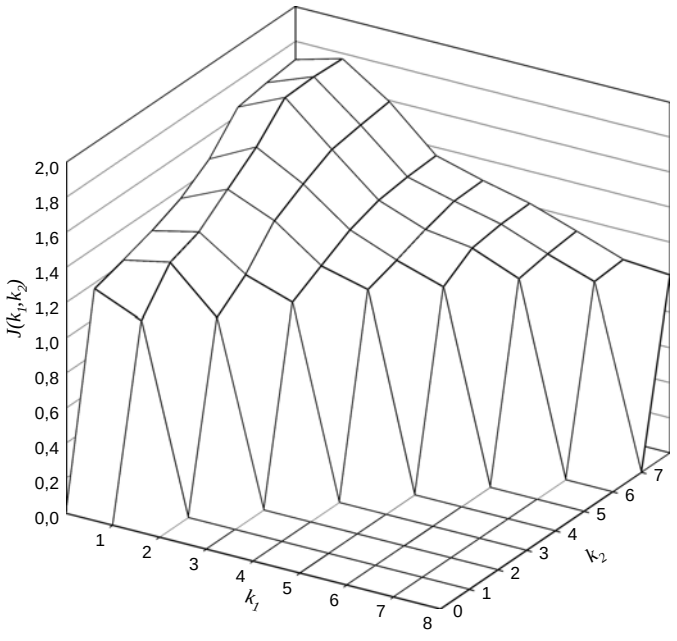


Figura 11. Función discriminativa $J(k_1, k_2)$ vs. (k_1, k_2) en $G[k]$.

Los valores obtenidos para $R(k)$ son presentados en la Figura 12 y en la Tabla 1. Estos fueron estimados usando tres diferentes configuraciones: el sistema *baseline* sin ningún filtrado temporal de características; el sistema *baseline* con CMN; y, el sistema *baseline* con RASTA. Como se puede ver en la Figura 12 y Tabla 1, algunos componentes en el dominio del espectro LFBE muestran una medida de importancia relativa $R(k)$ mucho mayor que otros. De hecho, algunos componentes k muestran un valor negativo para $R(k)$. Este resultado sugiere que algunas bandas en el espectro LFBE deben ser atenuadas o eliminadas con objeto de aumentar la capacidad discriminativa del sistema. Además, el resultado del análisis de importancia relativa con la función discriminante J definida en (39) sugiere, los límites inferiores y superiores de $G[k]$, k_L y k_H, respectivamente, pueden ser definidos en componentes en las que la medida de importancia relativa toma valores negativos o cercanos a cero. Como consecuencia de este análisis, se decide usar el filtro G definido en (40) con las siguientes frecuencias de corte: $k_L = 1$ y $k_H = 6$ en los casos sin filtrado de trayectorias de características y con CMN, y $k_L = 0$ y $k_H = 6$ con RASTA. Después de que k_L y k_H han sido definidos, las ganancias correspondientes a cada frecuencia de corte w_L y w_H, definidas en (40), pueden ser ajustadas maximizando el desempeño del sistema. De esta forma el procedimiento de ajuste de parámetros propuesto en esta tesis puede ser considerado como "guiado" por un análisis previo *data-driven* o basado en una base de datos auxiliar. Observe que no hay una relación numérica directa entre la salida de RIA

mostrada en la Figura 12 y los parámetros estimados para el filtro $G[k]$ en (40). En conclusión, este resultado valida el uso del dominio del espectro LFBE para obtener una representación concisa del efecto de distorsión de canal. De hecho, la Figura 12 claramente muestra que el efecto de la distorsión se acentúa mucho más en algunas bandas del espectro LFBE que en otras.

Tabla 1. $R(k)$ vs. componente k del dominio del espectro LFBE considerando y sin considerar filtrado temporal de características (es decir, CMN o RASTA).

	$R(k)$		
k	*baseline*	CMN	Rasta
0	-0.0704	0.1187	0.1905
1	0.307	0.5401	0.3095
2	0.1785	0.6065	0.2338
3	0.083	0.1303	0.208
4	0.0925	0.242	0.1981
5	0.0984	0.4856	0.1603
6	0.1072	0.134	0.0658
7	0.0261	0.0259	0.0254
8	0.0115	-0.1086	0.0048

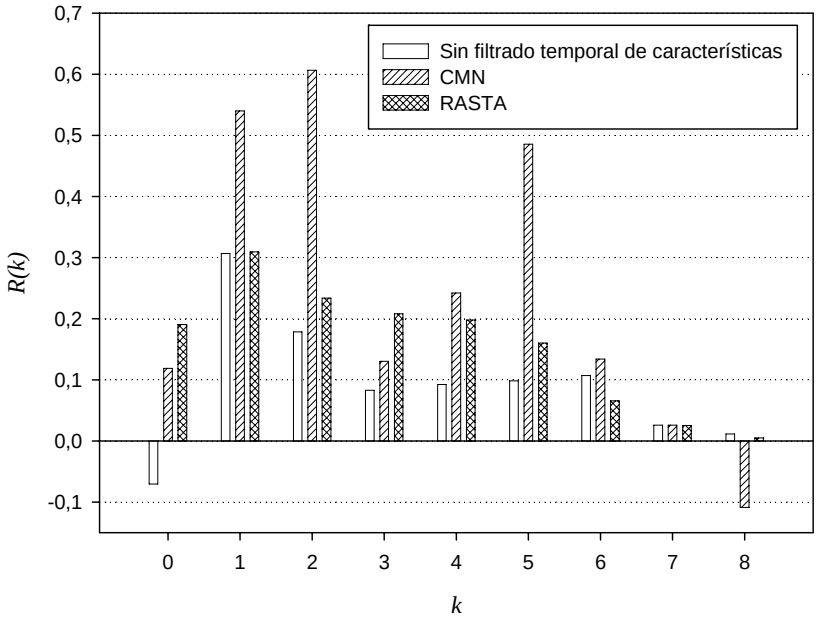


Figura 12. $R(k)$ vs. componente k del dominio del espectro LFBE considerando y sin considerar filtrado temporal de características (es decir, CMN y RASTA).

3.4 Experimentos

Se utilizó el sistema de verificación de locutor texto-dependiente del Laboratorio de Procesamiento y Transmisión de Voz de la Universidad de Chile (TD-SV-LPTV). Las señales de voz fueron divididas en *frames* de 25 [ms] con 12.5 [ms] de traslape y cada *frame* fue procesado con una ventana de Hamming. Para calcular la envolvente espectral de cada *frame* se utilizaron 14 filtros Mel DFT que cubren la banda de 300 a 3400 [Hz]; se calculó el logaritmo de la energía a la salida de cada filtro. A continuación, el esquema de filtrado de características propuesto en esta tesis es aplicado como se

muestra en (3). Luego, se estima la energía de cada *frame* y diez coeficientes cepstrales estáticos, y finalmente, se agregan al vector de observación la primera y segunda derivada temporal de las características. Los modelos HMM fueron estimados utilizando el alineamiento entregado por el algoritmo de Viterbi. Las unidades acústico-fonéticas usadas fueron trifonemas modelados con una topología izquierda-derecha de tres estados sin transiciones con salto de estado. Los modelos acústico-fonéticos consideraron matrices de covarianza diagonales. En la etapa de verificación cada elocución fue procesada con el algoritmo forzado de Viterbi. Dado un intento de verificación donde la identidad del locutor s es declarada, y O denota la secuencia de observación asociada de la elocución correspondiente al intento de verificación, el score de salida del sistema, $\log L(O)$, se calculó como un logaritmo de verosimilitud normalizado por una selección o *cohort* de locutores impostores:

$$\log L\left(O\right) = \log L\left(O\middle|\lambda_S\right) - \overline{\log L\left(O\middle|\lambda_{\bar{S}}\right)} \tag{42}$$

donde $\log L\left(O\middle|\lambda_S\right)$ es el logaritmo de la verosimilitud de la hipótesis de cliente y λ_S es el modelo acústico del locutor s; y, $\overline{\log L\left(O\middle|\lambda_{\bar{S}}\right)}$ es el logaritmo de la verosimilitud de la hipótesis de impostor, calculado como el logaritmo del promedio de las verosimilitudes asociadas a cada locutor del *cohort* de impostores.

Los resultados presentados en este capítulo fueron conseguidos utilizando una versión telefónica de la base de datos YOHO (Campbell y Higgins, 1994). En particular

para el trabajo de investigación realizado se usó un subconjunto de 70 locutores (40 hombres y 30 mujeres). Los locutores fueron divididos de la siguiente manera: 40 locutores (20 hombres y 20 mujeres) fueron utilizados para entrenar los modelos de impostor; y, 30 locutores (20 hombres y 10 mujeres) fueron usados como usuarios en los experimentos de verificación. Para cada locutor seleccionado se consideró una sesión de enrolling de 24 elocuciones. En la etapa de test se emplearon cuatro sesiones de verificación por cada locutor considerado. En cada sesión se seleccionaron cuatro elocuciones. Cada elocución fue grabada simulando una llamada telefónica en una línea telefónica real empleando un parlante de alta calidad acoplado acústicamente a un micrófono telefónico. Se consideraron siete auriculares telefónicos (`hset1`, `hset2`, ... , `hset7`). Las señales grabadas fueron muestreadas a 8 [kHz] utilizando 16 bits por muestra. Las grabaciones telefónicas fueron post-procesadas con un filtro FFT ecualizador apropiado con objeto de compensar la distorsión espectral causada por el parlante y la interfaz de captura de audio utilizada. El auricular `hset1` se etiqueto como canal de referencia o *matched*. A continuación, la base de datos telefónica se dividió en tres grupos: Y1 (base de datos auxiliar), compuesta por las señales de entrenamiento de los 40 locutores impostores grabadas con los auriculares `hset2`, `hset3` y `hset4`; Y2, compuesta de las señales de entrenamiento de los 40 locutores impostores grabadas con el auricular `hset1`; e, Y3, compuesta de las señales de entrenamiento y test de los 30 locutores de test grabadas con el auricular `hset1` y los auriculares `hset1`, `hset5`, `hset6` y `hset7`, respectivamente.

Como se explicó en la sección 3.3, Y3 se utilizó para estimar la función discriminativa $J(k_1,k_2)$ y la medida de análisis relativo $R(k)$. La base de datos Y2 fue usada para calcular las verosimilitudes asociadas al conjunto de impostores en la normalización del scores de cada intento de verificación en (42). Los modelos de usuarios fueron generados haciendo uso de las elocuciones de entrenamiento de Y3 grabadas con el auricular `hset1`. Los intentos de verificación fueron realizados empleando las señales de test de Y3 grabadas con los auriculares `hset1`, `hset5`, `hset6` y `hset7`. Consecuentemente, las curvas de falso rechazo fueron estimadas con 4 auriculares x 30 locutores/auricular x 16 señales de verificación/locutor = 1920 experimentos. Las curvas de falsa aceptación fueron estimadas con 4 auriculares x 29 impostores/auricular x 6 señales de verificación/impostor x 30 locutores = 20.880 experimentos. Observar que, como es sugerido por varios autores (Kajarekar et al., 2009; Bimbot et al., 2004; Reynolds, 2000), la razón entre la cantidad de test de clientes y de impostores debe ser aproximadamente igual a 1/10. El sistema *baseline* entrega un EER igual a 2.71% en condiciones *matched*, es decir, considerando sólo los intentos de verificación grabados en el auricular `hset1.` Cuando la base de datos de test Y3 es usada (`hset1`, `hset5`, `hset6` y `hset7`) el sistema *baseline* arroja un EER igual a 3.99%. Notar que en el sistema *baseline* no se aplica ninguna técnica de filtrado de trayectoria temporal de características ni compensación a la distorsión de canal.

Tabla 2. EER(%) Obtenido con las siguientes condiciones: sistema *baseline* sin considerar y considerando la transformación propuesta; RASTA sin considerar y considerando la transformación propuesta; CMN sin considerar y considerando la transformación propuesta; y, CMVN.

	baseline	RASTA	CMN	CMVN
EER(%)	3.99	3.27	2.60	5.2
EER(%)con transformación	3.64	3.13	2.35	—

3.5 Resultados y discusiones

Como se menciona en la sección 3.3, la Figura 11 muestra que la supresión o atenuación de alguna componentes en el dominio del espectro del vector LFBE pueden llevar a valores de $J(k_1,k_2)$ mayores que los obtenidos con el sistema *baseline* en una región no despreciable del dominio (k_1,k_2). La Figura 12 muestra la medida de importancia relativa R considerando y sin considerar filtrado temporal de características. Como se puede ver en la Figura 2, algunos componentes del espectro del vector LFBE proveen de mayor información discriminante en la presencia de *mismatch* de canal. Los valores más bajos de $R(k)$ corresponden a las mayores y menores componentes espectrales $k = 0$ y $k = 7$ o $k = 8$, respectivamente, sin considerar filtrado de la trayectoria temporal de características. Sin embargo, cuando CMN o RASTA son aplicados, la función de importancia relativa aumenta considerablemente en las primeras componentes espectrales, especialmente en $k = 0$. Esto puede deberse al hecho de que los métodos convencionales de filtrado de la trayectoria temporal de características reducen también

el efecto de la distorsión de canal en las componentes más bajas en el dominio del espectro LFBE. Como resultado, se requerirá una menor atenuación en la componente de corte de baja cuando se aplique la técnica propuesta en combinación con filtrado de trayectoria temporal de características en comparación a cuando se ésta aplica de manera aislada. Consecuentemente, $G[k]$ fue evaluado con las siguientes configuraciones: $k_L = 1$, $k_H = 6$, $\omega_L = 0.4$ y $\omega_H = 0.0$ cuando no se aplica filtrado de trayectoria temporal de características; $k_L = 1$, $k_H = 6$, $\omega_L = 0.8$ y $\omega_H = 0.0$ al utilizar CMN; y, $k_L = 0$, $k_H = 6$ y $\omega_H = 0.0$ al emplear filtrado RASTA. De acuerdo a la Figura 13 y la Tabla 2, la técnica propuesta de filtrado de características puede llevar a reducciones en el EER tan grandes como 8.6% al compararse con el sistema base. Además, al ser combinada con CMN y RASTA, el método presentado en esta tesis puede llevar a reducciones adicionales en el EER de 4.3% y 9.7%, respectivamente. Más aún, cuando el esquema de filtrado propuesto es aplicado en combinación con CMN y RASTA, se puede llegar a reducciones totales en el EER de 21.6% y 41.1%, respectivamente al compararse son el sistema *baseline*, el que no considera filtrado de trayectoria de características, respectivamente. La Tabla 2 además muestra que el uso de CMVN en una tarea de verificación de locutor texto-dependiente con elocuciones cortas puede aumentar el EER hasta en un 30.3% al compararse con el sistema *baseline*, esto debido a la estimación inexacta de la varianza a nivel cepstral cuando se cuenta con datos limitados en las etapas de entrenamiento y/o verificación.

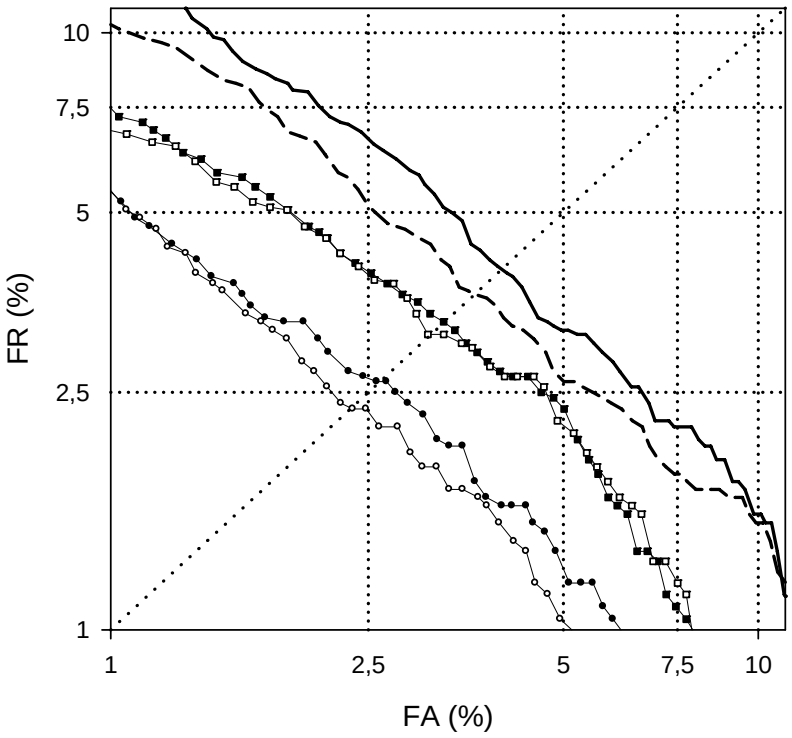


Figura 13. Curvas DET obtenidas bajo las siguientes condiciones: sistema *baseline* sin considerar (—) y considerando la transformación propuesta (---); RASTA sin considerar (■) y considerando la transformación propuesta (□); y, CMN sin considerar (●) y considerando la transformación propuesta (○).

3.6 Conclusiones

Este capítulo presentó una novedosa transformación de características *frame*-por-*frame* para robustez a la variabilidad de canal de comunicaciones en verificación de locutor texto-dependiente. La transformación es aplicada como un filtro pasa-banda a lo largo del vector de características que representa la envolvente del logaritmo de las energías

del banco de filtros Mel. El objeto de este filtrado es el de reducir el efecto variable en el tiempo de la componente de distorsión de canal en el dominio cepstral, el que a su vez es generado por la dependencia de la respuesta del canal de comunicaciones en la señal de voz de entrada. La transformación presentada en este capítulo es aplicada en una base *frame*-por-*frame*. Consecuentemente, se consigue una compensación de canal dependiente del *frame*. Además, a diferencia de las técnicas convencionales de filtrado de trayectoria temporal de características, la propuesta en este capítulo no considera el efecto de la distorsión de canal como una componente constante o de variación lenta en la trayectoria temporal del vector de características espectral y/o cepstral. El filtro es definido empleando análisis de importancia relativa en combinación con una función discriminativa basada en la razón de dispersión intra-locutor/inter-locutor. Los resultados presentados en este capítulo muestran que el espectro del vector de energías del banco de filtros Mel provee de una representación concisa del efecto de la distorsión de canal. Por su flexibilidad y por operar en el bloque de parametrización, el esquema propuesto puede ser aplicado a cualquier tarea de reconocimiento de patrones en procesamiento de voz. Los experimentos realizados con un sistema de verificación de locutor texto-dependiente muestran que la transformación propuesta puede llevar a significativas mejoras en el EER cuando se aplica de manera aislada o en combinación con filtrados de trayectorias temporales de características, tales como CMN o RASTA. Finalmente, es posible proponer como trabajo de investigación futuro la aplicación de la

técnica propuesta en este capítulo en combinación con otros métodos de remoción de canal o a otras tareas de reconocimiento de patrones en procesamiento de voz.

Capítulo 4

Compensación de la distorsión de canal usando un modelo de aproximación polinomial en el dominio de las energías del banco de filtros Mel

4.1 Introducción

Como se ha reiterado en esta tesis, cuando sistema de reconocimiento de patrones de voz funciona sobre una plataforma real comercial, se puede observar que las condiciones de operación imponen dos grandes restricciones: la cantidad de datos, ya sea para etapas de entrenamiento como de prueba, es limitada por motivos de usabilidad; y, las señales de voz generalmente están distorsionadas por el canal de comunicación, ya sea un canal telefónico o una grabación realizada con micrófonos de baja calidad. Escenarios de datos

limitados conducen a modelos pobremente entrenados y a estimaciones inexactas de la distorsión de canal para su cancelación en la etapa de test. Esto trae como consecuencia sistemas reconocimiento de patrones de voz extremadamente vulnerable al *mismatch* por condiciones de canal.

En intentos previos en el estado-del-arte para abordar el problema de *mismatch* de canal, ya sea mediante adaptación de modelos o compensación de características, el modelo más considerado ha sido el de la distorsión como una componente constante en el dominio MFCC o LFBE (*Log Filter-Bank Energy*), el que se deriva de considerar las hipótesis H1 y H2, descritas en la sección 2.5.1. En general estás técnicas se basan en el modelo de canal básico mostrado en (24) donde $x(t)$, $n(t)$, $h(t)$ y $y(t)$ representan, respectivamente, la señal de voz limpia, el ruido aditivo, la respuesta impulsiva lineal invariante del canal y la señal distorsionada, tal como también se modela en (Jiang, 2001; Rahim y Juang, 1996). Para enfocar la modelación en la distorsión de canal, se puede asumir ausencia de ruido aditivo y descartar $n(t)$, simplificando también el análisis. Así, basándose en (24), si las señales son procesadas con un banco de filtros Mel de tipo DFT y a la salida de cada filtro la energía de $x(t)$ es calculada, además, si la respuesta en frecuencia de $h(t)$ es considerada constante, el logaritmo de la energía de la señal distorsionada a la salida de cada filtro m en un instante de tiempo i puede ser modelada como (Huerta, 2002; Jiang, 2001; Rahim y Juang, 1996):

$$\log\left[\overline{y_{i,m}^2}\right] = \log\left[\overline{x_{i,m}^2}\right] + \log\left[\overline{H^2[\omega_m]}\right] \tag{43}$$

donde: $\overline{x_{i,m}^2}$ e $\overline{y_{i,m}^2}$ representan la energía de las señales limpia y distorsionada a la salida del filtro m en el *frame* i, respectivamente; $H[\omega_m]$ es la respuesta en frecuencia del filtro $h(t)$ que modela la distorsión de canal; donde $1 \le m \le M$ con M siendo el número total de filtros Mel que componen el banco; y, ω_m es la frecuencia central discreta del filtro Mel m. Como resultado, la señal observada en el dominio MFCC puede ser modelada como (Huerta, 2002; Jiang, 2001; Rahim y Juang, 1996):

$$Y_{i,n}^C = X_{i,n}^C + H_n^C \tag{44}$$

donde $X_{i,n}^C$ y $Y_{i,n}^C$ denotan, respectivamente, el coeficiente cepstral estático n en el *frame* i de $x(t)$ y $y(t)$; con $1 \le n \le N$, donde N es el número total de coeficientes cepstrales estáticos; y, H_n^C es el *bias* cepstral asociado a la distorsión de canal en la dimensión n.

A pesar de que el modelo en (43) y (44) ha sido ampliamente usado por muchos autores y que el canal es modelado usualmente como un filtro lineal invariante en el tiempo, la continuidad de la respuesta en frecuencia de $H(\omega)$ no ha sido explorada exhaustivamente en la literatura especializada. En otras palabras, la componente aditiva $\log\left[\overline{H^2[\omega_m]}\right]$ en (43) es usualmente tratado y estimada sin considerar que $H[\omega_m]$ corresponde a un muestreo de la curva continua $H(\omega)$.

Este capítulo abarca el problema de la compensación de la distorsión de canal desde el punto de vista de los métodos de compensación de características basados en modelos, explicados en la sección 2.5.3. Este grupo de técnicas utiliza prioritariamente un modelo de referencia de voz para estimar la componente constante de canal mostrada

en (26) empleando criterios de máxima verosimilitud o MAP. El modelo más ampliamente adoptado corresponde al de mezclas Gaussianas (*GMM, Gaussian Mixture Model*) construido a partir de datos de voz de referencia (Yiu et al., 2006; Mak, 2004; Huerta, 2002; Jiang, 2001; Acero et al., 2000; Afify et al., 1998; Rahim y Juang, 1996), donde la corrección en el espacio de las características puede ser estimada utilizando el algoritmo EM. Las técnicas basadas en el algoritmo pueden llevar a reducciones significativas del *mismatch* por distorsión de canal, sin embargo el algoritmo es computacionalmente muy costoso. La compensación en el dominio de las características puede ser también estimada con un criterio de máxima verosimilitud combinado con una estimación basada en búsqueda de vecino más cercano, donde cada *frame* observado es asociada a la unida acústica del modelo de referencia que más se le asemeja. La búsqueda de vecino más cercano puede realizarse siguiendo dos estrategias (Rahim y Juang, 1996): (a) usar la distribución Gaussianas más probable dentro de un GMM para cada *frame* procesado; y (b) usar un esquema basado en el alineamiento entregado por el algoritmo forzado, donde cada *frame* es asociado con la mezcla Gaussiana, estado y modelo acústico fonético que maximizan su verosimilitud. A pesar de que los métodos basados en el algoritmo EM generalmente proveen de mayores mejoras, la carga computacional requerida por los algoritmos basados en búsqueda de vecino más cercano, ya sea con su versión GMM o Viterbi forzado, es sustancialmente menor.

En las técnicas convencionales de sustracción de la componente constante de canal que hacen uso de las hipótesis H1 y H2 la componente de corrección en (2) y (3) es

usualmente estimada de forma independiente en cada dimensión del vector de características. En este contexto, una arista interesante es el número de parámetros requerido para estimar el ruido convolucional: el lógico suponer que mientras más parámetros necesite el modelo de distorsión para estimar, mayor será la cantidad de datos requeridos para que este se estime confiablemente. En general, las técnicas convencionales de cancelación de canal estiman, por lo menos, tantos parámetros como dimensiones tenga el vector de características cepstrales estáticas, como se ve en (44). Si el número de parámetros en el modelo de distorsión de canal se reduce, se pueden lograr grandes mejoras en la estimación de la distorsión cuando se cuenta con señales cortas o datos limitados de *test*.

Este capítulo propone un método para mejorar la exactitud de las estimaciones de la distorsión de canal basadas en búsqueda de vecino más cercano. La idea de la técnica propuesta es no aumentar los requerimientos computacionales significativamente. Dado este escenario la estrategia más directa es reducir el número de parámetros requeridos en la estimación del modelo de distorsión. Para lograr este objetivo, la técnica aquí descrita modela la distorsión de canal como una componente constante en el dominio LFBE como una función polinomial de m. Esto se inspira en la condición de que la curva $\log\left[\overline{H^2[\omega_m]}\right]$ representa muestras de una curva continua, $\log\left[\overline{H^2(\omega)}\right]$, la que puede ser modelada como una función polinomial de ω. Como se mencionó anteriormente ω_m es la frecuencia central del filtro m del banco de filtros Mel, así la curva $\log\left[\overline{H^2[\omega_m]}\right]$

puede también ser modelada como una función polinomial de m en vez de of ω_m. Consecuentemente, a la distorsión aditiva de canal se ajusta un polinomio de orden P, donde $P \leq M$ siendo M el número total de filtros en el banco Mel. Como resultado, la distorsión de canal también es modelada como una componente aditiva en el dominio cepstral, pero en contraste a (44), la componente constante es estimada como una suma ponderada, cuyos pesos son función de los coeficientes polinomiales. Si $P \leq N$, donde N es el número de características cepstrales estáticas, el número de parámetros requerido por la técnica presentada en este capítulo será menor que el requerido por las técnicas ordinarias de sustracción de componentes constantes en el dominio cepstral estático, como en (44). Tal como se muestra aquí, la función polinomial que modela $\log\left[\overline{H^2[\omega_m]}\right]$ en el dominio LFBE lleva a una función lineal en el dominio MFCC. Cabe destacar que las funciones polinomiales han sido empleadas exitosamente en modelamiento acústico y robustez a ruido aditivo en ASR y SV debido a su forma simple, flexibilidad de ajuste y baja carga computacional (Cui y Alwan, 2005; Acero et al., 2000).

En este capítulo el modelo propuesto de distorsión de canal polinomial en el dominio LFBE es empleado en combinación con esquemas de estimación basados en búsqueda de vecino más cercano, usando GMM y/o alineamiento de Viterbi forzado. El modelo es probado con dos tareas de procesamiento de patrones acústicos basada en señales de *test* de corta duración: un sistema de verificación de locutor texto dependiente basado en HMM que opera con voz telefónica y un sistema de evaluación de

pronunciación del Ingles basado en ASR, el que opera con grabaciones realizadas con micrófonos de escritorio de baja calidad.

4.2 Modelo polinomial para la distorsión en el dominio de energías del banco de filtros Mel

Como se menciona en la sección 4.1 la principal innovación del modelo es el hecho de tomar en consideración que las componentes aditivas generadas por la distorsión del canal de comunicaciones en el dominio LBFE de acuerdo a (43) son muestras de una curva continua de respuesta en frecuencia en ω. Por ejemplo, si una línea telefónica analógica está compuesta básicamente por cables trenzados de cobre y el auricular telefónico o *handset*, es razonable pensar que un modelo simple para el canal telefónico puede estar compuesto de un filtro pasa-bajos con una curva continua de respuesta en frecuencia. Claramente, el efecto del canal de comunicaciones en cada componente espectral no es independiente de la ganancia introducida por éste en el resto de las componentes. Consecuentemente, el ruido convolucional debe ser modelado con una función paramétrica en el espectro de forma de ligar o "amarrar" el efecto del canal entre las distintas componentes espectrales. Además, una función paramétrica permite reducir el número de parámetros requeridos para estimar la distorsión de canal. Al hacer esto, la componente aditiva mostrada en (43) y (44) puede ser estimada de forma más confiable en situaciones donde se cuenta con datos limitados o elocuciones cortas. Con el

fin de simplificar la notación se propone definir $Y_{i,m} = \log\left[\overline{y_{i,m}^2}\right]$, $X_{i,m} = \log\left[\overline{x_{i,m}^2}\right]$ y $G_m = \log\left[\overline{H^2[\omega_m]}\right]$. En este capítulo, G_m se modela como una función polinomial en m:

$$G_m = \sum_{p=0}^{P} a_p \cdot m^p \tag{45}$$

donde a_p es la $p-ésima$ componente polinomial, P es el orden polinomial y $A = \{a_p\}_{p=0}^{P}$. Empleando la transformada coseno discreta (DCT) y (43), la $n-ésima$ componente cepstral distorsionada en el *frame* i, $Y_{i,n}^C$, se puede expresar como:

$$\begin{aligned} Y_{i,n}^C &= \sum_{m=1}^{M} Y_{i,m} \cdot \cos\left(\frac{\pi \cdot n}{M} \cdot \left(m - 0.5\right)\right) \\ &= \sum_{m=1}^{M} \left(X_{i,m} + G_m\right) \cdot \cos\left(\frac{\pi \cdot n}{M} \cdot \left(m - 0.5\right)\right) \end{aligned}. \tag{46}$$

Reemplazando G_m con la aproximación polinomial en (45), $Y_{n,t}^C$ puede re-escribirse como:

$$\begin{aligned} Y_{i,n}^C = & \sum_{m=1}^{M} \left\{ X_{i,m} \cdot \cos\left(\frac{\pi \cdot n}{M} \cdot \left(m - 0.5\right)\right) \right\} \\ & + \sum_{p=0}^{P} \left\{ a_p \cdot \sum_{m=1}^{M} m^p \cdot \cos\left(\frac{\pi \cdot n}{M} \cdot \left(m - 0.5\right)\right) \right\} \end{aligned}. \tag{47}$$

En una aplicación real, $Y_{i,n}^C$ es la $n-ésima$ característica cepstral observada en el *frame* i. Definiendo:

$$W_{p,n} = \sum_{m=0}^{M} m^p \cdot \cos\left(\frac{\pi \cdot n}{M} \cdot \left(m - 0.5\right)\right), \tag{48}$$

$Y_{i,n}^{C}$ en (47) puede expresarse como:

$$Y_{i,n}^{C} = X_{i,n}^{C} + \sum_{p=0}^{P} a_p \cdot W_{p,n} \tag{49}$$

Notar que $W_{p,n}$ depende solo de $0 \le p \le P$ y las constantes asociadas al cálculo de la DCT $\cos\left(\frac{\pi \cdot n}{M} \cdot \left(m - 0.5\right)\right)$. De acuerdo a (49) la componente aditiva de canal en el dominio MFCC $G_n^C(A) = \sum_{p=0}^{P} a_p \cdot W_{p,n}$, es una combinación lineal de los coeficientes polinomiales a_p ponderados por los factores $W_{p,n}$.

4.3 Estimación de la distorsión de canal modelada como una función polinomial usando búsqueda de vecino más cercano

Como se explica en la introducción de este capítulo, el modelo polinomial para la distorsión de canal de comunicaciones propuesto es usado en combinación con algoritmos de búsqueda de vecino más cercano. Considere la secuencia observada de vectores de características $Y^C = \{Y_i^C\}_{i=0}^{I-1}$, donde $Y_i^C = \{Y_{i,n}^C\}_{n=0}^{N-1}$ corresponde al *frame* en el instante i e I es el número total de *frames* en la elocución. Al usar la filosofía de vecino más cercano para estimar la compensación de la distorsión de canal, el *frame* Y_i^C debe ser asociado con una de las unidades acústicas s_k que pertenece a un modelo acústico fonético de referencia λ (generalmente entrenado con elocuciones limpias o sin ruido de canal), donde $1 \le k \le K$ con K siendo el número total de unidades acústicas

en λ. Consecuentemente, en el caso de la estimación de vecino más cercano usando un *codebook* (nn-GMM), λ y s_k representan una componente Gaussiana y el índice k asociado a esta, respectivamente. Cuando se usa la estimación basada en el alineamiento entregado por el algoritmo forzado de Viterbi (forced-Viterbi), λ y s_k representan, respectivamente, la secuencia de HMMs de fonemas contexto-dependientes, y una distribución Gaussiana en un estado dentro de este HMM compuesto. Finalmente, en ambas estimaciones, GMM y forced-Viterbi, se generará una salida denotada por $S = \{s_{k(i)}\}_{i=0}^{I-1}$ la que estará alineada a la secuencia de vectores de observación de entrada Y^C, donde $s_{k(i)}$ representa la unida acústica asociada al *frame* Y_i^C. Así, el método propuesto en esta tesis envuelve tres pasos principales:

Paso 1. Dada una secuencia de entrada de vectores de características Y^C, S es obtenido empleando una búsqueda de vecino más cercano usando nn-GMM o forced-Viterbi.

Paso 2. La corrección en el demonio de las características se computa empleando el modelo de aproximación polinomial de acuerdo a (45). Como resultado, el vector de parámetros polinomiales A es estimado

Paso 3. Finalmente, la secuencia de *frames* compensados $\hat{X}^C = \{\hat{X}_i^C\}_{i=0}^{I-1}$ es obtenida de acuerdo a:

$$\hat{X}_{i,n}^C = Y_{i,n}^C - \sum_{p=0}^{P} a_p \cdot W_{p,n} \,. \tag{50}$$

Dentro del paso 2, el vector de parámetros de la función polinomial A, puede ser estimado usando el criterio de máxima verosimilitud o ML:

$$\hat{A} = \arg\max_{A}\{p(Y^C \mid \lambda, S, A)\} \tag{51}$$

donde $\hat{A} = \{\hat{a}_p\}_{p=0}^{P}$ es el vector de parámetros óptimo que define la función polinomial en (45). La función de distribución de probabilidad de la unidad acústica s_k puede ser modelada con una función Gaussiana con vectores de medias $\mu_k = \{\mu_{k,n}\}_{n=0}^{N-1}$ y matriz de covarianza diagonal Σ_k. El set de parámetros de la función Gaussiana se define como $\phi_k = (\mu_k, \Sigma_k)$. Los elementos de la diagonal de Σ_k son denotados por $\sigma_k^2 = \{\sigma_{k,n}^2\}_{n=0}^{N-1}$. En este caso, la verosimilitud $p\left(Y_i^C \middle| \phi_{k(i)}, A\right)$ es definida como:

$$p\left(Y_i^C \middle| \phi_{k(i)}, A\right) = \frac{1}{(2\pi)^{\frac{N}{2}} \mid \Sigma_{k(i)} \mid^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}\sum_{n=0}^{N-1} \frac{\left[Y_{i,n}^C - \left(\sum_{p=0}^{P} a_p \cdot W_{p,n}\right) - \mu_{k(i),n}\right]^2}{\sigma_{k(i),n}^2}} . \tag{52}$$

donde $\phi_{k(i)} = (\mu_{k(i)}, \Sigma_{k(i)})$ es set de parámetros de la función Gaussiana de $s_{k(i)}$, unidad acústica asociada al *frame* Y_i^C. El vector óptimo de coeficientes polinomiales $\hat{A}$ puede ser estimado maximizando la siguiente función objetivo, basada en el logaritmo de la verosimilitud $\log\left[p(Y^C \mid \lambda, S, A)\right]$:

$$\begin{aligned} \hat{A} &= \arg\max_{A}\left\{\log\left[p(Y^C \mid \lambda, S, A)\right]\right\} \\ &= \arg\max_{A}\left\{\sum_{i=0}^{I-1}\log\left[p(Y_i^C \mid \phi_{k(i)}, A)\right]\right\}. \end{aligned} \tag{53}$$

Reemplazando (52) en (53), la optimización puede ser escrita como:

$$\hat{A} = \arg\max_{A} \left\{ \begin{array}{l} \sum_{i=0}^{I} \log\left[\left((2\pi)^{\frac{N}{2}} \mid \Sigma_{k(i)} \mid^{\frac{1}{2}} \right)^{-1} \right] \\ -\frac{1}{2} \cdot \sum_{i=0}^{I-1} \sum_{n=0}^{N-1} \frac{1}{\sigma_{k(i),n}^{2}} \left[Y_{i,n}^{C} - \left(\sum_{p=0}^{P} a_p \cdot W_{p,n} \right) - \mu_{k(i),n} \right]^2 \end{array} \right\} \tag{54}$$

donde $\sum_{i=0}^{I} \log\left[\left((2\pi)^{\frac{N}{2}} \mid \Sigma_{k(i)} \mid^{\frac{1}{2}} \right)^{-1} \right]$ no depende de A y es descartado. Como resultado, $\hat{A}$ es estimado calculando las derivadas parciales de (54) con respecto a a_q, donde $0 < q \leq P$, e igualándolas a cero. Luego, la optimización en (54) lleva a un sistema lineal de P+1 ecuaciones y P+1 variables incógnitas:

$$\begin{array}{l} \sum_{p=0}^{P} \hat{a}_p \cdot \left\{ \sum_{i=1}^{I} \sum_{n=1}^{N} \left[\left(W_{q,n} \cdot W_{p,n} \right) \Big/ \sigma_{k(i),n}^{2} \right] \right\} \\ \quad = \sum_{i=1}^{I} \sum_{n=1}^{N} \left[\left(W_{q,n} \Big/ \sigma_{k(i),n}^{2} \right) \cdot \left(Y_{i,n}^{C} - \mu_{k(i),n} \right) \right]. \end{array} \tag{55}$$

Si $\beta_{q,p} = \sum_{i=1}^{I} \sum_{n=1}^{N} \left(W_{q,n} \cdot W_{p,n} \right) \Big/ \sigma_{k(i),n}^{2}$ y

$\gamma_q = \sum_{i=1}^{I} \sum_{n=1}^{N} \left[\left(W_{q,n} \Big/ \sigma_{k(i),n}^{2} \right) \cdot \left(Y_{i,n}^{C} - \mu_{k(i),n} \right) \right]$, el sistema lineal en (55) puede ser expresado como:

$$\sum_{p=0}^{P} \hat{a}_p \cdot \beta_{p,q} = \gamma_q \,. \tag{56}$$

Considerando $\Gamma = \left\{\gamma_q\right\}_{q=0}^{P}$ y $\mathrm{B} = \left\{\beta_{p,q}\right\}_{(P+1)\times(P+1)}$ la solución para el sistema en (55) puede ser fácilmente re-escrita como:

$$\hat{A} = \mathrm{B}^{-1} \cdot \Gamma \tag{57}$$

Los parámetros Γ y B son estimados considerando todos los *frames* de la elocución empleando el procedimiento recién descrito.

4.4 Experimentos

4.4.1 Verificación de locutor texto-dependiente

Al igual que en el capítulo 3, se utilizó el sistema de verificación de locutor texto-dependiente del Laboratorio de Procesamiento y Transmisión de Voz de la Universidad de Chile (TD-SV-LPTV). El procedimiento de pre-procesamiento y parametrización de las señales de voz es el mismo que el descrito en la sección 3.4. La metodología de entrenamiento de los modelos HMM es también la explicada en la sección 3.4. A diferencia del procedimiento usado en el capítulo 3, se utiliza un procedimiento distinto para la normalización la de la verosimilitud de cada señal de test luego de ser procesada con el algoritmo forzado de Viterbi. Dado un intento de verificación donde la identidad del locutor s es declarada, y O denota la secuencia de observación asociada de la elocución correspondiente al intento de verificación, el score de salida del sistema,

$\log L(O)$, se calculó como un logaritmo de verosimilitud normalizado por una selección o *cohort* de locutores impostores:

$$\log L\left(O\right) = \log L\left(O\middle|\lambda_S\right) - \log L\left(O\middle|\lambda_{\overline{S}}\right) \tag{58}$$

donde $\log L\left(O\middle|\lambda_S\right)$ es el logaritmo de la verosimilitud de la hipótesis de cliente y λ_S es el modelo acústico del locutor s; y, $\log L\left(O\middle|\lambda_{\overline{S}}\right)$ es el logaritmo de la verosimilitud de la hipótesis de impostor, calculado utilizando un modelo universal o *speaker-independent* (SD), λ_{SI}, entrenado con un conjunto de locutores impostores.

El método polinomial de compensación de la distorsión de canal presentado en este capítulo fue probado utilizando una versión telefónica de la base de datos YOHO (Campbell y Higgins, 1994) diferente a la empleada en los experimentos de mostrados en la sección 3.4. En particular para el trabajo de investigación de este capítulo se usó un subconjunto de 70 locutores (40 hombres y 30 mujeres). Los locutores fueron divididos de la siguiente manera: 40 locutores (20 hombres y 20 mujeres) fueron utilizados para entrenar el HMM *speaker-independent*; y, 30 locutores (20 hombres y 10 mujeres) fueron usados como usuarios en los experimentos de verificación. Para cada locutor seleccionado se consideró una sesión de enrolling de 24 elocuciones. En la etapa de test se emplearon cuatro sesiones de verificación por cada locutor considerado, en cada sesión se seleccionaron cuatro elocuciones. Cada elocución fue grabada simulando una llamada telefónica en una línea telefónica real tal como se explica en la sección 3.4. Se

consideraron siete auriculares telefónicos (`hset1`, `hset2`, ... , `hset7`). El auricular `hset1` se etiquetó como canal de referencia o *matched*. Con las grabaciones de enrolling del canal `hset1` se entrenaron los modelos HMM *speaker-dependent* de todos los locutores de la base de datos. A su vez, los experimentos de verificación de locutor se realizaron utilizando las señales de test grabadas en los siete *handsets*. Con esto, se consiguen experimentos en los que ha *matching* de canal con el modelo HMM *speaker - dependent* y experimentos en los que no existe *matching* de canal con éste. Consecuentemente, las curvas de falso rechazo fueron estimadas con 7 auriculares x 30 locutores/auricular x 16 señales de verificación/locutor = 3.360 experimentos. Las curvas de falsa aceptación fueron estimadas con 7 auriculares x 29 impostores/auricular x 6 señales de verificación/impostor x 30 locutores = 36.540 experimentos. Observar que, como es sugerido por varios autores (Kajarekar et al., 2009; Bimbot et al., 2004; Reynolds, 2000), la razón entre la cantidad de test de clientes y de impostores debe ser aproximadamente igual a 1/10. Bajo las condiciones experimentales descritas, el sistema *baseline* entrega un EER igual a 3.33% en condiciones *matched*, es decir, considerando sólo los intentos de verificación grabados en el auricular `hset1.` Cuando la base de datos de test es usada de forma completa (hset1, hset2, ..., hset7) el sistema *baseline* arroja un EER igual a 5.77%. Cabe destacar que el sistema *baseline* no emplea ninguna técnica de filtrado de trayectoria temporal de características ni compensación a la distorsión de canal.

El modelo polinomial propuesto en este capítulo es usado en combinación con estimaciones de vecino más cercano basadas en GMM y alineamiento forzado de Viterbi, denotadas por nn-GMM-Poly y Viterbi-Poly, respectivamente. El modelo acústico GMM de referencia usado en la estimación de vecino más cercano basada está compuesto por 256 componentes Gaussianas y fue generado usando exactamente los mismos datos de empleados para entrenar el HMM SI, λ_{SI}, como se explica en esta sección. En el caso de la estimación de vecino más cercano basada en alineamiento de Viterbi forzado, el estado óptimo del alineamiento es obtenido usando λ_{SI} como modelo de referencia. En este caso la componente Gaussiana más verosímil $s_{k(i)}$ es elegida de un set compuesto por las ocho Gaussianas que componente el estado alineado del modelo λ_{SI} asociado a Y_i^C más la Gaussiana en el estado correspondiente en el HMM λ_{SD}. El modelo de estimación polinomial de la distorsión de canal propuesto en este capítulo es comparado con una técnica convencional de remoción de la componente constante de canal o *Signal Bias Removal* (SBR) (Rahim y Juang, 1996) la que hace uso del modelo ordinario de *bias* de canal mostrado en (43) y (44). SBR también es probado en combinación con los dos tipos de estimación de vecino más cercano descritos en este capítulo: nn-GMM y forced-Viterbi. Estas variantes se denotarán como: nn-GMM-SBR y Viterbi-SBR, respectivamente.

4.4.2 Evaluación de pronunciación basada en reconocimiento de voz

Para los experimentos de CAPT basado en ASR se empleó el sistema de evaluador de pronunciación de palabras del idioma inglés para estudiantes de habla hispana del Laboratorio de Procesamiento y Transmisión de Voz de la Universidad de Chile (LPTV). Este sistema está basado en el motor de reconocimiento automático de voz del LPTV. El procedimiento de pre-procesamiento y parametrización de las señales de voz es el mismo que el descrito en la sección 3.4, salvo que en esta tarea se considera el uso de CMN, tal como se describe en la sección 2.5.3, en el bloque parametrizador del sistema.

Como se describe en (Molina et al., 2009), los modelos HMM fueron entrenados usando elocuciones de usuarios nativos de lengua inglesa y española. Para esto se utilizaron las basas de datos CSR-I WSJ0 (Garafalo *et al.*, 1993) para las elocuciones en lengua inglesa y LATINO40 (LDC, 1995) para el idioma español. En CSR-I WSJ0 las señales fueron grabadas usando micrófonos de alta calidad y fueron muestreadas a 16 [KHz]. Todas las señales de entrenamiento de esta base de datos, en total 20.055, fueron utilizadas para entrenar los CD-HMMs. LATINO40 fue empleada para entrenar las unidades fonéticas en español. Estas unidades fueron utilizadas para modelar los CD-HMMs. Correspondientes a las versiones españolizadas de los fonemas del idioma inglés, como se describe brevemente en la sección 2.2. Esta base de datos, al igual que la primera, está compuesta por voz continua. LATINO40 fue grabada por 40 locutores de

América latina y cada locutor leyó y grabó 125 frases obtenidas de lecturas de noticias. La base de entrenamiento está compuesta por 4.500 señales derivadas de 36 locutores. El vocabulario de esta base de datos está compuesto de 6.000 palabras. Se adoptó un modelo de lenguaje de forma plana compuesto por el vocabulario competitivo generado de acuerdo a lo indicado en la sección 2.2. Cabe mencionar que el vocabulario competitivo de cada palabra-objetivo utilizada en este trabajo fue seleccionado desde el *corpus* de CSR-I WSJ0.

La base de datos de prueba está compuesta por un vocabulario de 15 palabras objetivo: *Against*, *Boyfriend*, *Chocolate*, *College*, *Example*, *Handsome*, *Hospital*, *Mouth*, *Should*, *Special*, *Student*, *Thirty two*, *Tourism*, *Vegetable* y *Yesterday*. Estas palabras fueron seleccionadas por expertos (adultos) en fonética y lenguaje inglés con el objetivo de obtener un balance fonético entre las palabras a probar. Luego, para cada una de las palabras se definieron cuatro categorías de errores de pronunciación, con dos o tres ejemplos por cada categoría. Las categorías de errores de pronunciación fueron definidas por los mismos expertos en fonética que seleccionaron las palabras de la base de entrenamiento. La idea que siguieron fue representar los errores frecuentes en pronunciación que comenten los usuarios no-nativos. Estos errores en la pronunciación van desde los más sutiles a los más significativos que corresponde a la pronunciación de la palabra objetivo siguiendo las reglas fonéticas del español. Como resultado, se obtienen 5 niveles de calidad de pronunciación. Los ejemplos de la palabra objetivo

pronunciados fueron indexados de 1 a 5: donde el puntaje 5 corresponde a la correcta pronunciación de la palabra objetivo; y el puntaje 1 denota la peor posible pronunciación, es decir, pronunciar la palabra objetivo siguiendo reglas fonéticas del español. Estas transcripciones de evaluación fueron grabadas por nueve expertos en el idioma inglés usando dos diferentes micrófonos de escritorio de baja calidad. La tasa de muestreo fue de 16 [KHz]. Una vez grabada la base de datos (que en total resultaron 3.811 señales), estás fueron re-etiquetadas (evaluadas en su nivel de calidad de pronunciación) por los mismos expertos en idioma inglés aplicando las reglas de pronunciación explicadas anteriormente con los 5 niveles. El total de estas grabaciones fue dividido en una base que denominaremos evaluación y otra que denominaremos de test. La base de datos de evaluación fue empleada para entrenar las curvas utilizadas para mapear el output del motor ASR al score de cinco niveles discretos utilizados por la aplicación CAPT, esta base utiliza 2.959 señales del total. Por otro lado, la base de datos de prueba o de test se compuso de 632 señales y fue utilizada para hacer pruebas de desempeño del sistema CAPT, usado como medida de desempeño la correlación entre el *score* de salida del sistema y el score con que los expertos en idioma inglés etiquetaron cada señal.

La medida usada como score de salida del motor ASR es la posición en la lista de N-mejores hipótesis de la hipótesis con mayor probabilidad que contenga la palabra objetivo (palabra correctamente pronunciada).

4.5 Resultados y discusiones

4.5.1 Verificación de locutor texto-dependiente

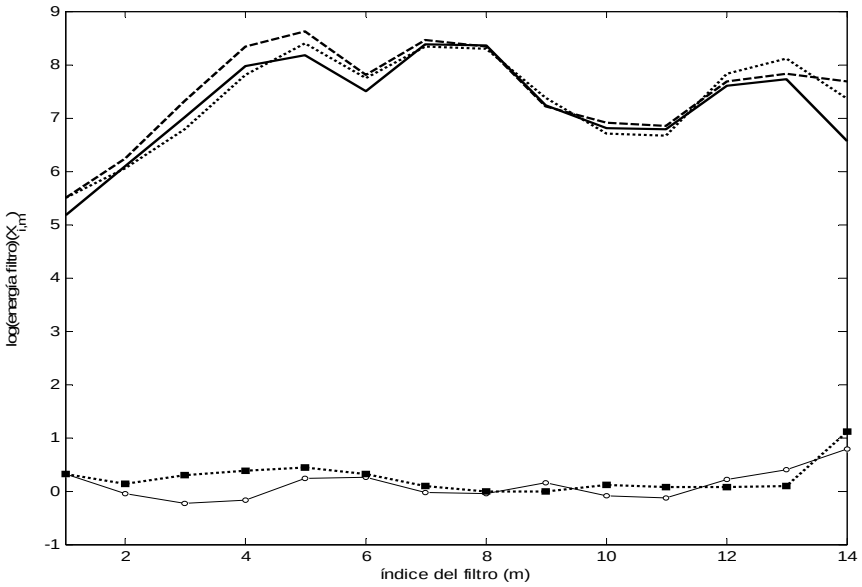
La Figura 14 muestra un ejemplo de vector de características dado en el dominio LFBE grabado con el canal telefónico `hset1` (canal de referencia), `hset2` y `hset3`. La Figura 14 también sugiere fuertemente que la distorsión de canal, i.e. la diferencia entre los vectores de características LFBE es una curva continua, la que puede ser modelada fácilmente usando una función polinomial.

Tabla 3. EER(%) obtenido con el sistema *baseline*, RASTA, CMN y CMVN.

	baseline	RASTA	CMN	CMVN
EER(%)	5.77	5.54	5.51	5.97

La Tabla 3 muestra los resultados utilizando técnicas convencionales de filtrado de trayectoria de características: RASTA, CMN y CMVN para propósitos de comparación. Cuando se comparan con el sistema *baseline*, las reducciones relativas en el EER provistas por RASTA y CMN son iguales a 4.0% y 4.5%, respectivamente. Además, la Tabla 3 muestra que el uso de CMVN puede incrementar el error relativo en 3.5% cuando se compara con el sistema *baseline*. Esto puede deberse a la inexacta estimación de las estadísticas cepstrales (en particular la varianza cepstral) cuando se cuenta con datos limitados o elocuciones de corta duración, ya sea en la etapa de

entrenamiento o testeo del sistema. Esta falta de datos lleva a una estimación poco confiable de la distorsión de canal, y por ende a un consecuente aumento de la tasa de error del sistema



.

Figura 14. Las curvas superiores muestran una representación grafica del vector de características LFBE para un *frame* de voz dado grabado con: el canal de referencia hset1 (——); hset2 (– – –) y hset3 (· · ·). Las curvas inferiores muestran la diferencia en el dominio LFBE entre: hset2 y el canal de referencia hset1 (· · ■ · ·); y, hset3 y el canal de referencia hset1 (– ∘ –).

De acuerdo a la Figura 15, los procedimientos nn-GMM-SBR y Viterbi-SBR pueden llevar a mejoras relativas en el EER tan grandes como 5.7% y 15.5%, respectivamente, al compararse con el sistema *baseline*. El hecho de que Viterbi-SBR provea de mayores mejoras que nn-GMM-SBR claramente se debe a que Viterbi-SBR

emplea información temporal sobre la secuencia de fonemas en las elocuciones de testeo, mientras que nn-GMM-SBR no lo hace. Además en la Figura 15, Viterbi-Poly puede reducir el EER en 22.7% y 8.4% cuando se compara al sistema base y al método de Viterbi-SBR, respectivamente, al aplicarlo con $P = 8$. Si se realiza un test de significancia utilizando el test de McNamar (Gillick y Cox, 1989) es posible mostrar que estas mejoras son estadísticamente significativas ($p < 0.034$). Más aún, nn-GMM-Poly puede llevar a mejoras de hasta 11.5% y 6.3% en el EER cuando se compara al sistema base y al método de nn-GMM-SBR, respectivamente, al aplicarlo con $P = 6$. Estos resultados son también estadísticamente significativos ($p <0.01$). Es interesante destacar que el modelo propuesto de distorsión de canal basado en una función polinomial lleva claramente a mayores mejoras que los métodos basados en SBR empleando ordenes polinomiales de $6 \leq P \leq 8$. Notar que el modelo SBR ordinario requiere estimar tantas componentes de distorsión de canal como coeficientes cepstrales estáticos consideres el vector de características (e.g. diez coeficientes cepstrales estáticos en esta tesis). Consecuentemente, se consigue una clara reducción en la cantidad de parámetros requeridos en la estimación de la distorsión del canal de comunicaciones.

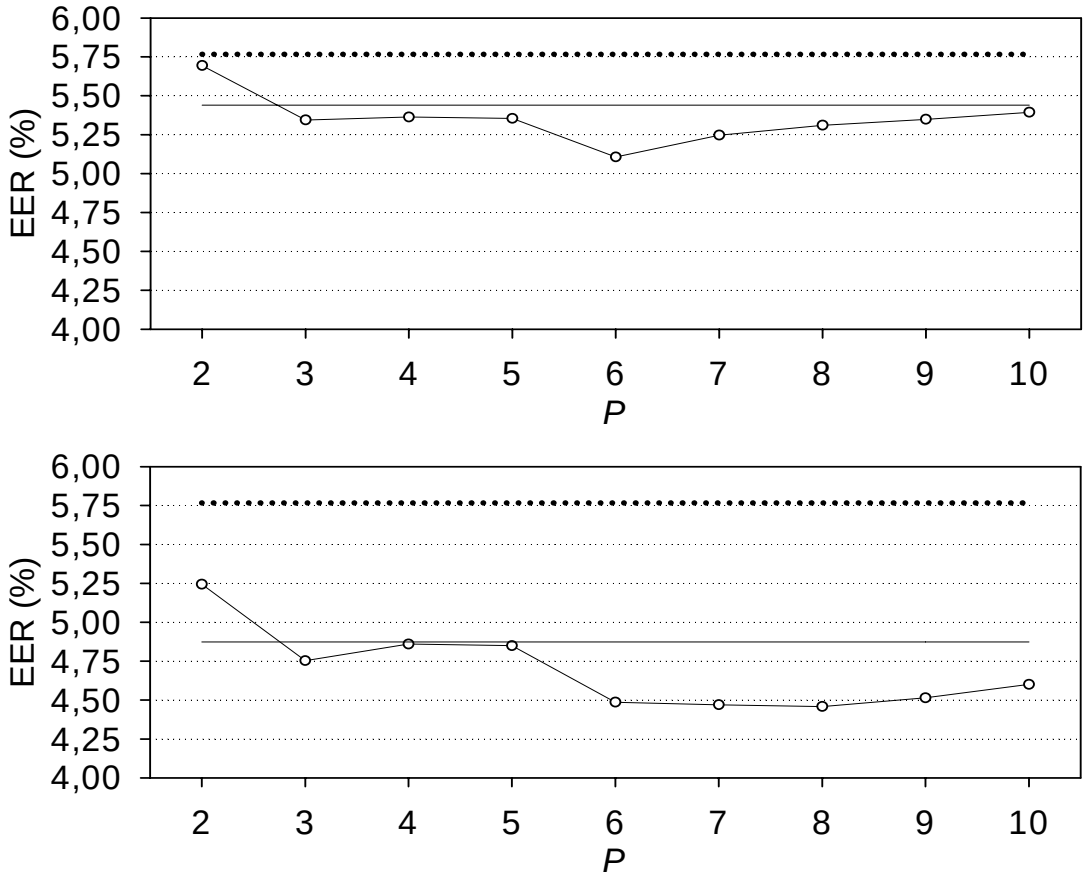


Figura 15. EER (%) vs. orden de la función polinomial P: (a) nn-GMM-Poly $(-\circ-)$, sistema *baseline* $(\cdot\ \cdot\ \cdot)$, y nn-GMM-SBR (——); y, (b) Viterbi-Poly $(-\circ-)$, sistema *baseline* $(\cdot\ \cdot\ \cdot)$ y Viterbi-SBR (——).

La Figura 16 y la Figura 17 presentan las curvas DET obtenidas con el sistema *baseline*, el modelo ordinario de *Signal Bias Removal*, denotado por SBR, y el modelo propuesto de distorsión de canal basado en aproximación polinomial. Como se puede ver en las Figura 16 y Figura 17, nn-GMM-Poly $(P = 6)$ y Viterbi-Poly $(P = 6)$ entregan reducciones en el área bajo la curva DET superiores que las alcanzadas usando nn-

GMM-SBR y Viterbi-SBR, respectivamente. Esta evidencia muestra que ambas variantes para aplicar el modelo propuesto pueden reducir de forma importante el EER en las proximidades del TEER y consecuentemente aumentar la habilidad discriminativa del sistema. Vale la pena mencionar que, debido al hecho de que el modelo polinomial de distorsión de canal es empleado con estrategias de búsqueda de vecino más cercano, el aumento en la carga computacional de nn-GMM-Poly y Viterbi-Poly al ser comparado con nn-GMM-SBR y Viterbi-SBR, respectivamente, es despreciable en la práctica. Por ejemplo, la estimación del vector de parámetros polinomiales A requiere un tiempo de procesamiento solo un 18.8% mayor que el requerido para estimar la componente constante de canal H^C definida en el modelo convencional de SBR en (44), esta etapa a su vez representa solamente un 22.2% del tiempo total de procesamiento requerido por el motor de TD-SV para procesar una señal de verificación. Consecuentemente, el esquema de compensación polinomial de canal propuesto en este capítulo incrementa el tiempo total de procesamiento del sistema en sólo un 4.2% por cada intento de verificación al ser comparado con el método estándar de SBR.

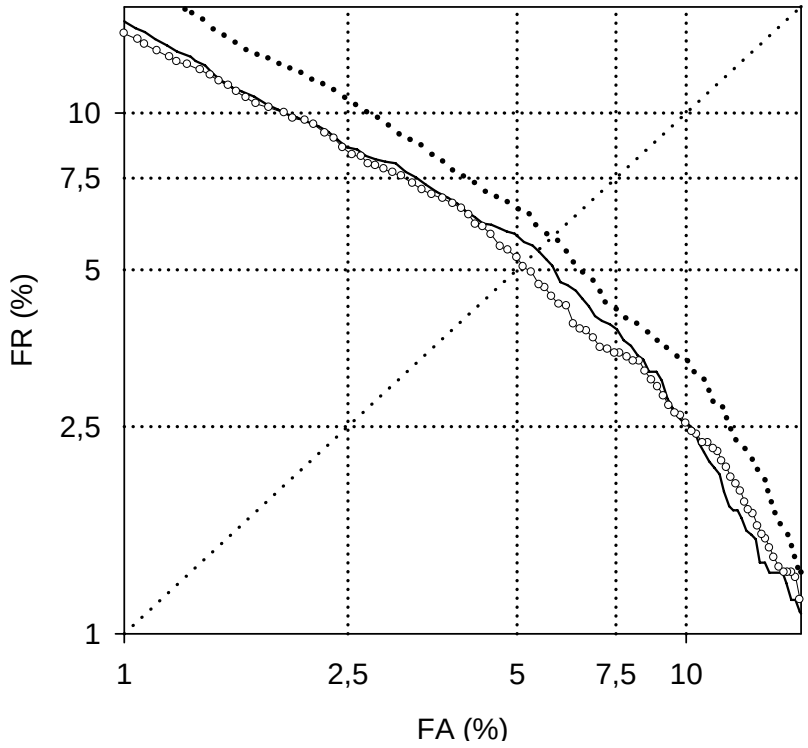


Figura 16. Curvas DET obtenidas con el sistema *baseline* (· · ·), nn-GMM-SBR (——) y nn-GMM-Poly, usando un orden polinomial P igual a 6 (– ∘ –).

Cabe notar que la mayor mejora toma lugar al usar $P = 6$ y $P = 8$ empleando las estimaciones basadas en nn-GMM y forced-Viterbi, respectivamente. A pesar del hecho de que la efectividad del método propuesto está estrechamente ligada al orden polinomial P, los resultados muestran que es posible alcanzar reducciones relevantes en las tasas de error del sistema en el rango $6 \leq P \leq 8$ con ambas variantes del método nn-GMM y forced-Viterbi. Este resultado se confirma al observar la Tabla 4, la que muestra los resultados obtenidos con Viterbi-Poly calculando el error del sistema con

cada *handset* en condiciones de *mismatch* (`hset2`, `hset3`,..., `hset7`) probado individualmente. De este resultado se puede ver que al emplear la técnica propuesta en este capítulo en cada experimento usando un único *handset* en condiciones de *mismatch* puede llevar al más bajo EER en cuatro de seis casos al usar P en el rango $6 \leq P \leq 8$. En los otros dos casos, el EER alcanzado en el intervalo $6 \leq P \leq 8$ es sólo 0.7% y 0.3% mayor que el más bajo ERR alcanzado en los intervalos $2 \leq P \leq 5$ y/o $9 \leq P \leq 10$.

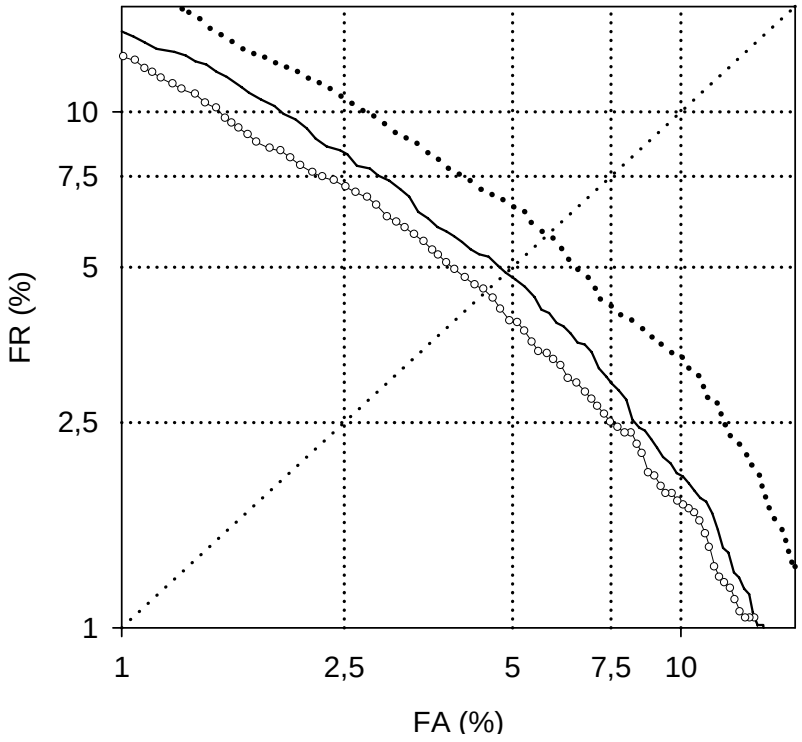


Figura 17. Curvas DET obtenidas con el sistema *baseline* (· · ·), Viterbi-SBR (——) y Viterbi-Poly, usando un orden polinomial P igual a 6(– ∘ –).

Tabla 4. EER(%) obtenidos con el sistema *baseline* y Viterbi-Poly con cada *handset* en condiciones de *mismatch* probado individualmente. En negrita, el menor EER alcanzado en cada canal.

		Canal					
		hset2	hset3	hset4	hset5	hset6	hset7
baseline		3,54	4,06	10,83	3,60	5,66	6,67
Viterbi-poly	$P = 2$	**3,33**	4,38	9,34	3,41	5,42	5,83
	$P = 3$	3,33	4,17	8,56	3,65	5,08	4,32
	$P = 4$	3,59	3,96	8,75	3,37	5,20	4,47
	$P = 5$	3,75	3,96	8,78	3,54	5,21	4,28
	$P = 6$	3,56	**3,73**	7,87	**2,96**	**4,83**	4,30
	$P = 7$	3,34	3,87	**7,48**	3,54	5,20	4,37
	$P = 8$	3,40	4,02	7,58	3,54	5,00	4,32
	$P = 9$	3,35	3,88	7,63	3,61	5,16	**4,27**
	$P = 10$	3,51	3,91	7,60	3,75	5,42	3,99

La Tabla 5 muestra resultados con $(6 \leq P \leq 8)$ aplicado de forma aislada y en combinación con RASTA, CMN y CMVN. Como se puede ver en la Tabla 5, Viterbi-Poly es capaz de reducir el ERR alcanzado con RASTA, CMN y CMVN en 15.2%, 12.9% y 4.2%, respectivamente. Sin embargo, el mejor EER es obtenido cuando Viterbi-Poly es aplicado sin ser combinado con técnicas de filtrado de trayectoria temporal de

características. Este resultado puede deberse al hecho de que RASTA, CMN y CMVN tienden a perder efectividad cuando son aplicados en tareas de datos limitados o elocuciones cortas.

Tabla 5. EER(%) obtenido con Viterbi-Poly con $P = \{6,7,8\}$ al ser aplicado de forma aislada y en combinación con RASTA, CMN y CMVN.

	Viterbi-Poly $P = 6$	Viterbi-Poly $P = 7$	Viterbi-Poly $P = 8$
Aislado	4.49	4.47	4.46
RASTA	4.60	4.77	4.70
CMN	4.82	4.82	4.80
CMVN	5.98	5.97	5.72

4.5.2 Evaluación de pronunciación basada en reconocimiento de voz

La Figura 18 y la Tabla 6 muestran los resultados de correlación promedio entre los puntajes subjetivos-objetivos en el sistema CAPT empleado en este capítulo empleando el sistema *baseline* (i.e. sin aplicar ninguna compensación a la distorsión de canal), y empleando las técnicas Viterbi-SBR y Viterbi-Poly $\left(P = \{6,7,8,9\}\right)$, propuesta en este capítulo. Como se puede ver en la Figura 18 y la Tabla 6, Viterbi-SBR puede aumentar significativamente la correlación subjetiva-objetiva de la salida del sistema CAPT de

0.705 a 0.720. Sin embargo, los mejores valores de correlación son los observados al emplear el modelo polinomial presentado en esta tesis, al utilizar un rango amplio de ordenes polinomiales la correlación subjetiva-objetiva promedio puede ser aumentada hasta 0.726.

Tabla 6. Correlación promedio entre los puntajes subjetivos-objetivos en el sistema CAPT empleado en este capítulo, obtenidos con los sistemas *baseline*, Viterbi-SBR y Viterbi-Poly

	baseline	Viterbi-SBR	Viterbi-Poly $P = 6$	Viterbi-Poly $P = 7$	Viterbi-Poly $P = 8$	Viterbi-Poly $P = 9$
Correlación	0.705	0.720	0.721	0.717	0.723	0.726

Al igual que en el caso de verificación de locutor texto-dependiente, se puede notar en la Figura 18 y la Tabla 6 que el método propuesto depende fuertemente del orden polinomial P. A pesar de este hecho, los resultados muestran que es posible alcanzar aumentos muy relevantes en la correlación subjetiva-objetiva del sistema en el rango $6 \leq P \leq 10$ con la variante forced-Viterbi del método, al compararse al sistema *baseline* y a Viterbi-SBR.

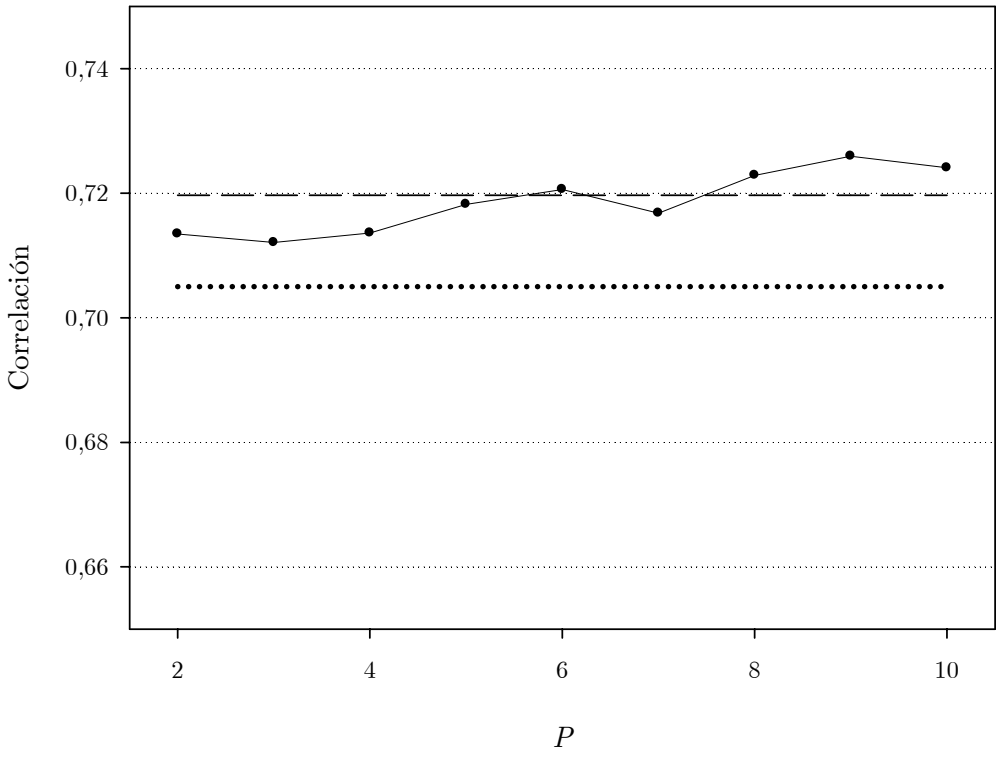


Figura 18. Correlación promedio entre los puntajes subjetivos-objetivos en el sistema CAPT empleado en este capítulo vs. orden del la función polinomial empleada en Viterbi-Poly. La curva se contrasta con los resultados obtenidos con los sistemas *baseline* y Viterbi-SBR.

4.6 Conclusiones

Una novedosa técnica de compensación de la distorsión de canal en el dominio de las características, basada en una aproximación polinomial fue propuesta en este capítulo. El método presentado modela la distorsión empleando una función polinomial en el dominio del logaritmo de las energías del banco de filtros Mel. El método descrito en

este capítulo considera en su modelamiento la continuidad de la respuesta en frecuencia del canal. Por el hecho de usar un modelo paramétrico, el esquema propuesto puede reducir el número de parámetros requeridos en la estimación de la distorsión al imponer restricciones apropiadas.

Los resultados con verificación de locutor texto-dependiente muestran que al cancelar la distorsión de canal con el modelo propuesto se pueden llegar a reducciones relativas en el EER tan grandes como 22% y 8%, respectivamente, al compararse con el sistema *baseline* y con una técnica convencional de cancelación de la componente constante en el dominio cepstral, siempre usando datos limitados en cada intento de verificación. Al observar los resultados obtenidos con el sistema de evaluación de pronunciación basado en ASR con verificación de locutor texto-dependiente muestran que al suprimir el efecto de la distorsión de canal generada por micrófonos de baja calidad empleando el modelo propuesto es posible aumentar la correlación subjetiva-objetiva del sistema desde 0.705 a 0.726. Resultado significativo según las magnitudes de desempeño encontradas en la literatura. Los resultados conseguidos con ambos sistemas sugieren que al usar métodos de búsqueda de vecino más cercano en la etapa de estimación, la carga computacional se mantiene controlada tanto en la versión basada en nn-GMM como la basada en forced-Viterbi del esquema aquí presentado. Por ejemplo, el tiempo de procesamiento total del sistema de TD-SV es aumentado solo en un 4.2%

Finalmente, es posible proponer como trabajo de investigación futuro la aplicación del esquema de cancelación de canal presentado en este capítulo en combinación con otros métodos de remoción de canal y/o a otras tareas en el área de procesamiento de voz, como lo es el reconocimiento automático de voz.

Capítulo 5

Conclusiones

5.1 Análisis y discusiones finales

La presente tesis ha abordado el problema del *mismatch* por distorsión de canal de comunicaciones en sistemas de procesamiento de patrones de voz, con especial énfasis en dos aplicaciones que operan con dispositivos de captura de baja calidad y con señales cortas en las etapas de entrenamiento y test: un sistema telefónico de verificación de locutor texto-dependiente y una plataforma para evaluación automática de pronunciación basada en reconocimiento de voz. La motivación principal de este trabajo de investigación fue la de contribuir a la generación de aplicaciones de procesamiento de voz robustas a los efectos del canal de comunicaciones y que a la vez cumplan con requerimientos de usabilidad, con objeto de facilitar su masificación.

Esta tesis enfrentó el problema de la distorsión de canal proponiendo dos métodos de cancelación: una técnica de parametrización robusta y un esquema de compensación de distorsión en el espacio de las características. Para modelar la distorsión por canal telefónico y micrófonos de baja calidad en señales cortas, ambas propuestas incluyeron la hipótesis de dependencia de la respuesta del canal con respecto a la señal de entrada.

En primer lugar esta tesis propuso una novedosa transformación de características aplicada a verificación de locutor texto-dependiente con datos limitados para la robustez a la variabilidad de canal de comunicaciones. La transformación es aplicada en un esquema *frame*-por-*frame* como un filtro pasa-banda a lo largo del vector de características de energías del banco de filtros Mel. Así, es posible reducir el efecto variable en el tiempo de la componente de distorsión de canal en el dominio cepstral, generado por la dependencia de la respuesta del canal en la señal de entrada. El esquema presentado emplea análisis de importancia relativa en combinación con una función discriminativa basada en la razón intra-locutor/inter-locutor para definir los parámetros del filtrado. Los resultados obtenidos con verificación de locutor texto-dependiente muestran que la técnica propuesta entrega significativas mejoras en el EER al ser empleada de manera aislada o en combinación con filtrados de trayectorias temporales de características como CMN o RASTA.

Luego, la presente tesis presentó una nueva estrategia para robustez a canal mediante compensación de características aplicada a verificación de locutor texto-

dependiente y evaluación de pronunciación basada en reconocimiento de voz. La técnica propuesta modela la distorsión de canal empleando una función polinomial en el dominio del logaritmo de las energías del banco de filtros Mel. Al usar este modelo paramétrico aparecen dos efectos positivos: se considera la continuidad de la respuesta en frecuencia del canal de comunicaciones y es posible reducir el número de parámetros requeridos en la estimación de la distorsión. Los resultados alcanzados con verificación de locutor texto-dependiente y evaluación automática de pronunciación basada en reconocimiento de voz muestran que, al ser comparado con el sistema *baseline* y con técnicas convencionales de compensación de componente constante de canal en el dominio cepstral, el método presentado mejora notablemente el desempeño de ambos sistemas en condiciones adversas de canal usando señales cortas de prueba.

5.2 Trabajo Futuro

Como trabajo futuro es posible proponer la aplicación de las técnicas presentadas en otras tareas de procesamiento de patrones acústicos, no consideradas en la presente tesis, y que se vean enfrentados al problema de la distorsión de canal en situaciones prácticas reales, tales como: reconocimiento de voz, reconocimiento de locutor, separación de locutores, verificación de locutor texto-independiente, entre otras. Además, el uso de los métodos propuestos en esta tesis en combinación con otros esquemas de cancelación de canal y/o para robustez a otros efectos que deterioran el

rendimiento de sistemas de procesamiento de patrones de voz tales como el ruido aditivo y variabilidad de locutor puede ser propuesto como trabajo futuro.

Referencias

- Abdou Sh., Hamid S., Rashwan M., Samir A., Abdel-Hamid O., Shahin M. y Nazih W., 2006. Computer Aided Pronunciation Learning System Using Speech Recognition Techniques, in Proceedings Interspeech 2006, Pittsburgh, PA, EE.UU.
- Acero A., Li Deng, Kristjansson T. y Zhang J., HMM adaptation using vector Taylor series for noisy speech recognition, In Proceedings ICSLP 2000, Beijing, China, pp. 869-872, 2000.
- Acero A., 1993. Acoustical and environmental robustness in automatic speech recognition. Kluwer Academic Pub., Dordrecht.
- Afify M., Gong Y. y Haton J., 1998. A general joint additive and convolutive bias compensation approach applied to noise Lombard speech recognition. IEEE Transactions on Speech and Audio Processing, 6 (6), pp. 524-538.
- Becchetti C. y Prina L., 1999. Speech recognition, theory and c++ implementation. Wiley E. London, Reino Unido.
- Becerra Yoma N. y Villar M., 2002. Verificación de identidad de individuos mediante la voz. Revista Ciencia Abierta. Universidad de Chile, 19.
- Bimbot F., Bonastre J.-F., Fredouille C., Gravier G., Magrin-Chagnolleau I., Meignier S., Merlin T., Ortega-Garcia J., Petrovska-Delacretaz P. y Reynolds D., 2004. A Tutorial on Text-Independent Speaker Verification. EURASIP Journal on Applied Signal Processing, 2004 (4), pp. 430-451.
- Bonaventura P., Herron D. y Menzel W., 2000. Phonetic rules for diagnosis of pronunciation errors, in Proceedings of Konvens 2000 (Conference on Natural Language Processing), pp. 225-230, Ilmenau, Alemania.
- Campbell J. y Higgins A., 1994. YOHO Speaker Verification. Linguistic Data Consortium, Philadelphia.

- Chen C.-P. y Bilmes J.-A., 2007. MVA Processing of Speech Features. IEEE Transactions on Speech and Audio Processing, 15 (1), pp. 257-270.
- Cook G.-D., Kershaw D.-J., Christie J.D.M., Seymour C.-W. y Waterhouse S.-R., 1997. Transcription of broadcast television and radio news: the 1996 abbot system. In Proceedings ICASSP 1997, Munich, Alemania.
- Cucchiarini C., Strik H., y Boves L., 1998. Automatic pronunciation grading for dutch. In Proceedings Still, ESCA Workshop, pp. 95-99.
- Cui X. y Alwan A, 2005. Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR," IEEE Transactions on Speech and Audio Processing, 13(6), pp.1161-1172.
- Damper R.-I. y Higgins J.-E., 2003. Improving speaker identification in noise by subband processing and decision fusion. Pattern Recognition Letters, 24 (13), pp. 2167-2173.
- De la Torre A., Peinado A.-M., Segura J.-C., Perez-Cordoba J.-L., Benitez M.-C. y Rubio A.-J., 2005. Histogram equalization of speech representation for robust speech recognition. IEEE Transactions on Speech and Audio Processing, 13 (3), pp. 355-366.
- Doddington G.R., 1985. Speaker recognition: identifying people by their voices. Proceedings of the IEEE, 73 (11), pp. 1651-1664.
- Duda R. y Hart P., 1973. Pattern Classification and Scene Analysis, Wiley-Interscience, Nueva York.
- Forsyth M., 1995. Discriminating observation probability (DOP) HMM for speaker verification. Speech Communication, 17, pp. 117-129.
- Franco H., Neumeyer, L. Kim Y. y Ronen O., 1997. Automatic pronunciation scoring for language instruction, In Proceedings ICASSP 1997, Munich, Alemania, pp. 1471-1474.
- Furui S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Speech and Audio Processing, 29 (2), pp.254-272.
- Furui S., 1994. An overview of speaker recognition technology. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9.

- Furui S., 2005. Recent progress in corpus-based spontaneous speech recognition. IEICE Transactions on Information and Systems, E88-D(3), pp. 366-375.
- Gales M., 1998. Maximum-likelihood linear transformation for HMM-based speech recognition," Computer Speech and Language, 12, pp. 75-98.
- Gales M. y Young S., 2008. The Application of Hidden Markov Models in Speech Recognition, Now Publishers.
- Garofalo J., Graff D., Paul D., y Pallett D., 1993. Continous Speech Recognition (CSR-I) Wall Street Journal (WSJ0) news. Linguistic Data Consortium, Philadelphia.
- Gauvain J.L. y Lee C.H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains. IEEE Transactions on Speech Audio Processing, 2, pp. 291-298.
- Gillick L y Cox S.J., 1989. Some statistical issues in the comparison of speech recognition algorithms, in Proceedings. ICASSP 1989, vol. 1, pp. 532-535.
- Hamada H., Miki S., y Nakatsu R., 1993. Automatic evaluation of English pronunciation based on speech recognition techniques. IEICE Trans. Inf. and Sys., E76-D(3), pp. 352-359.
- Hacker C., Cincarek T., Maier A, Hebler, A. y Noth E., 2007. Boosting of Prosodic and Pronunciation Features to Detect Mispronunciations of Non-Native Children. In: Proc. ICASSP 2007, Honolulu, Hawaii, USA.
- Hamid S. y Rashwan M., 2004. Automatic Generation of Hypotheses for Automatic Diagnosis of Pronunciation Errors, in Proceedings of NEMLAR Conference on Arabic Language Resources and Tools, Cairo, Egypt.
- Heck L.-P., Konig Y., Kemal Sönmez M. y Weintraub M., 2000. Robustness to telephone handset distortion in speaker recognition by discriminative feature design. Speech Communication, 31(2-3), pp. 181-192.
- Hermansky H. y Morgan N., 1994. RASTA processing of speech. IEEE Transactions Speech and Audio Processing, 2 (4), pp. 578-589.

- Hermansky, H., Morgan, N., Bayya, A. y Kohn, P., 1992. RASTA-PLP speech analysis technique. in Proc. ICASSP 92, vol. I, pp. 121-124, San Francisco, March 1992.
- Huerta J.M., 2002. Alignment-based codeword-dependent cepstral normalization, IEEE Trans. on SAP, 10 (7), pp. 451-459.
- Hung J.-W. y Lee L.-S., 2006. Optimization of temporal filters for constructing robust features in speech recognition. IEEE Transactions on Speech and Audio Processing, 14 (3), pp. 808-832.
- Jelinek F., 1997. Statistical methods for speech recognition. Massachusetts Institute of Technology. Chapter 1-5. pp. 1-90.
- Jiang H., Soong F. K., y Lee C.-H., 2001. Hierarchical stochastic feature matching for robust speech recognition, in Proceedings. ICASSP, Salt Lake City, UT, pp. 217-220.
- H.Y. Jung, 2004. Filtering of Filter-Bank Energies for Robust Speech Recognition, ETRI Journal, 26 (3), pp.273-276, 2004.
- Kajarekar S. S., Scheffer N., Graciarena M., Shriberg E., Stolcke A., Ferrer L., y Bocklet T., 2009. The SRI NIST 2008 Speaker Recognition Evaluation System, In Proceedings ICASSP 2009, Taipei, Taiwan, pp. 4205-4208.
- Kanedera, N., Arai T., Hermansky H. y Pavel M., 1999. On the relative importance of various components of the modulation spectrum for automatic speech recognition, Speech Communication, 28(1), pp. 43-55.
- Lamel L.F., Rabiner L.R., Rosenberg A.E. y Wilpon J.G., 1981. An improved endpoint detector for isolated word recognition. IEEE Transactions on Acoustics Speech and Signal Processing, 29, pp. 777-785.
- Laurila K., Vasilache M. y Viikki O., 1998. A combination of discriminative and maximum likelihood techniques for noise robust speech recognition. IEEE Conference on Acoustics, Speech and Signal Processing. pp. 12-15.
- Leggetter C.J. y Woodland P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language, 9 (4),806-814.

- Mahadeva Prasanna S.R., Zachariah, J.-M. y Yegnanarayana, B., 2004. Neural network models for combining evidence from spectral and suprasegmental features for text-dependent speaker verification. In: Proc. ICISIP 2004, Chennai, India.
- Mak M-.W., Tsang C.-L., y Kung S.-Y., 2004, Stochastic feature transformation with divergence-based out-of-handset rejection for robust speaker verification. EURASIP J. on Applied Signal Processing, 4, pp. 452-465.
- Mak M.-W., Yiu K.-K. y Kung S.-Y., 2007. Probabilistic feature-based transformation for speaker verification over telephone networks. Neurocomputing, 71 (1-3), pp. 137-146.
- Martin A., Doddington G., Kamm T., Ordowski M. y Przybocki M., 1997. The DET curve in assessment of detection task performance. Proceedings of Eurospeech, Rodas, Grecia, pp 1895-1898.
- Meng X., Wu Z., Huang P., Zhan S. y Zhang B., 2008. Automatic detection of pronunciation errors in CAPT systems based on confidence measure. In: Proc. ICIA 2008, ZhangJiaJie, Hunan, China.
- Molina C., Becerra Yoma N., Wuth J. y Vivanco H., 2009, ASR based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion. Speech Communication, 51(6), pp. 485-498.
- Moreno P. J., Raj B., Gouvêa E. y Stern R., 1995. Multivariate-Gaussian-Based Cepstral Normalization for Robust Speech Recognition. Proc. of the ICASSP, Detroit, MI, EE.UU.
- Moustroufas, N. y Digalakis, V., 2007. Automatic pronunciation evaluation of foreign speakers using unknown text, Computer Speech and Language 21(1), pp. 219-230.
- Nadeu C., Macho D. y Hernando J., 2001. Time and frequency filtering of filter-bank energies for robust HMM speech recognition, Speech Communication, 34 (1-2), pp. 93-114, 2001.
- National Institute of Standards and Technology (NIST), 2006. The NIST Year 2006 Speaker Recognition Evaluation Plan (http://www.nist.gov/speech/tests/spk/2006/).

- Nealand J.-H., Pelecanos, J.-W., Zilca, R.-D. y Ramaswamy, G.-N, 2005. A study of the relative importance of temporal characteristics in text-dependent and text-constrained speaker verification. In: Proc. ICASSP 2005, Philadelphia, PA, USA.
- Neumeyer L., Franco H., Weintraub M. y Price. P., 1996. Automatic text-independent pronunciation scoring of foreign language student speech. In: Proc. ICSLP '96, Vol. 3, Philadelphia, PA, pp. 1457-1460.
- Neustein, A., 2010. Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, Springer.
- Openshaw J.P., Sun S.P. y Mason J.S., 1993. A comparison of composite features under degraded speech in speaker recognition. Proceedings of ICASSP, Minneapolis, EE.UU., 2, pp. 371-374.
- Oppenheim A.-V., Willsky A.-S. y Nawab S.-H., 1997. Signals and Systems, Prentice Hall.
- Picone J., 1993. Signal modeling techniques in speech recognition. Proceedings of the IEEE, 81 (9), pp. 1215-1247.
- Preti A., Ravera B, Capman F. y Bonastre J.-F., 2008. An application constrained front end for speaker verification, in Proceedings of 16th European Signal Processing Conference, Lousanne, Suiza.
- Rabiner L. R., Juang B. H. y Lee C. H., 1996. An overview of automatic speech recognition. Automatic Speech and Speaker Recognition: Advanced Topics, C. H. Lee, F. K. Soong and K. K. Paliwal editores, Kluwer Academic Publisher, pp. 1-30.
- Rabiner L.R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77 (2), pp. 257-286.
- Rahim M.G. y Huang B.H., 1996. Signal bias removal by maximum likelihood for robust telephone speech recognition. IEEE Transactions on Speech and Audio Processing, 4 (1), pp. 19-30.
- Reynolds D.A., 2003. Channel robust speaker verification via feature mapping. In Proceedings ICASSP 2003, Hong Kong, China, pp. 53-56.

- Reynolds D.A., Quatieri T.F. y Dunn R.B., 2000. Speaker verification using adapted Gaussian mixture models, Digital Signal Processing, 10(1), pp. 19-41.
- Reynolds D.A., Zissman M.A., Quatieri T.F., O'Leary G.C. y Carlson, B.A., 1995. The effects of telephone transmission degradations on speaker recognition performance. In Proceedings ICASSP 1995, Detroit, MI, USA, pp. 329-332.
- Rong Zheng, Shuwu Zhang y Bo Xu, 2005. A Comparative Study of Feature and Score Normalization for Speaker Verification, Lecture Notes in Computer Science 3832, pp. 531-538.
- Savoji M.H., 1989. A robust algorithm for accurate endpointing of speech signals. Speech Communication, 8 (1), pp. 45-60.
- Schwartz R., Chow Y. L., Kimbal O., Roucos S., Krasner M., y Makhoul J., 1985. Context-dependent modelling for acoustic-phonetic recognition of continuous speech. Proceedings of ICASSP, pp. 1205-08, Trampa, FL.
- Sevenster B., de Krom G., y Bloothoof G., 1998. Evaluation and training of second-language learners' pronunciation using phoneme-based HMMs," In Proceedings STiLL, pp 91-94, Marholmer, Sweden, ESCA Workshop.
- Sivakumaran P., Ariyaeeinia A.-M. y Loomes M.-J., 2003. Sub-band based text-dependent speaker verification, Speech Communication 41(2-3), pp. 485-509.
- Skosan M. y Mashao D., 2006. Modified Segmental Histogram Equalization for robust speaker verification. Pattern Recognition Letters, 27 (5), pp. 479-486.
- Sorell M., 2009. Forensics in Telecommunications, Information and Multimedia: Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, Vol. 8.
- Souilmi Y., Rigazio L., Nguyen P., Kryze D. y Junqua J.-C., 2002. Blind channel estimation based on speech correlation structure. In: Proc. ICASSP 2002, Orlando, FL, USA.
- Thomas S., Ganapathy S., Hermansky H., 2008. Recognition of Reverberant Speech Using Frequency Domain Linear Prediction. IEEE Signal Processing Letters, 15, pp. 681-684.

- Tufekci Z., 2007. Convolutional Bias Removal Based on Normalizing the Filterbank Spectral Magnitude. IEEE Signal Processing Letters, 14 (7), pp. 485-488.
- Subramanian M., Mohan S. y Mahajan A., 2010. Speaker recognition, Lap Lambert Academic Publishing.
- van Vuuren S. y Hermansky H., 1998. On the importance of components of the modulation spectrum for speaker verification, In Proceedings ICSLP 1998, pp. 3205-3208.
- Wang L., Kitaoka N. y Nakagawa S., 2007. Robust distant speech recognition by combining position-dependent CMN with conventional CMN, in Proceedings ICASSP 2007, pp. 817-820.
- Wolfel M., 2009. Enhanced Speech Features by Single-Channel Joint Compensation of Noise and Reverberation. IEEE Transactions on Audio, Speech, and Language Processing, 17 (2), pp. 312-323.
- Yang X., Millar J.B. y Macleod I., 1996. On the sources of inter- and intra-speaker variability in the acoustic dynamics of speech. Proceedings of ICSLP, Philadelphia, EE.UU., pp. 1792-1795.
- Yiu K.K., Mak M.W. y Kung S.Y., 2006. Blind stochastic feature transformation for channel robust speaker verification, Journal of VLSI Signal Proc., 42 (2), pp. 117-126.
- Zheng-Hua Tan, y Boerge Lindberg, 2010. Automatic Speech Recognition on Mobile Devices and Over Communication Networks, Advances in Pattern Recognition Series, Springer.

Anexo

Publicaciones del autor

Las siguientes son las publicaciones generadas como parte del trabajo de tesis:

Publicaciones en revistas ISI como primer autor

1.- Claudio Garretón y Néstor Becerra Yoma, “Telephone channel compensation in speaker verification using a polynomial approximation in the log-filter-bank energy domain,” Aceptado para su publicación en IEEE Transactions on Audio, Speech and Language Processing. 2011.

2.- Claudio Garretón, Néstor Becerra Yoma y Matías Torres. “Channel robust feature transformation based on filter-bank energy filtering,” IEEE Transactions on Audio, Speech and Language Processing, Vol. 18, No. 5, pp. 1082 - 1086. 2010.

Publicaciones en congresos internacionales como primer autor

1.- Claudio Garretón y Néstor Becerra Yoma, "On enhancing feature sequence filtering with filter-bank energy transformation in speaker verification with telephone speech", En proceedings INTERSPEECH 2010, Makuhari, Japan, Septiembre 26-30, pp. 1461-1464.