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

RECONOCIMIENTO ROBUSTO DE PATRONES ACÚSTICOS BASADOS EN EL SISTEMA AUDITIVO PERIFÉRICO

TESIS PARA OPTAR AL GRADO DE DOCTOR EN INGENIERÍA ELÉCTRICA

VÍCTOR HERNÁN POBLETE RAMÍREZ

PROFESOR GUÍA:

NÉSTOR BECERRA YOMA

MIEMBROS DE LA COMISIÓN:

JOHN ATKINSON ABUTRIDY
CARLOS BUSSO RECABARREN
JORGE SILVA SÁNCHEZ

Este trabajo ha sido parcialmente financiado por el Programa de Becas para estudios de Doctorado año 2010 de CONICYT

SANTIAGO DE CHILE
Julio 2014

Reconocimiento robusto de patrones acústicos basados en el sistema auditivo periférico

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

Víctor Hernán Poblete Ramírez
Profesor guía: Néstor Becerra Yoma

Santiago de Chile, Julio de 2014.

La verificación de locutor (SV) por biometría de voz, se ha integrado en diversas aplicaciones como interfaz de comunicación entre personas y máquinas. Sin embargo, su principal inconveniente es enfrentar variabilidades o *mismatch* entre condiciones de entrenamiento y prueba. La robustez es la propiedad que le permite mantener su funcionamiento superando perturbaciones. En SV existe necesidad de extraer parámetros espectrales propios del locutor y robustos a ruido y a efectos de distorsión de canal. Varios métodos de extracción se inspiran en la fisiología periférica auditivo y en teorías de codificación neuronal de fibras del nervio auditivo (AN). El término "periférico" se utiliza para dar a entender aquella parte del sistema auditivo que es externo al sistema nervioso central (CNS). La salida del sistema auditivo periférico es la actividad del AN. A su vez, esta actividad es la entrada al CNS.

Esta tesis aborda dos novedosos métodos inspirados en la periferia auditiva que contribuyen a la robustez de sistemas de SV, ante condiciones de *mismatch* por ruido aditivo y por variabilidad en el canal acústico de transmisión, en una tarea de reconocimiento de patrones acústicos: verificación de locutor de texto-independiente (TI-SV). El primero, es una función sigmoidal óptima, tasa-nivel, que es una componente de muchos modelos del sistema auditivo periférico. La optimización usa criterios definidos exclusivamente sobre la base de atributos físicos del sonido de entrada inspirados en evidencia fisiológica. Estos criterios, discriminan entre una señal de voz degradada, y ruido, para preservar la máxima cantidad de información en la región lineal de la curva sigmoidal y para minimizar la distorsión en regiones de saturación. El desempeño de la función sigmoidal se valida con experimentos de TI-SV, con señales degradadas por ruido aditivo a diferentes *SNRs*. Los resultados, comparados con el sistema *baseline* MFCC, muestran que el método propuesto, en combinación con normalización de varianza cepstral (CVN), conduce a reducciones relativas en EER, tan grandes como 40 % en ciertos SNRs.

A continuación, se presenta un nuevo conjunto de *features*, llamados Coeficientes Cepstrales Localmente-Normalizados (LNCCs), que se basan en el Detector de Sincronía Generalizada (GSD) de Stephanie Seneff. El método propuesto, motivado perceptualmente, permite normalizar en forma instantánea *features* de voz. La efectividad de los LNCCs se demuestra en una tarea de TI-SV a lo largo de una variedad de condiciones de *tilt* espectral en el canal acústico de transmisión. Los resultados, comparados con el sistema *baseline* MFCC y con MFCC+CMN, muestran que los LNCCs se caracterizan por requerir de baja complejidad computacional y por compensar más ampliamente el *tilt* espectral que los coeficientes MFCCs. Además, LNCCs no requieren el cálculo y almacenamiento de un promedio móvil de valores de *features*, proporcionando reducciones relativas en EER tan altas como 32 % y 35 % cuando se comparan con MFCC y MFCC+CMN, con *tilt* espectral variable, respectivamente. Es interesante destacar que los *features* LNCC pueden llegar a ser una alternativa a MFCC y MFCC+CMN, en cualquier situación donde es difícil estimar confiablemente la media cepstral. Ambas estrategias propuestas en esta tesis, comparadas con el sistema *baseline* MFCC, consiguen robustez del sistema TI-SV mejorando su desempeño frente a diversas condiciones de *mismatch*.

... Dedicado a mis amores Fernanda, Carolina y Marcia.

Agradecimientos

En primer lugar, quisiera agradecer a mi querida Marcia y a mis vidas Carolina y Fernanda, por su paciencia estos años, por su valor, comprensión e infinito amor que han ayudado a sobrellevar la distancia entre Valdivia y Santiago, incluso en los momentos más difíciles. Muchas gracias por existir.

También, quisiera extender mis agradecimiento al Profesor Néstor Becerra Yoma, por haberme dado la oportunidad de estudiar y desarrollar mi trabajo de doctorado en el Laboratorio de Procesamiento y Transmisión de Voz (LPTV). Por su valiosa dedicación, apoyo y consejos durante estos años.

Agradezco también al Profesor Richard M. Stern y al Profesor Simon King por haber compartido su tiempo, sus experiencias y conocimientos durante sus estadías aquí en el LPTV.

Este trabajo no habría sido posible sin la importante ayuda entregada por las autoridades de la Universidad Austral de Chile y del Instituto de Acústica de la misma Universidad.

Para el desarrollo de esta tesis se contó con el financiamiento de los Programas de Beca para estudios de Doctorado Nacional año 2010 de la Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) del Gobierno de Chile, y de la Beca para Gastos Operacionales del Proyecto de Tesis Doctoral año 2013 de CONICYT, para asistencia a exponer un trabajo al Congreso Internacional Interspeech 2013 realizado en la ciudad de Lyon, Francia en agosto de ese mismo año. Además, parte de este trabajo se realizó en el marco de los proyectos CONICYT-ANILLO ACT 1120 y CONICYT-FONDEF 1100195.

Quiero agradecer infinitamente a cada uno de los compañeros y ex-compañeros del LPTV por su ejemplo de amistad y solidaridad y por la increible ayuda recibida durante cada día. Y a todos los que de una u otra forma ayudaron a hacer realidad este trabajo.

Tabla de Contenido

Índice de figuras **x**

Glosario **xiii**

1. Introducción **1**
1.1. Verificación de locutor 1
1.2. Motivación 3
1.3. Definición de la propuesta de investigación 5
1.3.1. Descripción de la propuesta 5
1.3.2. Hipótesis 6
1.3.3. Objetivos generales y específicos 7
1.3.4. Metodología 8
1.4. Estructura de la tesis 9
1.5. Contribuciones de la tesis 10

2. Robustez en sistemas de verificación de locutor **12**
2.1. Introducción 12
2.2. Estructura de un sistema de verificación de locutor 13
2.2.1. Medida del Desempeño de un Sistema de Verificación de Locutor . . 14
2.2.2. Parametrización acústica de una señal de voz 16
2.2.3. Verificación de locutor de texto-independiente, con modelo GMM-UBM 18
2.2.3.1. Modelos de mezclas de Gaussianas 20
2.2.3.2. Sistema GMM adaptado 21
2.2.3.3. Normalización 22
2.3. Extracción robusta de características 23
2.3.1. Procesamiento auditivo de la voz 25
2.3.2. Componentes del sistema auditivo 26
2.3.3. Membrana basilar 27
2.4. Proceso activo en la cóclea 29
2.4.1. Amplificación activa 29
2.4.2. *Tuning* neuronal 30
2.4.3. No linealidad compresiva 30

2.4.4. Emisiones espontáneas otoacústicas 31
2.4.5. Curva *tuning* 31
2.5. Propiedades de las salidas del nervio auditivo 32
2.5.1. Respuesta síncrona en el nervio auditivo 32
2.5.2. Dinámica de las salidas del nervio auditivo 33
2.5.3. Relación no lineal entre la tasa de descarga versus el nivel de intensidad 33
2.6. Codificación de voz en el nervio auditivo: vocales 34
2.6.1. Codificación de vocales usando información temporal 34
2.7. Modelo auditivo de Seneff 34
2.7.1. Estimación de la tasa de descarga en las fibras del nervio auditivo . . 36
2.7.2. Estimación de sincronía en las fibras del nervio auditivo 36
2.7.3. Estrategias de procesamiento de señales de voz inspiradas en la fisiología de la audición para extracción de características 38

3. Optimización de parámetros de las funciones sigmoidales tasa-nivel basada en características acústicas 41
3.1. Introducción 42
3.1.1. Procesamiento neuronal de señales de voz 42
3.1.2. Extracción de características para verificación de locutor 44
3.1.3. La función sigmoidal tasa-nivel 45
3.2. Desarrollo de los criterios de optimización para la función sigmoidal 48
3.2.1. Especificación matemática de la función sigmoidal 49
3.2.2. Especificación de la función objetivo usada para optimizar la no linealidad sigmoidal 50
3.2.2.1. Criterio 1: Distorsión no lineal en la región lineal 50
3.2.2.2. Criterio 2: Potencia del ruido 50
3.2.2.3. Criterio 3: Similitud entre la entrada de voz limpia y la voz degradada 51
3.2.2.4. Criterio 4: Varianza de la señal de voz degradada por ruido después del procesamiento de la función sigmoidal 51
3.2.3. Especificación de la función objetivo completa 51
3.3. Implementación de la función sigmoidal tasa-nivel 52
3.4. Resultados experimentales 57
3.4.1. Dependencia general sobre el SNR y la presencia de no linealidad sigmoidal 59
3.4.2. Comparación de entrenamiento de parámetros de no linealidad a SNRs fijos versus *matched* 61
3.4.3. Comparación de los resultados de Chiu, Raj & Stern (2012) 62
3.4.4. Impacto de la estimación, específica del canal, de las no linealidades sigmoidales 63

3.4.5. Comparaciones para funciones sigmoidales óptimas entrenadas y probadas con diferente tipo de ruido 64
3.4.6. Comentarios generales 65
3.5. Conclusiones 67
3.6. Apéndice 68

4. Coeficientes cepstrales basados en la normalización local del espectro 69
4.1. Introducción 70
4.1.1. Motivación 70
4.1.2. La necesidad de *features* robustos de voz 71
4.1.2.1. Canales variables en el tiempo 72
4.1.2.2. Escenarios de aplicación 72
4.1.3. Alcances de este capítulo 74
4.2. Desarrollo del método propuesto a partir de un modelo auditivo 74
4.2.1. Modelamiento auditivo 75
4.2.1.1. Representaciones de tasa-media y la envolvente espectral . . 75
4.2.1.2. Tasa promedio de sincronía localizada (ALSR) 76
4.2.1.3. Del ALSR al detector generalizado de sincronía (GSD) . . . 76
4.2.2. El potencial de los *features* tipo-GSD para reconocimiento de voz . . 78
4.2.2.1. Previos intentos para usar este modelo 78
4.2.2.2. Un análisis en el dominio de la frecuencia del GSD 79
4.2.2.3. Respuestas espurias del GSD 79
4.3. Coeficientes cepstrales localmente normalizados 81
4.3.1. Del GSD a un modelo el el dominio de frecuencia apropiado para tecnología de voz 81
4.3.1.1. Respuesta de frecuencia del par propuesto de filtros autonormalizados 84
4.3.1.2. Robustez a *mismatch* de canal 84
4.4. Experimentos de verificación de locutor 87
4.4.1. Sistema de verificación de locutor 88
4.4.2. Extracción de *features* 88
4.4.3. Base de datos YOHO 89
4.4.4. Experimentos inciales: sensibilidad a ajustes de parámetros 89
4.4.4.1. Número de canales LNCC y ancho de banda de los filtros . . 90
4.4.4.2. Valor central mínimo del denominador ($d_{\min}$) 91
4.4.5. Experimento 1: micrófono distante simulado 91
4.4.5.1. Procesamiento de voz para simular la respuesta de frecuencia de un micrófono distante 91
4.4.5.2. Resultados 92
4.4.6. Experimento 2: Canales que varían rapidamente 92

4.4.6.1. Procesamiento de voz para simular una respuesta de canal variable en el tiempo 92
4.4.6.2. Resultados 93
4.4.7. Resumen de resultados 93
4.5. Conclusiones 94

5. Conclusiones **96**
5.1. Análisis y discusiones finales 96
5.2. Trabajo futuro 97

Bibliografía **98**

A. Publicaciones del autor **122**

Índice de figuras

2.1. *Componentes generales de un sistema verificador de locutor.* 13
2.2. *Curvas de falsa aceptación (FA) y falso rechazo (FR).* 15

3.1. *Diagrama en bloques del método de extracción de características propuesto.* . 53
3.2. *Diagrama en bloques para obtener parámetros óptimos $\hat{\omega}_j$ y $\hat{\mu}_j$ de la función sigmoidal.* 53
3.3. *La función objetivo $J(\omega_j, \mu_j)$ dibujada como una función de los parámetros de la función sigmoidal: (a) pendiente ω_j, y (b) posición sigmoidal μ_j. Los valores óptimos de $\hat{\omega}_j$ y $\hat{\mu}_j$ se indican por los círculos abiertos para cada uno de los 35 canales del banco de filtros.* 54
3.4. *Función sigmoidal óptima (línea sólida) y mapeo lineal correspondiente (línea punteada). Histogramas de potencia se representan para frames de voz degradada (barras rellenas) y frames solamente ruido (barras en blanco).* 55
3.5. *Funciones sigmoidales graficadas como una función de SNR.* 55
3.6. *Gráficos tridimensionales de funciones sigmoidales tasa-nivel entrenadas con voz degradada por ruido babble a SNR igual a 20 dB y 5 dB. El gráfico se rota para mostrar diferencias en pendiente y desplazamiento horizontal entre ambos conjuntos de funciones.* 56
3.7. *Comparación de funciones sigmoidales tasa-nivel entrenadas con ruidos restaurant y car, a SNR igual a 20 dB (derecha) y a 5 dB (izquierda).* 57
3.8. *EER durante verificación de locutor como función del SNR de información de prueba y SNR utilizado para desarrollar los parámetros de la función sigmoidal.* 60
3.9. *Comparación de EER como una función de SNR para voz en ruido babble, car y restaurant, respectivamente.* 61
3.10. *Comparación de EER como una función de SNR para voz en ruido babble, usando la función sigmoidal combinada con CVN.* 62
3.11. *Comparación de EERs(%) conseguidos utilizando el método con parámetros de no linealidad dados por Chiu, Raj & Stern (2012), obtenidos con CVN, y el método propuesto combinado con CVN, para voz en ruido car, babble, y restaurant, respectivamente.* 63

3.12. *Comparación de EER(%) obtenido utilizando el método propuesto, combinado con CVN, empleando diferentes funciones sigmoidales por canal, y el mismo valor medio para todos los canales, para voz en ruido car, babble y restaurant, respectivamente.* 64
3.13. *Comparación de EER(%) entre método propuesto con parámetros de no linealidad entrenados (a SNR igual a 10 dB) y probados con el mismo tipo de ruido; y método sugerido usando parámetros sigmoidales entrenados con ruido rosa (a SNR igual a 10 dB).* 65
3.14. *Comparación del espectro promedio de potencia para 50 elocuciones de voz limpia, con el espectro de potencia del ruido. Los errores cuadráticos medio (MSE), entre la voz y ruidos restaurant, babble y car, son 58.7, 70.1 y 95.4, respectivamente.* 66

4.1. *Respuesta de frecuencia del numerador y denominador de un GSD sintonizado a 692 Hz. El numerador en la Ecuación 4.2 se muestra como una línea sólida y el denominador en la Ecuación 4.3 se presenta con líneal discontinua.* . . . 80
4.2. *Respuesta de frecuencia de un canal GSD a f_i^c = 692 Hz.* 80
4.3. *Magnitud logarítmica de la respuesta del numerador y denominador del GSD, sintonizado a f_i^c = 692 Hz. El numerador (Equación 4.2) se muestra con una línea gruesa, en tanto que el recíproco del denominador (Equación 4.3) se presenta con línea fina.* 81
4.4. *Formas de la magnitud de los filtros numerador (línea sólida) y denominador (línea discontinua), para un único canal del banco de filtros propuesto auto-normalizado.* 82
4.5. *Respuesta de frecuencia del numerador y denominador sintonizados en f_i^c = 515 Hz.* 83
4.6. *Diagrama de flujos para extracción de features LNCC (izquierda) y MFCC (derecha).* 85
4.7. *Respuesta de frecuencia del numerador y denominador, separadamente, ambos sintonizados en f_i^c = 515Hz, sobre una escala logarítmica.* 86
4.8. *Respuesta de frecuencia del numerador dividido por el denominador, ambos sintonizados en f_i^c = 515Hz, sobre una escala logarítmica.* 86
4.9. *Envolventes espectrales para un único frame de voz sonora, utilizando un banco de filtros tradicional en escala Mel (línea sólida) y para el banco propuesto LNCC (línea discontinua).* 87
4.10. *Envolventes espectrales para un único frame de voz sonora, utilizando un banco de filtros convencional en escala Mel (figura superior), y empleando el banco de filtros propuestos LNCC (figura inferior).* 87
4.11. *Sensibilidad al ancho de banda del filtro. Ambos con d_{min}=0.001. (a) 14 canales, (b) 28 canales.* 90
4.12. *Sensibilidad a parámetro d_{min}. LNCC con 28 canales, B=3 Barks.* 91

4.13. *Desempeño durante tilt espectral constante. Los features LNCC se calculan utilizando 28 canales,* d_{min}*=0.001 y* B*= 3 Barks.* 92
4.14. *Desempeño durante tilt espectral variable.* $VaryingST_1$*: 0 dB/oct* $\rightarrow -6$ *dB/oct.* $VaryingST_2$*: 0 dB/oct* $\rightarrow -6$ *dB/oct* $\rightarrow$ *0 dB/oct.* $VaryingST_3$*: 0 dB/oct* $\rightarrow -6$ *dB/oct* $\rightarrow$ *0 dB/oct* $\rightarrow -6$ *dB/oct.* 94

Glosario

ALSR:	*Average Localized Synchrony Rate*
AN:	*Auditory Nerve*
ASR:	*Automatic Speech Recognition*
CF:	*Characteristic Frequency*
CMN:	*Cepstral Mean Normalization*
CMVN:	*Cepstral Mean and Variance Normalization*
CNS:	*Central Nervous System*
DCT:	*Discrete Cosine Transform*
DFT:	*Discrete Fourier Transform*
EM:	*Expectation-Maximization Algorithm*
EER:	*Equal Error Rate*
FA:	*False Acceptance*
FR:	*False Rejection*
GMMs:	*Gaussian Mixture Models*
GSD:	*Generalized Synchrony Detector*
IHCs:	*Inner Hair Cells*
LNCCs:	*Locally-Normalized Cepstral Coefficients*
MAP:	*Maximum A Posteriori*
MFCCs:	*Mel Frequency Cepstral Coefficients*
OHCs:	*Outer Hair Cells*
PLP:	*Perceptual Linear Prediction*
RASTA:	*Relative Spectral Transform*
SNRs:	*Signal-to-Noise Ratios*
SPL:	*Sound Pressure level*
SD:	*Speaker Dependent*
SI:	*Speaker Independent*
SV:	*Speaker Verification*
TI-SV:	*Text-Independent Speaker Verification*
TD-SV:	*Text-Dependent Speaker Verification*
UBM:	*Universal Background Model*
VAD:	*Voice Activity Detector*
WER:	*Word Error Rate*

Capítulo 1

Introducción

Un sistema de autentificación de la identidad de una persona basado en características biométricas como la voz, el iris o la huella digital, se considera más seguro y más personal, que otro sistema que se base en *password* o bien, en tarjeta magnética. Esto se debe al hecho que tales características biométricas pertencen a la propia persona y ellas no se pueden olvidar o extraviar, lo cual sí puede suceder con los otros métodos. Además, la voz, opuesta a las anteriores características biométricas, permite que el reconocimiento se realice en forma remota, siendo también, fácil de transmitir a través de un canal de comunicación tal como un canal telefónico (Fazel & Chakrabartty, 2011). En los últimos años, las tecnologías de voz, se han integrado como interfaz de comunicación entre las personas y las máquinas, en diversas aplicaciones, por ejemplo, en sistemas de telecomunicaciones, robótica y multimedia (por ejemplo, Becerra Yoma *et al.* (2013b)). Esto permite que las tecnologías de voz evolucionen, en velocidad de funcionamiento, confiabilidad y eficiencia, junto con la propia evolución de las otras tecnologías de telecomunicaciones y multimedia (Wang *et al.*, 2011).

1.1. Verificación de locutor

La tarea de reconocer a una persona por su voz, corresponde a determinar automáticamente la identidad de aquella persona usando solamente información específica incluida en la señal de voz (Rose, 2002; Furui, 1994).

La identificación automática de locutor consiste en identificar a una persona a partir de una base de datos finita de locutores registrados conocidos. El locutor desconocido es comparado con aquellos registrados en esta base de datos y el mejor locutor que se iguale, se regresa como la decisión de identificación (Kinnunen *et al.*, 2011).

Por otro lado, la tarea de verificación de locutor consiste en decidir si una muestra dada de voz (una elocución), producida por un locutor corrresponde a la identidad que demanda (¿es

la persona quien dice ser?). La persona desconocida demanda una identidad la cual podría ser aceptada o rechazada por el sistema. La señal de voz emitida por la persona desconocida se compara con un modelo de voz del usuario registrado cuya identidad está siendo demandada. Si ese modelo de voz de la persona registrada y la muestra de voz desconocida, coinciden dentro de un cierto límite permitido (umbral de decisión), la identidad será aceptada, o bien, en caso contrario, rechazada por el sistema (por ejemplo, Becerra Yoma & Villar (2002)).

Existe varios tipos de sistemas de verificación de locutor. Dependiendo del método de operación o la aplicación, estos sistemas se pueden clasificar en texto-dependiente (TD-SV), o bien texto-independiente (TI-SV) (Furui, 1994; Garretón, 2011).

Un sistema de verificación de locutor, de texto dependiente, (TD-SV), requiere que una persona pronuncie una frase o secuencia de palabras, previamente determinadas por el sistema. Al contrario, un sistema de verificación de locutor, de texto independiente, (TI-SV), se prepara para verificar la identidad de la persona cualquiera sea la frase o secuencia de palabras usadas (Bimbot *et al.*, 2004).

Dentro de estos tipos de sistemas es posible también, distinguir aquellos que corresponden a sistemas de palabra aislada o sistemas de pronunciación continua. En un sistema de verificación de locutor de palabra aislada, la persona debe pronunciar una secuencia de palabras aisladas separadas por un corto intervalo de silencio para evitar que palabras sean influenciadas por la palabra inmediatamente pronunciada antes o después (Furui, 1986; Zhao, 1994). Así, en un sistema de verificación de locutor de pronunciación continua, la persona puede pronunciar un conjunto de palabras de cualquier forma que quiera.

Los sistemas de verificación de locutor TI-SV, se centran en aplicaciones comerciales tales como, contraseñas de voz para acceso telefónico. Mientras que sistemas de verificación de locutor TI-SV, están más ligados a la tarea de identificación forense (Fazel & Chakrabartty, 2011; Rose, 2002).

Un sistema de verificación de locutor necesita un conjunto de usuarios registrados llamados también, clientes enrolados. Para esto, se hace necesario desarrollar sesiones de entrenamiento con los usuarios que quieran emplear el sistema. El número de sesiones y el tiempo de separación entre ellas, va a influir sobre la exactitud y efectividad final del sistema. Esto se debe a que mientras más alto sea el número de sesiones de entrenamiento del cliente (mayor número de pronunciaciones), mejor resultará el modelo acústico representativo del cliente (Pradhan & Prasanna, 2011).

Por otra parte, las sesiones para el mismo cliente debieran ser separadas en tiempo para incluir variaciones propias de la voz ya sea por salud o estado emocional. Quien administre

el sistema, debe decidir por sesiones más cortas (no más de tres minutos) pero más frecuentes.

El número de sesiones de entrenamiento queda determinado por el nivel de seguridad que requiera la aplicación. Por lo que un sistema con seguridad máxima debe invertir tiempo importante y recursos en sesiones de entrenamiento. En casos como éste, las partes interesadas (clientes y administrador) concuerdan un correcto desempeño del sistema con alta seguridad.

Las diferencias en las condiciones existentes entre una prueba de verificación y aquellas de las sesiones de entrenamiento, influyen negativamente en la precisión y exactitud del sistema. Tales diferencias se denominan *mismatch* entre entrenamiento y prueba, y se relacionan con el ruido de fondo acústico, ruido en el canal de transmisión, líneas telefónicas, entre otras (Lei & Hansen, 2011).

Si las sesiones de entrenamiento y de verificación se realizan, por ejemplo, bajo diferentes condiciones acústicas (ruido ambiental), o de micrófonos, o bien, por canales de transmisión variables en el tiempo, va a existir una distorsión (por ejemplo, (Garretón, Becerra Yoma & Torres, 2010; Becerra Yoma *et al.*, 2013a) que se asocia a esas señales y, por lo tanto, ellas serán diferentes. En aplicaciones prácticas de un sistema automático de verificación de locutor, estas diferencias van a existir y por consiguiente, se hace necesario incorporar métodos que las atenúen (Jin, 2007).

1.2. Motivación

Desde un punto de vista auditivo, la comunicación verbal cotidiana entre los seres humanos se desarrolla en ambientes no silenciosos y en extremo variables acústicamente. El ruido de fondo y las características del canal acústico de transmisión está presente casi permanentemente, lo que degrada la señal acústica e interfiere con la transcripción neuronal hacia las vías auditivas superiores (Moore, 2003b; Parbery-Clark, Anderson & Kraus, 2013). El rol del sistema auditivo es informar al oyente sobre el ambiente acústico y sobre la presencia de señales de comunicación muy específicas (Syka, 2002). Este es un rol dominante en el ser humano dado que un aspecto que lo distingue de las demás especies es su habilidad para aprender y usar el lenguage (Tremblay *et al.*, 1997; Hauser, Chomsky & Fitch, 2002).

El sistema auditivo cumple extraordinariamente bien la función comunicativa. La audición se adapta a condiciones de información incompleta, a la degradación de la señal acústica, incluso a múltiples señales que compiten por su atención (Barbour, 2011). Aquellos sistemas, ya sean organismos biológicos, o bien, complejos sistemas de ingeniería, que son capaces de mantener sus funciones extrayendo continuamente la información relevante a pesar de perturbaciones externas e internas, se les llama robustos (von Bertalanffy, 1976; Carlson & Doyle,

2002; Kitano, 2004). Esta robustez es la propiedad fundamental del sistema auditivo, por lo que es altamente deseable de aplicar en sistemas de ingeniería, que utilizan procesamiento de señales de voz, por ejemplo, en áreas como robótica, reconocimiento de voz/locutor, entre otras (Jeon & Juang, 2007).

Para el ser humano reconocer rápidamente y sin esfuerzo a un individuo a partir de los sonidos de la voz, incluso sin recibir la información del rostro, le resulta extremadamente fácil lo cual se contrapone a los procesos cerebrales asombrosamente complejos que sirven de fundamento a este acto de reconocimiento (Atal, 1976; Larson, Billimoria & Sen, 2009). El reconocimiento de patrones, entendido como la tarea de asignar un evento a una categoría previamente especificada (Duda, Hart & Stork, 2001), ha sido durante años un objetivo de investigación en sistemas muy diversos de ingeniería (Lebedev & Nicolelis, 2006; Brumberg *et al.*, 2010).

El interés de investigar sobre algoritmos para reconocimiento de patrones acústicos y sistemas inspirados biológicamente, para reconocer en forma artificial señales de voz, se debe a la necesidad de poder desarrollar una interacción más natural con las máquinas y que ésta sea lo más parecida a las interacciones humanas (Cohen & Oviatt, 1995; Stern & Morgan, 2012a). Sin embargo, la tarea de reconocer artificialmente voces, o identificar individuos a partir de su voz, es un verdadero desafío y existen muchos problemas todavía que están sin resolver y posibilidades que explorar para que estos sistemas puedan funcionar a cabalidad.

En las útimas décadas a nivel mundial, las tecnologías de voz se han desarrollado intensivamente, primero a nivel experimental (condiciones de laboratorio) y posteriormente a nivel de mercado (aplicaciones reales), generalizándose en sistemas de reconocimiento de voz/locutor que han tenido impacto, en áreas tan diversas, como entretenimiento, telecomunicaciones, automóviles, ayuda a personas con discapacidad auditiva, medicina, educación, entre otros (Moore & Cutler, 2001). Sin embargo, un reconocedor de voz/locutor, puede ser robusto en un ambiente pero inapropiado para otro. La razón es que el desempeño del sistema de reconocimiento, que asume *a priori* un ambiente tranquilo (libre de ruido), se degrada con rapidez en presencia de fuentes de ruido y distorsión (Hansen, 1996; Becerra Yoma *et al.*, 2013a).

Así, para los investigadores en tecnologías de lenguaje hablado, es un desafío lograr que estos sistemas mantengan su funcionalidad y desempeño (sistemas robustos) en presencia de perturbaciones extremas, por ejemplo, en presencia de altos niveles de ruido acústico en el ambiente, perturbaciones que surgen imprevistamente, distorsiones generadas por el canal de transmisión, distintas calidades de micrófonos, variaciones de distancia entre micrófono y locutor, bloqueos entre locutor y micrófono, o bien, distorsiones propias de un canal telefónico. Cada una de estas perturbaciones, genera cambios en las características propias de la señal

de voz, cambios que degradan el funcionamiento de los sistemas de verificación de locutor (Pearce & Hirsch, 2000).

Frente a este escenario real, el desafío es desarrollar estrategias que identiquen estos contratiempos y permitan al sistema adaptarse robustamente a los cambios en los ambientes de operación, conservando su funcionamiento y desempeño, sin depender de las características imprevistas que encuentra en los ambientes en que debe funcionar.

Muchos métodos se han propuesto en la literatura para reducir o eliminar estos factores que afectan a los sistemas de reconocimiento de patrones acústicos. Sin embargo, la investigación todavía está lejos de hallar la solución definitiva a estos problemas. La importancia comercial final es que las tecnologías de procesamiento de voz, alcancen confiablemente un área del mercado y se adapten como ocurre en los sistemas biológicos, superando fragilidades y funcionando robustamente ante cualquier condición adversa o inesperada (Kitano, 2004).

Varios investigadores han formulado métodos de extracción de características basados en los mecanismos del sistema auditivo, para aplicarlos a sistemas de reconocimiento de locutor/voz (Hermansky, Cohen & Stern, 2013). Estos métodos, inspirados biológicamente, intentan reducir este desajuste o *mismatch*, compensando variabilidades inesperadas, entre las condiciones de grabación de las señales de voz para entrenamiento (ambiente tranquilo libre de ruido), y aquellas durante una prueba o aplicación real (Becerra Yoma *et al.*, 2008; Garretón & Becerra Yoma, 2012).

1.3. Definición de la propuesta de investigación

1.3.1. Descripción de la propuesta

La presente tesis frente al problema de *mismatch* en sistemas de verificación de locutor (SV), propone métodos que utilizan procesamiento de patrones acústicos basado en modelos del sistema auditivo periférico, los que puedan otorgar robustez al funcionamiento del sistema de SV, en ambientes adversos, ya sea por ruido aditivo, o bien, por variabilidades en el tiempo del canal de transmisión (canales con respuestas de frecuencia variables en el tiempo). Los métodos propuestos son evaluados en una plataforma de reconocimiento de patrones acústicos: en un sistema de verificación de locutor, texto-independiente (TI-SV).
0

En particular, se describe la elaboración de una función objetivo que optimiza la etapa de no linealidad sigmoidal tasa-nivel que forma parte de los modelos del sistema auditivo periférico (Pickles, 2008). Por otro lado, para abordar la robustez de patrones acústicos frente a *mismatch* por variabilidad de canal acústico de transmisión, se propone un método motivado

por el modelo de sincronismo del sistema auditivo periférico y que se inspira en el Detector de Sincronía Generalizado (GSD) de Seneff (Seneff, 1988).

El desafío de desarrollar aplicaciones de procesamiento de patrones acústicos basado en modelos del sistema auditivo periférico, robustas a pesar de condiciones de *mismatch*, y aplicadas en una tarea de verificación de locutor, es de interés actual y los métodos propuestos son originales y no han sido publicadas previamente en la literatura. Esto queda demostrado por las publicaciones en revistas ISI logradas como resultado del trabajo de esta tesis (ver anexo A)[1]

1.3.2. Hipótesis

Las hipótesis que justifican los objetivos y respaldan la investigación y metodología propuesta son las siguientes:

H1) El uso de modelos del sistema auditivo periférico en tecnologías de verificación de locutor no ha sido suficientemente explorado. Las propiedades del sistema auditivo demuestran un alto potencial y aplicabilidad.

H2) El principio de no linealidad sigmoidal tasa-nivel puede contribuir a la robustez de un sistema de verificación de locutor bajo condiciones de ruido aditivo.

H3) El principio de sincronismo o *phase-locking* en el nervio auditivo puede dar robustez a un sistema de verificación de locutor bajo condiciones de variación en el canal de transmisión.

De acuerdo a la literatura, en reconocimiento automático de voz han sido aplicados diversos métodos basados en modelos del sistema auditivo periférico con éxitos relativos pero que validan la importancia que tiene su aplicación. Sin embargo, en verificación de locutor la utilización de estos modelos no ha sido suficientemente bien explorada abriendo la posibilidad para desarrollar su potencialidad y aplicabilidad.

La optimización de la no linealidad sigmoidal hace uso de un conjunto de criterios que se definen sobre la base de atributos físicos de la señal de entrada, sin utilizar discriminación de fonemas (Chiu, Raj & Stern, 2012). La optimización está motivada por evidencia fisiológica auditiva de adaptación de la función tasa-nivel a cambios en los niveles sonoros, en diferentes

[1] **Poblete, V., Becerra Yoma, N., Stern, R. M.**, 2014. Optimization of the parameters characterizing sigmoidal rate-level functions based on acoustic features. Speech Communication (Elsevier). 56, 19-34.
Poblete, V., Espic, F., King, S., Stern, R. M., Huenupán, F., Becerra Yoma, N., 2014. A perceptually-motivated low-complexity instantaneous channel normalization technique applied to speaker verification. Submitted to Computer Speech and Language (Elsevier). February 2014.

especies de mamíferos (Wen *et al.*, 2009).

Estas hipótesis se sustentan en propiedades de la actividad neuronal en el nervio auditivo en diferentes especies de mamíferos. Particularmente, tales propiedades demuestran que en las fibras nerviosas el contenido espectral de las vocales (caracterizadas por sus formantes) se representa por la tasa de descarga de las fibras, o bien, por el sincronismo temporal de las descargas neuronales (Sachs & Young, 1979; Young & Sachs, 1979). Young & Sachs (1979) demostraron que una medida de sincronismo como función de la frecuencia, llega a ser mucho más robusta a cambios de niveles sonoros que la correspondiente medida de tasa de descarga. Estos resultados sugieren que la información síncrona es robusta a cambios en el nivel sonoro y potencialmente, más robusta a otros tipos de variabilidad de la señal, o degradación, que la propia medida de tasa de descarga. Además, se sustenta en la evidencia del modelo de Seneff de Detector de Sincronía Generalizada (GSD) (Seneff, 1984, 1988). Sin embargo, el GSD presenta un problema potencial cual es producir *peaks* espurios en armónicos de la frecuencia detectada, lo que puede explicar por qué diversos intentos previos para usarlo directamente en aplicaciones de reconocimiemto de voz, mostraron sólo mejoras limitadas en precisión (Jankowski & Lippmann, 1992; Ohshima & Stern, 1994; Jankowski, Vo & Lippmann, 1995; Ali, Van Der Spiegel & Mueller, 2000, 2002; Kim, Chiu & Stern, 2006; Stern & Morgan, 2012a). A pesar de estas observaciones, el comportamiento del GSD aun tiene propiedades deseables y potenciales de aplicar en verificación de locutor.

1.3.3. Objetivos generales y específicos

Los principales objetivos desarrollados en esta Tesis son:

Objetivo general:

- Mejorar la robustez de *features* de voz bajo condiciones de *mismatch* por ruido aditivo o por variabilidad en el canal acústico de transmisión, en una tarea de reconocimiento de patrones acústicos basado en modelos del sistema auditivo periférico.

Objetivos específicos:

- Generar una función objetivo que permita optimizar la etapa de nolinealidad sigmoidal tasa-nivel presente en el sistema auditivo periférico

- Definir criterios de optimización en base a atributos físicos de la señal de entrada sin considerar discriminación de fonemas.

- Generar una medida de sincronismo de una señal basada en el procesamiento auditivo periférico que tome en consideración los grados de concentración y dispersión de la energía en torno a una frecuencia específica.

- Desarrollar métodos de extracción de características basadas en el sistema auditivo periférico, robustos a *mismatch* por ruido aditivo o por variabilidad de canal.

- Mejorar la robustez de un sistema de verificación de locutor de texto-independiente (TI-SV) al evaluar su desempeño en condiciones de *mismatch* por ruido aditivo y por variabilidad en el canal acústico de transmisión, utilizando por una parte, un método de la función sigmoidal tasa-nivel óptima y por otro, un método inspirado en la propiedad de sincronismo auditivo, y comparar sus desempeños con el método clásico de extracción de coeficientes cepstrales en escala *Mel* (sistema *MFCC*).

1.3.4. Metodología

En general, para evaluar los desempeños, tanto de la no linealidad sigmoidal óptima, así como también del conjunto propuesto de *features* denominados Coeficientes Cepstrales Localmente-Normalizados (LNCC), se utilizará una plataforma que comprende una tarea de reconocimiento de patrones acústicos y que consiste en un sistema biométrico de verificación de locutor de texto-independiente, con la tasa de igual error (EER), utilizada como la principal figura de mérito. Los resultados ha describir se obtendrán al usar la base de datos YOHO (Campbell & Higgins, 1994).

Los parámetros óptimos de la no linealidad sigmoidal se estimarán empleando un subconjunto de elocuciones (base de datos de desarrollo), extraído de YOHO. Las elocuciones utilizadas para entrenar la función sigmoidal no se incluyen en la información de prueba durante el experimento principal de verificación de locutor. De la base de datos AURORA se selecionarán tres tipos de ruidos (Hirsch and Pearce, 2000). Estos ruidos se sumaron artificialmente a la base YOHO para generar versiones ruidosas de las elocuciones en varios SNRs. Durante todos los experimentos de verificaciÓn, el sistema se entrena con voz limpia.

Aunque siempre se espera obtener mejores desempeños cuando los parámetros óptimos de la no linealidad sigmoidal se estimen en condiciones ambientales que se igualen (*matched*) al ambiente de prueba, se intentará cuantificar la magnitud del ambiente que se espera, al experimentar algunas de las condiciones con los SNRs para estimación de parámetros, igualados a los SNRs utilizados en los propios experimentos de verificación de locutor. Se compararán los EERs para verificación de locutor cuando las sigmoides son entrenadas en el SNR de prueba con los correspondientes EERs obtenidos cuando los parámetros son siempre estimados a partir de señales a un SNR que resulte ser el mejor SNR único de entrenamiento.

Se compararán los resultados aquí propuestos de estimación de la no linealidad sigmoidal tasa-nivel basado en las características de la forma de onda, con aquellos que describen Chiu, Raj & Stern (2012), para reconocimiento de voz, donde la no linealidad sigmoidal es formada empleando discriminación basada en clases de fonemas.

Asimismo, se realizarán experimentos de verificación de locutor para evaluar la capacidad de los *features* propuestos para normalizar durante canales acústicos variables, sobre voz degradada por varios canales. Los canales serán simulados imponeniendo *tilt* espectral que imita el efecto de micrófonos bloqueados o fuera del eje, al igual que características de *tilt* espectral que varíen dentro de una elocución. Por razones de control experimental y de replicabilidad, se simularán las respuestas del canal. En todos los experimentos, el sistema se entrenará utilizando solamente voz limpia. La voz de prueba se degradará con respecto a la información de entrenamiento al imponer *tilt* espectral estático y variable en el tiempo.

1.4. Estructura de la tesis

Esta tesis ha sido estructurada de tal forma de guiar al lector de modo gradual hacia el problema que se aborda, comenzando por conceptos, definiciones y antecedentes del contexto general, necesarios para comprender en detalle los métodos propuestos. De esta manera, la tesis comienza con una introducción sobre los sistema de verificación de locutor, sobre el sistema auditivo y el procesamiento de la voz, y también se introduce a las aplicaciones y métodos de extracción de característica inspirados en evidencia auditiva fisiológica, los que se han abordado en la literatura especializada. A continuación, se presenta una descripción en detalle de la propuesta de esta tesis que considera también resultados de experimentos y comparaciones con otras técnicas en el estado del arte disponibles en las referencias bibliográficas. Esta tesis ha sido dividida en cinco capítulos describiéndose brevemente a continuación la estructura de cada uno de ellos.

El capítulo 2 presenta los aspectos fundamentales sobre la robustez en sistemas de verificación de locutor, los procesos auditivos básicos involucrados en el análisis de sonidos de la voz, los modelos auditivos y las tecnologías de procesamiento de voz, e introduce al problema de *mismatch* que surge debido a las variaciones entre las condiciones de entrenamiento y de prueba en aplicaciones reales. En este capítulo se muestran además los métodos en el estado del arte inspirados en evidencia experimental de la fisiología auditiva y los algoritmos que los conectan al procesamiento de señales de voz y a los sistemas de verificación de locutor, así como también se presentan las metodologías de evaluación. También, en este capítulo se hace una mención especial al sistema TI-SV que se ha desarrollado en el Laboratorio de Procesamiento y Transmisión de Voz (LPTV) de la Universidad de Chile.

El capítulo 3 describe el desarrollo de una técnica novedosa para optimizar la función tasa-nivel, sigmoidal óptima, que es una parte de muchos modelos del sistema auditivo periférico. En particular, la optimización hace uso de un conjunto de criterios definidos exclusivamente sobre la base de atributos físicos del sonido de entrada que están inspirados por evidencia

fisiológica. Los criterios desarrollados intentan discriminar entre una señal de voz degradada y ruido para preservar la máxima cantidad e información en la región lineal de la curva sigmoidal, y minimizar los efectos de distorsión en las regiones de saturación. El desempeño de la función sigmoidal óptima es validado por experimentos de verificación de locutor de texto independiente, con señales degradadas por ruido aditivo a diferentes razones de señal-a-ruido. Los resultados experimentales sugieren que el método presentado en combinación con normalización de varianza cepstral puede conducir a reducciones relativas importantes en *EER* cuando se compara con el uso de los coeficientes baseline *MFCC* para ciertas razones señal-a-ruido.

En el capítulo 4 se extiende la idea de aplicar propiedades del sistema auditivo periférico como las presentadas en el capítulo 3, al problema de verificación de locutor y *mismatch* en el canal acústico de transmisión, proponiéndose un nuevo conjunto de *features* que se denominan Coeficientes Cepstrales Localmente-Normalizados (LNCC). Estos nuevos coeficientes se basan en el Detector de Seneff de Sincronía Generalizada (GSD). En este contexto, el presente capítulo proporciona resultados de verificación de locutor que demuestran que los *features* LNCC compensan más el *tilt* espectral que los coeficientes convencionales MFCC. Por último, el capítulo 5 describe las principales conclusiones y análisis final de los métodos propuestos en esta tesis. Además, se describen las principales direcciones de trabajo futuro.

1.5. Contribuciones de la tesis

Esta tesis presenta dos novedosas estrategias inspiradas en el sistema auditivo periférico para obtener patrones acústicos robustos. En primer lugar, un método de optimización de la función sigmoidal tasa-nivel para modelamiento auditivo y, en segundo lugar, un método de representación espectral, basado en el detector *GSD* de sincronismo en procesamiento auditivo. Con respecto a la estrategia de optimización de la función sigmoidal tasa-nivel, se mencionan las siguientes contribuciones: un método de optimización de la función no lineal sigmoidal usada típicamente en modelamiento auditivo, que se basa en atributos físicos de la señal acústica y no en discriminación fonética; un esquema que intenta simultáneamente minimizar la potencia del ruido, minimizar la distorsión no lineal, maximizar la similitud entre la voz limpia y la voz degradada de entrada, y maximizar la varianza de la señal degradada por ruido después del proceso de la función sigmoidal; un método que usando la nolinealidad sigmoidal óptima puede conducir, en experimentos de verificación de locutor de texto independiente, a reducciones relativas promedio en *EER* significativas comparadas al procesamiento baseline con voz degradada por ruido aditivo y que demuestra la potencialidad de las nolinealidades que están presentes en el procesamiento auditivo periférico humano; un método donde las funciones sigmoidales óptimas tasa-nivel son estimadas separadamente para cada canal del banco de filtros; un método que es aplicable a cualquier tarea de procesamiento de voz ya que todo el análisis se lleva a cabo a nivel de la señal acústica. Como

resultado, se generó la publicación citada en Anexo A.

Por otra parte, en relación con el nuevo conjunto de *features* llamados LNCC, que se basan en el detector de sincronía de Seneff, GSD, se mencionan las contribuciones más importantes: se sugiere una nueva estrategia, motivada perceptualmente y extremadamente simple, pero no menos efectiva, para normalizar instantáneamente *features* de voz; los *features* propuestos LNCC demuestran su efectividad en una tarea de verificación de locutor a los largo de una variedad de condiciones de canal; los *features* LNCC además, no requieren calcular, ni almacenar, ningún promedio móvil de valores de *features*; los *features* LNCC proporcionan reducciones promedio relativas, durante *tilt* espectral variable y *tilt* espectral estático, superiores si se comparan con los *features* MFCC; los *features* propuestos LNCC demuestran ser una alternativa a MFCC y MFCC + CMN, en cualquier situación donde es difícil estimar la media cepstral confiablemente. Como resultado, se originó la siguiente publicación referida en Anexo A.

Capítulo 2

Robustez en sistemas de verificación de locutor

2.1. Introducción

La verificación de locutor (SV) es un método de autentificación biométrica de una persona basado solamente en su voz. Este método es muy atractivo ya que no requiere contacto directo (es decir, se puede llevar a cabo en forma remota) con la persona, evitando así, la sensación de invasividad propia de otros métodos biométricos, como reconocimiento de rostros, iris y huella digital (Majekodunmi & Idachaba, 2011). Otra ventaja de la verificación de locutor por sobre los otros métodos antes mencionados, es que ésta no requiere un hardware especializado para interfaz entre la persona y la máquina. Sólo se requiere de un micrófono el cual hoy se encuentra disponible en la mayor parte de los equipos de telefonía celular y en los computadores personales (Fazel & Chakrabartty, 2011; Kinnunen & Li, 2010).

La verificación de locutor hace posible el uso de la voz de una persona para control de accesos a servicios restringidos, por ejemplo, al dar comandos de voz a computadores, acceso vía teléfono a cuentas de banco, a servicios de bases de datos, compras o mensajes de voz, y acceso seguro a equipamiento (Wang *et al.*, 2011).

Aplicaciones en sistemas biométricos demuestran resultados exitosos en ambientes de bajo ruido de fondo. Desafortunadamente, la aplicación de estos sistemas es limitada cuando el ambiente acústico no es el mismo al utilizado en entrenamiento y durante las pruebas. Sería deseable tener un sistema que funcione bien sin considerar las condiciones de grabación natural, es decir, por ejemplo, en diferentes salas, distintos micrófonos (canal de comunicación) y con niveles variables de sonido (ruido de fondo acústico) (Nemala & Elhilali, 2011; Deshpande & Holambe, 2011). Una solución a este problema ha sido reducir el *mismatch* al entrenar o reentrenar el sistema de reconocimiento bajo condiciones ruidosas representativas del ambiente de aplicación (Kinnunen & Li, 2010; Paliwal, Wojcicki & Shannon, 2011).

2.2. Estructura de un sistema de verificación de locutor

Un sistema de SV debe proporcionar una base de datos de clientes registrados. Estos locutores se llaman "clientes". La base de datos consiste de modelos acústicos que representan las características de la voz de los clientes. Los modelos se generan en sesiones de entrenamiento en los cuales el cliente debe pronunciar varias frases.

En un sistema de SV, el extractor de características (Figura 2.1) transforma la señal de voz original (sin procesamiento) en un nuevo formato de señal, preservando las características específicas del locutor. La transformación resulta en una secuencia de vectores de características representativos de la señal de voz (Garretón, 2011).

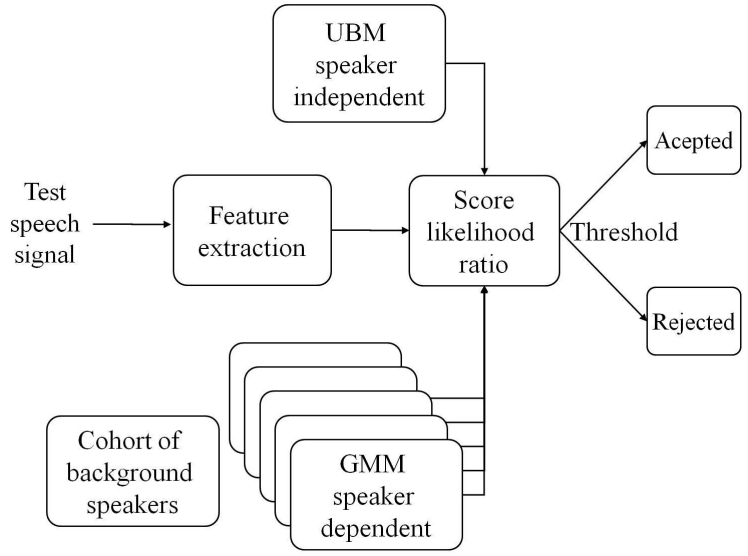

Figura 2.1: *Componentes generales de un sistema verificador de locutor.*

Desde un punto de vista de usuarios, la tarea de verificar a un locutor consiste de dos modos de operación:

1. Entrenamiento, llamado también Enrolamiento.

2. Prueba, o bien Verificación.

En el modo de entrenamiento, el usuario proporciona sus muestras de voz al sistema. Luego, la información de enrolamiento es utilizada para entrenar a un clasificador el cual debe ser representativo del locutor enrolado. Por lo general, la información de enrolamiento se puede reunir usando sistemas o equipos, tales como, teléfonos celulares o micrófonos, con diferentes ruidos de fondo (Lei & Hansen, 2011).

Por otro lado, en el modo de verificación, un usuario proporciona una voz de muestra demandando una identidad, y un clasificador es usado para determinar si la información de prueba fue pronunciada por el locutor enrolado demandado en la prueba. Si la información es capturada bajo condiciones que incluyen diferentes canales y ruidos, significa que el ambiente, o condiciones, (ruido de canal y ruido de fondo acústico), entre el enrolamiento y la información de verificación o *test*, son diferentes y reflejan *mismatch*.

En general, el objetivo de la tarea de verificación de locutor es ya sea, aceptar o rechazar, una identidad demandada por un locutor. La pronunciación de un usuario desconocido se compara con el modelo del locutor cuya identidad está siendo demandada (Tchorz & Kollmeier, 1999; Pradhan & Prasanna, 2011).

Si el *score* S de comparación, está por arriba de un cierto umbral T, se verifica la identidad demandada. Un valor de umbral T, muy alto, hace difícil que los impostores sean aceptados por el sistema, pero a riesgo de que se rechacen clientes. Por el contrario, un valor de umbral T muy bajo, asegura que el cliente es aceptado sistemáticamente pero, a riesgo de aceptar impostores. Esto quiere decir que una vez que se ha calculado el *score* S, de una pronunciación, este *score* se compara con el umbral T. En seguida, el sistema debe tomar una determinación de acuerdo a la siguiente regla: $S > T$, aceptar al locutor (la persona es quien demandaba ser), y por otro lado, $S < T$, rechazar al locutor (es decir, la persona es un impostor).

En el estado del arte las tecnologías de verificación de locutor pueden alcanzar una tasa de error entre 0,1 % a 5 %, utilizando señales en condiciones controladas a niveles bajos de ruido. Durante los modos de operación de entrenamiento y prueba, los sistemas de SV emplean elocuciones entre 10 a 30 segundos y entre 2 a 10 segundos de duración, respectivamente (por ejemplo, Becerra Yoma & Villar (2002)). Dadas estas tasas de error, un sistema de SV se presenta como una alternativa interesante si se usa en combinación con, por ejemplo, números de identificación personal.

2.2.1. Medida del Desempeño de un Sistema de Verificación de Locutor

Cuando un sistema de verificación llega a la etapa de decisión, éste muestra dos opciones: aceptar o rechazar a un usuario de prueba. Ambas opciones resultarán en cuatro casos posibles: dos opciones correctas y dos opciones equivocadas. Los casos correctos son:

(a) aceptar a un locutor verdadero (cliente), y

(b) rechazar a un locutor falso (impostor).

En tanto que los casos equivocados corresponden a:

(c) aceptar a un impostor, y

(d) rechazar a un cliente.

En los casos (a) y (b) el sistema toma decisiones correctas, mientras que en los otros dos casos, (c) y (d), el sistema toma decisiones incorrectas. Estos errores se llaman Falsa Aceptación (FA) y Falso Rechazo (FR), respectivamente.

Es posible establecer un umbral de decisión que minimice ambos tipos de errores. Cuando este umbral óptimo se encuentra, existirá un equilibrio entre estos errores. Para encontrar el punto óptimo se definen dos curvas: la curva de falso rechazo y la curva de falsa aceptación. El valor en el cual se igualan estos niveles de error se llama *Equal Error Rate* (EER [%]). EER es utilizado generalmente para medir el desempeño de un sistema de verificación de locutor así como también, de otros sistemas biométricos. El desempeño del sistema se puede representar gráficamente al generar curvas de FA y FR, como una función del umbral de decisión (Furui, 1997) (Figura 2.2).

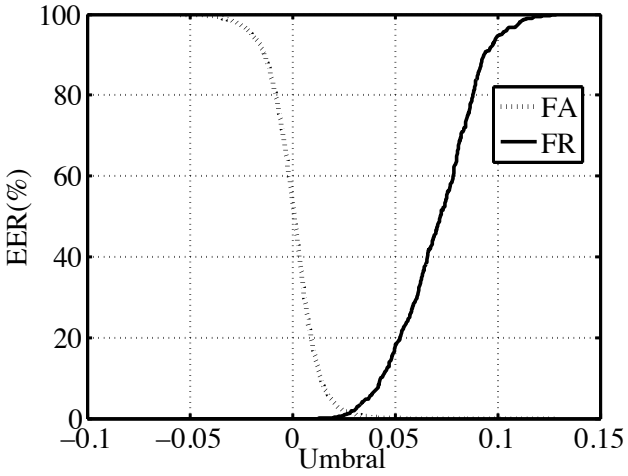

Figura 2.2: *Curvas de falsa aceptación (FA) y falso rechazo (FR).*

Los sistemas de verificación de locutor que operen en aplicaciones prácticas pueden considerar umbrales de decisión obtenidos a posteriori en sesiones de entrenamiento y luego utilizar un límite que según sea el nivel de seguridad del sistema, permita la falsa aceptación de una fracción de impostores y, a su vez, un falso rechazo de un porcentaje de clientes. Este límite de decisión se puede encontrar graficando la curva de operación del receptor (ROC) que corresponde al gráfico de la curva de falsa aceptación frente a la curva de falso rechazo. El valor del área bajo la curva ROC sirve como indicador de la habilidad discriminativa del sistema bajo el rango completo de valores de umbral de decisión en el que éste es probado. Mientras mejor sea el desempeño presentado por el sistema, menor será el área bajo la curva ROC (Campbell, 1997).

2.2.2. Parametrización acústica de una señal de voz

Para caracterizar una señal acústica de voz, es necesario tener en consideración dos importantes puntos asociados a esta tarea (Garretón, 2011):

1. La señal de voz es aproximada como semi-estacionaria (estacionaria a nivel segmental),
2. Las variaciones temporales entre señales contienen la misma información fonética acústica.

En una señal de voz, la variabilidad temporal se puede deber a aspectos relacionados con:

1. Locutor;
2. Ambiente acústico de fondo (por ejemplo, ruidos acústicos, música), y canal acústico de transmisión;
3. La fuente o medios de grabación de la señal de voz (entre los que se pueden mencionar el micrófono de teléfono celular o micrófono directo).

La variabilidad del locutor es llamada también, variabilidad intra-locutor. Esta variabilidad describe las variaciones de la información fonético acústica, entre pronunciaciones de la misma persona, las cuales son extraídas a partir de una señal de voz (Cooke, Hershey & Rennie, 2010). Asimismo, se desprende el concepto de variabilidad entre locutores. La variabilidad entre locutores se relaciona con las variaciones entre pronunciaciones que pertencen a un grupo grande (llamado "cohorte") de locutores o universo de locutores (Bimbot *et al.*, 2004).

La variabilidad tanto en el ambiente como en el canal acústico de transmisión, introducen variaciones no deseadas durante un proceso de extracción de paramétros de una señal de voz. Esta variabilidad representa la cantidad de ruido y la propia variabilidad del canal a través del tiempo (Fazel & Chakrabartty, 2011). Asimismo, la variabilidad de la fuente corresponde a aquella debido al medio de grabación o canal de comunicación (por ejemplo, telefónico). Esta variabilidad lleva a uno de los aspectos de distorsión más relevantes en pronunciaciones con información fonético acústica desde una misma persona.

El proceso completo de extracción de características acústicas a partir de una señal de voz, se basa en calcular los coeficientes cepstrales en escala Mel (Furui, 1981). El análisis de una señal de voz en el domino cepstral permite robustecer las componentes asociadas a los formantes del tracto vocal, incluso bajo condiciones de señales ruidosas (Damper & Higgins, 2003).

Previo a extraer los parámetros o características de la voz (vectores de *features* o simplemente *features*), se prepara previamente la señal. La conversión análoga-digital es el primer

paso. Luego, se aplica una detección de inicio y fin (*end-point detection*), la que elimina de la señal los períodos de silencio antes del inicio del primer pulso de voz y después del último (Lamel *et al.*, 1981; Savoji, 1989). Seguidamente, se caracteriza la señal mediante secuencias estacionarias, o casi estacionarias. Este paso se realiza a través de un proceso de segmentación, en el que cada uno de los segmentos se denomina *frame*. Pare ello, se emplea una ventana (por ejemplo, Hamming) (Picone, 1993), y a continuación se lleva a cabo un análisis espectral por cada frame, en el cual la señal se procesa por la transformada discreta de Fourier (DFT). En base al comportamiento de la cóclea y su función biológica de "filtrar" las frecuencias del sonido de entrada al sistema auditivo periférico, se emplea un banco de filtros pasabanda, el que tiene filtros espaciados sobre una escala de frecuencia no lineal (Allen, 1980; Robles & Ruggero, 2001).

Dado que la respuesta en frecuencia del sistema auditivo humano es no lineal, se utiliza una escala en la cual se concentran frecuencias producto del proceso de filtrado, simulando la capacidad de discriminabilidad del sistema auditivo. Una de las escalas usada habitualmente para estos propósitos es la escala de frecuencia *Mel* (derivada de la palabra *Mel*ody). Esta escala Mel es motivada percepción auditiva (Skowronski & Harris, 2004):

$$Mel(f) = 2595 \cdot log_{10}\left(1 + \frac{f}{700}\right) \tag{2.1}$$

El banco de filtros consiste de un conjunto de funciones triangulares simétricas, de ganancia unitaria para la frecuencia central, con solapamiento de 50 % y un ancho de banda constante en escala Mel. Se calcula el logaritmo de la energía en cada filtro y luego, usando la transformada discreta del coseno (DCT) (Ahmed, Natarajan & Rao, 1974) a la salida de cada filtro, se obtienen los coeficientes cepstrales en escala Mel (por brevedad, MFCCs) (Davis & Mermelstein, 1980). En procesamiento de voz, se calcula un vector de parámetros o vector de *features* espectrales, para cada *frame* que se analiza, es decir, una elocución se caracteriza como una secuencia de vectores de observación en el dominio MFCC, $O = [O_1, O_2, \ldots, O_t, \ldots, O_T]$, donde T es el número total de frames y O_t es el vector de observación para el frame t. Este método de parametrización de señales de voz, basado en el uso de coeficientes cepstrales en escala Mel, es ampliamente utilizado en verificación de locutor (Shannon & Paliwal, 2003; Furui, 2005; Bimbot *et al.*, 2004). El analizar una señal en el dominio cepstral realza las componentes asociadas a los formantes del trayecto vocal, incluso en señales degradadas por la presencia de ruido (Forsyth, 1995).

No obstante, la tarea de producir parámetros robustos frente a diversas condiciones muy adversas para el sistema, es todavía un problema abierto y de gran interés. Otros métodos de parametrización robusta, basados en el sistema auditivo, se describen con mayor detalle en el capítulo 3.

Para capturar la naturaleza dinámica de la voz, es común aumentar la dimensión del vector de características con los paramétros "deltas" (Furui, 1986). Los coeficientes cepstrales representan información importante acerca de las características específicas de un locutor, y los coeficientes cepstrales delta, aportan al sistema de verificación de locutor una robustez mayor cuando existen distorsiones causadas por variabilidades del canal de comunicación (Campbell, 1997). En general, en procesamiento de voz, la dimensión de un vector de parámetros acústicos o vector de *features*, depende de la duración en frames que tenga la señal. Frecuentemente, cada vector de parámetros se compone de 33 coeficientes cepstrales para el caso del sistema MFCC.

2.2.3. Verificación de locutor de texto-independiente, con modelo GMM-UBM

En sistemas de SV, el modelo de *mezcla de Gaussianas-modelo de referencia universal* (GMM-UBM), es un método ampliamente utilizado para verificación de locutor de texto-independiente (Reynolds, Quatieri & Dunn, 2000). En este método, la elocución de voz de un locutor incógnita, se modela como un GMM y los impostores se modelan como un UBM.

El GMM-UBM se comprende como un detector de razón de verosimilitud: el UBM se entrena para representar la distribución de caracteríticas independiente del locutor (modelo *speaker independent*, SI); mientras que el GMM se adapta a partir del UBM para ilustrar las características individuales propias de un locutor (modelo *speaker dependent*, SD) (Bimbot *et al.*, 2004).

En SV, dada una elocución Y y un locutor hipotético S, la tarea es determinar si Y fue hablada por S. Esta tarea a menudo se denomina como detección. Un supuesto en la tarea es que Y contiene voz de un solo locutor. La tarea se establece como una prueba de hipótesis entre:

H0: Y es del locutor hipotético S.
H1: Y *no* es del locutor hipotético S.

La prueba óptima para decidir entre estas dos hipótesis es una prueba de razón de verosimilitud (LR), dada por

$$\frac{p(Y|H0)}{p(Y|H1)} \begin{cases} > \theta, & \text{acepta } H0 \\ < \theta, & \text{acepta } H1, \end{cases} \tag{2.2}$$

donde $p(Y|H0)$ es la función de densidad de probabilidad para la hipótesis $H0$, evaluada para la elocución observada Y, también denominada como la verosimilitud de la hipótesis $H0$ dada la elocución. Asimismo, la función de verosimilitud para $H1$ es $p(Y|H1)$. El umbral

para aceptar o rechazar $H0$ se representa por θ.

Si dada una elocución, se considera que el *frame* en el instante t se ilustra por un vector de parámetros espectrales observados (esto es, un vector de *características* observadas) $O_t = [O_{t,1}, O_{t,2}, \ldots, O_{t,n}, \ldots, O_{t,N}]$, donde N es el número total de parámetros y $O_{t,n}$ es el n-ésimo parámetro en el *frame* t, entonces una elocución se representa por una secuencia de vectores O:

$$O = [O_1, O_2, \ldots, O_T] \tag{2.3}$$

donde T es la duración en *frames* de la elocución. Estos vectores de características se usan entonces para calcular las verosimilitudes de $H0$ y $H1$. Matemáticamente, un modelo, expresado como λ_{SD}, representa a $H0$, el que caracteriza, en el espacio de características O, al locutor hipotético S, o también llamada hipótesis del cliente.

En SV, se asume que para $H0$ una distribución Gaussiana representa de mejor forma la distribución de los vectores de características, de modo que λ_{SD} contiene los parámetros de la distribución Gaussiana: el vector de media y la matriz de covarianza. Por otro lado, el modelo λ_{SI}, ilustra la hipótesis alternativa, $H1$, o también referida como hipótesis del impostor. La razón de las verosimilitudes se expresa en la forma $L(O) = p(O|\lambda_{SD})/p(O|\lambda_{SI})$. Frecuentemente, se utiliza el logarítmo de $L(O)$,

$$LL(O) = \log p(O|\lambda_{SD}) - \log p(O|\lambda_{SI}) \tag{2.4}$$

donde el término $LL(O)$ se denomina verosimilitud logarítmica (*loglikelihood*) normalizada. La probabilidad que la secuencia de vectores de características O, corresponda al modelo del locutor $p(O|\lambda_{SD})$, se calcula estimando la verosimilitud de O en el modelo SD. Por otra parte, la probabilidad $p(O|\lambda_{SI})$, denominado término normalizador, corresponde a la verosimilitud calculada con respecto a un modelo general de impostores o modelo SI (Campbell, 1997). Este modelo, idealmente se entrena con elocuciones que pertenecen a una gran cantidad de usuarios que no se encuentran registrados en el sistema. Para independizar el cálculo de $LL(O)$ de la duración de las elocuciones, se divide el resultado por el número de frames total de la elocución, T:

$$LL(O)' = \frac{LL(O)}{T} \tag{2.5}$$

El uso de normalización de la verosimilitud, demuestra una reducción significativa del error provocado por ejemplo, por la presencia de ruido convolucional al usar distintos tipos de micrófonos. Existen variadas formas adicionales de aplicar una normalización a la verosimilitud o *score* de una elocución de verificación. Cada una se diseña con algún objetivo particular (por ejemplo, eliminar la dependencia al locutor, compensación de *mismatch* de

canal, etc.).

Aunque el modelo para $H0$, λ_{SD}, se define bien y se puede estimar utilizando elocuciones de entrenamiento del locutor S, el modelo para λ_{SI} es menos definido dado que éste potencialmente debe representar el espacio completo de posibles alternativas para el locutor hipotético. Un método para modelar la hipótesis del impostor, es el modelo de referencia universal (UBM) (Reynolds, 1997). Este método reune una gran cantidad de elocuciones, de varios locutores, representativas de la población de locutores esperados durante la verificación y entrena un modelo único, λ_{ubm}. La principal ventaja de este método, es que un modelo SI único se entrena una vez y luego se usa para todos los locutores hipotéticos (Bimbot *et al.*, 2004).

2.2.3.1. Modelos de mezclas de Gaussianas

Un paso importante en la implementación de un detector de la razón de las verosimilitudes es la selección de la función $p(O|\lambda)$. En TI-SV, donde no existe conocimiento a priori de lo que dirá el locutor, se utilizan ampliamente los modelos de mezclas de Gaussianas GMMs.

Para un vector de características, $O(t)$, de dimensión N, la densidad de mezcla usada para la función de verosimilitud, se define de la forma (Reynolds, Quatieri & Dunn, 2000; Bimbot *et al.*, 2004)

$$p(O(t)|\lambda) \;=\; \sum_{m=1}^{M} \omega_m \cdot p_m(O(t)). \tag{2.6}$$

La densidad es una combinación lineal de M densidades Gaussianas unimodales $p_m(O(t))$, cada una parametrizada por un vector de media de $N \times 1$, μ_m, y por una matriz de covarianza de $D \times D$, Σ_m:

$$p_m(O(t)) \;=\; \frac{1}{\sqrt{(2\pi)^N |\Sigma_m|}} \cdot e^{-(1/2)(O(t)-\mu_m)^T \Sigma_m^{-1} (O(t)-\mu_m)} \tag{2.7}$$

Los pesos de las mezclas ω_m satisfacen la condición $\sum_{m=1}^{M} \omega_m = 1$. En conjunto, los parámetros del modelo se denotan como $\lambda = (\omega_m, \mu_m, \Sigma_m)$, $m = (1, \cdots, M)$. Aunque el modelo general establece matrices completas, es decir, con todos sus elementos, por lo general, se usan solamente matrices diagonales. Esto se hace dado que GMMs de matrices diagonales son más eficientes computacionalmente, que los GMMs de matrices completas para entrenamiento ya que no se requiere inversión de una matriz de $N \times N$. Además, empíricamente se ha observado que GMMs de matrices diagonales superan en desempeño a los GMMs de matrices completas.

Si se tiene un conjunto de vectores de entrenamiento, los parámetros del modelo de verosimilitud máxima se estiman utilizando el algoritmo de máxima esperanza (EM) (Dempster, Laird & Rubin, 1977). Este algoritmo perfecciona los parámetros del GMM, aumentando la verosimilitud del modelo estimado para los vectores de características observadas, esto es, por iteraciones k y $k+1$, $p(O|\lambda^{k+1}) \geq p(O|\lambda^{k})$ (Reynolds, 1995; Duda, Hart & Stork, 2001).

La ventaja de usar GMMs como la función de verosimilitud, para tareas de TI-SV, radica en su bajo costo computacional, es insensible a aspectos temporales de la voz, es decir, modela solamente las distribuciones de las observaciones acústicas y no las secuencias de palabras pronunciadas, y también, los GMMs no consideran información de clases de fonemas (Huenupán, 2010).

2.2.3.2. Sistema GMM adaptado

De acuerdo a lo observado anteriormente, el método para modelar la hipótesis de impostor, se basa en un modelo único de referencia, independiente de los locutores, UBM, para representar $p(O|\lambda_{SI})$. Utilizando un GMM como la función de verosimilitud, el modelo único de referencia es un GMM grande, entrenado para representar las distribuciones de características que son independientes de los locutores. Específicamente, se seleccionan elocuciones que reflejen la voz de la hipótesis de impostor encontrada durante un experimento de verificación de locutor. Por lo general, el orden del modelo (o número de componentes Gaussianas en el GMM), para el modelo de referencia universal varía desde 64 a 2048, dependiendo de los *features* escogidos y su dimensionalidad, del número de vectores de entrenamiento y el modelo de clustering seleccionado, por ejemplo, GMM (Bimbot *et al.*, 2004; Kinnunen & Li, 2010; Kinnunen *et al.*, 2011). Modelos de mezclas de orden bajo a menudo se emplean cuando se trabaja con voz restringida o elocuciones de corta duración (por ejemplo, dígitos), en tanto que, mezclas de 2048 se utilizan cuando de trabaja con elocuciones no restringidas o de larga duración (tal como, conversaciones).

Además, para entrenar un modelo UBM no existe una medida objetiva que determine el número correcto de locutores o la cantidad de elocuciones que se deben usar. Experimentalmente, existe evidencia en estudios de SV que no hay pérdida en el desempeño del sistema si se utiliza un UBM entrenado con una hora de elocuciones a uno que se entrene con seis horas de voz (Reynolds, Quatieri & Dunn, 2000; Bimbot *et al.*, 2004).

Para el modelo del cliente (modelo SD), un GMM se puede entrenar usando el algoritmo de clustering EM sobre la base de datos de entrenamiento del locutor (Kinnunen *et al.*, 2011). Sin embargo, otra manera consiste en derivar el modelo del cliente, al adaptar los parámetros del modelo UBM usando elocuciones de entrenamiento del locutor y el método de estimación máxima *a posteriori*, MAP (Gauvain & Lee, 1994). La idea en este método de adaptación

es derivar el modelo del cliente actualizando los parámetros a través del modelo UBM. Esto proporciona un acoplamiento ajustado entre el modelo del cliente y el modelo del impostor, acoplamiento que no solamente produce un mejor desempeño del sistema de SV, que los modelos desacoplados, sino que además ofrece una técnica para obtener la verosimilitud de manera rápida.

MAP estima las estadísticas de información de entrenamiento del locutor para cada mezcla en el UBM. Luego, durante la adaptación, se calculan nuevas estimaciones de estadísticas, las que se combinan con las *antiguas* de los parámetros de mezcla del UBM, usando coeficientes de mezcla que son dependientes de la información. Así, dado un modelo UBM y vectores de entrenamiento del locutor hipotético (cliente), se determina el alineamiento probabilístico de los vectores de entrenamiento en las componentes de mezcla del modelo UBM. Por consiguiente, para una mezcla m en el modelo UBM, se calcula:

$$P_r(m \,|\, O(t)) \;=\; \frac{\omega_m \cdot p_m(O(t))}{\sum_{j=1}^{M} \omega_j \cdot p_j(O(t))} \tag{2.8}$$

A continuación, se utiliza $P_r(m \,|\, O(t))$ y $O(t)$ para calcular las estadísticas para los parámetros peso, media y varianza (Bimbot et al., 2004). Por último, estas nuevas estadísticas se usan para actualizar las estadísticas del modelo antiguo UBM durante la mezcla m, para crear los parámetros adaptados peso, media y varianza (Gauvain & Lee, 1994).

El método GMM adaptado permite calcular la verosimilitud de manera rápida. El cálculo de $LL(O)$ requiere obtener la verosimilitud para el modelo del cliente y para el modelo del impostor, para cada vector de características. Dado que el modelo del cliente se adaptó del modelo del impostor, UBM, los valores de verosimilitud se pueden aproximar bastante usando pocas componentes de mezcla, lo que representa un cálculo más rápido. El segundo efecto observado, es que las componentes del GMM adaptado conservan una correspondencia con las mezclas del modelo UBM, de modo que los vectores de características cercanos a una mezcla particular en el modelo UBM también se encuentran próximos a la mezcla correspondiente en el modelo del cliente (Bimbot *et al.*, 2004).

2.2.3.3. Normalización

El último paso en la tarea de SV es el proceso de decisión. Este proceso consiste en comparar la verosimilitud resultante de la comparación con un umbral de decisión, entre el modelo del cliente demandado y la elocución que ingresa. Si la verosimilitud es más alta que el umbral, se aceptará la identidad demandada del cliente, de lo contrario, se rechazará.

El ajuste del umbral es un aspecto problemático en SV. La variabilidad de este umbral viene de distintas fuentes, por ejemplo, la naturaleza de la información de entrenamiento, la que pude variar entre locutores, además, el contenido fonético, la duración, el ruido, así como

también, la calidad del entrenamiento del modelo de cliente.

Por otra parte, también son fuente de variabilidad, el posible *mismatch* entre la condición de entrenamiento (etapa en la que se modela al cliente) y la de prueba. Dos factores pueden contribuir a este *mismatch*. El primero, la variabilidad intra-locutor (Yang, Millar & Macleod, 1996) que se describe como la variación, entre elocuciones de un mismo individuo, de la información acústico fonética, que se extrae de la elocución. De forma análoga, se desprende el concepto de variabilidad inter-locutor, que se relaciona con las variaciones entre elocuciones que pertenecen a un grupo amplio (o universo) de locutores. El segundo factor, lo representan todos los cambios que puedan suceder en la condición ambiental al momento de parametrizar la elocución, como cambios en el canal de transmisión, o bien presencia de ruido de fondo en el ambiente acústico (Garretón, 2011).

Las técnicas de normalización de score se introducen explícitamente para abordar esta variabilidad en el score y hacer más fácil, independiente del locutor, el ajuste del umbral de decisión (Li & Porter, 1988). La normalización basada en cohorte (Higgins, Bahler & Porter, 1991; Matsui & Furui, 1994), toma scores normalizados en la forma de razón de verosimilitudes logarítmicas:

$$\tilde{L}_{\lambda}(O) = \frac{LL_{\lambda}(O)}{LL_{\overline{\lambda}}(O)} \tag{2.9}$$

En este método, que permite mejorar el desempeño del sistema de SV, la verosimilitud $LL_{\overline{\lambda}}(O)$ se estima a partir de una cohorte de modelos de locutores. Esta cohorte se denomina cohorte impostores. Esta normalización forma parte fundamental del modelo de referencia universidal, UBM (Carey, Parris & Bridle, 1991).

2.3. Extracción robusta de características

En verificación de locutor la extracción robusta de características acústicas (por ejemplo, formantes), es un problema de amplio interés (Dimitriadis, Maragos & Potamianos, 2011). A pesar de años de progreso y de los diversos métodos de robustez que se han propuesto, la aplicación práctica de los sistemas de reconocimiento de voz y de verificación de locutor continúa siendo afectada debido a los problemas que surgen por variabilidades entre las condiciones de grabación de las señales de entrenamiento y de evaluación. En ambientes silenciosos y bien controlados, el desempeño de estos sistemas es exitoso. No obstante, su desempeño decae cuando varían las condiciones en que se captura y se transmite la señal de voz. Una consecuencia práctica de la presencia de distorsiones no esperadas, es su efecto sobre las características que son extraídas las cuales son utilizadas en las etapas posteriores en estos sistemas y que afectan en su desempeño final.

Para los seres humanos con audición normal, la voz puede ser asombrosamente bien reconocida, incluso la identidad de un locutor se puede verificar casi sin esfuerzo bajo condiciones adversas ya sea en ambientes acústicos con intenso ruido, o bajo múltiples voces de locutores, hasta por variaciones en los canales de transmisión de la voz (por ejemplo, teléfono) (Darwin, 2008; Moore, Tyler & Marslen-Wilson, 2008b). Diversos métodos motivados en propiedades del sistema auditivo humano se han propuesto tendientes a minimizar el *mismatch* debido a las variabilidades, acústicas o de canal de transmisión, que afectan a las condiciones de grabación de la voz. Por ejemplo, el uso de bancos de filtros auditivos en sistemas de verificación de locutor, se ha inspirado por el proceso de audición humana (Seneff, 1988; Ghitza, 1994). En estos filtros auditivos, se estima la energía de cada componente de frecuencia del espectro de la voz. Además, se han realizado esfuerzos adicionales por incorporar a estos sistemas otras propiedades fundamentales que ocurren en la periferia auditiva, al igual que también, propiedades presentes en las zonas superiores de la corteza auditiva, aunque muchas de estas últimas, todavía son tema de investigación y aún no están bien comprendidadas (Moore, 2014).

La comprensión de los procesos auditivos llevados a cabo en la periferia, principalmente en la cóclea, evolucionó a través de los años presentando tres épocas de investigación bien marcadas. La primera, de 1850 en adelante, estuvo dominada por el trabajo de Helmholtz quien sugirió que cada parte de la membrana basilar entra en resonancia en una frecuencia particular y este análisis espectral se caracteriza por un orden espacial. Dicho período fue revisado por Wever (Wever, 1949).

La segunda época, que duró desde finales de 1940 hasta principios de los años 70, fue dominada por la descripción de von Békésy quien demostró lo sugerido por Helmholtz que el primer paso para el análisis de sonido en el sistema auditivo lo realiza un proceso mecánico en la cóclea (von Békésy, 1947, 1953). Para su investigación, von Békésy adoptó un método experimental para medir sobre los huesos temporales de oídos de cadáveres humanos, la respuesta vibratoria de la membrana basilar a sonidos con niveles de presión sonora extremadamente altos, a partir de lo cual obtuvo gráficas de las curvas de resonancia mecánica de la membrana basilar (Robles & Ruggero, 2001). La onda viajera *pasiva* de von Békésy y su rol sobre la discriminación de frecuencias del oído, significaron una de las contribuciones más importante a la ciencia auditiva (von Békésy, 1960; Moore, 2014).

El trabajo de von Békésy mostró que diferentes frecuencias de un sonido de entrada excitan distintas partes de la cóclea. Las frecuencias altas causan deflexión en la base de la cóclea (el extremo más cercano a la entrada de la cóclea), mientras que las frecuencias bajas producen amplitudes máximas en el extremo más elejado de la entrada a la cóclea llamado *apex* (Rhode, 1971). La descripción de von Békésy fue confirmada por Emadi, Richter & Dallos (2004).

La tercera época que corresponde a la actual, evolucionó con la hipótesis de Gold (1948), un paradigma asombroso y fundamental, aunque muy diferente a la descripción hecha por von Békésy de una cóclea *pasiva*. La tesis de Gold (Ren, 2002) postula que la onda de von Békésy aumenta su amplitud por un proceso de amplificación biológicamente activo, el cual involucra a las células auditivas externas las que actúan tanto como elementos sensoriales, así como también, como mecanismos de realimentación positiva. Esta hipótesis, nombrada por Davis como el amplificador coclear (Davis, 1983), depierta interés. Es asombroso que existan células auditivas que operan como receptores sensoriales, que utilizan un motor biológico para amplificar el movimiento oscilatorio en su entrada y que esta ganancia sea la responsable de aumentar la sensibilidad auditiva, contrarrestando la pérdida de energía debido a la disipación viscosa que acompaña el movimiento de las células (Gold, 1948; Lighthill, 1981; Fisher *et al.*, 2012)

2.3.1. Procesamiento auditivo de la voz

La voz es el sonido más interesante y uno de los sonidos más complejos con los que trata el sistema auditivo (Young, 2008). La representación neuronal de la voz requiere capturar o extraer, aquellas características acústicas de la señal sobre las cuales el cerebro depende para la comunicación del lenguaje. Fundamental para esta representación es la propiedad de las neuronas auditivas periféricas de responder solamente a un rango estrecho de frecuencias, lo que se conoce con el nombre de *frequency tuning* o simplemente *tuning* (Joris *et al.*, 2011). Esta propiedad de *tuning* neuronal cumple un rol crítico en la habilidad para distinguir y separar diferentes sonidos. Mediciones directas del *tuning* son posibles de realizar unicamente con animales de laboratorio y el comportamiento del *tuning* neuronal en el ser humano se infiere a partir de las mediciones obtenidas de estos animales pues tales resultados se consideran comparables (Ruggero *et al.*, 1997; Ruggero & Temchin, 2005).

Uno de los avances trascendentes de las últimas décadas en cuanto a conocimiento del sistema auditivo, particularmente el sistema periférico y cómo éste procesa los sonidos complejos, ha sido llegar a comprender los mecanismos fundamentales de selectividad auditiva de frecuencias y la importancia que esta selectividad tiene en el proceso de audición normal, al igual que también en el proceso de una audición dañada (por ejemplo, usuarios con implantes cocleares) (Moore, 2003b; Zhou & Pfingst, 2014). La selectividad de frecuencias explica diversas características de las respuesta de los filtros auditivos. Así, esta selectividad tienen consecuencias relevantes para la comprensión de la voz.

La voz se caracteriza por bandas de energía que abarcan múltiples frecuencias (por ejemplo, las consonantes y las vocales). Estas bandas de energía cambian en el tiempo tanto en frecuencia como en amplitud, y en particular cambian también las frecuencias formantes de las vocales. El oído tiene al menos tres tareas que realizar sobre los sonidos de voz (Moore,

Tyler & Marslen-Wilson, 2008b). La primera consiste en separar las componentes individuales de frecuencia de las múltiples frecuencias presentes simultáneamente (por ejemplo, separar las frecuencias formantes). Esta tarea denominada selectividad de frecuencia o, análisis o resolución de frecuencia, se define como la habilidad del oído para descomponer un sonido complejo en sus componentes individuales de frecuencia (Joris *et al.*, 2011). Una vez analizada o resuelta la complejidad espetral del sonido, se logra el factor decisivo que es la percepción del timbre del sonido (Moore, 2003a; Pickles, 2008). La resolución de frecuencia en conjunto con la información de los tiempos, son importantes también para determinar el *pitch* del sonido de la voz (es decir, la frecuencia laríngea percibida) (Moore, 2008c). Esta información es fundamental y permite al sistema auditivo diferenciar entre locutores (Mesgarani & Chang, 2012). La segunda tarea para el oído es mejorar los contrastes (Rabinowitz *et al.*, 2011) en espectro y tiempo de las componentes de frecuencia resueltas para compensar bajas razones señal-a-ruido (*SNR*) y ayudar en la tercera tarea, que es extraer las características más significativas a partir de los resultados de la resolución espectral periférica (Nelken, 2008).

2.3.2. Componentes del sistema auditivo

Una de las principales componentes de sistema auditivo es la audición periférica (Moore, 2008a). En el sistema periférico se encuentra el oído externo y el oído medio los cuales actúan como acondicionadores de la señal, es decir, enfatizan aquellas frecuencias que son de mayor relevancia para cada especie en particular (Moore, 2003a; Pickles, 2008). Así, en el ser humano, por ejemplo, se enfatizan frecuencias entre 1 a 3 kHz; para gatos y roedores como conejillos de india (también conocidos como *guinea-pigs*) (Rhode & Cooper, 1993), se enfatizan aproximadamente 5 a 15 kHz (Russel & Nilsen, 1997; Dean, Harper & McAlpine, 2005; Wen *et al.*, 2009). El oído interno o cóclea actúa como el analizador de frecuencias del sistema periférico (Robles & Ruggero, 2001). La membrana basilar y el órgano de Corti en conjunto realizan la función de un banco de filtros que se distribuye a través de la longitud de la membrana basilar (Yates, Winter & Robertson, 1990). A partir de este banco de filtros se proyectan diversos canales auditivos con propiedades de *tuning*, bien definidas y que se reflejan en la actividad de las neuronas en las fibras del nervio auditivo (AN) (Kiang & Moxon, 1974; Eggermont, 1977; Ruggero & Temchin, 2005; Joris *et al.*, 2011). Estas fibras codifican en la forma de un mensaje abstracto, las respuestas filtradas en términos de sus patrones espacio-temporal en un mapa tonotópico (Moore, 2003a).

Por otro lado, las respuestas filtradas y codificadas en actividad neuronal en las fibras del nervio auditivo, son transmitidas hacia el núcleo coclear a través del nervio auditivo. Desde el núcleo coclear esta información codificada se envía a su vez hacia la corteza auditiva por vías de procesamiento paralelo e independientes hacia un número de subsistemas en el cerebro. Sin embargo, aún no se logra aclarar hasta dónde en la corteza auditiva estos subsistemas se mantienen separados mientras ascienden en su camino hacia el cerebro (Young, 1997; Nelken, 2008; Moore, 2014).

Conviene mencionar que esta descripción realizada de las componentes del sistema auditivo es una representación aproximada del sistema global, dado que se ha descrito desde el punto de vista monoaural de procesamiento auditivo y no binaural como es el sistema en la realidad. Por lo tanto, hay que tener presente que en las tareas de separación fuentes de sonido y análisis de sonidos tanto el proceso monoaural como también el binaural, ambos están involucrados simultáneamente (Pickles, 2008). Usando una analogía musical, las características de las respuestas de la periferia auditiva se parecerían a las de un piano, en tanto que las características de las respuestas en la corteza auditiva se aproximarían a aquellas de una orquesta completa (Parbery-Clark, Anderson & Kraus, 2013). Así, en esta tesis el análisis se restringe a procesamiento periférico monoaural de las funciones y propiedades auditivas.

2.3.3. Membrana basilar

La habilidad que le permite al ser humano distinguir las diferentes frecuencias de los sonidos de la voz depende en gran medida de la función que realiza la membrana basilar dentro de la cóclea (Rhode, 1971; Ren, 2002). Esta membrana que separa dos tubos adyacentes llenos de fluidos viscosos, es una banda elástica tensa, muy vibratoria, que tiene una longitud aproximada de 35 mm (ver por ejemplo, Pickles (2008) su Figura 3.1). En los mamíferos, la longitud de la membrana basilar está correlacionada con el peso del cuerpo. Por ejemplo, para el ratón, gato, elefante y algunas ballenas estas longitudes son 7, 25, 60 y 100 mm, respectivamente (Robles & Ruggero, 2001).

En su parte superior la membrana basilar permanece en contacto mecánico con el órgano de Corti, un cordón de tejido nervioso que contiene dentro a las células auditivas, tanto las internas (IHC) como las externas (OHC), y que son las encargadas de transformar las vibraciones mecánicas de la membrana basilar en señales eléctricas (Ren, 2002).

El movimiento de pistón que ejerce el estribo en la ventana oval o también llamada la entrada a la cóclea o base, crea diferencias de presión en el fluido viscoso lo que origina una fuerza que actúa sobre la membrana en dirección perpendicular y que provoca una oscilación mecánica que se propaga longitudinalmente a lo largo de la membrana basilar, desde la base o extremo más cercano a la ventana oval (frecuencias altas), al extremo más alejado llamado *apex* (frecuencias bajas). El desplazamiento transversal de la membrana basilar alcanza un valor extremo máximo o *peak* en una región muy angosta. A esta región de la membrana se le da el nombre de frecuencia característica (CF) y su ubicación varía dependiendo de la frecuencia del sonido de entrada (Robles & Ruggero, 2001). En el mismo instante en que la membrana basilar alcanza el valor *peak* moverá de manera máxima al órgano de Corti causando un incremento máximo en el movimiento de las células auditivas IHC y OHC (Pickles, 2008). Las CFs están ordenadas de acuerdo a un mapa espacial (tonotopía) de tal manera que éstas disminuyen progresivamente desde la base hacia el *apex*, de la misma manera que

disminuye la rigidez a lo largo de la longitud.

Dado que cada lugar de la membrana basilar tiene una CF en la que su respuesta alcanza un *peak*, se considera que la cóclea funciona como un analizador de frecuencias que separa las componentes individuales de un sonido de entrada (Joris *et al.*, 2011). En ingeniería eléctrica esta tarea de análisis de frecuencias se consigue con el uso de filtros selectivos, es decir, filtros que tienen anchos de banda suficientemente angostos para que la información en una de las bandas pueda ser seleccionada, y que en canales de frecuencias adyacentes pueda ser rechazada. La selectividad de frecuencias de la cóclea en analogía con un mecanismo de filtros en ingeniería ha sido ampliamente aceptada (Evans, 1992).

Las OHCs cumplen una función vital en el proceso de amplificación activa de la cóclea. No obstante, su función en el proceso de transmitir información hacia el cerebro es escasa, o tal vez nula (Pickles, 2008). Por otro lado, las IHCs desempeñan la importante función de extraer las características de la información acústica de entrada, codificada en la forma de patrones de descargas eléctricas, y transmitir dichas características hacia el cerebro a través del nervio auditivo. Cada una de las IHCs se considera un canal de frecuencia que transmite información acerca del sonido de entrada pero unicamente dentro de una región espectral específica y muy angosta (*tuning* neuronal) (Ruggero & Temchin, 2005).

Debido al proceso de amplificación activa de la cóclea, la respuesta vibratoria de la membrana basilar muestra un *tuning* que depende del nivel de intensidad del sonido de entrada (Rhode, 1978). Por ejemplo, el *tuning* es muy puntiagudo para bajos niveles de intensidad sonora (niveles de umbral) y presenta pendientes muy pronunciadas, ordenadas casi simétricamente, a ambos lados de la CF. Sin embargo, a altos niveles de intensidad sonora, el *tuning* llega a ser crecientemente asimétrico, mostrando una región de resonancia bastante ancha en la cual además, el valor *peak* se hace menos visible (más redondeado) y se desplaza hacia las frecuencias más bajas (Pickles, 2008).

El proceso de amplificación activo de la cóclea también tienen otra consecuencia importante sobre la vibración de la membrana basilar y que es la no linealidad de las funciones de entrada-salida (Moore, 2003a; Pickles, 2008; Barbour, 2011). La entrada al sistema es el nivel de presión sonora (SPL) en dB que se convierte a velocidad del estribo, mientras que la salida corresponde a la velocidad de la membrana basilar (m/s) (Meddis & Lopez-Poveda, 2002). Estas funciones muestran una respuesta altamente compresiva, es decir, la velocidad de la membrana basilar no es proporcional al nivel de presión sonora, en los niveles de intensidad de entrada medios entre 40 a 80 dB de SPL, con una razón de compresión aproximada de $0.2 - 0.3$ dB/dB (Moore, 1996; Oxenham & Bacon, 2003). Asimismo, las funciones de entrada-salida llegan a ser aproximadamente lineales para niveles de intensidad de entrada bajo < 40 dB de SPL, y también para niveles de intensidad muy altos > 90 dB de SPL

(Rhode, 1971; Yates, Winter & Robertson, 1990; Moore, 2003a; Pickles, 2008). La compresión ocurre solamente para frecuencias que son muy cercanas a la CF (Robles & Ruggero, 2001).

Este comportamiento compresivo no lineal de las funciones entrada-salida para las respuestas vibratorias de la membrana basilar, es el que le permite a la cóclea convertir el inmenso rango dinámico de 120 dB de SPL de entrada auditiva, en un rango de vibraciones que varían entre 30-40 dB apropiados para la transducción de las IHCs (Yates, Winter & Robertson, 1990). La compresión no lineal de las funciones entrada-salida también cumple otro rol importante en muchos aspectos de la percepción auditiva que incluye: discriminación de intensidades, enmascaramiento, sonoridad y percepción del timbre de un sonido (Moore, 2002).

La respuesta a tonos con frecuencia bien por debajo de la CF es aproximadamente lineal. Cuando existe una pérdida de audición en la cóclea, por ejemplo, daño en las células auditivas, las funciones entrada-salida de la membrana basilar llegan a ser muy lineales para todas las frecuencias. De hecho, la diferencia entre personas con audición normal y personas con daño auditivo coclear está en la pérdida de la compresión no lineal de las funciones entrada-salida de la membrana basilar que afecta a las personas con daño auditivo y, por consecuencia, en ellos parece no estar operando el mecanismo *activo* de la cóclea (Moore, 2003b).

2.4. Proceso activo en la cóclea

El funcionamiento de la cóclea es en gran medida mejorado debido a un proceso de amplificación activo definido por cuatro características: 1) amplificación; 2) agudeza en la selectividad de frecuencia; 3) no linealidad compresiva; y 4) emisiones espontáneas otoacústicas. En otras palabras, el efecto del proceso activo no es tanto la suma de energía para una membrana basilar disispativa, sino que la eliminación efectiva de la disipación misma (Hudspeth, 2008).

2.4.1. Amplificación activa

Uno de las propiedades auditivas más significativas y, tal vez, la más importante que ocurre en la cóclea, es el proceso de amplificación activo (Manley, 2000, 2001). Las células auditivas, IHC y OHC, deben amplificar las señales que ingresan para separar la información del ruido de fondo. Tal amplificación constituye una realimentación positiva que aumenta la sensibilidad de la audición y que contrarresta la pérdida de energía producto del amortiguamiento viscoso presente en la cóclea. El movimiento de los grupos de células auditivas no está libre de fricción pues se ve afectado por la presencia de fuerzas viscosas de amortiguamiento.

En general, en el sistema auditivo la amplificación activa sólo ocurre cuando la salida de energía excede a la de su entrada. Cuando esto sucede, el principio de conservación de

energía implica que el amplificador, en este caso, la célula auditiva, ha contribuido al balance de energía (Lighthill, 1981). Desde un punto de vista mecánico de la membrana basilar, el proceso activo se puede comprender al analizar la frecuencia natural de cada segmento de la membrana. Esto es, la frecuencia a la cual aquel segmento va a resonar. Aunque la membrana basilar es más elaborada que un simple resonador mecánico, sus componentes fundamentales son muy similares. Posee elasticidad, caracterizada por su rigidez k; tiene una constante de inercia representada por su masa m; y posee amortiguamiento viscoso que se representa por el coeficiente de arrastre ξ. La impedancia mecánica, o resistencia al movimiento, ofrecida por la membrana basilar durante una estimulación a una frecuencia ω tiene una magnitud:

$$|Z| = \left|\frac{F}{v}\right| = \sqrt{\xi^2 + \left(\omega \cdot m - \frac{k}{\omega}\right)^2} = \sqrt{\xi^2 + (\omega^2 - \omega_0^2)^2 \left(\frac{m}{\omega}\right)^2}$$

en la cual F es la amplitud de la fuerza senoidal de entrada y v la velocidad resultante de la membrana basilar (Lighthill, 1991). La resonancia ocurre en ω_0, la frecuencia natural del sistema que es especificada por $\omega_0 = \sqrt{k/m}$. En resonancia $Z = \xi$ de tal modo que la pregunta es si los grupo de células auditivas pueden producir fuerzas suficientes para contrarrestar el efecto de la viscosidad. Afortunadamente, la respuesta es sí. Las células auditiva localmente sí pueden producir fuerzas activas que aumentan la vibración, vencen el arrastre viscoso y como consecuencia producen una ganancia que acentúa el peak de la onda (Hudspeth, 2008; Fisher *et al.*, 2012).

2.4.2. *Tuning* neuronal

La segunda característica del proceso activo es su capacidad de acentuar el *tuning* neuronal (Ruggero & Temchin, 2005; Joris *et al.*, 2011). En una cóclea sana se refleja el rango angosto de frecuencias que excitan a una célula auditiva dada, especialmente para estimulación cercana al umbral auditivo (Moore, 2003a; Pickles, 2008). Sin embargo, cuando se interrumpe el proceso activo, por ejemplo una cóclea con daño auditivo, o en un oído de un cadáver, la disminución de la sensibilidad es acompañada por una severa degradación del *tuning* neuronal (Eggermont, 1977).

2.4.3. No linealidad compresiva

La no linealidad compresiva es la tercera característica del proceso activo de la cóclea (Yates, Winter & Robertson, 1990; Oxenham & Bacon, 2003). En la periferia auditiva, un tono de umbral 0 dB (ref. presión sonora umbral $2 \times 10^{-5}(N/m^2)$) provoca una oscilación en la membrana basilar cercana a $\pm 0.1(nm)$ (Hudspeth, 1989). El sonido tolerado más alto, por ejemplo, un oído al lado de una turbina de una avión, puede llegar a provocar 120 dB lo que moverá a la membrana basilar solamente $\pm 100(nm)$ (Gillespie, 2004). En otras palabras, una entrada de un millón de veces la presión sonora umbral produce una oscilación solamente de unas mil veces la respuesta umbral, lo que indica que el crecimiento de la salida

es enormemente comprimido relativo a aquél de la entrada. Por consiguiente, la sensibilidad de la membrana basilar está caracterizada por una relación no lineal (Ruggero *et al.*, 1997).

2.4.4. Emisiones espontáneas otoacústicas

Las bases del concepto de función coclear activa fueron desarrolladas en los años 70 y 80 (Dallos, 1992). El comportamiento de una cóclea pasiva o muerta como la describió von Békésy, no tiene características como amplificación o sensibilidad, agudeza del *tuning* y no linealidad compresiva. La naturaleza mecánica del proceso coclear activo fue sugerido al detectar en ambientes muy silenciosos, sonidos en el canal auditivo debido a oscilaciones espontáneas, aparentemente de origen coclear, retransmitido por el oído medio (He & Ren, 2013). Estas oscilaciones espontáneas, llamadas emisiones otoacústicas, (Wilson, 1980), son la evidencia más fuerte disponible que vibraciones pueden ser producidas en la cóclea. Las emisiones de sonido desde el oído que se originan en la cóclea en respuesta a entradas acústicas también apoyan la validez de procesos activos que proporcionan amplificación local (Moore, 2003a; Pickles, 2008).

2.4.5. Curva *tuning*

Fundamental para comprender la selectividad auditiva de frecuencias es la curva *tuning* de una fibra individual del nervio auditivo (Kiang & Moxon, 1974; Eggermont, 1977), la que describe la intensidad umbral de sonido requerido para obtener una respuesta mínima de la fibra del nervio auditivo (un aumento en la tasa de descarga eléctrica) como una función de la frecuencia (Sachs & Young, 1979; Young & Sachs, 1979; Yates, Winter & Robertson, 1990; Wen *et al.*, 2009). Bajo condiciones normales la respuesta *tuning* de las fibras del nervio auditivo es muy puntiaguda, esto es, la fibras representan filtros con *tuning* muy fino y con anchos de banda del orden de un sexto de una octava (Winslow & Sachs, 1987; Taberner & Liberman, 2005).

Existe evidencia que el *tuning* es fisiológicamente vulnerable, es decir, por ejemplo, una reducción progresiva en el suministro de oxígeno a la cóclea, puede cambiar el tuning de puntiagudo y con umbral bajo, a una curva sin punta definida más bien redondeada y con umbral alto (Moore, 2003a; Pickles, 2008). Esta es una de las líneas de investigación que aportó evidencia para validar la hipótesis de la existencia de un proceso biológicamente activo que define claramente el *tuning* neuronal y que lo distingue de la mecánica pasiva.

La curva *tuning* que representa la energía o nivel de intensidad del sonido versus la frecuencia de ese sonido (ver por ejemplo, Young, 2008, su Figura 1.(a)), se caracteriza por tener forma de-V, siendo la parte puntiaguda, o bien, más "aguda", la que indica la frecuencia a la cual una fibra es más sensible (Eggermont, 1977).

2.5. Propiedades de las salidas del nervio auditivo

Toda la información acerca del sonido es transportada al cerebro a través de las células ciliares y del nervio auditivo (Holmberg, Gelbart & Hemmert, 2007; Eggermont, 2001; Cariani, 1999). De esta manera, la energía del estímulo sonoro se convierte en mensajes electroquímicos, llamados *spikes*. Estos mensajes afectan al sistema nervioso central (CNS) dando origen a experiencias sicológicas, es decir, produciendo sensaciones y percepciones (Moore, Tyler & Marslen-Wilson, 2008b). Alrededor de 3.500 células ciliares internas conectan a 30.000 neuronas auditivas (Pickles, 2008). Por lo general, tanto en animales como en seres humanos, es posible medir los spikes o mensajes electroquímicos que se generan en respuesta al estímulo sonoro (Pickles, 2008; Uysal, Sathyendra & Harris, 2008). Son comunes dos tipos de medidas de *spikes* (ver por ejemplo, Rose *et al.* (1971) sus figuras: Fig. 8 y Fig. 10):

1. **Tasa de descarga**: Consite en medir el número de impulsos en el nervio auditivo en un tiempo dado en una sola neurona. Dado que una neurona podría no emitir a menudo, es normal calcular un histograma de tiempo, post-estímulo, al medir tiempo de llegada de los *spikes* para repeticiones del mismo estímulo.

2. **Estadísticas del tiempo de *spike***: Consiste en obtener un histograma de tiempo entre *spikes* adyacentes en una neurona sola. Nuevamente, esto es calculado usando repeticiones del mismo estímulo.

2.5.1. Respuesta síncrona en el nervio auditivo

Para un estímulo sonoro monofrecuencial, el intervalo de tiempo dominante entre descargas sucesivas del nervio auditivo corresponde a la inversa de la frecuencia del estímulo (Young & Sachs, 1979, su Figura 2). La respuesta síncrona es una propiedad importante que posee el sistema periférico auditivo (Engel, Fries & Singer, 2001; Varela *et al.*, 2001; Kim, Chiu & Stern, 2006). Esta sincronía demuestra que las fibras del nervio auditivo están cerradas en fase (*phase-locked*) al estímulo sonoro, es decir, la respuesta del nervio auditivo ocurre durante un segmento restringido de un ciclo del estímulo monofrecuencial (Sinex *et al.*, 2003). Sin embargo, las fibras pueden no ser capaces de emitir mensajes o *spikes* cada vez (Moore, 2003a; Pickles, 2008).

De esta manera, frecuencia y/o intensidad, ambas información espectrales muy importantes de un locutor, pueden ser codificadas por tiempo y no por tasa de descarga. Es significativo observar que este cierre de fase (*phase-locking*) se rompe progresivamente en frecuencias medias y altas y se termina aproximadamente en 4000 Hz (Shamma, 1985).

2.5.2. Dinámica de las salidas del nervio auditivo

Las características dinámicas de las salidas del nervio auditivo se han medido usando histogramas de tiempo, post-estímulo, en respuesta a repeticiones del sonido. Estos histogramas revelan diversas características (Pickles, 2008, su Figura 6.4 (a) y (b)):

- **Tasa espontánea**: En ausencia de estimulación la salida está por encima de cero.
- **Respuesta onset**: La neurona es más probable que descargue en el onset (inicio del estímulo) más que en otro tiempo.
- **Respuesta adaptada**: La tasa de descarga rápidamente se adapta a un nivel entre el onset y la tasa espontánea.
- **Respuesta offset**: La tasa de descarga disminuye por debajo de la tasa espontánea y puede tomar algún tiempo recuperarse.

2.5.3. Relación no lineal entre la tasa de descarga versus el nivel de intensidad

Está limitado el número de *spikes* que una fibra del nervio auditivo puede generar en un tiempo dado (Pickles, 2008). Esto se puede ilustrar al usar una función sigmoide. La tasa de descarga se entiende como una función del estímulo sonoro (Moore, 2003, su Figura 1.17).

- En bajas intensidades de sonido la fibra continúa emitiendo a la tasa espontánea.
- En intensidades medias de sonido la fibra sigue casi un comportamiento lineal.
- A altas intensidades de sonido satura la tasa de descarga de la fibra.

Dado que la tasa de descarga de las fibras del nervio auditivo cambia como función del nivel del estímulo, las curvas resultantes se denominan "funciones tasa-nivel". Las neuronas auditivas con alta tasa espontánea conforman el 61 % del nervio auditivo. Por sobre un cierto valor de intensidad, la neurona ya no responde al incremento del nivel del estímulo y la neurona se dice que está saturada (Viemeister, 1988).

El rango de los niveles de sonido entre umbral y nivel de saturación se denomina "rango dinámico" (Dean *et al.*, 2008; Barbour, 2011). Para neuronas con alta tasa espontánea, este rango es muy pequeño, aproximadamente 15 a 30 dB (Moore, 2003a; Zilany & Carney, 2010). Las neuronas auditivas con tasa espontánea media conforman el 23 % del nervio auditivo. Para estas neuronas, el umbral es ligeramente mayor y el rango dinámico es un poco más ancho. Las neuronas con tasa espontánea baja conforman aproximadamente el 16 % del nervio auditivo. El umbral es mayor y, en principio, la tasa de descarga aumenta bastante rápido con el incremento del nivel de intensidad, pero luego la tasa disminuye.

2.6. Codificación de voz en el nervio auditivo: vocales

Para resumir las propiedades de las fibras del nervio auditivo, se considera la representación de la voz a través de sonidos de vocales, como una función del nivel de intensidad, o como una función de la cantidad de ruido adicionado (Sachs & Young, 1980, su Figura 5). A bajas intensidades, las amplitides de los formantes vocálicos son claramente representadas por la tasa de descarga como una función de la frecuencia. Cuando aumenta la intensidad, el perfil de tasa se aplana, y los formantes ya no son más visibles. Todavía un auditor puede identificar vocales con claridad a esas intensidades. Es probable que la saturación y la supresión de dos tonos sean los responsables (Sachs & Young, 1979). También, es posible que las fibras con rango dinámico mayor hagan un mejor trabajo de preservar el perfil espectral hasta este punto (Sachs & Young, 1979; Young & Sachs, 1979; Young, 2008).

2.6.1. Codificación de vocales usando información temporal

La capacidad de las fibras del nervio auditivo de descargar síncronamente (*phase-locked*) con el estímulo sonoro es fundamental para la codificación de los sonidos de las vocales (Sachs & Young, 1979; Young & Sachs, 1979; Sachs & Young, 1980; Cowper-Smith & Dingle, 2010), (Sachs & Young, 1980, su Figura 10).

Las respuestas de las fibras del nervio auditivo se caracterizan como descargas (*spikes*) cerradas en fase, (*phase-locked*), esto es, ocurren durante un segmento del ciclo de la forma de onda, con la componente más cercana a la del estímulo vocálico (Sachs & Young, 1979; Young & Sachs, 1979). Cuando aumenta la intensidad del sonido de entrada, el phase-locking se mantiene incluso aunque la tasa media de descarga ya comience a saturar (Shamma, 1985).

2.7. Modelo auditivo de Seneff

El modelo auditivo de Seneff es un modelo conocido en procesamiento auditivo periférico (Stern & Morgan, 2012a,b). Se utiliza en este trabajo debido a su estructura simple y porque facilita un análisis etapa por etapa del modelo. Seneff (1988) propuso un modelo del sistema auditivo periférico que se divide en dos fases (Seneff, 1988, su Figura 1 (a) y (b)). Cada fase refleja las principales transformaciones que experimenta el sonido de entrada cuando se transmite por la periferia auditiva. En primer lugar, el modelo contempla la vibración en la membrana basilar (Robles & Ruggero, 2001). A continuación, la transducción neuronal de las células auditivas en las fibras del nervio auditivo (Meddis & Lopez-Poveda, 2002). La señal de sonido pasa a través de un banco de filtros auditivos pasabanda el que representa la función de separación de frecuencias en la cóclea (Moore, 2003a; Pickles, 2008). Este banco de filtros contiene canales de frecuencia con bandas relativamente angostas en la región de baja frecuencia, en tanto que éstas se hacen más anchas para los canales de frecuencias más altas (Moore, 2003b).

Seneff (1988) propuso un modelo para las células auditivas el que supone la transducción electroquímica desde la vibración en la membrana basilar, representada por las salidas del banco de filtros, pasando por la tasa media de descarga eléctrica neuronal de las fibras del nervio auditivo, la que varía durante el tiempo. Las principales etapas de su modelo son:

a) **Rectificación de media onda con compresión no lineal**: Esta etapa representa la naturaleza positiva de la tasa de generación de descargas eléctricas y la relación entrada-salida, entre intensidad sonora y tasa de descarga eléctrica.

   A continuación del banco de filtros auditivo, se observa una etapa de rectificación de media onda, con una compresión no lineal y saturación. La descripción de esta etapa viene dada por (Seneff, 1988):

$$\begin{aligned} y &= 1 + A \cdot tan^{-1} Bx \quad x > 0 \\ &= e^{ABx} \quad x \leq 0 \end{aligned} \tag{2.10}$$

   donde x es la entrada, y es la salida, A y B son valores constantes (10 y 65, respectivamente). La función es exponencial para entradas negativas, lineal para valores pequeños de entrada y compresiva para señales más grandes.

b) **Adaptación *short term***: Esta etapa modela el proceso electroquímico que generan los *spikes* o descargas eléctricas, originadas en la respuesta de la cóclea. El modelo describe dos mecanismos separados que influyen en la concentración de los neurotransmisores. Una membrana permite el flujo desde una fuente a una tasa proporcional al gradiente a través de la membrana, con una constante de proporcionalidad μ_a. Cuando el gradiente es negativo, esto es, la concentración en la fuente es pequeña, los canales en la membrana se cierran y los neurotransmisores se pierden por decaimiento natural a una tasa que es proporcional a su propia concentración dentro de la región, con una constante de proporcionalidad μ_b. Matemáticamente, esto se expresa en la forma (Seneff, 1988):

$$\begin{aligned} \frac{dC(t)}{dt} &= \mu_a \cdot [S(t) - C(t)] - \mu_b \cdot C(t), \quad C(t) < S(t) \\ &= -\mu_b \cdot C(t), \quad C(t) \geq S(t) \end{aligned} \tag{2.11}$$

   donde $C(t)$ es la concentración de neurotransmisores dentro de la fuente y $S(t)$ es la concentración en la entrada. La salida del sistema es representada por el flujo sobre la

membrana: $\mu_a[S(t) - C(t)]$. Las constantes μ_a y μ_b son 8.3 $[s^{-1}]$ and 58.3 $[s^{-1}]$, respectivamente (Seneff, 1985).

c) **Filtro pasabajos**: Representa la pérdida de sincronía en altas frecuencias. Esta etapa se usa para modelas la supresión de sincronismo que ocurre en frecuencias altas debido a latencias neuronales. Esto atenúa la capacidad de *phase-locking* arriba de 4000 [Hz] (Seneff, 1985; Ali, Van Der Spiegel & Mueller, 2002).

d) **Control automático de ganancia**: Modela el límite de los spikes impuesto por la incapacidad de generar spikes en cortas sucesiones de tiempo. Esta componente final está definida por:

$$y[n] = \frac{x[n]}{1 + K_{AGC} \cdot \langle x[n] \rangle} \tag{2.12}$$

donde K_{AGC} es una constante, y $< >$ simboliza el valor esperado (Seneff, 1985).

2.7.1. Estimación de la tasa de descarga en las fibras del nervio auditivo

La representación de la transmisión del sonido hacia las etapas superiores del sistema auditivo, es decir, analizado desde el punto de vista del ser humano y su percepción frente al sonido, se ilustra como codificación de información, tanto en frecuencia (información espectral), como en tiempo (información temporal), por el número de descargas eléctricas (o tasa de descarga), dentro de un corto intervalo de tiempo, en respuesta al sonido, cuando este es proporcional al nivel intensidad sonora (Sachs & Young, 1979; Young & Sachs, 1979). Si la señal de sonido de entrada, se mantiene dentro de un nivel apropiado para evitar saturación en las fibras del nervio auditivo, el patrón de descarga puede preservar el contenido de frecuencia y su representación se transmite hacia las etapas superiores del sistema auditivo (cerebro auditivo). Dado que las salidas del modelo auditivo se miden en *spikes*/segundo, se considera que la tasa de descarga puede ser descrita por el número de *spikes* dentro de un cierto intervalo de tiempo Moore (2003a); Pickles (2008).

2.7.2. Estimación de sincronía en las fibras del nervio auditivo

Sachs & Young (1979) usaron sonidos de vocales como estímulo de entrada para medir la respuesta de las fibras del nervio auditivo (AN) en gatos anesteciados y así, obtener una representación espectral del sonido vocálico. Estos autores demuestran que el método de estimación tasa de descarga no es suficiente, ya que es fuertemente dependiente del nivel de intensidad del sonido y se ve afectado por la saturación de las fibras (Sachs & Young, 1980; Sinex *et al.*, 2003; Taberner & Liberman, 2005). Para altos niveles de intensidad, el efecto

negativo es que las frecuencias de los formantes tienden a desaparecer cuando se utiliza este método, lo que significa que los formantes ya no son visibles (Young & Sachs, 1979; Sachs & Young, 1980).

Young & Sachs (1979) usaron gatos anesteciados, para estimar las respuestas de las fibras del nervio auditivo a sonidos de vocales, utilizando un método temporal de sincronía. Estos autores demuestran que a altos niveles de intensidad sonora, se mantiene la información espectral del primer formante vocálico y las de sus armónicos, aumentando su amplitud a pesar del incremento del nivel de intensidad.

Estos resultados se consideran importantes para una aplicación de un método basado en sincronía en una tarea de verificación de locutor. Aun cuando tales estudios hayan sido obtenidos en animales (Cariani, 1999; Wen *et al.*, 2009), también debiera ser posible suponer su relación directa con la percepción y procesamiento de los sonidos vocálicos en los seres humanos (Moore, 2003b; Pickles, 2008).

La estimación de sincronía en las fibras del AN se utiliza para detectar las características temporales de las respuestas *phase-locked* con la forma de onda del sonido de entrada. (Seneff, 1988) propuso un detector de sincronía generalizado llamado GSD, con el fin de realzar los *peaks* espectrales más prominentes en las resonancias de las frecuencias formantes y así mejorar la resolución espectral. El GSD de Seneff, busca detectar periodicidades en las respuestas temporales y no utiliza el método de tasa de descarga. Seneff (1988) genera una razón limitada suave, del valor de la magnitud esperada de la suma y diferencia de la salida de cada filtro y una versión retardada de ésta (Seneff, 1985). El retardo de cada GSD debe corresponder con su respectiva frecuencia central del canal que se analiza. Esto significa que el retardo es igual a la inversa de la frecuencia central (Ali, Van Der Spiegel & Mueller, 2002). La siguiente ecuación representa esta idea:

$$\mathrm{GSD}_i(y) = A_s \arctan\left[\frac{1}{A_s}\left(\frac{\langle|y[n] + y[n-n_i]|\rangle - \delta}{\langle|y[n] - \beta^{n_i} y[n-n_i]|\rangle}\right)\right] \tag{2.13}$$

donde $y[n]$ es la entrada al GSD (esto es, salida de la etapa de Control Automático de Ganancia) en el tiempo n, GSD_i es la salida de sincronía del canal i-ésimo, es decir, sintonizado al filtro i, al hacer $n_i = f_S/f_i$, donde f_i es la frecuencia central del i-ésimo filtro y f_S es la frecuencia de sampleo, $< >$ representa la envolvente, y A_S, β y δ son constantes.

La suma y diferencia se construye a partir de la salida de la etapa de Control Automático de Ganancia. La constante β se hace un valor ligeramente menor que 1,0. Un umbral pequeño δ se resta del numerador para suprimir la respuesta a señales de pequeñas amplitud. Su valor se escoge ligeramente mayor que la tasa espontánea. La nolinealidad de saturación se usa

como límite suave de las salidas y para prevenir respuestas infinitas. A pequeñas amplitudes la respuesta es casi lineal y luego satura para entradas de gran amplitud. El rango lineal de la entrada se controla con el valor A_S (Seneff, 1985; Ali, Van Der Spiegel & Mueller, 2002).

2.7.3. Estrategias de procesamiento de señales de voz inspiradas en la fisiología de la audición para extracción de características

En el área de reconocimiento automático robusto de locutor, varios han sido los métodos propuestos que se basan en principios de la fisiología auditiva periférica y en los tipos de codificación de la actividad neuronal en la fibras del AN, para seres humanos como también para otros mamíferos (Rhode & Cooper, 1993). El término "periférico" se utiliza para dar a entender periférico al sistema nervioso central (CNS): la salida de la periferia es la actividad del nervio auditivo (AN), la que es además, la entrada al CNS (Moore, 2003a; Pickles, 2008).

La mayoría de los métodos de extracción de características basados en la audición, se focalizan en la periferia auditiva, pues el nivel de conocimiento de los principios de funcionamiento del cerebro auditivo no está desarrollado con un nivel de detalle comparado con la periferia auditiva (Moore, 2014).

Existe evidencia que la implementación cuidadosa de estrategias de procesamiento de señales de voz, que se inspiran fisiológicamente, pueden conducir a un incremento importante de la robustez de los vectores de características que se utilizan en situaciones en las cuales la señal de voz se degrada por efectos de fuentes de ruido o variabilidades del canal de transmisión (Stern & Morgan, 2012a).

Sin embargo, la extracción robusta basada en principios biológicos, todavía permanece como un problema abierto el cual puede llegar a contribuir a mejorar el desempeño de los sistema en general. Fundamentalmente, la mayor parte de la información fisiológica empleada en tecnologías de procesamiento de voz, proviene de estudios sobre animales (Kiang & Moxon, 1974; Sachs & Young, 1979; Young & Sachs, 1979; Sinex *et al.*, 2003; Taberner & Liberman, 2005; Dean *et al.*, 2008; Wen *et al.*, 2009; Rabinowitz *et al.*, 2011). Específicamente, la cóclea en el oído interno de los mamíferos, es aproximadamente la misma, con la diferencia principal, el rango de frecuencias sobre el cual funcionan las componentes mecánicas (Robles & Ruggero, 2001; Pickles, 2008).

Dos teorías de codificación dominan los principios neuronales a la salida de la cóclea: la codificación por medio de la tasa de descarga y la codificación por sincronía (Eggermont, 2001). Ambas teorías, sirven para diferenciar aspectos de la audición y conservar características acústicas importantes de los sonidos de la voz (Young, 2008).

No obstante, bajo ambas teorías, existe actualmente controversia sobre si el cerebro auditivo es capaz o no, de utilizar combinaciones de *features* que provengan de codificaciones diferentes del sonido a nivel del nervio auditivo (Moore, 2014), y si tal conjunto de combinaciones permite o no, al cerebro obtener una representación complementaria que proporcione más información acerca de las características de un sonido de voz (Nelken, 2008).

Liberman (1980) fue el primero que demostró que el nervio auditivo (AN) no consiste en una población homogénea de fibras: al menos existen dos subpoblaciones, una población pequeña de baja tasa espontánea de descarga pero que tiene umbrales más elevados y rangos dinámicos más anchos. La otra subpoblación tiene las tasa de descarga más altas, umbrales más bajos y rangos dinámicos restringidos (Liberman, 1980; Evans, 1992). El umbral de una neurona es el nivel presión sonora (SPL) más bajo frente al cual se mide la respuesta de la neurona (Moore, 2008c; Pickles, 2008).

La cóclea representa el analizador de frecuencias del sistema auditivo periférico (Robles & Ruggero, 2001). La membrana basilar y el órgano de Corti, juntas realizan la función de un banco de filtros distribuido, con canales sintonizados en bandas muy agudas, que emergen de las respuestas de las fibras del AN (Pickles, 2008). Estas fibras codifican las respuestas filtradas en términos de patrones de descargas, representados en espacio (o "lugar" sobre la membrana basilar) y tiempo, en una forma de mapa tonotópico, de actividad de la población de fibras que nacen en la cóclea (Liberman, 1980). Esta es la teoría de codificación "tasa-lugar" o simplemente "tasa de descarga".

Sin embargo, cuando la cóclea está dañada, se produce una codificación "borrosa" del espectro de la señal en su tasa de descarga. Esta es la razón por la que para las personas con daño coclear (sordas o casi sordas), cualquier amplicación lineal, como la que ofrece la ayuda auditiva convencional, aunque hace audible la voz, de ninguna manera la hace más clara o inteligible (Moore, 2008c; Pickles, 2008). Este desorden está en la cóclea, más bien que en el CNS y el nervio auditivo puede estar prácticamente intacto (Moore, 2003b, 2008c; Zhou & Pfingst, 2014). En este grupo de personas, se puede crear una sensación de sonido, estimulando eléctricamente en forma directa el nervio auditivo, debido al hecho que el AN está conectado directamente al CNS (Pickles, 2008).

Similar "borrosidad" de la codificación de tasa de descarga, ocurre como resultado de severas no linealidades en las respuestas de las fibras del nervio auditivo, en especial, debido a lo restringido de los rangos dinámicos de la mayoría de las fibras (Liberman, 1980). El rango de intensidades sonoras capaces de producir cambios en la tasa de descarga debido a cambios en el nivel del sonido, es limitado aproximadamente 30-40 dB en la mayoría de las fibras del AN (Sachs & Abbas, 1974; Zilany & Carney, 2010; Sumner & Palmer, 2012).

Producto de esta restricción, el mapeo de actividad en términos de la tasa de descarga en respuesta de la mayoría de las fibras del AN, es borroso casi completamente, a niveles moderados y altos de nivel sonoro (Young, 2008). Es decir, las fibras del nervio auditivo no son capaces de codificar el espectro de la señal de voz a niveles de sonido, moderados y altos (Young, 2008; Pickles, 2008). Así, en la búsqueda de alternativas para codificación de tasa, el potencial para codificar sonidos de voz sobre un rango dinámico más amplio se explora en términos de los patrones temporales de descarga, en las fibras del nervio auditivo (Johnson, 1980; Sachs, 1984; Kayser *et al.*, 2009).

El punto importante aquí en esta nueva alternativa de codificación, es que el tiempo de las descargas a frecuencias hasta 4kHz aproximadamente (Shamma, 1985), se relaciona al período o a múltiplos del período de la señal de voz. Este sincronismo o *phase-locking* de las descargas con respecto a la forma de onda de la señal, representa el aspecto fundamental de la teoría de codificación temporal de los sonidos de la voz, en especial, las vocales (Delgutte & Kiang, 1984; Seneff, 1988; Ghitza, 1994; Ali, Van Der Spiegel & Mueller, 2002; Kim, Chiu & Stern, 2006; Kayser *et al.*, 2009).

Capítulo 3

Optimización de parámetros de las funciones sigmoidales tasa-nivel basada en características acústicas

En este capítulo se describe el desarrollo de una función sigmoidal óptima tasa-nivel que es una componente de muchos modelos del sistema auditivo periférico. Es importante señalar que para este desarrollo se consideran las siguientes hipótesis formuladas anteriormente:

H1) El uso de modelos del sistema auditivo periférico en tecnologías de verificación de locutor no ha sido suficientemente explorado. Las propiedades del sistema auditivo demuestran un alto potencial y aplicabilidad.

H2) El principio de no linealidad sigmoidal tasa-nivel puede contribuir a la robustez de un sistema de verificación de locutor bajo condiciones de ruido aditivo.

Consecuentemente, dadas estas hipótesis y considerando los objetivos planteados al inicio de esta Tesis, la optimización hace uso de un conjunto de criterios definidos exclusivamente sobre la base de atributos físicos del sonido de entrada los que se inspiran en evidencia fisiológica auditiva. Los criterios desarrollados intentan discriminar entre una señal de voz degradada y ruido para preservar la máxima cantidad de información en la región lineal de la curva sigmoidal, y para minimizar los efectos de distorsión en las regiones de saturación. En consecuencia, y de acuerdo a la metodología propuesta, el desempeño de la función sigmoidal óptima propuesta se valida con experimentos de verificación de locutor de texto-independiente, con señales degradadas por ruido aditivo a diferentes relaciones-señal-ruido (SNRs). Los resultados experimentales sugieren que el método presentado en combinación con normalización de varianza cepstral (CVN) puede conducir a reducciones relativas en la tasa de error (EER) tan grandes como 40 % cuando se compara con el uso del sistema *baseline* de los coeficientes cepstrales para ciertas SNRs.

3.1. Introducción

Los sonidos de la voz son ondas de presión que varían como una función del tiempo. Estos sonidos atraviesan el sistema auditivo periférico antes de ser convertidos en actividad neuronal eléctrica en el nervio auditivo (Pickles, 2008). Una descomposición espectral se realiza en la cóclea la que separa los sonidos de la voz en sus componentes constituyentes de frecuencia y la información es transmitida al nervio auditivo, al tronco cerebral y finalmente, a la corteza auditiva, a través de canales que permanecen dependientes de la frecuencia. En un ambiente natural, tanto los sonidos de la voz objetivo, como también el ruido, ingresan juntos al sistema auditivo periférico. Sin embargo, una de las características más impresionantes del sistema auditivo es su habilidad para responder y distinguir a cualquier sonido de voz ante ruido de fondo (Darwin, 2008). En este capítulo se introduce una nueva manera de mejorar la precisión en tareas de verificación de locutor, al incorporar un tipo particular de adaptación en una representación utilizada para extracción de características que se basa en procesamiento en la periferia auditiva.

3.1.1. Procesamiento neuronal de señales de voz

El procesamiento neuronal de la señal de voz se representa por patrones temporales de impulsos neuronales (o bien, descargas de "*spikes*") trasnmitidos a lo largo de las fibras del nervio auditivo, los cuales varían en el tiempo en respuesta al sonido de entrada. La dependencia del número promedio de *spikes* por segundo sobre la intensidad de la señal de entrada, en una región de frecuencia particular, se resume por las curvas llamadas funciones tasa de descarga-versus-nivel de intensidad, o bien, simplemente, funciones tasa-nivel (Moore, 2003a; Pickles, 2008). Estas funciones muestran una variedad de formas, aunque ellas son por lo general, sigmoidales (por ejemplo, Sachs & Abbas (1974); Johnson (1980); Yates, Winter & Robertson (1990)). Bajo estas condiciones, las funciones tasa-nivel se pueden caracterizar por cuatro atributos: (1) umbral de descarga; (2) tasa de descarga máxima; (3) tasa de descarga espontánea; y (4) rango dinámico (Nizami, 2005). Según lo describe Young (2008), el rango dinámico en este contexto se refiere a "el rango de niveles sonoros sobre los cuales la fibra cambia su tasa cuando cambia el nivel."

La mayoría de las fibras del nervio auditivo muestran un rango dinámico menor a 35 dB cuando son estimuladas con tonos en su frecuencias característica (May & Sachs, 1992). Por el contrario, en los seres humanos el rango dinámico de percepción de sonoridad es tan grande como 100 dB de nivel de presión sonora (Winslow & Sachs, 1987). A través de los años, un gran número de hipótesis se han formulado sobre cómo los humanos pueden percibir los cambios en sonoridad sobre un rango dinámico de percepción tan amplio, a pesar de que el rango dinámico intrínseco de las fibras del nervio auditivo se limita a 20-35 dB. Estas hipótesis han incluido la consideración de las distribuciones de los umbrales de fibras individuales del nervio auditivo, la extensión de la excitación de las fibras sobre la frecuencia,

y la posible codificación de la sonoridad basada en respuesta síncrona, al menos a frecuencias bajas ($< 3 - 4$ kHz) (Shamma, 1985).

En los últimos años, la atención se ha centrado en la habilidad potencial de la respuesta en el nervio auditivo para desarrollar funciones tasa-nivel que varíen de acuerdo a la distribución de los niveles de intensidad del estímulo (Barbour, 2011; Dean, Harper & McAlpine, 2005; Dean *et al.*, 2008). Por ejemplo, experimentos con gatos han demostrado que el rango dinámico en sus neuronas auditivas, frente a estímulos de tonos puros y ruido, se adapta a la distribución de los niveles sonoros. Esta adaptación es caracterizada por desplazamiento del rango dinámico hacia los niveles que ocurren más frecuentemente (Wen *et al.*, 2009, 2012). Las funciones tasa-nivel en los conejillos de india (*guinea pigs*) muestran además un rango dinámico limitado y cambiable. En estos animales, las respuestas neuronales se ajustan rápidamente y tienden a mejorar la codificación de los niveles sonoros (Dean, Harper & McAlpine, 2005). Las fibras del nervio auditivo en el ratón también muestran un comportamiento similar aunque con diferencias en los rangos de frecuencia (Taberner & Liberman, 2005). Más adelante, en la siguiente sección, se elaboran resultados y consecuencias potenciales basados en estas observaciones fisiológicas de la audición en animales.

Durante muchos años las propiedades del sistema auditivo han atraido el interés de investigadores en procesamiento de voz, lo que incluye el uso de modelos del sistema auditivo como parte del proceso de extracción de características en reconocimiento automático de voz, verificación de locutor, etc. Parte de ese trabajo se revisa en Stern & Morgan (2012a,b) y los primeros modelos computacionales del sistema auditivo periférico que se han desarrollado incluyen el trabajo de Allen (1985), Ghitza (1986, 1994), Lyon (1982), Seneff (1988), Shamma (1988), y Cohen (1989). La mayoría de estos modelos se inician con un banco de filtros sintonizado a diferentes frecuencias centrales, el cual modela la descomposición espectral de los sonidos de entrada a la cóclea, seguido de un modelo de transducción que incluye la nolinealidad sigmoidal del proceso de transducción auditiva que transforma el movimiento mecánico que ocurre en la cóclea, en producción de *spikes* en el nervio auditivo. Como un ejemplo de este último mecanismo, el modelo de Seneff incluye una representación de las células auditivas internas que consiste de cuatro etapas: (1) una no linealidad tasa-nivel que limita las respuestas de componentes de la señal en una frecuencia particular con amplitudes muy pequeñas y muy grandes; (2) una adaptación en corto plazo (*short-term*) que modela la liberación de neurotransmisores durante la etapa de sinápsis; (3) un filtro pasa bajos que modela la pérdida de sincronía en respuesta a componentes de alta frecuencia; y (4) un control automático de ganancia que mantiene una presencia de sonidos de alta intensidad cuando ha saturado el nervio auditivo. Seneff (1988) propuso dos caminos paralelos, no interactuantes, para analizar las salidas de esta representación. Un camino mide la energía *short-term* total instantánea presente a la salida de cada canal, y el otro camino desarrolla un representación espectral basada en la extensión a la cual la señal de salida es sincronizada a la frecuen-

cia característica de la respuesta de la fibra. Durante años numerosos grupos han utilizado modelos auditivos tales como los que se han mencionado más arriba para desarrollar métodos de extracción de características para ser empleados en reconocimiento de voz e identificación de locutor, entre otras tecnologías, (por ejemplo, Kim, Chiu & Stern (2006); Kim, & Stern (2012)).

3.1.2. Extracción de características para verificación de locutor

En verificación de locutor el propósito es determinar si una señal de voz dada pertenece, o no, a una persona demandada, basado solamente en una muestra de voz (Reynolds, 1995). Generalmente, un sistema de verificación de locutor comprende tres secciones: 1) extracción de características; 2) modelamiento del locutor; y 3) etapa de decisión (Kinnunen & Li, 2010). La sección de extracción de características se diseña para suministrar al sistema suficiente información discriminativa a partir de la señal de voz para permitir que el locutor sea verificado (Li & Huang, 2011). El desarrollo de características relevantes es claramente de importancia para discriminar un locutor de otro en una forma que mantenga la exactitud de verificación en ambientes que son diferentes del ambiente original de entrenamiento (Shao & Wang, 2008; Li & Huang, 2011; Kinnunen *et al.*, 2012). Diferencias en el ambiente pueden surgir de diversas fuentes que incluyen ruido aditivo de interferencia (Ming *et al.*, 2007) y por variaciones en las condiciones del canal de transmisión sobre el cual la voz está siendo grabada (Wu *et al.*, 2007). La resolución de los desjustes (*mismatches*) entre ambientes de entrenamiento y de prueba se mantiene como uno de los problemas más desafiantes por resolver para que la verificación de locutor sea exitosa en aplicaciones reales (Saedi *et al.*, 2010; Hasan & Hansen, 2013).

Las características más comunmente utilizadas para verificación de locutor han sido coeficientes cepstrales *short-term* tales como los coeficientes cepstrales de frecuencia en escala Mel (MFCC) (Ajmera, Jadhav & Holambe, 2011; Wang *et al.*, 2011; Hanilçi *et al.*, 2012). El método estándar MFCC se desempeña razonablemente bien cuando los ambientes de entrenamiento y prueba están ajustados (*matched*), no obstante, la precisión de verificación se degrada seriamente bajo ambientes ruidosos, en especial, cuando las condiciones de los ambientes de entrenamiento y prueba están desajustadas (Kinnunen *et al.*, 2012; Li & Huang, 2011). La degradación más grande en desempeño de verificación se observa cuando la señal de voz es degradada por ruido aditivo a un baja SNR, especialmente, cuando el sistema se entrena usando voz limpia (Hanilçi *et al.*, 2012; Kinnunen *et al.*, 2012).

La extracción de características inspirada en la fisiología del sistema auditivo periférico se ha propuesto con el fin de mejorar el desempeño de verificación de locutor bajo condiciones desajustadas (*mismatch*) (por ejemplo, Li & Huang (2010, 2011); Shao & Wang (2008); Shao, Srinivasan & Wang (2007). De hecho, Shao & Wang (2008) propusieron características basadas en el sistema auditivo conocidas como coeficientes cepstrales de frecuencia Gammatone (GFCC), las cuales efectivamente reemplazan a la ponderación de frecuencia triangular usada

en el método MFCC por el uso de filtros Gammatone (Shao & Wang, 2008) para conseguir selectividad de frecuencia. Los filtros Gammatone se utilizan ampliamente en modelos del sistema auditivo y se desarrollaron para imitar el filtrado coclear (Patterson, Holdsworth & Allerhand, 1992). Shao, Srinivasan & Wang (2007) demostraron que las características GFCC pueden proporcionar reconocimiento robusto de locutor en presencia de ruido aditivo sobre un rango amplio de SNRs, y además que el desempeño se puede mejorar realizando en forma complementaria un anlisis del escenario auditivo (*auditory scene analysis*) (Shao *et al.*, 2010).

Del mismo modo, Li & Huang (2011) propusieron el uso de coeficientes cepstrales de filtro coclear (CFCC) para la identificación robusta de locutor en condiciones de *mismatch* (Li & Huang, 2010, 2011). Las características CFCC propuestas se basan en una transformación de frecuencia y tiempo, llamada Transformada Auditiva que incluye varias componentes que imitan el procesamiento en el sistema auditivo periférico humano (Li & Huang, 2011). Además, las características CFCC mejoran la exactitud de la identificación de locutor, comparada con el procesamiento estándar MFCC, cuando se prueba bajo condiciones de *mismatch* (Li & Huang, 2010). Otras características recientes basadas en el sistema auditivo, incluye a los coeficientes cepstrales de energía Teager (TECC), propuestos por Dimitriadis, Maragos & Potamianos (2011).

3.1.3. La función sigmoidal tasa-nivel

En un estudio previo Chiu & Stern (2008) examinaron las contribuciones de cada etapa del modelo auditivo clásico presentado por Seneff (1988). Analizaron sus impactos en la mejora de la precisión en experimentos de reconocimiento de voz bajo condiciones de ruido aditivo. Experimentalmente encontraron que la mejora más importante en la exactitud la proporciona la etapa de no linealidad tasa-nivel (Chiu & Stern, 2008). Esta etapa, de acuerdo a la literatura, se incluye en la mayoría de los modelos del sistema auditivo periférico, justo después del filtrado pasa banda (tipicamente lineal), que modela el movimiento de la membrana basilar en la cóclea (Kiang *et al.*, 1965; Johnson, 1980; Robles & Ruggero, 2001). Dicha no linealidad tiene una forma aproximada a una "S" y cuenta con tres regiones importantes (ver por ejemplo, Moore (2003a) su Figura 1.17): (1) un rango de intensidades de entrada que están "bajo el umbral" en el cual la salida de la función es aproximadamente constante a un nivel bajo; (2) un rango de intensidades de entrada para el cual la salida de la función es aproximadamente lineal con respecto a la intensidad de entrada en decibeles; y (3) una región saturada en la cual la salida de la función es aproximadamente constante a un nivel más elavado.

Resultados recientes de estudios auditivos fisiológicos, describen e intentan explicar varios tipos de adaptación dinámica de las funciones tasa-nivel con respecto a la intensidad del sonido de entrada, intensidad del ruido de fondo, y el contraste entre ruido y señal de voz degradada (Dean, Harper & McAlpine, 2005; Zilany & Carney, 2010). Estas adaptaciones

posibilitan que el rango dinámico de las funciones tasa-nivel, el cual es por naturaleza limitado, cubra un rango mucho más amplio de niveles sonoros. En general, elevados niveles de sonido de entrada tienden a mover las curvas tasa-nivel hacia la derecha (es decir, hacia los niveles más elevados), mientras aumentan también sus pendientes máximas (Gao *et al.*, 2009; Bureš *et al.*, 2010). Por ejemplo, en gatos el ruido de fondo produce en las funciones tasa-nivel, un desplazamiento del rango dinámico hacia las intensidades más altas. Además, se ha observado que el nivel de ruido donde se inicia este desplazamiento puede ser dependiente de la frecuencia (Costalupes, Young & Gibson, 1984), y que la pendiente de las funciones tasa-nivel puede aumentar en presencia de ruido (May & Sachs, 1992) junto con el incremento de los niveles de entrada.

Resultados similares en hurones (*ferrets*) han caracterizado el realce del contraste espectro-temporal en el ambiente acústico como otra consecuencia importante de la adaptación de la no linealidad sigmoidal (Rabinowitz *et al.*, 2011; Wang & Shamma, 1994). Este es similar al realce en el contraste espectro-temporal que se produce en la retina de lo vertebrados (Ohzawa, Sclar & Freeman, 1985; Werblin, Jacobs & Teeters, 1996). Como un ejemplo, Rabinowitz *et al.* (2011), decriben el procesamiento auditivo que mejora las fluctuaciones en la envolvente de respuesta a señales deseadas en presencia de ruido. Esto no se puede conseguir por un simple control de ganancia el cual en forma simultánea, amplifica tanto la voz degradada como las componentes de ruido, sino más bien por correspondientes formas de control de ganancia no lineal ajustables que aumentan el rango dinámico de la voz degradada mientras suprime las fluctuaciones producidas por el ruido (Schneider *et al.*, 2011).

Un mayor pronunciamiento de la pendiente de las funciones tasa-nivel de las neuronas auditivas, se ha observado también en respuesta al incremento en el nivel sonoro (Middlebrooks, 2004; Kang *et al.*, 2010; Pfingst *et al.*, 2011) y en respuesta a ruido (Bureš *et al.*, 2010; Gao *et al.*, 2009). De acuerdo a García-Lázaro *et al.* (2009), quienes investigaron neuronas auditivas de ratas, las curvas observadas tasa-nivel, son aproximadamente de forma sigmoidal, con un cambio en la inclinación de la función tasa-nivel interpretado como un cambio en la "ganancia de la respuesta neuronal". Estudios sobre neuronas auditivas de monos titíes (*marmoset monkeys*), demuestran que la pendiente en la función tasa-nivel es una medida de la discriminabilidad del nivel sonoro (Watkins & Barbour, 2011).

Otros estudios sobre las neuronas auditivas, en respuesta a estímulos sonoros continuos, dinámicos, demuestran un desplazamiento horizontal de la función tasa-nivel, trasladando la región dinámica de la función hacia el nivel sonoro promedio, resultando en una precisión superior de codificación de los niveles de intensidad (Dean, Harper & McAlpine, 2005; Wen *et al.*, 2009; Miller *et al.*, 2011; Schneider *et al.*, 2011). Al expandir o comprimir la respuesta auditiva ante un sonido de entrada en varios grados, el control de ganancia de contraste en la audición humana puede servir a dos funciones: (1) proteger el sistema sensorial de sobre-

carga, y (2) mejorar la discriminabilidad entre los estímulos seleccionados (Schneider et al., 2011). Idealmente, la función tasa-nivel aumentaría su pendiente para correspondientemente mejorar el contraste en su respuesta a pequeñas amplitudes de la señal de voz degradada por encima del ruido (bajo contraste entre las señales de voz degradada y ruido). Estos propósitos, en combinación con la reducción de la distorsión no lineal de la voz degradada y la reducción de las diferencias entre la voz original limpia y la voz degradada, descrita en Chiu, Raj & Stern (2012), conducen a la definición de los cuatro criterios que se describirán más adelante.

Con estos ejemplos fisiológicos en mente, Chiu, Raj & Stern (2012), propusieron que la adaptación dinámica de la no linealidad tasa-nivel, podría además mejorar la exactitud de reconocimiento para voz bajo la presencia de ruido. Particularmente, estos autores modelaron la no linealidad tasa-nivel por un conjunto de funciones logísticas dependientes de la frecuencia y desarrollaron un procedimiento que optimizaba los parámetros que especificaban la forma de la no linealidad sigmoidal para un ambiente particular de ruido aditivo, usando una función objetivo basada en la maximización de la discriminabilidad fonética. Estos autores establecieron que el empleo de una función no lineal tasa-nivel, reduce las diferencias entre las formas de las distribuciones espectrales de la voz limpia versus la voz en ruido y demostraron que la adaptación de la no linealidad mejora la exactitud de reconocimiento de voz en presencia de ruido.

En este capítulo, se describe una nueva forma para optimizar la función sigmoidal tasa-nivel, la que se basa en los atributos físicos de la señal acústica, en lugar de la discriminación fonética que era la base del método de Chiu, Raj & Stern (2012). El nuevo esquema intenta discriminar entre la señal de voz degradada y ruido, preservar la máxima información en la región lineal de la curva sigmoide, y minimizar los efectos de distorsiones en la regiones de saturtación. El método propuesto se aplica a una tarea de verificación de locutor de texto-independiente, con señales de voz que se encuentran degradadas por ruido aditivo a diferentes SNRs.

El desarrollo de la adaptación basado en el análisis de señal (en lugar de análisis fonético), es motivado por varias consideraciones. Primero, y la más importante, el entrenamiento discriminativo utilizado por Chiu, Raj & Stern (2012), se basa en reconocimiento de voz a nivel de fonemas para generar la representación fonética verdadera (*ground-truth*) del conjunto de información. Se asume en el presente capítulo, que la optimización de parámetros basada en reconocimiento de voz, puede no ser lo mejor para tareas de voz, otras que no sea esta tarea, tales como, la verificación de locutor, considerada en el presente capítulo. Además, se ha descrito más arriba, que en diversas especies de mamíferos, diferentes del ser humano, así como también en otras modalidades sensoriales, existe una adaptación no lineal tipo sigmoidal, similar a aquellas modeladas en este capítulo. Esto sugiere que se debiera investigar un método viable de adaptación de la no linealidad que esté basado en algo más que la sola

discriminación fonética de los seres humanos.

Desde el punto de vista computacional, el entrenamiento discriminativo descrito en Chiu, Raj & Stern (2012), requiere una cantidad abundante de información *a priori* y aumenta considerablemente el cálculo necesario, comparado con el procesamiento de señal basado en el método descrito en este capítulo. De igual forma, el método basado en el procesamiento de la señal es mucho más flexible a una implementación adaptiva o en línea (*online*), que el método de Chiu, Raj & Stern (2012), en el cual el entrenamiento discriminativo se debe realizar fuera de tiempo real (*offline*) basado en información de entrenamiento.

Una razón no asociada hasta ahora para revisar el concepto de adaptar la no linealidad sigmoidal, es que Chiu, Raj & Stern (2012), desarrollaron sus cálculos usando los mismos parámetros para todos los canales de análisis de frecuencia. En este capítulo se analiza hasta qué punto el desempeño sería mejorado por la adaptación y a la cual se le permitiese variar sobre la base de un canal a otro, lo cual parece razonable dado que el SNR efectivo varía de canal en canal.

En principio, se reafirma que el método propuesto en este capítulo es aplicable a cualquier tarea de procesamiento de voz puesto que todo el análisis tiene lugar a nivel de la señal acústica. Además, las funciones sigmoidales son estimadas separadas para cada canal.

En la próxima sección se describe el método de optimización propuesto y específicamente, el desarrollo de cuatro criterios que optimizan la función sigmoidal tasa-nivel, basados en atributos acústicos de las señales de entrada. Posteriormente, se explica la presente implementación de la no linealidad sigmoidal tasa-nivel, así como también, se describen los resultados experimentales que validan la utilidad del método.

3.2. Desarrollo de los criterios de optimización para la función sigmoidal

En esta sección se analiza el desarrollo de los criterios de optimización para la función sigmoidal tasa-nivel. Se inicia con una especificación matemática de la nolinealidad sigmoidal y posteriormente, se proporciona una descripción matemática de las cuatro componentes de la función objetivo utilizada para optimizar la no linealidad tasa-nivel. Se enfatiza que el propósito de la adaptación es modificar, tanto la ubicación como también la pendiente de la no linealidad sigmoidal, de modo que esta función capture lo mejor posible las fluctuaciones de intensidad de las componentes de voz de la señal en cada canal y mejore el contraste cuando la voz de entrada alcance un alto nivel de degradación debido al ruido aditivo.

3.2.1. Especificación matemática de la función sigmoidal

Se representa la no linealidad tasa-nivel en transducción auditiva por la función sigmoidal $g(l)$ dada por:

$$g(l) = \frac{1}{1 + e^{\omega(l-\mu)}} \tag{3.1}$$

donde μ and ω corresponden a la posición (desplazamiento) y la pendiente de $g(l)$, respectivamente. Esta función permite modelar la respuesta no lineal. El parámetro de posición μ corresponde a la posición a lo largo del eje horizontal en el cual la curva sigmoidal $g(l)$ es igual a $1/2$. La pendiente de $g(l)$ es igual a $-\omega/4$ cuando $l = \mu$. En consecuencia, la posición μ y la pendiente ω de la función sigmoidal son los parámetros que se desea estimar.

Se considera a continuación, la salida de un canal particular del banco de filtros inicial pasabanda que es la primera estapa de todo modelo. Se representa la señal de entrada de voz degradada $x_{j,k}$ a la salida del filtro j en el índice del tiempo k como:

$$x_{j,k} = s_{j,k} + n_{j,k} \tag{3.2}$$

donde $s_{j,k}$ y $n_{j,k}$ representan las señales de voz limpia y de ruido, respectivamente. Si la señal completa $x_{j,k}$ se divide en N_f cuadros (*frames*) de W muestras por cuadro, con $50\,\%$ de traslape, la energía $E_{j,i}$, en escala logarítmica, en el *frame* i, en el filtro j, se puede escribir como:

$$E_{j,i} = 10 \cdot log_{10}\left(\sum_{k \,\in\, frame\ i} w_{i-k}^2 \cdot x_{j,k}^2\right) \tag{3.3}$$

donde w_k la respuesta de la función de ventana de duración finita. Histogramas de las energías-logarítmicas $E_{j,i}$, en el filtro j, en el *frame* i, son generadas para discriminar entre frames de ruido y de voz degradada, utilizando el detector de actividad de voz (VAD) propuesto por Shin et al., (2008), como se analiza más adelante. Por consiguiente, los frames se dividen en dos subconjuntos, uno que se considera que contiene frames de voz degradada, y el segundo subconjunto que representa a los frames que se supone contienen solamente ruido. Se usan los símbolos N_f^{sn} y N_f^n, donde $N_f = N_f^{sn} + N_f^n$, indica el número de frames que se asume contiene voz degradada por ruido, y el número de frames que se supone contiene sólo ruido, respectivamente. Finalmente, se usan los símbolos $E_{j,i}^x$, $(1 \leq i \leq N_f)$; $E_{j,m}^{sn}$, $(1 \leq m \leq N_f^{sn})$; y $E_{j,r}^{sn}$, $(1 \leq r \leq N_f^n)$, para representar las energías en el filtro j, y en los frames i, m y r, para frames que se asumen pertenecen a la entrada original, al subconjunto de frames que contiene voz degradada por ruido y el subconjunto de frames de entrada que contiene sólo ruido, respectivamente. (Se reitera que cada frame de la entrada se clasifica como conteniendo ya sea voz degradada o ruido puro). Además, la media y la varianza de la energía en los frames de voz degradada se definen como $\mu_{j,sn}$ y $\sigma_{j,sn}^2$, respectivamente; en tanto que, las correspondientes medias y varianzas de la energía en los frames que se consideran que contienen sólo frames de energía del ruido, se representan como $\mu_{j,n}$ y $\sigma_{j,n}^2$, respectivamente.

3.2.2. Especificación de la función objetivo usada para optimizar la no linealidad sigmoidal

Basándose en la discusión anterior, se selecciona una función objetivo para la no linealidad sigmoidal que (1) minimiza la distorsión no lineal en la región lineal; (2) minimiza la potencia del ruido; (3) maximiza la similitud entre energía en los frames que se asumen representan voz degradada, y la energía de la voz sola en aquellos frames; y (4) maximiza la energía en la señal de salida que se presume está dominada por voz.

3.2.2.1. Criterio 1: Distorsión no lineal en la región lineal

La pendiente ω y la posición μ se debieran seleccionar de tal manera que la voz degradada permanezca en la parte lineal de la curva sigmoidal. Por lo tanto, una vez que la función sigmoidal se aplique, la distorsión no lineal, en la voz degradada debiera estar minimizada. Esta distorsión no lineal, $D_j^{non-linear}$, se define como:

$$D_j^{non-linear}(\omega_j, \mu_j) = \frac{\mathbf{E}[A_j E_{j,m}^{sn} + B_j - g(E_{j,m}^{sn})]^2}{\mathbf{E}[(E_{j,m}^{sn})^2]} \tag{3.4}$$

donde (como antes) $E_{j,m}^{sn}$ se refiere a la energía de los frames de voz degradada en el índice de frame m, para el filtro j, $g(\cdot)$ representa la función sigmoidal y $\mathbf{E}[\cdot]$ es el operador de esperanza. Los parámetros A_j and B_j corresponden a una transformación lineal que permite la comparación de $E_{j,m}^{sn}$ y $g(E_{j,m}^{sn})$ (como se desarrolla en el apéndice). Al aproximar el valor esperado a la media muestreal, $D_j^{non-linear}(\omega_j, \mu_j)$, se puede reescribir como:

$$D_j^{non-linear}(\omega_j, \mu_j) = \frac{\frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} [A_j E_{j,m}^{sn} + B_j - g(E_{j,m}^{sn})]^2}{\frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} (E_{j,m}^{sn})^2} \tag{3.5}$$

donde N_f^{sn} es el número total de frames que contienen voz degradada.

3.2.2.2. Criterio 2: Potencia del ruido

La función sigmoidal se puede utilizar para atenuar el ruido en la señal de voz, debido al hecho que los frames de baja energía se pueden asociar con ruido. La potencia del ruido, $P_j^{noise}(\omega_j, \mu_j)$, después que éste pasa a través de la no linealidad sigmoidal, $g(l)$, está dada por:

$$P_j^{noise}(\omega_j, \mu_j) = \mathbf{E}[g^2(E_{j,r}^{n})] \tag{3.6}$$

donde $E_{j,r}^{n}$ corresponde a la energía de los frames ruido en el frame r, para el filtro j; $g(\cdot)$ representa la función sigmoidal; y $\mathbf{E}[\cdot]$ es el operador de esperanza. La no linealidad sigmoidal

$g(\cdot)$, debiera minimizar $P_j^{noise}(\omega_j, \mu_j)$ para reducir el efecto de la energía del ruido. Al estimar el valor esperado como la media muestreal, $P_j^{noise}(\omega_j, \mu_j)$, se puede escribir como:

$$P_j^{noise}(\omega_j, \mu_j) = \frac{1}{N_f^n} \sum_{r=1}^{N_f^n} g^2(E_{j,r}^n) \tag{3.7}$$

donde N_f^n es el número de frames que se supone contienen solamente ruido.

3.2.2.3. Criterio 3: Similitud entre la entrada de voz limpia y la voz degradada

De acuerdo a Chiu, Raj & Stern (2012), el uso de una función no lineal tasa-nivel debiera reducir las diferencias entre la respuesta de frecuencia promedio de la voz limpia y la respuesta de frecuencia promedio de la señal de entrada degradada, ambas evaluadas después de la no linealidad sigmoidal. En consecuencia, la diferencia entre la energía de la voz limpia y la entrada de voz degradada, se representa por:

$$D_j^{clean-noisy}(\omega_j, \mu_j) = \sum_{i=1}^{N_f} [g(E_{j,i}^s) - g(E_{j,i}^x)] \tag{3.8}$$

donde $E_{j,i}^s$ y $E_{j,i}^x$, corresponden a la energía de la voz limpia y la energía de la voz de entrada degradada, respectivamente, en el frame i, para el filtro j, y $g(\cdot)$ es la función sigmoidal.

3.2.2.4. Criterio 4: Varianza de la señal de voz degradada por ruido después del procesamiento de la función sigmoidal

Para evitar compresión extrema o saturación, la varianza de la voz degradada resultante después de la función sigmoidal debiera ser maximizada. Esta varianza de la voz degrada $V_j(\omega_j, \mu_j)$, se expresa como:

$$V_j(\omega_j, \mu_j) = \sigma^2[g(E_{j,m}^{sn})] \tag{3.9}$$

donde $E_{j,m}^{sn}$ es la energía de los frames de voz degradada en el frame m, en el filtro j; y $g(\cdot)$ es la función sigmoidal. El expandir la expresión de la varianza, $V_j(\omega_j, \mu_j)$ se puede reescribir de la forma:

$$V_j(\omega_j, \mu_j) = \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} g^2(E_{j,m}^{sn}) - \left[\frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} g(E_{j,m}^{sn}) \right]^2 \tag{3.10}$$

donde N_f^{sn} es el número de frames que contienen frames de voz degradada.

3.2.3. Especificación de la función objetivo completa

Basándose en los cuatro criterios descritos anteriormente, se adopta la función $J(\omega_j, \mu_j)$, que se define como:

$$J(\omega_j, \mu_j) = D_j^{non-linear}(\omega_j, \mu_j) + P_j^{noise}(\omega_j, \mu_j) +$$

$$+ \ D_j^{clean-noisy}(\omega_j, \mu_j) - V_j(\omega_j, \mu_j) \tag{3.11}$$

Por consecuencia, la pendiente óptima, $\hat{\omega}_j$, de la función sigmoidal se estima como:

$$\hat{\omega}_j = \underset{\omega_j}{argmin} \{J(\omega_j, \mu_j)\} \tag{3.12}$$

En (3.12), la posición, μ_j, de la función sigmoidal se establece por $\mu_j = \mathbf{E}[E_{j,m}^{sn}]$, (esto es, centrada sobre la media de la energía de los frames de voz degradados $E_{j,m}^{sn}$).

Finalmente, la posición óptima, $\hat{\mu}_j$, de la función sigmoidal se estima de acuerdo a:

$$\hat{\mu}_j = \underset{\mu_j}{argmin} \{J(\omega_j, \mu_j)\} \tag{3.13}$$

En (3.13), ω_j corresponde a la pendiente sigmoidal óptima $\hat{\omega}_j$.

Aunque se reconoce que la definición de la función objetivo $J(\omega_j, \mu_j)$ como la suma simple de los cuatro criterios presentados anteriormente, es un caso especial de una combinación lineal más general,

$$J(\omega_j, \mu_j) = a \cdot D_j^{non-linear}(\omega_j, \mu_j) + b \cdot P_j^{noise}(\omega_j, \mu_j) +$$

$$c \cdot D_j^{clean-noisy}(\omega_j, \mu_j) - d \cdot V_j(\omega_j, \mu_j)$$

se adopta la función de (3.11) por simplicidad, en ausencia de evidencia convincente que las otras combinaciones de los cuatro criterios proporcionaran mejor desempeño.

3.3. Implementación de la función sigmoidal tasa-nivel

En esta sección se describe el procedimiento adaptivo basado en análisis de señal que se utiliza para optimizar la función sigmoidal tasa-nivel. Se presenta la Fig. 3.1 para una descripción del método completo de extracción de características y la Fig. 3.2 para una descripción del procedimiento para obtener los parámetros óptimos $\hat{\omega}_j$ y $\hat{\mu}_j$. Los valores específicos de los parámetros sigmoidales $\hat{\omega}_j$ y $\hat{\mu}_j$, como se definen en Eqs. (3.12) y (3.13), respectivamente, se determinaron usando una base de datos de desarrollo voz, degradada por ruido *babble* (o balbuceo, en adelante se usa babble) a un SNR igual a 10 dB, como se analiza más adelante.

Los valores óptimos de los parámetros $\hat{\omega}_j$ y $\hat{\mu}_j$, utlizados en el presente capítulo, varían de canal en canal, al contrario del método presentado por Chiu, Raj & Stern (2012), en el cual los parámetros de la no linealidad son los mismos para todos los filtros. Esto es beneficioso

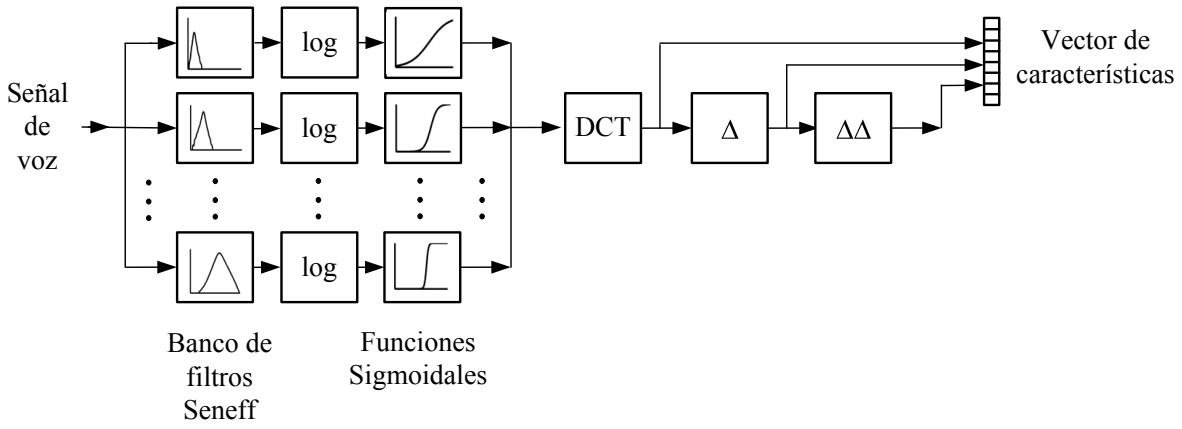

Figura 3.1: *Diagrama en bloques del método de extracción de características propuesto.*

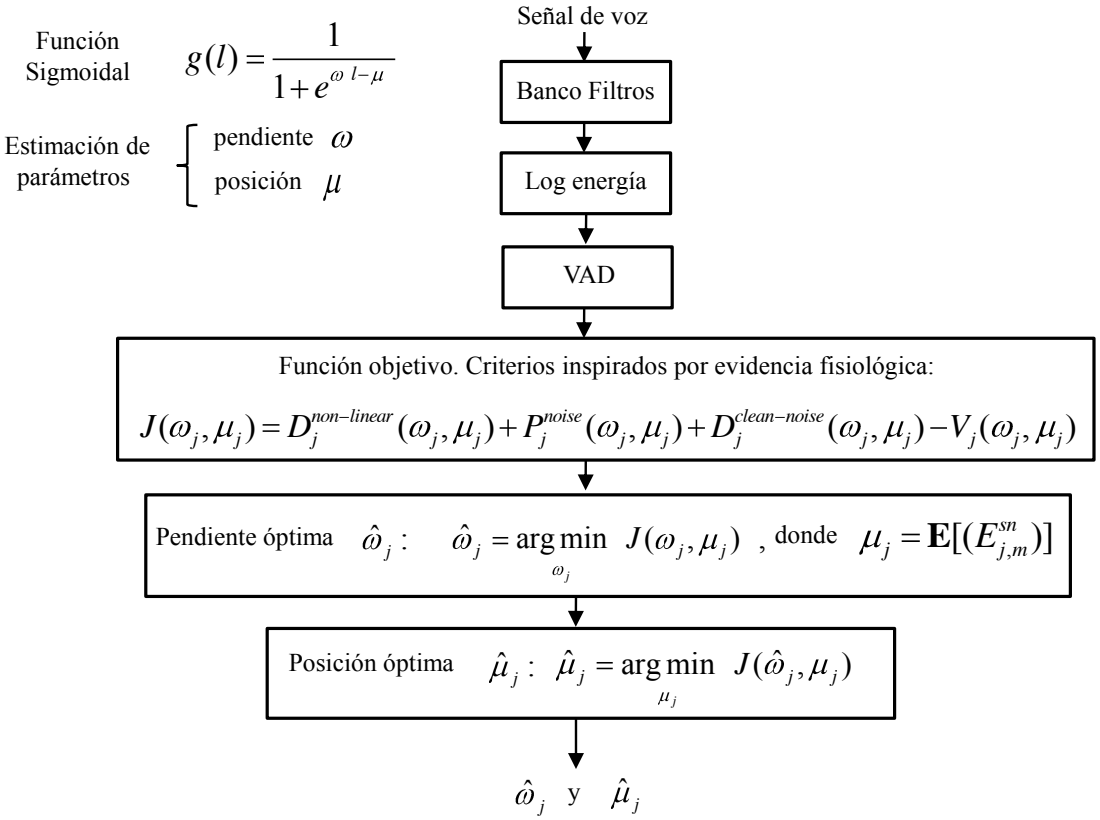

Figura 3.2: *Diagrama en bloques para obtener parámetros óptimos $\hat{\omega}_j$ y $\hat{\mu}_j$ de la función sigmoidal.*

dado que el SNR varía de un filtro a otro. Como se menciona anteriormente, se emplea el detector de actividad de voz (VAD) propuesto por Shin *et al.* (2008), para discriminar entre voz degradada y ruido. Dos subconjuntos de frames se definen basados en los resultados del

VAD, uno que representa a frames que contienen voz degradada, y aquel que representa a los frames que se supone contienen solamente ruido, respectivamente.

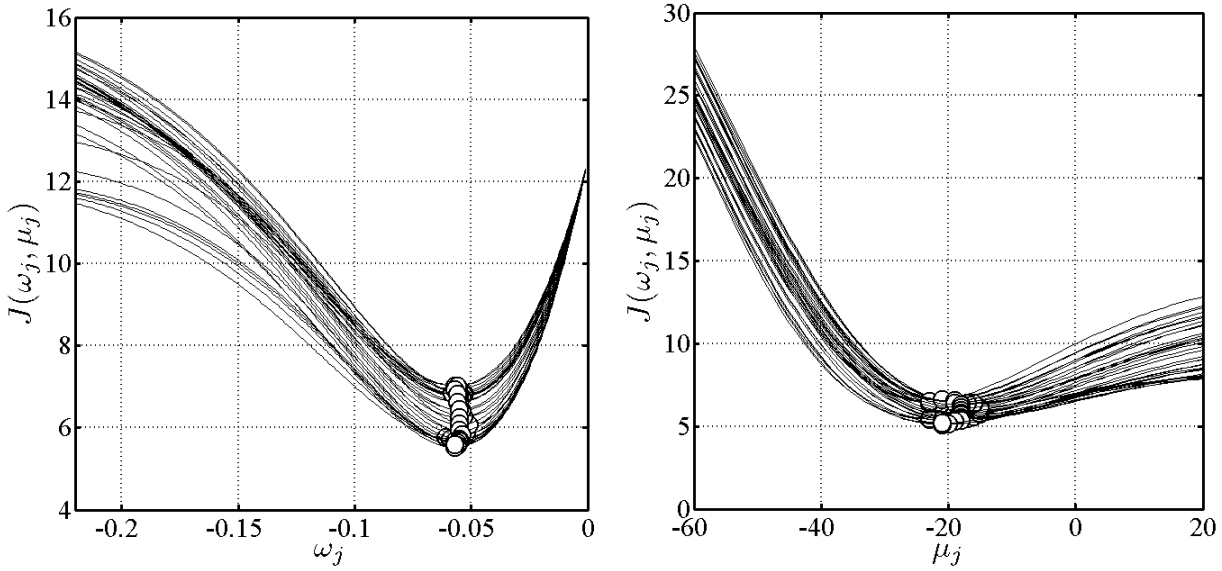

Figura 3.3: *La función objetivo $J(\omega_j, \mu_j)$ dibujada como una función de los parámetros de la función sigmoidal: (a) pendiente ω_j, y (b) posición sigmoidal μ_j. Los valores óptimos de $\hat{\omega}_j$ y $\hat{\mu}_j$ se indican por los círculos abiertos para cada uno de los 35 canales del banco de filtros.*

Fig. 3.3 describe la dependencia de la forma de la función objetivo, de los parámetros ω_j y μ_j, los que especifican la pendiente y la posición, respectivamente, para cada una de las 35 bandas de análisis j. Fig. 3.4 representa un ejemplo representativo de una función sigmoidal óptima (línea sólida), y un mapeo lineal correspondiente (línea punteada) con una pendiente igual a la sigmoide en su punto central. La curva sigmoidal se obtiene con una pendiente óptima $\hat{\omega}_j$ y posición $\hat{\mu}_j$ de la función sigmoidal para el filtro $j = 8$. Fig. 3.4 también representa los histogramas extraídos de una elocución de prueba para el filtro $j = 8$, con ruido babble a SNR igual a 10 dB. Resultados para el filtro $j = 8$ se grafican con parámetros óptimos: $\hat{\omega}_j = -0.071$ y $\hat{\mu}_j = -14$. Al comparar las líneas sólida y punteada, en Fig. 3.4, se puede observar que la función sigmoidal comprime el ruido en la región de no linealidad, mientras la mayor parte de los frames que contienen voz degradada se sitúan en la parte lineal de la función sigmoidal.

Fig. 3.5 muestra cuatro funciones sigmoidales entrenadas con ruido babble a SNRs iguales a 20 dB, 15 dB, 10 dB y 5 dB, junto con una quinta función sigmoidal que se entrena con voz limpia. Resultados para $j = 17$ se grafican para parámetros óptimos. Como se observa en la Fig. 3.5, tanto la pendiente óptima $\hat{\omega}_j$ como la posición $\hat{\mu}_j$ de la función sigmoidal, dependen del SNR en el cual se entrene la función sigmoidal: cuando el SNR aumenta, las curvas

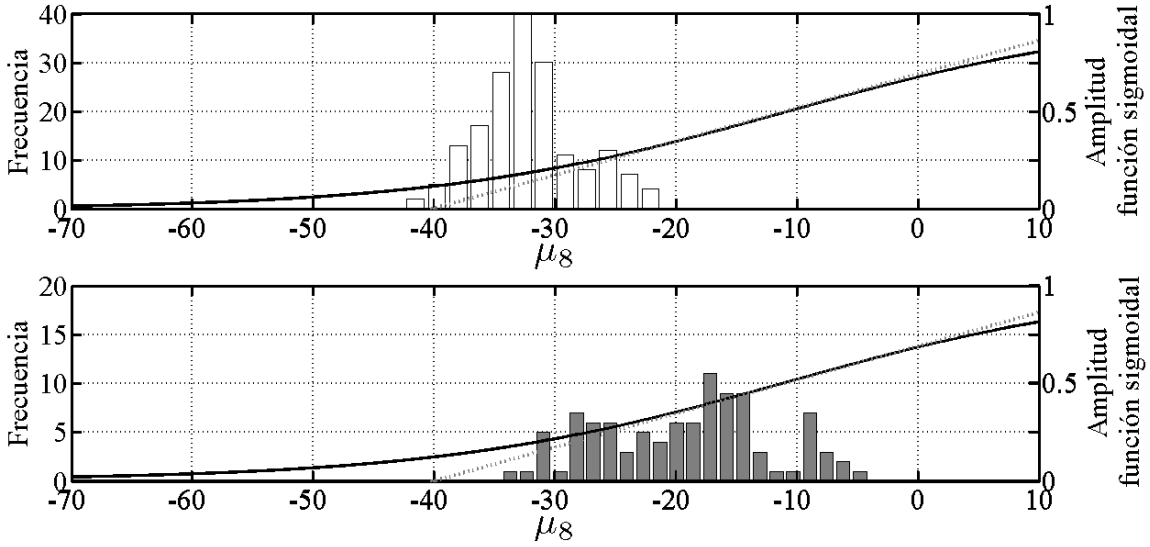

Figura 3.4: *Función sigmoidal óptima (línea sólida) y mapeo lineal correspondiente (línea punteada). Histogramas de potencia se representan para frames de voz degradada (barras rellenas) y frames solamente ruido (barras en blanco).*

en Fig. 3.5 se desplazan hacia la derecha y llegan a ser más empinadas. Por consiguiente, la optimización de la pendiente sigmoidal $\hat{\omega}_j$ y la posición sigmoidal $\hat{\mu}_j$, proporcionan una adaptación en la función sigmoidal que compensa durante las variaciones en SNR.

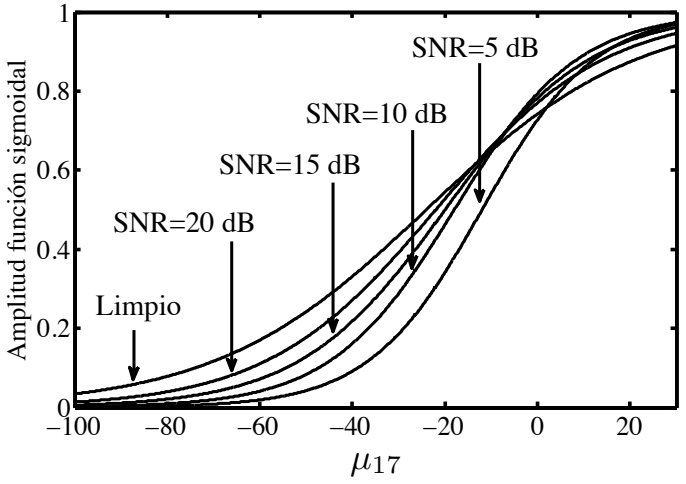

Figura 3.5: *Funciones sigmoidales graficadas como una función de SNR.*

Fig. 3.6 es una combinanción de todas las funciones sigmoidales tasa-nivel, graficadas como una función de SNR con parámetros óptimos para todos los 35 canales y entrenadas sobre voz degradada con ruido babble. Se observa que las funciones sigmoidales se adaptan suavemente para cada canal en cada SNR. Específicamente, cuando el SNR disminuye, las

curvas se desplazan hacia la derecha, y sus pendientes llegan a ser más empinadas en los puntos medios. En conjunto estos fenómenos modifican las no linealidades para asegurar que la mayor parte de la energía de la voz se sitúe sobre la parte relativamente lineal de la curva, lo que representa a los *features* más robustos en contra de cambios en SNR, cuando las condiciones de entrenamiento y prueba están desajustadas (*mismatched*).

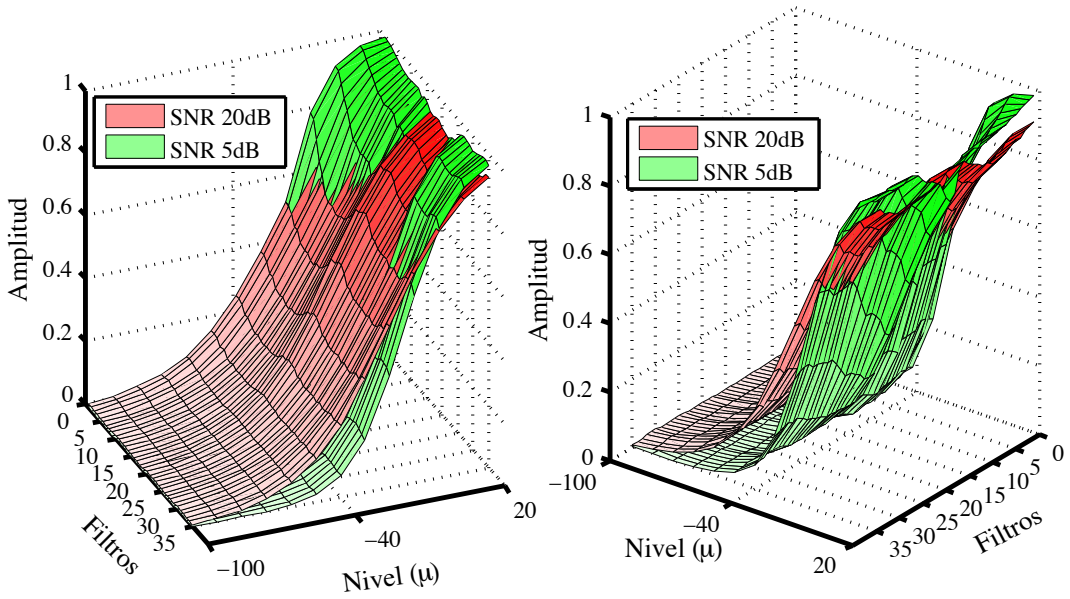

Figura 3.6: *Gráficos tridimensionales de funciones sigmoidales tasa-nivel entrenadas con voz degradada por ruido babble a SNR igual a 20 dB y 5 dB. El gráfico se rota para mostrar diferencias en pendiente y desplazamiento horizontal entre ambos conjuntos de funciones.*

Fig. 3.7 compara un grupo de funciones sigmoidales tasa-nivel entrenadas con diferentes tipos de ruido (ruidos *restaurant* y *car*), pero al mismo SNR. Resultados son similares a las curvas de Fig. 3.6 en que las curvas se desplazan hacia la derecha y sus pendientes aumentan cuando el SNR disminuye, con variaciones en las respuestas individuales observadas de filtro en filtro. Los patrones entrelazados generados por las funciones sigmoidales, entrenadas sobre los dos tipos de ruido, indican que la forma de la no linealidad óptima depende de la distribución espectral del ruido enmascarador.

Esta dependencia de la ubicación y de la pendiente de las curvas sigmoidales en Fig. 3.5, Fig. 3.6 y Fig. 3.7, es consistente con los resultados experimentales en la literatura fisiológica descrita anteriormente. Como se observó, la pendiente de las funciones tasa-nivel de las neuronas auditivas, es congruente con los resultados de numerosos estudios que describen la no linealidad en transducción sensorial (por ejemplo, Middlebrooks (2004); García-Lázaro *et al.* (2009); Gao *et al.* (2009); Kang *et al.* (2010); Bureš *et al.* (2010); Watkins & Barbour (2011); Pfingst *et al.* (2011).

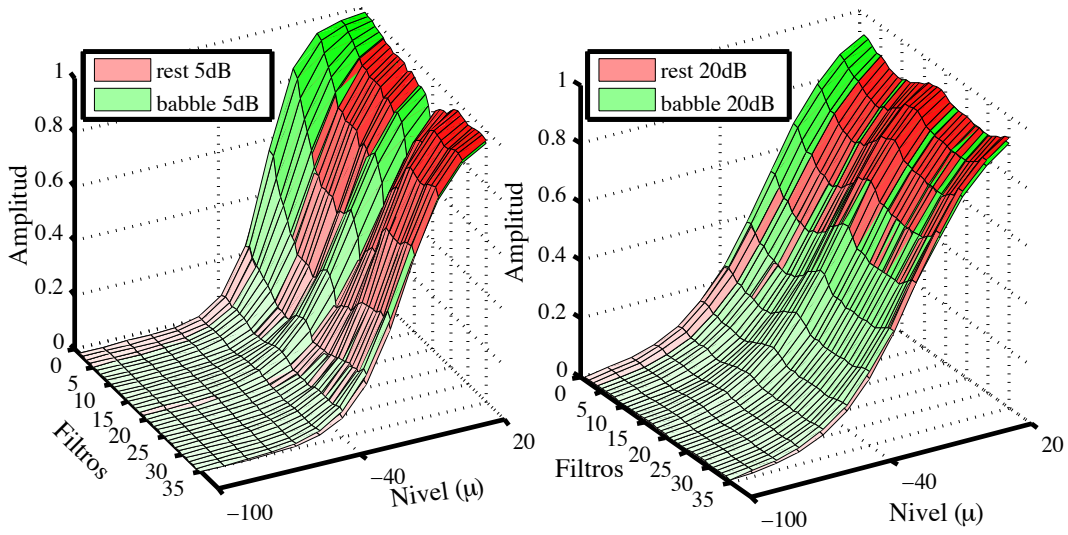

Figura 3.7: *Comparación de funciones sigmoidales tasa-nivel entrenadas con ruidos restaurant y car, a SNR igual a 20 dB (derecha) y a 5 dB (izquierda).*

Se hace notar que en este capítulo, se describe separadamente una etapa de compresión logarítmica, seguida de la no linealidad sigmoidal. Se toma en cuenta la compresión logarítmica debido a que es una aproximación de la ganancia compresiva de la membrana basilar (Robles & Ruggero, 2011). La posterior función tasa-nivel sigmoidal, tiene la intención de representar una aproximación tanto para la existencia de actividad espontánea en el nervio auditivo (AN) a bajos niveles de entrada de sonido, como también representar la saturación de las fibras del AN a elavados niveles de entrada de sonido, probablemente, debido a una combinación de supresión, tal como la supresión de dos tonos (Rhode & Cooper, 1993), como también a un límite presuntamente, sobre la tasa de descarga con la cual un fibra del AN es capaz de codificar la intensidad del sonido.

3.4. Resultados experimentales

La utilidad de la no linealidad sigmoidal óptima, se evaluó empleando una tarea de verificación de locutor de texto-independiente, con la tasa de igual error (EER), utilizada como la principal figura de mérito. Los resultados descritos se obtuvieron al usar la base de datos YOHO (Campbell & Higgins, 1994), la que contiene segmentos de desarrollo, entrenamiento y prueba, de sistemas de verificación de locutor. El vocabulario, en Inglés, se compone de números de dos dígitos hablados contínuamente en conjuntos de tres (por ejemplo, “62-31-53” se pronuncia como “sixty-two thirty-one fifty-three”). La base de datos se divide en secciones de “entrenamiento” y “verificación”. Cada sección contiene información de 138 locutores. En este capítulo, se utiliza un subconjunto de 70 locutores (43 hombres y 27 mujeres). Estos locutores, se dividen en 40 locutores impostores de referencia (28 hombre y 12 mujeres), para entrenar los modelos de referencia; y 30 locutores de prueba (15 hombre y 15 mujeres) se

emplearon en ensayos de verificación. Por cada locutor, se considera una sesión de entrenamiento de 96 elocuciones. Curvas de falso rechazo se estiman con 30 locutores $\times$ 40 señales de verificación por cliente = 1200 elocuciones. Curvas de falsa aceptación se obtienen con 30 locutores $\times$ 29 impostores $\times$ 12 señales de verificación por impostor = 10440 experimentos. Además, un subconjunto compuesto de 50 locutores y una elocución por locutor (base de datos de desarrollo), extraída de YOHO, se usó para entrenar los parámetros óptimos $\hat{\omega}_j$ y $\hat{\mu}_j$, de las funciones sigmoidales. Las elocuciones utilizadas para entrenar la función sigmoidal no se incluyen en la información de prueba durante el experimento principal de verificación de locutor. De la base de datos AURORA se selecionaron tres tipos de ruidos (ruidos de balbuceos de voz (*babble*); automóvil (*car*); y restaurante (*restaurant*)) (Hirsch and Pearce, 2000). Estos ruidos se sumaron artificialmente a la base YOHO para generar versiones ruidosas de las elocuciones en varios SNRs: 20 dB, 15 dB, 10 dB, 5 dB y 0 dB. Durante todos los experimentos de verificación, el sistema se entrena con voz limpia.

En este capítulo, el banco de filtros auditivo (Etapa I en el modelo auditivo de Seneff (1985, 1988)), se obtuvo directamente del Toolbox Auditivo de Malcolm Slaney, ampliamente utilizado (Slaney, 1998), que implementa 35 filtros con frecuencias centrales espaciadas de acuerdo a la escala Bark desde 200 a 3300 Hz. Cada filtro se rediseñó al reducir la tasa de muestreo a 8 kHz. Finalmente, la señal de entrada se normaliza, dividiendo las muestras por la amplitud absoluta máxima. Luego del filtrado, las señales se dividen en frames de 25-ms con 12.5-ms de traslape entre frames, utilizando ventanas Hamming. La energía-logarítmica se calcula a la salida de cada filtro.

A continuación, en cada frame, una función sigmoidal óptima específica de cada canal, estimada utilizando la base de desarrollo y el procedimiento explicado en la Sección 3.2, se aplica a la energía-logarítmica de la salida de cada filtro, tanto en los conjuntos de entrenamiento como también de prueba. Por último, la energía-logarítmica más diez coeficientes cepstrales estáticos, y sus derivadas temporales, primera y segunda, se estimaron de forma similar al procesamiento MFCC (ver Fig. 3.1).

Cuatro configuraciones se consideran: (1) un sistema baseline, corresponde a energías-logarítmicas de la salida del banco de filtros de Seneff; (2) el sistema baseline con normalización de varianza cepstral (CVN); (3) el sistema baseline con normalización cepstral de media y varianza (CMVN); y (4) el método propuesto en este capítulo que utiliza la función sigmoidal, combinada con CVN. Si la función completa se mapea en la región lineal de la función sigmoidal, la estrategia propuesta se podría considerar equivalente al algoritmo CVN. Por lo tanto, el impacto de la no linealidad, proporcionado por la función sigmoidal, puede ser deducido al comparar los resultados obtenidos con las configuraciones (2) y (4), como se ha descrito más arriba.

En el procedimiento de verificación, se estima la verosimilitud-logarítmica normalizada. Dado un intento de verificación, en el que la identidad de un locutor s se demanda, O representa la secuencia de observación correspondiente a la elocución del demandante. La calificación de salida del sistema es una verosimilitud logarítmica de cohorte-normalizada, $\log L(O)$:

$$\log L(O) = \log L(O/\lambda_s) - \overline{\log L(O/\lambda_{\bar{s}})} \tag{3.14}$$

donde $\log L(O/\lambda_s)$ es el logarítmo de la verosimilitud de la hipótesis del cliente, λ_s es el modelo s del locutor, y $\overline{\log L(O/\lambda_{\bar{s}})}$ es el promedio de la verosimilitud logarítmica de la cohorte de modelos impostores. Un modelo universal de referencia, UBM (*Universal Background Model*), es entrenado al utilizar los locutores impostores de referencia. El orden del modelo UBM fue de 256 Gaussianas. Un modelo de mezclas de Gausssianas GMM es generado por cada locutor empleando adaptación MAP (Reynolds et al., 2000). Al hacer esto, se mantiene la correspondencia de las Gaussianas dentro de cada GMM dependiente de locutor, con aquellas en el GMM de referencia (Reynolds, Quatieri & Dunn, 2000).

3.4.1. Dependencia general sobre el SNR y la presencia de no linealidad sigmoidal

La Fig. 3.8 describe resultados proporcionados al utilizar las funciones sigmoidales óptimas para la tarea de verificación de locutor, en presencia de tres tipos de ruido de fondo: babble, car y restaurant, todas como una función en conjunto del SNR al cual las funciones sigmoidales fueron entrenadas, y el SNR de la voz de entrada. Las funciones sigmoidales óptimas se entrenaron con ruido babble y a SNRs iguales a 20 dB, 15 dB, 10 dB, 5 dB y 0 dB. La información obtenida de cada SNR empleado para pruebas, se representa en cada gráfico, con el SNR de prueba indicado en la parte superior. Como se puede observar en la Fig. 3.8, los EERs más bajos se consiguen cuando la función sigmoidal se entrena con 10 dB para todas las condiciones de entrenamiento. Se enfatiza que estos resultados son consistentes con lo encontrado por Chiu, Raj & Stern (2012), donde la función sigmoidal óptima fue entrenada a un SNR de 10 dB empleando un criterio basado en discriminación de fonemas. A diferencia de Chiu, Raj & Stern (2012), la función objetivo $J(\omega_j, \mu_j)$ se basa completamente en características físicas (y en especial, la distribución de potencia), de la voz de entrada, y no toma en cuenta contenido fonético. En consecuencia, el hecho que los mejores resultados de verificación de locutor se consigan con la función óptima entrenada con señales a SNR 10 dB, significa solamente que para esos SNRs los beneficios proporcionados por la eliminación del ruido son más importantes que la distorsión introducida por saturación a niveles más altos. La mayor parte de los resultados que se describen, se efectúan utilizando funciones sigmoidales entrenadas a un SNR igual a 10 dB.

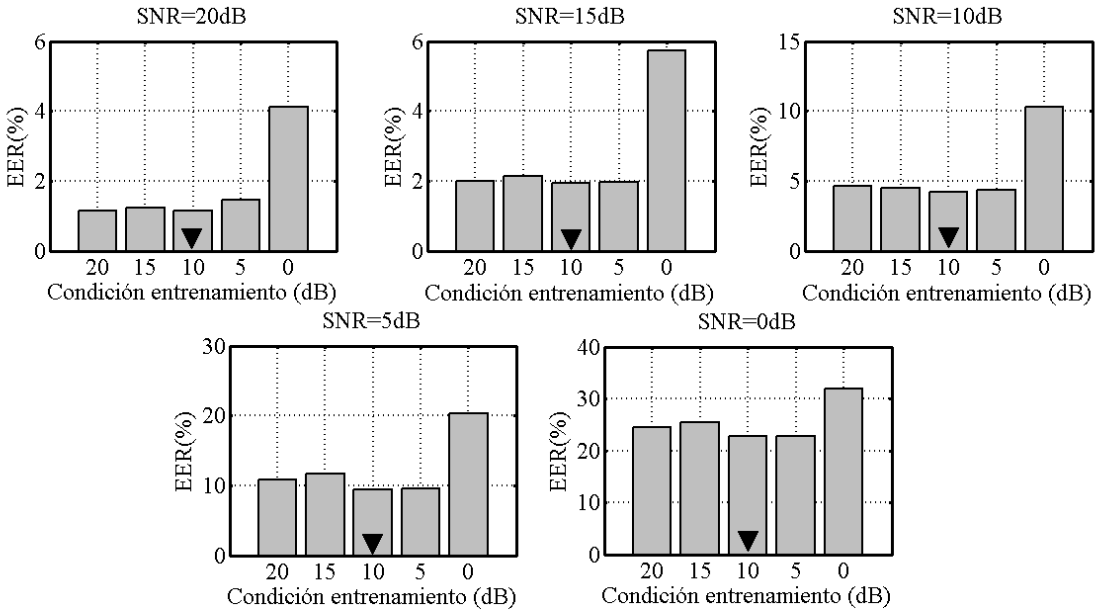

Figura 3.8: *EER durante verificación de locutor como función del SNR de información de prueba y SNR utilizado para desarrollar los parámetros de la función sigmoidal.*

Además, se observa que la función sigmoidal entrenada a 0-dB de SNR, proporciona pobre desempeño en verificación de locutor, con voz en todos los SNRs de prueba. Es posible que ésto sea una consecuencia del hecho que a 0-dB de SNR, las distribuciones de voz y ruido estén efectivamente traslapadas. Por consiguiente, las distribuciones, voz-más-ruido y solo-ruido, no sean separables debido al SNR, y ninguna compresión no lineal, aplicada al ruido, será aplicada también a la voz. A su vez, la estimación de los parámetros óptimos, ω_j y μ_j, por la distribución sigmoidal, es menos confiable.

Fig. 3.9 describe el EER obtenido como una función del SNR para voz en presencia de tres tipos de interferencia de fondo: *babble*; *car* y *restaurant*. Los resultados se comparan para el sistema *baseline* (o de referencia); el sistema *baseline* combinado con CVN; el sistema *baseline* con normalización cepstral de media y varianza; y el método propuesto que combina la no linealidad sigmoidal óptima sugerida y CVN, como se describe más arriba. Las funciones sigmoidales se entrenan y prueban en condiciones ruidosas "igualadas" (matched) a un SNR de 10 dB. El uso de la función sigmoidal óptima, en combinación con CVN, mejora el SNR efectivo del sistema en los tres tipos de ruido, típicamente 1-2 dB. Los porcentajes máximos relativos de mejoras en SNR, comparados al *baseline* en SNRs seleccionados, son aproximadamente 31.7 %, 40.6 % y 28.4 % para los tres tipos de ruidos de fondo. El mejor desempeño siempre se obtiene utilizando la no linealidad sigmoidal óptima propuesta, aunque al menos para ruido car, el desempeño de un sistema con CVN sólo se aproxima.

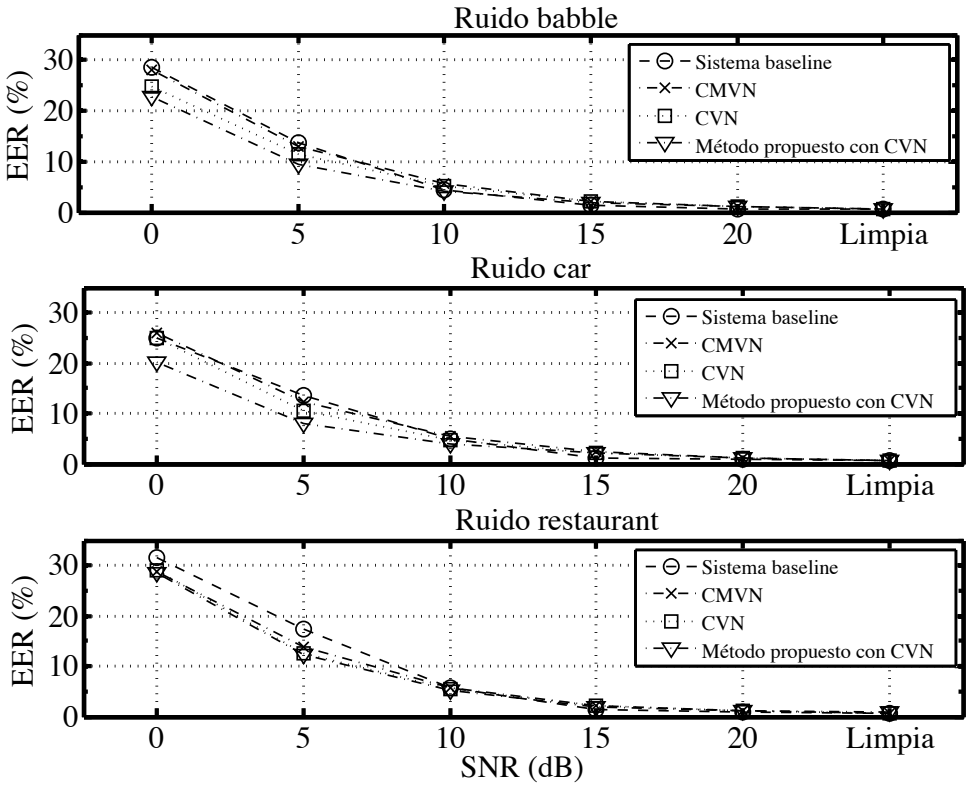

Figura 3.9: *Comparación de EER como una función de SNR para voz en ruido babble, car y restaurant, respectivamente.*

3.4.2. Comparación de entrenamiento de parámetros de no linealidad a SNRs fijos versus *matched*

Como se observó anteriormente, los resultados analizados en Fig. 3.9, se obtienen al estimar los parámetros que caracterizan las funciones sigmoidales usando voz en presencia de ruido *babble*, a un SNR de +10 dB, que parece ser el mejor SNR único de entrenamiento, de acuerdo a los resultados descritos en la Fig. 3.8. Sin embargo, siempre se va a esperar obtener mejores desempeños cuando los parámetros que caracterizan a las funciones sigmoidales, se estimen en condiciones ambientales que se igualen (*matched*) al ambiente de prueba. En un esfuerzo por cuantificar la magnitud del ambiente que se espera, se repitieron algunas de las condiciones con los SNRs para estimación de parámetros igualadas a los SNRs utilizados en los propios experimentos de verificación de locutor. Fig. 3.10 compara EERs para verificación de locutor cuando las sigmoides son entrenadas en el SNR de prueba con los correspondientes EERs obtenidos cuando los parámetros son siempre estimados a partir de señales a 10-dB SNR. En una curva la sigmoide entrenada con SNR igual a 10 dB se aplicó tanto a las elocuciones de entrenamiento como también de prueba. En la segunda curva, elocuciones de entrenamiento y prueba se procesaron con la sigmoide entrenada al mismo SNR utilizado en la elocución. Como se puede ver en la figura, el método proporciona un muy pobre desempeño cuando

los parámetros se estiman en un SNR *matched* (probablemente porque la información es muy ruidosa para entregar estimación confiable de parámetros). Por esta razón, se sigue utilizando sigmoides entrenadas en un SNR de 10 dB para todas las elocuciones, ya que ellas proporcionan desempeños similares a sigmoides entrenadas para que coincida (*matched*) la información de prueba en la mayoría de los SNRs, y sustancialmente, proporcione mejor desempeño en 0-dB SNR.

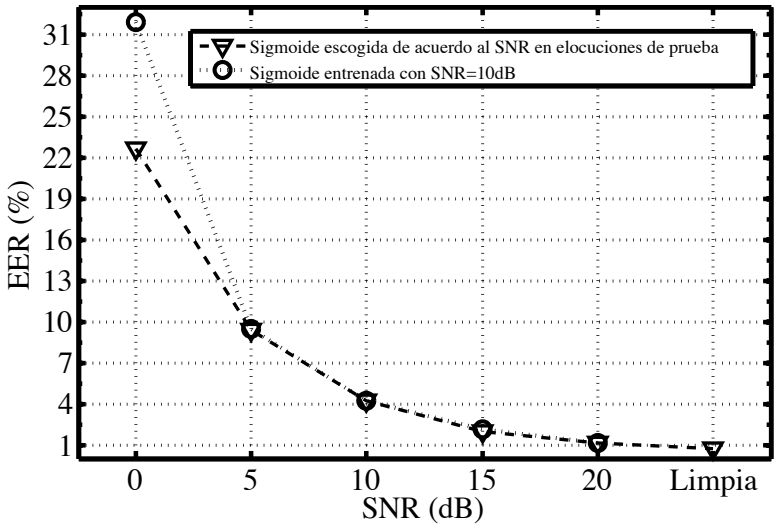

Figura 3.10: *Comparación de EER como una función de SNR para voz en ruido babble, usando la función sigmoidal combinada con CVN.*

3.4.3. Comparación de los resultados de Chiu, Raj & Stern (2012)

Como se analizó anteriormente, Chiu, Raj & Stern (2012) describen un método de adaptación de la función sigmoidal tasa-nivel, que emplea un criterio que se basa en análisis de discriminación a nivel fonético. Fig. 3.11 compara resultados obtenidos empleando el método descrito en este capítulo, con resultados conseguidos usando el método propuesto por Chiu, Raj & Stern (2012), con CVN incluido en la obtención de ambos conjuntos de resultados. Los parámetros obtenidos para la función sigmoidal de Chiu, Raj & Stern (2012), fueron $\alpha = 0.05$; $\omega_0 = 0.613$; y $\omega_1 = 0.521$. Resultados se analizan para tres tipos de ruido: *babble*, *car* y *restaurant*, a SNRs iguales a 20 dB, 15 dB, 10 dB, 5 dB y 0 dB, respectivamente. Los resultados experimentales mostrados en la Fig. 3.11, indican que tanto el método propuesto en este capítulo como la estrategia sugerida por Chiu, Raj & Stern (2012), son efectivas al mantener un buen desempeño casi en todos los SNRs, pero el método propuesto en el presente capítulo se desempeña algo mejor para los tres tipos de ruido en los SNRs más bajos.

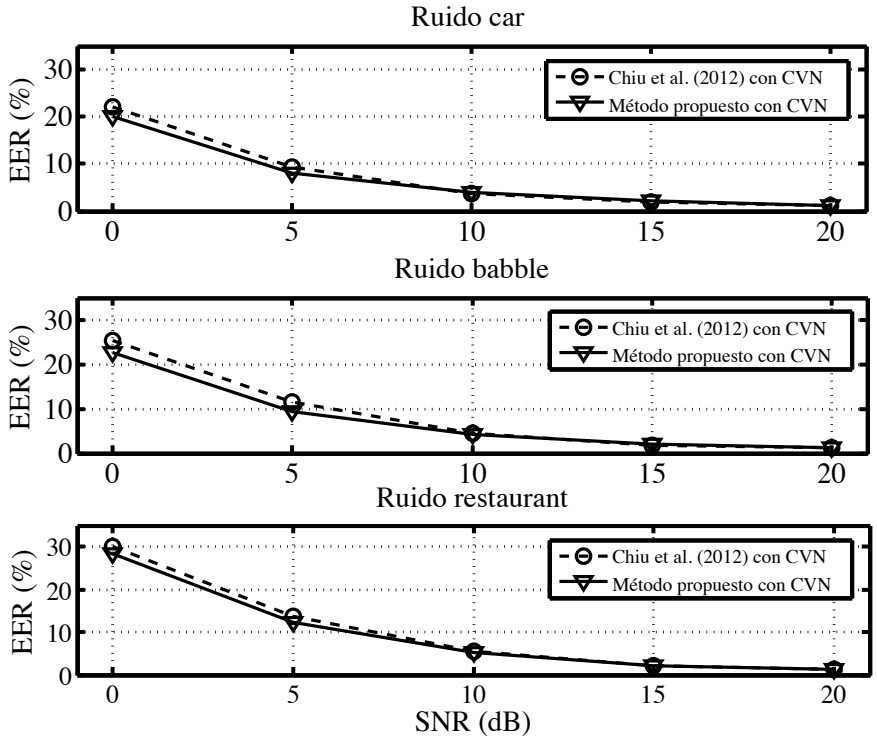

Figura 3.11: *Comparación de EERs(%) conseguidos utilizando el método con parámetros de no linealidad dados por Chiu, Raj & Stern (2012), obtenidos con CVN, y el método propuesto combinado con CVN, para voz en ruido car, babble, y restaurant, respectivamente.*

3.4.4. Impacto de la estimación, específica del canal, de las no linealidades sigmoidales

Fig. 3.12 compara resultados obtenidos usando las no linealidades sigmoidales estimadas sobre una base específica de canal, como se describe en este capítulo, con resultados en que se utiliza una única no linealidad para los 35 canales de frecuencia en total. Se puede observar que el permitir que las no linealidades sigmoidales varíen de un canal a otro, tiene ventaja a SNRs de 0 dB y +5 dB, lo que probablemente se debe a que los SNRs locales muestran una variación mayor de canal en canal, mientras más bajo sean los SNRs. De este modo, a SNRs bajos, el desempeño de la optimización mejora la precisión de verificación de locutor, debido al hecho que la adaptación permite aumentar el rango dinámico de la voz degradada por encima del ruido y minimiza las distorsiones no lineales en la región lineal, mientras suprime fluctuaciones producidas por ruido.

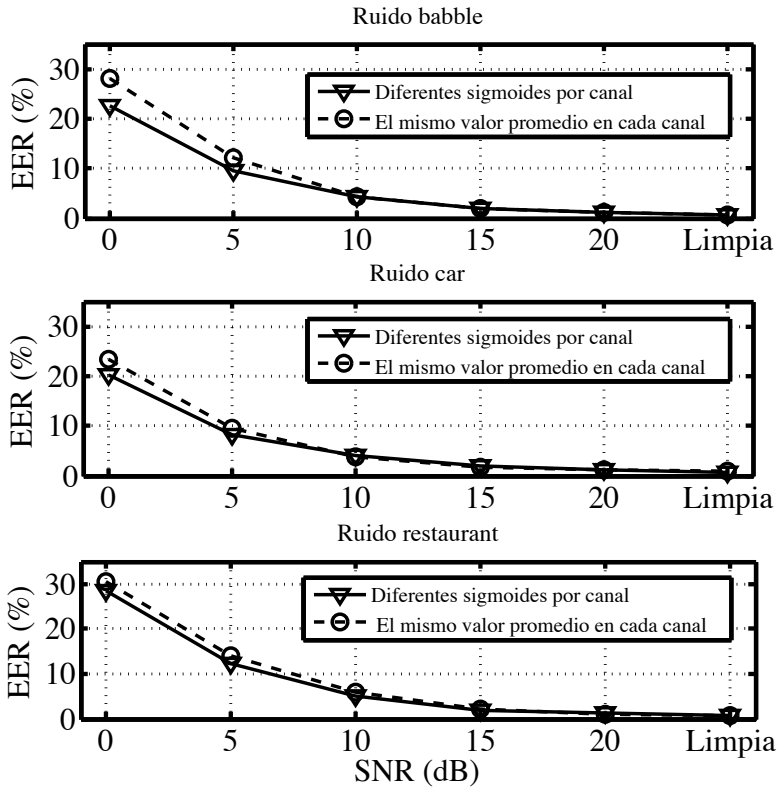

Figura 3.12: *Comparación de EER(%) obtenido utilizando el método propuesto, combinado con CVN, empleando diferentes funciones sigmoidales por canal, y el mismo valor medio para todos los canales, para voz en ruido car, babble y restaurant, respectivamente.*

3.4.5. Comparaciones para funciones sigmoidales óptimas entrenadas y probadas con diferente tipo de ruido

Fig. 3.13 compara resultados obtenidos al utilizar las no linealidades sigmoidales con ruido rosa a SNR igual a 10 dB, con resultados mostrados en la Fig. 3.9, donde el mismo tipo de ruido se empleó en entrenamiento y prueba. Cuando se compara al procesamiento baseline, las funciones sigmoidales entrenadas con ruido rosa, en combinación con CVN, conducen a reducciones relativas promedio en EER iguales a 23.5 % y 13 %, en SNR igual a 5 dB y 0 dB, respectivamente, con ruido car, babble y restaurant. Este resultado valida fuertemente el método de optimización propuesto. No obstante, las reducciones más altas en EER se obtienen cuando las no linealidades se entrenan y prueban, con el mismo tipo de ruido, excepto con ruido restaurant a SNR igual a 0 dB, donde ambas funciones sigmoidales proporcionaron casi el mismo resultado.

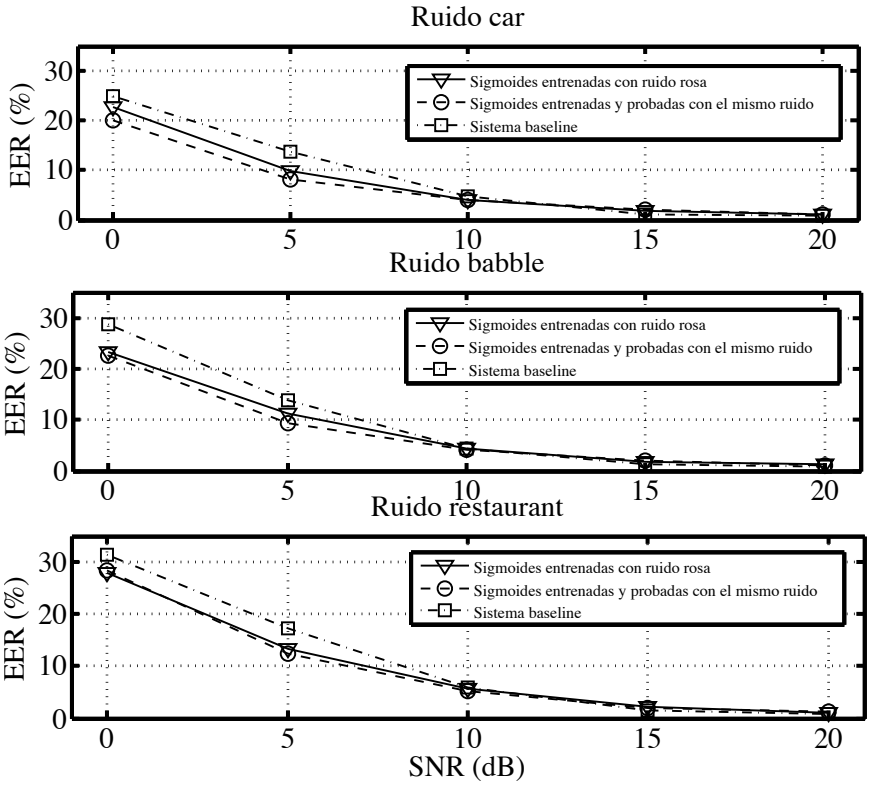

Figura 3.13: *Comparación de EER(%) entre método propuesto con parámetros de no linealidad entrenados (a SNR igual a 10 dB) y probados con el mismo tipo de ruido; y método sugerido usando parámetros sigmoidales entrenados con ruido rosa (a SNR igual a 10 dB).*

3.4.6. Comentarios generales

Las mejoras producidas por la utilización de la función sigmoidal, son consistentes con los resultados de otros estudios en reconocimiento de voz, basados en procesamiento auditivo. Las estrategias en base a la audición, proporcionan por lo general, mejoras relativas importantes en bajos SNRs, pero a SNRs más altos, ellas pueden conseguir desempeño que no es mejor (o incluso peor), que el desempeño que se observa usando procesamientro de señal convencional basado en *features* MFCC o PLP (por ejemplo, Ghitza (1986); Jankowski & Lippmann (1992); Kim, Lee & Kil (1999); Chiu & Stern (2008); Chiu, Raj & Stern (2012). Se reitera que los resultados presentados en este capítulo, son consistentes con aquellos que describen Chiu, Raj & Stern (2012), donde la función tasa-nivel, sigmoidalmente formada, ha sido identificada como una componente importante de los sistemas para reconocimiento de voz con extracción de características basados en la audición.

Sin embargo, el uso de la no linealidad entrenada a 0-dB SNR, falló en producir una mejora importante. Como se analiza más arriba, en 0-dB SNR las distribuciones de potencia de los frames que contienen voz degradada, se traslapan con las distribuciones de potencia de frames que se supone contienen sólo ruido. Además, al parecer el espectro de potencias de las señales

de voz y del ruido de fondo, son más similares en el caso de ruido *restaurant* que en los casos de los otros dos tipos de ruido considerados. Esto se ilustra en la Fig. 3.14, la que muestra el espectro promedio de potencia de 50 elocuciones sacadas de voz limpia y el espectro de potencia de los ruidos *restaurant*, *babble* y *car*, en cada uno de los tres paneles. Los espectros se estiman al utilizar una FFT con 2^{15} puntos. Un filtro promediador se aplica para suavizar los espectros de voz y ruido. Finalmente, para propósitos de comparación, cada espectro se normaliza de acuerdo a su energía. Se observan diferencias el error cuadrático medio (MSE) entre los espectros de voz y ruido, iguales a 58.7, 70.1, y 95.4, para ruidos restaurant, babble y car, respectivamente. Las diferencias correspondientes entre EERs obtenidos usando la no linealidad sigmoidal, combinada con CVN, comparada con el uso de CVN solo, son 0.57 %, 2.1 % y 4.9 %, respectivamente, en SNR igual a 0 dB. Por lo tanto, se cree que la no linealidad sigmoidal falla en mejorar el EER durante la tarea de verificación de locutor, en ruido *restaurant* a 0-dB SNR, debido a que los espectros de voz y ruido, son muy similares, haciendo que las curvas de potencia para voz degradada y ruido se traslapen en la totalidad de las frecuencias.

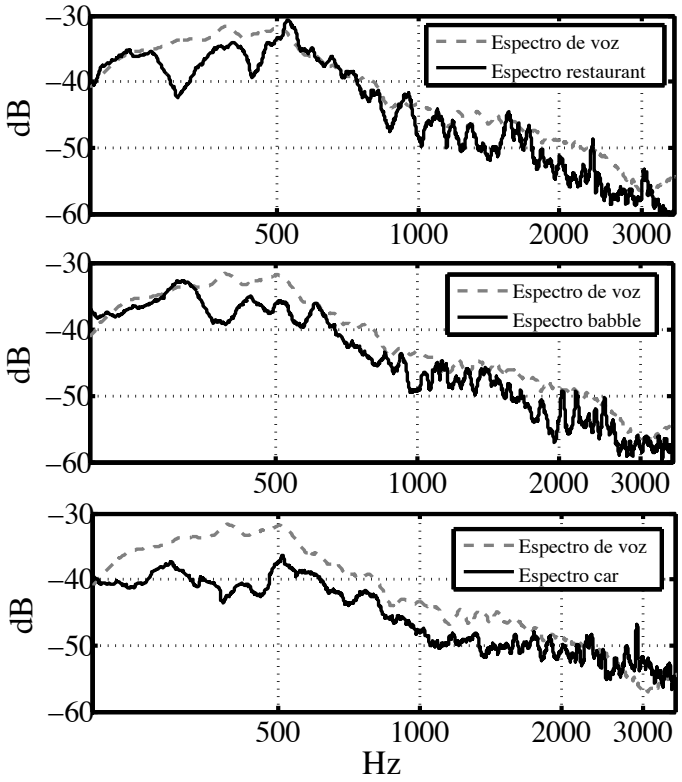

Figura 3.14: *Comparación del espectro promedio de potencia para 50 elocuciones de voz limpia, con el espectro de potencia del ruido. Los errores cuadráticos medio (MSE), entre la voz y ruidos restaurant, babble y car, son 58.7, 70.1 y 95.4, respectivamente.*

3.5. Conclusiones

Este capítulo, describe un método que se puede utilizar para desarrollar una función rectificadora, no lineal, sigmoidal óptima, para modelamiento auditivo, la que se basa únicamente en la distribución de la potencia en los frames de voz degradada y la potencia en los frames que solamente contienen ruido. La función objetivo que se describe, intenta simultáneamente minimizar la potencia del ruido, minimizar la distorsión no lineal, maximizar la similitud entre voz limpia y entrada de voz degradada, y maximizar la varianza de la señal de voz degradada por el ruido, después de ser procesada por la función sigmoidal. Las funciones sigmoidales óptimas obtenidas son dependientes de la frecuencia, dado que el SNR de salida de los canales del banco de filtros pasabanda inicial, varía de un canal a otro. Por último, se observa que el método sugerido difiere de normalización cepstral de media y varianza (CMVN), la cual de hecho, produce una función lineal que relaciona la entrada y la salida, similar a las aproximaciones lineales de la Fig. 3.4. Las mejoras observadas en precisión de verificación de locutor, obtenidas usando la no linealidad sigmoidal óptima (comparada a los resultados obtenidos con CMVN), demuestran el potencial de las no linealidades que son parte del procesamiento auditivo humano.

Se demuestra que las no linealidades sigmoidales resultantes, presentan una ubicación y pendiente las que cambian en función del SNR, en una forma que es consistente con las dependencias correspondientes descritas en la literatura fisiológica. La utilidad de las no linealidades sigmoidales óptimas, derivadas en esta forma, se considera en una serie de experimentos al medir la precisión de verificación de locutor, utilizando la base de datos YOHO. Estos resultados evidencian que el uso de una no linealidad sigmoidal, definida estrictamente a partir de las características físicas de la entrada (en contraposición a discriminación de fonemas), puede conducir a reducciones promedio relativas en EER, comparadas al procesamiento baseline, tan grandes como 12 %, 33.6 % y 16.6 %, en SNR igual a 10 dB, 5 dB y 0 dB, respectivamente, con voz degradada por ruido babble, car y restaurant. La no linealidad sigmoidal produce beneficios más pequeños en SNRs mayores, consistente con experimentos previos con modelos auditivos en reconocimiento de voz. La consistencia de resultados entre las dos estrategias de optimización (empleando discriminación basada en clases de fonemas y discriminación basada en las características de la forma de onda), confirma la idea que las funciones sigmoidales óptimas pueden reducir la desigualdad (*mismatch*), entre la condición de entrenamiento y prueba. En principio, los resultados obtenidos parecen ser genéricos, sugiriendo que este método de optimización puede ser aplicable a cualquier sistema de reconocimiento de sonidos o imágenes, en los que la extracción de características utilice una función no lineal, basada en respuestas tasa-nivel.

3.6. Apéndice

En este Apéndice, se desarrollan los parámetros A_j y B_j del factor de distorsión no lineal $D_j^{non-linear}$, los que se definen en la Sección 3.2.2.

Los parámetros A_j y B_j, se estiman de acuerdo con:

$$(A_j, B_j) = \underset{A_j, B_j}{argmin} \left\{ D_j^{\text{non-linear}} (\omega_j, \mu_j) \right\} \tag{A.1}$$

En primer lugar, se estima la derivada parcial de $D_j^{\text{non-linear}} (\omega_j, \mu_j)$, con respecto a A_j:

$$\frac{\partial D_j^{\text{non-linear}}}{\partial A_j} = \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} 2 \cdot [A_j E_{j,m}^{sn} + B_j - g(E_{j,m}^{sn})] \cdot E_{j,m}^{sn} \tag{A.2}$$

Después, el resultado obtenido en (A.2) se hace igual a cero:

$$\frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} 2 \cdot [A_j E_{j,m}^{sn} + B_j - g(E_{j,m}^{sn})] = 0$$

$$A_j \cdot \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} \left(E_{j,m}^{sn} \right)^2 + B_j \cdot \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} \left(E_{j,m}^{sn} \right) =$$

$$= \frac{1}{N_f^{sn}} \cdot \sum_{m=1}^{N_f^{sn}} E_{j,m}^{sn} \cdot g \left(E_{j,m}^{sn} \right) \tag{A.3}$$

$$A_j \cdot \mathbf{E} \left[\left(E_{j,m}^{sn} \right)^2 \right] + B_j \cdot \mathbf{E} \left[E_{j,m}^{sn} \right] = \mathbf{E} \left[E_{j,m}^{sn} \cdot g \left(E_{j,m}^{sn} \right) \right]$$

De manera similar, al estimar la derivada de $D_j^{\text{non-linear}} (\omega_j, \mu_j)$, con respecto a B_j, y haciendo el resultado igual a cero, se obtiene la siguiente ecuación:

$$A_j \cdot \mathbf{E} \left[\left(E_{j,m}^{sn} \right)^2 \right] + B_j = -\mathbf{E} \left[g \left(E_{j,m}^{sn} \right) \right] \tag{A.4}$$

Al combinar (A.3) y (A.4), y haciendo uso de las expresiones:

$$\mu_j = \mathbf{E} \left[E_{j,m}^{sn} \right] \quad \text{y} \quad \sigma_j^2 = \mathbf{E} \left[\left(E_{j,m}^{sn} \right)^2 \right] - \left\{ \mathbf{E} \left[E_{j,m}^{sn} \right] \right\}^2$$

los parámetros A_j y B_j, se encuentran que son iguales a:

$$A_j = \frac{1}{\sigma_j^2} \cdot \left\{ \mathbf{E} \left[E_{j,m}^{sn} \cdot g \left(E_{j,m}^{sn} \right) \right] - \mu_j \cdot \mathbf{E} \left[g \left(E_{j,m}^{sn} \right) \right] \right\},$$

$$B_j = \mathbf{E} \left[g \left(E_{j,m}^{sn} \right) \right] - \mu_j \cdot A_j \tag{A.5}$$

Capítulo 4

Coeficientes cepstrales basados en la normalización local del espectro

En este capítulo se propone un nuevo conjunto de *features* llamados Coeficientes Cepstrales Localmente-Normalizados (LNCC), que se basan en el Detector de Seneff de Sincronía Generalizada (GSD). Es importante enfatizar que en este capítulo se consideran las siguientes hipótesis formuladas anteriormente:

H1) El uso de modelos del sistema auditivo periférico en tecnologías de verificación de locutor no ha sido suficientemente explorado. Las propiedades del sistema auditivo demuestran un alto potencial y aplicabilidad.

H3) El principio de *phase-locking* en el nervio auditivo puede dar robustez a un sistema de verificación de locutor bajo condiciones de variación en el canal de transmisión.

En consecuencia, dadas estas hipótesis y considerando los principales objetivos planteados al inicio de esta Tesis, en primer lugar, se proporciona un análisis de la respuesta de frecuencia del GSD para demostrar que éste produce peaks espurios en armónicos de la frecuencia detectada. A continuación, la respuesta de frecuencia del GSD se modela como un cociente de dos filtros en la frecuencia detectada. El numerador es un filtro pasabanda triangular alrededor de una frecuencia particular, similar a los filtros tradicionales Mel. El término del denominador, es un filtro que responde en forma máxima a componentes de frecuencia sobre ambos lados del filtro numerador. Como resultado, se realiza una normalización local sin los peaks espurios del GSD original. Resultados de verificación de locutor demuestran que los *features* propuestos LNCC requieren baja complejidad computacional y compensan más ampliamente el *tilt* espectral que los coeficientes MFCC convencionales. Los *features* LNCC no requieren el cálculo y almacenamiento de un promedio móvil de valores de *features*, y ellos proporcionan reducciones relativas en EER tan altas como 32 % y 35 % cuando se comparan con MFCC y MFCC+CMN, con *tilt* espectral variable, respectivamente.

4.1. Introducción

4.1.1. Motivación

La utilización de *features* perceptualmente-motivados, se encuentran ampliamente distribuidos a través de la tecnología de lenguage hablado, con escalas de frecuencia no lineales y compresión del rango dinámico de la energía espectral (por ejemplo, al tomar el logaritmo o la raíz cúbica de la salida del banco de filtros). En reconocimiento automático de voz (Gales, 1998; Tchorz & Kollmeier, 1999; Kim, Lee & Kil, 1999), diarización de locutor (Tranter & Reynolds, 2006), y en verificación de locutor (Reynolds, 1995; Campbell, 1997), los Coeficientes Cepstrales de Frecuencia en escala Mel (MFCCs) (Davis & Mermelstein, 1980), o los coeficientes de Predicción Lineal Perceptual (PLPs) (Hermansky, 1990), son *features* ampliamente utilizados, así como también en *features* estadísticos paramétricos de síntesis de voz en escala Mel (Tokuda *et al.*, 2000).

Por supuesto, el sistema auditivo humano realiza operaciones por lejos más complejas que la transformación de escala de frecuencia y compresión del rango dinámico, pero estas operaciones se encuentran con menos frecuencia en aplicaciones de procesamiento de voz. En este capítulo, se explota una propiedad a veces ignorada de muchos modelos del sistema auditivo: su potencialidad de producir representaciones que son relativamente invariantes a cambios en el canal. Se inicia este capítulo con el punto de partida en el modelo auditivo de Seneff (1985, 1988) y sus dos representaciones paralelas en el nervio auditivo, las cuales no tienen interacción entre sí. La primera de estas dos representaciones es la tasa-media instantánea de descarga de neuronas en fibras individuales en el nervio auditivo (Kiang *et al.*, 1965; Young & Sachs, 1979; Sachs & Young, 1979; Tchorz & Kollmeier, 1999), cuya contraparte es la usual envolvente espectral, utilizada en los convencionales *features* de voz como son los MFCCs (Davis & Mermelstein, 1980), donde se implementa como un banco de filtros (Bimbot *et al.*, 2004). La segunda representación captura la sincronía y se piensa que ésta es menos variante en presencia de ruido (Young & Sachs, 1979; Johnson, 1980; Sachs & Young, 1980; Seneff, 1984, 1985, 1988; Engel, Fries & Singer, 2001; Ali, Van Der Spiegel & Mueller, 2002; Young, 2008; Kayser *et al.*, 2009), y posiblemente también ante cambios en el canal de transmisión (Rosen, 1992; Watkins & Makin, 1996; Tchorz, Kleinschmidt & Kollmeier, 1996).

En las próximas secciones se presenta el desarrollo de la idea propuesta en este capítulo. Esta idea se inspira en ciertas propiedades teóricas y en el comportamiento empírico observado en el modelo de Seneff de Detector de Sincronía Generalizada (GSD) (Seneff, 1984, 1988). Se propone aquí un tipo de normalización local de energía espectral para compensar variaciones en la respuesta de frecuencia del canal acústico de transmisión. Se identifica y se ofrece una solución para un problema potencial con el comportamiento del modelo GSD de Seneff, solución que puede explicar por qué ciertos intentos previos para usar este modelo

directamente en aplicaciones de reconocimiemto de voz, mostraron sólo mejoras limitadas en precisión (Jankowski & Lippmann, 1992; Ohshima & Stern, 1994; Jankowski, Vo & Lippmann, 1995; Ali, Van Der Spiegel & Mueller, 2000, 2002; Kim, Chiu & Stern, 2006; Stern & Morgan, 2012a). Más adelante se presenta un resumen de estas aplicaciones.

Además, se explica cómo el método propuesto de extracción de características se puede realizar dentro de una etapa de un procedimiento típico basado en frames, con muy bajo costo computacional. La extracción de características que se propone es sin memoria y no involucra retardo de tiempo o *look-ahead*, y por lo tanto, no agrega ninguna latencia al sistema. En principio, los *features* resultantes no necesitan ninguna alteración para modelos estadísticos aprendidos a partir de ellos.

Para demostrar la efectividad de los *features* aquí propuestos, se presentan resultados para una tarea de verificación de locutor, texto-independiente, en la cual se observa que los *features* "auto-normalizados" son capaces de compensar variaciones en la respuesta de frecuencia del canal acústico, en forma casi tan efectiva como los *features* tradicionales (MFCCs) que se utlizan en combinación con la técnica estándar de normalización de canal Cepstral Mean Normalization (CMN) (Furui, 1981). Sin embargo, CMN para su buen funcionamiento, requiere la estimación de modo confiable de la media cepstral en la vecindad de cada frame que está siendo normalizado, por ejemplo, la elocución actual completa. En el caso del método propuesto, éste realiza una normalización instantáneamente dentro de cada frame, sin ninguna referencia externa.

La eliminación del requerimiento que la media cepstral local deba ser estimada, es una ventaja en aplicaciones donde el canal acústico puede variar en forma muy rápida. En esos casos, puede llegar a ser difícil escoger la dimensión de la ventana deslizante sobre el cual se debe estimar el valor medio (por ejemplo, Hsu & Lee (2009)). Además, este hecho conduce a una implementación frame a frame, simple y conveniente, que puede ser atractiva en algunas situaciones. Se proporcionan resultados experimentales que demuestran que el método propuesto puede llegar a ser tan competitivo como CMN en ciertos escenarios probados, y superior en el caso de canales acústicos que cambian en el tiempo en forma rápida.

4.1.2. La necesidad de *features* robustos de voz

En este capítulo, el foco de atención se restringe a tratar con canales (mayormente lineales), cuya respuesta de frecuencia puede diferir entre condiciones de entrenamiento y prueba, que pueden variar de una elocución de prueba a la siguiente, o de hecho, dentro de una misma elocución, desconociéndose además en qué tiempo de la prueba. El propósito es extraer *features* a partir de la señal de voz que sean robustos, con esto se quiere dar a entender, *features* que sean invariantes a cambios en el canal (por ejemplo, Wang, Kitaoka & Nakagawa (2007); Wölfel & McDonough (2009c); Hori *et al.* (2012)). Específicamente, el propósito se centra en

las variaciones en la respuesta de frecuencia, por ejemplo, de canales acústicos que surgen a consecuencia de la posición relativa entre el locutor y el micrófono (Nakano, Nakagawa & Yamamoto, 2010).

4.1.2.1. Canales variables en el tiempo

Un gran número de métodos se han descrito en la literatura para mejorar la robustez de sistemas de reconocimiento automático de locutor y de voz, bajo condiciones de canales acústicos variables en el tiempo. No se intenta hacer un resumen de estos métodos en este capítulo, pero se puede orientar, por ejemplo, mencionando los trabajos de Seltzer, Raj & Stern (2004); Buchner, Benesty & Kellermann (2005); Morales *et al.* (2009); Meyer & Kollmeier (2011), y también Lu, Unoki & Nakamura (2011). Normalmente, tales métodos intentan mejorar la precisión de reconocimiento para casos donde la información de entrenamiento y prueba ha sido adquirida bajo condiciones acústicas diferentes, por ejemplo, para que el sistema sea capaz de abordar el problema de cambios de micrófonos. Ciertos métodos proponen extraer *features* invariantes, mientras que otros, intentan ajustar el modelo estadístico. El método propuesto en el presente capítulo es del primer tipo, pero en principio éste podría ser combinado con técnicas de compensación de modelos.

4.1.2.2. Escenarios de aplicación

En numerosas aplicaciones reales, el canal entre el locutor y el sistema de reconocimiento automático de voz (o bien, de verificación de locutor, de diarización de locutor, ...), puede variar en el tiempo. A continuación, se mencionan unos pocos ejemplos de dichas aplicaciones.

Transcripción de reuniones. La tarea de transcribir en forma detallada, interacciones hombre-hombre, ha recibido considerable atención durante las últimas décadas (Hori *et al.*, 2012; Yokoyama *et al.*, 2013), particularmente para el escenario de pequeñas reuniones de negocio con alrededor de cuatro participantes (Hain *et al.*, 2006, 2007; Renals, Hain & Bourlard , 2007; Hain *et al.*, 2012). Un problema fundamental en este dominio es tratar con micrófonos que se ubican en forma distante del locutor, tales como aquellos sobre la cubierta de una mesa, en dispositivos portátiles aleatoriamente ubicados, o que comprenden arreglos de micrófonos. Las tareas que se realizan usando voz capturada bajo tales condiciones, varían en un rango que va desde detección de voz, diarización de locutor y transcripción de palabras, hasta análisis a nivel superior de vinculación entre contenidos (Sangwan *et al.*, 2013; Malionek *et al.*, 2013).

En este dominio de tareas como las que se describen anteriormente, el uso de arreglos de micrófonos es amplio, debido a la gran capacidad que tienen para direccionar los ejes de los micrófonos y así aislar en algo, la señal de un locutor objetivo, ya sea de la voz de otros locutores o bien, de otras fuentes de ruido. No obstante, las propiedades físicas del canal acústico entre el locutor y el micrófono (o arreglo de micrófonos), aún son altamente

variables causando degradación de la señal de voz, por ejemplo, en la precisión de las transcripciones. Las fuentes de variabilidad en este canal incluyen aspectos tales como la variación en el tiempo de la distancia entre el locutor y el micrófono, así como también, la oclusión del camino directo entre el locutor y el micrófono, cuando intervienen objetos como pantallas de computadores o cualquier otro obstáculo como una puerta o pared divisoria (ver por ejemplo, Wölfel & McDonough (2009c), en su Figura 1.1). Se presenta en este capítulo una componente para este complejo problema y, como se justificará en las próximas secciones, se modela la situación como una inclinación espectral o *tilt* espectral, potencialmente variable en el tiempo, impuesto sobre las grabaciones de prueba.

Transcripción de clases. Otra tarea que ha recibido un nivel creciente de atención, es la transcipción de clases (Trancoso, Nunes & Neves, 2006; Bell *et al.*, 2013). Por lo general, esta tarea se realiza a partir de grabaciones hechas con micrófono de solapa, lo que se prefieren por ser relativamente más discretos, comparados con los de tipo auricular con micrófono de proximidad a la boca. Desafortunadamente, el uso de estos micrófono de solapa conduce a cambios muy frecuentes y rápidos en el canal acústico entre el locutor y el micrófono, debido a la cabeza puede estar girando y en constante movimiento. Mientras que, en buenas condiciones, son posibles tasas de error aceptablemente bajas, cuando estas condiciones acústicas se degradan, la Tasa de Error de Palabra (*WER*) puede aumentar 40-45 % (Leeuwis, Federico & Cettolo, 2003; Park, Hazen & Glass, 2005; Hsu & Glass, 2006; Glass *et al.*, 2007). La alternativa a micrófonos de solapa es emplear micrófonos, o arreglos de micrófonos, distantes, pero éstos se encuentran sujetos a problemas similares a los descritos anteriormente.

Interacción hombre-máquina. Las dos acciones tan cotidianas como hablar y escuchar, suelen ocurrir en situaciones donde el ambiente acústico no es constante. También, los locutores se pueden ver afectados por la entrada auditiva del ambiente, por las voces de otros locutores, así como también, por la realimentación de sus propias voces (Cooke *et al.*, 2013b). Más aún, el ruido de fondo produce que los propios locutores ajusten el nivel de intensidad de su voz en una amplia variedad de maneras, (por ejemplo, ver Cooke *et al.* (2014); Cooke, Mayo & Valentini-Botinhao (2013a) para una revisión completa), que incluyen la tan conocida voz "lombard" en la cual uno de los cambios principales, además de aumentar la intensidad, es una reducción en el *tilt* espectral, que conduce a un espectro global más plano, comparado con el espectro normal (Cooke & Lecumberri, 2012). También, con frecuencia los locutores y los oyentes se pueden encontrar en movimiento relativo entre sí, donde cada uno de ellos puede estar ajustando en forma continua su estilo de habla y la posición de su cabeza, para intentar compensar los cambios en el canal acústico.

Las máquinas que escuchan, sean estas robots móviles socialmente interactivos, que pueden operar en espacios públicos, tales como supermercados, museos y exposiciones (Jensen *et al.*, 2005; Ishi *et al.*, 2010), o bien, los tradicionales sistemas estáticos que utilizan

arreglos de micrófonos, orientados según la direccionalidad del eje de los micrófonos (*beamforming*) (Wölfel & McDonough, 2009c), se ven enfrentados a los mismos problemas de canales variables en el tiempo y a estilos diferentes de locutor. Por ejemplo, como se ha mencionado, el *tilt* espectral de la voz del locutor variará tanto con su esfuerzo al hablar, como también, debido a los cambios de distancia entre el locutor y el "oyente" (un robot o un arreglo de micrófonos). Asimismo, la respuesta de frecuencia de un micrófono direccional variará (por lo general, con un aumento del *tilt* espectral debido al filtrado pasa bajos), cuando el locutor se ubique fuera del eje, comparada con la respuesta de frecuencia si el locutor se encuentra en el propio eje. En las próximas secciones este efecto se verificará experimentalmente.

4.1.3. Alcances de este capítulo

Los *features* basados en el modelo auditivo que se introducen en las próximas secciones, se diseñan específicamente para que sean fundamental e instantáneamente, robustos a la respuesta de frecuencia de un canal acústico desconocido y potencialmente variable en el tiempo, presente en aplicaciones tan diversas como las que se analizaron anteriormente. Consecuentemente, este capítulo se limita a la investigación experimental para esos escenarios y no se orienta a otros aspectos de robustez, tales como ruido aditivo y reverberación.

4.2. Desarrollo del método propuesto a partir de un modelo auditivo

Los modelos del sistema auditivo intentan capturar varios comportamientos del sistema natural que ellos están imitando (Stern & Morgan, 2012b). Algunos de estos comportamientos pueden ser útiles para extracción de características de voz, por lo que en este sección, se presenta la motivación a los *features* propuestos, comenzando desde los modelos auditivos. Se identifica un comportamiento que actúa como una normalización, localizada e instantánea, y que no está actualmente como parte de los *features* convencionales de voz, inspirados perceptuamente, que se utilizan en aplicaciones de reconocimiento de patrones.

Un problema de los *features* convencionales, tales como MFCCs, o PLPs, es que ellos capturan no sólo importantes características de la voz, tales como frecuencias de formantes, sino que además, propiedades del canal, como el *tilt* espectral global (Hansen & Varadarajan, 2009a; Kinnunen & Li, 2010). Por supuesto, un amplio arreglo de técnicas de robustez a ruido está disponible para ser aplicadas, ya sea a estos *features*, o a los modelos aprendidos de ellos (Li & Huang, 2010, 2011; Lei & Hansen, 2011; Kumar, Kim & Stern, 2011; Kinnunen *et al.*, 2012). Los *features* que se proponen son menos variantes a diferencias de canal que los MFCCs.

4.2.1. Modelamiento auditivo

En tecnologías de voz los *features* más ampliamente utilizados son representaciones de la envolvente del espectro de potencia (Reynolds & Rose, 1996; Wölfel , 2009a,b). Por otro lado, en modelamiento auditivo periférico, se sabe por años que el sistema auditivo hace uso no sólo de la envolvente espectral, sino que también de información relacionada a la sincronía entre las respuestas en diferentes fibras del nervio auditivo (Johnson, 1980; Sachs, 1984; Seneff, 1988; Eggermont, 1998; Dreyer & Delgutte, 2006). Esta información relacionada con la sincronía es más invariante a diferencias de nivel de señal que la representación tradicional tasa de descarga de la energía espectral (Young & Sachs, 1979; Sachs & Young, 1979, 1980; Young, 2008). Además, a través de la sincronía es posible capturar señales periódicas, incluso en presencia de ruido (Smith, Delgutte & Oxenham, 2002; Moore, 2003b; Heinz & Swaminathan, 2009). Esta notable capacidad sistema auditivo periférico de poder representar información espectral, bajo una amplia variedad de condiciones de audición (por ejemplo, ruido aditivo o distorsión de canal), es uno de los fundamentos que explica la increible robustez del sistema auditivo (Costalupes, Young & Gibson, 1984; Delgutte & Kiang, 1984; Evans, 1992; Moore, 2003b; Kitano, 2004; Young, 2008; Darwin, 2008).

4.2.1.1. Representaciones de tasa-media y la envolvente espectral

La mayor parte de las estrategias de extracción de *features* (tales como coeficientes MFCC y PLP), se basan en energía en segmentos de corta duración (*short-term*) en un conjunto de bandas de frecuencia, estrategia que está más directamente relacionada a la tasa-media que a la sincronía temporal en la respuesta fisiológica del sistema auditivo (Davis & Mermelstein, 1980; Hermansky, 1990, 1994; Dimitriadis, Maragos & Potamianos, 2011). Por ejemplo, el banco de filtros en escala Mel, a partir del cual se derivan los MFCCs, captura solamente la envolvente espectral (Davis & Mermelstein, 1980). Obviamente, la envolvente espectral transmite información acerca de la señal de voz y el canal de transmisión, así como además, acerca de cualquier ruido aditivo (Kuwabara & Sagisaka, 1995; Watkins & Makin, 1996; Zilovic, Ramachandran & Mammone, 1998; Parikh & Loizou, 2005; Kinnunen & Li, 2010; Miettinen *et al.*, 2011).

Separar estos aspectos de la información después de la extracción de características, es un problema de separación ciega y, por lo tanto, solamente solucionable al hacer algunas suposiciones. Una suposición típica sería suponer que el canal cambia más lentamente que el espectro de voz (Stockham, Cannon & Ingebretsen, 1975; Hermansky, 1994; Gaubitch, Brookes & Naylor, 2013). Esto conduce a un método en el cual un promedio relativamente de larga duración, se resta en el dominio cepstral - Cepstral Mean Normalization (CMN) (Atal, 1974; Furui, 1981; Schwartz *et al.*, 1993; Liu *et al.*, 1993; Hermansky, 1994). La desventaja de este tipo de normalización es que ésta requiere la estimación del cepstrum promedio sobre alguna ventana (por ejemplo, todos los frames de la elocución, o los N frames previos)

(Soong & Rosenberg, 2002; Rose & Reynolds, 1990). Si se utiliza una ventana demasiado corta entonces la media estimada contendrá cierta información de voz, no sólo información del canal. Si la suposición acerca de que el canal cambia más lentamente, relativo al tamaño de la ventana seleccionada, no es correcta, entonces la media estimada no va a reflejar con precisión la respuesta del canal y la normalización será menos efectiva (Bořil & Hansen, 2010; Nakano, Nakagawa & Yamamoto, 2010; Wang *et al.*, 2011).

4.2.1.2. Tasa promedio de sincronía localizada (ALSR)

Junto con las representaciones de tasa-media, se sabe que el sistema auditivo hace uso de otra representación que captura información temporal, aunque precisamente de qué forma estas dos representaciones se combinan en el cerebro, se mantiene como una pregunta abierta (Moore, 2014). En tanto que la codificación temporal es importante para localización binaural de sonido (Stern, Wang & Brown, 2006; Joris & Yin, 2007), también puede desempeñar un papel importante en la interpretación robusta de señales a partir de oídos individuales (Young, 2008).

Por ejemplo, Young & Sachs (1979) demostraron que la tasa promedio de sincronía localizada (ALSR), que se deriva de las descargas en el nervio auditivo, es mucho más robusta a cambios en intensidad de los sonidos tipo vocales, que la correspondiente tasa-media de respuesta, como una función de la respuesta de frecuencia característica (CF). La ALSR describe la extensión a la cual la respuesta neuronal en una CF dada, se sincroniza al armónico más cercano a la frecuencia fundamental de la vocal. Estos resultados sugieren que la información de tiempo asociada con la respuesta a componentes de baja frecuencia de una señal, pueden ser en realidad más robustas a variaciones en intensidad (y potencialmente a varios otros tipos de variabilidad de señal y/o a degradación, tales como canales variables o ruido aditivo), que la tasa-media de la respuesta neuronal.

Un extenso conjunto de modelos auditivos que incluyen detección de sincronía se han propuesto (por ejemplo, para revisiones de valiosa ayuda Jankowski & Lippmann (1992); Jankowski, Vo & Lippmann (1995); Ali, Van Der Spiegel & Mueller (2002); Kim, Chiu & Stern (2006)), y por tanto, no es la finalidad entregar una revisión de todos ellos. De hecho, este capítulo se centra en particular sobre un modelo que fue la inspiración para los *features* que se proponen.

4.2.1.3. Del ALSR al detector generalizado de sincronía (GSD)

El modelo auditivo de Seneff (Seneff, 1988; Stern & Morgan, 2012a), consiste de 40 filtros lineales recursivos, implementados en forma de cascada los cuales cubren un rango de frecuencia desde 130 Hz a 6400 Hz. El ancho de banda de los canales es 0.5 Bark (Seneff, 1988). Estos filtros imitan las respuestas de frecuencia nominales del nervio auditivo, tal como se

describen por Kiang *et al.* (1965), y otros fisiólogos contemporáneos (Liberman, 1978; Young & Sachs, 1979; Sachs & Young, 1979; Sinex & Geisler, 1983; Delgutte & Kiang, 1984; Evans, 1992; Pickles, 2008). El modelo de Seneff utiliza un "modelo de célula auditiva interna", que incluye cuatro etapas: (1) rectificación no lineal de media onda usando una función inversa de tangente para entradas positivas y una función exponencial para entradas negativas; (2) adaptación *short-term* que modela la libración de transmisores en la sinápsis; (3) un filtro pasa bajos con frecuencia de corte aproximadamente de 1 kHz para suprimir la respuesta síncrona a freecuencias de entrada superiores; y (4) una etapa de control automático de ganancia (AGC) para mantener una tasa de respuesta aproximadamente constante a intensidades de entrada superiores, cuando una fibra del nervio auditivo está nominalmente en saturación.

Reflejando el hecho que el sistema auditivo hace uso de dos representaciones espectrales, Seneff propuso dos módulos paralelos, no interactuantes, que operan sobre las salidas del modelo de célula auditiva interna (Seneff, 1988; Moore, 2014). El primero de estos módulos de operación fue un detector de envolvente que produce una estadística que intenta modelar la tasa-media instantánea de la respuesta de una fibra individual dada. El segundo módulo fue llamado GSD, motivado por la medida ALSR de Young & Sachs (1979), y cada canal i se modela (Seneff, 1985; Ali, Van Der Spiegel & Mueller, 2002) como en la Ecuación 4.1, donde $y[n]$ es el valor de la forma de onda de la voz, en la muestra n:

$$\mathrm{GSD}_i(y) = A_s \arctan\left[\frac{1}{A_s}\left(\frac{\langle|y[n]+y[n-n_i]|\rangle - \delta}{\langle|y[n]-\beta^{n_i}y[n-n_i]|\rangle}\right)\right] \tag{4.1}$$

La salida de la célula auditiva interna para este canal i se compara con una versión de sí misma, detardada por el recíproco de la frecuencia central f_i^c del filtro en cada canal (n_i en Ecuación 4.1), y los promedios *short-term* (esto es, detección de envolvente, denotados por $< \ldots >$ en Ecuación 4.1), de las magnitudes (repesentadas por $|\ldots|$ en Ecuación 4.1), de las sumas y diferencias de estas dos cantidades se dividen una por otra. Un umbral δ se introduce para suprimir la respuesta a señales de baja intensidad, y el cociente resultante se pasa a través de un rectificador saturante, de media onda ($arctan[\ldots]$, en Ecuación 4.1), para limitar la magnitud (Seneff, 1985). Un valor ligeramente menor que 1 se utiliza para la constante β en el denominador, mientras que la constante δ en el numerador tiene un valor más bien pequeño Seneff (1985). El parámetro A_s representa un control en el rango lineal para la forma de onda de la voz de entrada (Seneff, 1985; Ali, Van Der Spiegel & Mueller, 2002).

Con los limitados recursos computacionales disponibles en aquel tiempo, Seneff pudo solamente comparar visualmente la respuesta de tasa-media y el GSD, para entradas seleccionadas. Los GSD demostraron de hecho proporcionar una representación útil de las componentes espectrales, incluyendo el ruido (ver por ejemplo, Seneff (1985); Chigier & Leung (1992); Jankowski & Lippmann (1992); Ohshima & Stern (1994)).

Por supuesto, modelos más nuevos y más complejos que el modelo de Seneff se han

propuesto en años más recientes (ver por ejemplo, para una revisón más completa, Moore (2003a); Pickles (2008); Stern & Morgan (2012a); Moore (2014)). Sin embargo, estos modelos más nuevos no son relevantes en el presente capítulo, ya que se está haciendo uso de una propiedad particular del modelo de Seneff como la inspiración del método propuesto, más bien que implementar el modelo completo.

4.2.2. El potencial de los *features* tipo-GSD para reconocimiento de voz

4.2.2.1. Previos intentos para usar este modelo

El modelo de detector generalizado de sincronía propuesto por Seneff (1985), corresponde a uno de los primeros intentos por desarrollar una representación espectral a partir de la codificación temporal que ocurre en las fibras del nervio auditivo (en lugar de sus códigos de tasa) para ser utilizada como módulos de procesamiento (*front-ends*) en sistemas de reconocimiento automático de voz (Seneff, 1986b; Stern & Morgan, 2012a). Seneff reportó fuertes evidencias que las representaciones basadas en la audición son interesantes y es valioso su estudio en sistemas de análisis de voz. De acuerdo a Seneff (1988), resultados preliminares de las dos distintas representaciones espectrales para la señal de voz, una basada en la tasa de descarga (*rate coding*) de las fibras del nervio auditivo, y la otra representación basada en la respuesta síncrona de las fibras (*synchrony coding*), indicaron que las salidas de la respuesta de tasa son exitosas para ubicar fronteras acústicas. Del mismo modo, las salidas de sincronía aplicadas a reconocimiento de vocales, independientes del locutor, en voz continua, demostraron un desempeño superior. No obstante, no hubo ninguna explicación sobre los mecanismos de interacción neuronal entre codificación de tasa versus codificación de sincronía, y cómo el cerebro auditivo utiliza partes de la información de estas dos representaciónes en situaciones de comunicación real (Smith, Delgutte & Oxenham, 2002; Moore, 2008c).

Aunque el GSD de Seneff se ha empleado como un método de extracción de características para reconocimiento de voz, tal como la detección de frecuencias formantes (Seneff, 1984, 1986a; Kim, Lee & Kil, 1999), su desempeño ante los *features* inspirados en la tradicional tasa-media, tales como los MFCCs (Jankowski, Vo & Lippmann, 1995; Ali, Van Der Spiegel & Mueller, 2002), ha sido regular. En general, en voz limpia los *features* GSD proporcionan precisiones de reconocimiento que no son mejores que aquellos proporcionados por los *features* convencionales MFCC o PLP (y en algunos casos, su desempeño empeora), pero en ruido aditivo ellos pueden ser de utilidad (por ejemplo, Chiu & Stern (2008)). Una extensión del GSD de Seneff fue proporcionada por Ali, Van Der Spiegel & Mueller (2002). Ese método conocido como Detección Promedio Localizado de Sincronía también produce un espectro de sincronía y proporciona mejores resultados de reconocimiento bajo condiciones de ruido que el detector GSD original de Seneff.

Más aún, los GSDs deben sintonizarse perfectamente a las frecuencias formantes para obtener una salida limpia (Seneff, 1988). Este fue uno de los principales problemas del algoritmo original GSD (Seneff, 1985; Ali, Van Der Spiegel & Mueller, 2002).

4.2.2.2. Un análisis en el dominio de la frecuencia del GSD

El GSD original de Seneff está definido en el dominio del tiempo por la Ecuación 4.1. Dado que en reconocimiento de voz/locutor, es más conveniente realizar extracción de características en el dominio de la frecuencia, se propone realizar un análisis del GSD de Seneff en el dominio de frecuencia. Para esto, se pasan tono puros (sinusoides) a diferentes frecuencias, barriendo el espectro total, usando el filtro en el dominio del tiempo de la Ecuación 4.1, lo cual es efectivamente, una forma de análisis de frecuencia.

Se considera solamente la respuesta de magnitud de los filtros GSD, y se ignora la fase ya que se asume que ésta probablemente sería menos importante en aplicaciones de reconocimiento de voz/locutor. Ecuación 4.1 es la razón de dos términos, numerador y denominador, los cuales se pueden analizar en forma separada. Ecuaciones 4.2 y 4.3, representan estos términos, que son calculados para cada canal i en cada *frame* de análisis.

$$\text{Numerator}_{\text{GSD}} = \langle |y[n] + y[n - n_i]| \rangle - \delta \tag{4.2}$$

$$\text{Denominator}_{\text{GSD}} = \langle |y[n] - \beta^{n_i} y[n - n_i]| \rangle \tag{4.3}$$

Se considera, a modo de ejemplo, la respuesta de los términos del numerador y denominador, de uno de los canales GSD (posterior a un filtro pasabanda), sintonizado a una frecuencia central f_i^c de $692Hz$, para 1024 tonos puros que expanden el rango de frecuencia de $60Hz$ a $3500Hz$, como se muestra en la Figura 4.1.

4.2.2.3. Respuestas espurias del GSD

Las respuestas de frecuencia del numerador y denominador, mostradas en la Figura 4.1, parecen prometedoras inicialmente, estando centradas en la frecuencia sintonizada de 692 Hz, como se esperaba, y con el ancho de banda del denominador siendo ligeramente más ancho que aquél del numerador. Los valores utilizados para las constantes son $\delta = 1 \times 10^{-5}$ and $\beta = 0.999$. Sin embargo, si se examina la respuesta GSD - el numerador dividido por el denominador - como se grafica en la Figura 4.2, se observan peaks adicionales en frecuencias más altas, junto con el peak deseado en 692 Hz. Seneff describe esta limitación del GSD (Ecuación 4.1), estableciendo que el GSD genera peaks espurios en armónicos de la frecuencia detectada (Seneff, 1985). A pesar de estas observaciones, el comportamiento de cada canal GSD en la región en torno a su frecuencia central, aun tiene propiedades deseables, y éstas se explotan en los *features* propuestos en el presente capítulo, los cuales se describen en las

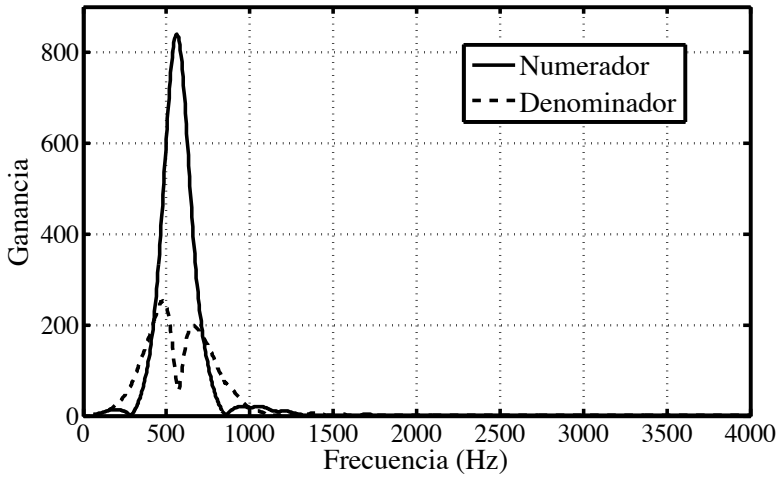

Figura 4.1: *Respuesta de frecuencia del numerador y denominador de un GSD sintonizado a 692 Hz. El numerador en la Ecuación 4.2 se muestra como una línea sólida y el denominador en la Ecuación 4.3 se presenta con líneal discontinua.*

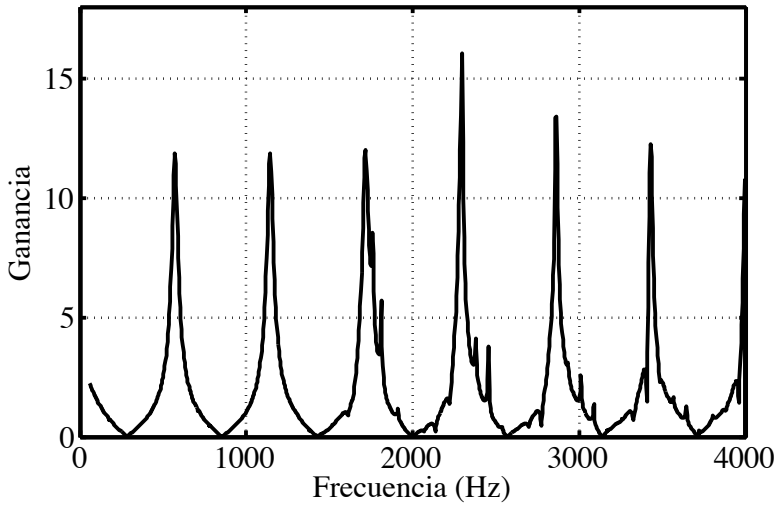

Figura 4.2: *Respuesta de frecuencia de un canal GSD a* $f_i^c = 692$ *Hz.*

próximas secciones.

Al examinar las respuestas del numerador y denominador, graficadas sobre una escala logarítmica como en la Figura 4.3, se observa la causa de este comportamiento. En la figura, el denominador es dibujado como su recíproco para entender más claramente su relación con el numerador. En la siguiente sección, se construye un canal tipo-GSD que conserva el comportamiento de la normalización deseada, proporcionado por el término del denominador, pero que no produce respuestas espúrias fuera de su banda de paso nominal.

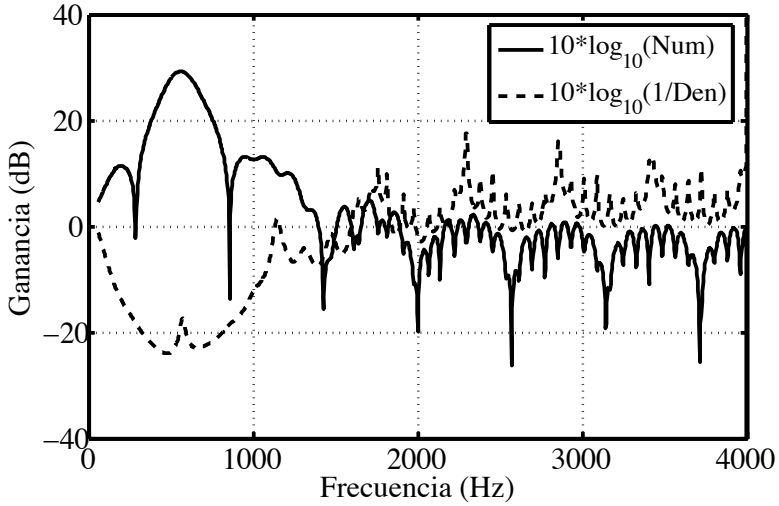

Figura 4.3: *Magnitud logarítmica de la respuesta del numerador y denominador del GSD, sintonizado a $f_i^c = 692$ Hz. El numerador (Equación 4.2) se muestra con una línea gruesa, en tanto que el recíproco del denominador (Equación 4.3) se presenta con línea fina.*

4.3. Coeficientes cepstrales localmente normalizados

Las aplicaciones objetivo descritas anteriormente, involucran canales con respuestas de frecuencia variables en el tiempo, las que incluyen situaciones en las cuales puede variar el ordenamiento físico entre locutor y micrófono. Por lo tanto, se buscan *features* que sean relativamente invariantes a cambios en la respuesta de frecuencia del canal. Los *features* propuestos logran este propósito al utilizar una forma de normalización local inspirada por la razón entre los términos del numerador y denominador del GSD de Seneff, dados en Ecuación 4.2 y Ecuación 4.3. En este capítulo se refiere a estos *features* como LNCC, *Locally-Normalized Cepstral Coefficients*.

4.3.1. Del GSD a un modelo el el dominio de frecuencia apropiado para tecnología de voz

Al examinar el comportamiento del GSD, se puede identificar algunos atributos deseables que no se encuentran en los tradicionales *features* tales como los MFCCs. Se observa a partir de la forma de la Ecuación 4.1 y de la Figura 4.3 (ignorando por el momento las respuestas espúrias de alta frecuencia), que la parte del numerador actúa como un filtro pasabanda centrada en torno a una frecuencia particular, y que su salida se divide por (es decir, normalizada por), un término denominador que es un filtro que responde a energía sobre ambos lados del filtro numerador. En otras palabras, una normalización local está siendo realizada: la salida de un canal GSD, se relaciona con la cantidad de energía en una banda de frecuencia particular, *relativa* a la enegía en las regiones vecinas (frecuencias bajas y altas).

Con un ancho de banda del filtro, apropiadamente seleccionado, el efecto es tal que conserva los *peaks* espectrales (los que se relacionan a la voz), mientras estos peaks son relativamente invariantes, por ejemplo, a *tilt* espectral global. El concepto de respuesta en una región localizada, que es suprimida (o inhibida) por la respuesta en un rango más ancho de frecuencia, se encuentra también en el sistema visual (Werblin, Jacobs & Teeters, 1996), y en el sistema auditivo (Sachs & Kiang, 1968; Houtgast, 1972). Wang & Shamma (1994), entre otros, han comentado sobre la utilidad de este tipo de mecanismo para reconocimiento de voz.

Se puede conseguir un comportamiento similar directamente en el dominio de la frecuencia, al diseñar filtros simples para el numerador y denominador, respectivamente (Figura 4.4, donde f_c es la frecuencia central del canal, $d_{\min}$ el valor centrado mínimo del denominador, y B su ancho de banda. En este capítulo, las frecuencias se definen sobre la escala Bark (Zwicker, 1961)). Tal par de filtros desempeñarán, en el dominio de la frecuencia, una normalización local similar a aquella desempeñada en el dominio del tiempo por el GSD (Ecuación 4.1).

Al trabajar en el dominio de la frecuencia (tal como en los MFCCs tradicionales), los filtros pueden diseñarse con facilidad, para que así respondan sólo dentro de la banda de paso principal, eliminando los *peaks* espurios como se ven en la Figura 4.2.

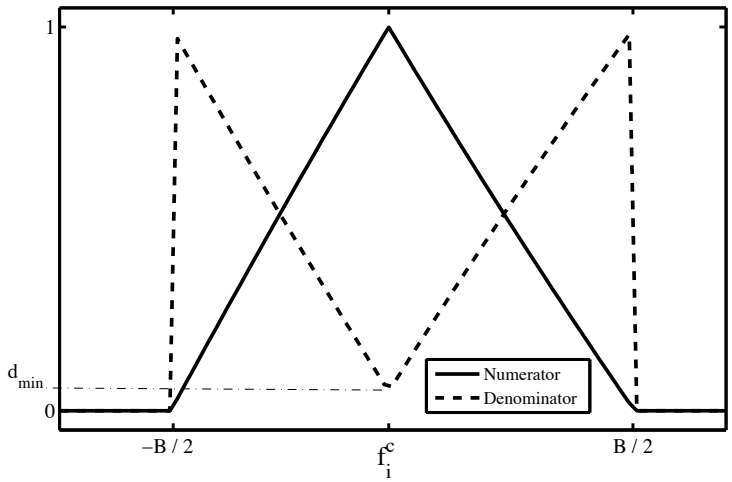

Figura 4.4: *Formas de la magnitud de los filtros numerador (línea sólida) y denominador (línea discontinua), para un único canal del banco de filtros propuesto auto-normalizado.*

El par de filtros para un canal como el descrito anteriormente, se diseñó a través de experimentación informal. Las respuestas del par de filtros configurados a una frecuencia central $f_i^c = 515$ Hz, se presentan en la Figura 4.5. El filtro numerador es fundamentalmente el mismo que el filtro triangular utilizado normalmente en el banco de filtros usado para derivar los MFCCs (Davis & Mermelstein, 1980), y se describe en el dominio de la frecuencia por la Ecuación 4.4 para cada canal i con frecuencia central f_i^c. El filtro denominador captura

energía sobre los dos lados del filtro numerador; este filtro se describe por la Ecuación 4.5.

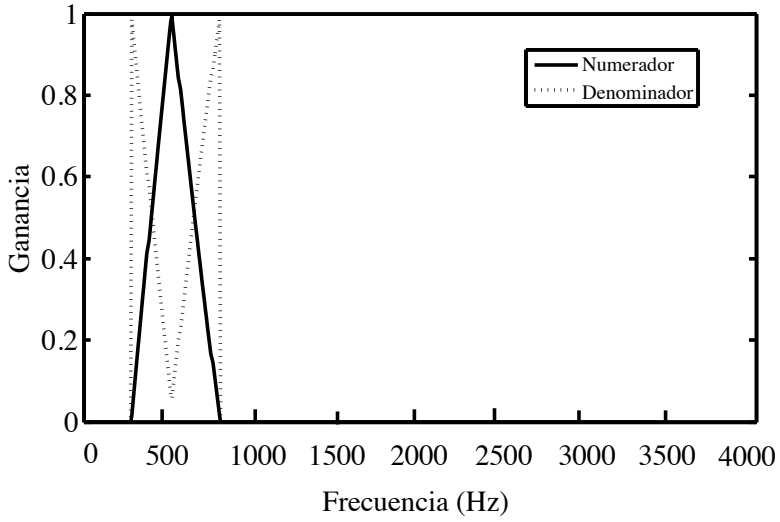

Figura 4.5: *Respuesta de frecuencia del numerador y denominador sintonizados en f_i^c = 515 Hz.*

$$\text{Numerador}_{\text{LNCC}}(f) = \begin{cases} -\frac{2}{B}|f - f_i^c| + 1 & \text{cuando } |f - f_i^c| \leq \frac{B}{2} \\ 0 & \text{en otro caso} \end{cases} \tag{4.4}$$

$$\text{Denominador}_{\text{LNCC}}(f) = \begin{cases} \frac{2}{B}(1 - d_{\min})|f - f_i^c| + d_{\min} & \text{cuando } |f - f_i^c| \leq \frac{B}{2} \\ 0 & \text{en otro caso} \end{cases} \tag{4.5}$$

En tanto que ambos filtros $\text{Numerator}_{LNCC}(f)$ y $\text{Denominator}_{LNCC}(f)$ tienen una respuesta distinta de cero solamente para frecuencias en el rango de $-\frac{2}{B} \leq |f - f_i^c| \leq \frac{2}{B}$, es fácilmente observable que la respuesta de $\text{Numerator}_{LNCC}(f)$ es mayor para un rango estrecho de frecuencias en torno de $f = f_i^c$, mientras que $\text{Denominator}_{LNCC}(f)$ responde a actividad en las regiones de frecuencias de los alrededores. Al montar un banco de tales pares de filtros, se puede extraer una representación de un banco de filtros localmente-normalizado de la señal, representación que se puede utilizar seguidamente para calcular los *features* cepstrales, siguiendo los mismos pasos que se utilizan para derivar los MFCCs de las salidas del banco de filtros convencional (Davis & Mermelstein, 1980). En todos los experimentos presentados en este capítulo, los filtros se construyeron sobre una escala Bark.

Es sencillo reemplazar el banco de filtros utilizado habitualmente en extracción de características MFCC, por este banco de pares de filtros auto-normalizados. Al espaciar los filtros sobre una escala perceptual (tal como la escala Bark), seguida por una compresión logarítmica y una transformada coseno truncada, se derivan *features* de voz que tendrán propiedades

muy similares a los MFCCs tradicionales (por ejemplo, ellos son estadísticamente decorrelacionados), pero con la incorporación de la normalización local durante la etapa del banco de filtros. El efecto global combina filtrado con un banco de filtros (el que elimina detalles finos del espectro tales como armónicos de la frecuencia fundamental F_0), y normalización local (la cual quita variaciones muy poco finas en la forma espectral, tal como un *tilt* global que se supone surge generalmente de variabilidad de canal).

En otras palabras, los *features* propuestos se pueden utilizar como una sustitución de "paso" directo a MFCCs sin cambios ninguno al modelo estadístico, por ejemplo. La Figura 4.6, describe la secuencia completa de pasos requerida para extraer *features* LNCC y muestra para comparación, los pasos correspondientes para la extración de *features* MFCC convencional. Se observa la similitud entre ambas secuencias, siendo la única diferencia la normalización de las salidas del banco de filtros en LNCC. Es común agregar los coeficientes delta y delta-delta; éstos no se muestran en los diagramas.

4.3.1.1. Respuesta de frecuencia del par propuesto de filtros auto-normalizados

Las respuestas de frecuencia del los filtros propuestos, numerador y denominador, definidos en las Ecuaciones 4.4 y 4.5 se presentan en la Figura 4.7 la que grafica las respuestas individuales del par de filtros numerador y denominador, respectivamente, sobre una escala logarítmica. La respuesta combinada del numerador dividido por el denominador, se grafica en la Figura 4.8. Este gráfico pone a descubierto que, cuando se combinan, el par de filtros en LNCC muestra una respuesta más "aguda" que los filtros triangulares en un banco de filtros tradicionales MFCC.

Mientras la Figura 4.8 presenta la respuesta a tonos puros, es más útil examinar la respuesta a una señal de banda ancha (esto es, una vocal), para observar el efecto de normalización. La Figure 4.9 grafica las envolventes espectrales (esto es, salidas del banco de filtros graficadas inmediatamente después del paso de compresión logarítmica en la Figura 4.6), estimadas por el banco propuesto de filtros normalizados, y lo compara a la respuesta correspondiente de un banco de filtros convencional, tal como los filtros habitualmente utilizados para derivar los coeficientes MFCCs. Para facilitar la compresnión de las envolventes, la línea sólida se ha desplazado alrededor de +15 dB. Se observa que el banco propuesto de filtros, auto-normalizado, conserva importante información de la forma espectral, tales como peaks espectrales, pero elimina el *tilt* espectral global.

4.3.1.2. Robustez a *mismatch* de canal

En la Figura 4.10, se observa la respuesta del banco de filtros LNCC cuando la voz es filtrada por un canal con una respuesta de frecuencia no-plana, en este caso, un *tilt* espectral de -6 dB/octava (envolventes espectrales que expresan salidas del banco de filtros graficadas

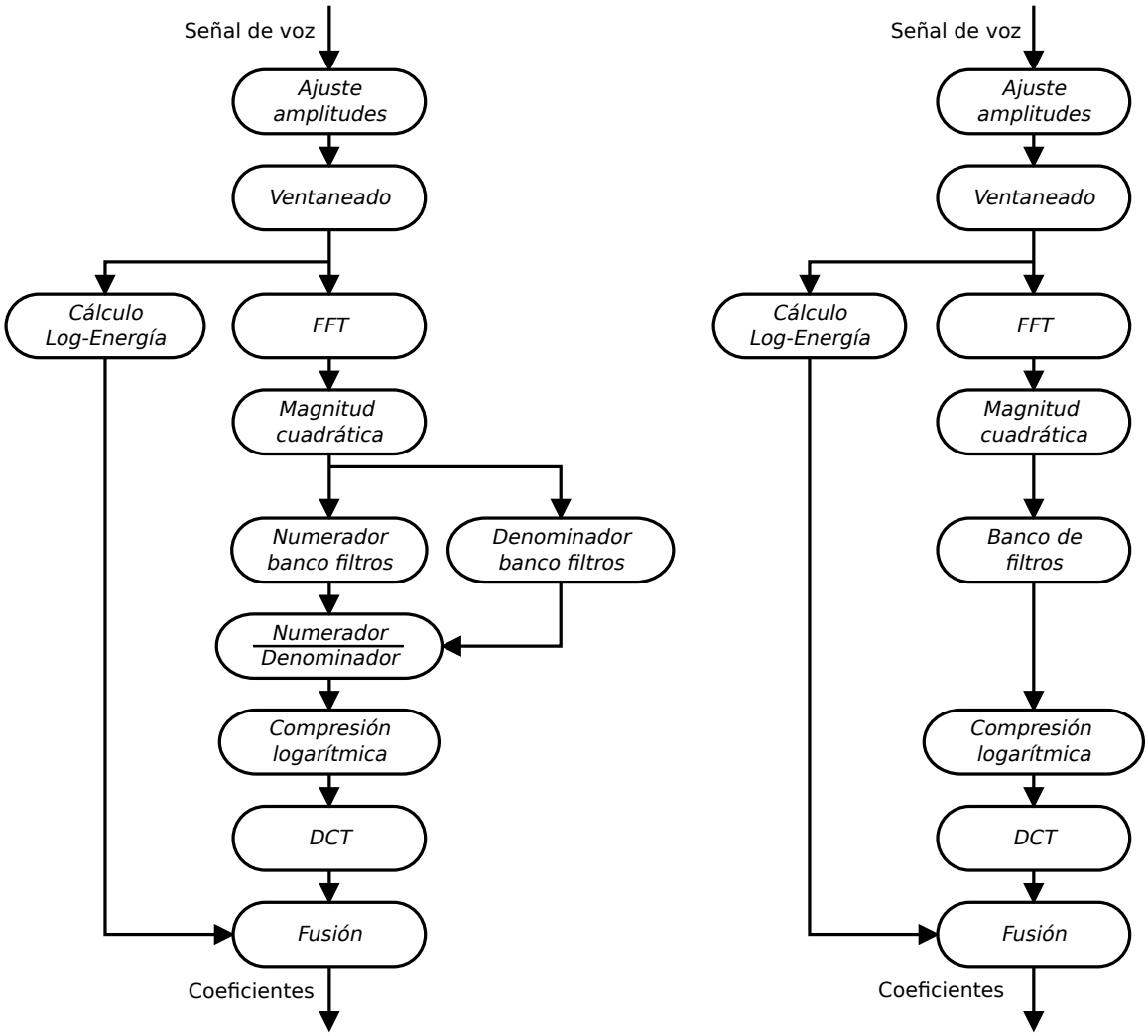

Figura 4.6: *Diagrama de flujos para extracción de features LNCC (izquierda) y MFCC (derecha).*

justo después de las etapas de compresión logarítmica en la Figura 4.6). Se observa que los *features* propuestos son invariantes al *tilt* espectral del canal, en tanto que, las salidas del banco de filtros convencional, son extremadamente sensibles a este *tilt*. Las respuestas a voz no modificada se muestran en líneas sólidas y las respuestas a voz filtrada a través de un canal que impone un *tilt* espectral de -6 dB/octava, se presentan en líneas discontinuas. El banco de filtros clásico conserva la respuesta del canal en su salida, mientras mantiene las propiedades claves relacionadas a la voz tales como los peaks espectrales.

En todos los experimentos presentados en este capítulo, se compara los *features* propuestos con los *features* MFCCs tradicionales, que están opcionalmente normalizados utilizando Normalización Media Cepstral (CMN) (Atal, 1974; Furui, 1981; Wang, Kitaoka & Nakagawa, 2007). Técnicas más avanzadas tales como filtrado RASTA (Hermansky *et al.*, 1991a,b; Her-

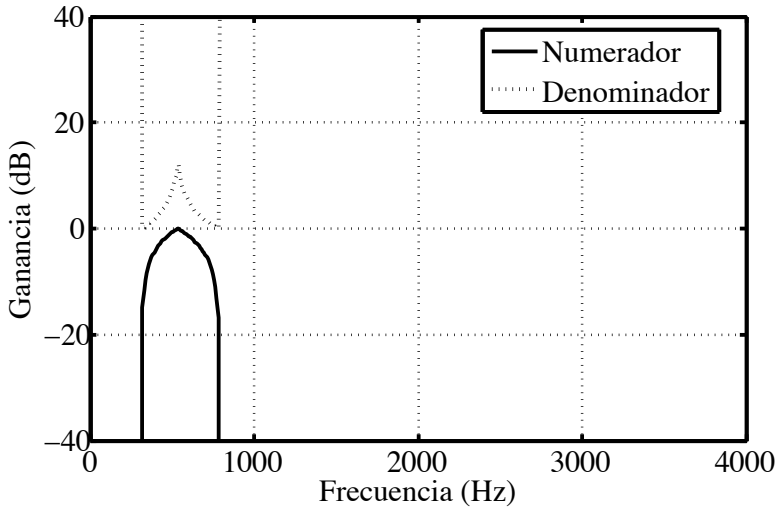

Figura 4.7: *Respuesta de frecuencia del numerador y denominador, separadamente, ambos sintonizados en $f_i^c = 515Hz$, sobre una escala logarítmica.*

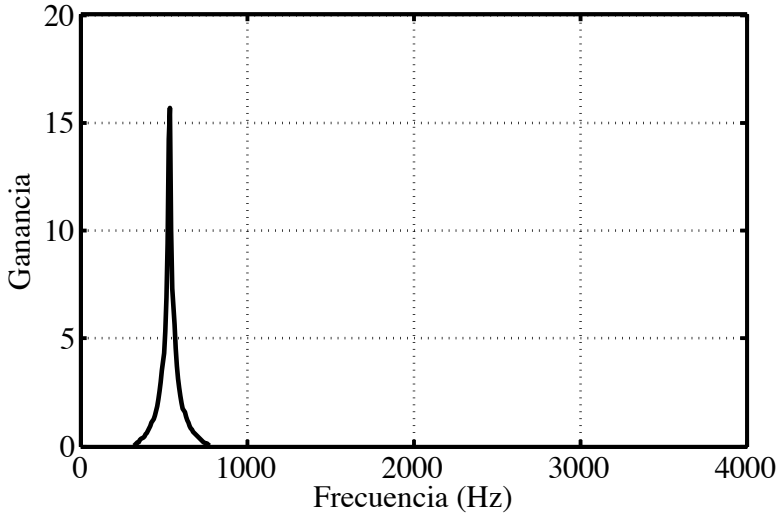

Figura 4.8: *Respuesta de frecuencia del numerador dividido por el denominador, ambos sintonizados en $f_i^c = 515Hz$, sobre una escala logarítmica.*

mansky, 1994), están por supuesto disponibles, pero la comparación (o bien, de hecho la combinación) con aquellas otras técnicas, se deja como trabajo futuro. Es importante observar que todas estas otras estrategias requieren información fuera del frame actual que está siendo procesado y, por lo tanto, son menos efectivas para canales de variación rápida (Leus & Moonen, 2003; Leus, 2004). Por ejemplo, CMN requiere una estimación precisa de la media cepstral, la que puede ser defícil de obtener confiablemente en algunos casos (Qi Li *et al.*, 2002); RASTA hace una suposición equivalente, que el canal cambia sustancialmente en forma más lenta que la envolvente espectral de la voz (Hermansky, 1994).

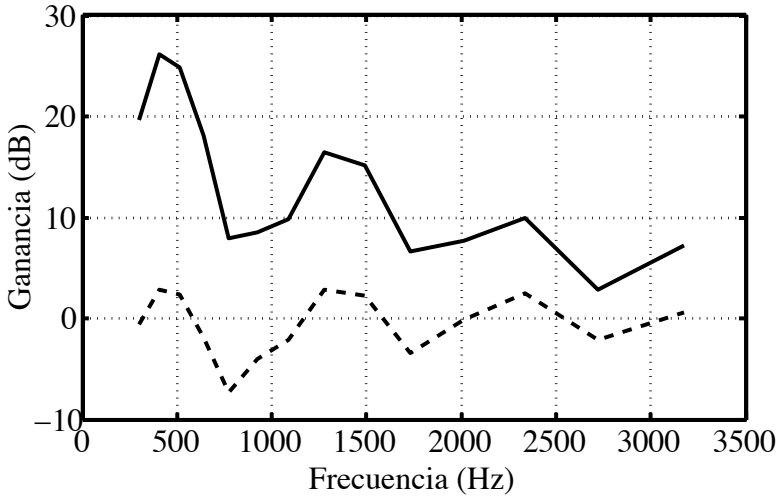

Figura 4.9: *Envolventes espectrales para un único frame de voz sonora, utilizando un banco de filtros tradicional en escala Mel (línea sólida) y para el banco propuesto LNCC (línea discontinua).*

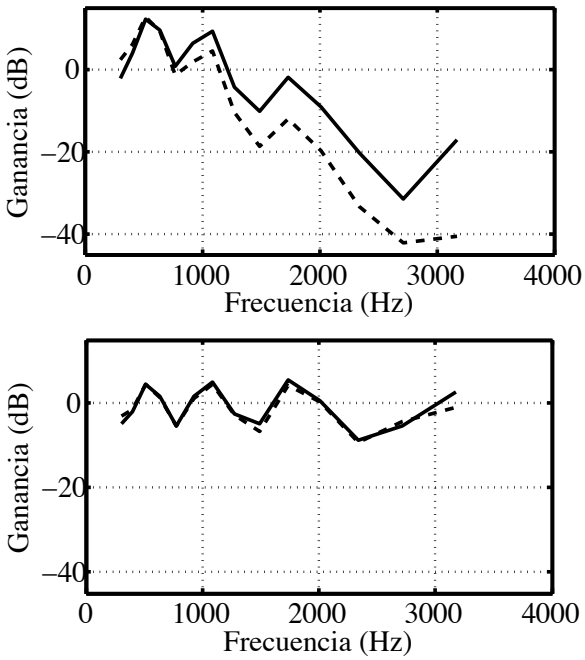

Figura 4.10: *Envolventes espectrales para un único frame de voz sonora, utilizando un banco de filtros convencional en escala Mel (figura superior), y empleando el banco de filtros propuestos LNCC (figura inferior).*

4.4. Experimentos de verificación de locutor

Para investigar la capacidad de los *features* propuestos para normalizar durante canales variables, se realizaron series de experimentos de verificación de locutor sobre voz degra-

dada por varios canales. Estos involucran canales simulados imponeniendo *tilt* espectral que imita el efecto de micrófonos bloqueados o fuera del eje (Sección 4.4.5) al igual que características de *tilt* espectral que varíen dentro de una elocución (Section 4.4.6). Por razones de control experimental y de replicabilidad, se simularon las respuestas del canal. En todos los experimentos, el sistema se entrenó utilizando solamente voz limpia. La voz de prueba se degradó con respecto a la información de entrenamiento al imponer *tilt* espectral estático y variable en el tiempo.

4.4.1. Sistema de verificación de locutor

Los experimentos se realizaron con sistema de verificación de locutor de texto-independiente basado en modelos de mezclas de Gaussianas (GMMs) (Reynolds, Quatieri & Dunn, 2000; Bimbot *et al.*, 2004), con un modelo universal de referencia, UBM (*Universal Background Model*), que se entrena usando locutores impostores. El orden del modelo UBM fue de 256 Gaussianas. Para cada uno de los locutores se genera un GMM dependiente del locutor al utilizar adaptación MAP (Reynolds, Quatieri & Dunn, 2000). Al hacer esto, se mantiene la correspondencia de las Gaussianas dentro de cada GMM dependiente de locutor, con aquellas en el GMM de referencia (Reynolds, Quatieri & Dunn, 2000). Dada una prueba de verificación donde la identidad del locutor s es demandada, O representa la secuencia de observación correspondiente a la elocución que se demanda. La calificación de salida del sistema es una verosimilitud logarítmica de cohorte-normalizada, $\log L(O)$:

$$\log L(O) = \log L(O/\lambda_s) - \overline{\log L(O/\lambda_{\bar{s}})} \tag{4.6}$$

donde $\log L(O/\lambda_s)$ representa el logarítmo de la verosimilitud de la hipótesis del cliente, λ_s es el modelo s del locutor, y $\overline{\log L(O/\lambda_{\bar{s}})}$ es el promedio de la verosimilitud logarítmica de la cohorte de modelos impostores.

Según describe Becerra Yoma & Villar (2002), los frames con más alto SNR por segmento local, proporcionan información más confiable que aquellos con bajo SNR segmental. Además, sonidos sonoros (por ejemplo, vocales), muestran una capacidad mucho mayor de discriminación de locutor que sonidos fricativos. En consecuencia, se descartan todos los frames cuya energía normalizada con respecto a la energía máxima del frame de la elocución, es inferior a un umbral dado.

4.4.2. Extracción de *features*

Los *features* fueron extraídos utilizando procesamiento LNCC y MFCC, como se describe en la Figura 4.6. La longitud del frame en todos los casos fue 25 ms, con un traslape de 50 %. Se cubrió un rango de frecuencia de 200 a 3860 Hz por 14 filtros triangulares, uniformemente ordenados, sobre una escala Bark en el caso de los *features* propuestos LNCC. Si un canal

LNCC va más allá del rango 0Hz a la frecuencia de Nyquist, este es simplemente truncado. La DCT, en ambos casos, fue truncada en 11 coeficientes, luego el primer coeficiente fue reemplazado por el logaritmo de la energía del frame. Por último, los 11 coeficientes resultantes, aumentan con deltas y delta-delta, conformando así el vector de *features* de dimensión 33 para cada frame.

4.4.3. Base de datos YOHO

En todos los experimentos se utilizó la base de datos YOHO Speaker Verification Corpus, que comprende voz en Inglés grabada de alta calidad a una tasa de sampleo de 8kHz (Campbell & Higgins, 1994). YOHO mantiene el desarrollo, entrenamiento y prueba de sistemas de verificación con un vocabulario que comprende números de dos dígitos hablados continuamente, en conjuntos de tres (por ejemplo, "62-31-53" pronunciados como "*sixty-two thirty-one fifty-three*").

La base de datos se divide en secciones de entrenamiento y verificación. Cada una de estas secciones, contiene información de 138 locutores. En los experimentos, se utilizó un subconjunto de 70 locutores. Estos locutores se dividen como se explica a continuación: 40 locutores impostores de referencia para entrenar los modelos de referencia; 30 locutores clientes para prueba, que se emplean en los intentos de verificación. Por cada locutor, se empleó una sesión de entrenamiento de 24 elocuciones. Curvas de falso rechazo se estiman con 30 locutores $\times$ 16 señales de verificación por cliente = 480 elocuciones. Curvas de falsa aceptación se obtienen con 30 locutores $\times$ 29 impostores $\times$ 6 señales de verificación por impostor = 5220 experimentos.

4.4.4. Experimentos inciales: sensibilidad a ajustes de parámetros

Experimentos preliminares se realizaron para determinar cuán sensibles son los *features* propuestos a los diferentes parámetros que deben ser escogidos: el ancho de banda de los filtros (todos los filtros tienen el mismo ancho de banda sobre una escala Bark), el número de canales (el número de filtros además determina su espaciamiento, como un ancho de banda que es demasiado angosto dejaría un "gap" en la respuesta global del banco de filtros), y el parámetro $d_{\min}$ que previene división por cero en la frecuencia central de cada par de filtros numerador y denominador. Como se describe en la Sección 4.3.1.1, los filtros LNCC muestran una respuesta más "aguda" que los filtros triangulares en el banco de filtros MFCC. Por lo tanto, se obtienen normalmente mejores desempeños con un número mayor de filtros (por ejemplo, 28) que en el banco de filtros MFCC (el que comprende 14 filtros). Todos los experimentos en relación con sensibilidad de parámetros se realizaron con voz limpia, y con voz procesada a través de un canal con una respuesta de frecuencia de -6 dB/octava de *tilt* espectral.

4.4.4.1. Número de canales LNCC y ancho de banda de los filtros

Como se observa en la Figura 4.11, el desempeño sobre voz limpia es relativamente no afectado por el ancho de banda hasta que este llega a ser demasiado angosto, esto es debido presumiblemente porque a anchos de banda angostos con un número constante de canales, *gaps* comienzan a aparecer entre los filtros y la información de voz entonces se pierde. Para voz espectralmente con *tilt*, el mismo efecto se observa con anchos de banda angostos, pero además se observa un deterioro del desempeño a anchos de banda demasiado anchos. Esto se supone sea una consecuencia de la normalización local que llega a ser "menos local" y, por lo tanto, menos efectiva.

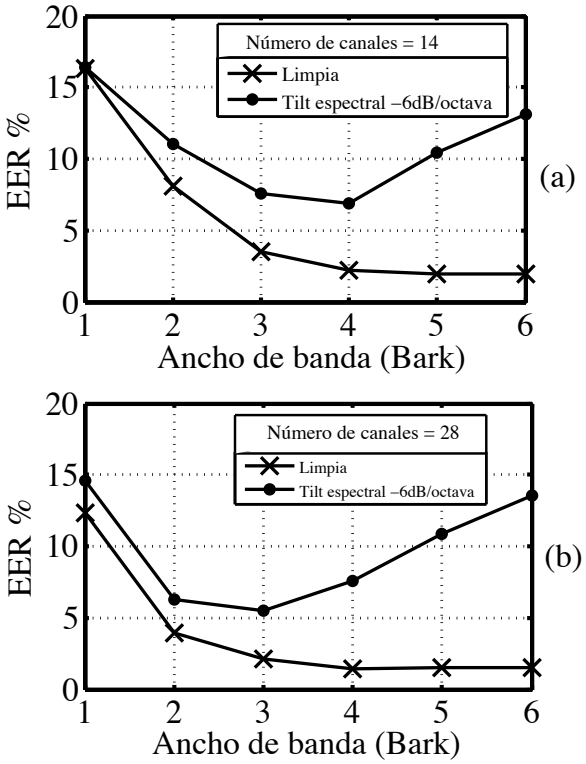

Figura 4.11: *Sensibilidad al ancho de banda del filtro. Ambos con d_{min}=0.001. (a) 14 canales, (b) 28 canales.*

Además, en la Figura 4.11 se observa que 28 canales LNCC conducen a valores de EER menores que para 14 filtros. Aunque no se presenta en este capítulo, experimentos adicionales se realizaron con 20 y 56 canales LNCC. Sin embargo, aquellas configuraciones no guiaron a mejoras significativas en tasas de igual error (EER), cuando se compararon con 28 canales. Es valioso enfatizar que el ancho de banda óptimo con *tilt* espectral de -6 dB/octava, se desplaza a partir de B igual a 4 Barks hacia 3 Barks cuando el número de canales LNCC aumenta desde 14 a 28. Este resultado debe ser debido al hecho que a mayor número de canales LNCC, menor tiende a ser el *gap* entre los filtros.

4.4.4.2. Valor central mínimo del denominador ($d_{\min}$)

La Figura 4.12 describe EER como una función de $d_{\min}$ para voz limpia y voz degradada por *tilt* espectral de -6 dB/octava. Los coeficientes LNCC se calculan utilizando 28 canales y un ancho de banda de B=3 Barks. De acuerdo a la Figura 4.12, existe un rango amplio de valores para $d_{\min}$ ($0 \leq d_{\min} \leq 0.01$) para los cuales EER muestra poca variación.

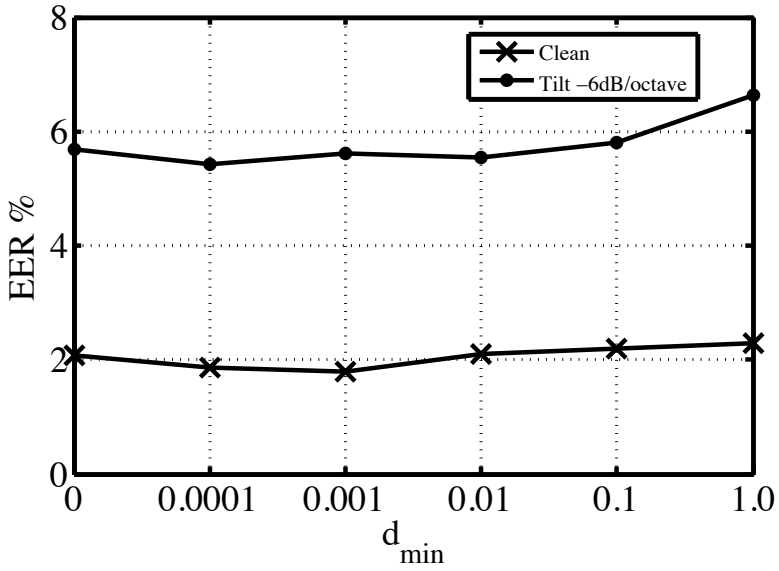

Figura 4.12: *Sensibilidad a parámetro d_{min}. LNCC con 28 canales, B=3 Barks.*

4.4.5. Experimento 1: micrófono distante simulado

4.4.5.1. Procesamiento de voz para simular la respuesta de frecuencia de un micrófono distante

Como se menciona en la Sección 4.1.2.2, una de las consecuencias de utilizar un micrófono bloqueado, o distante, o fuera del eje de direccionalidad, o bien un arreglo de micrófonos para capturar la voz, es que ciertas formas espectrales desconocidas van a imponerse sobre la voz debido al canal. La voz resultante de un aumento del esfuerzo vocal puede además variar el *tilt* espectral con respecto a voz limpia. El efecto global es que la voz de prueba tiene una forma espectral promedio distinta a la voz limpia de entrenamiento. Esto se simuló empleando un filtro simple que obliga a un *tilt* espectral de -3 dB/octava o -6 dB/octava; se llegó a estos valores particulares a través de experimentos informales en los cuales se volvió a grabar la voz reproducida por un altavoz, con el micrófono fuera del eje de direccionalidad, o lugares de bloqueos entre el altavoz y el micrófono.

4.4.5.2. Resultados

La Figura 4.13 compara los resultados con MFCC, MFCC+CMN, LNCC and LNCC+CMN. Los *features* propuestos LNCC conducen a una tasa de error EER menor que MFCC+CMN y ligeramente peor que MFCC con voz limpia. Cuando la voz se degrada con un *tilt* espectral de -3 dB/octava, LNCC proporciona el más bajo EER, que es 24 % (p <0.01) menor que aquél conseguido a través de MFCC+CMN. Con -6 dB/octava, tanto MFCC+CMN como LNCC, dramáticamente compensan durante esta distorsión y dan reducciones en EER tan altas como 87 % y 79 %, respectivamente. Por otra parte, MFCC+CMN da un EER 2.1 % menor (absoluto) que LNCC. No obstante, LNCC es sin memoria, y no requiere calcular ni almacenar el promedio móvil necesitado por CMN. Se observa que LNCC+CMN no muestra ninguna mejora con respecto a LNCC. Este resultado sugiere que CMN no ayuda a LNCC a compensar durante *tilt* espectral, e introduce una distorsión debido a la estimación estadística de las medias de los *features*.

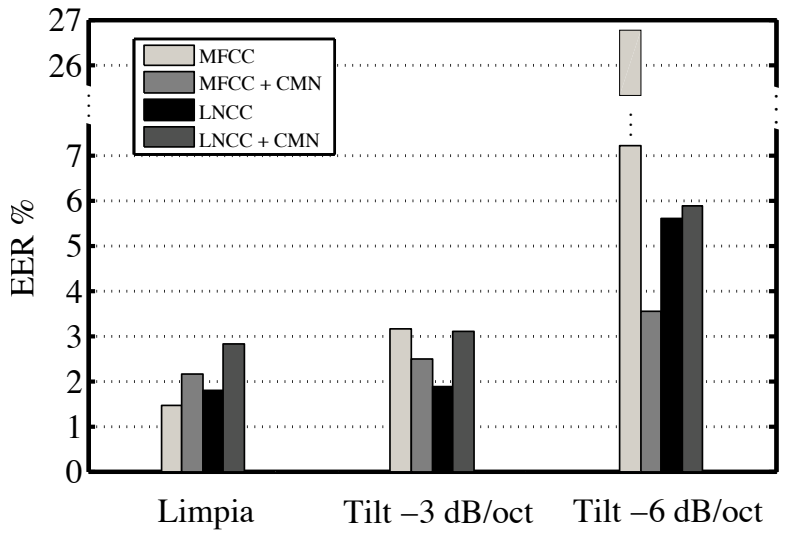

Figura 4.13: *Desempeño durante tilt espectral constante. Los features LNCC se calculan utilizando 28 canales,* d_{min}*=0.001 y B= 3 Barks.*

4.4.6. Experimento 2: Canales que varían rapidamente

4.4.6.1. Procesamiento de voz para simular una respuesta de canal variable en el tiempo

Un filtro dinámico se diseñó para modificar el *tilt* espectral a través del tiempo. Este es un filtro FFT aplicado a un frame sobre una base frame-a-frame. La pendiente del *tilt* objetivo varía linealmente entre 0 dB/octava y −6 dB/octava dentro de cada elocución. Luego, la energía de cada etapa de filtro se normaliza para compensar durante la atenuación producida por el *tilt* espectral frame-a-frame. Para asegurar la efectividad del filtro, éste se aplica

entre estimaciones de punto de inicio y término de la voz. Tres *tilts* espectrales variables en el tiempo fueron aplicados: (1) $VaryingST_1$, el *tilt* espectral cambia linealmente con el tiempo desde 0 dB/octava a −6 dB/octava, desde el comienzo hasta el final de la elocución; (2) $VaryingST_2$, el *tilt* espectral cambia linealmente con el tiempo, a una tasa absoluta constante, desde 0 dB/octava a −6 dB/octava, y a continuación, desde −6 dB/octava a 0 dB/octava a partir de el principio hasta el final de la elocución; y (3) $VaryingST_3$, el *tilt* espectral cambia linealmente con el tiempo, a una tasa absoluta constante, desde 0 dB/octava a −6 dB/octava, luego desde −6 dB/octava a 0 dB/octava y, finalmente, desde 0 dB/octava a −6 dB/octava a partir del inicio hasta el término de la elocución.

4.4.6.2. Resultados

Como se observa en la Figura 4.14, LNCC proporciona EER más bajo que los otros *features*, con valores medios de EER de 2.65 %, 2.79 %, 1.93 %, 3.33 % para MFCC, MFCC+CMN, LNCC, LNCC+CMN, respectivamente. Los *features* LNCC se calculan utilizando 28 canales, $d_{\min}$=0.001 y B= 3 Barks. Además, se observa que LNCC proporciona reducciones en EER iguales a 12.1 % (p <0.271), 41.2 % (p $<1 \times 10^{-7}$), 42.2 % (p $<3.5 \times 10^{-7}$) para las condiciones $VaryingST_1$, $VaryingST_2$, y $VaryingST_3$, respectively, cuando se compara con MFCC+CMN, así como también proporciona reducciones en EER iguales a 30.4 % (p $<1.9 \times 10^{-4}$), 32.2 % (p $<7 \times 10^{-5}$), y 42.2 % (p $<1.8 \times 10^{-8}$), cuando se compara con los *features* estándar MFCC. Además, CMN no mejora MFCC en la mayoría de estas condiciones, lo que demuestra una falta de robustez sobre la parte de CMN en respuesta a *tilt*s espectrales variables en el tiempo. Este resultado es un reflejo del hecho que los coeficientes CMN se estiman sobre un intervalo de tiempo durante el cual las estadísticas de la señal cambian y las medias de los *features* ya no se calculan de manera confiable. Vale la pena volver a enfatizar que los coeficientes LNCC proporcionan un desempeño más constante en EER durante voz limpia y condiciones espectralmente variables, lo cual se puede apreciar de las desviaciones estándar 0.796, 0.660, 0.151, 0.486, producidas por MFCC, MFCC+CMN, LNCC y LNCC+CMN, respectivamente.

4.4.7. Resumen de resultados

A través de todos los experimentos, se observa que los *features* propuestos LNCC son competitivos tanto con los *features* MFCC o MFCC+CMN. También, se aprecia que la mejor elección de si utilizar CMN con los *features* MFCC, depende de las condiciones ambientales, en tanto que los *features* LNCC proporcionan consistentemente buen desempeño a lo largo de todas las condiciones y nunca sufren de tasas de error extremadamente altas, lo que se notó en algunos casos cuando se emplearon los MFCCs.

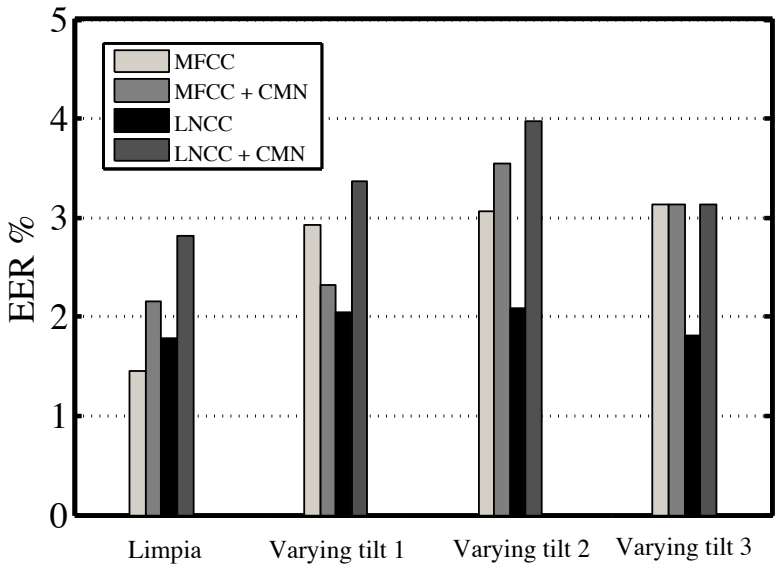

Figura 4.14: *Desempeño durante tilt espectral variable.* $VaryingST_1$*: 0 dB/oct* $\rightarrow -6$ *dB/oct.* $VaryingST_2$*: 0 dB/oct* $\rightarrow -6$ *dB/oct* $\rightarrow$ *0 dB/oct.* $VaryingST_3$*: 0 dB/oct* $\rightarrow -6$ *dB/oct* $\rightarrow$ *0 dB/oct* $\rightarrow -6$ *dB/oct.*

	Tasa de Igual Error (EER) %			
	Baselines		Propuestos	
Test data	MFCC	MFCC+CMN	LNCC	LNCC+CMN
Limpia	1.46	2.15	1.79	2.82
Tilt espectral −3 dB/octava	3.15	2.48	1.88	3.11
Tilt espectral −6 dB/octava	26.7	3.54	5.61	5.88
Tilt variable 0 to −6 dB/octava	2.93	2.32	2.04	3.37
Tilt variable 0 to -6 to 0 dB/octava	3.07	3.54	2.08	3.97
Tilt variable 0 to -6 to 0 to −6 dB/octava	3.13	3.13	1.81	3.14

Resumen de resultados. LNCCs calculados con 28 canales, d_{min}*=0.001 y B=3 Barks.*

4.5. Conclusiones

En este capítulo se propone una estrategia, motivada perceptualmente, efectiva y extremadamente simple, para normalizar instantáneamente *features* de voz. La efectividad de los *features* propuestos se demuestra para una tarea de verificación de locutor a lo largo de una variedad de condiciones de canal. Los Coeficientes Cepstrales Localmente-Normalizados, LNCCs no requieren el cálculo ni almacenamiento de un promedio móvil de los valores de *features*, y ellos proporcionan en algunos casos, reducciones relativas en EER tan altas como 32 % y 35 % cuando se comparan con MFCC y MFCC+CMN durante *tilt* espectral variable, respectivamente. Durante un *tilt* espectral estático de -6 dB/octava los coeficientes propuestos localmente-normalizados, dan una dramática reducción en EER tan alta como 79 % cuando

se compara con los coeficientes estándar MFCC. Aun cuando otras evaluaciones son necesarias de realizar como trabajo futuro, para llegar a resultados más consistentes, en principio, los *features* propuestos LNCC pueden ser una alternativa a MFCC y MFCC+CMN, en cualquier situación donde es difícil estimar confiablemente la media cepstral. Otras aplicaciones podrían incluir escenarios donde se desee una muy baja latencia o baja complejidad, donde calcular y almacenar el promedio móvil requerido por CMN pueda llegar a ser inconveniente. Como trabajo futuro también, se considera evaluar los *features* propuestos durante una tarea de reconocimiento automático de voz (ASR), aunque es posible que el banco de filtros auto-normalizado pueda eliminar una pequeña contidad de información fonética junto con la información del canal, así que ciertas modificaciones podrían ser necesarias para limitar la cantidad de normalización que se realice. En esa dirección, una línea obvia de investigación sería combinar LNCC con MFCCs o PLPs utilizando tanto combinación de *features* o combinación de sistemas.

Capítulo 5

Conclusiones

5.1. Análisis y discusiones finales

La presente tesis aborda dos tipos de problemas que involucran la robustez de un sistema de verificación de locutor texto-independiente, bajo condiciones de *mismatch* por ruido aditivo y por variabilidad en el canal de acústico de transmisión, en una tarea de reconocimiento de patrones acústicos basados en el sistema auditivo periférico. Se ha hecho uso de tres hipótesis formuladas en este trabajo las cuales han permitido, junto con los principales objetivos y la metodología, correspondientes a cada problema, poder responder a las preguntas de investigación incialmente formuladas. La motivación principal ha sido la de contribuir con dos novedosas estrategias, motivadas por el comportamiento de la audición en el ser humano en particular, y en los mamíferos en general, para obtener patrones acústicos robustos los que a su vez, al ser evaluados en un sistema de verificación de locutor, mejoren el desempeño del sistema en diversas condiciones de *mismatch*.

En primer lugar, esta tesis propuso una función sigmoidal óptima, tasa-nivel, que es una componente de muchos modelos del sistema auditivo periférico. La optimización hace uso de un conjunto de criterios definidos exclusivamente sobre la base de atributos físicos del sonido de entrada los que se inspiran en evidencia fisiológica. Los criterios desarrollados intentan discriminar entre una señal de voz degradada, y ruido para preservar la máxima cantidad de información en la región lineal de la curva sigmoidal y para minimizar los esfectos de distorsión en las regiones de saturación. El desempeño de la función sigmoidal óptima propuesta se valida con experimentos de verificación de locutor de texto-independiente, con señales degradadas por ruido aditivo a diferentes relaciones-señal-ruido (SNRs). Los resultados experimentales muestran que el método presentado en combinación con normalización de varianza cepstral (CVN) puede conducir a reducciones relativas en la tasa de error (EER) tan grandes como 40 % cuando se compara con el uso del sistema *baseline* de los coeficientes cepstrales para ciertas SNRs.

Luego, se propuso en esta tesis un nuevo conjunto de *features* llamados Coeficientes Cepstrales Localmente-Normalizados (LNCCs), que se basan en el Detector de Seneff de Sincronía Generalizada (GSD). Esta estrategia, motivada perceptualmente y extremadamente simple, pero no menos efectiva, permite normalizar instantáneamente *features* de voz. La efectividad de estos *features* se demuestra para una tarea de verificación de locutor a lo largo de una variedad de condiciones de canal. Los resultados alcanzados con verificación de locutor, texto independiente, muestran que, al ser comparados con el sistema *baseline* MFCC y con MFCC+CMN, los *features* propuestos LNCC se caracterizan por requerir de baja complejidad computacional y por compensar más ampliamente el *tilt* espectral que los coeficientes MFCC convencionales. Además, los *features* LNCC no requieren el cálculo y almacenamiento de un promedio móvil de valores de *features* proporcionando reducciones en EER tan altas como 32 % y 35 % cuando se comparan con MFCC y MFCC+CMN, con *tilt* espectral variable, respectivamente. Es interesante destacar que los *features* propuestos LNCC pueden llegar a ser una alternativa a MFCC y MFCC+CMN, en cualquier situación donde es difícil estimar confiablemente la media cepstral. Asimismo, otras posibles aplicaciones podrían suponer escenarios donde se desea una muy baja latencia o baja complejidad, esto es, donde calcular y almacenar el promedio móvil requerido por CMN, llegue a ser un serio inconveniente.

5.2. Trabajo futuro

Como trabajo futuro es posible proponer la evaluación de los *features* LNCCs durante una tarea de reconocimiento automático de voz (ASR), aunque se asume previamente que es posible que el banco de filtros auto-normalizado pueda eliminar una pequeña contidad de información fonética, junto con la información del canal, por lo cual, ante esta suposición inicial, podrían ser necesarias ciertas modificaciones para limitar la cantidad de normalización que se realice. En esa misma dirección, una línea natural de investigación podría ser llegar a combinar LNCC con otros *features* como por ejemplo, MFCCs o PLPs, utilizando tanto combinación de *features* como combinación de sistemas.

Bibliografía

Ahmed, N., Natarajan, T., Rao, K. R., 1974. Discrete cosine transform. IEEE Transactions on Computers. C-23(1), 90-93.

Ajmera, P. K., Jadhav, D. V., Holambe, R. S., 2011. Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram. Pattern Recognition. 44(10-11), 2749-2759.

Ali, A. M., Van Der Spiegel, J., Mueller, P., 2000. Auditory-based speech processing based on the average localized synchrony detection. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 1623-1626. Instanbul.

Ali, A. M., Van Der Spiegel, J., Mueller, P., 2002. Robust auditory-based speech processing using the average localized synchrony detection. IEEE Transactions on Speech and Audio Processing. 10(5), 279-292.

Allen, J. B., 1980. Cochlear micromechanics: A physical model of transduction. Journal of the Acoustical Society of America. 68(6), 1660-1670.

Allen, B. S., 1985. Cochlear modeling. IEEE ASSP Magazine. 2(1), 3-29.

Anderson, S., Skoe, E., Chandrasekaran, B., Kraus, N., 2010. Neural timing is linked to speech perception in noise. Journal of Neuroscience. 30, 4922-4926.

Atal, B. S., 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. Journal of the Acoustical Society of America. 55(6), 1304-1312.

Atal, B. S., 1976. Automatic recognition of speakers from their voices. Proceedings of the IEEE. 64(4), 460-475.

Barbour, D. L., 2011. Intensity-invariant coding in the auditory system. Neuroscience and Biobehavioral Reviews. 35, 2064-2072.

Becerra Yoma, N., Villar, M., 2002. Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm. IEEE Transactions on Speech and Audio Processing. 10(3), 158-166.

Becerra Yoma, N., Garretón, C., Molina, C., Huenupán, F., 2008. Unsupervised intra-speaker variability compensation based on Gestalt and model adaptation in speaker verification with telephone speech. Speech Communication. 50(11-12), 953-964.

Becerra Yoma, N., Garretón, C., Huenupán, F., Catalán, I., Wuth, J., 2013a. On reducing harmonic and sampling distortion in vocal tract length normalization. IEEE Transactions on Audio, Speech and Language Processing. 21(1), 108-119.

Becerra Yoma, N, Benavides, L., Wuth, J., Vivanco, H., 2013. Multicriteria-based computer-aided pronunciation quality evaluation of sentences. ETRI Electronics and Telecommunications Research Institute Journal. 35(1), 89-99.

Bell, P., Yamamoto, H., Swietojanski, P., Wu, Y., McInnes, F., Hori, C., Renals, S., 2013. A lecture transcription system combining neural network acoustic and language models. Proceedings of Interspeech 2013. 3087-3091. Lyon.

von Békésy, G., 1947. The variations of phase along the basilar membrane with sinusoidal vibrations. Journal of the Acoustical Society of America. 19, 452-460.

von Békésy, G., 1953. Description of some mechanical properties of the organ of Corti. Journal of the Acoustical Society of America. 25, 770-785.

von Békésy, G., 1960. Experiments in Hearing. McGraw-Hill, New York.

von Bertalanffy, L., 1976. General System Theory: Foundations, Development, Applications. George Braziller Inc., New York.

Bimbot, F., Bonastre, J., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega, J., Petrovska, D., Reynolds, D. A., 2004. A tutorial on text-independent speaker verification. Journal on Applied Signal Processing. 4, 430-451.

Bořil, H., Hansen, J. H. L., 2010. Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments. IEEE Transactions on Audio, Speech, and Language Processing. 18(6), 1379-1393.

Brandstein, M. and Ward, D., 2010. Microphone Arrays: Signal Processing Techniques and Applications. Digital Signal Processing. Springer.

Brumberg, J. S., Nieto-Castanon, A., Kennedy, P. R., Guenther, F. H., 2010. Brain-computer interfaces for speech communication. Speech Communication. 52(4), 367-379.

Buchner, H., Benesty, J., Kellermann, W., 2005. Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication. Signal Processing. 85(3), 549-570.

Bureš, Z., Grécová, J., Popelář, J., Syka, J., 2010. Noise exposure during early development impairs the processing of sound intensity in adult rats. European Journal of Neuroscience. 32(1), 155-164.

Campbell, Jr., J. P., Higgins, A., 1994. YOHO speaker verification. Linguistic Data Consortium. Philadelphia, PA.

Campbell, Jr, J. P., 1997. Speaker Recognition: A Tutorial. Proceedings of the IEEE. 85(9), 1437-1462.

Carey, M., Parris, E., Bridle, J., 1991. A speaker verification system using alpha-nets. In Proceedings of IEEE International Conference on Acoustics and Speech Signal Processing. 1, 397-400. Toronto.

Cariani, P., 1999. Temporal coding of periodicity pitch in the auditory system: An overview. Neural Plasticity. 6(4), 147-173.

Carlson, J. M., Doyle, J., 2002. Complexity and robustness. Proceedings of the National Academy of Sciences of the United States of America. 99(1), 2538-2545.

Chigier, B., Leung, H. C., 1992. The effects of signal representations, phonetic classification techniques, and the telephone network. In Proceedings of the Second International Conference on Spoken Language Processing. 97-100. Banff, Alberta.

Chiu, Y.-H. B., Stern, R. M., 2008. Analysis of physiologically-motivated signal processing for robust speech recognition. In Proceedings of Interspeech 2008, Brisbane, Australia, 1000-1003.

Chiu, Y. B., Raj, B., Stern, R. M., 2012. Learning-based auditory encoding for robust speech recognition. IEEE Transactions on Audio, Speech, and Language Processing. 20(3), 900-914.

Cooke, M., Hershey, J. R., Rennie, S. J., 2010. Monaural speech separation and recognition challenge. Computer Speech and Language. 24, 1-15.

Cohen, J. R., 1989. Application of an auditory model to speech recognition. Journal of the Acoustical Society of America. 85(6), 2623-2629.

Cohen, P. R., Oviatt, S. L., 1995. The role of voice input for human-machine communication. Proceedings of the National Academy of Sciences of the United States of America. 92(22), 9921-9927.

Cooke, M. and Lecumberri, M. L., 2012. The intelligibility of Lombard speech for non-native listeners. Journal of the Acoustical Society of America. 132(2), 1120-1129.

Cooke, M. and Mayo, C. and Valentini-Botinhao, C., 2013a. Intelligibility-enhancing speech modifications: the Hurricane Challenge. In Proceedings of Interspeech 2013. 3552-3556. Lyon.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Sauert, B., Tang, Y., 2013b. Evaluating the intelligibility benefit of speech modifications in known noise conditions. Speech Communication. 55, 572-585.

Cooke, M., King, S., Garnier, M., Aubanel, V., 2014. The listener talker: A review of human and algorithmic context-induced modifications of speech. Computer Speech and Language. 28(2), 543-571.

Costalupes, J. A., Young, E. D., Gibson, D. J., 1984. Effects of continuous noise backgrounds on rate response of auditory nerve fibers in cat. Journal of Neurophysiology. 51, 1326-44.

Cowper-Smith, C. D., Dingle, R. N., 2010. Synchronous auditory nerve activity in the carboplatin chinchilla model of auditory neuropathy. Journal of the Acoustical Society of America. 128(1), 56-62.

Dallos, P., 1992. The active cochlea. The Journal of Neuroscience. 12(12), 4575-85.

Damper, R., Higgins, J., 2003. Improving speaker identification in noise by subband processing and decision fusion. Pattern Recognition Letters. 24(13), 2167-2173.

Darwin, C.J., 2008. Listening to speech in the presence of other sounds. Philosophical Transactions of Royal Society B. Biological Science. 363(1493), 1011-1021.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing. 28(4), 357-366.

Davis, H., 1983. An active process in cochlear mechanics. Hearing Research. 9, 79-90.

Dean, I., Harper, N. S., McAlpine, D., 2005. Neural population coding of sound level adapts to stimulus statistics. Nature Neuroscience. 8(12), 1684-89.

Dean, I., Robinson, B. L., Harper, N. S., McAlpine, D., 2008. Rapid Neural Adaptation to Sound Level Statistics. Journal of Neuroscience. 28(25), 6430-6438.

Delgutte, B., Kiang, N. Y. S., 1984. Speech coding in the auditory nerve: I. Vowels-like sounds. Journal of the Acoustical Society of America. 75(3), 866-876.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B. 39(1), 1-38.

Deshpande, M., R. Holambe. 'AM-FM Based Robust Speaker Identification in Babble Noise', 2nd International Conference and workshop on Emerging Trends in Technology (ICWET), Proceedings published by International Journal of Computer Applications, 1-12, 2011.

Dimitriadis, D., Maragos, P., Potamianos, A., 2011. On the effects of filterbank design and energy computation on robust speech recognition. IEEE Transactions on Audio, Speech, and Language Processing. 19(6), 1504-1516.

Dreyer, A., Delgutte, B., 2006. Phase locking of auditory-nerve fibers to the envelopes of high frequency sounds: Implications for sound localization. Journal of Neurophysiology. 96(5), 2327-2341.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern Classification. Second Edition. A Wiley-Interscience Publication, New York.

Eggermont, J., 1977. Compound action potential tuning curves in normal and pathological human ears. Journal of the Acoustical Society of America. 62(5), 1247-1251.

Eggermont, J. J., 1998. Is there a neural code? Neuroscience and Biobehavioral Reviews. 22(2), 355-370.

Eggermont, J., 2001. Between sound and perception: reviewing the search for a neural code. Hearing Research. 157, 1-42.

Ehkan, P., Allen, T., Quigley, S. F., 2011. FPGA Implementation for GMM-Based Speaker Identification. International Journal of Reconfigurable Computing. 1-8, Volume 2011.

Emadi, G., Richter, C.P., Dallos, P., 2004. Stiffness of the gerbil basilar membrane: radial and longitudinal variations. Journal of Neurophysiology. 91, 474-488.

Engel, A. K., Fries, P., Singer, W., 2001. Dynamic predictions: Oscillations and synchrony in top-down processing. Nature. 2, 704-716.

Evans, E. F., 1992. Auditory Processing of Complex Sounds: An Overview. Journal of Neurophysiology. Philosophical Transactions of the Royal Society of London. Serie B. 336(1278), 295-306.

Fazel, A., Chakrabartty, S., 2011. An overview of statistical pattern recognition techniques for speaker verification. IEEE Circuits ans Systems Magazine. 62-81, Second Quarter.

Fisher, J. A., Nin, F., Reichenbach, T., Uthaiah, R., Hudspeth, A. J., 2012. The Spatial Pattern of Cochlear Amplification. Neuron. 76(5), 989-997.

Forsyth, M., 1995. Discriminating observation probability (DOP) HMM for speaker verification. Speech Communication. 17, 117-129.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech, and Signal Processing. 29(2), 254-272.

Furui, S., 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Transactions on Acoustics, Speech and Signal Processing. 34(1), 52-59.

Furui, S., 1994. An overview of the speaker recognition technology. Workshopon Automatic Speaker Recognition, Identification, and Verification. 2-9, Switzerland.

Furui, S., 1997. Recent advances in speaker recognition. Pattern Recognition Letters. 18(9), 859-872.

Furui, S., 2005. Recent progress in corpus-based spontaneous speech recognition. IEICE Transactions on Information and Systems. E88-D(3), 366-375.

Gales, M. J. F., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language. 12(2), 75-98.

Gao, F., Zhang, J., Sun, X., Chen, L., 2009. The effect of postnatal exposure to noise on sound level processing by auditory cortex neurons of rats in adulthood. Physiology & Behaviour. 97, 369-373.

García-Lázaro, J. A., Ho, S. S., Nair, A., Schnupp, J. W., 2007. Shifting and scaling adaptation to dynamic stimuli in somatosensory cortex. European Journal of Neuroscience. 26(8), 2359-68.

Garretón, C., Becerra Yoma, N., Torres, M., 2010. Channel robust feature transformation based on filter-bank energy filtering. IEEE Transactions on Audio, Speech and Language Processing. 18(5), 1082-1086.

Garretón, C., 2011. Robustez a la variavilidad de canal en reconocimiento de patrones acústicos con aplicaciones en enseñanza de idiomas y biometría. Universidad de Chile (Tesis de Doctorado).

Garretón, C., Becerra Yoma, N., 2012. Telephone channel compensation in speaker verification using a polynomial approximation in the log-filter-bank energy domain. IEEE Transactions on Audio, Speech and Language Processing. 20(1), 336-341.

Gaubitch, N. D., Brookes, M., Naylor, P. A., 2013. Blind channel magnitude response estimation in speech using spectrum classification. IEEE Transactions on Audio, Speech, and Language Processing. 21(10), 2162-2171.

Gauvain, J. L., Lee, C. H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing. 2, 291-298.

Ghitza, O., 1986. Auditory nerve representation as a front-end for speech recognition in a noisy environment. Computer Speech and Language. 1, 109-130.

Ghitza, O.,1994. Auditory models and human performance in tasks related to speech coding and speech recognition. IEEE Transactions on Speech and Audio Processing. 2(1), 115-132.

Gillespie, P. G., 2004. Myosin I and adaptation of mechanical transduction by the inner ear. Philosophical Transactions of the Royal Society of London. 359, 1945-51.

Gillick, L. and Cox, S., 1989. Some statistical issues in the comparison of speech recognition algorithms. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 532-535. Glasgow.

Glass, J., Hazen, T. J., Cyphers, S., Malioutov, I., Huynh, D., Barzilay, R., 2007. Recent progress in the MIT spoken lecture processing project. In Proceedings of Interspeech 2007. 2553-2556. Antwerp.

Gold, T., 1948. Hearing. The physical basis of the action of the cochlea. Proceedings of the Royal Society of Edinburgh. Biological Science. 135, 492-498.

Hain, T., Burget, L., Karafiat, M., Garau, G., Lincoln, M., Renals, S., Dines, J., Moore, D., McCowan, I., Vepa, J., Wan, V., Oerdelman, R., van Leeuwen, D., 2006. The AMI meeting transcription system. In Proceedings of the NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation Workshop.

Hain, T., Burget, L., Dines, J., Garau, G., Wan, V., Karafiat, M., Vepa, J., Lincoln, M., 2007. The AMI system for the transcription of speech in meetings. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 357-360. Honolulu

Hain, T., Burget, L., Dines, J., Garner, P. N., Grezl, F., Hannani, A. E., Huijbregts, M., Karafiat, M., Lincoln, M., Wan, V., 2012. Transcribing meetings with the AMIDA system. IEEE Transactions on Audio, Speech, and Language Processing. 20(2), 486-498.

Hanilçi, C., Kinnunen, T., Ertaş, F., Saeidi, R., Pohjalainen, J., Alku, P. Regularized all-pole models for speaker verification under noisy environments. IEEE Signal Processing Letters. 19(3), 163-166.

Hansen, J. H. L., 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech Communication. 20,151-173.

Hansen, J. H. L., Varadarajan, V., 2009a. Analysis and compensation of Lombard speech acroos noise type and levels with application to In-Set/Out-of-Set speaker recognition. IEEE Transactions on Audio, Speech, and Language Processing. 17(2), 366-378.

Hansen, J. H. L., Womack, B. D., 2009b. Feature analysis and neural network-based classification on speech under stress. IEEE Transactions on Speech and Audio Processing. 2(4), 307-313.

Hasan, T., Hansen, J. H. L., 2013. Acoustic factor analysis for robust speaker verification. IEEE Transactions on Audio, Speech, and Language Processing. 21(4), 842-853.

Hauser, M. D., Chomsky, N., Fitch, H., 2002. The faculty of language: What is it, who has it, and how did it evolve? Science. 298, 1569-1579.

He, W., Ren, T., 2013. Basilar membrane vibration is not involved in the reverse propagation of otoacoustic emissions. Scientific Reports. 3, 1-7.

Heinz, M. G., Colburn, H. S., Carney, L. H., 2001. Rate and timing cues associated with cochlear amplifier: level discrimination based on monoaural cross-frequency coincidence detection. Journal of the Acoustical Society of America. 110(4), 2065-2084.

Heinz, M. G. and Swaminathan, J., 2009. Quantifying envelope and fine-structure coding in auditory-nerve responses to chimaeric speech. Journal of the Association for Research in Otolaryngology. 10(3), 407-423.

Hermansky, H., 1990. Perceptual linear predictive PLP analysis of speech. Journal of the Acoustical Society of America. 87(4), 1738-1752.

Hermansky, H., Morgan, N., Bayya, A., Khon, P., 1991a. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In Proceedings of Eurospeech. 1367-1370. Genova.

Hermansky, H., Morgan, N., Bayya, A., Khon, P., 1991b. (RASTA-PLP) speech analysis technique. In Proceedings International Conference on Acoustics, Speech, and Signal Processing. 121-124. San Francisco.

Hermansky, H., 1994. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing. 2(4), 578-589.

Hermansky, H., Cohen, J., Stern, R. M., 2013. Perceptual properties of current speech recognition technology. Proceedings of the IEEE. 101(9), 1968-1985.

Higgins, A., Bahler, L., Porter, J., 1991. Speaker verification using randomized phrase prompting. Digital Signal Processing. 1(2), 89-106.

Holmberg, M., Gelbart, D., Hemmert, H., 2007. Speech encoding in a model of peripheral auditory processing: Quantitative assessment by means of automatic speech recognition. Speech Communication. 49, 917-932.

Hori, T., Fujimoto, M., Ogawa, A., Kinoshita, K., Nakamura, A., 2012. Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera. IEEE Transactions on Audio, Speech, and Language Processing. 20(2), 499-513.

Houtgast, T., 1972. Psychophysical evidence for lateral inhibition in hearing. Journal of the Acoustical Society of Amererica. 68, 1885-1894.

Hsu, B. J., Glass, J., 2006. Style and topic language model adaptation using HMM-LDA. In Proceedings of the Conference on Empirical Methods in Natural Processing. 373-381. Sydney.

Hsu, C. W., Lee, L. S., 2009. Higher order Cepstral moment normalization for improved robust speech recognition. IEEE Transactions on Audio, Speech and Language Processing. 17(2), 205-220.

Hudspeth, A. J., 1989. How the ear's works work? Nature. 341, 397-404.

Hudspeth, A. J., 2008. Making an effort to listen: Mechanical amplification in the ear. Neuron. 59, 530-545.

Huenupán, F., 2010. Fusión de múltiples clasificadores en verificación de locutor. Universidad de Chile (Tesis de Doctorado).

Ishi, C. T., Matsuda, S., Kanda, T., Jitsuhiro, T., Ishiguro, H., Nakamura, S., Hagita, N., 2008. A robust speech recognition system for communication robots in noisy environments. IEEE Transactions on Robotics. 24(3), 759-763.

Jankowski, C. R., Lippmann, R. P., 1992. Comparison of auditory model for robust speech recognition. In Proceedings of the Workshop on Speech and Natural Language, Stroudsburg, PA, pp. 453-454.

Jankowski, C. R., Vo, H. D., Lippmann, R. P., 1995. A comparison of signal processing front ends for automatic word recognition. IEEE Transactions on Speech and Audio Processing. 3, 286-293.

Jensen, B., Tomatis, N., Drygajlo, A., Siegwart, R., 2005. Robots meet human interaction in public spaces. IEEE Transactions on Industrial Electronics. 52(6), 1530-1546.

Jeon, W., Juang, B. H., 2007. Speech analysis in a model of the central auditory system. IEEE Transactions on Audio, Speech, and Language Processing. 15(6), 1802-1817.

Jin, Q., 2007. Robust Speaker Recognition. Carnegie Mellon University, Pittsburgh (Ph.D. Thesis).

Johnson, D., 1980. The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. Journal of the Acoustical Society of America. 68(4), 1115-1122.

Joris, P., Yin, T., 2007. A matter of time: Internal delays in binaural processing. Trends in Neuroscience. 30(2), 70-78.

Joris, P. X., Bergevin, C., Kalluri, R., Mc Laughlin, M., Michelet, P., van der Heijden, M., Shera, C. A., 2011. Frequency selectivity in Old-World monkeys corroborates sharp cochlear tuning in humans. Proceedings of the National Academy of Sciences of the United States of America. 108(42), 17516-20.

Kang, S. Y., Colesa, D. J., Swiderski, D. L., Su, G. L., Raphael, Y., Pfingst, B. E., 2010. Effects of hearing preservation on psychophysical responses to cochlear implant stimulation. Journal of the Association for Research in Otolaryngology. 11(2), 245-65.

Kayser, C., Montemurro, M. A., Logothetis, N. K.,Panzeri, S., 2009. Spike-phase coding boost and stabilizes information carried by spatial and temporal spike patterns. Neuron. 61(4), 597-608.

Kiang, N. Y. S., Watanabe, T., Thomas, E. C., Clark, L. F., 1965. Discharge patterns of single fibers in the cat's auditory nerve. Research Monograph No. 35, MIT Press, Cambridge, MA.

Kiang, N. Y. S., Moxon, E. C., 1974. Tails of tuning curves of auditory-nerve fibers. Journal of the Acoustical Society of America. 55(3), 620-630.

Kim, D. S., Lee, S. Y., Kil, R. M., 1999. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. IEEE Transactions on Speech and Audio Processing. 7(1), 55-69.

Kim, C., Chiu, Y., Stern, R. M., 2006. Physiologically motivated synchrony based processing for robust automatic speech recognition. In Proceedings of Interspeech 2006. 1483-1486.

Kim, C., Stern, R. M., 2012. Power normalized cepstral coefficients (PNCC) for robust speech recognition. In Proceedings Acoustics, Speech and Signal Processing. 4101-4104.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: from features to supervectors. Speech Communication. 52(1), 12-40.

Kinnunen, T., Sidoroff, I., Tuononen, M., Franti, P., 2011. Comparison of clustering methods: A case study of text-independent speaker modeling. Pattern Recognition Letters. 32, 1604-1617.

Kinnunen, T., Saeidi, R., Sedlak, F., Lee, K. A., Sandberg, J., Hansson-Sandsten, M., Li, H., 2012. Low-variance multitaper MFCC features: A case study in robust speaker verification. IEEE Transactions on Audio, Speech, and Language Processing. 20 (7), 1990-2001.

Kitano, H., 2004. Biological robustness. Nature Reviews Genetics. 5(11), 826-837.

Kumar, K., Kim, C. & Stern, R. M., 2011. Delta-spectral cepstral coefficients for robust speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 1-4. Prague, Czech Republic.

Kumaresan, R., Rao, A., 2011. Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications. Journal of the Acoustical Society of America. 105(3), 1912-1924.

Kuwabara, H., Sagisaka, Y., 1995. Acoustics characteristics of speaker individuality: control and conversion. Speech Communication. 16(2), 165-173.

Lamel, L. F., Rabiner, L. R., Rosenberg, A. E., Wilpon. J. G., 1981. An improved endpoint detection detector for isolated word recognition. IEEE Transactions on Acoustics Speech and Language Proceesing. 29, 777-785.

Larson, E., Billimoria, C. P., Sen, K., 2009. A Biologically plausible computational model for auditory object recognition. Journal of Neurophysiology. 101, 323-331.

Lebedev, M. A., Nicolelis, M. A. L., 2006. Brain-machine interfaces: past, present and future Trends in Neurosciences. 29(9), 536-546.

Leeuwis, E., Federico, M., Cettolo, M., 2003. Language modeling and transcription of the TED corpus lecture. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 232-235. Toulouse

Lei, Y., Hansen, J. H. L., 2011. Mismatch modeling and compensation for robust speaker verification. Speech Communication. 53(2), 257-268.

Leus, G., Moonen, M., 2003. Deterministic subspace based blind channel estimation for doubly-selective channels. In Proceedings of the 4th IEEE Workshop on Signal Processing Advances in Wireless Communications. 210-214.

Leus, G., 2004. On the estimation of rapidly time-varying channels. In Proceedings of European Signal Processing Conference. 2227-2230. Vienna

Li, K. P., Porter, J. E., 1988. Normalizations and selection of speech segments for speaker recognition scoring. In Proceedings of IEEE International Conference on Acoustics and Speech Signal Processing. 1, 595-598.

Li, Q., Huang, Y., 2010. Robust speaker identification using an auditory based feature. In Proceedings of IEEE International Conference on Acoustics and Speech Signal Processing. 19(6), 4514-4517.

Li, Q., Huang, Y., 2011. An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. IEEE Transactions on Audio, Speech, and Language Processing. 19(6), 1791-1801.

Liberman, M. C., 1978. Auditory nerve response from cats raised in a low noise chamber. Journal of the Acoustical Society of America. 63(2), 442-455.

Liberman, M. C., 1980. Morphological differences among radial afferent fibers in the cat cochlea: An electron-microscopic study of serial sections. Hearing Research. 3,45-63.

Lighthill, J., 1981. Energy flow in the cochlea. Journal of Fluid Mechanics. 106,149-213.

Lighthill, J., 1991. Biomechanics of hearing sensitivity. Journal of Vibrations and Acoustics. 113,1-13.

Liu, F., Stern, R. M., Huang, X., Acero, A., 1993. Efficient cepstral normalization for robust speech recognition. In Proceedings DARPA Speech and Natural Language Workshop. 69-74. Cambridge.

Lu, X., Unoki, M., Nakamura, S., 2011. Sub-band temporal modulation envelopes and their normalization for automatic speech recognition in reverberant environments. Computer Speech and Language. 25(3), 571-584.

Lyon, R., 1982. A computational model of filtering, detection and compression in the cochlea. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. 1282-1285, Paris.

Majekodunmi, T, Idachaba, F., 2011. A review of the fingerprint, speaker recognition, face recognition and iris recognition based on biometric identification technologies. Proceedings of the World Congress on Engineering, Vol II, 1-7, London.

Mak, B. K. W., Tam, Y. C., Qi Li, P., 2004. Discriminative auditory-based features for robust speech recognition. IEEE Transactions on Audio, Speech and Language Processing. 12(1), 27-36.

Malionek, J., Oard, D. W., Sangwan, A., Hansen, J. H. L., 2013. Linking Transcribed Conversational Speech. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. 961-964. Dublin.

Manley, G. A., 2000. Cochlear mechanisms from a phylogenetic viewpoint. Proceedings of the National Academy of Sciences of the United States of America. 97, 11736-43.

Manley, G. A., 2001. Evidence for an active process and a cochlear amplifier in nonmammals. Journal of Neurophysiology. 86, 541-549.

Matsui, T., Furui, S., 1994. Similarity normalization methods for speaker verification based on posteriori probability. In Proceedings First ESCA Workshop on Automatic Speaker Recognition, Identification and Verification. 59-62, Martigny, Switzerland.

May, B. J., Sachs M. B., 1992. Dynamic range of neural rate responses in the ventral cochlear nucleus of awake cats. Journal of Neurophysiology. 68(5), 1589-1602.

Meddis, R., Lopez-Poveda, E. A., 2010. Auditory Periphery: From Pinna to Auditory Nerve. In Computational Models of the Auditory System. Springer Verlag. 7-38.

Mesgarani, N., Chang, E. F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature. 485, 232-237.

Meyer, B., Kollmeier, B., 2011. Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. Speech Communication. 53(5), 753-767.

Middlebrooks, J. C., 2004. Effects of cochlear-implant pulse rate and inter-channel timing on channel interactions and thresholds. Journal of the Acoustical Society of America. 116, 452-468.

Miettinen, I., Alku, P., Salminen, N., May, P. J. C., Tiitinen, H., 2011. Responsiveness of the human auditory cortex to degraded speech sounds: Reduction of amplitude resolution vs. additive noise. Brain Research. 1367, 298-309.

Miller, C. A., Woo, J., Abbas, P. J., Hu, N., Robinson, B. K., 2011. Neural Masking by Subthreshold Electric Stimuli: Animal and Computer Model Results. Journal of the Association for Research in Otolaryngology. 12(2), 219-232.

Ming, J., Hazen, T. J., Glass, J. R., Reynolds, D. A., 2007. Robust speaker recognition in noisy conditions. IEEE Transactions on Audio, Speech, and Language Processing. 15(5), 1711-1723.

Moore, R.K., Cutler, A., 2001. Constraints on theories of human vs. machine recognition of speech. In Proceedings of the SPRAAC Workshop on Human Speech Recognition as Pattern Classification. Max-Planck-Institute for Psycholinguistics, Nijmegen, 145-150.

Moore, B. C. J., 1996. Perceptual consequences of cochlear hearing loss and their implications for the design of hearing aids. Ear & Hearing. 17, 133-161.

Moore, B. C. J., 2002. Interference effects and phase sensitivity in hearing. Philosophical Transactions of the Royal Society of London. Serie A. 360, 833-858.

Moore, B. C. J., 2003a. An Introduction to the Psychology of Hearing. Fifth Edition. Academic Press. Elsevier Science, USA.

Moore, B. C. J., 2003b. Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants. Otology & Neurotology. 24, 243-254.

Moore, B. C. J., 2008a. Basic auditory processes involved in the analysis of speech sounds. Philosophical Transactions of the Royal Society of London. Serie B. Biological Science. 363(1493), 947-963.

Moore, B. C. J., Tyler, L. K., Marslen-Wilson, W., 2008b. The perception of speech: from sound to meaning. Philosophical Transactions of the Royal Society. Serie B. Biological Science. 363(1493), 917-921.

Moore, B. C. J., 2008c. The rol of temporal fine structure processing in pitch perception, masking, and speech perception for normal hearing and hearing-impaired people. Journal of the Association for Research in Otolaryngology. 9, 399-406.

Moore, B. C. J., 2014. Auditory Processing of Temporal Fine Structure: Effects of Age and Hearing Loss. Audiology and Otology, World Scientific Publishing CO PTE LTD, UK.

Morales, N., Toledano, D., Hansen, J. H. L., Garrido, J., 2009. Feature compensation techniques for ASR on band-limited speech. IEEE Transactions on Audio, Speech and Language Processing. 17(4), 758-774.

Nakano, A. Y., Nakagawa, S., Yamamoto, K., 2010. Distant speech recognition using a microphone array network. E93.D(9), 2451-2462.

Nelken, I., 2008. Processing of complex sounds in the auditory system. Current Opinion in Neurobiology. 18, 413-417.

Nemala, S. & Elhilali, M., 2011. Multistream Robust Speaker Recognition Based on Speech Intelligibility. Annual Conference on Information Sciences and Systems, 1-5. Baltimore, MD.

Nizami, L., 2005. Dynamic range relations for auditory primary afferents. Hearing Research. 208(1-2), 26-46.

Ohshima, Y., Stern, R. M., 1994. Environmental robustness in automatic speech recognition using physiologically-motivated signal processing. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 1-4. Adelaide

Ohzawa, I., Sclar, G., Freeman, R. D., 1985. Contrast gain control in the cat's visual system. Journal of Neurophysiology. 54(3), 651-67.

Oxenham, A. J., Bacon, S. P., 2003. Cochlear compression: Perceptual measures and implications for normal and impaired hearing. Ear & Hearing. 24(5), 352-366.

Paliwal, K. K., Wojcicki, K., Shannon, B. J., 2011. The importance of phase in speech enhancement. Speech Communication. 53, 465-494.

Parbery-Clark, A., Anderson, S., Kraus, N., 2013. Musicians change their tune: How hearing loss alters the neural code. Hearing Research. 302, 121-131.

Parikh, G., Loizou, P. C., 2005. The influence of noise of vowel and consonant cues. Journal of the Acoustical Society of America. 18(6), 3874-3888.

Park, A., Hazen, T., Glass, J., 2005. Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 497-500. Philadelphia.

Patterson, R. D., Holdsworth, J., Allerhand, M., 1992. Auditory Models as Preprocessors for Speech Recognition. In Marten Egbertus Hendrik Schouten. The auditory processing of speech: From sounds to words. Walter de Gruyte.

Pearce, D., Hirsch, H., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In Proceedings of the International Conference on Spoken Language Processing (ICSLP). 4, 29-32, Beijing.

Pfingst, B. E, Bowling, S. A., Colesa, D. J., Garadat, S.N., Raphael, Y., Shibata, S. B., Strahl, S. B., Su, G. L., Zhou N., 2011. Cochlear infrastructure for electrical hearing. Hearing Research. 281(1-2), 65-73.

Pickles, L., 2008. An Introduction to the Physiology of Hearing. Third Edition. Emerald Group Publishing Limited, UK.

Picone, J., 1993. Signal modeling techniques in speech recognition. Proceedings of the IEEE. 81, 1215-1247.

Poblete, V., Becerra Yoma, N., Stern, R. M., 2014. Optimization of the parameters characterizing sigmoidal rate-level functions based on acoustic features. Speech Communication (Elsevier). 56, 19-34.

Poblete, V., Espic, F., King, S., Stern, R. M., Huenupán, F., Becerra Yoma, N., 2014. A perceptually-motivated low-complexity instantaneous channel normalization technique applied to speaker verification. Submitted to Computer Speech and Language (Elsevier), February 2014.

Pradhan, G., Prasanna, S., 2011. Speaker verification under degraded condition: a perceptual study. International Journal of Speech Technology. 14, 405-417.

Qi Li, P., Zheng, J., Tsai, A., Zhou, Q., 2002. Robust end-point detection and energy normalization for real-time speech and speaker recognition. IEEE Transactions on Speech and Audio Processing. 10(3), 146-157.

Qi Li, P., Huang, Y., 2010. Robust speaker identification using an auditory-based feature. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 4514-4517. Dallas.

Qi Li, P., Huang, Y., 2011. An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. IEEE Transactions on Audio, Speech and Language Processing. 19(6), 1791-1801.

Qin, L., Wang, J. Y., Sato, Y., 2008. Representations of Cat Meows and Human Vowels in the Primary Auditory Cortex of Awake Cats. Journal of Neurophysiology. 99, 2305-2319.

Rabinowitz, N. C., Willmore, B., Schnupp, J., King, A. J., 2011. Contrast Gain Control in Auditory Cortex. Neuron. 70(6), 1178-91.

Ren, T., 2002. Longitudinal pattern of basilar membrane vibration in the sensitive cochlea. Proceedings of the National Academy of Sciences of the United States of America. 99, 17101-17106.

Renals, S., Hain, T., Bourlard, H., 2007. Recognition and understanding of meetings: The AMI and AMIDA Projects. In Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU) 2007. 238-247. Kyoto

Reynolds, D. A., 1995. Speaker identification and verification using Gaussian mixture speaker models. Speech Communication. 17, 91-108.

Reynolds, D. A., Rose, R. C., 1996. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing. 3(1), 72-83.

Reynolds, D. A., 1997. Comparison of background normalization methods for text-independent speaker verification. In Proceedings of the European Conference on Speech Communication and Technology, 963-966, Rhodes, Greece.

Reynolds, D. A., Quatieri, T. F., Dunn, R. B., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing 10, 19-41.

Robles, L., Ruggero, M., 2001. Mechanics of the mammalian cochlea. Physiological Reviews. 81(3), 1305-1352.

Rhode, W., 1971. Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique. Journal of the Acoustical Society of America. 49, 1218-1231.

Rhode, W., 1978. Some observations on cochlear mechanics. Journal of the Acoustical Society of America. 64(1), 158-176.

Rhode, W. S., Cooper, N. P., 1993. Two-tone suppression and distortion production on the basilar membrane in the hook region of cat and guinea pig cochleae. Hearing Research. 66(1), 31-45.

Rose, J. E., Hind, J. E., Anderson, D. J., Brugge, J. F., 1971. Some effects of stimulus intensity on response of auditory nerve fibers in the squirrel monkey. Journal of Neurophysiology. 34, 685-699.

Rose, R. C., Reynolds, D. A., 1990. Text-independent speaker identification using automatic acoustic segmentation. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 293-296. Albuquerque.

Rose, P., 2002. Forensic Speaker Identification. Taylor and Francis, London.

Rosen, S., 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. Philosophical Transactions of the Royal Society of London. Serie B. 336, 367-373.

Ruggero, M. A., Rich, N. C., Recio, A. & Narayan, S. S., 1997. Basilar-membrane responses to tones at the base of the chinchilla cochlea. Journal of the Acoustical Society of America. 101, 2151-2163.

Ruggero, M. A., Temchin, A. N., 2005. Unexceptional sharpness of frequency tuning in the human cochlea. Proceedings of the National Academy of Sciences of the United States of America. 102, 18614-19.

Russel, I. J., Nilsen, K. E., 1997. The location of the cochlear amplifier: Spatial representation of a single tone on the guinea pig basilar membrane. Proceedings of the National Academy of Science of the United States of America 94, 2660-2664.

Sachs, M. B., Kiang, N. Y. S., 1968. Two-tone inhibition in auditory-nerve fibers. Journal of the Acoustical Society of America. 43, 1120-1128.

Sachs, M. B., Abbas, P. J., 1974. Rate versus level functions for auditory-nerve fiber in cats: tone burst stimuli. Journal of the Acoustical Society of America. 56(6), 1835-47.

Sachs, M. B., Young, E. D., 1979. Encoding of steady state vowels in the auditory nerve: Representation in terms of discharge rate. Journal of the Acoustical Society of America. 66(2), 470-479.

Sachs, M. B., Young, E. D., 1980. Effects of nonlinearities on speech encoding in the auditory nerve. Journal of the Acoustical Society of America. 68(3), 858-875.

Sachs, M. B., 1984. Neural coding of complex sounds: Speech. Annual Review of Physiology. 46, 261-273.

Saeidi, R., Pohjalainen, J., Kinnunen, T., Alku, P., 2010. Temporally Weighted Linear Prediction Features for Tackling Additive Noise in Speaker Verification. IEEE Signal Processing Letters. 17(6), 599-602.

Sangwan, A., Kaushik, L., Yu, C., Hansen, J. H. L., Oard, D. W., 2013. Houston, we have a solution: using NASA Apollo Program to advance speech and language procesing technology. In Proceedings of Interspeech 2013. 1135-1139. Lyon.

Savoji, M. H., 1989. A robust algorithm for accurate endpointing of speech signals. Speech Communication. 8(1), 45-60.

Schneider, T. R., Lorenz, S., Senkowski, D., Engel, A. F., 2011. Gamma-band activity as a signature for cross-modal priming of auditory object recognition by active haptic exploration. The Journal of Neuroscience. 31(7), 2502-2510.

Schwartz, R., Anastasakos, T., Kubala, F., Makhoul, J., Nguyen, L., Zavaliagkos, G., 1993. Comparative experiments on large vocabulary speech recognition. In Proceedings of the Workshop on Human Language Technology. 75-80. Princeton.

Seltzer, M. L., Raj., B., Stern, R. M., 2004. Likelihood-maximizing beamforming for robust hands-free speech recognition. IEEE Transactions on Speech and Audio Processing. 12, 489-498.

Seneff, S., 1984. Pitch and spectral estimation of speech based on an auditory synchrony model. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. 1-4. San Diego.

Seneff, S., 1985. Pitch and spectral analysis of speech based on an auditory synchrony model. PhD. Dissertation, Massachusetts Institute of Technology, Cambridge.

Seneff, S., 1986a. A computational model for the peripheral auditory system: application to speech recognition research. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. 1983-1986. Tokyo.

Seneff, S., 1986b. Characterizing formants through straight line approximations without explicit formant tracking. In Proceedings of the First Montreal Symposium on Speech Recognition. 21-27. Montreal.

Seneff, S., 1987. Vowel recognition based on line-formants derived from an auditory-based spectral. In Proceedings of the 11th International Congress of Phonetic Sciences. 1-4. Tallin.

Seneff, S., 1988. A joint synchrony/mean rate model of auditory speech processing. Journal of Phonetics. 16, 55-76.

Shamma, S. A., 1985. Speech processing in the auditory system: The representation of speech sounds in the responses of the auditory nerve. Journal of the Acoustical Society of America. 78(5), 1612-1621.

Shamma, S. A., 1985. The acoustic features of speech sounds in a model of auditory processing: vowels and voiceless fricatives. Journal of Phonetics. 16, 77-91.

Shannon, B. J., Paliwal, K. K., 2003. A Comparative study of filter bank spacing for speech recognition. In Proceedings of the Microelectronic Engineering Research Conference. 1-3. Brisbane, Australia.

Shao, Y., Srinivasan, S., Wang, D. L., 2007. Incorporating auditory feature uncertainties in robust speaker identification. In Proceedings of IEEE International Conference on Acoustics and Speech Signal Processing. 4, 277-280.

Shao, Y., Wang, D. L., 2008. Robust speaker identification using auditory features and computational auditory scene analysis. In Proceedings of IEEE International Conference on Acoustics and Speech Signal Processing. 1589-1592.

Shao, Y., Srinivasan, S., Jin, Z., Wang, D. L., 2010. A computational auditory scene analysis system for speech segregation and robust speech recognition. Computer Speech and Language. 24(1), 77-93.

Shin, J. W., Kwon, H. J., Jin, S. H., Kim, N. S., 2008. Voice activity detection based on conditional MAP criterion," IEEE Signal Processing Letters. 15(2), 257-260.

Sinex, D. G., Geisler, D., 1983. Responses of primary auditory fibers to consonant-vowel syllables. Journal of the Acoustical Society of America. 602-615.

Sinex, D., Guzik, H., Li, H., Sabes, J., 2003. Responses of auditory nerve fibers to harmonic and mistuned complex tones. Hearing Research. 182, 130-139.

Skowronski, M., Harris, J., 2004. Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition. Journal of the Acoustical Society of America. 116(3), 1774-1780.

Slaney, M., 1998. Auditory toolbox. Version 2. Technical report 1998-010. Interval Research Corporation.

Smith, Z. M., Delgutte, B., Oxenham, A. J., 2002. Chimaeric sounds reveal dichotomies in auditory perception. Nature. 416, 87-90.

Soong, F. K., Rosenberg, A. E., 1988. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Transactions on Acoustics, Speech and Signal Processing. 36(6), 871-879.

Sumner, C. J., Palmer, A. R., 2012. Auditory nerve fibre responses in the ferret. European Journal of Neuroscience. 36(4), 2428-39.

Stern, R. M., Wang, D. L., Brown, G. J., 2006. Binaural sound localization. Computational Auditory Scene Analysis, Chapter 5. Wang, D. L. and Brown, G. J. Editors. Wiley-IEEE Press.

Stern, R. M., Morgan, N., 2012. Hearing is believing. Biologically inspired methods for robust automatic speech recognition. IEEE Signal Processing Magazine. 29(6), 34-43.

Stern, R. M., Morgan, N., 2012b. Features based on auditory physiology and perception. In Techniques for Noise Robustness in Automatic Speech Recognition, Virtanen, T., Raj, B. and Singh, R., Eds., New York, NY, USA: Wiley. 207-243.

Stockham, T. G., Cannon, T. N., Ingebretsen, R. B., 1975. Blind deconvolution through digital signal processing. Proceedings of the IEEE. 63(4), 678-693.

Syka, J., 2002. Plastic changes in the central auditory system after hearing loss, restoration of function, and during learning. Physiological Review. 82, 601-636.

Taberner, A. M., Liberman, M. C., 2005. Response properties of single auditory nerve fibers in the mouse. Journal of Neurophysiology. 93(1), 557-569.

Tchorz, J., Kleinschmidt, M., Kollmeier, B., 1996. A psychoacoustical model of the auditory periphery as a front end for ASR. Journal of the Acoustical Society of America. 105(2), 1157-1157.

Tchorz, J., Kollmeier, B., 1999. A model of auditory perception as front end for automatic speech recognition. Journal of the Acoustical Society of America. 108(4), 2040-2050.

Togneri, R., Pullella, D., 2011. An overview of speaker identification: Accuracy and robustness issues. IEEE Circuits and Systems Magazine. 11(2), 23-61.

Tokuda, K., Yoshimura, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. 1315-1318. Instanbul.

Trancoso, I., Nunes, R., Neves, L., 2006. Classroom lecture recognition. Computational Processing of the Portuguese Language, Proceedings Book Series: Lecture Notes in Artificial Intelligence. 3960, 190-199.

Tranter, S., D. A. Reynolds, 2006. An overview of automatic speaker diarization systems. IEEE Transactions on Audio, Speech and Language Processing. 14(5), 1557-1565.

Tremblay, K., Kraus, N., Carrell, T. D., McGee, T., 1997. Central auditory system plasticity: Generalization to novel stimuli following listening training. Journal of the Acoustical Society of America. 102(6), 3762-3773.

Uysal, I., Sathyendra, H., Harris, J., 2008. Can spike based speech recognition systems outperform conventional approaches?. The Neuromorphic Enginee, A Publication of INEWEB.ORG. 1-3.

Varela, F., Lachaux, J. P., Rodriguez, E., Martinerie, J., 2001. The brainweb: Phase synchronization and large-scale integration. Nature Reviews, Neuroscience. 2, 229-239.

Viemeister, N. F., 1988. Intensity coding and the dynamic range problem. Hearing Research. 34(3), 267-74.

Wang, K., Shamma, S., 1994. Self-normalization and noise-robustness in early auditory representations. IEEE Transactions on Speech and Audio Processing. 2(3), 421-435.

Wang, L., Kitaoka, N., Nakagawa, S.,2007. Robust distant speech recognition by combining position-dependent CMN with conventional CMN. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. 817-820. Honolulu.

Wang, N., Ching, P., Zheng, N., Lee, T., 2011. Robust speaker recognition using denoised vocal source and vocal tract features. IEEE Transactions on Speech and Audio Processing. 19(1), 196-205.

Watkins, A. J., Makin, S. J., 1996. Some effects of filtered contexts on the perception of vowels and fricatives. Journal of the Acoustical Society of America. 99(1), 588-594.

Watkins, P. V., Barbour, D. L., 2011. Level-tuned neurons in primary auditory cortex adapt differently to loud versus soft sounds. Cerebral Cortex. 21(1), 178-90.

Wen, B., Wang, G. I., Dean, I., Delgutte, B., 2009. Dynamic range adaptation to sound level statistics in the auditory nerve. The Journal of Neuroscience. 29(44), 13797-13808.

Wen, B., Wang, G. I., Dean, I., Delgutte, B., 2012. Time course of dynamic range adaptation in the auditory nerve. Journal of Neurophysiology. 108(1), 69-82.

Werblin, F. S., Jacobs, A., Teeters, J., 1996. The computational eye. IEEE Spectrum. 33(5), 30-37.

Wever, E. G., 1949. Theory of Hearing. Dover Publications: New York, 1970 edition.

Wilson, J. P., 1980. Evidence for a cochlear origin for acoustic -re-emissions, threshold fine-structure and tonal tinnitus. Hearing Research. 2, 233-252.

Winslow, R. L., Sachs, M. B., 1987. Effect of electrical stimulation of the crossed olivocochlear bundle on auditory nerve response to tones in noise. Journal of Neurophysiology. 57(4), 1002-1021.

Wölfel, M., 2009a. Enhanced speech features by single-channel joint compensation of noise and reverberation. IEEE Transactions on Audio, Speech and Language Processing. 17(2), 312-323.

Wölfel, M., 2009b. Signal adaptive spectral envelope estimation for robust speech recognition. Speech Communication. 51(6), 551-561.

Wölfel, M., McDonough, J., 2009c. Distant Speech Recognition. Wiley. Chichester, UK.

Wu, W., Zheng, T. F., Xu, M. X., Soong, F. K., 2007. A cohort based speaker model synthesis for mismatched channels in speaker verification. IEEE Transactions on Audio, Speech, and Language Processing. 15(6), 1893-1903.

Yang, X., Millar, J. B., Macleod, I., 1996. On the sources of inter and intra speaker variability in the acoustic dynamics of speech. In Proceedings of IEEE International Conference on Acoustics and Speech Signal Processing. 1792-1795. Philadelphia.

Yates, G. K., Winter, I. M., Robertson, D., 1990. Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range. Hearing Research. 45(3), 203-219.

Yokoyama, R., Nasu, Y., Iwano, K., Shinoda, K., 2013. Detection of overlapped speech using lapel microphones in meeting. Speech Communication. 55(10), 941-949.

Young, E. D., Sachs, M. B., 1979. Representation of steady state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers. Journal of the Acoustical Society of America. 66(5), 1381-1403.

Young, E. D., 1997. Parallel processing in the nervous system: evidency from sensory maps. Proceedings of the National Academy of Sciences of the United States of America. 94, 933-934.

Young, E. D., 2008. Neural representation of spectral and temporal information in speech. Philosophical Transactions of the Royal Society of London. Serie B. 363, 923-945.

Zhao, Y., 1994. An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition. IEEE Transactions on Speech and Audio Processing. 2(3), 380-394.

Zhou, N., Pfingst, B. E., 2014. Effects of site-specific level adjustments on speech recognition with cochlear implants. Ear and Hearing. 35(1), 30-40.

Zilany M. S., Carney, L. H., 2010. Power-law dynamics in an auditory-nerve model can account for neural adaptation to sound-level statistics. Journal of Neuroscience. 30(31), 10380-90.

Zilovic, M. S., Ramachandran, R. P., Mammone, R. J., 1998. Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer function. IEEE Transactions on Speech and Audio Processing. 6(3), 260-267.

Zwicker, E., 1961. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). Journal of the Acoustical Society of America. 33, 248.

Anexos

Anexo A

Publicaciones del autor

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

Publicaciones en revistas ISI como primer autor

1. **Poblete, V.**, Becerra Yoma, N., Stern, R. M., (2014), "Optimization of the parameters characterizing sigmoidal rate-level functions based on acoustic features," Speech Communication (Elsevier), Volume 56, January 2014, pp. 19-34.

2. **Poblete, V.**, Espic, F., King, S., Stern, R. M., Huenupán, F., Becerra Yoma, N., (2014), "A perceptually-motivated low-complexity instantaneous channel normalization technique applied to speaker verification," Submitted to Computer Speech and Language (Elsevier), February 2014.

Publicaciones en congresos internacionales como primer autor

1. **Poblete, V.**, Becerra Yoma, N., Stern, R. M., (2013), "Optimization of sigmoidal rate-level function based on acoustic features," in Proceedings of Interspeech 2013, 896-900, August 2013, Lyon, France.

Publicaciones

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 56 (2014) 19–34

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Optimization of the parameters characterizing sigmoidal rate-level functions based on acoustic features

Víctor Poblete [a,c], Néstor Becerra Yoma [a,*], Richard M. Stern [b]

[a] Speech Processing and Transmission Laboratory, Universidad de Chile, Av. Tupper 2007, P.O. Box 412-3, Santiago, Chile
[b] Department of Electrical and Computer Engineering and Language Technologies Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA
[c] Institute of Acoustics, Universidad Austral de Chile, Av. General Lagos 2086, P.O. Box 5111187, Valdivia, Chile

Received 1 February 2013; received in revised form 14 June 2013; accepted 20 July 2013
Available online 29 July 2013

Abstract

This paper describes the development of an optimal sigmoidal rate-level function that is a component of many models of the peripheral auditory system. The optimization makes use of a set of criteria defined exclusively on the basis of physical attributes of the input sound that are inspired by physiological evidence. The criteria developed attempt to discriminate between a degraded speech signal and noise to preserve the maximum amount of information in the linear region of the sigmoidal curve, and to minimize the effects of distortion in the saturating regions. The performance of the proposed optimal sigmoidal function is validated by text-independent speaker-verification experiments with signals corrupted by additive noise at different SNRs. The experimental results suggest that the approach presented in combination with cepstral variance normalization can lead to relative reductions in equal error rate as great as 40% when compared with the use of baseline MFCC coefficients for some SNRs.

Keywords: Sigmoidal function; Auditory systems; Optimization; Acoustic features; Speech enhancement

1. Introduction

Speech sounds are pressure waves that vary as a function of time. These sounds are passed through the peripheral auditory system before being converted into electrical neural activity in the auditory nerve (Pickles, 2008). A spectral decomposition is performed in the cochlea that separates the incoming speech sounds into their constituent frequency components, and information is passed on through the auditory nerve, the brainstem, ultimately to the auditory cortex through channels that remain frequency dependent. In a natural environment the target speech sounds and background noise enter the peripheral auditory system together. Nevertheless, one of the most compelling characteristics of the auditory system is its ability to respond to and distinguish speech sounds from background noise (Darwin, 2008). In this paper we introduce a new way to improve accuracy in speaker verification tasks by incorporating a particular type of adaptation in a representation used for feature extraction that is based on processing in the auditory periphery.

1.1. Neural processing of speech signals

Neural processing of speech is represented by the temporal patterns of neural impulses (or "spikes") transmitted along the auditory-nerve fibers, which vary in time in response to the incoming sound. The dependence of the average number of spikes per second on incoming signal intensity in a particular frequency region is summarized by curves called rate-versus-level functions (e.g. Moore, 2003; Pickles, 2008). Rate-level functions display a variety of forms, although they are usually sigmoidal (e.g. Sachs

* Corresponding author. Tel.: +56 2 29784205.
E-mail address: nbecerra@ing.uchile.cl (N.B. Yoma).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.07.006

and Abbas, 1974; Yates et al., 1990). Under these circumstances, rate-level functions can be characterized by four attributes: (1) discharge threshold; (2) maximum discharge rate; (3) spontaneous discharge rate; and (4) dynamic range (Nizami, 2005). As described by Young (2008), dynamic range in this context refers to "the range of sound levels over which the fiber changes its rate when the input changes in level."

Most of the auditory-nerve fibers exhibit a dynamic range of less than 35 dB when stimulated with tones at their characteristic frequency (May and Sachs, 1992). In contrast, the dynamic range of loudness perception for humans is as great as 100 dB of sound pressure level (Winslow and Sachs, 1987). Through the years there have been a number of hypotheses concerning how humans can perceive loudness changes over such a wide dynamic range while the intrinsic dynamic range of auditory-nerve fibers is limited to 20–35 dB. These speculations have included consideration of the distributions of the thresholds of individual auditory-nerve fibers, the spread of excitation of the fibers over frequency, and possible loudness coding based on synchronous response, at least at low frequencies (<3–4 kHz) (Shamma, 1985).

In recent years, attention has also focused on the potential ability of the response at the auditory nerve to develop rate-level functions that vary according to the distribution of stimulus levels (Barbour, 2011; Dean et al., 2005, 2008). For example, experiments with cats have shown that the dynamic range in their auditory neurons is adapted for tone and noise stimuli to the distribution of sound levels. This adaptation is characterized by shifting towards the most frequently occurring level (Wen et al., 2009, 2012). Rate-level functions in the guinea pig also exhibit a restricted and mutable dynamic range. In these animals, neural responses are rapidly adjusted and tend to improve coding of the sound levels (Dean et al., 2005). Auditory-nerve fibers in the mouse also display similar behavior, although with differences in frequency ranges (Taberner and Liberman, 2005). We elaborate on these results and some of their potential consequences in Section 2.

For many years the properties of the auditory system have attracted the interest of researchers in speech processing, including the use of models of the auditory system as part of the feature extraction process for automatic speech recognition, speaker verification, etc. Some of this work has been reviewed in Stern and Morgan (2012a,b), and the earliest computational models of the peripheral auditory system that have been developed include the work of Allen (1985), Ghitza (1986, 1994), Seneff (1988), Lyon (1982), Shamma (1988), and Cohen (1989). Most of these models begin with a bank of filters tuned to different center frequencies to model the spectral decomposition of incoming sounds into the cochlea, followed by a model of transduction that includes the sigmoidal nonlinearity of the auditory transduction process that transforms the mechanical motion in the cochlea to the production of auditory-nerve spikes. As an example of the latter mechanism, the Seneff model includes a representation of the inner hair cells that consists of four stages: (1) a rate-level nonlinearity that limits the responses to signal components of a particular frequency with very small and very large amplitudes, (2) short-term adaptation that models the release of neurotransmitters during the synapse stage, (3) a low-pass filter that models the loss of synchrony in response to signal components of high frequency, and (4) an automatic gain control that maintains a presence of high-intensity sounds when the auditory nerve is saturated. Seneff (1988) proposed two parallel paths to analyze the outputs of this representation. One path measures the instantaneous overall short-time energy appearing each channel output, and in the other develops a spectral representation based on the extent to which the output signal is synchronized to best frequency of response of the fiber. Over time numerous groups have used auditory models such as the ones listed above to develop features for use in speech recognition and speaker identification, among other technologies (e.g. Kim et al., 2006; Kim and Stern, 2012).

1.2. Feature extraction for speaker verification

In speaker verification the aim has been to determine whether a given speech signal belongs to a claimed person or not based only on a voice sample (Reynolds, 1995). Usually, a speaker verification system comprises three sections: feature extraction, speaker modeling (performed from the extracted features), and decision making (Kinnunen and Li, 2010). The feature extraction section is designed to provide enough discriminative information from the speech signal to enable the speaker to be verified (Li and Huang, 2011). The development of relevant features is clearly important to discriminate one speaker from another in a fashion that preserves verification accuracy in environments that are different from the original training environment (Kinnunen et al., 2012; Li and Huang, 2011; Shao and Wang, 2008). Differences in the environment may arise from various sources, including additive interfering noise (Ming et al., 2007) and variations in the transmission channel conditions over which the speech is being recorded (Wu et al., 2007). Resolving mismatches between training and testing environments remains one of the most challenging problems to be solved for successful speaker verification in real applications (Hasan and Hansen, 2013; Saeidi et al., 2010).

The most commonly used features for speaker verification have been short-time cepstral coefficients such as Mel-frequency cepstral coefficients (MFCC) (Ajmera et al., 2011; Hanilçi et al., 2012; Wang et al., 2011). The standard MFCC method performs reasonably well when training and testing environments are matched but verification accuracy degrades seriously under noisy environments, especially when training and testing conditions are mismatched (Kinnunen et al., 2012; Li and Huang, 2011). The greatest degradation in verification performance is

observed when the speech signal is degraded by additive noise at a low SNR, especially when the system is trained on clean speech (Hanilçi et al., 2012; Kinnunen et al., 2012).

Feature extraction inspired by the physiology of the human peripheral auditory system has also been proposed to improve speaker verification performance under mismatched conditions (e.g. Li and Huang, 2010, 2011; Shao and Wang, 2008; Shao et al., 2007). For instance, Shao and Wang (2008), proposed auditory-based features known as Gammatone frequency cepstral coefficients (GFCC), which effectively replace the triangular frequency weighting used in the MFCC method by the use of Gammatone filters (Shao and Wang, 2008) to achieve frequency selectivity. Gammatone filters are widely used in models of the auditory system and were developed to mimic cochlear filtering (Patterson et al., 1992). Shao et al. (2007) have demonstrated that GFCC features can provide robust speaker recognition in the presence of additive noise over a wide range of SNRs, and performance can be further improved by adding complementary auditory scene analysis (Shao et al., 2010). Similarly, Li and Huang (2011) proposed the use of cochlear filter cepstral coefficients (CFCC) for robust speaker identification in mismatched conditions (Li and Huang, 2010, 2011). The proposed CFCC features are based on a time–frequency transform called the Auditory Transform that includes several components that mimic the processing in the human peripheral auditory system (Li and Huang, 2011). CFCC features also improve speaker identification accuracy compared to conventional MFCC processing when tested under mismatched conditions (Li and Huang, 2010). Other recent auditory-based features include the Teager energy cepstrum coefficients (TECC), developed by Dimitriadis et al. (2011).

1.3. The sigmoidal rate-intensity function

In an earlier study, Chiu and Stern (2008) examined the contributions of each stage of the classic auditory model by Seneff (1988) to analyze their impact in improving recognition accuracy for speech in the presence of noise and found that the best improvement in speech-recognition accuracy is provided by the rate-level nonlinearity stage that most models of the peripheral auditory system include just after the (typically linear) bandpass filtering that models the motion of the basilar membrane in the cochlea. This nonlinearity is roughly S-shaped, and has three major regions: (1) a range of input intensities that are "below threshold" in which the function output is roughly constant at a low level, (2) a range of input intensities for which the function output is roughly linear with respect to the input intensity in decibels, and (3) a "saturated" region in which the function output is roughly constant at a higher level.

Results from recent physiological studies describe and attempt to explain various types of dynamic adaptation of the rate-level functions with respect to the intensity of the incoming sound, background noise intensity, and the contrast between noise and the degraded speech signal (Dean et al., 2005; Zilany and Carney, 2010). These adaptations enable the dynamic range of the rate-level functions, which intrinsically is rather limited, to cover a much broader range of sound levels. In general, higher input sound levels tend to move the rate-level curves to the right and increase their maximum slope (Bureš et al., 2010; Gao et al., 2009). For example, in cats the background noise causes in the rate-level functions a shift of the dynamic range to higher intensities. It has also been noted that the noise level where this shift begins can be frequency dependent (Costalupes et al., 1984), and that the slope of the rate-level functions can increase in the presence of noise (May and Sachs, 1992) in addition to increased input levels.

Similar research in ferrets has characterized enhancement of spectro-temporal contrast in the acoustic environment as another important consequence of the adaptation of the sigmoidal nonlinearity (Rabinowitz et al., 2011; Wang and Shamma, 1994). This is similar to the enhancement in spatio-temporal contrast that is developed in the vertebrate retina (Ohzawa et al., 1985; Werblin et al., 1996). As an example, Rabinowitz et al. (2011), describe auditory processing that enhances local fluctuations in the envelope of response to desired signals in the presence of noise. This cannot be accomplished by a simple gain control which simultaneously amplifies both the degraded speech and the noise components, but rather a form of adjustable nonlinear gain control corresponds that increases the dynamic range of the degraded speech while suppressing the fluctuations produced by the noise (Schneider et al., 2011).

A steepening of rate-level functions of auditory neurons has also been observed in response to increment in sound level (Kang et al., 2010; Middlebrooks, 2004; Pfingst et al., 2011) and in response to noise (Bureš et al., 2010; Gao et al., 2009). According to Garcia-Lazaro et al. (2007) who investigated rat auditory neurons, the observed rate-level curves were more or less sigmoidal in shape, with a change in the steepness of the rate-level function interpreted to be a change in the "neural response gain". Reports on auditory neurons of marmoset monkeys have shown that the slope in the rate-level function is a measure of the sound level discriminability. A steeper slope would allow greater discrimination of sound level (Watkins and Barbour, 2011). Other studies on auditory neurons in response to continuous, dynamic sound stimuli have shown a horizontal displacement of the rate-level function relocating the dynamic region of the function toward the mean sound level, resulting in higher coding precision of the levels (Dean et al., 2005; Wen et al., 2009), (Miller et al., 2011; Schneider et al., 2011). By expanding or compressing the auditory response to incoming sound in varying degrees, contrast gain control in human audition can serve two functions: (1) to protect sensory systems from overload and (2) to enhance discriminability among selected stimuli (Schneider et al., 2011). Ideally, the rate-level function

would increase its slope to correspondingly enhance contrast in its response to small amplitude of the degraded speech signal above the noise (low contrast between degraded speech and noise signals). These goals, in combination with reducing the nonlinear distortion of the degraded speech and reducing the differences between the original clean speech and the degraded speech described in Chiu et al. (2012), lead to the definition of the four optimization criteria described in Section 2.2.

With these physiological examples in mind, Chiu et al. (2012) postulated that dynamic adaptation of the rate-level nonlinearity could also improve speech recognition accuracy for speech in the presence of noise. In particular, they modeled the rate-level nonlinearity by a set of frequency-dependent logistic functions, and developed a procedure that optimized the parameters that specified the form of the sigmoidal nonlinearity for a particular additive-noise environment using an objective function based on maximizing phonemic discriminability. These authors demonstrated that the use of an adapted nonlinear rate-level function reduces differences between the shapes of spectral distributions of clean speech versus speech in noise, and they showed that the adaptation of the nonlinearity improves speech recognition accuracy in the presence of noise.

In this paper, we describe a new approach for optimizing the sigmoidal rate-level function that is based on physical attributes of the acoustical signal, rather than the phoneme discrimination that was the basis for the approach of Chiu et al. (2012). The method attempts to discriminate between the degraded speech signal and noise, preserve maximum information in the linear region of the sigmoidal curve, and minimize the effects of distortions in the saturation regions. The proposed method is applied to a text-independent speaker verification task with speech signals that are corrupted by additive noise at different SNRs.

The development of adaptation based on signal analysis (rather than phonetic analysis) is motivated by several considerations. First, and foremost, the discriminative training used by Chiu et al. (2012) is based on speech recognition at the phoneme level in order to develop the ground-truth phoneme representation of the data. We believe that parameter optimization based on speech recognition may not be best for speech tasks other than speech recognition, such as the speaker verification considered in the present paper. We also had described above the existence of adaptation of sigmoidal nonlinearities in several nonhuman species and in other sensory modalities that is similar to those modeled in the present paper. This suggests that we should search for a viable approach to adaptation of the nonlinearity that is based on something other than human phonetic discrimination.

From the computational standpoint, the discriminative training described in Chiu et al. (2012) requires a substantial amount of *a priori* information, and increases the computation needed substantially compared to the signal-processing-based approach described in this paper. Similarly, the signal-processing-based approach is much more amenable to an online or adaptive implementation than the approach of Chiu et al., in which the discriminative training must be performed offline based on training data.

An unrelated reason for revisiting the issue of adapting the sigmoidal nonlinearity is that Chiu et al. perform their computations for all the channels of frequency analysis using the same parameters. We examine in this paper the extent to which performance would be further improved by adaptation that is allowed to vary on a channel-by-channel basis, which seems reasonable as the effective SNR varies from channel to channel.

We reiterate that in principle the approach proposed in this paper is applicable to *any* speech processing task because all analysis takes place at the level of the acoustic signal. Also, the sigmoidal functions are estimated separated for each channel.

In Section 2 we describe our optimization approach and specifically the development of four criteria that optimize the sigmoidal rate-level function based on acoustic attributes of the incoming signals. In Section 3 we discuss the actual implementation of the sigmoidal rate-level nonlinearity, and in Section 4 we describe the experimental results that validate the utility of the approach.

2. Development of the optimization criteria for the sigmoidal function

In this section we describe the development of the optimization criteria for the sigmoidal rate-level function. We begin with a mathematical specification of the sigmoidal nonlinearity, and subsequently we provide a mathematical description of the four components of the objective function used to optimize the rate-level nonlinearity. We remind the reader that the goal of the adaptation is to modify the location and the slope of the sigmoidal nonlinearity so that it is best able to capture the intensity fluctuations of the speech components of the signal in each channel, and to enhance the contrast when the input speech incurs a high level of degradation from additive noise.

2.1. Mathematical specification of the sigmoidal function

Let us represent the rate-level nonlinearity in auditory transduction by the sigmoidal function $g(l)$ given by:

$$g(l) = \frac{1}{1 + e^{\omega(l-\mu)}} \tag{1}$$

where μ and ω correspond to the offset and the slope of $g(l)$, respectively. This function allows modeling the nonlinear response. The offset parameter μ corresponds to the location along the horizontal axis at which the sigmoidal curve $g(l)$ equals 1/2. The slope of $g(l)$ equals $-\omega/4$ when $l = \mu$. Consequently, the position μ and the slope ω of the sigmoidal function are the parameters to be estimated.

Let us now consider the output of a particular channel of the initial bandpass filter bank that is the first stage of every model. We represent the degraded input speech signal $x_{j,k}$ at the output of filter j at the discrete-time index k by:

$$x_{j,k} = s_{j,k} + n_{j,k} \tag{2}$$

where $s_{j,k}$ and $n_{j,k}$ denote the clean speech and noise signals, respectively. If the entire signal $x_{j,k}$ is divided into N_f frames of W samples per frame with 50% overlap, the log-energy at frame i at filter j, $E_{j,i}$, can be written as:

$$E_{j,i} = 10 \cdot \log\left(\sum_{k\in \text{ frame } i} w_{i-k}^2 x_{j,k}^2\right) \tag{3}$$

where w_k represents the response of the finite-duration window function. Histograms of the log-energies $E_{j,i}$ at filter j and at frame i are generated in order to discriminate between noise and degraded speech frames by using the voice activity detector (VAD) proposed by Shin et al. (2008), as discussed in Section 3. Hence the frames are divided into two subsets, one believed to contain degraded speech and the second subset representing frames that are assumed to contain only noise.We use the symbols N_f^{sn} and N_f^n, where $N_f = N_f^{sn} + N_f^n$, to indicate the number of frames that are assumed to contain speech degraded by noise and the number of frames that are assumed to contain noise alone, respectively. Finally, we use the symbols $E_{j,i}^x$ $(1 \leqslant i \leqslant N_f)$, $E_{j,m}^{sn}$ $\left(1 \leqslant m \leqslant N_f^{sn}\right)$ and $E_{j,r}^n$ $\left(1 \leqslant r \leqslant N_f^n\right)$ to represent the energies at filter j and at frames i, m and r for frames that are considered to belong to the original input, the subset of frames that contain degraded speech and the subset of input frames that contain noise alone, respectively. (Recall that each frame of the input is classified as containing either degraded speech or pure noise.) In addition, the mean and variance of the energy in the degraded-speech frames are defined to be $\mu_{j,sn}$ and $\sigma_{j,sn}^2$, respectively, while the corresponding mean and variances of the energy in the frames that are assumed to contain only noise energy frames are $\mu_{j,n}$ and $\sigma_{j,n}^2$, respectively.

2.2. Specification of the objective function used to optimize the sigmoidal nonlinearity

Based on the discussion above, we choose an objective function for the sigmoidal nonlinearity that (1) minimizes nonlinear distortion in the linear region, (2) minimizes noise power, (3) maximizes the similarity between energy in the frames that are believed to represent degraded speech and the energy of the speech alone in those frames, and (4) maximizes the energy in the output signal which is presumed to be dominated by speech.

2.2.1. Criterion 1: Nonlinear distortion in the linear region

The slope ω and the position μ of the sigmoidal function should be chosen in such a way that the degraded speech lies in the linear part of the sigmoidal curve. Therefore, once the sigmoidal function is applied, the nonlinear distortion in the degraded speech would be minimized. This nonlinear distortion, $D_j^{non-linear}(\omega_j, \mu_j)$, is defined as:

$$D_j^{non-linear}(\omega_j, \mu_j) = \frac{\mathbf{E}\left\{\left[A_j E_{j,m}^{sn} + B_j - g\left(E_{j,m}^{sn}\right)\right]^2\right\}}{\mathbf{E}\left[\left(E_{j,m}^{sn}\right)^2\right]} \tag{4}$$

where (as before) $E_{j,m}^{sn}$ refers to the energy of frames of degraded speech at frame index m for filter index j, $g(\cdot)$ represents the sigmoidal function and $\mathbf{E}[\cdot]$ is the expectation operator. The parameters A_j and B_j correspond to a linear transformation that allows the comparison of $E_{j,m}^{sn}$ and $g\left(E_{j,m}^{sn}\right)$ (as developed in the Appendix). By approximating the expected value by the sample mean, $D_j^{non-linear}(\omega_j, \mu_j)$ can be rewritten as:

$$D_j^{non-linear}(\omega_j, \mu_j) = \frac{\frac{1}{N_f^{sn}}\sum_{m=1}^{N_f^{sn}}\left[A_j E_{j,m}^{sn} + B_j - g\left(E_{j,m}^{sn}\right)\right]^2}{\frac{1}{N_f^{sn}}\sum_{m=1}^{N_f^{sn}}\left(E_{j,m}^{sn}\right)^2} \tag{5}$$

where N_f^{sn} is the total number of frames containing degraded speech.

2.2.2. Criterion 2: Noise power

The sigmoidal function can be employed to attenuate the noise in the speech signal due to the fact that low-energy frames can be associated with noise. The power of the noise, after it is passed through the sigmoidal function, $P_j^{noise}(\omega_j, \mu_j)$, is given by:

$$P_j^{noise}(\omega_j, \mu_j) = \mathbf{E}\left[g^2\left(E_{j,r}^n\right)\right] \tag{6}$$

where $E_{j,r}^n$ corresponds to energy of noise frames at frame r for filter j, $g(\cdot)$ represents the sigmoidal function, and $\mathbf{E}[\cdot]$ is the expectation operator. The sigmoidal function should minimize $P_j^{noise}(\omega_j, \mu_j)$ in order to reduce the effect of noise energy. By estimating the expected value as the sample mean, $P_j^{noise}(\omega_j, \mu_j)$ can be rewritten as:

$$P_j^{noise}(\omega_j, \mu_j) = \frac{1}{N_f^n}\sum_{r=1}^{N_f^n} g^2\left(E_{j,r}^n\right) \tag{7}$$

where N_f^n is the number frames that are assumed to contain noise only.

2.2.3. Criterion 3: Similarity between clean speech and the degraded speech input

According to Chiu et al. (2012), the use of a nonlinear rate-level function should reduce the differences between the average frequency response of clean speech and the average frequency response of the degraded input signal, both assessed after the sigmoidal nonlinearity. Consequently, the difference between the energy of the clean speech and the degraded speech input is represented by:

$$D_j^{clean-noise}(\omega_j,\mu_j)=\sum_{i=1}^{N_f}\left[g\left(E_{j,i}^{s}\right)-g\left(E_{j,i}^{x}\right)\right]^2 \tag{8}$$

where $E_{j,i}^{s}$ and $E_{j,i}^{x}$ correspond to the energy of clean speech and the energy of the degraded input speech, respectively, at frame i for filter j, and $g(\cdot)$ is the sigmoidal function.

2.2.4. Criterion 4: Signal variance of speech degraded by noise after processing by sigmoidal function

To avoid extreme compression or saturation, the variance of the resulting degraded speech after the sigmoidal function should be maximized. This variance of the degraded speech, $V_j(\omega_j,\mu_j)$, is expressed as:

$$V_j(\omega_j,\mu_j)=\sigma^2\left[g\left(E_{j,m}^{sn}\right)\right] \tag{9}$$

where $E_{j,m}^{sn}$ is energy of the frames of degraded speech at frame m at filter j and $g(\cdot)$ is the sigmoidal function. By expanding the expression of the variance, $V_j(\omega_j,\mu_j)$ can be rewritten as:

$$V_j(\omega_j,\mu_j)=\frac{1}{N_f^{sn}}\sum_{m=1}^{N_f^{sn}}g^2\left(E_{j,m}^{sn}\right)-\left[\frac{1}{N_f^{sn}}\sum_{m=1}^{N_f^{sn}}g\left(E_{j,m}^{sn}\right)\right]^2 \tag{10}$$

where N_f^{sn} is the number of frames containing degraded speech frames.

2.2.5. Specification of the complete objective function

Based on the four criteria described above, we adopt for this study the objective function $J(\omega_j,\mu_j)$ that is defined as:

$$J(\omega_j,\mu_j)=D_j^{non-linear}(\omega_j,\mu_j)+P_j^{noise}(\omega_j,\mu_j)+D_j^{clean-noise}(\omega_j,\mu_j)-V_j(\omega_j,\mu_j) \tag{11}$$

Consequently, the optimal slope, $\hat{\omega}_j$, of the sigmoidal function is estimated as:

$$\hat{\omega}_j=\arg\min_{\omega_j}\{J(\omega_j,\mu_j)\} \tag{12}$$

In (12), the position $\mu_j$of the sigmoidal function is set to $\mu_j=\mathbf{E}\left[\left(E_{j,m}^{sn}\right)\right]$ (*i.e.* centered on the mean of the energy of the degraded speech frames $E_{j,m}^{sn}$).

Finally, the optimal position, $\hat{\mu}_j$, of the sigmoidal function is estimated according to:

$$\hat{\mu}_j=\arg\min_{\mu_j}\{J(\omega_j,\mu_j)\} \tag{13}$$

In (13), ω_j corresponds to the optimal sigmoidal slope $\hat{\omega}_j$.

While we recognize that the definition of the objective function $J(\omega_j,\mu_j)$ as the simple sum of the four criteria above is a special case of the more general linear combination

$$J(\omega_j,\mu_j)=a\cdot D_j^{non-linear}(\omega_j,\mu_j)+b\cdot P_j^{noise}(\omega_j,\mu_j)+c\cdot D_j^{clean-noise}(\omega_j,\mu_j)-d\cdot V_j(\omega_j,\mu_j)$$

we adopted the function of (11) for simplicity in the absence of compelling evidence that other combinations of the four criteria would provide better performance.

3. Implementation of the sigmoidal rate-level function

In this section we describe the adaptive procedure based on signal analysis that is used to optimize the sigmoidal rate-level function. We refer the reader to Fig. 1 for a depiction of the complete feature extraction scheme, and Fig. 2 for a depiction of the procedure for obtaining the optimal parameters $\hat{\omega}_j$ and $\hat{\mu}_j$. The specific values of the sigmoidal parameters $\hat{\omega}_j$ and $\hat{\mu}_j$, as defined in Eqs. (12) and (13), respectively, were determined using a development database of speech corrupted by babble noise at an SNR equal to 10 dB, as discussed in Section 4.

The optimal values of the parameters $\hat{\omega}_j$ and $\hat{\mu}_j$ used in our work vary from channel to channel, in contrast to the approach of Chiu et al. (2012), in which the parameters of the nonlinearity are the same for all the filters. This is helpful because the SNR varies from one filter to the other. As mentioned above, we used the voice activity detector (VAD) proposed by Shin et al. (2008) to discriminate between degraded speech and noise. Two subsets of frames are defined based on the VAD results, representing frames that contain degraded speech, and the representing frames that are assumed to contain only noise, respectively.

Fig. 3 describes the dependence of the shape of the objective function on the parameters ω_j and μ_j, which describe the slope and position, respectively, for each the 35 analysis bands j. Fig. 4 depicts a representative example of an optimal sigmoidal function (solid line) and a corresponding linear mapping (dotted line) with a slope equal to the sigmoid at its center point. The sigmoidal curve

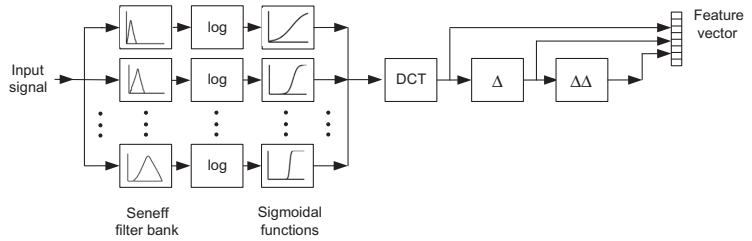

Fig. 1. Block diagram of the proposed feature extraction scheme.

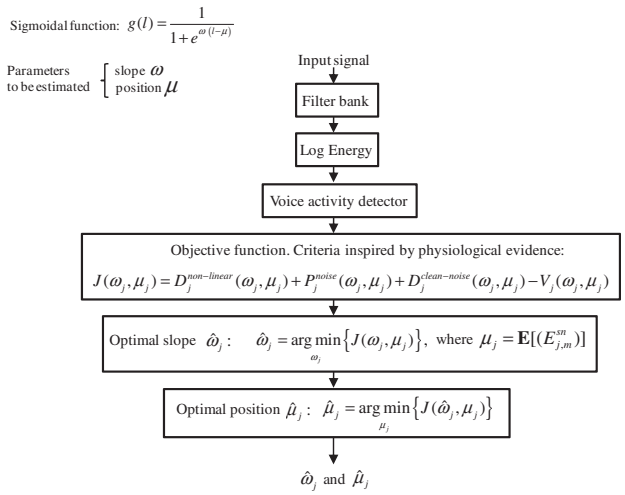

Fig. 2. Block diagram for obtaining optimal parameters $\hat{\omega}_j$ and $\hat{\mu}_j$ of the sigmoidal function.

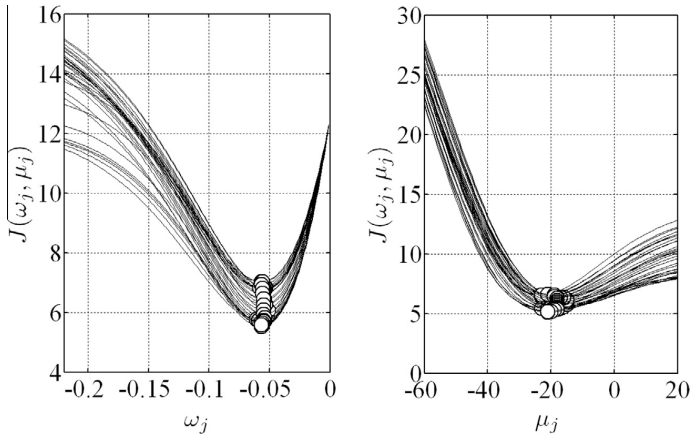

Fig. 3. The objective function $J(\omega_j,\mu_j)$ plotted as a function of the sigmoidal function parameters: (a) sigmoidal slope ω_j and (b) sigmoidal position μ_j. The optimal values of $\hat{\omega}_j$ and $\hat{\mu}_j$ are indicated by the open circles for each of the 35 channels of the filter bank.

was obtained with the optimal slope $\hat{\omega}_j$ and position $\hat{\mu}_j$ of the sigmoidal function for the filter $j=8$. Fig. 4 also depicts the histograms extracted from a testing utterance for filter $j=8$ with babble noise at SNR equal to 10 dB. By comparing the solid and dotted lines in Fig. 4 it can be seen that the sigmoidal function compresses the noise in the nonlinearity region, while most of the frames containing degraded speech lie within the linear part of the sigmoidal function.

Fig. 5 shows four sigmoidal functions trained with babble noise at SNRs equal to 20 dB, 15 dB, 10 dB and 5 dB, along with a fifth sigmoidal function that was trained with clean speech. As shown in Fig. 5, both optimal slope $\hat{\omega}_j$ and position $\hat{\mu}_j$ of the sigmoidal function depend on the SNR at which the sigmoidal function was trained: as SNR is increased, the curves in Fig. 5 shift to the right and become steeper. Consequently, the optimization of sigmoidal slope $\hat{\omega}_j$ and the sigmoidal position $\hat{\mu}_j$ provides an adaptation in the sigmoidal function that compensates for variations in SNR.

Fig. 6 is a composite of all of the sigmoidal rate-level functions, plotted as a function of SNR with optimal parameters for all 35 channels, trained on speech degraded with babble noise. We observe that the sigmoidal functions adapt slightly for each channel at each SNR. Specifically, as the SNR decreases, the curves are displaced toward the right, and their slopes become steeper at the midpoints. Collectively these phenomena modify the nonlinearities to ensure that most of the speech energy falls on the relatively

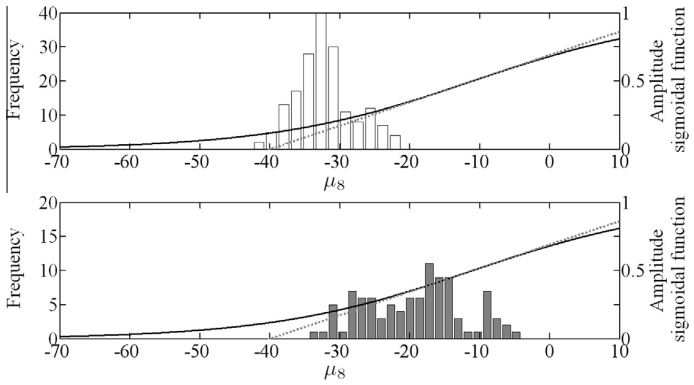

Fig. 4. Example of optimal sigmoidal function (solid line) and the corresponding linear mapping (dotted line). The training conditions for the sigmoid were babble noise at SNR = 10 dB. Results for filter $j = 8$ are plotted with optimal parameters: $\dot{\omega}_8 = -0.071$ and $\dot{\mu}_8 = -14$. In addition, histograms of power are depicted for frames containing degraded speech (filled bars) and frames assumed to contain noise only (open bars).

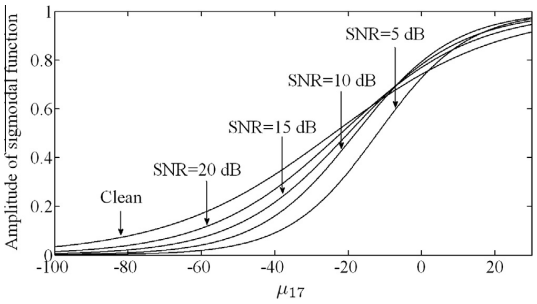

Fig. 5. Sigmoidal functions plotted as a function of SNR. Training conditions: clean speech and speech degraded by babble noise at SNRs equal to 20 dB, 15 dB, 10 dB and 5 dB. Results for filter $j = 17$ are plotted with optimal parameters.

linear part of the curve, which renders the features more robust against changes in SNR when training and testing conditions are mismatched.

Fig. 7 compares an ensemble of sigmoidal rate-level functions that were trained with different types of noise (restaurant and car noise) but at the same SNR. Results are similar to the curves of Fig. 6 in that the curves shift to the right and their slopes increase as SNR decreases, with variations in the individual responses observed from filter to filter. The interlaced patterns generated by the sigmoidal functions trained on the two types of noise indicate that the shape of the optimal non-linearity depends on the spectral distribution of the masking noise.

This dependence of the location and steepness of the sigmoidal curves in Fig. 5, Fig. 6 and Fig. 7, is consistent with the experimental results in the physiological literature described above. As noted, the steepening of rate-level functions of auditory neurons is consistent with the results of numerous physiological studies describing nonlinearity in sensory transduction (*e.g.* Bureš et al., 2010; Gao et al., 2009; Garcia-Lazaro et al., 2007; Kang et al., 2010; Middlebrooks, 2004; Pfingst et al., 2011; Watkins and Barbour, 2011).

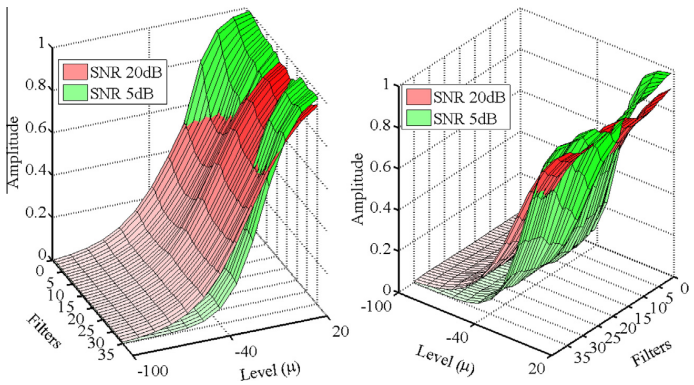

Fig. 6. Three-dimensional graphs of the sigmoidal rate-level functions trained with speech degraded by babble noise at SNR equal to 20 dB and 5 dB. The plot is rotated to show the difference in slope and horizontal displacement between both set of functions.

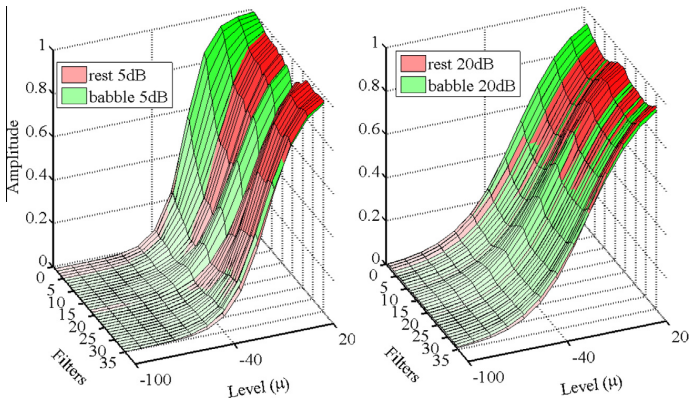

Fig. 7. Comparison of optimal sigmoidal rate-level functions trained with restaurant and car noise at SNR equal to 20 dB (right) and 5 dB (left).

We note that in this paper we describe separately a stage of logarithmic compression followed by the sigmoidal nonlinearity. We consider the logarithmic compression to be an approximation of the compressive gain of the basilar membrane (Robles and Ruggero, 2011). The subsequent sigmoidal rate-level function is intended to represent an approximation to both the existence of spontaneous AN activity at low input sound levels and the saturation of AN fibers at high sound levels, presumably because of a mixture of suppression (such as two-tone suppression (Rhode and Cooper, 1993), and a presumed limit on the discharge rate with which a fiber is capable of coding intensity.

4. Experimental results

The utility of the optimal sigmoidal nonlinearity was evaluated using a text-independent speaker verification task, with equal error rate (EER) employed as the major figure of merit. The results presented here were obtained using the YOHO database (Campbell and Higgins, 1994), which supports the development, training, and testing of speaker verification systems. The vocabulary is composed of two-digit numbers spoken continuously in sets of three (*e.g.*, "62-31-53" is pronounced as "sixty-two thirty-one fifty-three"). The database is divided into "enrollment" and "verification" segments. Each segment contains data from 138 speakers. In this paper a subset of 70 speakers (43 males and 27 females) was employed. These speakers were divided as follows: 40 background impostor speakers (28 males and 12 females) to train the background models; and 30 testing speakers (15 males and 15 females) were used in verification attempts. For each speaker, one 96-utterance enrollment session was considered. False rejection curves were estimated with 30 speakers × 40 verification signals per client = 1200 utterances. False acceptance curves were obtained with 30 speakers × 29 impostors × 12 verification signals/per impostor = 10,440 experiments. In addition, a subset composed of 50 speakers and one utterance per speaker (development database) extracted from YOHO was employed to train the optimal parameters $\hat{\omega}_j$ and $\hat{\mu}_j$ of the sigmoidal functions. The utterances used to train the sigmoidal function were not included in the testing data for the main speaker verification experiment. Three types of noise (babble, car, and restaurant) were selected from the AURORA database (Hirsch and Pearce, 2000). These noises were artificially added to the YOHO corpus to generate noisy versions of the utterances at various SNRs: 20 dB, 15 dB, 10 dB, 5 dB and 0 dB. For all the speaker verification experiments, the system was trained with clean speech.

In this paper, the auditory filter bank (Stage I in the Seneff auditory model (Seneff, 1988) was obtained directly from Malcolm Slaney's widely-used Auditory Toolbox (Slaney, 1998), which implements 35 filters with center frequencies spaced according to the Bark scale from 200 to 3300 Hz. Each filter was redesigned by reducing the sampling frequency to 8 kHz. Finally, the input signal was normalized by dividing the samples by the maximum absolute amplitude. After filtering, the signals were divided into 25-ms frames with 12.5-ms overlap between frames using Hamming windows. The log energy was computed at the output of each filter. Then, in each frame, a channel-specific optimal sigmoidal function, estimated using the development data set and the procedure explained in Section 2, was applied to the log-energy of the output of each filter, both in the training and testing data sets. Finally, the log-energy plus ten static cepstral coefficients, and their first and second cepstral time derivatives were estimated in a fashion that is similar to MFCC processing (see Fig. 1). Four configurations were considered: (1) a baseline system, which corresponds to the log-energies of the Seneff filter bank output; (2) the baseline system with cepstral variance normalization (CVN); (3) the baseline system with cepstral mean and variance normalization (CMVN), and (4) the method proposed in this paper using the optimal sigmoidal function, combined with CVN. If the entire signal were mapped into the linear region of the sigmoidal

function, the proposed scheme could be considered equivalent to the CVN algorithm. Therefore, the impact of the nonlinearity provided by the sigmoidal function may be inferred by comparing results obtained with the configurations (2) and (4) as described above.

In the verification procedure, the normalized log likelihood is estimated. Given a verification attempt in which the identity of Speaker s is claimed, O denotes the observation sequence corresponding to the claimant's utterance. The output score of the system is a cohort-normalized log likelihood, log $L(O)$:

$$logL(O) = logL(O/\lambda_s) - \overline{logL(O\lambda_{\bar{s}})} \quad (14)$$

where log $L(O/\lambda_s)$ is the log likelihood of the client hypothesis and λ_s is the speaker s model, and $\overline{logL(O/\lambda_{\bar{s}})}$ is the averaged log likelihood of the cohort of impostor models. A universal background model (UBM) is trained by using the background impostor speakers. A speaker-dependent Gaussian mixture model GMM is generated for each speaker by employing MAP adaptation (Reynolds et al., 2000). By doing so, the correspondence of the Gaussians within each speaker-dependent GMM with those in the background GMM is preserved (Reynolds et al., 2000).

4.1. General dependence on SNR and the presence of the sigmoidal nonlinearity

Fig. 8 describes results provided using the optimal sigmoidal functions for the speaker verification task in the presence of three types of background noise: speech babble, car noise, and restaurant noise, all as a joint function of the SNR at which the sigmoidal functions were trained and the SNR of the incoming speech. The optimal sigmoidal functions were trained with babble noise and SNRs equal to 20 dB, 15 dB, 10 dB, 5 dB and 0 dB. As can be seen in Fig. 8, the lowest EERs are achieved when the sigmoidal function is trained with 10 dB for all testing conditions. We note that these results are consistent with similar findings observed by Chiu et al. (2012) where the optimal sigmoidal function was trained at an SNR of 10 dB by using a criterion based on phoneme discrimination. In contrast, the objective function $J(\omega_j, \mu_j)$ is based entirely on the physical characteristics (and especially the power distribution) of the incoming speech, and does not take phonetic content into account at all. Consequently, the fact that the best speaker verification results are achieved with the sigmoidal function trained with signals at SNR 10 dB means only that for these SNRs the benefits provided by noise suppression are more significant than the distortion introduced by saturation at higher levels. Most of the results we described below are carried out using sigmoidal functions trained at an SNR equal to 10 dB.

We also note that the sigmoidal function trained at 0-dB SNR provides poor performance in speaker verification of speech at all testing SNRs. This is most likely a consequence of the fact that at 0-dB SNR the speech and noise distributions are nominally overlapping. Hence, the speech-plus-noise and noise-alone distributions are not separable by SNR, and any nonlinear compression applied to the noise will be applied to the speech as well. The estimation of the optimal parameters $\hat{\omega}_j$ and $\hat{\mu}_j$ for the sigmoidal distribution is less reliable as well.

Fig. 9 describes EER results obtained as a function of SNR for speech in the presence of three types of background interference: speech babble, car noise, and restaurant noise. Results are compared for the baseline system, the baseline system combined with CVN, the baseline

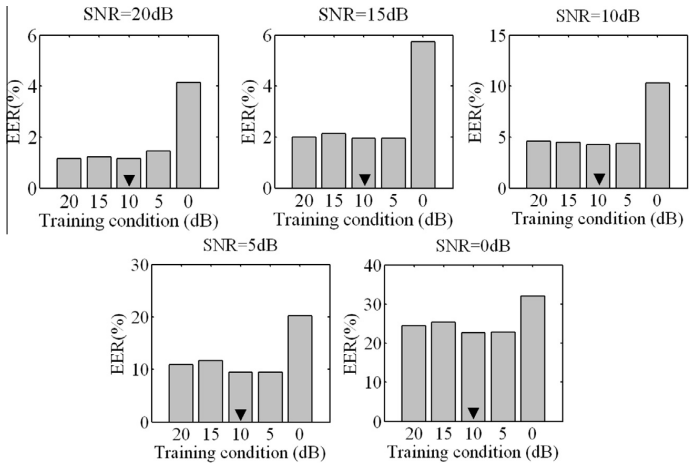

Fig. 8. EER for speaker verification as a function of the SNR of the testing data and the SNR used to develop the parameters for the sigmoidal function. The data obtained from each SNR used for testing are described in a single graphic, with the testing SNR indicated at the top. The optimal sigmoidal functions were trained with babble noise and SNRs equal to 20 dB, 15 dB, 10 dB, 5 dB and 0 dB, as indicated by the scale at the bottom of each panel.

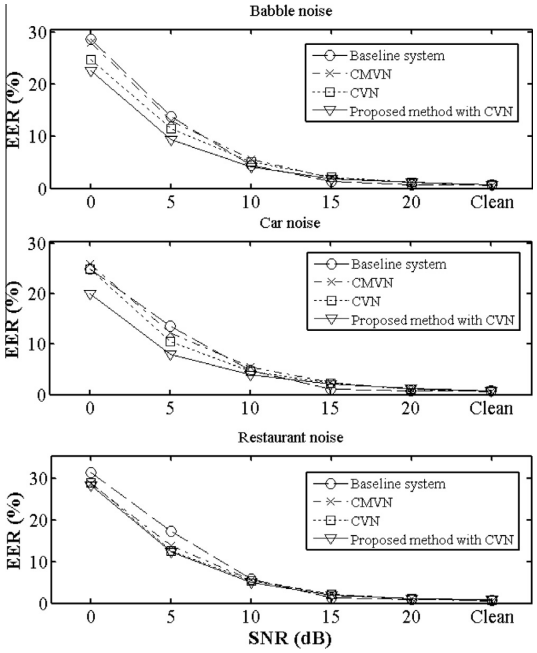

Fig. 9. Comparison of EER as a function of SNR for speech in babble, car and restaurant noise, respectively. Depicted separately are results for the baseline system, the baseline system with CMVN, the baseline system with CVN; and the system using the optimal sigmoidal function combined with CVN. The optimal sigmoidal functions were trained and tested in matched noisy condition at an SNR of 10 dB.

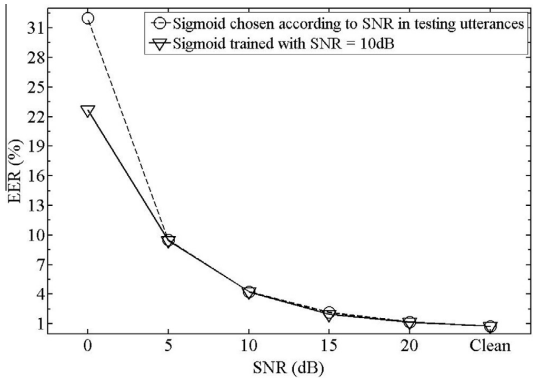

Fig. 10. Comparison of EER as a function of SNR for speech in babble noise using the sigmoidal function combined with CVN. In one curve the sigmoid trained with SNR equal to 10 dB was applied to both training and testing utterances. In the second curve, training and testing utterances were processed with the sigmoid trained at the same SNR used in the utterance.

system with cepstral mean and variance normalization (CMVN) and the proposed method combining the proposed optimal sigmoidal nonlinearity and CVN, as

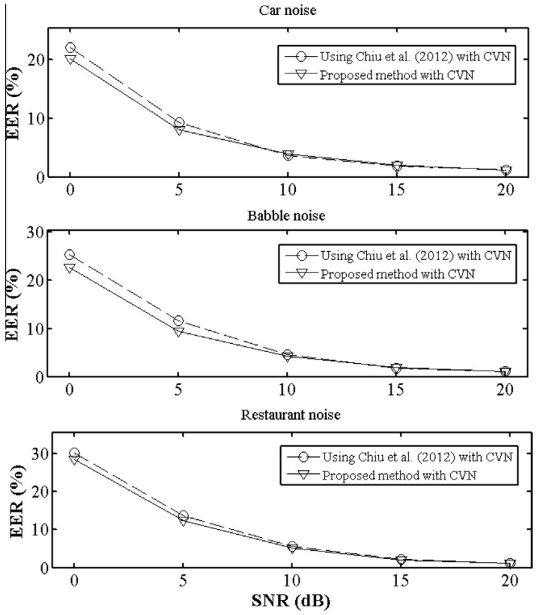

Fig. 11. Comparison of EER (%) obtained using the method with parameters of the nonlinearity given by Chiu et al. (2012), combined with CVN, and the proposed method in this paper combined with CVN, for speech in car, babble and restaurant noise, respectively.

described above. The use of the optimal sigmoidal function in combination with CVN improves the effective SNR of the system in all three types of noise, typically on the order of 1–2 dB. Maximum relative percentage improvements in SNR compared to baseline at selected SNRs are approximately 31.7%, 40.6%, and 28.4% for the three types of background noise. Best performance is always obtained using the proposed optimal sigmoidal nonlinearity, but at least for car noise, the performance of a system with CVN only comes close.

4.2. Comparison of training the nonlinearity parameters at fixed versus matched SNRs

As noted above, the results depicted in Fig. 9 were obtained by estimating the parameters characterizing the sigmoidal functions using speech in the presence of speech babble at an SNR of +10 dB, which appears to the best single training SNR according to the results depicted in Fig. 8. Nevertheless, we would always expect better performance to be obtained when the parameters characterizing the sigmoidal functions are estimated in environmental conditions that match the testing environment. In an effort to quantify the magnitude of the improvement to be expected, we repeated some of the conditions with the SNRs for parameter estimation matched to the SNRs used in the speaker verification experiment itself. Fig. 10 compares EERs for speaker

verification when the sigmoids are trained at the testing SNR with the corresponding EERs obtained when the parameters are always estimated from signals at 10-dB SNR. As can be seen from the figure, very little difference in results is observed, except for the lowest SNR, 0 dB, which actually provides very poor performance when the parameters are estimated at a matched SNR (presumably because the data are too noisy to provide reliable parameter estimation). For this reason we continue to use sigmoids trained at an SNR of 10 dB for all utterances, because they provide similar performance to sigmoids trained to match the testing data at most SNRs, and substantially better performance at 0-dB SNR.

4.3. Comparison to the results of Chiu et al.

As noted above, Chiu et al. (2012) described a method of adapting the sigmoidal rate-level function using a criterion based on discrimination analysis at phonetic level. Fig. 11 compares results obtained using the method described in this paper with results obtained using the method described by Chiu et al., with CVN included in obtaining both sets of results. The parameters obtained for the sigmoidal functions of Chiu et al. (2012) were $\alpha = 0.05$; $\omega_0 = 0.613$; and $\omega_1 = 0.521$. Results are presented for three types of noise:

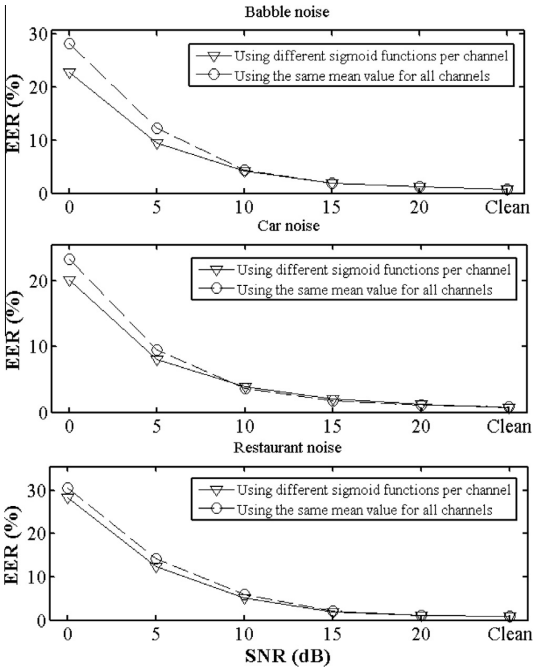

Fig. 12. Comparison of EER (%) obtained using the proposed method in this paper combined with CVN, using different sigmoidal functions per channel and the same mean value for all channels, for speech in car, babble and restaurant noise, respectively.

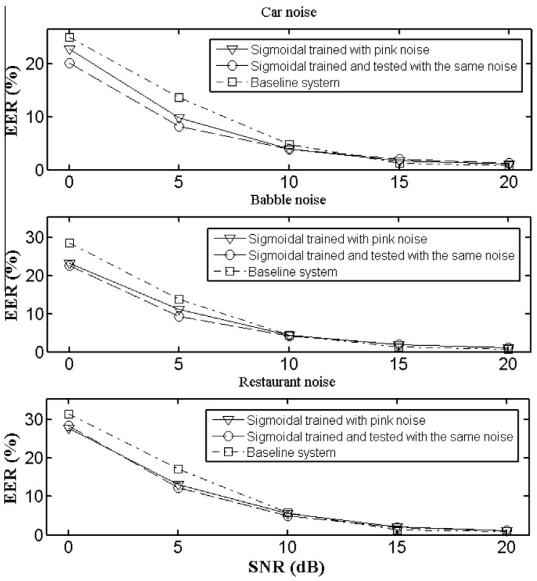

Fig. 13. Comparison of EER (%) obtained using the proposed method with parameters of the nonlinearity trained (at SNR equal to 10 dB) and tested with the same type of noise, and the proposed optimization method with the parameters of the sigmoidal function trained with pink noise (at SNR equal to 10 dB). In both cases the speech was degraded with car, babble and restaurant noise, and the sigmoidal function was combined with CVN.

babble, car, and restaurant at SNRs equal to 20 dB, 15 dB, 10 dB, 5 dB and 0 dB, respectively. The experimental results shown in Fig. 11 indicate that both the method proposed in this paper and the method proposed by Chiu et al. are effective in maintaining good performance at most SNRs, but the method proposed in the present paper performs somewhat better for all three noise types at the lower SNRs.

4.4. Impact of channel-specific estimation of sigmoidal nonlinearities

Fig. 12 compares results obtained using the sigmoidal nonlinearities estimated on a channel-specific basis as described in this paper with results obtained using a single nonlinearity for all 35 frequency channels. It can be seen that the allowing the sigmoidal nonlinearities to vary from channel to channel is advantageous at SNRs of 0 and +5 dB, most likely because the local SNRs exhibit greater variation from channel to channel at the lower SNRs. Thus, at lower SNRs, the performance of the optimization improves speaker verification accuracy, due to the fact that the adaptation enables to increase the dynamic range of the degraded speech above the noise and minimizes nonlinear distortions in the linear region while suppressing fluctuations produced by noise.

4.5. Comparisons to optimal sigmoidal functions trained and tested with different type of noise

Fig. 13 compares results obtained by using the sigmoidal nonlinearities estimated with pink noise at SNR equal to 10 dB with results shown in Fig. 9 where the same kind of noise was employed in training and testing. When compared to baseline processing, the sigmoidal functions trained with pink noise in combination with CVN leads to average relative reductions in EER equal to 23.5% and 13% at SNR equal to 5 dB and 0 dB, respectively, with car, babble and restaurant noise. This result strongly validates the proposed optimization method. However, the highest reductions in EER are obtained when the sigmoidal nonlinearities are trained and tested with the same noise, except with restaurant noise at SNR equal to 0 dB where both sigmoidal functions provide almost the same result.

4.6. General comments

The improvements provided by the use of the sigmoidal function are consistent with results from other studies in speech recognition based on auditory processing. Auditory-based approaches typically provide significant improvements at lower SNRs, but at higher SNRs they may achieve performance that is no better than (or worse than) the performance that is observed using conventional signal processing based on MFCC or PLP features (e.g. (Ghitza, 1986), (Chiu and Stern, 2008; Chiu et al., 2012), (Jankowski and Lippmann, 1992; Kim et al., 1999). We reiterate that the results presented here are consistent with those described in Chiu et al. (2012) where the sigmoidally-shaped rate-intensity function has been identified as an important component of auditory-based feature extraction systems for speech recognition.

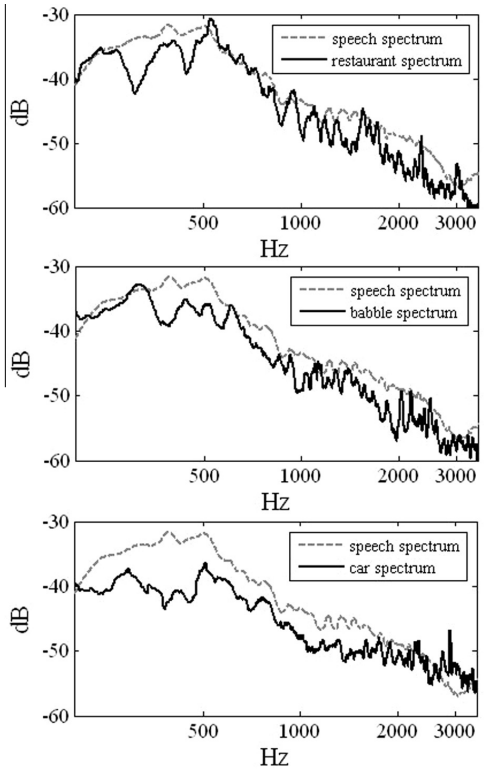

Fig. 14. Comparison of the average power spectrum for 50 utterances of clean speech with the power spectra of noise. The mean-squared errors (MSE) between the speech and restaurant, babble and car noise spectra are 58.7, 70.1, and 95.4, respectively.

Nevertheless, the use of the sigmoidal nonlinearity trained at 0-dB SNR failed to provide a significant improvement. As we have noted above, at 0-dB SNR the distributions of power of the frames containing degraded speech overlap with the power distributions of the noise-only frames. In addition, it appears that the power spectra of the speech signals and background noise are more similar in the case of restaurant noise than in the cases of the other two noise types considered. This is illustrated in Fig. 14, which shows the average power spectrum of 50 utterances extracted from clean speech and the power spectrum of restaurant, babble and car noises in each of the three panels. Spectra were estimated by using a FFT with 2^{15} points. An averaging filter was applied to smooth the speech and noise spectra. Finally, for comparison purposes, each spectrum was normalized according to its energy. We observed differences in mean-squared error (MSE) between the speech and noise spectra equal to 58.7, 70.1 and 95.4 for restaurant, babble and car noises, respectively. The corresponding differences between EERs obtained using the sigmoid nonlinearity combined with CVN compared with the use of CVN alone are 0.57%, 2.1%, and 4.9%, respectively, at SNR equal to 0 dB. Hence, we believe that the sigmoidal nonlinearity fails to improve EER for the speaker-verification task in restaurant noise at 0-dB SNR because the spectra of speech and noise are very similar, causing the power curves for degraded speech and noise to overlap at all frequencies.

5. Conclusions

This paper describes a method that can be used to develop an optimal sigmoidal nonlinear rectifier function for auditory modeling that is based solely on the distribution of power in the degraded speech frames and the power in the frames containing noise only. The objective function that is described attempts to simultaneously minimize noise power, minimize nonlinear distortion, maximize the similarity between clean speech and the degraded speech input, and maximize the signal variance of the speech degraded by noise after processing by sigmoidal function. The optimal sigmoidal functions obtained are frequency dependent because the output SNR from the channels of the initial

bandpass filter bank varies from one channel to the other. Finally, we note the proposed approach differs from cepstral mean and variance normalization (CMVN), which in effect produces a linear function that relates input and output, similar to the linear approximations of Fig. 4. The observed improvements in speaker identification accuracy obtained using the optimal sigmoidal nonlinearity (compared to results obtain with CMVN) demonstrate the potential of the nonlinearities that are part of human auditory processing.

The resulting sigmoidal nonlinearities are demonstrated to exhibit a location and slope that change as a function of SNR in a fashion that is consistent with the corresponding dependencies that are described in the physiological literature. The utility of the optimal sigmoidal nonlinearities derived in this fashion is considered in a series of experiments measuring speaker verification accuracy using the YOHO database. Our results indicate that the use of a sigmoidal nonlinearity defined strictly from the physical characteristics of the input (as apposed to phoneme discrimination) can lead to average relative reductions in EER compared to baseline processing as great as 12%, 33.6% and 16.6% at SNR equal to 10 dB, 5 dB and 0 dB, respectively, with speech degraded by babble, car and restaurant noise. The sigmoidal nonlinearity provides smaller benefit at higher SNRs, consistent with previous experiments with auditory models in speech recognition. The consistency of results between the two optimization schemes (using discrimination based on phoneme classes and discrimination based on waveform characteristics), reinforces the notion that optimal sigmoidal functions can reduce the mismatch between the training conditions and testing. In principle, our results obtained appear to be generic suggesting that this optimization approach could be applicable to any image or sound recognition system in which the feature extraction employs a nonlinear function based on rate-level responses.

Acknowledgements

This research was funded by Conicyt-Chile under grants Fondecyt 1100195 and Team Research in Science and Technology ACT 1120.

Appendix A

In this Appendix we develop the parameters A_j and B_j of the nonlinear distortion factor $D_j^{non-linear}(\omega_j, \mu_j)$, which are defined in Section 2.3.

The parameters A_j and B_j are estimated according to:

$$(A_j, B_j) = \arg\min_{A_j, B_j} \left\{ D_j^{non-linear}(\omega_j, \mu_j) \right\} \tag{A1}$$

First, the partial derivative of $D_j^{non-linear}(\omega_j, \mu_j)$ with respect to A_j is estimated:

$$\frac{\partial D_j^{non-linear}}{\partial A_j} = \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} 2 \cdot \left[A_j E_{j,m}^{sn} + B_j - g\left(E_{j,m}^{sn}\right) \right] \cdot E_{j,m}^{sn} \tag{A2}$$

Then, the result obtained in (A2) is set to zero:

$$\frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} 2 \cdot \left[A_j E_{j,m}^{sn} + B_j - g\left(E_{j,m}^{sn}\right) \right] \cdot E_{j,m}^{sn} = 0$$

$$A_j \cdot \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} \left(E_{j,m}^{sn}\right)^2 + B_j \cdot \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} E_{j,m}^{sn} = \frac{1}{N_f^{sn}} \sum_{m=1}^{N_f^{sn}} E_{j,m}^{sn} \cdot g\left(E_{j,m}^{sn}\right) \tag{A3}$$

$$A_j \cdot \mathbf{E}\left[\left(E_{j,m}^{sn}\right)^2\right] + B_j \cdot \mathbf{E}\left[E_{j,m}^{sn}\right] = \mathbf{E}\left[E_{j,m}^{sn} \cdot g\left(E_{j,m}^{sn}\right)\right]$$

Similarly, by estimating the derivative of $D_j^{non-linear}(\omega_j, \mu_j)$ with respect to B_j and setting the result to zero, the following equation is obtained:

$$A_j \cdot \mathbf{E}\left[\left(E_{j,m}^{sn}\right)^2\right] + B_j = -\mathbf{E}\left[g\left(E_{j,m}^{sn}\right)\right] \tag{A4}$$

By combining (A3) and (A4) and making use of the expressions $\mu_j = \mathbf{E}\left[E_{j,m}^{sn}\right]$ and $\sigma_j^2 = \mathbf{E}\left[\left(E_{j,m}^{sn}\right)^2\right] - \left\{\mathbf{E}\left[E_{j,m}^{sn}\right]\right\}^2$, the parameters A_j and B_j are found to be equal to:

$$\begin{aligned} A_j &= \tfrac{1}{\sigma_j^2}\left\{\mathbf{E}\left[E_{j,m}^{sn} \cdot g(E_{j,m}^{sn})\right] - \mu_j \cdot \mathbf{E}\left[g\left(E_{j,m}^{sn}\right)\right]\right\} \\ B_j &= \mathbf{E}\left[g\left(E_{j,m}^{sn}\right)\right] - \mu_j \cdot A_j \end{aligned} \tag{A5}$$

References

Ajmera, P.K., Jadhav, D.V., Holambe, R.S., 2011. Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram. Pattern Recognition 44 (10–11), 2749–2759.

Allen, J.B., 1985. Cochlear modeling. IEEE ASSP Magazine 2 (1), 3–29.

Barbour, D.L., 2011. Intensity-invariant coding in the auditory system. Neuroscience and Biobehavioral Reviews 35 (10), 2064–2072.

Bureš, Z., Grécová, J., Popelář, J., Syka, J., 2010. Noise exposure during early development impairs the processing of sound intensity in adult rats. European Journal of Neuroscience 32 (1), 155–164.

Campbell, J., Higgins, A., 1994. YOHO speaker verification. Linguistic Data Consortium, Philadelphia, PA.

Chiu, Y.-H.B., Stern, R.M., 2008. Analysis of physiologically-motivated signal processing for robust speech recognition. In: Proceedings of Interspeech, Brisbane, Australia, pp. 1000–1003.

Chiu, Y.-H.B., Raj, B., Stern, R.M., 2012. Learning-based auditory encoding for robust speech recognition. IEEE Transactions on Audio, Speech and Language Processing 20 (3), 900–914.

Cohen, J.R., 1989. Application of an auditory model to speech recognition. Journal of the Acoustical Society of America 85 (6), 2623–2629.

Costalupes, J.A., Young, E.D., Gibson, D.J., 1984. Effects of continuous noise backgrounds on rate response of auditory nerve fibers in cat. Journal of Neurophysiology 51 (6), 1326–1344.

Darwin, C.J., 2008. Listening to speech in the presence of other sounds. Philosophical Transactions of Royal Society B: Biological Science 363 (1493), 1011–1021.

Dean, I., Harper, N.S., McAlpine, D., 2005. Neural population coding of sound level adapts to stimulus statistics. Nature Neuroscience 8 (12), 1684–1689.

Dean, I., Robinson, B.L., Harper, N.S., McAlpine, D., 2008. Rapid neural adaptation to sound level statistics. Journal of Neuroscience 28 (25), 6430–6438.

Dimitriadis, D., Maragos, P., Potamianos, A., 2011. On the effects of filterbank design and energy computation on robust speech recognition. IEEE Transactions on Audio, Speech and Language Processing 19 (6), 1504–1516.

Gao, F., Zhang, J., Sun, X., Chen, L., 2009. The effect of postnatal exposure to noise on sound level processing by auditory cortex neurons of rats in adulthood. Physiology & Behavior 97, 369–373.

Garcia-Lazaro, J.A., Ho, S.S., Fair, A., Schnupp, J.W., 2007. Shifting and scaling adaptation to dynamic stimuli in somatosensory cortex. European Journal of Neuroscience 26 (8), 2359–2368.

Ghitza, O., 1986. Auditory nerve representation as a front-end for speech recognition in a noisy environment. Computer Speech & Language 1 (2), 109–131.

Ghitza, O., 1994. Auditory models and human performance in tasks related to speech coding and speech recognition. IEEE Transactions on Speech and Audio Processing 2 (1), 115–132.

Hanilçi, C., Kinnunen, T., Ertaş, F., Saeidi, R., Pohjalainen, J., Alku, P., 2012. Regularized all-pole models for speaker verification under noisy environments. IEEE Signal Processing Letters 19 (3), 163–166.

Hasan, T., Hansen, J.H.L., 2013. Acoustic factor analysis for robust speaker verification. IEEE Transactions on Audio, Speech and Language Processing 21 (4), 842–853.

Hirsch, H.G., Pearce, D., 2000. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condition. In: ISCA ASR2000-Automatic Speech Recognition: Challenges for the Next Millennium, Paris, pp. 181–188.

Jankowski, C.R., Lippmann, R.P., 1992. Comparison of auditory model for robust speech recognition. In: Proceedings of the Workshop on Speech and Natural Language, Stroudsburg, PA, pp. 453–454.

Kang, S.Y., Colesa, D.J., Swiderski, D.L., Su, G.L., Raphael, Y., Pfingst, B.E., 2010. Effects of hearing preservation on psychophysical responses to cochlear implant stimulation. Journal of the Association for Research in Otolaryngology 11 (2), 245–265.

Kim, D.S., Lee, S.Y., Kil, R.M., 1999. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. IEEE Transactions on Speech and Audio Processing 7 (1), 55–69.

Kim, C., Chiu, Y.-H.B., Stern, R.M., 2006. Physiologically-motivated synchrony-based processing for robust speech recognition. In: Proceedings of Interspeech, Pittsburgh, Pennsylvania, pp. 1975–1978.

Kim, C., Stern, R.M., 2012. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. In: Proceedings Acoustics, Speech and, Signal Processing, pp. 4101–4104.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: from features to supervectors. Speech Communication 52 (1), 12–40.

Kinnunen, T., Saeidi, R., Sedlák, F., Lee, K.A., Sandberg, J., Hansson-Sandsten, M., Li, H., 2012. Low-variance multitaper MFCC features: a case study in robust speaker verification. IEEE Transactions on Audio, Speech and Language Processing 20 (7), 1990–2001.

Li, Q., Huang, Y., 2010. Robust speaker identification using and auditory-based feature. In: Proceedings of Acoustics Speech and, Signal Processing, pp. 4514–4517.

Li, Q., Huang, Y., 2011. An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. IEEE Transactions on Audio, Speech and Language Processing 19 (6), 1791–1801.

Lyon, R.F., 1982. A computational model of filtering, detection, and compression in the cochlea. In: Proceedings of the IEEE-International Conference on Acoustics, Speech, and Signal Processing, Paris, pp. 1282–1285.

May, B.J., Sachs, M.B., 1992. Dynamic range of neural rate responses in the ventral cochlear nucleus of awake cats. Journal of Neurophysiology 68 (5), 1589–1602.

Middlebrooks, J.C., 2004. Effects of cochlear-implant pulse rate and inter-channel timing on channel interactions and thresholds. Journal of the Acoustical Society of America 116 (1), 452–468.

Miller, C.A., Woo, J., Abbas, P.J., Hu, H., Robinson, B.K., 2011. Neural masking by sub-threshold electric stimuli: animal and computer model results. Journal of the Association for Research in Otolaryngology 12 (2), 219–232.

Ming, J., Hazen, T.J., Glass, J.R., Reynolds, D.A., 2007. Robust speaker recognition in noisy conditions. IEEE Transactions on Audio, Speech and Language Processing 15 (5), 1711–1723.

Moore, B.C.J., 2003. An Introduction to the Psychology of Hearing, 5th ed. Academic Press, London, pp. 39–41.

Nizami, L., 2005. Dynamic range relations for auditory primary afferents. Hearing Research 208 (1–2), 26–46.

Ohzawa, I., Sclar, G., Freeman, R.D., 1985. Contrast gain control in the cat's visual system. Journal of Neurophysiology 54 (3), 651–658.

Patterson, R.D., Holdsworth, J., Allerhand, M., 1992. Auditory models as preprocessors for speech recognition. In: Schouten, M.E.H. (Ed.), The Auditory Processing of Speech: From Sounds to Words. Mouton de Gruyter, Berlin, Germany, pp. 67–83 (Chapter 1).

Pfingst, B.E., Bowling, S.A., Colesa, D.J., Garadat, S.N., Raphael, Y., Shibata, S.B., Strahl, S.B., Su, G.L., Zhou, N., 2011. Cochlear infrastructure for electrical hearing. Hearing Research 281 (1–2), 65–73.

Pickles, J.O., 2008. An Introduction to the Physiology of Hearing, 3rd ed. Emerald Group, Bingley, England, ch. 4.

Rabinowitz, N.C., Willmore, B., Schnupp, J., King, A.J., 2011. Contrast gain control in auditory cortex. Neuron 70 (6), 1178–1191.

Reynolds, D., 1995. Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17 (1–2), 91–108.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian Mixture Models. Digital Signal Processing 10 (1–3), 19–41.

Rhode, W.S., Cooper, N.P., 1993. Two-tone suppression and distortion production on the basilar membrane in the hook region of cat and guinea pig cochleae. Hearing Research 66, 31–45.

Robles, L., Ruggero, M., 2011. Mechanics of the mammalian cochlea. Physiological Reviews 81 (3), 1305–1352.

Sachs, M.B., Abbas, P.J., 1974. Rate versus level functions for auditory-nerve fiber in cats: tone burst stimuli. Journal of the Acoustical Society of America 56 (6), 1835–1847.

Saeidi, R., Pohjalainen, J., Kinnunen, T., Alku, P., 2010. Temporally weighted linear prediction features for tackling additive noise in speaker verification. IEEE Signal Processing Letters 17 (6), 599–602.

Schneider, B.A., Parker, S., Murphy, D., 2011. A model of top down gain control in the auditory system. Attention, Perception and Psychophysics 73 (5), 1562–1578.

Seneff, S., 1988. A joint synchrony/mean-rate model of auditory speech processing. Journal of Phonetics 16 (1), 55–76.

Shamma, S.A., 1985. Speech processing in the auditory system I: the representation of speech sounds in the responses of the auditory nerve. Journal of the Acoustical Society of America 78 (5), 1612–1621.

Shamma, S.A., 1988. The acoustics features of speech sounds in a model of auditory processing: vowels and voiceless fricatives. Journal of Phonetics 16, 77–91.

Shao, Y., Srinivasan, S., Wang, D.L., 2007. Incorporating auditory feature uncertainties in robust speaker identification. In: Proceedings of Acoustics Speech and Signal Processing, vol. IV, pp. 277–280.

Shao, Y., Wang, D.L., 2008. Robust speaker identification using auditory features and computational auditory scene analysis. In: Proceedings of Acoustics Speech and, Signal Processing, pp. 1589–1592.

Shao, Y., Srinivasan, S., Jin, Z., Wang, D.L., 2010. A computational auditory scene analysis system for speech segregation and robust speech recognition. Computer Speech & Language 24 (1), 77–93.

Shin, J.W., Kwon, H.J., Jin, S.H., Kim, N.S., 2008. Voice activity detection based on conditional MAP criterion. IEEE Signal Processing Letters 15, 257–260.

Slaney, M., Auditory Toolbox, Version 2, Technical Report No. 1998–010, Interval Research Corporation, 1998.

Stern, R.M., Morgan, N., 2012a. Features based on auditory physiology and perception. In: Virtanen, T., Raj, B., Singh, R. (Eds.), . In: Techniques for Noise Robustness in Automatic Speech Recognition. Wiley.

Stern, R.M., Morgan, N., 2012b. Hearing is believing: biologically-inspired feature extraction for robust speech recognition. IEEE Signal Processing Magazine 20 (6), 34–43.

Taberner, A.M., Liberman, M.C., 2005. Response properties of single auditory nerve fibers in the mouse. Journal of Neurophysiology 93 (1), 557–569.

Wang, K., Shamma, S., 1994. Self-normalization and noise-robustness in early auditory representations. IEEE Transactions on Speech and Audio Processing 2 (3), 421–435.

Wang, N., Ching, P.C., Zheng, N., Lee, T., 2011. Robust speaker recognition using denoised vocal source and vocal tract features. IEEE Transactions on Audio, Speech and Language Processing 19 (1), 196–205.

Watkins, P.V., Barbour, D.L., 2011. Level-tuned neurons in primary auditory cortex adapt differently to loud versus soft sounds. Cerebral Cortex 21 (1), 178–190.

Wen, B., Wang, G.I., Dean, I., Delgutte, B., 2009. Dynamic range adaptation to sound level statistics in the auditory nerve. Journal of Neuroscience 29 (44), 13797–13808.

Wen, B., Wang, G.I., Dean, I., Delgutte, B., 2012. Time course of dynamic range adaptation in the auditory nerve. Journal of Neurophysiology 108 (1), 69–82.

Werblin, F.S., Jacobs, A., Teeters, J., 1996. The computational eye. IEEE Spectrum 33 (5), 30–37.

Winslow, R.L., Sachs, M.B., 1987. Effect of electrical stimulation of the crossed olivocochlear bundle on auditory nerve response to tones in noise. Journal of Neurophysiology 57 (4), 1002–1021.

Wu, W., Zheng, T.F., Xu, M.-X., Soong, F.K., 2007. A cohort-based speaker model synthesis for mismatched channels in speaker verification. IEEE Transactions on Audio, Speech and Language Processing 15 (6), 1893–1903.

Yates, G.K., Winter, I.M., Robertson, D., 1990. Basilar membrane nonlinearity determines auditory nerve rate-intensity functions and cochlear dynamic range. Hearing Research 45 (3), 203–219.

Young, E.D., 2008. Neural representation of spectral and temporal information in speech. Philosophical Transactions of Royal Society B: Biological Science 363 (1493), 923–945.

Zilany, M.S., Carney, L.H., 2010. Power-law dynamics in an auditory-nerve model can account for neural adaptation to sound-level statistics. The Journal of Neuroscience 30 (31), 10380–10390.

A perceptually-motivated low-complexity instantaneous channel normalization technique applied to speaker verification

Victor Poblete[a,b], Felipe Espic[a], Simon King[c], Richard M. Stern[d], Fernando Huenupán[e], Nestor Becerra Yoma[a,*]

[a] *Speech Processing and Transmission Laboratory, Electrical Engineering Department, University of Chile, Santiago, Chile*
[b] *Institute of Acoustics, Universidad Austral de Chile, Valdivia, Chile*
[c] *Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK*
[d] *Department of Electrical and Computer Engineering and Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA*
[e] *Departamento de Ingeniería Eléctrica, Universidad de la Frontera, Temuco, Chile*

Abstract

This paper proposes a new set of speech features called Locally-Normalized Cepstral Coefficients (LNCC) that are based on Seneff's Generalized Synchrony Detector (GSD). First, an analysis of the GSD frequency response is provided to show that it generates spurious peaks at harmonics of the detected frequency. Then, the GSD frequency response is modeled as a quotient of two filters centered at the detected frequency. The numerator is a triangular band pass filter centered around a particular frequency similar to the ordinary Mel filters. The denominator term is a filter that responds maximally to frequency components on either side of the numerator filter. As a result, a local normalization is performed without the spurious peaks of the original GSD. Speaker verification results demonstrate that the proposed LNCC features require low computational complexity and far more effectively compensate for spectral tilt than ordinary MFCC coefficients. LNCC features do not require the computation and storage of a moving average of the feature values, and they provide reductions in EER as high as 32% and 35% when compared with MFCC and MFCC+CMN with variable spectral tilt, respectively.

Keywords: channel robust feature extraction; auditory models; spectral local normalization; synchrony detection.

1. Introduction

1.1. Motivation

The use of perceptually-motivated features is widespread across spoken language technology, with non-linear frequency scales and compression of the dynamic range of the spectral energy (*e.g.* by taking the logarithm or cube root of filterbank outputs) being ubiquitous. In automatic speech recognition (Gales, 1998; Hermansky et al., 2013), speaker diarization (Tranter & Reynolds, 2006) and speaker verification (Reynolds & Rose, 1995; Campbell, 1997),

*Corresponding author
Email addresses: vpoblete@ing.uchile.cl (Victor Poblete), fespic@ing.uchile.cl (Felipe Espic), Simon.King@ed.ac.uk (Simon King), rms@cs.cmu.edu (Richard M. Stern), fhuenu@ufro.cl (Fernando Huenupán), nbecerra@ing.uchile.cl (Nestor Becerra Yoma)

generally at least as good as CMN in all scenarios tested, and superior in the case of rapidly changing channels.

1.2. The need for robust speech features

In the current work, we restrict ourselves to dealing with (mostly linear) channels whose frequency response may differ between training and testing data, that may vary from one test utterance to the next or indeed within an utterance, and that is unknown at test time. We aim to extract features from the speech signal that are robust – by which we mean invariant – to changes in the channel (*e.g.* Togneri & Pullella (2011)). Specifically, we target variations in the frequency response of either acoustic channels which are a consequence of the relative position of the speaker with respect to the microphone.

1.2.1. Time-varying channels

A great number of approaches have been described in the literature to enhance the robustness of automatic speech and speaker recognition systems with respect to changes in the channel. We do not attempt a survey of these methods here, but refer the reader to, for example Seltzer et al. (2004); Buchner et al. (2005); Morales et al. (2009); Meyer & Kollmeier (2011); Lu et al. (2011). Typically, such methods attempt to improve recognition accuracy for cases where the training and testing data have been acquired under different acoustic conditions – for example, in order to enable the systems to cope with changes in microphone. Some methods aim to extract invariant features, while others attempt to adjust the statistical model. Our proposed method is of the former type, but could in principle be combined with model compensation techniques.

1.2.2. Application scenarios

In numerous real applications, the channel between the speaker and the automatic speech recognition (or speaker verification, speaker diarization, ...) system may vary over time. A few examples of such applications are mentioned below.

Meeting transcription. The task of richly transcribing human-human interactions has received considerable attention over the last decade (Hori et al., 2012; Yokoyama et al., 2013), particularly for the scenario of small business meetings with around 4 participants (Hain et al., 2006, 2007; Renals et al., 2007; Hain et al., 2012). A key problem in this domain is dealing with distantly-positioned microphones, such as those on the table-top, in randomly-positioned portable devices, or comprising microphone arrays. Tasks that are performed on speech captured in this way range from speech detection, speaker diarization, and transcription of the words, to higher-level analysis such as content-linking (Sangwan et al., 2013; Malionek et al., 2013).

The use of microphone arrays is widespread in this task domain, because of their ability to beamform, and thus to somewhat isolate the signal of a target speaker from other speakers or noise sources. Nevertheless, the properties of the physical acoustic channel between speaker and microphone (or microphone array) are still highly variable and are a cause of degradation in, for example, accuracy of transcriptions. Sources of variability in this channel include

distance between speaker and microphone, beamforming arrays which low-pass filter off-axis signals (Brandstein & Ward, 2010, their Figure 1.1), occlusion of the direct path by intervening objects such as laptop screens (Wölfel & McDonough, 2009, their Figure 1.1), and so on.

We address one component of this complex puzzle, and – as will be justified in Section 4.5.1 – we will model the situation as an unknown and potentially time-varying spectral tilt imposed on the test recordings.

Lecture transcription. Another task that has received a growing level of attention recently is that of transcribing lectures (Trancoso et al., 2006; Bell et al., 2013). Typically, this task is performed from recordings made with lapel-microphones, which are used because they are relatively discrete compared to close-talking headsets. Unfortunately, this leads to frequent and rapid changes in the acoustic channel between speaker and microphone, due to head turning. While acceptably low error rates are possible in good conditions, when acoustic conditions degrade, the Word Error Rate (WER) can increase to 40-45% (Leeuwis et al., 2003; Park et al., 2005; Hsu & Glass, 2006; Glass et al., 2007). The alternative to lapel microphones is the use of distant microphones or arrays, but these are subject to similar problems, as described above.

Human-machine interaction. Speaking and hearing take place in situations where the acoustic environment is not constant and where speakers are affected by auditory input from the environment, other speakers, and feedback of their own speech (Cooke et al., 2013b). Background noise causes speakers to adjust their speech in a variety of ways (see Cooke et al. (2014, 2013a) for a comprehensive review) including so-called 'Lombard' speech (Cooke & Lecumberri, 2012), in which one of the principal changes in addition to increased intensity is a reduction in the spectral tilt, leading to an overall flatter spectrum compared to normal. Often, speakers and listeners are also mobile, with each making continuous adjustments to speaking style and head position to compensate for the changing channel.

Machines that listen, whether they are socially-interactive mobile robots operating in public spaces such as supermarkets, museums, and expositions (Jensen et al., 2005; Ishi et al., 2008) or static systems using beamforming microphone arrays (Wölfel & McDonough, 2009) are faced with the same challenges of varying channel and speaking style – for example, the spectral tilt of the speaker's speech will varying with their speaking effort, which will change with the physical distance between speaker and 'listener' (robot or microphone array); the frequency response of a directional microphone will vary (typically with increased spectral tilt due to low-pass filtering) when the speaker is off-axis, compared to being on-axis (see Section 4.5.1 for experimental verification of this effect).

1.3. Scope of our work

The auditory model-inspired features that we will introduce in Section 3 are specifically designed to be inherently and instantaneously robust to unknown and potentially time-varying channel frequency response, as present in the applications described in Section 1.2.2. Therefore, we limit the experimental investigation reported in Section 4 to such a scenario and do not address other aspects of robustness, such as additive noise or reverberation.

2. Development of the proposed approach from an auditory model

Models of the auditory system attempt to capture various behaviors of the natural system that they are mimicking (Stern & Morgan, 2012b). Some of these behaviors may be useful for speech feature extraction, so in this section we motivate our proposed features by starting from auditory models. We will identify a behavior which acts as a localised and instantaneous normalization, and which is not currently part of typical perceptually-motivated speech features used in pattern recognition applications.

One problem with such typical features, such as MFCCs or PLPs, is that they capture not only important speech features such as the frequencies of formants, but also channel properties too such as overall spectral tilt (Hansen & Varadarajan, 2009). Of course, a vast array of noise-robustness techniques is available to be applied either to these features, or to models learned from them, such approaches. The features we propose are inherently less variant to channel differences than MFCCs.

2.1. Auditory modelling

In speech technology, the most widely-used features are (usually decorrelated) representations of the envelope of the power spectrum (Wölfel, 2009a,b). On the other hand, auditory modelling has long known that the auditory system makes use not only of the spectral envelope but also information related to the synchrony between the responses in different nerve fibres (Johnson, 1980; Sachs, 1984; Eggermont, 1998; Dreyer & Delgutte, 2006). This synchrony-related information is more invariant to signal level differences than the rate-place representation of the spectral energy, is able to capture periodic signals even in the presence of noise, and so is thought (Smith et al., 2002; Moore, 2008; Heinz & Swaminathan, 2009) to be one of the reasons for the auditory system's incredible robustness to a wide variety of listening conditions, such as additive noise or channel distortion (Ghitza, 1994; Shao et al., 2010; Anderson et al., 2010).

2.1.1. Mean rate representations and the spectral envelope

Most conventional feature extraction schemes (such as MFCC and PLP coefficients) are based on short-time energy in a set of frequency bands, which is more directly related to mean-rate than temporal synchrony in the physiological responses of the auditory system (Davis & Mermelstein, 1980; Hermansky, 1990, 1994; Dimitriadis et al., 2011). For example, the Mel-scaled filterbank, from which MFCCs are derived, captures the spectral envelope only (Kumaresan & Rao, 1999). The spectral envelope obviously carries information about both the speech signal and the transmission channel and any additive noise (Kuwabara & Sagisaka, 1995; Watkins & Makin, 1996; Zilovic et al., 1998; Parikh & Loizou, 2005; Miettinen et al., 2011). Separating these out *after* feature extraction is a blind separation problem and therefore only solvable by making some assumptions. A typical assumption would be that the channel changes more slowly than the speech spectrum (Stockham et al., 1975; Hermansky, 1994; Gaubitch et al., 2013); this leads to a method in which a relatively long-term average is subtracted in the cepstral domain – Cepstral Mean Normalization (CMN) (Atal, 1974; Furui, 1981; Schwartz et al., 1993; Liu et al., 1993; Hermansky, 1994). The

disadvantage of this type of normalization is that it requires the estimation of the average cepstrum over some window (*e.g.* , all frames of the current utterance, or the previous N frames) (Soong & Rosenberg, 1988; Rose & Reynolds, 1990); if too short a window is used, then the estimated mean will contain some speech information, not just channel information. If the assumption about the channel changing slowly relative to the selected window/batch size is not correct, then the estimated mean will not accurately reflect the channel response and the normalization will be less effective (Bořil & Hansen, 2010; Nakano et al., 2010; Wang et al., 2011).

2.1.2. Average localized synchrony rate (ALSR)

In addition to mean rate representations, the auditory system is known to make use of another representation which captures temporal information, although precisely how the two are combined in the brain remains an open question (Moore, 2014). While temporal coding is clearly important for binaural sound localization (Stern et al., 2006; Joris & Yin, 2007), it may also play a role in the robust interpretation of signals from individual ears as well (Young, 2008).

For example, Young & Sachs (1979) demonstrated that the average localized synchrony rate (ALSR) that is derived from auditory nerve firing is much more robust to changes in intensity of vowel-like sounds than the corresponding mean-rate of response as a function of characteristic frequency (CF). The ALSR describes the extent to which the neural response at a given CF is synchronized to the nearest harmonic of the fundamental frequency of the vowel. These results suggest that the timing information associated with the response to low-frequency components of a signal can be substantially more robust to variations in intensity (and potentially various other types of signal variability and/or degradation such as varying channel or additive noise) than the mean-rate of the neural response.

A vast array of auditory models which include synchrony detection have been proposed (*e.g.* Jankowski & Lippmann (1992); Jankowski et al. (1995); Ali et al. (2002); Kim et al. (2006) for helpful reviews), and so we do not offer a survey of them here. Instead, we focus on the particular model that was the inspiration for the features we propose.

2.1.3. From ALSR to Generalized Synchrony Detector (GSD)

Seneff's auditory model (Seneff, 1988) consists of 40 recursive linear filters implemented in cascade form which cover a frequency range from 130 to 6400 Hz. The bandwidth of the channels is 0.5 Bark (Seneff, 1988). These filters mimic the nominal auditory-nerve frequency responses as described by Kiang et al. (1965) and other contemporary physiologists (Liberman, 1978; Young & Sachs, 1979; Sachs & Young, 1979; Sinex & Geisler, 1983; Delgutte & Kiang, 1984; Pickles, 2008). Seneff's model employs an "inner hair cell model" that includes four stages: (1) nonlinear half-wave rectification using an inverse tangent function for positive inputs and an exponential function for negative inputs, (2) short-term adaptation that models the release of transmitter in the synapse, (3) a lowpass filter with cutoff frequency of approximately 1 kHz to suppress synchronous response at higher input frequencies, and (4) a rapid automatic gain control (AGC) stage to maintain an approximately-constant response rate at higher input intensities when an auditory-nerve fibre is nominally in saturation.

Reflecting the fact that the auditory system makes use of two representations, Seneff proposed two non-interacting

parallel modules that operate on the hair-cell model outputs. The first of these was an envelope detector, which produced a statistic intended to model the instantaneous mean-rate of response of a given fibre. The second operation was called a GSD, motivated by the ALSR measure of Young & Sachs (1979) and each channel i is modelled (Seneff, 1985; Ali et al., 2002) as in Equation 1, where $y[n]$ is the speech waveform value at sample n.

$$\mathrm{GSD}_i(y) = A_s \arctan\left[\frac{1}{A_s}\left(\frac{\langle|y[n] + y[n - n_i]|\rangle - \delta}{\langle|y[n] - \beta^{n_i} y[n - n_i]|\rangle}\right)\right] \tag{1}$$

The hair-cell output for this channel i is compared to itself delayed by the reciprocal of the centre frequency f_i^c of the filter in each channel (n_i in Equation 1), and the short-time averages (i.e., envelope detection, denoted by $\langle \ldots \rangle$ in Equation 1) of the magnitudes (denoted by $|\ldots|$ in Equation 1) of the sums and differences of these two quantities are divided by one another. A threshold δ is introduced to suppress response to low-intensity signals and the resulting quotient is passed through a saturating half-wave rectifier ($\arctan[\ldots]$ in Equation 1) to limit the magnitude (Seneff, 1985). A value slightly less than 1 is used for the constant β in the denominator while the constant δ in the numerator has a rather small value (Seneff, 1985). The parameter A_s represents a control in the linear range for the input speech waveform (Seneff, 1985; Ali et al., 2002).

With the limited computational resources available at that time, Seneff could only compare the mean-rate and GSD response visually for selected inputs. The GSD display did indeed to provide a useful representation of the spectral components, including in noise (see *e.g.* Seneff (1985); Chigier & Leung (1992); Jankowski & Lippmann (1992); Ohshima & Stern (1994))

Newer and more sophisticated models than Seneff's have of course been proposed in more recent times (see Moore (2003); Pickles (2008); Stern & Morgan (2012a); Moore (2014) for a comprehensive review). Nevertheless, these newer models are not relevant to the work described in this paper because we are using a particular property of Seneff's model as the *inspiration* for our proposed method, rather than implementing the complete model.

2.2. The potential of GSD-like features for speech recognition

2.2.1. Previous attempts to use this model

The generalized synchrony detector model proposed by Seneff (Seneff, 1985) corresponds to one of the first attempts for developing a spectral representation from the temporal coding that occurs in the auditory nerve fibres (instead of their rate codes) for use as front ends to automatic speech recognition systems (Seneff, 1986b; Stern & Morgan, 2012a). Seneff reported strong evidences that auditory based representations are interesting and worthy of study in speech analysis systems. According to Seneff (1988) preliminary results of the two distinct spectral representations for the speech signal, one based on the discharge rate (rate coding) of the auditory nerve fibres and the other based on the synchronous response of the fibres (synchrony coding), indicated that the rate response outputs are successful for locating acoustic boundaries. Similarly, the synchrony outputs applied to speaker-independent vowel recognition in continuous speech showed superior performance (Seneff, 1987). However, there was no explanation on

the neural interaction mechanisms between rate versus synchrony coding and how the auditory brain uses some of the information of theses two representations in real communication situations (Smith et al., 2002; Moore, 2008).

Although Seneff's GSD has been used as a feature extraction method for speech recognition, such as the detection of formant frequencies (Seneff, 1984, 1986a; Kim et al., 1999), its performance to conventional mean-rate inspired features such as MFCCs (Jankowski et al., 1995; Ali et al., 2002) has been mixed. In general, in clean speech, GSD features provide recognition accuracies that are no better than what is provided conventional MFCC or PLP features (and in some cases their performance is worse), but in additive noise, they can be helpful (*e.g.* Chiu & Stern, 2008). An extension of the Seneff GSD was proposed by Ali et al. (2002). This approach known as Average Localized Synchrony Detection also produces a synchrony spectrum and provides better recognition results under noise conditions than the Seneff's original GSD detector.

Furthermore, the GSDs must be perfectly tuned to the formant frequencies in order to obtain a clean output (Seneff, 1988). This was a major problem of the original GSD algorithm (Ali et al., 2002).

2.2.2. A frequency-domain analysis of GSD

Seneff's original GSD is defined in the time domain by Eq. 1. Because it is more convenient to perform feature extraction for speech/speaker recognition in the frequency domain, we perform a frequency-domain analysis of Seneff's GSD by passing pure tones (sinusoids) at different frequencies sweeping the entire spectrum, using the time-domain filter of Equation 1, which is effectively a form of frequency analysis. We consider only the magnitude response of the GSD filters and neglect phase since we assume that phase is unimportant in automatic speech/speaker recognition applications. Equation 1 is the ratio of two terms, numerator and denominator, which can be analyzed separately. Equations 2 and 3 give these terms, which are computed for each channel i at each analysis frame.

$$\text{Numerator}_{\text{GSD}} = \langle |y[n] + y[n - n_i]| \rangle - \delta \tag{2}$$

$$\text{Denominator}_{\text{GSD}} = \langle |y[n] - \beta^{n_i} y[n - n_i]| \rangle \tag{3}$$

Consider, as an example, the response of the numerator and denominator terms of one GSD channel (after a bandpass filter) tuned to a center frequency f_i^c of 692 Hz, to 1024 pure tones spanning the frequency range 60 Hz to 3500 Hz, as shown in Figure 1.

2.2.3. Spurious responses of the GSD

The frequency responses of numerator and denominator shown in Figure 1 initially look promising, being centered at the tuned frequency of 692Hz as expected, and with the denominator bandwidth being slightly wider than that of the numerator. Nevertheless, if we examine the final GSD response – the numerator divided by the denominator – as plotted in Figure 2, we observe additional peaks at higher frequencies, along with the desired peak at 692 Hz. Seneff

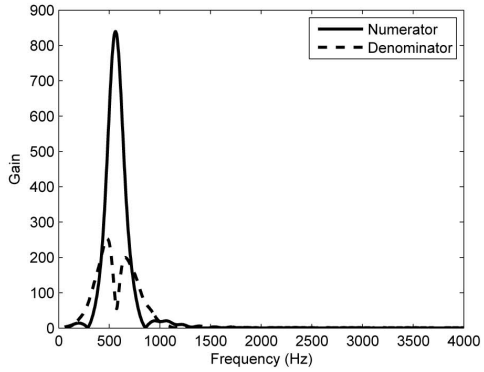

Figure 1: Frequency response of the numerator and denominator of a GSD tuned at 692 Hz. The numerator in Equation 2 is shown as a solid line and the denominator in Equation 3 is shown as a dashed line. The values used for the constants are $\delta = 1 \times 10^{-5}$ and $\beta = 0.999$.

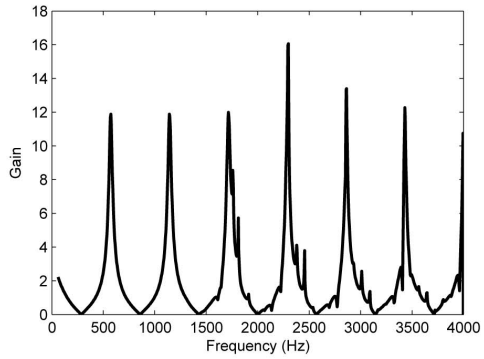

Figure 2: Frequency response of a single GSD channel at $f_i^c = 692$ Hz.

herself describes this limitation of the GSD (Equation 1), stating that it produces spurious peaks at harmonics of the detected frequency. These observations notwithstanding, the behavior of each GSD channel in the region around its center frequency still has desirable properties, and we will exploit these in our proposed features described in Section 3 below.

Examining the numerator and denominator responses plotted on a logarithmic scale as in Figure 3 reveals the cause of this behavior. In the figure, the denominator is plotted as its reciprocal to understand more clearly its relationship with the numerator. In the next section, we construct a GSD-like channel that preserves the desirable normalization behavior provided by the denominator term, but that does not produce spurious responses outside its nominal "passband".

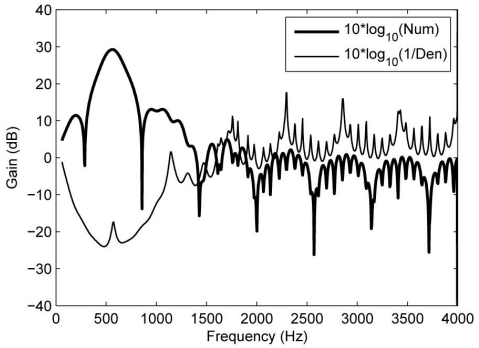

Figure 3: Log magnitude of the frequency response of the numerator and denominator of the GSD tuned at f_i^c = 692 Hz. The numerator (Equation 2) is shown with a thick line and the reciprocal of the denominator (Equation 3) with a thin line.

3. The proposed features

Our target applications, described in Section 1.2.2, involve channels with time-varying frequency responses, including situations in which the physical arrangement of speaker and microphone may vary. We therefore seek features that are relatively invariant to changes in the channel frequency response. The proposed features achieve this by using a form of local normalization inspired by the ratio between the numerator and denominator terms of Seneff's GSD, given in Equations 2 and 3. We refer to these features as LNCC, for Locally-Normalized Cepstral Coefficients.

3.1. From the GSD to a frequency-domain model suitable for speech technology

By examining the behavior of the GSD, we can identify some desirable attributes that are not found in typical features such as MFCCs. We note from the form of Equation 1 and from Figure 3 (ignoring for the moment the spurious higher-frequency responses)that the numerator part acts as a band-pass filter centered around a particular frequency, and that its output is divided by (i.e. normalized by) a denominator term which is a filter that responds to energy on either side of the numerator filter. In other words, a local normalization is being performed: the output of a GSD channel relates to the amount of energy in a particular frequency band *relative* to the energy in neighboring (lower and higher frequency) regions. With an appropriately-selected filter bandwidth, the effect is one of preserving spectral peaks (which are speech-related) while being relatively invariant to overall spectral tilt, for example. We note that the concept of a response in a localized central region being inhibited or suppressed by a response over a broader range of space or frequency is commonly encountered in vision (*e.g.* Werblin et al., 1996) and audition (*e.g.* Sachs & Kiang, 1968; Houtgast, 1972), and Wang & Shamma (1994), among others, have commented on the utility of this type of mechanism for speech recognition. We can achieve a similar behavior directly in the frequency domain, by designing simple filters for the numerator and denominator respectively (Figure 4). Such a pair of filters will perform, in the frequency domain, a similar local normalization to that performed in the time-domain by GSD (Equation 1).

By working in the frequency domain (just as in conventional MFCCs), the filters can easily be designed so as to only respond within the main passband, eliminating the spurious higher-frequency peaks seen in Figure 2.

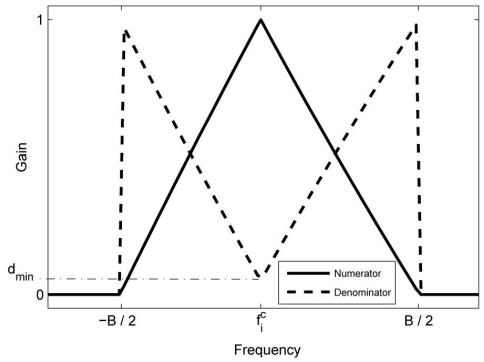

Figure 4: The shapes of the magnitude of the numerator (solid line) and denominator (dashed line) filters, for a single channel of the proposed self-normalizing filterbank. f_c is the center frequency of the channel, $d_{\min}$ the minimum centered value of the denominator, and B its bandwidth. In our work, these frequencies are defined on the Bark scale (Zwiker, 1961).

The pair of filters for one such channel were designed through informal experimentation. Responses of the pair of actual filters configured at a center frequency f_i^c of 515 Hz are shown in Figure 5. The numerator filter is essentially the same as the triangular filter commonly employed in the filterbank used to derive MFCCs (Davis & Mermelstein, 1980) and is described in the frequency domain by Equation 4 for each channel i with center frequency f_i^c. The denominator filter captures energy on either side of the numerator filter; it is described by Equation 5.

$$\text{Numerator}_{\text{LNCC}}(f) = \begin{cases} -\frac{2}{B}|f - f_i^c| + 1 & \text{when } |f - f_i^c| \leq \frac{B}{2} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$\text{Denominator}_{\text{LNCC}}(f) = \begin{cases} \frac{2}{B}\left(1 - d_{\min}\right)|f - f_i^c| + d_{\min} & \text{when } |f - f_i^c| \leq \frac{B}{2} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

While both filters $\text{Numerator}_{LNCC}(f)$ and $\text{Denominator}_{LNCC}(f)$ have a nonzero response only for frequencies in the range of $-\frac{2}{B} \leq |f - f_i^c| \leq \frac{2}{B}$, it is easily seen that the response of $\text{Numerator}_{LNCC}(f)$ is greatest for a narrow range of frequencies about $f = f_i^c$, while $\text{Denominator}_{LNCC}(f)$ is responsive to activity in the surrounding frequency regions. By assembling a bank of such filter pairs, we can extract a locally-normalized filterbank representation of the signal, which can then be used subsequently to compute cepstral features, following the same steps as for deriving MFCCs from conventional filterbank outputs (Davis & Mermelstein, 1980). In all experiments presented in this paper, the filters are constructed on a Bark scale.

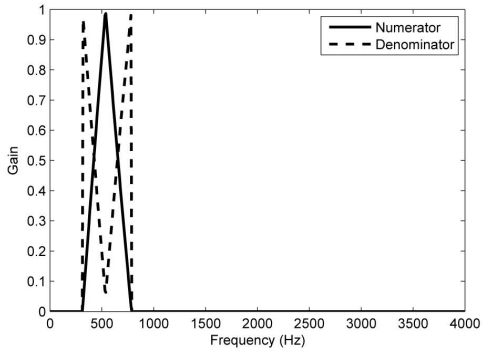

Figure 5: Frequency response of the numerator and denominator tuned at $f_i^c = 515$ Hz.

It is trivial to replace the filterbank normally used in MFCC feature extraction with this bank of self-normalizing filter pairs. By spacing the filters on a perceptual scale (such as the Bark scale used in our work) followed by logarithmic compression and a truncated cosine transform, we derive speech features that will have very similar properties to conventional MFCCs (*e.g.* they are statistically decorrelated), but with the addition of the local normalization during the filterbank stage. The overall effect combines filtering with a filterbank (which removes fine detail from the spectrum such as harmonics of the fundamental frequency F_0) and local normalization (which removes very coarse variations in the spectral shape, such as overall tilt, which we assume arise mostly from channel variability).

In other words, the proposed features can be used as a straightforward "drop-in" replacement for MFCCs without any changes to the statistical model, for example. Figure 6 describes the complete sequence of steps required to extract LNCC features, and shows the corresponding steps for conventional MFCC feature extraction for comparison.

3.1.1. Frequency response of the proposed self-normalizing filter pairs

The frequency responses of the proposed numerator and denominator filters defined in Equations 4 and 5 are illustrated in Figure 7 which plots the individual responses of one pair of numerator and denominator filters, on a logarithmic scale. The combined response of the numerator divided by the denominator is plotted in Figure 8. This plot reveals that, when combined, the pair of filters in LNCC exhibits a sharper response than the triangular filters in a typical MFCC filterbank.

While Figure 8 shows the response to pure tones, it is more useful to examine the response to a broadband signal (i.e. a vowel) in order to observe the normalization effect. Figure 9 plots the spectral envelopes estimated by the proposed normalized filterbank, and compares this to the corresponding response of a conventional filterbank such as the filters typically used to derive MFCCs.

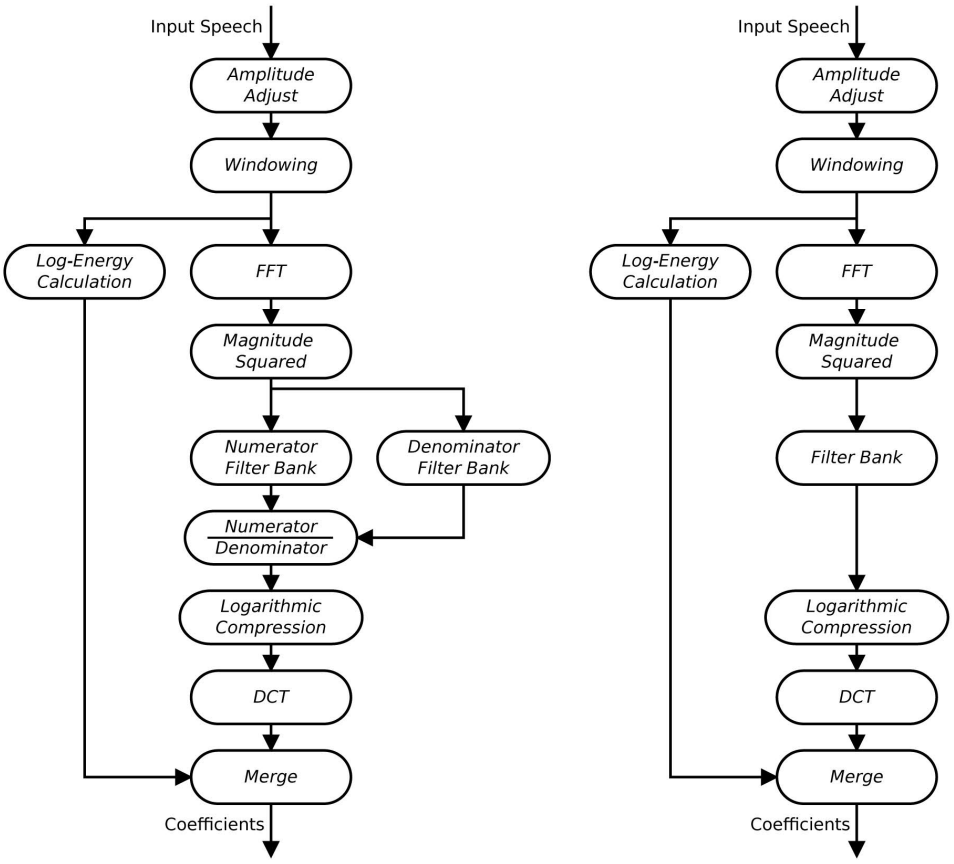

Figure 6: Flowcharts for LNCC (left) and MFCC (right) feature extraction: note the similarity between the two, with the only difference being the normalization of the filterbank outputs in LNCC . It is common to append delta and delta-delta co-efficients; this is not shown in the diagrams.

3.1.2. Robustness to channel mismatch

In Figure 10 we observe the response of the LNCC filterbank when speech is filtered by a channel with a non-flat frequency response, in this case, a spectral tilt of -6 dB/octave. The classical filterbank preserves the channel response in its output, whereas the normalized filterbank exhibits a response that is almost invariant to the channel response, while preserving key speech-related properties such as the spectral peaks.

In the experiments presented in this paper, we compare the proposed features with conventional MFCCs which are optionally normalized using Cepstral Mean Normalization (Atal, 1974; Furui, 1981; Wang et al., 2007). More advanced schemes such as RASTA filtering (Hermansky et al., 1991a,b; Hermansky, 1994) are of course available, but comparison (or indeed combination) with those is left as future work. It is important to note that all these other

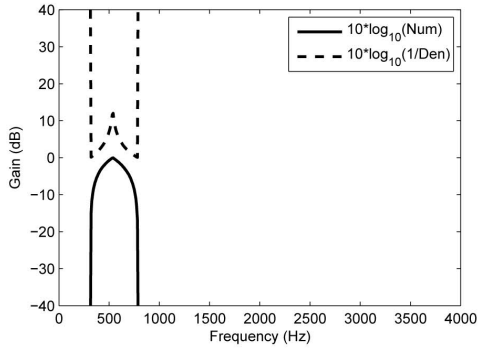

Figure 7: Frequency response of the numerator and denominator separately, both tuned at f_i^c = 515Hz, on a logarithmic scale.

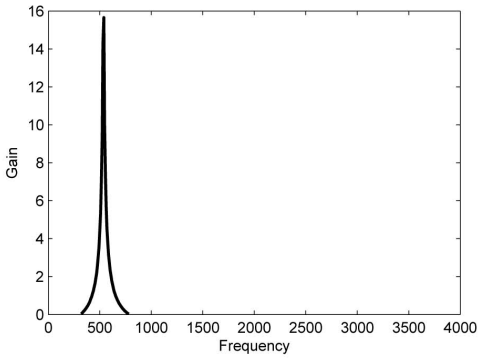

Figure 8: Frequency response of the numerator divided by the denominator, both tuned at f_i^c = 515Hz, on a logarithmic scale.

schemes require information outside the current frame being processed, and so are less effective for rapidly-varying channels (Leus & Moonen, 2003; Leus, 2004). For example: CMN requires an accurate estimate of the cepstral mean, which may be hard to obtain reliably in some cases (Qi Li et al., 2002); RASTA makes an equivalent assumption, that the channel changes substantially more slowly than the speech spectral envelope (Hermansky, 1994).

4. Speaker verification experiments

To investigate the ability of the proposed features to normalize for varying channels, we conducted a sequence of speaker verification experiments on speech degraded by various channels. These involve simulated channels imposing spectral tilt which mimics the effect of off-axis or occluded microphones (Section 4.5) as well as spectral tilt characteristics that vary within a utterance (Section 4.6). For reasons of experimental control and repeatability, channel responses were simulated. In all experiments, the system was trained using only clean speech. Test speech was degraded with respect to the training data by imposing static and time varying spectral tilt.

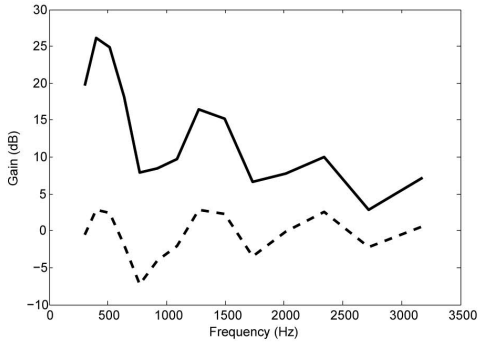

Figure 9: Spectral envelopes (i.e. filterbank outputs plotted immediately after the logarithmic compression step in Figure 6) for a single frame of voiced speech using a conventional Mel-scale filterbank (solid line), and for the proposed LNCC filterbank (dashed line). To aid readability, the solid line has been shifted by +15 dB. Observe that the proposed self-normalizing filterbank preserves important spectral shape information, such as the spectral peaks, but removes overall spectral tilt.

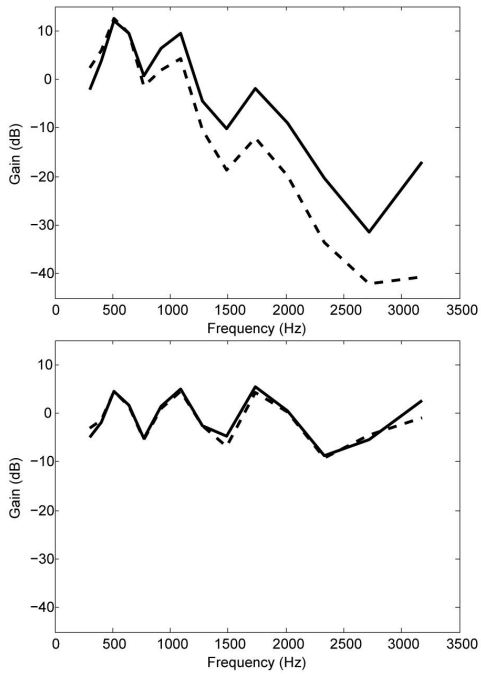

Figure 10: Spectral envelopes (i.e., filterbank outputs plotted just after the logarithmic compression steps in Figure 6) for for a single frame of voiced speech using a conventional Mel-scale filterbank (upper figure), and using the proposed LNCC filterbank (lower figure). The responses to unmodified speech are shown in solid lines and the responses to speech filtered through a channel that imposes a −6 dB/octave spectral tilt are shown in dashed lines. Observe that the proposed features are invariant to the channel's spectral tilt, whereas the conventional filterbank outputs are highly sensitive to it.

4.1. Speaker verification system

The experiments were carried out with a text-independent speaker verification system based Gaussian mixture models (GMMs) (Reynolds et al., 2000; Bimbot et al., 2004), with a universal background model (UBM) that was trained using background impostor speakers. A speaker-dependent GMM is generated for each speaker by employing MAP adaptation (Reynolds et al., 2000). By doing so, the correspondence of the Gaussians within each speaker-dependent GMM with those in the background GMM is preserved (Reynolds et al., 2000). Given a verification attempt where the identity of the speaker s is claimed, O denotes the observation sequence corresponding to the claimant's utterance. The output score of the system is a cohort-normalized log likelihood, $\log L(O)$:

$$\log L(O) = \log L(O/\lambda_s) - \overline{\log L(O/\lambda_{\bar{s}})} \tag{6}$$

where $\log L(O/\lambda_s)$ is the log likelihood of the client hypothesis and λ_s is the speaker s model, and $\overline{\log L(O/\lambda_{\bar{s}})}$ is the averaged log likelihood of the cohort of impostor models.

As described by Yoma & Villar (2002), frames with higher local segmental SNR provide more reliable information than those with low segmental SNR. Also, voiced sounds (*e.g.* vowels) show much higher speaker discrimination ability than fricative sounds. Accordingly, all the frames whose normalized energy with respect to the maximum utterance frame energy was lower than a given threshold are discarded.

4.2. Feature extraction

Features were extracted using LNCC and MFCC processing, as described by Figure 6. The frame length in all cases was 25 msec with a 50% overlap. A frequency range from 200 to 3860 Hz was covered by 14 triangular filters uniformly arranged on a Bark scale, in the case of MFCCs, or 28 pairs (unless otherwise noted) of numerator and denominator filters uniformly arranged on a Bark scale in the case of the proposed LNCC features. If an LNCC channel goes beyond the range 0Hz to Nyquist frequency, it is simply truncated. The DCT was truncated at 11 coefficients in both cases, then the first coefficient was replaced by the log frame energy. Finally, the resulting 11 coefficients are augmented with deltas and delta-delta to make up the final feature vector of dimension 33 for each frame.

4.3. Task

All experiments used the YOHO Speaker Verification Corpus, which comprises high quality recorded speech at 8kHz sampling rate (Campbell & Higgings, 1994). YOHO supports the development, training, and testing of speaker verification systems with a vocabulary comprising two-digit numbers spoken continuously in sets of three (*e.g.* "62-31-53" pronounced as "sixty-two thirty-one fifty-three").

The database is divided into enrollment and verification portions. Each of these contains data from 138 speakers. In our experiments a subset of 70 speakers was used. These speakers were divided as follows: 40 background impostor speakers to train the background models; 30 test client speaker for use in verification attempts. For each speaker, one

24-utterance enrolment session was used. False rejection curves were estimated with 30 speakers × 16 verification signals per client = 480 utterances. False acceptance curves were obtained with 30 speakers × 29 impostors × 6 verification signals per impostor = 5220 experiments.

4.4. Initial experiments: sensitivity to parameter settings

Preliminary experiments were performed to determine how sensitive the proposed features are to the various parameters which must be chosen: the bandwidth of the filters (all filters have the same bandwidth on a Bark scale), the number of channels (the number of filters also determines their spacing, as a bandwidth that is too narrow would leave a "gap" in the filterbank's overall response), and the parameter $d_{\min}$ which prevents division by zero at the centre frequency of each pair of numerator and denominator filters. As we described in Section 3.1.1, the LNCC filters exhibit a sharper response than the triangular filters in the MFCC filterbank. Therefore, we typically obtain best performance with a larger number of filters (*e.g.* , 28) than in the MFCC filterbank (which comprises 14 filters). All experiments regarding parameter sensitivity were performed with clean speech, and with speech processed through a channel with a −6-dB/octave spectrally-tilted frequency response.

4.4.1. Number of LNCC channels and filter bandwidth

As can be seen in Figure 11, performance on clean speech is relatively unaffected by the bandwidth until it becomes too narrow – this is presumably because at narrow bandwidths with a constant number of channels, "gaps" start to appear between filters and speech information is then missed. For spectrally-tilted speech, the same effect is seen with narrow bandwidths, but we also observe a worsening of performance at wider bandwidths. This is presumed to be a consequence of the local normalization becoming "less local" and therefore less effective.

Also in Figure 11 it is observed that 28 LNCC channels leads to lower EER than 14 filters. Although not presented in this paper, further experiments were carried out with 20 and 56 LNCC channels. However, those configurations did not lead to significant improvements in equal error rates when compared to 28 channels. It is worth emphasizing that the optimal bandwidth with -6 dB/octave spectral tilt is shifted from B equal to 4 Barks to 3 Barks when the number of LNCC channels is increased from 14 to 28. This result must be due to the fact that the higher number of channels, the lower the "gap" between the filters tends to be.

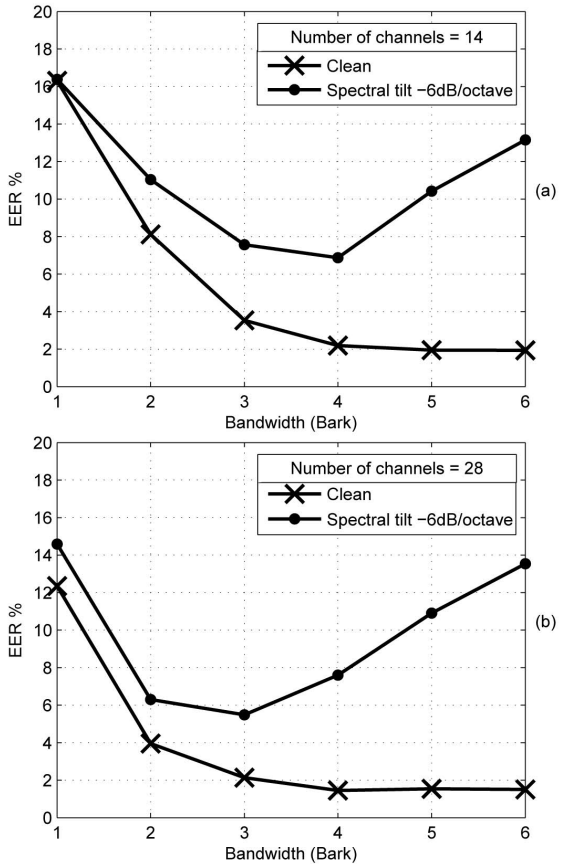

Figure 11: Sensitivity to the filter bandwidth. Both with d_{min}=0.001. (a) 14 channels, (b) 28 channels.

4.4.2. Denominator Minimum Centre Value (d_{min})

Figure 12 describes EER as a function of $d_{\min}$ for clean speech and speech corrupted by -6 dB/octave spectral tilt. The LNCC coefficients were computed using 28 channels, and a bandwidth B=3 Barks. According to Figure 12 there is a wide range of values for $d_{\min}$ ($0 \leq d_{\min} \leq 0.01$) for which EER shows little variation.

4.5. Experiment 1: simulated distant microphone

4.5.1. Speech processing to simulate the frequency response of a distant microphone

As mentioned in Section 1.2.2 one of the consequences of using a distant, off-axis, or occluded microphone or microphone array to capture speech is that some unknown spectral shaping will be imposed on the speech by the channel. Speech produced with increased vocal effort may also vary the spectral tilt with respect to clean speech.

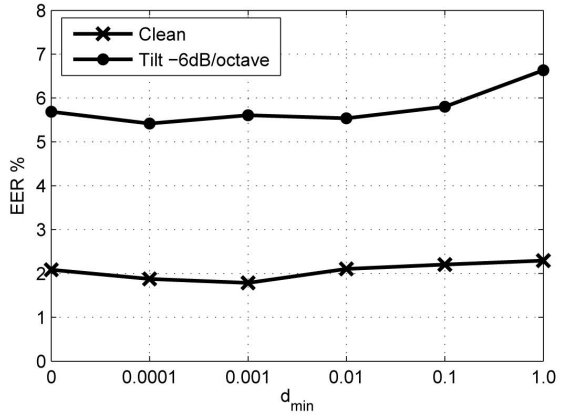

Figure 12: Sensitivity to the $d_{\min}$ parameter. LNCC with 28 channels, B=3 Barks.

Overall the effect is one where the test speech has a different average spectral shape to the clean training speech. We simulated this using a simple filter that imposes −3 dB/octave or −6 dB/octave spectral tilt; these particular values were arrived at through informal experiments in which we re-recorded speech reproduced over a loudspeaker, with the microphone set off-axis, or occlusions place between loudspeaker and microphone.

4.5.2. Results

Figure 13 compares results with MFCC, MFCC+CMN, LNCC and LNCC+CMN. The proposed LNCC features leads to a lower error rate than MFCC+CMN and slightly worse than MFCC with clean speech. When the speech was degraded with a -3 dB/octave spectral tilt, LNCC provides the lowest EER, which is 24% ($p < 0.01$) lower than the one achieved with MFCC+CMN. With -6 dB/octave both MFCC+CMN and LNCC dramatically compensate for this distortion and give reductions in EER as high as 87% and 79%, respectively. On the other hand, MFCC+CMN gives an EER 2.1% lower (absolute) than LNCC. However, LNCC is memoryless, and does not require computing nor storing the moving average needed by CMN. Observe that LNCC+CMN does not show any improvement with respect to LNCC. This result suggests that CMN does not help LNCC in compensating for spectral tilt and introduces a distortion due to the statistical estimation of the feature means.

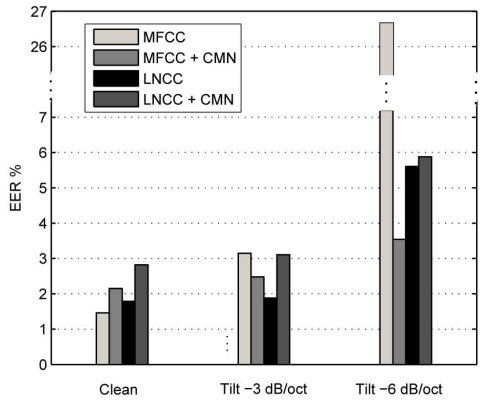

Figure 13: Performance for constant spectral tilt. LNCC features are computed using 28 channels, $d_{\min}$=0.001 and B= 3 Barks.

4.6. Experiment 2: rapidly-varying channels

4.6.1. Speech processing to simulate a time-varying channel response

A dynamic filter was designed in order modify the spectral tilt over time. This is an FFT filter applied on a frame-by-frame basis. The slope of the target tilt linearly varies between 0 dB/octave and −6 dB/octave within each utterance. Then, the energy of each filter step is normalized to compensate for the attenuation produced by the spectral tilt frame-by-frame. To ensure the effectiveness of the filter, this is applied between estimates of starting and ending

points of speech. Three time-varying spectral tilts were applied: $VaryingST_1$, the spectral tilt changes linearly with time from 0 dB/octave to -6 dB/octave from the beginning to the end of the utterance; $VaryingST_2$, the spectral tilt changes linearly with time, at an absolute constant rate, from 0 dB/octave to -6 dB/octave, and then from -6 dB/octave to 0 dB/octave from the beginning to the end of the utterance; and $VaryingST_3$, the spectral tilt changes linearly with time, at an absolute constant rate, from 0 dB/octave to -6 dB/octave, then from -6 dB/octave to 0 dB/octave and, finally, from from 0 dB/octave to -6 dB/octave from the beginning to the end of the utterance.

4.6.2. Results

As can be seen in Figure 14, LNCC provides lower EER than the other features considered, with mean EER values of 2.65%, 2.79%, 1.93%, 3.33% for MFCC. MFCC+CMN, LNCC, LNCC+CMN, respectively. We also note that LNCC provides reductions in EER equal to 12.1% ($p < 0.271$), 41.2% ($p < 1\times10^{-7}$), 42.2% ($p < 3.5\times10^{-7}$) for the conditions $VaryingST_1$, $VaryingST_2$, $VaryingST_3$, respectively, when compared with the MFCC+CMN, as well as provides reductions in EER equal to 30.4% ($p < 1.9\times10^{-4}$), 32.2% ($p < 7\times10^{-5}$), 42.2% ($p < 1.8\times10^{-8}$) when compared with standard MFCC features. In addition, CMN does not improve MFCC in most of these conditions, which shows a lack of robustness on the part of CMN in response to time-varying spectral tilts. This result is a reflection of the fact that CMN coefficients are estimated over a time interval during which the statistics of the signal change and the feature means are not reliably computed. It is worthwhile to reemphasize that LNCC coefficients provide more consistent performance in EER for clean speech and spectrally-varying conditions, which can be observed from the standard deviations 0.796, 0.660, 0.151, 0.486 provided by MFCC, MFCC+CMN, LNCC and LNCC+CMN, respectively.

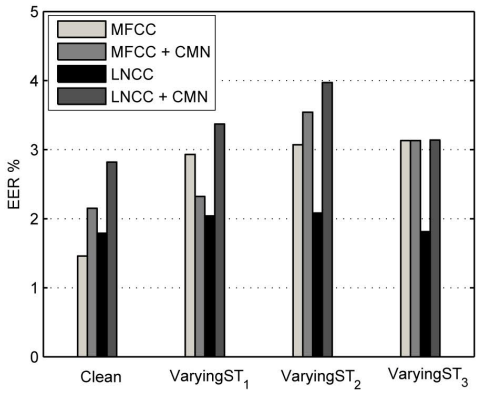

Figure 14: Performance for time varying spectral tilt. $VaryingST_1$: 0 dB/oct $\rightarrow$ -6 dB/oct. $VaryingST_2$: 0 dB/oct $\rightarrow$ -6 dB/oct $\rightarrow$ 0 dB/oct. $VaryingST_3$: 0 dB/oct $\rightarrow$ -6 dB/oct $\rightarrow$ 0 dB/oct $\rightarrow$ -6 dB/oct. LNCC features are computed using 28 channels, $d_{\min}$=0.001 and B= 3 Barks.

4.7. Summary of results

Across all experiments we observe that the proposed features are competitive with either MFCC or MFCC+CMN features. We also note that the best choice of whether to use CMN with MFCC features depends on environmental conditions, while the proposed LNCC features provide consistently good performance across all conditions and never suffer from extremely high error rates, which are observed in some cases when MFCCs are used.

	Equal Error Rate (EER) %			
	Baselines		Proposed	
Test data	MFCC	MFCC+CMN	LNCC	LNCC+CMN
Clean	1.46	2.15	1.79	2.82
Spectral tilt −3 dB/octave	3.15	2.48	1.88	3.11
Spectral tilt −6 dB/octave	26.7	3.54	5.61	5.88
Varying tilt 0 to −6 dB/octave	2.93	2.32	2.04	3.37
Varying tilt 0 to -6 to 0 dB/octave	3.07	3.54	2.08	3.97
Varying tilt 0 to -6 to 0 to −6 dB/octave	3.13	3.13	1.81	3.14

Table 1: Summary of results. LNCC features are computed using 28 channels, $d_{\min}$=0.001 and B=3 Barks.

5. Conclusions

In this paper a perceptually-motivated and extremely simple, but nevertheless effective, way to *instantaneously* normalize speech features is proposed. The effectiveness of the proposed features is demonstrated for a speaker verification task across a variety of channel conditions. The Locally-Normalized Cepstral Coefficients do not require the computation and storage of a moving average of the feature values, and they provide reductions in EER as high as 32% and 35% when compared with MFCC and MFCC+CMN with variable spectral tilt, respectively. With a static -6 dB/octave spectral tilt the proposed locally-normalized coefficients give a dramatic reduction in EER as high as 79% when compared with the ordinary MFCC. Consequently, the proposed LNCC features are an attractive alternative to MFCC and MFCC+CMN in any situation where it is difficult to estimate the cepstral means accurately. Other applications might include scenarios where very low latency or low complexity is desired, where computing and storing the moving average required by CMN may become inconvenient. As future work we plan to evaluate the proposed features for an automatic speech recognition (ASR) task, although it is possible that the self-normalizing filterbank may remove a small amount of phonetic information along with the channel information, so some modifications may be necessary to limit the amount of normalization that is performed. In that direction, an obvious line of investigation would be to combine LNCC with MFCCs or PLPs using either feature combination or system combination.

Acknowledgements

The research leading to these results was funded by CONICYT-ANILLO project ACT 1120 and CONICYT-FONDEYT project 1100195. S. King was partly funded by EPSRC grant EP/I031022/1 (Natural Speech Technology). Richard Stern was partially funded by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20024. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch.

References

Ali, A. M., Van Der Spiegel, J., & Mueller, P. 2000. Auditory-based speech processing based on the average localized synchrony detection. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Instanbul, pp. 1623–1626.

Ali, A. M., Van Der Spiegel, J., & Mueller, P. 2002. Robust auditory-based speech processing using the average localized synchrony detection. IEEE Transactions on Speech and Audio Processing, 10, 279–292.

Anderson, S., Skoe, E., Chandrasekaran, B., & Kraus, N. 2010. Neural timing is linked to speech perception in noise. Journal of Neuroscience, 30, 4922–4926.

Atal, B. 1974. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. Journal of the Acoustical Society of America, 55, 1304–1312.

Bell, P., Yumamoto, H., Swietojanski, P., Wu, Y., McInnes, F., Hori, C., & Renals, S. 2013. A lecture transcription system combining neural network acoustic and language models. In Proceedings of Interspeech 2013, Lyon, pp. 3087–3091.

Bimbot, F., Bonastre, J. F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I. M., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., & Reynolds, D. A. 2004. A tutorial on text-independent speaker verification. EURASIP Journal on Applied Signal Processing, 52, 430–451.

Bořil, H., & Hansen, J. H. L. 2010. Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments. IEEE Transactions on Audio, Speech, and Language Processing, 18, 1379–1393.

Brandstein, M., & Ward, D. 2010. Microphone Arrays: Signal Processing Techniques and Applications. Digital Signal Processing. Springer.

Buchner, H., Benesty, J., & Kellermann, W. 2005. Generalized multichannel frequency-domain adaptive filtering: efficient realization and application to hands-free speech communication. Signal Processing, 85, 549–570.

Campbell, J. P. 1997. Speaker recognition: A tutorial. Proceedings of the IEEE, 85, 1437–1462.

Campbell, J. P., & Higgings, A. 1994. YOHO Speaker Verification. Linguistic Data Consortium, Philadelphia, PA.

Chigier, B., & Leung, H. C. 1992. The effects of signal representations, phonetic classification techniques, and the telephone network. In Proceedings of the Second International Conference on Spoken Language Processing, Banff, Alberta, pp. 97–100.

Chiu, Y.-H., & Stern, R. M. 2008. Analysis of physiologically-motivated signal processing for robust speech recognition. In Proceedings of Interspeech 2008, Brisbane, pp. 1000–1003.

Cooke, M., King, S., Garnier, M., & Aubanel, V. 2014. The listener talker: A review of human and algorithmic context-induced modifications of speech. Computer Speech and Language, 28, 543–571.

Cooke, M., & Lecumberri, M. L. 2012. The intelligibility of lombard speech for non-native listeners. Journal of the Acoustical Society of America, 132, 1120–1129.

Cooke, M., Mayo, C., & Valentini-Botinhao, C. 2013a. Intelligibility-enhancing speech modifications: the Hurricane Challenge. In Proceedings of Interspeech 2013, Lyon, pp. 3552–3556.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Sauert, B., & Tang, Y. 2013b. Evaluating the intelligibility benefit of speech modifications in known noise conditions. Speech Communication, 55, 572–585.

Davis, S., & Mermelstein, P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, 28, 357–366.

Delgutte, B., & Kiang, N. Y. S. 1984. Speech coding in the auditory nerve: I. vowels-like sounds. Journal of the Acoustical Society of America, 75, 866–876.

Dimitriadis, D., Maragos, P., & Potamianos, A. 2011. On the effects of filterbank design and energy computation on robust speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 19, 1504–1516.

Dreyer, A., & Delgutte, B. 2006. Phase locking of auditory-nerve fibers to the envelopes of high frequency sounds: Implications for sound localization. Journal of Neurophysiology, 96, 2327–2341.

Eggermont, J. J. 1998. Is there a neural code? Neuroscience and Biobehavioral Reviews, 22, 355–370.

Furui, S. 1981. Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech, and Signal Processing, 29, 254–272.

Gales, M. J. F. 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language, 12, 75–98.

Gaubitch, N. D., Brookes, M., & Naylor, P. A. 2013. Blind channel magnitude response estimation in speech using spectrum classification. IEEE Transactions on Audio, Speech, and Language Processing, 21, 2162–2171.

Ghitza, O. 1994. Auditory models and human performance in tasks related to speech coding and speech recognition. IEEE Transactions on Speech and Audio Processing, 2, 115–132.

Glass, J., Hazen, T., Cyphers, S., Malioutov, I., Huynh, D., & Barzilay, R. 2007. Recent progress in the MIT spoken lecture processing project. In Proceedings of Interspeech 2007, Antwerp, pp. 2553–2556.

Hain, T., Burget, L., Dines, J., Garau, G., Karafit, M., Lincoln, M., & Wan, V. 2006. The AMI meeting transcription system. In Proceedings of the NIST Rich Transcription 2006 Spring Meeting Recognition Evaluation Workshop.

Hain, T., Burget, L., Dines, J., Garau, G., Wan, V., Karafiat, M., Vepa, J., & Lincoln, M. 2007. The AMI system for the transcription of speech in meetings. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2007, Honolulu, pp. 357–360.

Hain, T., Burget, L., Dines, J., Garner, P., Grezl, F., Hannani, A., Huijbregts, M., Karafiat, M., Lincoln, M., & Wan, V. 2012. Transcribing meetings with the AMIDA system. IEEE Transactions on Audio, Speech, and Language Processing, 20, 486–498.

Hansen, J. H. L., & Varadarajan, V. 2009. Analysis and compensation of Lombard speech acroos noise type and levels with application to In-Set/Out-of-Set speaker recognition. IEEE Transactions on Audio, Speech, and Language Processing, 17, 366–378.

Heinz, M. G., & Swaminathan, J. 2009. Quantifying envelope and fine-structure coding in auditory-nerve responses to chimaeric speech. Journal of the Association for Research in Otolaryngology, 10, 407–423.

Hermansky, H. 1990. Perceptual linear predictive PLP analysis of speech. Journal of the Acoustical Society of America, 87, 1738–1752.

Hermansky, H. 1994. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing, 2, 578–589.

Hermansky, H., Cohen, J. R., & Stern, R. M. 2013. Perceptual properties of current speech recognition technology. Proceedings of IEEE, 101, 1968–1985.

Hermansky, H., Morgan, N., Bayya, A., & Khon, P. 1991a. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In Proceedings of Eurospeech, Genova, pp. 1367–1370.

Hermansky, H., Morgan, N., Bayya, A., & Khon, P. 1991b. (RASTA-PLP) speech analysis technique. In Proceedings International Conference on Acoustics, Speech, and Signal Processing, San Francisco, pp. 121–124.

Hori, T., Fujimoto, M., Ogawa, A., Kinoshita, K., & Nakamura, A. 2012. Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera. IEEE Transactions on Audio, Speech, and Language Processing, 20, 499–513.

Houtgast, T. 1972. Psychophysical evidence for lateral inhibition in hearing. Journal of the Acoustical Society of Amererica, 51, 1885–1894.

Hsu, B. J., & Glass, J. 2006. Style and topic language model adaptation using HMM-LDA. In Proceedings of the Conference on Empirical Methods in Natural Processing, Sydney, pp. 373–381.

Hsu, C. W., & Lee, L. S. 2009. Higher order cepstral moment normalization for improved robust speech recognition. IEEE Transactions on Audio,

Speech and Language Processing, 17, 205–220.

Ishi, C. T., Matsuda, S., Kanda, T., Jitsuhiro, T., H., I., S., N., & Hagita, N. 2008. A robust speech recognition system for communication robots in noisy environments. IEEE Transactions on Robotics, 24, 759–763.

Jankowski, C. R., & Lippmann, R. P. 1992. Comparison of auditory models for robust speech recognition. In Proceedings of the DARPA Speech and Natural Language Workshop, New York, pp. 453–454.

Jankowski, C. R., Vo, H. D., & Lippmann, R. P. 1995. A comparison of signal processing front ends for automatic word recognition. IEEE Transactions on Speech and Audio Processing, 3, 286–293.

Jensen, B., Tomatis, N., Drygajlo, A., & Siegwart, R. 2005. Robots meet human interaction in public spaces. IEEE Transactions on Industrial Electronics, 52, 1530–1546.

Johnson, D. 1980. The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. Journal of the Acoustical Society of America, 68, 1115–1122.

Joris, P., & Yin, T. 2007. A matter of time: Internal delays in binaural processing. Trends in Neuroscience, 30, 70–78.

Kayser, C., Montemurro, M. A., Logothetis, N. K., & Panzeri, S. 2009. Spike-phase coding boost and stabilizes information carried by spatial and temporal spike patterns. Neuron, 61, 597–608.

Kiang, N. Y. S., Watanabe, T., Thomas, E. C., & Clark, L. F. 1965. Discharge Patterns of Single Fibers in the Cat's Auditory Nerve. Cambridge, MA: MIT Press.

Kim, C., Chiu, Y. B., & Stern, R. M. 2006. Physiologically-motivated synchrony-based processing for robust automatic speech recognition. In Proceedings of Interspeech 2006, Pittsburgh, pp. 1483–1486.

Kim, D. S., Lee, S. Y., & Kil, R. M. 1999. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. IEEE Transactions on Speech and Audio Processing, 7, 55–69.

Kumaresan, R., & Rao, A. 1999. Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications. Journal of the Acoustical Society of America, 105, 1912–1924.

Kuwabara, H., & Sagisaka, Y. 1995. Acoustics characteristics of speaker individuality: control and conversion. Speech Communication, 16, 165–173.

Leeuwis, E., Federico, M., & Cettolo, M. 2003. Language modeling and transcription of the TED corpus lecture. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006, Toulouse, pp. 232–235.

Leus, G. 2004. On the estimation of rapidly time-varying channels. In Proceedings of European Signal Processing Conference, Vienna, pp. 2227–2230.

Leus, G., & Moonen, M. 2003. Deterministic subspace based blind channel estimation for doubly-selective channels. In Proceedings of the 4th IEEE Workshop on Signal Processing Advances in Wireless Communications pp. 210–214.

Liberman, M. C. 1978. Auditory nerve response from cats raised in a low noise chamber. Journal of the Acoustical Society of America, 63, 442–455.

Liu, F., Stern, R. M., Huang, X., & Acero, A. 1993. Efficient cepstral normalization for robust speech recognition. In Proceedings DARPA Speech and Natural Language Workshop, Cambridge, pp. 69–74.

Lu, X., Unoki, M., & Nakamura, S. 2011. Sub-band temporal modulation envelopes and their normalization for automatic speech recognition in reverberant environments. Computer Speech and Language, 25, 571–584.

Malionek, J., Oard, D. W., Sangwan, A., & Hansen, J. H. L. 2013. Linking transcribed conversational speech. In Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, pp. 961–964.

Meyer, B., & Kollmeier, B. 2011. Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. Speech Communication, 53, 753–767.

Miettinen, I., Alku, P., Salminen, N., May, P. J. C., & Tiitinen, H. 2011. Responsiveness of the human auditory cortex to degraded speech sounds: Reduction of amplitude resolution vs. additive noise. Brain Research, 1367, 298–309.

Moore, B. C. J. 2003. An Introduction to the Psychology of Hearing. Academic Press, Elsevier Science.

Moore, B. C. J. 2008. The rol of temporal fine structure processing in pitch perception, masking, and speech perception for normal hearing and hearing-impaired people. Journal of the Association for Research in Otolaryngology, 9, 399–406.

Moore, B. C. J. 2014. Auditory Processing of Temporal Fine Structure: Effects of Age and Hearing Loss. Audiology and Otology. World Scientific Publishing CO PTE LTD.

Morales, N., Toledano, D., Hansen, J. H. L., & Garrido, J. 2009. Feature compensation techniques for ASR on band-limited speech. IEEE Transactions on Audio, Speech and Language Processing, 17, 758–774.

Nakano, A. Y., Nakagawa, S., & Yamamoto, K. 2010. Distant speech recognition using a microphone array network. IEICE Transactions on Information and Systems, E93.D, 2451–2462.

Ohshima, Y., & Stern, R. M. 1994. Environmental robustness in automatic speech recognition using physiologically-motivated signal processing. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994, Adelaide, pp. 1–4.

Parikh, G., & Loizou, P. C. 2005. The influence of noise of vowel and consonant cues. Journal of the Acoustical Society of America, 118, 3874–3888.

Park, A., Hazen, T., & Glass, J. 2005. Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005, Philadelphia, pp. 497–500.

Pickles, J. O. 2008. An Introduction to the Physiology of Hearing. Emerald Group Publishing Limited.

Qi Li, P., Zheng, J., Tsai, A., & Zhou, Q. 2002. Robust end-point detection and energy normalization for real-time speech and speaker recognition. IEEE Transactions on Speech and Audio Processing, pp. 146–157.

Qin, L., Wang, J. Y., & Sato, Y. 2008. Representations of cat meows and human vowels in the primary auditory cortex of awake cats. Journal of Neurophysiology, 99, 2305–2319.

Renals, S., Hain, T., & Bourlard, H. 2007. Recognition and understanding of meetings: The AMI and AMIDA projects. In Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU) 2007, Kyoto, pp. 238–247.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. 2000. Speaker verification using adapted Gaussian Mixture Models. Digital Signal Processing, 10, 19–41.

Reynolds, D. A., & Rose, R. C. 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing, 3, 72–83.

Rose, R. C., & Reynolds, D. A. 1990. Text-independent speaker identification using automatic acoustic segmentation. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1990, Albuquerque pp. 293–296.

Rosen, S. 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. Philosophical Transactions of the Royal Society B, 336, 367–373.

Sachs, M. B. 1984. Neural coding of complex sounds: Speech. Annual Review of Physiology, 46, 261–273.

Sachs, M. B., & Kiang, N. Y.-S. 1968. Two-tone inhibition in auditory-nerve fibers. Journal of the Acoustical Society of America, 43, 1120–1128.

Sachs, M. B., & Young, E. D. 1979. Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. Journal of the Acoustical Society of America, 66, 470–479.

Sangwan, A., Kaushik, L., Yu, C., Hansen, J. H. L., & Oard, D. W. 2013. Houston, we have a solution: using NASA Apollo Program to advance speech and language procesing technology. In Proceedings of Interspeech 2013, Lyon pp. 1135–1139.

Schwartz, R., Anastasakos, T., Kubala, F., Makhoul, J., Nguyen, L., & Zavaliagkos, G. 1993. Comparative experiments on large vocabulary speech recognition. In Proceedings of the Workshop on Human Language Technology, Princeton, pp. 75–80.

Seltzer, M. L., Raj., B., & Stern, R. M. 2004. Likelihood-maximizing beamforming for robust hands-free speech recognition. IEEE Transactions on Speech and Audio Processing, 12, 489–498.

Seneff, S. 1984. Pitch and spectral estimation of speech based on an auditory synchrony model. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1984, San Diego, pp. 1–4.

Seneff, S. 1985. Pitch and spectral analysis of speech based on an auditory synchrony model. PhD. Dissertation, Massachusetts Institute of Technology, Cambridge.

Seneff, S. 1986a. Characterizing formants through straight line approximations without explicit formant tracking. In Proceedings of the First Montreal Symposium on Speech Recognition, Montreal, pp. 21–27.

Seneff, S. 1986b. A computational model for the peripheral auditory system: application to speech recognition research. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1986, Tokyo, pp. 1983–1986.

Seneff, S. 1987. Vowel recognition based on line-formants derived from an auditory-based spectral. In Proceedings of the 11th International Congress of Phonetic Sciences, Tallin,.

Seneff, S. 1988. A joint synchrony/mean-rate model of auditory speech processing. Journal of Phonetics, 16, 55–76.

Shao, Y., Srinivasan, S., Jin, Z., & Wang, D. 2010. A computational auditory scene analysis system for speech segregation and robust speech recognition. Computer Speech and Language, 24, 77–93.

Sinex, D. G., & Geisler, D. 1983. Responses of primary auditory fibers t6o consonant-vowel syllables. Journal of the Acoustical Society of America, 73, 602–615.

Smith, Z. M., Delgutte, B., & Oxenham, A. J. 2002. Chimaeric sounds reveal dichotomies in auditory perception. Nature, 416, 87–90.

Soong, F. K., & Rosenberg, A. E. 1988. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Transactions on Acoustics, Speech and Signal Processing, 36, 871–879.

Stern, R. M., & Morgan, N. 2012a. Features based on auditory physiology and perception. In Techniques for Noise Robustness in Automatic Speech Recognition, Virtanen, T., Raj, B. and Singh, R., Eds., New York, NY, USA: Wiley, pp. 207–243.

Stern, R. M., & Morgan, N. 2012b. Hearing is believing: Biologically inspired methods for robust automatic speech recognition. Signal Processing Magazine IEEE, pp. 34–43.

Stern, R. M., Wang, D., & Brown, G. J. 2006. Binaural sound localization. In D. Wang, & G. J. Brown (Eds.), Computational Auditory Scene Analysis chapter 5. Wiley-IEEE Press.

Stockham, T. G., Cannon, T. N., & Ingebretsen, R. B. 1975. Blind deconvolution through digital signal processing. Proceedings of the IEEE, 63, 678–693.

Tchorz, J., Kleinschmidt, M., & Kollmeier, B. 1996. A psychoacoustical model of the auditory periphery as a front end for ASR. Journal of the Acoustical Society of America, 105, 1157–1157.

Tchorz, J., & Kollmeier, B. 1999. A model of auditory perception as front end for automatic speech recognition. Journal of the Acoustical Society of America, 106, 2040–2050.

Togneri, R., & Pullella, D. 2011. An overview of speaker identification: Accuracy and robustness issues. IEEE Circuits and Systems Magazine, 11, 23–61.

Tokuda, K., Yoshimura, T., Kobayashi, T., & Kitamura, T. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2000, Instanbul, pp. 1315–1318.

Trancoso, I., Nunes, R., & Neves, L. 2006. Classroom lecture recognition. Computational Processing of the Portuguese Language, Proceedings Book Series: Lecture Notes in Artificial Intelligence, 3960, 190–199.

Tranter, S., & Reynolds, D. A. 2006. An overview of automatic speaker diarization systems. IEEE Transactions on Audio, Speech and Language Processing, 14, 1557–1565.

Wang, K., & Shamma, S. A. 1994. Self-normalization and noise-robustness in early auditory representations. IEEE Trans. on Speech and Audio Processing, 2, 421–435.

Wang, L., Kitaoka, N., & Nakagawa, S. 2007. Robust distant speech recognition by combining position-dependent CMN with conventional CMN. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2007, Honolulu, pp. 817–820.

Wang, L., Kitaoka, N., & Nakagawa, S. 2011. Distant-talking speech recognition based on spectral subtraction by multi-channel LMS algorithm. IEICE Transactions on Information and Systems, E.94.D, 659–667.

Watkins, A. J., & Makin, S. J. 1996. Some effects of filtered contexts on the perception of vowels and fricatives. Journal of the Acoustical Society of America, 99, 588–594.

Werblin, F. S., Jacobs, A., & Teeters, J. 1996. The computational eye. IEEE Spectrum, 33, 30–37.

Wölfel, M. 2009a. Enhanced speech features by single-channel joint compensation of noise and reverberation. IEEE Transactions on Audio, Speech and Language Processing, 17, 312–323.

Wölfel, M. 2009b. Signal adaptive spectral envelope estimation for robust speech recognition. Speech Communication, 51, 551–561.

Wölfel, M., & McDonough, J. 2009. Distant Speech Recognition. Chichester, UK: Wiley.

Yokoyama, R., Nasu, Y., Iwano, K., & Shinoda, K. 2013. Detection of overlapped speech using lapel microphones in meeting. Speech Communication, 55, 941–949.

Yoma, N. B., & Villar, M. 2002. Speaker verification in noise using a stochastic version of the weighted viterbi algorithm. IEEE Transactions on Speech and Audio Processing, 10, 158–166.

Young, E. D. 2008. Neural representation of speech spectral and temporal information in speech. Philosophical Transactions of the Royal Society B, 363, 923–945.

Young, E. D., & Sachs, M. B. 1979. Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers. Journal of the Acoustical Society of America, 66, 1381–1403.

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

Zwiker, E. 1961. Subdivision of the audible frequency range into critical bands (frequenzgrupenn). Journal of the Acoustical Society of America, 33, 248–249.