Comparación de diversas parametrizaciones para reconocimiento de habla robusto en entorno telefónico

Ascensión Gallardo Antolín, Javier Macías Guarasa, Rubén San Segundo, Javier Ferreiros López, Ricardo de Córdoba y José Manuel Pardo Muñoz*

Grupo de Tecnología del Habla. Departamento de Ingeniería Electrónica.
Universidad Politécnica de Madrid
*Departamento de Teoría de la Señal y Comunicaciones.
Universidad Carlos III de Madrid
gallardo@tsc.uc3m.es, {macias, lapiz, jfl, cordoba, pardo}@die.upm.es

Resumen

En este artículo, investigamos el funcionamiento de distintos tipos de rasgos acústicos en un sistema de reconocimiento automático de habla (SRAH) en entorno telefónico. En concreto, exploramos dos alternativas distintas para el diseño del módulo parametrizador. En la primera de ellas, las características de dicho módulo son elegidas de forma empírica o basándose en conocimiento psicoacústico. En la segunda, dichas características son determinadas mediante la extracción discriminativa de rasgos que permiten una optimización conjunta del parametrizador y clasificador. Ambas estrategias han sido aplicadas a parametrizadores basados en la transformada ondicular dando lugar a mejoras significativas en la tasa de reconocimiento del sistema en comparación con las parametrizaciones convencionales basadas en la transformada de Fourier.

1. Introducción

Los SRAH convencionales constan básicamente de dos módulos: el módulo de parametrización y el módulo de clasificación o reconocedor propiamente dicho.

La principal propiedad que debe poseer el módulo de parametrización es, por una parte, la capacidad de extraer un conjunto limitado de coeficientes que representen de la manera mejor posible los aspectos relevantes del espectro a corto plazo de la señal de voz y por otra parte, la de conseguir que dichos parámetros sean lo suficientemente discriminativos como para poder distinguir la señal de voz a la que representan de otras fuentes de distorsión como el ruido aditivo y convolutivo.

Las propiedades de "relevancia" y "capacidad discriminativa" que se exige a los parámetros de la señal de voz están estrechamente ligados a las características, estructura y complejidad del módulo clasificador posterior. Por ejemplo, de todos es conocido que las técnicas de realce de la señal de voz ("speech enhancement"), cuyo objetivo es la eliminación de ruido, pueden resultar satisfactorias si el clasificador posterior es humano (puesto que reducen significativamente la presencia de ruido de forma audible) y sin embargo, pueden producir resultados no tan satisfactorios si el clasificador posterior es un sistema de reconocimiento automático (puesto que introducen distorsiones que no son molestas al oído humano, pero que sí son "perceptibles" por el reconocedor). En cualquier caso, es evidente que el funcionamiento del reconocedor global se verá muy influido por la calidad de dichos parámetros. De aquí, que la investigación sobre parametrizaciones más robustas que las convencionales a la presencia de diferentes tipos de distorsión es de gran vigencia en la actualidad. Este artículo es precisamente, un estudio comparativo de la robustez al ruido convolutivo (canal telefónico) de distintas parametrizaciones convencionales y otras obtenidas mediante la extracción discriminativa de rasgos.

La organización del artículo es la siguiente. En la sección 2, haremos una breve descripción de la representación paramétrica de la señal de voz. En la sección 3 describiremos las diferentes parametrizaciones basadas en métodos empíricos utilizadas en la experimentación. En la sección 4 mostraremos una alternativa a dichas parametrizaciones basadas en la optimización conjunta del parametrizador basado en la transformada ondicular y el clasificador. A continuación, mostraremos los resultados más significativos. Finalmente, en la sección 6 se enumerarán las principales conclusiones y líneas futuras de trabajo.

2. Representación paramétrica

El proceso básico de extracción de los rasgos de la señal de voz consta básicamente de dos subprocesos. El primero de ellos consiste en la estimación del espectro a corto plazo de la señal. En el segundo, se aplica algún tipo de transformación lineal sobre los parámetros espectrales calculados previamente con el objetivo de proporcionar un grado alto de decorrelación entre las diversas componentes frecuenciales.

Para la primera de estas etapas se han propuesto en la literatura diversas técnicas, que podemos clasificar en:

- Técnicas basadas en el espectro de predicción lineal: los coeficientes de predicción lineal (LPC), los de reflexión (RC) y los LPC-cepstrum (LPCC) son ejemplos de este grupo. Una descripción exhaustiva de estas técnicas puede encontrarse en [1].
- Técnicas basadas en la transformada de Fourier: p. ej., los coeficientes mel-cepstrum (MFCC) [2] y las log-energías en bandas filtradas (FFLFBE) [3].
- Técnicas basadas en la transformada ondicular: p. ej. [4], [5] y [6].

En cuanto a la segunda fase del proceso referente a la decorrelación de los parámetros espectrales, también han sido propuestos diversos métodos entre los que destacamos:

- Aplicación de la transformada coseno discreta (DCT): los MFCC son el ejemplo más característico [2].
- Aplicación de filtrado sobre las componentes espectrales: los FFLFBE [3] son un ejemplo.
- Aplicación de la transformada ondicular: p. ej. [6].

Tras estas dos fases básicas, los coeficientes pueden ser mejorados aplicando un postproceso de filtrado de las trayectorias temporales como en las técnicas CMN [7], RASTA [8] o filtrados más complejos [9]. También, cabe la posibilidad de utilizar dichos parámetros filtrados como complemento de los originales, como es el caso de los clásicos parámetros dinámicos (primera derivada o deltas).

A continuación, describiremos las combinaciones de estas técnicas que hemos considerado y que están representadas en la Figura 1.

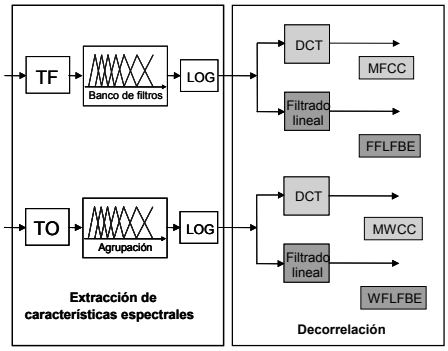

Figura 1: Resumen de parametrizaciones utilizadas

3. Parametrizaciones empíricas y/o basadas en conocimientos psicoacústicos

En la mayoría de los casos, el parametrizador se diseña haciendo uso de los conocimientos científicos y/o heurísticos disponibles de la señal de habla. Esta estrategia es la que hemos denominado parametrizaciones empíricas. Por ejemplo, el número de bandas críticas y distribución de las mismas es un claro ejemplo de parámetro de configuración cuya elección se realiza de esta forma.

3.1. Parámetros basados en la transformada de Fourier

En este caso, las características espectrales de la señal de voz se derivan del análisis de Fourier enventanado o a corto plazo (STFT) definido por la siguiente expresión:

$$S_x(t,f)=\int_{-\infty}^{\infty} x(\tau+t)\,w(\tau)\,e^{-j2\pi f\tau}\,d\tau \qquad (1)$$

en la que $x(\tau)$ es la señal de voz y $w(\tau)$ representa la función de la ventana de análisis (por ejemplo, la ventana de Hamming). De esta forma se realiza un análisis localizado de la señal, esto es, la señal enventanada por $w(\tau)$ alrededor del instante de tiempo *"t"* es analizada a todas las frecuencias consideradas *"f"*.

Una vez que la ventana, $w(\tau)$ (tanto el tipo como la longitud) ha sido elegida para la realizar la STFT, la resolución temporal y frecuencial es fija sobre todo el plano tiempo-frecuencia puesto que se utiliza la misma ventana para extraer las características espectrales a todas las frecuencias.

A continuación, las log-energías en banda se calculan filtrando el espectro de potencia, $|S_x(\tau, f)|^2$ con un banco de filtros distribuidos según la escala mel y calculando el logaritmo correspondiente. A partir de las log-energías en banda pueden generarse dos tipos de parámetros distintos de entrada al reconocedor como describimos a continuación.

3.1.1. Parametrización MFCC

En el caso de los parámetros mel-cepstrum o MFCC, la decorrelación de las log-energías en banda se realiza utilizando la transformada coseno (DCT). A partir de ellos, se calculan los ΔMFCC (primeras derivadas).

3.1.2. Parametrización FFLFBE

Para el cálculo de la representación paramétrica de la señal de voz basada en el filtrado frecuencial de log-energías en bandas (FFLFBE), la transformación lineal que se utiliza para la decorrelación es un filtrado paso banda realizado mediante el filtro FIR de segundo orden, $H(z) = z - z^{-1}$, propuesto en [3]. Los ΔFFLFBE son las derivadas correspondientes.

3.2. Parámetros basados en la transformada ondicular

En este caso, para la extracción de las características espectrales de la señal de voz se utiliza la transformada ondicular definida por:

$$W_x(t,a)=\frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} x(\tau)\,\Psi^*\left(\frac{\tau-t}{a}\right)d\tau \qquad (2)$$

en la que $x(\tau)$ es la señal de voz y $\Psi(\tau)$ es la ondícula prototipo desplazada "t" unidades y escalada por el factor "a". Hemos elegido como ondícula prototipo la de Morlet, que es una función gaussiana modulada. De este modo, la ecuación (2) puede expresarse como:

$$W_x(t,a)=\frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} x(\tau+t)\,g\left(\frac{\tau}{a}\right)e^{-j2\pi\frac{f_o}{a}\tau}\,d\tau \qquad (3)$$

en la que $g(\tau)$ es una ventana gaussiana. Aunque dicha función es de longitud infinita, en la práctica truncamos los valores que están por debajo de un umbral establecido en 0,07 (este valor de umbral se eligió para aproximar el valor en los extremos de la ventana de Hamming), por lo que podemos determinar que la longitud temporal de la ondícula prototipo $\Psi(\tau)$ es $\lambda_{prot} \approx 4{,}625$ s. Dicha ondícula es un filtro paso banda cuya frecuencia central es $f_{prot} = f_o$ y con un ancho de banda determinado, bw_{prot}. Las versiones escaladas presentan una longitud proporcional a la escala ($\lambda_s = a.\lambda_{prot}$). También son filtros paso banda cuya frecuencia central es inversamente proporcional a la escala ($f_s = f_{prot}/a$) y cuyo ancho de banda también lo es ($bw_s = bw_{prot}/a$).

Si comparamos las ecuaciones (1) y (3), podemos observar que ambas expresiones son similares con $f = f_o/a$ y $w(t) = g(t/a)$ [10]. De hecho, en el caso de la TO, a la escala *"s-ésima"*, la señal es enventanada con una función gaussiana, $g(t/a_s)$ (con longitud $\lambda_s = a_s.\lambda_{prot}$), y entonces analizada a la frecuencia $f_s = f_{prot}/a_s$. Es decir, las longitudes

de las ventanas y la frecuencia central de cada subbanda son inversamente proporcionales entre sí.

La principal diferencia con respecto a la STFT clásica es el tamaño de la ventana de análisis: constante para todas las frecuencias en la STFT y variable con el factor de escala *"a"* (y por tanto, con la frecuencia) en la TO. De este modo, la TO ofrece diferentes resoluciones tiempo-frecuenciales. Para escalas pequeñas, el análisis ondicular tiene una buena resolución temporal para altas frecuencias (mientras que λ_s disminuye, f_s y bw_s aumentan) y, para valores más grandes de la escala *"a"*, se obtiene una buena resolución en frecuencia para bajas frecuencias (mientras que λ_s aumenta, f_s y bw_s disminuyen).

3.2.1. Elección de las escalas

Aunque existen otras posibilidades, la elección de las escalas se ha realizado de forma que las frecuencias correspondientes, f_s, tengan una distribución similar a la escala mel, aunque no exactamente la misma ([4], [5] y [6]). Del mismo modo, el valor concreto del número de escalas (N_s) es determinado experimentalmente.

En concreto, las frecuencias fueron calculadas según:

$$f_s = f_{\min}\left(\Delta_f\right)^s, \qquad 0 \le s \le N_s - 1 \tag{4}$$

en la que el incremento de frecuencias es:

$$\Delta_f = \left(\frac{f_{\max}}{f_{\min}}\right)^{\frac{1}{N_s+1}} \tag{5}$$

y f_{min} y f_{max} son la mínima y máxima frecuencia considerada (125 Hz y 4 KHz, respectivamente).

3.2.2. Elección de los desplazamientos

Otro aspecto importante a tener en cuenta en el diseño del parametrizador es el valor de los desplazamientos *"t"* en la ecuación (2), que determina la velocidad de extracción de los parámetros (tasa de trama). En el caso de la transformada de Fourier dicho desplazamiento es constante. Como en el caso de la transformada ondicular, las ventanas de análisis tienen distinta longitud para cada frecuencia, sería necesario establecer una tasa de trama distinta para cada una de dichas frecuencias. Este tipo de tasa de trama variable ha sido propuesto en [4] (reconocimiento de habla no uniforme). Sin embargo, en nuestro caso hemos optado por una tasa de trama fija como el propuesto en [5] y [6] aprovechando el alto grado de redundancia que presenta la señal de voz.

Una vez calculada la transformada ondicular y al igual que en el caso de la transformada de Fourier, en el que se aplica un banco de filtros en escala mel para reducir la dimensionalidad, también es necesario reducir la dimensionalidad de las *Ns* escalas consideradas. Para ello se realiza una agrupación de las escalas, de forma que el número de log-energías derivadas de la STFT y de la TO sea el mismo.

3.2.3. Parametrización MWCC

En este caso, se usa la DCT para la decorrelación de las log-energías. De este modo, se obtienen los coeficientes MWCC, a partir de los que se obtienen los delta, ΔMWCC.

3.2.4. Parametrización WFLFBE

El proceso de decorrelación se realiza mediante el filtrado lineal de las log-energías derivadas de la TO del mismo modo indicado en la sección 3.1.2, dando lugar a los parámetros WFLFBE y sus derivadas correspondientes ΔWFLFBE.

3.3. Combinación de parámetros

La combinación de diferentes tipos de parámetros puede dar lugar a importantes reducciones de la tasa de error del sistema si contienen información complementaria significativa. En nuestro caso, el hecho de que los parámetros MFCC y MWCC (o FFLFBE y WFLFBE) hayan sido derivados de distintas representaciones espectrales de la señal de voz (calculada a partir de la STFT en el primer caso, y a partir de la TO en el segundo), nos sugiere la posibilidad de plantear distintas combinaciones entre ellos que resulten ventajosas cara a mejorar la tasa de reconocimiento final. En la sección 5, mostraremos como efectivamente este tipo de parametrizaciones híbridas incrementa la tasa de reconocimiento del sistema.

4. Parametrizaciones basadas en la extracción discriminativa de rasgos (DFE)

Habitualmente, los dos módulos principales de un SRAH, parametrizador y clasificador, se diseñan de forma independiente. Sin embargo, este procedimiento no garantiza necesariamente una mejora de las tasas de reconocimiento del sistema global. De hecho, diversos trabajos muestran las ventajas de utilizar técnicas en las que se optimice de forma conjunta el parametrizador y el clasificador. Un ejemplo es la que se denomina genéricamente "Extracción Discriminativa de Rasgos" (DFE: "Discriminative Feature Extraction").

Recientemente, DFE se ha aplicado para el entrenamiento de bancos de filtros óptimos [11], de transformaciones lineales del espacio de los parámetros acústicos [12] o de los pesos de combinación de diferentes vectores de rasgos ("streams") [13]. En este artículo, lo utilizaremos para la mejora del parametrizador basado en transformada ondicular.

4.1. Ondículas gaussianas adaptativas

Podemos considerar las ondículas gaussianas adaptativas como una generalización de la de Morlet utilizada en la sección 3.2 con la diferencia de que la frecuencia de modulación f_o (antes constante), es una variable que se puede ajustar siguiendo un criterio de minimización especificado [14]. De esta forma, ya no están estrechamente ligadas la longitud y la frecuencia de análisis de cada subbanda, lo que posibilita su adaptación a las características de la señal de voz a analizar.

En este caso, la transformada ondicular viene definida por la siguiente expresión:

$$W_x(t,s) = \int_{-\infty}^{\infty} x(t+\tau)\, w_{g,s}(\tau) e^{-j2\pi f_s \tau} d\tau \tag{6}$$

en la que $w_{g,s}(\tau)$ es una ventana gaussiana adaptativa, con longitud λ_s (en s.) y modulada a la frecuencia f_s (en Hz).

4.1.1. Revisión del algoritmo MCE/GPD

El concepto principal de la técnica DFE es el entrenamiento discriminativo del módulo extractor de la representación paramétrica de la señal de voz con el objetivo de minimizar la tasa de error del sistema de reconocimiento global.

La aplicación práctica de las técnicas de extracción discriminativa de rasgos suele basarse en la aplicación del algoritmo de MCE/GPD ("Minimum Classification Error/Generalized Probabilistic Descent") [15]. Básicamente, consiste en la reestimación iterativa del conjunto de parámetros a entrenar, Φ, con el objetivo de minimizar una función de coste medio, $L(\Phi)$, que debe ser una buena aproximación del error de clasificación sobre el conjunto de datos de entrenamiento.

Para ello, se utiliza el algoritmo del gradiente descendente de modo que cada conjunto de parámetros, Φ, se actualiza iterativamente siguiendo la dirección del gradiente descendente. Por tanto, a cada iteración *"k"*, se calcula un nuevo conjunto de parámetros utilizando la siguiente expresión:

$$\Phi^k = \Phi^{k-1} - \eta \nabla L\left(\Phi^{k-1}\right) \qquad (7)$$

en la que η es el paso de adaptación y controla la velocidad de convergencia del algoritmo. El gradiente $\nabla L(\Phi)$ se obtiene calculando las derivadas parciales de $L(\Phi)$ mediante la aplicación de la regla de la cadena.

Para los parametrizadores basados en ondículas gaussianas adaptativas el conjunto de parámetros Φ que hemos considerado entrenables son dos tipos: las longitudes de las ventanas de análisis, $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_{Ns}\}$ y las frecuencias centrales de cada subbanda, $F = \{f_1, f_2, ..., f_{Ns}\}$.

En nuestro caso, las clases consideradas corresponden con los estados de los modelos de Markov que representan a cada una de las unidades alofónicas de la tarea. Para inicializar el algoritmo es necesario disponer de los datos de entrenamiento segmentados en estados y etiquetados. Este proceso se realiza de modo no supervisado mediante la aplicación del algoritmo de Viterbi con los HMM iniciales, lo que producirá errores en la segmentación que afectaran el comportamiento del algoritmo DFE.

Para evitar este inconveniente, se aplica una modificación del algoritmo DFE básico, conocido como SGDFE ("Single-Gaussian DFE") iterativo o segmental [12]. Se basa en el entrenamiento sucesivo de los parámetros ajustables del módulo extractor de características mediante DFE y de los parámetros de los modelos de Markov del clasificador final (medias, varianzas y matriz de transiciones) usando el criterio de máxima verosimilitud (o cualquier otro, como el MCE). A cada una de estas iteraciones se le denomina "iteración segmental". El proceso global acaba cuando la reducción de la función de coste medio o del error medio de clasificación en DFE es menor que un umbral predeterminado. En nuestro caso, el criterio de convergencia utilizado ha sido este último.

5. Resultados

5.1. Base de datos y sistema de referencia

Para la experimentación hemos utilizado la base de datos SpeechDat, un corpus de voz en castellano independiente del locutor grabado sobre la red telefónica pública española a 8 KHz (ley A). De todo el corpus hemos utilizado la parte correspondiente a palabras aisladas perteneciente a un conjunto de 1000 locutores distintos con un vocabulario de 1000 palabras. El conjunto de entrenamiento consiste en 5080 palabras y el de test de 2203 palabras.

El sistema de referencia es un reconocedor de palabras aisladas independiente del locutor basado en modelos ocultos de Markov continuos (CDHMM) [16]. El alfabeto de la tarea está compuesto por 45 unidades alofónicas cada una de ellas modelada por un HMM continuo de tres estados con topología de Bakis. Además, y con el objeto de reducir los posibles errores del detector de principio y fin (las palabras del conjunto de test han sido marcadas de forma automática) se han añadido dos unidades adicionales que representan el silencio inicial y final. De este modo, el número total de estados (clases diferentes en DFE) es de 141.

En todas las gráficas presentadas en este apartado, se muestra la tasa de reconocimiento del sistema junto con las bandas de fiabilidad para una confianza del 95 %.

5.2. Parametrizaciones empíricas

En este apartado consideraremos las parametrizaciones descritas en la sección 3.

En el caso de los parámetros derivados de la transformada de Fourier se utilizó una ventana de Hamming de 25 ms. y un banco de filtros de 17 subbandas en escala mel. De las log-energías se extrajeron dos tipos de parametrizaciones:

- Parametrización MFCC: 10 MFCC y sus derivadas.
- Parametrización FFLFBE: 17 FFLFBE y sus derivadas.

Para el caso de la transformada ondicular, se utilizaron 34 ventanas de análisis de tamaño variable (de 50 a 4 ms) tal y como están representadas en la Figura 4 (curva "TO inicial") en función de la frecuencia que analizan. El número de escalas fue determinado empíricamente. Del mismo modo, una serie de resultados preliminares sugirieron que era conveniente restringir el tamaño máximo de la ventana a 50 ms. Finalmente, se obtuvieron dos tipos de parámetros:

- Parametrización MWCC: 10 MWCC y sus derivadas.
- Parametrización WFLFBE: 17 WFLFBE y derivadas.

En todos los casos, el vector de parámetros fue complementado con la log-energía total y su derivada.

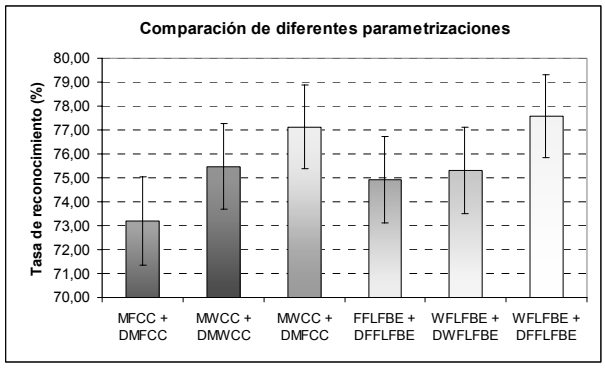

Figura 2: Resultados de reconocimiento para las parametrizaciones basadas en métodos empíricos.

En la Figura 2 están indicados los resultados obtenidos con los experimentos de referencia ("MFCC + ΔMFCC" y "FFLFBE + ΔFFLFBE") y los parámetros correspondientes ondiculares ("MWCC + ΔMWCC" y "WFLFBE + ΔWFLFBE") Podemos observar que ambos casos, los parámetros derivados de la TO ofrecen resultados mejores que las basadas en la STFT, aunque las diferencias en las tasas no son estadísticamente significativas En la misma gráfica, se pueden observar los resultados obtenidos con parametrizaciones híbridas, en las que se combinan parámetros estáticos derivados de la TO y dinámicos procedentes de la STFT ("MWCC + ΔMFCC" y "WFLFBE + ΔFFLFBE"). Como puede observarse estas combinaciones resultan en una mejora del funcionamiento del sistema comparado con el sistema base y el basado únicamente en parámetros derivados de la transformada ondicular. Por tanto, esta parametrización híbrida será la que utilicemos en el resto de la experimentación.

5.3. Parametrizaciones con DFE

En esta sección presentamos los resultados obtenidos con la aplicación de las técnicas de DFE sobre las parametrizaciones basadas en la transformada ondicular (ondículas gaussianas adaptativas).

En primer lugar, analizaremos en mayor detalle la aplicación de DFE para el entrenamiento de las longitudes de las ventanas de análisis en la parametrización híbrida MWCC + ΔMFCC. Los valores de inicialización para dichas longitudes corresponden con los utilizados en la experimentación anterior.

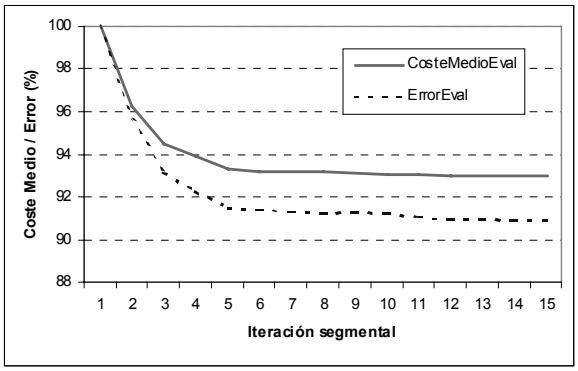

Figura 3: Evolución del coste medio y el error para MWCC + ΔMFCC (DFE con longitudes).

Para determinar la iteración segmental en la que el algoritmo converge, dividimos la base de datos de entrenamiento original en dos subconjuntos: los dos primeros tercios fueron usados para entrenar el parametrizador y el tercio restante fue utilizado como conjunto de validación. De este modo, podemos considerar que el proceso ha terminado cuando el error de clasificación cometido sobre el conjunto de validación no se reduce o la reducción de dicho error es tan pequeña que no supera un umbral predefinido.

En la Figura 3 se muestra la evolución del coste medio y error medio de clasificación del conjunto de evaluación en función del número de iteraciones segmentales. Para facilitar la visualización, cada una de estas medidas fue normalizada por su valor máximo (que corresponde con el de la primera iteración). Como podemos observar, el algoritmo disminuye tanto el coste como el error medio rápidamente en las primera iteraciones.

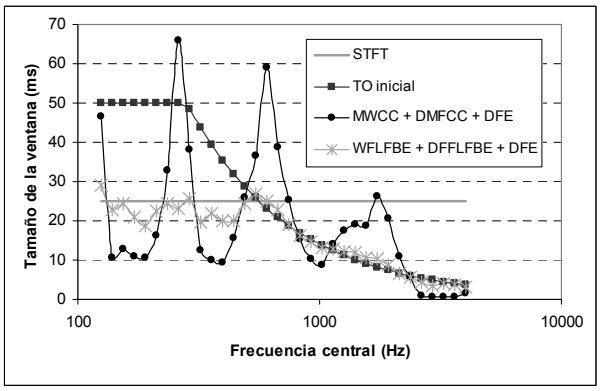

Figura 4: Longitudes de las ventanas de análisis en función de la frecuencia de cada subbanda

La Figura 4 muestra los tamaños de las ventanas resultantes en la iteración segmental 7. Para facilitar la comparación, esta figura también muestra los tamaños para el sistema convencional basado en STFT y la configuración inicial para los parámetros MWCC (curva "TO inicial"). Puede observarse que la longitud de las ventanas correspondientes a frecuencias en torno a 250, 750 y 1750 Hz (correspondientes a la posición de formantes) se incrementan notablemente cuando se aplica el procedimiento DFE. Por otra parte, las ventanas correspondientes a frecuencias sobre los 2250 Hz decrementan su longitud.

El mismo experimento se realizó para el sistema basado en la combinación de parámetros WFLFBE + ΔFFLFBE. En la Figura 4 están representadas las longitudes de las ventanas de análisis modificadas por el DFE para la iteración segmental 20. También se representan las de la iteración inicial ("TO inicial") y las de la STFT convencional. Es importante reseñar que las longitudes "óptimas" obtenidas son muy distintas de las conseguidas en el caso de parámetros cepstrales. De hecho, en este caso para frecuencias bajas, las longitudes obtenidas corresponden con las de la STFT mientras que para las altas corresponden con las de la TO original. Este hecho sugiere que los resultados no son extrapolables entre diferentes tipos de parametrizadores.

También se realizaron experimentos similares en los que los parámetros ajustables del parametrizador eran las frecuencias de análisis dejando fijas las longitudes.

La Figura 5 contiene los resultados de reconocimiento con DFE junto con los experimentos de referencia ("MFCC + ΔMFCC" y "FFLFBE + ΔFFLFBE"). Las tasas de reconocimiento obtenidas con DFE son el promedio sobre los cinco mejores resultados observados en cada iteración segmental. Como podemos observar, la aplicación de DFE para entrenamiento de longitudes y frecuencias siempre produce decrementos de la tasa de error del sistema con respecto a los experimentos base correspondientes. Sin embargo, las mejoras son estadísticamente significativas únicamente en el caso de entrenamiento de las longitudes de las ventanas de análisis para los dos casos de parámetros cepstrales y log-energías en banda.

Finalmente, con respecto a la comparación entre los parámetros cepstrales y los basados en log-energías filtradas, no podemos aseverar (estadísticamente) que sean más efectivos unos que otros, puesto que con ambos se obtienen resultados similares. Se requeriría disponer de más datos de evaluación para confirmar estadísticamente las diferencias producidas en la tasa de reconocimiento del sistema.

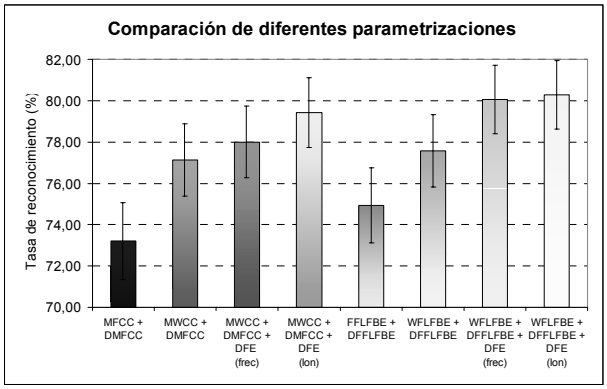

Figura 5: Resultados de reconocimiento para las parametrizaciones basadas en DFE.

Con este conjunto de experimentos hemos comprobado que las parametrizaciones basadas en transformada ondicular y con DFE producen tasas de reconocimiento significativamente superiores a las de las parametrizaciones convencionales.

6. Conclusiones

En este artículo, hemos comparado el funcionamiento de diversas parametrizaciones en entorno telefónico, que hemos clasificado en dos grandes grupos: empíricas y basadas en la extracción discriminativa de rasgos.

Con respecto al primer grupo, hemos analizado las ventajas que supone la utilización de la transformada ondicular. Esta alternativa produce resultados similares o ligeramente superiores a los obtenidos con el análisis de Fourier, aunque las mejoras no son estadísticamente significativas. Sin embargo, la utilización de parametrizaciones híbridas incrementa significativamente las tasas de reconocimiento con respecto a los experimentos base.

Con respecto al segundo grupo, hemos propuesto un análisis basado en ondículas gaussianas adaptativas que ofrece una gran flexibilidad en el diseño del parametrizador, de modo que la elección de las frecuencias de las subbandas y la longitud de las ventanas de análisis, pueden ser entrenados atendiendo al criterio de minimización del error de clasificación, mediante la extracción discriminativa de rasgos (DFE). Dicha técnica mejora de forma significativa la tasa de reconocimiento del sistema, con lo que hemos mostrado que es más efectivo realizar la optimización del parametrizador y clasificador de forma conjunta y no de forma independiente.

En la actualidad, estamos extendiendo nuestros resultados a modelado acústicos más complejos (modelos dependientes del contexto) y a entornos dominados por el ruido aditivo.

Finalmente, también estamos explorando la posibilidad de mejorar el proceso de combinación de parámetros mediante la inclusión de pesos de ponderación (entrenables) que determinen la contribución adecuada de cada tipo de rasgos.

7. Referencias

[1] Deller, J. R., Proakis J. G. and Hansen, J. H. L., *Discrete-Time Processing of Speech Signals,* Ed. Macmillan Publishing Company, New York, 1993.

[2] Davis, S. B. y Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. Acoustics, Speech and Signal Proc.*, 28 (4), pp. 357-366, agosto 1980.

[3] Nadeu C., Hernando J. and Gorricho, M., "On the Decorrelation of Filter-Bank Energies in Speech Recognition", *Eurospeech 1995*, pp. 1381-1384, 1995.

[4] Janer L., Martí J., Nadeu C. and Lleida Solano E., "Wavelet Transforms for Non-Uniform Speech Recognition Systems", *ICSLP 1996*, vol. 4, pp. 2348-2351, 1996.

[5] Gemello, R., Albesano D., Moisa L. and de Mori R., "Integration of Fixed and Multiple Resolution Analysis in a Speech Recognition System", *ICASSP 2001*, 2001.

[6] Sarikaya, R., Pellon B. L. and Hansen, J. H. L., "Wavelet Packet Transform Features with Application to Speaker Identification", *Eurospeech 2001*, 2001.

[7] Furui S., "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Trans. Acoustics, Speech and Signal Proc.*, vol. 29, pp. 254-272, 1981.

[8] Hermansky H. and Morgan, N., "RASTA Processing of Speech", *IEEE Trans. Speech and Audio Proc.*, vol. 2, nº 4, pp. 578-589, 1994.

[9] Nadeu, C., Macho D. and Hernando J, "Time and Frequency Filtering of Filter-Bank Energies for Robust HMM Speech Recognition", *Speech Communication*, 34, pp. 93-114, 2001.

[10] Unser, M., "Fast Gabor-Like Windowed Fourier and Continuous Wavelet Transforms", *IEEE Signal Processing Letters*, vol. 1, nº 5, pp. 76-79, mayo 1994.

[11] Biem A., Katagiri S., McDermott E. and Juang, B. H., "An Application of Discriminative Feature Extraction to Filter-Bank-Based Speech Recognition", *IEEE Trans. Speech and Audio Proc.*, 9 (2), pp. 96-110, febrero 2001.

[12] de la Torre A., Peinado A. M., Rubio A. J.and Segura, J. C., "Discriminative Improvement of the Representation Space for Continuous Speech Recognition", *Computational Models of Speech Pattern Processing*, NATO ASI Series, Springer Verlag, 1999.

[13] Chu S. M. and Zhao. Y., "Robust Speech Recognition Using Discriminative Stream Weighting and Parameter Interpolation", *ICSLP 1998*, 1998.

[14] Kadambe S. and Srinivasan, P.,"Applications of Adaptive Wavelets for Speech", *Journal of Optical Engineering. Special Issue on Wavelets*, julio 1994.

[15] Juang B. H. and Katagiri, S., "Discriminative Learning for Minimum Error Classification"., IEEE Trans. Signal Processing, vol. 40, nº 12, pp. 3043-3054, 1992.

[16] Ferreiros, J., "Aportación a los Métodos de Entrenamiento de Modelos de Markov para Reconocimiento de Habla Continua", *Tesis Doctoral*, ETSIT, Universidad Politécnica de Madrid, 1996.