

Procesado de Señal Vocal en Aplicaciones de Reconocimiento del Habla

Jon Ander Gómez

Departamento de Sistemas Informáticos y Computación

Universidad Politécnica de Valencia

ESPAÑA

jon@dsic.upv.es

Resumen

En el Reconocimiento Automático del Habla (RAH) mediante computadoras es necesario transformar la forma de onda que toma la señal vocal con el objeto de obtener características acústicas relevantes que permitan descifrar el mensaje contenido, y de paso reducir la cantidad de información a tratar.

Existen diversas técnicas para la extracción de características a partir de la forma de onda original. En todas ellas se obtiene una secuencia de vectores de parámetros que representa la señal vocal a lo largo del tiempo.

En el presente documento se describe una de las técnicas más utilizadas para la extracción de características en aplicaciones de RAH. Concretamente la que obtiene los *Coeficientes Cesptrales a partir del Banco de Filtros según Escala de Mel* (MFCC, del inglés *Mel Frequency Cepstral Coefficients*).

Palabras clave: Tratamiento Digital de la Señal, Extracción de características, Análisis Frecuencial, Reconocimiento del Habla.

1 Introducción

El principal objetivo del presente documento es plasmar como se realiza el *Preprocesado* de la señal vocal en la mayoría de los sistemas de RAH. Al preprocesado o extracción de características también se le conoce como *Parametrización*. Al presentar referencias de las técnicas sobre procesado de señal utilizadas junto a detalles de implementación, este documento pretende ser un punto de partida para cualquier discusión futura sobre el tema; así como evitar que se tenga que realizar todo el estudio de nuevo por parte de quien sólo necesite preprocesar señal vocal.

La forma de onda acústica es la manera en que se transmite la señal vocal a través del aire, su canal natural. Esta señal contiene el mensaje a transmitir entre

dos interlocutores, pero dicha forma de onda no es la mejor representación para el Reconocimiento Automático del Habla (RAH) mediante computadoras. Es necesario aplicar algún tipo de transformación a dicha señal en aras a obtener una representación más idonea para su análisis. Esta representación debe tener como objetivos principales: 1) reducir la cantidad de información a tratar, y 2) obtener características acústicas relevantes de la señal vocal que permitan descifrar el mensaje contenido.

Las distintas manifestaciones acústicas de la señal vocal se distinguen principalmente en que cada una concentra la energía a diferentes frecuencias. Por lo que el análisis en el dominio de la frecuencia es una buena manera de obtener representaciones de la señal vocal.

En el presente documento se muestra una de las técnicas más utilizadas para la extracción de características en aplicaciones de RAH. Aplicando cualquiera de las técnicas existentes se obtiene una secuencia de vectores de parámetros que representa la señal vocal a lo largo del tiempo. Para obtener cada vector se aplica la transformación elegida cada cierto intervalo de tiempo, y en función de la transformación utilizada los vectores contendrán unos parámetros u otros. La técnica expuesta aquí obtiene como parámetros los Coeficientes Cepstrales y la Energía. Como extensión a esta parametrización se utilizan también las primeras y las segundas derivadas de dichos parámetros.

La exposición del presente documento está organizada siguiendo el mismo orden en el que se procesa la señal vocal. Una pequeña introducción al muestreo de la señal y su conversión Analógico-Digital aparece en la sección 2. En la sección 3 se aborda el filtrado de la señal en el dominio del tiempo. En la sección 4 se explica como se realiza el análisis frecuencial. El cálculo del Banco de Filtros según Escala de Mel se explica en la sección 5, y la obtención de los Coefficients Cepstrales en la 6.

Toda la formulación utilizada en este documento ha sido extraída de la bibliografía expuesta al final. Todas las referencias son básicas para el tratamiento digital de la señal vocal, y en algunas se muestran técnicas alternativas de extracción de características.

2 Muestreo de señal. Conversión Analógico-Digital

La señal vocal producida por el habla humana es una señal continua en forma de onda acústica, y puede ser representada como función de una variable continua t , que representa el tiempo. En este documento se usa la notación $s(t)$ para representar la señal vocal analógica (o continua) a lo largo del tiempo.

Para su tratamiento mediante computadoras, sistemas discretos (digitales), la señal vocal ha de ser muestreada, no se puede tratar como una función continua. La forma de onda acústica es convertida en impulsos eléctricos que un convertidor Analógico-Digital (A/D) codificará en valores numéricos dentro de un rango determinado. Esto significa que como representación de la señal dispondremos de una secuencia de valores. Cada uno de estos valores es obtenido cada cierto intervalo de tiempo τ_m . A τ_m se le conoce como periodo de muestreo, aunque es más normal utilizar su inversa, la frecuencia de muestreo F_m . En definitiva, si F_m toma valor $16kHz$, como representación de un trozo de señal vocal dispondremos de una secuencia de valores a razón de 16000 por segundo.

La señal discretizada (muestreada) ya no es función de una variable continua, sino que es una secuencia de valores obtenidos cada intervalo de tiempo τ_m . Por ello cambia la notación, se utiliza $s(n\tau_m)$ en vez de $s(t)$. En muchos casos, por simplificación se utilizará $s(n)$.

Los valores que puede tomar F_m vienen determinados por el Teorema de Nyquist o del muestreo. La idea base de este teorema es que si vamos a considerar en nuestro sistema hasta al menos fHz , se debe de muestrear a $2fHz$ o frecuencias mayores.

Más detalles de lo expuesto en esta sección se pueden encontrar en [Rabiner, 1978] y [Oppenheim, 1975].

3 Filtrado de la señal vocal

Una vez se dispone de la señal vocal digitalizada $s(n)$, se suele aplicar un filtro de respuesta a impulsos con duración finita (FIR: Finite duration Impulse Response). Este filtro se aplica antes de realizar el análisis frecuencial y consiste en una transformada Z cuya función de transferencia es

$$H(z) = 1 - \alpha z^{-1}.$$

Aplicando esta función se consigue la siguiente ecuación diferencial de primer orden

$$x(n) = s(n) - \alpha s(n-1).$$

Donde $x(n)$ representa la señal vocal filtrada (o preenfatizada), y α es el coeficiente o factor de *preénfasis*. Los valores típicos para dicho factor oscilan entre 0.9 y 1.0, el más utilizado es 0.95.

El objeto de aplicar este filtro es realzar el espectro a altas frecuencias, ya que la mayor parte de la energía se concentra en las bajas frecuencias. Además tiene otro efecto importante y es que reduce considerablemente el problema de la componente

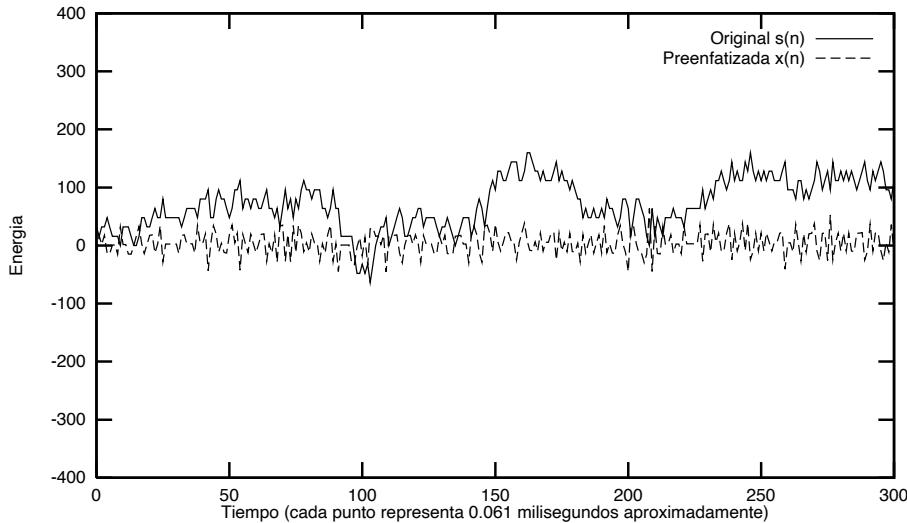


Figura 1: Segmento de señal vocal en el dominio del tiempo correspondiente a un silencio.

continua¹, lo cual evita tener que aplicar técnicas expresamente. En la figura 1 se aprecia el efecto de la componente continua en un segmento perteneciente a un silencio. Este problema afecta considerablemente a la estimación de la densidad de cruces por cero², y puede llegar a desvirtuar bastante el cálculo de la energía. Además, el valor de la componente continua varía de un sistema de adquisición a otro, lo que hace especialmente recomendable la aplicación de algún filtro.

Más detalles de lo expuesto en esta sección se pueden encontrar en [Rabiner, 1978], [Rabiner, 1993] y [Oppenheim, 1975].

4 Análisis frecuencial. Transformada de Fourier

Por la naturaleza de la señal vocal resulta muy apropiado el análisis en el dominio de la frecuencia, y en concreto la aplicación de la transformada de Fourier. Esta

¹La componente continua, inducida en el propio sistema de adquisición, provoca un desplazamiento de toda la señal vocal respecto del valor 0. Cero es el valor medio teórico de toda señal acústica.

²La densidad de cruces por cero es un parámetro importante para detectar los fonemas fricativos. Su cálculo se realiza contabilizando la veces que la señal cambia de signo. Si el valor medio de la señal no es 0 por culpa de la componente continua dicho cálculo será erróneo.

transformación nos permite conocer el valor de la energía contenida en la señal vocal como función de la frecuencia $S(\omega)$.

Dado que el RAH se lleva a cabo sobre sistemas discretos, de ahora en adelante utilizaremos $s(n)$, indistintamente de si ha sido filtrada, para referirnos a la señal vocal muestreada en el dominio del tiempo. Y para analizarla en el dominio de la frecuencia se utilizará la aproximación Discreta de la Transformada de Fourier (DFT, del inglés *Discrete Fourier Transform*). Concretamente una versión de la *Fast Fourier Transform (FFT)* [Press, 1994].

La FFT espera un vector con N elementos y devuelve un vector con $\frac{N}{2}$ elementos representando los valores que toma la energía en función de la frecuencia. N debe ser potencia de 2. Una vez más se dispone de una secuencia de valores y no de una función; así que los valores devueltos representan la energía a unas frecuencias determinadas. Por la naturaleza de la FFT, cada valor corresponde a un pequeño rango de frecuencias, también conocido como canal. La frecuencia central f_c de cada canal depende del número de puntos tomados de la señal vocal N y de la frecuencia de muestreo F_m . Así, que tras aplicar la FFT se obtiene un vector con $\frac{N}{2}$ valores representando los $\frac{N}{2}$ canales.

Cada f_c se calcula según la siguiente fórmula

$$f_c^k = \frac{k}{N} F_m = \frac{k}{N\tau_m} \quad \forall 0 \leq k < \frac{N}{2}.$$

El primer canal corresponde a $0Hz$ y el último a $(\frac{\frac{N}{2}-1}{N\tau_m})Hz$. Y como se puede apreciar, las f_c^k están equiespaciadas en el dominio de la frecuencia.

Para poder analizar la evolución de las características acústicas, y así distinguir los distintos fonemas emitidos a lo largo del tiempo, es necesario realizar el análisis frecuencial cada cierto intervalo de tiempo τ_s . A esto se le conoce como *submuestreo*, y τ_s es el periodo de submuestreo, expresado normalmente en milisegundos. En la bibliografía, al submuestreo también se le conoce como *Short-Term Processing of Speech* o *Short-Term Analysis* [Deller, 1993]. La motivación de esto parte de que el proceso de producción del habla puede ser considerado como un fenómeno quasi-estacionario. Por lo que aplicando el análisis frecuencial a un trozo de señal vocal en el que se considera que las características acústicas no varían, y repitiéndolo consecutivamente, se obtiene una representación de una pronunciación como una secuencia de vectores de parámetros.

El valor de la frecuencia de submuestreo f_s , inversa del periodo de submuestreo τ_s , debe de tomar valores que permitan el estudio de la evolución de las características acústicas. Valores demasiado pequeños de f_s (grandes de τ_s) impedirán que

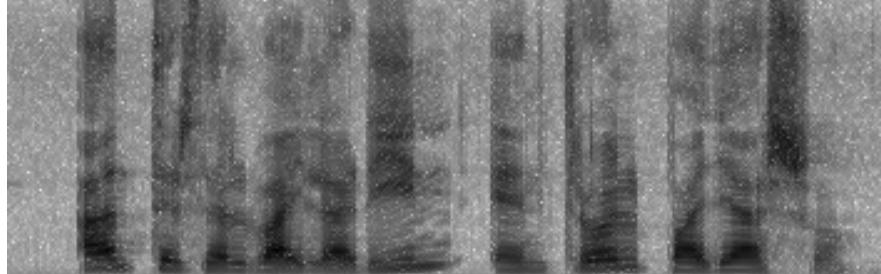


Figura 2: Espectograma de la frase “Antes del bailarin entró el payaso”.

se puedan detectar cambios espectrales importantes, y demasiado grandes (pequeños de τ_s) generará demasiada información a procesar sin reportar beneficios significativos. Asimismo, el valor de f_s vendrá en cierto modo limitado por la frecuencia de muestreo F_m . Valores típicos de F_m varían entre $8kHz$ y $20kHz$; f_s suele tomar valores alrededor de los $100Hz$, lo que equivale a realizar el sumbmuestreo cada 10 milisegundos. Discusiones y detalles se pueden encontrar en [Oppenheim, 1975], [Rabiner, 1978], [Rabiner, 1993] y [Deller, 1993]. Finalmente, el valor de N también está relacionado con F_m y f_s , su dependencia se basa en que debe de existir un solapamiento entre ventanas de análisis de al menos el 50%.

De lo descrito se deduce que el análisis frecuencial va a consistir en aplicar una transformación sobre la señal vocal, cuyo producto será una secuencia de vectores de parámetros. Estos parámetros serán finalmente otros, pero en esta sección vamos a hablar del espectro o representación de los valores de la energía en el dominio de la frecuencia.

Al tratar con señales discretas y aplicar la DFT sobre una ventana de análisis obtenemos una secuencia de vectores que representamos con $S(k\omega_c, f\tau_s)$, donde ω_c representa $\frac{1}{N\tau_m}$, y k toma valores desde 0 hasta $\frac{N}{2} - 1$. N es el número de puntos tomados para realizar la DFT. A la ventana de análisis se le conoce con el nombre *frame*³, que utilizaremos de ahora en adelante. Como simplificación de $S(k\omega_c, f\tau_s)$ se utilizará $S(k, f)$, donde k es el número de canal frecuencial y f es el número de *frame*, que en este caso representa el tiempo. En la figura 2 se puede apreciar el resultado de aplicar la FFT. Los niveles de gris representan la energía, el eje horizontal representa el tiempo (*frames*) y el vertical la frecuencia (canales).

Cada vez que tomamos una *frame* para aplicar la FFT se aplica una ventana

³Se adopta el término anglosajón por que denota perfectamente el significado que se quiere expresar en este contexto, ya que es ampliamente utilizado en RAH.

de suavizado. El objetivo es minimizar las discontinuidades de la señal al principio y fin de cada *frame*. De no aplicar ninguna ventana de suavizado estaríamos aplicando una ventana rectangular. La ventana más utilizada es la Hamming (vease figura 3), aunque existen otras como Blackman, Kaiser y Hanning [Rabiner, 1978], [Deller, 1993], [Rabiner, 1993]. Aplicar una ventana consiste en multiplicar $s(n)$ por una función ventana $w(n)$, $f(n; m) = s(n)w(m - n)$, donde $f(n; m)$ representa los N puntos de la señal vocal, desde el n hasta el m , que forman la ventana de análisis o *frame* sobre los que aplicar la FFT. $w(m - n)$ representa la ventana a aplicar.

La ventana rectangular se define como:

$$w(n) = \begin{cases} 1, & n = 0, 1, \dots, N - 1; \\ 0, & \text{cualquier otro valor de } n, \end{cases} \quad (1)$$

y la ventana de Hamming como:

$$w(n) = \begin{cases} 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N-1}\right), & n = 0, 1, \dots, N - 1; \\ 0 & \text{cualquier otro valor de } n. \end{cases} \quad (2)$$

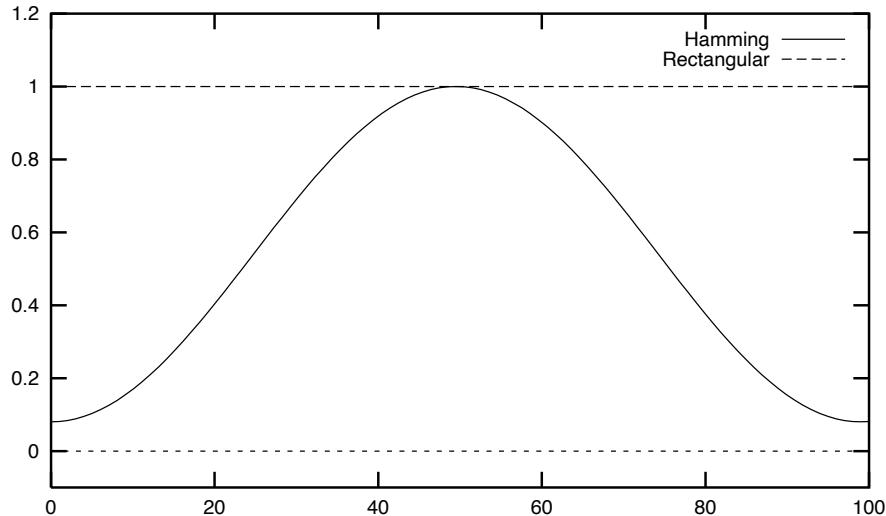


Figura 3: Ventana Hamming para $N=100$.

5 Banco de Filtros

Aplicando las técnicas descritas en las anteriores secciones se obtiene un espectograma como representación de la señal vocal (ver figura 2). Esto representa una gran cantidad de información por *frame* que en vistas a las siguientes fases de reconocimiento interesa reducir.

Cada *frame* está formada por $\frac{N}{2}$ canales. N se obtiene directamente de F_m y de f_s , que en la mayoría de los sistemas de RAH toman los valores $16kHz$ y $100Hz$ respectivamente. Esto se traduce en que se aplica una ventana de análisis cada 10 milisegundos, que sobre la señal discreta supone desplazar cada 162 puntos. Dado que entre dos ventanas de análisis consecutivas es aconsejable que exista un solapamiento del 50%, el número de puntos a tomar será de 324. Como la FFT espera un vector con un número de elementos potencia de 2, el número utilizado, en este caso, será 512. Al tomar de la señal 324 puntos, se llenará con ceros por la derecha (*Zero padding*) hasta 512 puntos [Press, 1994]. Por lo tanto, para el caso descrito cada *frame* está compuesta de 256 canales.

Una de las posibilidades para reducir la cantidad de canales es aplicar un *Banco de Filtros*. Esto es una secuencia de filtros paso-banda que agrupan los canales originales reduciendo la cantidad de parámetros a tratar. El Banco de Filtros que presentamos aquí, y que es ampliamente utilizado, está basado en la Escala de Mel [Rabiner, 1993], [Deller, 1993]. Esta escala consiste en agrupar los canales según unos filtros cuyo ancho de banda aumenta conforme la frecuencia. Las frecuencias centrales de dichos filtros se extraen aplicando la ecuación (3), que consiste en que las frecuencias centrales de dichos filtros estén equiespaciados según las frecuencias de Mel y no las normales en Hz . La relación entre f_{Hz} y f_{mel} aparece en la ecuación (3) y su inversa en (4) [Young, 1997]. La idea de esta escala es realizar el análisis frecuencial al igual que lo hace el oído humano, que es capaz de discriminar mejor los sonidos a bajas frecuencias.

$$f_{mel} = 2595 * \log_{10}(1 + \frac{f_{Hz}}{700}) \quad (3)$$

$$f_{Hz} = 700(10^{\frac{f_{mel}}{2595}} - 1) \quad (4)$$

Tomando la ecuación (3) y su inversa (4), se calculan las frecuencias centrales para cada filtro. Lo único que hay que determinar es entre qué rango de frecuencias se van a aplicar los filtros. Los valores utilizados por el autor son 100Hz para el

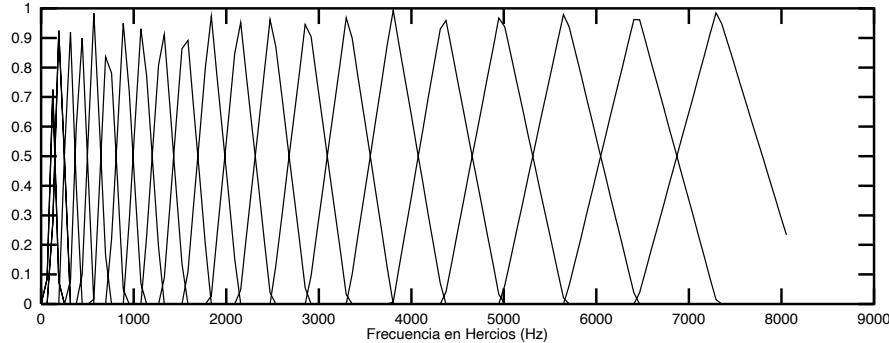


Figura 4: Banco de Filtros según Escala de Mel.

primer filtro y $0.45 * F_m$ para el último. La forma de cada filtro es triangular con un solapamiento del 50%, como se muestra en la figura 4.

La manera de obtener las frecuencias centrales en Hz se consigue con el siguiente algoritmo.

Algorithm 1 *Cálculo de los Filtros según Escala de Mel.*

Entrada:

f_{Hz}^1, f_{Hz}^{NF} - Frecuencias centrales para el primer y último filtros.

NF - Número de filtros.

Comienzo:

$$f_{mel}^1 = 2595 * \log_{10}(1 + \frac{f_{Hz}^1}{700})$$

$$f_{mel}^{NF} = 2595 * \log_{10}(1 + \frac{f_{Hz}^{NF}}{700})$$

$$\delta_{mel} = \frac{f_{mel}^{NF} - f_{mel}^1}{NF - 1}$$

$\forall 2 \leq k < NF \quad hacer$

$$f_{mel}^k = f_{mel}^{k-1} + \delta_{mel}$$

$$f_{Hz}^k = 700 * (10^{\frac{f_{mel}^k}{2595}} - 1)$$

\swarrow

Fin.

6 Coeficientes Cepstrales.

Los coeficientes cepstrales o *Mel-Frequency Cepstral Coefficients* (MFCC) se obtienen a partir de la salida del Banco de Filtros. Es otra reducción de la cantidad de información a tratar. Su cálculo se realiza aplicando una transformada coseno sobre el banco de filtros. Estos coeficientes recogen información sobre la forma del espectro, en otras palabras, de cómo se distribuye la energía a lo largo del espectro. En este caso se aplica sobre el Banco de Filtros según Escala de Mel, pero también se utiliza aplicada directamente sobre el espectro.

Cada coeficiente se obtiene aplicando la fórmula (5) sobre la salida del banco de filtros. El número de coeficientes cepstrales más utilizado ronda los 10. En la figura 5 se muestran las funciones de la transformada coseno para los 8 primeros.

$$c_n = \sqrt{\frac{2}{NF}} \sum_{k=1}^{NF} S(k) * \cos \left[n \left(k - \frac{1}{2} \right) \frac{\pi}{NF} \right] \quad (5)$$

NF es el número de filtros, n el orden del coeficiente y $S(k)$ representa en este caso el canal k -ésimo de la salida del banco de filtros. En las implementaciones que ha trabajado el autor el número de filtros es 21, y el número de coeficientes cepstrales 10.

Cuando hablamos de los MFCC como parametrización, no sólo se hace referencia a los 10 primeros coeficientes cepstrales, también se está utilizando la energía, de manera que esta parametrización constaría de 11 parámetros (Energía + 10 primeros CC's).

La energía se obtiene típicamente con la fórmula (6), pero para que sea independiente de las frecuencias de muestreo y submuestreo, de las que depende directamente el valor de N (número de puntos para la ventana de análisis), el autor utiliza la fórmula (7).

$$E = \log \sum_{n=1}^N s(n)^2 \quad (6)$$

$$E = \log \left(\frac{1}{N} \sum_{n=1}^N s(n)^2 \right) \quad (7)$$

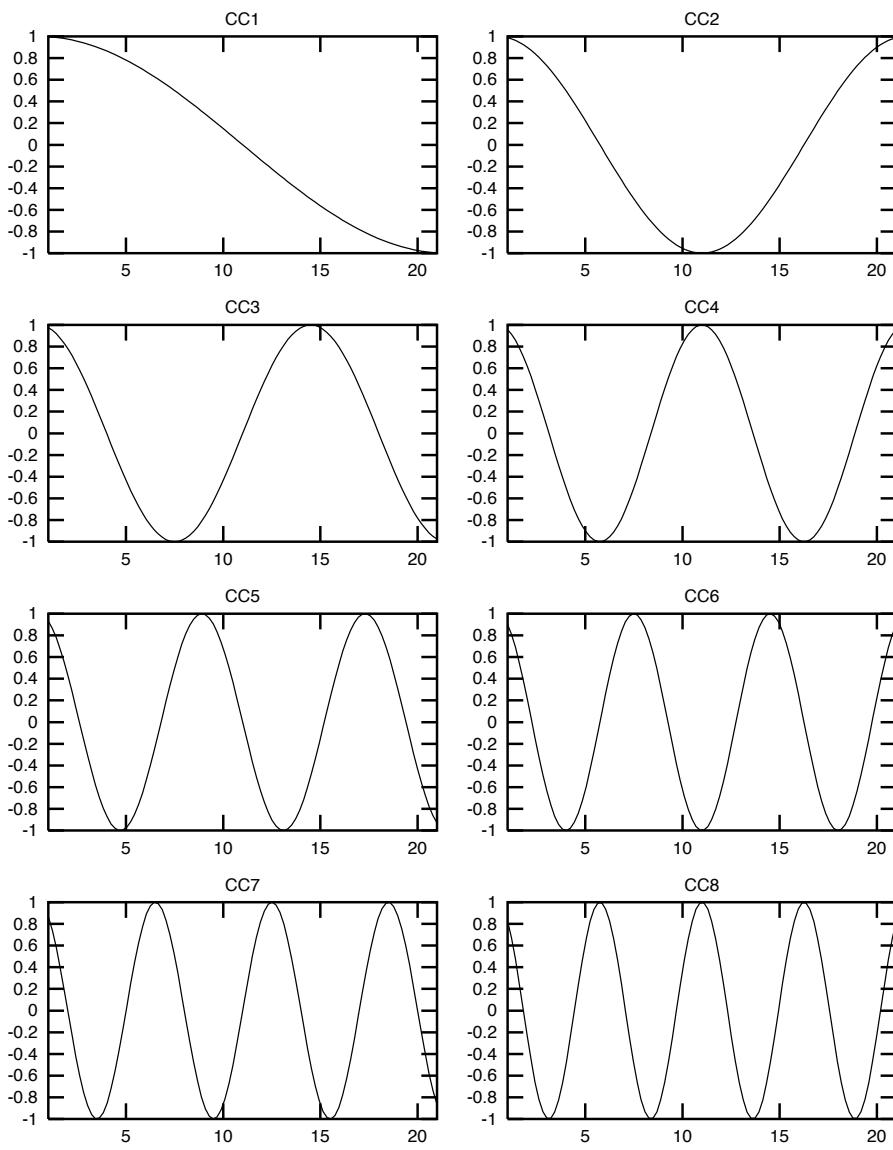


Figura 5: Aspecto de los 8 primeros Coeficientes Cepstrales para 21 filtros.

Como extensión de los MFCC, algunos sistemas de RAH utilizan las primeras y/o segundas derivadas de estos parámetros. Cuyo cálculo, por cada parámetro en particular, se obtiene utilizando la fórmula (8) [Young, 1996], [Young, 1997].

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2} \quad (8)$$

Donde d_t es la primera derivada de uno de los parámetros para el instante de análisis t , Θ es el ancho de la ventana para calcular las derivadas, y c_t representa el coeficiente en el instante t . Aplicando la misma fórmula sobre las primeras derivadas se obtienen las segundas.

Referencias

- [Brigham, 1974] E. Oran Brigham. “The Fast Fourier Transform”. Prentice-Hall, 1974
- [Brigham, 1988] E. Oran Brigham. “The Fast Fourier Transform and its Applications”. Prentice-Hall, 1988
- [Deller, 1993] John R. Deller, John G. Proakis, John H.L. Hansen. “Discrete-Time Processing of Speech Signals”. Macmillan Publishing Company, 1993
- [Press, 1994] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery. “Numerical Recipes in C”. Cambridge University Press, 1988-1994
- [Rabiner, 1978] L.R. Rabiner, R.W. Schafer. “Digital Processing of Speech Signals”. Prentice-Hall, 1978
- [Oppenheim, 1975] Alan V. Oppenheim, Ronald W. Schafer. “Digital Signal Processing”. Prentice-Hall, 1975
- [Rabiner, 1993] L.R. Rabiner, Biing-Hwang Juang. “Fundamentals of Speech Recognition”. Prentice-Hall, 1993
- [Young, 1996] Steve Young. “A Review of Large-Vocabulary Continuous-speech Recognition”. IEEE Signal Processing Magazine September 1996, 45-57
- [Young, 1997] Steve Young, Julian Odell, Dave, Ollason, Valtcho Valtchev, Phil Woodland. “The HTK Book”. Cambridge University, 1995, 1996, 1997