3 EXTRACCIÓN DE CARACTERÍSTICAS

En este capítulo se procederá a describir cómo se transforma el audio en una serie de parámetros que representan de forma compacta la información del sonido. Éste es el punto de partida no sólo de un sistema de verificación de locutor, sino de cualquier sistema de reconocimiento de voz, idioma, etc. Convendría aclarar que la parametrización es igual para los datos de entrenamientos que para los de evaluación. Por lo tanto, cualquier segmento de audio que se vaya a utilizar indistintamente para entrenar o evaluar deberá ser sometido previamente a un proceso de extracción de sus características. La técnica de extracción de parámetros que se ha empleado durante todo el proyecto es una de las más importantes y utilizadas a día de hoy en cualquier sistema de reconocimiento de voz y tiene como objetivo la obtención de los parámetros MFCC (Mel-Frequency Cepstral Coefficients). Dedicaremos todo este capítulo a analizar detenidamente todos los pasos necesarios para obtener dichos parámetros y adaptarlos correctamente para mejorar la eficiencia del sistema.

Los parámetros MFCC son un tipo particular de coeficientes cepstrales derivados de la aplicación del Cepstrum sobre una ventana de tiempo de la señal de voz. Analizando el Cepstrum desde un punto de vista matemático, podemos decir que se trata de un operador que transforma una convolución en el tiempo en una suma en el dominio espectral. De esta forma se consigue separar de una forma elegante las dos componentes de información de la señal de voz: la excitación y el tracto vocal.

En general, el Cepstrum se define como la transformada inversa de Fourier del logaritmo del espectro de la señal de voz, es decir:

$$Cepstrum(s[n]) = \hat{s}[n] = F^{-1}[log(|F[s[n]|)] = \frac{1}{2\pi}\int_{-\pi}^{\pi} log\left(\left|S(e^{j\omega})\right|\right)d\omega, \quad (3.1)$$

en donde s[n] representa la convolución entre la excitación y el tracto vocal:

$$s[n] = e[n] * h[n]. \quad (3.2)$$

Tal y como hemos afirmado anteriormente, se puede demostrar que después de la aplicación del Cepstrum, la convolución se transforma en una suma en el dominio cepstral (por esta razón se considera al Cepstrum una transformación homomórfica):

$$\hat{s}[n] = \hat{e}[n] + \hat{h}[n]. \quad (3.3)$$

Sin embargo, aunque casi todas las técnicas paramétricas de extracción de características de la señal de voz hacen uso del cepstrum, ésta raramente se utiliza directamente debido a su alta vulnerabilidad con los efectos del canal. Además, para mejorar la eficiencia del sistema, resultaría bastante útil intentar emular el comportamiento frecuencial del oído humano. De aquí surge el concepto de coeficientes MFCC que hacen uso de una nueva escala de frecuencia no lineal denominada MEL para imitar el comportamiento psicoacústico a tonos puros de distinta frecuencia dentro

del oído humano. De hecho, estudios dentro de esta ciencia han demostrado que el sistema auditivo humano procesa la señal de voz en el dominio espectral, caracterizándose por tener mayores resoluciones en bajas frecuencias y esto es precisamente lo que se consigue mediante la escala MEL, asignar mayor relevancia a las bajas frecuencias de forma análoga a como se hace en el sistema auditivo humano, en concreto en el oído interno.

3.1 MFCC (Mel-Frequency Cepstral coefficients)

La obtención de los coeficientes MFCC [S. B. Davis and P. Mermelstein, 1980] ha sido considerada como una de las técnicas de parametrización de la voz más importante y utilizada dentro del área de verificación de interlocutor. El objetivo de esta transformación es obtener una representación compacta, robusta y apropiada para posteriormente poder obtener un modelo estadístico del locutor con un alto grado de precisión. El esquema básico para la extracción de los MFCC aparece representado en la siguiente figura:

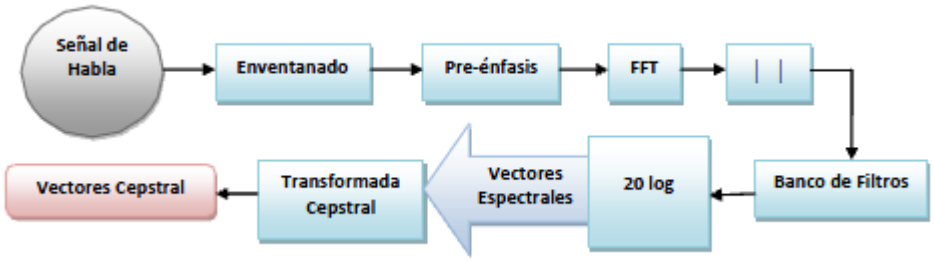

***Figura 8**. Diagrama de bloques de la parametrización mediante MFCC. Figura adaptada de [Eugenio Arévalo, 2011]*

Analizaremos con total detalle cada uno de los bloques que componen la figura 8 desde un punto de vista completamente teórico. En el capítulo 6, en el cual nos centraremos en el software y la base de datos necesarios para llevar a cabo las simulaciones, se hablará sobre los valores que hemos asignados a los parámetros en las pruebas para obtener los coeficientes cepstrales.

- **Pre-énfasis**. La señal de audio a parametrizar se pasa por un filtro de pre-énfasis. El objetivo es compensar la atenuación de aproximadamente 20 dB/década que se produce en el proceso fisiológico del mecanismo de producción del habla. Este paso es opcional pero altamente recomendable para enfatizar los formantes (frecuencias de resonancia de la cavidad acústica del tracto vocal) de alta frecuencia. La señal pre-enfatizada se obtiene a través del siguiente filtro:

$$y[n] = x[n] - ax[n-1], \tag{3.4}$$

donde a suele tomar valores en el intervalo [0.95, 0.98].

- **Enventanado**. La señal de voz es un proceso aleatorio y no estacionario. Esto supone un inconveniente a la hora de analizarla. No obstante, es posible salvar este problema si se tiene en cuenta que a corto plazo de tiempo (del orden de ms) la señal es casi-estacionaria. Este hecho da lugar a un tipo de análisis donde se obtienen segmentos o tramas de la señal de pocos ms denominado análisis localizado. A este proceso donde se generan tramas o segmentos consecutivos de señal se le denomina enventanado. Generalmente se suelen trabajar con ventanas de tipo Hamming y de Hanning, con un tamaño de 20 ms. Para mantener la continuidad de la información de la señal, es muy común realizar el enventanado con bloques de muestras solapados entre sí, de tal manera que no se pierde información en la transición entre ventanas. Generalmente el solapamiento se lleva a cabo con un desplazamiento entre ventanas de 10 ms, obteniéndose coeficientes MFCC cada 10 ms.

- Posteriormente al enventanado, se calcula la **DFT** de tamaño N de la señal enventanada aplicando:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk}, 0 \leq k \leq N. \tag{3.5}$$

A partir de este momento se descarta la fase y se trabaja con la envolvente de la señal de voz |X[k]|.

- **Filterbank**. La señal |X[k]| se multiplica por un banco de F filtros triangulares (de área unidad). Estos triángulos están espaciados de acuerdo con la escala de frecuencias MEL. Un ejemplo de este banco de filtros lo podemos ver en la siguiente figura:

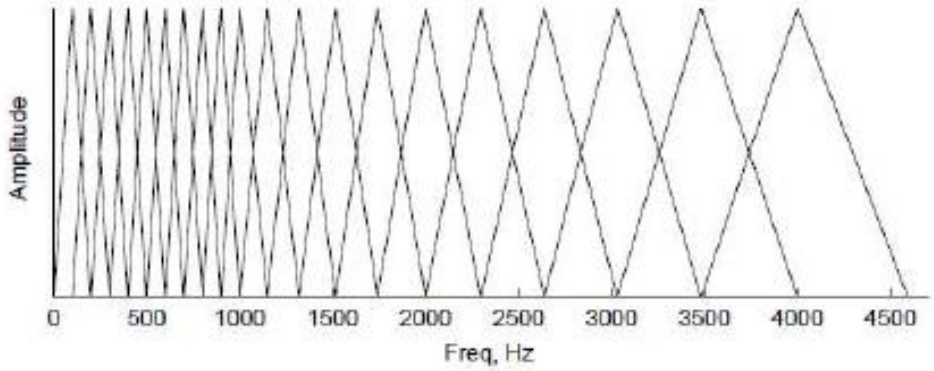

Figura 9. *Banco de filtros utilizado por [S. B. David and P. Mermelstein, 1980]*

Analizando la figura anterior, vemos que los puntos finales de cada filtro son definidos por las frecuencias centrales de los filtros adyacentes. En este caso concreto, el banco está formado por 20 filtros de los cuales, 10 están linealmente espaciados entre 100 y 1000 Hz, 5 logarítmicamente espaciados entre 1 y 2 kHz y otros 5 logarítmicamente espaciados entre 2 y 4 kHz.

El ancho de banda de los filtros triangulares viene determinado por la distribución de la frecuencia de centro de cada filtro, siendo ésta función de la frecuencia de muestreo y del número de filtros. Si el número de filtros en el banco aumenta, el ancho de banda de cada filtro decrece.

Además, tal y como se mencionó anteriormente, el comportamiento del sistema psico-acústico humano se aproxima mediante la escala de frecuencias MEL:

$$mel(f) = 2595 log_{10}\left(1 + \frac{f}{700}\right), \qquad (3.6)$$

donde f se corresponde con la frecuencia representada en el eje de escala líneal.

A continuación hemos representado la escala de frecuencias MEL con respecto a la lineal para ilustrar gráficamente la relación entre ambas:

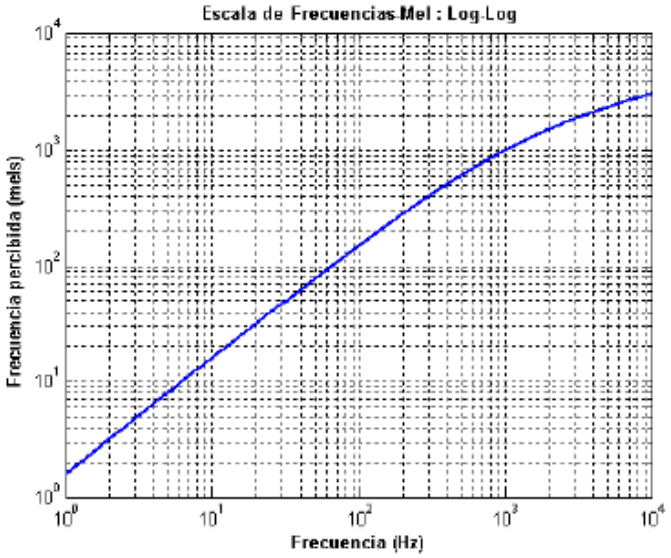

Figura 10. *Escala de frecuencias MEL*

Sin embargo, también resultaría bastante recomendable poder obtener una expresión matemática que nos permitiese calcular el valor de cada uno de estos filtros:

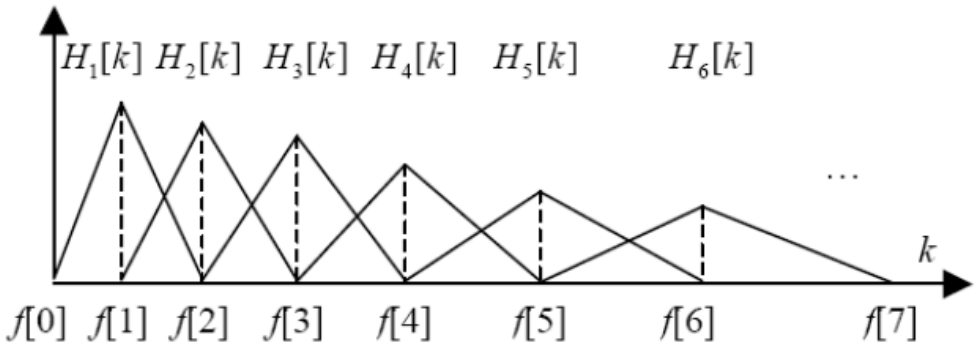

Figura 11. *Banco de filtros MEL*

$$H_m[k] = \begin{cases} 0 & , \quad k < f[m-1] \\ \dfrac{2(k - f[m-1])}{(f[m+1] - f[m-1])(f[m] - f[m-1])} & , \quad f[m-1] \le k \le f[m] \\ \dfrac{2(f[m+1] - k)}{(f[m+1] - f[m-1])(f[m] - f[m-1])} & , \quad f[m] \le k \le f[m+1] \\ 0 & , \quad k > f[m+1] \end{cases} \quad (3.7)$$

, donde $1 \leq m \leq F$, con F representando el número de filtros, y en donde además tenemos que:

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_1) + m\frac{B(f_h) - B(f_1)}{M+1}\right), \tag{3.8}$$

$$B^{-1}(b) = 700\left(e^{\frac{b}{2595}} - 1\right), \tag{3.9}$$

teniendo en cuenta que F_s es la frecuencia de muestreo. La función $f[m]$ determina los extremos de cada filtro triangular, siendo f_1 y f_h los extremos inferior y superior de cada uno de ellos.

- Una vez que la envolvente del espectro de la señal de voz es multiplicada por el banco, se calcula la **energía** correspondiente en cada uno de los filtros:

$$E_m = \sum_{k=0}^{N-1} |X[k]|^2 H_m[k] \qquad 1 \leq m \leq F. \tag{3.10}$$

Tras obtener la energía, debemos de calcularle el logaritmo, pasando por lo tanto al dominio de la potencia espectral logarítmica. El inconveniente de trabajar en este dominio es que los espectros de los filtros en las bandas adyacentes presentan un alto grado de correlación, originando coeficientes espectrales estadísticamente muy dependientes entre ellos.

- Para eliminar dicha dependencia o correlación estadística, se hace uso de la **Transformada Discreta del Coseno** (DCT), que lleva los coeficientes espectrales al dominio de la quefrencia convirtiéndolos en coeficientes cepstrales (MFCC):

$$c_{MFCC}[m] = \sum_{k=0}^{N-1} log(E_k)\, cos\left(m\left(k - \frac{1}{2}\right)\frac{\pi}{N}\right) \quad m = 1, \ldots, F. \tag{3.11}$$

3.2 MFCC-Delta y MFCC-Delta-Delta

Los MFCC representan la envolvente espectral de la señal de voz, obteniendo así importantes características identificadoras del habla. Por ejemplo, el primer coeficiente MFCC $c_{MFCC}[1]$ indica la energía de la señal, y se usa o no dependiendo de la aplicación. El segundo coeficiente $c_{MFCC}[2]$ tiene una razonable interpretación como indicador del balance global de la energía entre altas y bajas frecuencias.

Para obtener más información, como por ejemplo, la coarticulación de los fonemas, además de los MFCC se incluyen más coeficientes, que son los MFCC-Delta (o ΔMFCC) y los MFCC-Delta-Delta (o ΔΔMFCC) [Furui, 1981b], que representan la evolución temporal de los fonemas en su transición a otros fonemas, y en general permiten tener en cuenta la variabilidad de un interlocutor a la hora de hablar. Los ΔMFCC se le conocen como coeficientes de velocidad, ya que miden la variación de los coeficientes MFCC sobre un instante de tiempo. De la misma forma, a los ΔΔMFCC se les denomina coeficientes de aceleración, pues representan la variación de los ΔMFCC sobre un instante de tiempo. Si $c_{MFCCi}[m]$ representa la componente m del vector MFCC asociado a la trama o ventana i-ésima, la expresión que permite calcular el valor de los coeficientes dinámicos es la siguiente:

$$\begin{cases} \Delta c_{MFCCi}[m] = \sum_{k=-l}^{l} \dfrac{k c_{MFCC(i+l)}[m]}{|k|} \\ \Delta\Delta c_{MFCCi}[m] = \sum_{i=-l}^{l} \dfrac{k \Delta c_{MFCC(i+l)}[m]}{|k|} \end{cases} \quad 1 \leq m \leq F. \qquad (3.12)$$

l controla el número de tramas y consecuentemente, el intervalo de tiempo en el que la información dinámica está siendo calculada. El software de este proyecto trabaja con un valor de l correspondiente a 2.

Teniendo en cuenta que a lo largo de todo el proyecto haremos uso tanto de los coeficientes estáticos c_{MFCC} como de los dinámicos Δc_{MFCC} y $\Delta\Delta c_{MFCC}$, a partir de este momento, cuando hablemos de coeficientes cepstrales, implícitamente estaremos afirmando que tenemos un vector formado por estos tres tipos de coeficientes junto con tres componentes adicionales que representan sus correspondientes energías (sobre el tamaño de estos vectores se hablará más detalladamente en el capítulo 6). Así que a partir de este momento, $c_i[m]$ representa la componente m del vector de características asociado a la trama i-ésima, donde $1 \leq m \leq D$. Puesto que se tiene un coeficiente de velocidad y otro de aceleración por cada coeficiente MFCC y debido al hecho de que se añaden tres componentes adicionales correspondientes a las energías de los tres tipos de coeficientes (estático, velocidad y aceleración), los vectores cepstrales tendrán una dimensión de $D = 3F + 3$.

3.3 Speech Activity Detection (SAD)

En Verificación Automática de Locutor, cuando se analiza un segmento de entrenamiento o de evaluación, se trabaja con un número determinado de tramas en función de la duración de dicho segmento, y por cada trama, se obtiene su correspondiente vector MFCC junto con los deltas. Sin embargo, no todos estos vectores son útiles, puesto que muchos de ellos únicamente contienen silencio o ruido, por lo tanto, no aportan ninguna información útil a nuestro sistema.

Por esta razón, es necesario un mecanismo que nos permita detectar los vectores cepstrales menos energéticos que generalmente pertenecerán a tramos de silencio. En relación con el apartado anterior, se estableció que una de las *D* componentes cepstrales se corresponde con la energía de los coeficientes MFCC (estáticos). Dentro de este proyecto, las dos opciones que se van a barajar para la detección de energía [B. Fauve, 2009] parten del mismo principio, identificando a dicha componente como una variable aleatoria. El software con el que simularemos nuestros experimentos (ALIZE) representará dicha variable aleatoria como una combinación de 3 gaussianas. Tras procesar todas las tramas de un locutor, le asignará a cada una de estas tres componentes un peso determinado (la suma de los tres pesos debe ser 1). Dichos pesos dependerán del valor de la componente de energía de cada una de las tramas del locutor correspondiente. A partir de aquí, las dos técnicas intentarán obtener un umbral (de energía) que utilizaremos para aceptar o descartar las tramas del locutor cuya componente que representa la energía quede por debajo del umbral calculado.

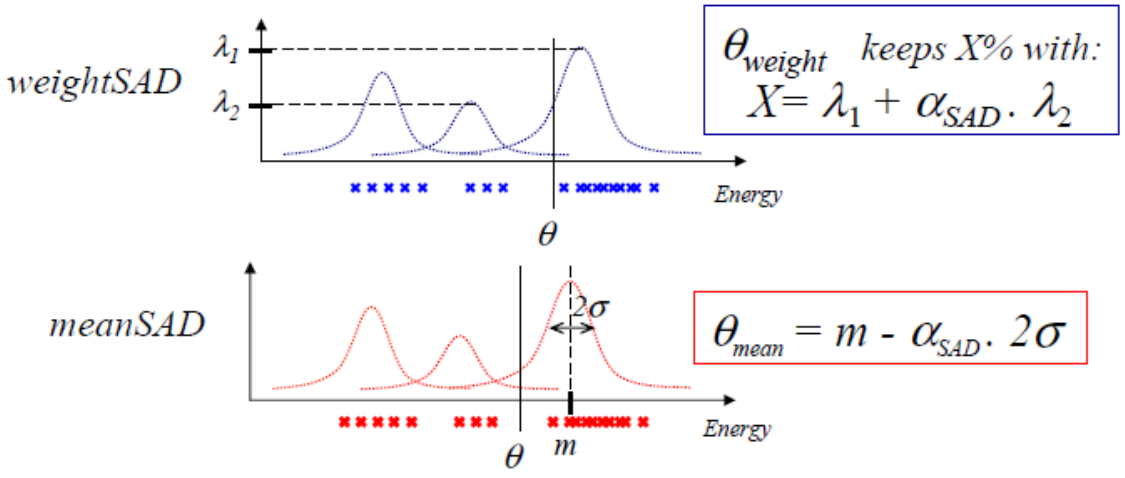

Figura 12. *Los dos ejemplos de actividad de detección de energía (SAD): weightSAD y meanSAD*

- **Weight SAD**. Aparece representado en la parte superior de la figura anterior. De las tres componentes gaussianas, seleccionamos la de mayor y menor peso,

calculando el máximo de sus correspondientes funciones de densidad de probabilidad, identificados en la figura mediante λ_1 y λ_2 respectivamente, lo cual nos permitirá obtener el siguiente parámetro:

$$X = \lambda_1 + \alpha_{SAD}\lambda_2. \tag{3.13}$$

Para calcular la expresión anterior, α_{SAD} debe ser definido (ya hablaremos sobre el valor del que hemos hecho uso en las simulaciones en los capítulos 6 y 7). El umbral seleccionado θ_{weight} será aquel que consiga que el X% de las energías correspondientes a todas las tramas del locutor queden por encima de dicho umbral.

- **Mean SAD** (parte inferior de la figura anterior). En este caso, seleccionamos la componente de mayor peso, calculando su media m y varianza σ:

$$\theta_{mean} = m - 2\alpha_{SAD}\sigma. \tag{3.14}$$

3.4 Normalización de los coeficientes cepstrales

Teniendo en cuenta que las diferentes condiciones de grabación de los segmentos de entrenamiento y evaluación, el ruido ambiente, el estado anímico del interlocutor en un momento determinado, las variaciones del canal, la variabilidad del interlocutor, el contenido lingüístico del mensaje, etc. son factores que pueden afectar en gran medida al rendimiento del sistema, sería bastante recomendable utilizar técnicas que modifiquen los vectores de características antes de realizar el entrenamiento o la evaluación, mitigando los efectos nocivos producidos por estas condiciones.

Dentro de este campo existen infinitas posibilidades, como RASTA (RelAtive SpecTrA), CMS (Cepstral Mean Substraction), CMVN (Cepstral Mean Variance Normalization) o Feature warping.

La técnica de normalización implementada en ALIZE es la correspondiente a CMVN [Koolwaaij and Boves, 2000]. Dado un segmento de audio proveniente de un locutor del que obtenemos sus vectores cepstrales, el procedimiento consiste en calcular la media y la varianza de cada una de las D componentes. A continuación, a cada uno de los vectores se le restará la media y luego se divide por su correspondiente varianza, tipificando por tanto cada una de las componentes cepstrales:

$$c_i{}^{CMVN}[m] = \frac{c_i[m] - \mu_m}{\sigma_m^2}, \tag{3.15}$$

$$\mu_m = \frac{1}{T}\sum_{i=1}^{T} c_i[m], \tag{3.16}$$

$$\sigma_m^2 = \frac{1}{T-1}\sum_{i=1}^{T} (c_i[m] - \mu_m)^2, \tag{3.17}$$

$$m = 1, \dots, D, \tag{3.18}$$

suponiendo un segmento de entrenamiento o evaluación con T tramas y vectores de características D-dimensionales.

De cualquier manera, para verificar los cambios en la distribución de los coeficientes cepstrales después de aplicar SAD y la correspondiente normalización, podemos observar las dos siguientes figuras:

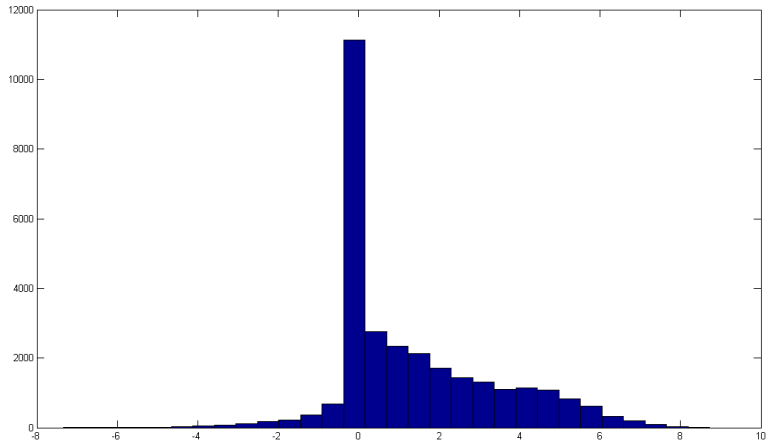

Figura 13. *Distribución de la primera componente cepstral de 2.5 minutos de conversación de un locutor antes de ser sometida a SAD*

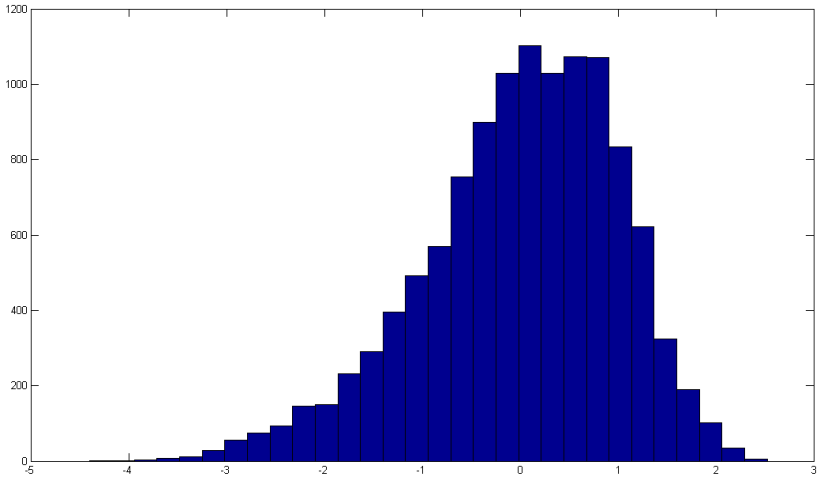

Figura 14. *Distribución de la primera componente cepstral de 2.5 minutos de conversación después de ser sometida a SAD y normalización mediante CMVN*

En la primera figura se ha representado la función de distribución de la primera componente de los vectores cepstrales de un segmento de voz de 2 minutos y medio correspondiente a un locutor antes de ser sometida a SAD, donde podemos observar que tiene un pico aproximadamente en el origen puesto que contienen tramas de silencio que todavía no han sido rechazadas. En cambio, la figura 14 muestra la distribución una vez aplicada SAD y su correspondiente normalización mediante la técnica CMVN. Podemos apreciar como finalmente su distribución se aproxima a la de una gaussiana o normal.