



Máster en
Tecnologías de la Información y Comunicación en
Redes Móviles

Ingeniería de Telecomunicación

Estudio de la parametrización de la señal de voz mediante modelado AM-FM para segmentación y clustering de locutores

Centro Politécnico Superior
UNIVERSIDAD DE ZARAGOZA
27 de agosto de 2009

Autor:
David Becerril Valle
Director:
Alfonso Ortega Giménez
Dr. Ingeniero de Telecomunicación

Resumen

Las señales de voz resonantes contienen información de modulación en amplitud y frecuencia. Mediante la utilización de operadores energía se consigue una demodulación eficiente. En esta tesis de máster se va a utilizar este método para obtener un conjunto de parámetros, los cuales han sido previamente utilizados para reconocimiento de fonemas, y que nosotros aplicaremos a la tarea de segmentación para la separación entre distintos fragmentos de audio, agrupación de los segmentos que tengan características comunes, y finalmente evaluaremos si estos parámetros nos aportan la información necesaria para mejorar el sistema de clustering de locutores y audio en general respecto a sistemas anteriores.

Agradecimientos

a Marta, mis padres y mi hermana,
siempre estáis ahí.

a Alfonso,
sus conocimientos, apoyo técnico y moral, dedicación y tesón durante estos meses han
sido esenciales para la consecución de este trabajo.

a Eduardo,
sus aportaciones se han visto reflejadas aquí.

a Carlos,
incansable y sufridor de mis dudas, compañero.

a Diego,
él re-abrió el baúl donde se encontraban los parámetros AM-FM.

a mis compañeros,
siempre han prestado la ayuda y el apoyo cuando era necesario.

There's *real* poetry in the *real* world. **Science is the poetry of reality.**

Richard Dawkins¹

¹The Enemies of Reason, "Slaves to Superstition"[1.01], 2007.

Índice general

1. Introducción	1
1.1. Contexto y Motivación	1
1.2. Precedentes	2
1.3. Objetivos	2
1.4. Estructura de la memoria	2
2. Metodología	5
2.1. Esquema del sistema de segmentación y clustering	5
2.2. Cálculo de los MFCC	6
2.3. Selección de tramas. VAD.	7
2.3.1. Selector de tramas	8
2.4. Segmentación	8
2.4.1. Segmentación basada en BIC	8
2.4.2. Segmentación basada en T^2	10
2.4.3. Ventajas y desventajas de BIC y T^2	11
2.5. Clustering	11
2.5.1. Clustering basado en BIC	11

2.5.2. Clustering basado en T^2	11
2.6. Evaluación del sistema	12
2.6.1. Medida de prestaciones: DER	12
2.7. Procedimiento del estudio	13
2.7.1. Base de datos para el test	13
2.7.2. Baseline	13
3. Extracción de características	15
3.1. Modelo AM-FM de la señal de voz	15
3.2. Algoritmo de separación de energía en tiempo discreto (DESA-1)	16
3.3. Parametrizador	17
3.3.1. Banco de filtros	17
3.3.2. Cálculo de DESA1 y características	19
3.3.3. Distribución estadística de las características	21
3.3.4. Regularización y normalización	22
3.4. Selección de parámetros	24
3.4.1. Matriz de covarianzas diagonal	24
3.4.2. PCA	26
4. Resultados de Segmentación y Clustering	29
4.1. Entramado sin solapamiento	29
4.2. Entramado con solapamiento	32
4.2.1. Prefiltrado multibanda	32
4.2.2. Postfiltrado multibanda	32
5. Conclusiones y líneas futuras	33
5.1. Conclusiones	33

5.2. Líneas futuras 34

Bibliografía **35**

Acrónimos **37**

Índice de figuras

2.1. Esquema general	5
2.2. Análisis tiempo-frecuencia	6
3.1. Esquema básico del parametrizador	17
3.2. Espectro filtro de gabor	18
3.3. Banco de filtros	20
3.4. Histograma de parámetros en f1	21
3.5. Histograma de parámetros, regularización	23
3.6. Separación ideal	25
3.7. Separación, mfcc	25
3.8. Separación, DESA	26
4.1. Esquema entramado. Prefiltrado	32
4.2. Esquema entramado. Postfiltrado	32

Índice de tablas

4.1. Resultados: MFCC-DESA	30
4.2. Resultados: Combinación MFCC-DESA	30
4.3. Resultados: λ	31
4.4. Resultados: DESA BIC- T^2	31
4.5. Resultados: PRE/POST filtrado	32

Capítulo 1

Introducción

En esta tesis se propone el estudio del comportamiento de un conjunto de parámetros sobre un sistema de segmentación y *clustering* obtenidos a partir de la demodulación de la señal de voz en amplitud y frecuencia. Este capítulo explicará el contexto y la motivación de la realización de la tesis, precedentes en los que se basa el trabajo, nuestros objetivos y la estructura de la memoria.

1.1. Contexto y Motivación

El trabajo ha sido realizado en el grupo de voz del departamento de Electrónica y Comunicaciones del Centro Politécnico Superior de la Universidad de Zaragoza. Dicho grupo tiene varias líneas de investigación, una de ellas es la identificación de eventos y transcripción enriquecida de señales de audio.

Un punto clave en este estudio es la separación y detección de fronteras en las señales, para etiquetar cada segmento y así poder identificar qué parte de la señal corresponde a un locutor o a otro, saber si es música, eventos acústicos aislados, ruido, etc.

Esta tarea no es trivial, y se están dedicando grandes cantidades de esfuerzo y recursos a buscar métodos para llevarla a cabo. El problema se está intentando atacar desde distintos puntos de partida, en el caso que nos atañe, buscando formas de parametrizar la señal de voz.

1.2. Precedentes

El uso del modelado AM-FM de la señal de voz surge de la necesidad de proporcionar información poco sensible al ruido y la variación del canal. Es decir, hacer frente a situaciones reales en las que la tarea de reconocimiento se hace muy difícil.

La parametrización mediante el modelado AM-FM de la señal de voz ha servido en trabajos anteriores para proporcionar robustez a la detección y reconocimiento de fonemas principalmente, aunque también se ha obtenido algún resultado para la tarea de verificación de locutor [JQR95, ER96].

Además, el estudio de segmentación y clustering se basará en un sistema implementando el criterio de información bayesiana (BIC), el cual nos proporcionará como salida la información sobre la relación entre cada segmento de la señal de voz y el locutor al que pertenece.

1.3. Objetivos

El principal objetivo de este proyecto es implementar un sistema de evaluación para realizar un estudio de segmentación y clustering mediante la extracción de nuevos parámetros de la señal de voz. Para ello, han de realizarse los siguientes hitos:

- Establecer un sistema de referencia (se basará en parametrización MFCC). Dicho sistema es un estándar y servirá para comparación de resultados con nuestro sistema propuesto.
- Extracción de parámetros a partir del modelo AM-FM. Existen varios métodos para la demodulación de una señal de AM-FM, de los cuales utilizaremos un método eficiente desarrollado por Teager y extendido por Kaiser. Para ello, utilizaremos un análisis multibanda de la señal.
- Análisis de la distribución de cada parámetro y compararla con los parámetros MFCC.
- Cálculo de la segmentación y clustering con los parámetros del modelo AM-FM, y comparación con la realizada con los parámetros MFCC.
- Propuesta de mejora y combinación de parámetros, para la posterior comparación de resultados de segmentación y clustering.
- Extracción de conclusiones y líneas futuras.

1.4. Estructura de la memoria

La memoria consta de las siguientes partes.

Introducción En este capítulo se introducirán los conceptos que dieron pie a la realización de esta tesis, el contexto, las motivaciones y los precedentes. Además se plantean los objetivos así como las subtarear a realizar. Para finalizar se comenta en este párrafo la estructura que tiene la memoria.

Metodología Descripción de cuál será la metodología a seguir para la obtención de los resultados y la consecución de los objetivos propuestos en la introducción.

Extracción de características Planteamiento de los conceptos básicos y el procedimiento para extraer las características de la señal de voz, así como el estudio de las distribuciones y el cálculo de los parámetros concretos.

Resultados de Segmentación y Clustering Explicación breve del proceso de segmentación mediante los parámetros obtenidos. Exposición de resultados.

Conclusiones y líneas futuras Extracción de las conclusiones por parte del autor, partiendo de los resultados. Se propondrán líneas futuras que servirán para una posible continuación del trabajo.

Capítulo 2

Metodología

En este capítulo se expondrá la metodología llevada a cabo para la consecución de los objetivos planteados en la introducción (ver 1.3)

2.1. Esquema del sistema de segmentación y clustering

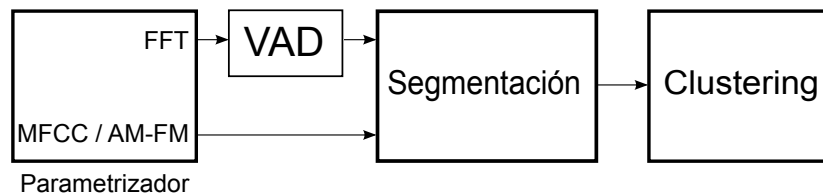


Figura 2.1: Esquema general

El sistema de selección de tramas a estudiar consta de cuatro bloques principales (fig. 2.1). El primer bloque realizará la parametrización de la señal. Dicho bloque incluye el análisis tiempo-frecuencia, se extraerá la FFT de cada trama (datos necesarios para el Voice Activity Detector (VAD)), se obtendrán los parámetros Mel-frequency cepstral coefficients (MFCC) [Sta00][Vaq08] y se calcularán los parámetros obtenidos a partir de la demodulación de la señal de voz, con el operador energía. En la sección 2.2 se describe brevemente la forma de calcular los MFCC, mientras que los coeficientes derivados de la demodulación AM-FM van a ser explicados con más detalle en el capítulo 3, especialmente dedicado a la descripción del Discrete Energy Separation Algorithm (DESA) y sus características.

En el bloque del VAD, se obtienen las tramas en las que hay actividad vocal, a partir de la FFT calculada en el bloque anterior.

La segmentación realiza la separación mediante las características obtenidas tanto con MFCCs como el bloque de demodulación AM-FM. El último bloque (clustering) agrupará los segmentos.

2.2. Cálculo de los MFCC

En el parametrizador se obtienen los MFCC de la señal a analizar. Se sigue el estándar ETSI de extracción de coeficientes MEL Cepstrum. Éstos coeficientes son una representación de la energía a corto plazo de la señal de audio, basada en la transformada lineal del coseno, de la log-potencia espectral en la escala MEL de frecuencias [Sta00].

De estas transformaciones se conservará cierta información de utilidad, como puede ser la transformada localizada de Fourier, necesaria para la selección de tramas o segmentación. El esquema del procesado tiempo-frecuencia es el de obtención de los coeficientes Mel-Cepstrum (MFCC) que se muestra en la figura 2.2.

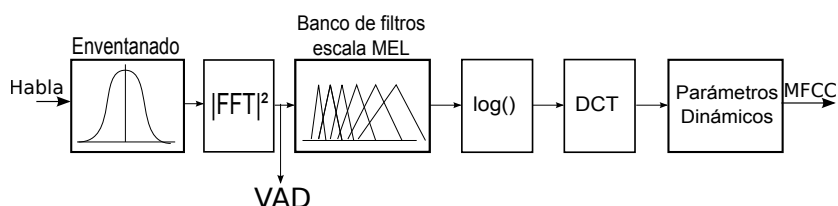


Figura 2.2: Estructura de proceso del análisis tiempo-frecuencia y extractor de parámetros acústicos.

Se utiliza una estructura clásica de extracción de parámetros acústicos, detallada a continuación:

Eventanado Para la parametrización, se utiliza un eventanado temporal de la señal, comúnmente mediante la ventana Hamming de duración 20 – 25ms, y un desplazamiento de 10ms.

FFT Una vez eventanada la señal, se realiza la Fast Fourier Transform (FFT) de ese fragmento y se guarda su módulo al cuadrado, es decir, se guarda la densidad espectral de potencia de cada trama $S_x = |FFT|^2$.

Banco de filtros A continuación se obtendrán las log-energías de cada uno de los 24 filtros del banco de filtros en escala perceptual MEL, adaptado a la frecuencia de trabajo¹.

Transformada discreta del coseno Para finalizar se realiza la Discrete Cosine Transform (DCT) para obtener los coeficientes MEL cepstrum a partir de las log-energías del banco de filtros.

¹Para el teléfono será de 8KHz

Los coeficientes MFCC son los parámetros estándar y más utilizados en el ámbito de procesado de señal de voz.

2.3. Selección de tramas. VAD.

El objetivo fundamental de la selección de tramas es discriminar entre aquellas que contengan información de interés para modelar y obtener las características presentes en el audio a tratar, y aquellas tramas que, bien por su baja energía o por su estadística, podrían confundir al sistema o nos aportan información errónea.

El VAD utilizado está basado en la estimación de la divergencia espectral a largo plazo (Long-Term Spectral Divergence estimation (LTSD)) [RSB⁺04]. Para calcular la LTSD es necesario conocer la envolvente espectral a largo plazo (Long-Term Spectral Envelope (LTSE)) y una estimación del ruido. La expresión para calcular la LTSD en función del número de trama, la LTSE y la estimación del ruido es la siguiente:

$$LTSD(n) = 10 \log_{10} \left(\frac{2}{N_{FFT}} \sum_{k=0}^{\frac{N_{FFT}}{2}-1} \frac{LTSE^2(k, n)}{N^2(k, n)} \right) \quad (2.1)$$

Donde n es el número de trama actual, N_{FFT} es el número de muestras de la FFT, $LTSE^2(k, n)$ es el cuadrado de la envolvente espectral a largo plazo en la frecuencia k de la FFT para la trama n y $N^2(k, n)$ es el cuadrado de la estimación del ruido en la frecuencia k de la FFT para la trama n .

La LTSE se define así:

$$LTSE(k, n) = \max \{Y(k, n + j)\}_{j=-N}^{j=N} \quad (2.2)$$

Siendo $Y(k, n + j)$ el valor de la FFT de la señal enventanada en la frecuencia k y la trama $n + j$, con $-N \leq j \leq N$, $j \in \mathbb{Z}$. N es el parámetro del VAD, el cual proporciona el control sobre la longitud de la ventana en la que se realiza la estimación de la envolvente espectral, ya que ésta toma, para cada frecuencia, el valor máximo de dicha frecuencia en una ventana de análisis de tamaño $2N + 1$ (de la ecuación (2.2)).

El método utilizado es, en definitiva, un seguimiento de máximos por la representación tiempo-frecuencia de la señal de entrada.

Por otra parte, la estimación de la potencia de ruido en la frecuencia k y la trama n en nuestro caso se realiza mediante la siguiente expresión:

$$N^2(k, n) = \alpha N^2(k, n - 1) + (1 - \alpha) \min \{Y^2(k, n), N_{st}^2(k, n)\} \quad (2.3)$$

Donde α es un factor de memoria que define la velocidad de adaptación de la estimación a las variaciones de ruido y la estabilidad de la estimación. Se recomienda utilizar valores de $\alpha \approx 0,1$. $Y^2(k, n)$ es el valor de la densidad espectral de potencia de la señal en la frecuencia k para la trama n y $N_{st}^2(k, n)$ es una estimación de potencia de ruido a corto plazo, definida de la siguiente manera:

$$N_{st}^2(k, n) = \begin{cases} \lambda N_{st}^2(k, n-1) + (1-\lambda)Y^2(k, n) & \text{Si la decisión } n-1 \text{ fue negativa} \\ \lambda N_{st}^2(k, n-1) & \text{Si la decisión } n-1 \text{ fue positiva} \end{cases} \quad (2.4)$$

Se pretende que esta estimación a corto plazo del ruido se adapte rápidamente a la estadística del ruido presente en la señal, de forma que se suelen tomar factores de memoria λ más bajos que para el caso de la estimación del ruido a largo plazo, en torno a 0,6. Como puede apreciarse en la ecuación (2.4), la estadística del ruido se actualiza en cuanto se decide que la trama no es de interés, y se mantiene inalterada mientras se detecte señal de interés.

2.3.1. Selector de tramas

Una vez calculado el LTSD, se procede a realizar la selección de tramas, en función de un umbral, de forma que si el LTSD en la trama n supera ese umbral, la trama será pasada al siguiente bloque, y si está por debajo, la trama será descartada:

$$\text{Decisión}(n) = \begin{cases} 1 & \text{LTSD}(n) \geq \text{umbral} \\ 0 & \text{LTSD}(n) < \text{umbral} \end{cases} \quad (2.5)$$

2.4. Segmentación

El objetivo de la segmentación es dividir la secuencia de audio en segmentos acústicamente homogéneos, para poder segregar los segmentos según características o patrones comunes. Una aplicación podría ser distinguir los segmentos asociados al locutor de interés frente a aquellas tramas relacionadas con eventos, ruidos transitorios y locutores esporádicos. Para realizar la segmentación, se utiliza el Criterio de Información Bayesiana (BIC).

2.4.1. Segmentación basada en BIC

Este método de segmentación, (Criterio de información Bayesiana), pretende dar una medida de validez de M , un modelo de una secuencia $\underline{x} = \{x_1, x_2, x_3, \dots, x_N\}$ de N

observaciones [CG][ZH05]. El BIC penaliza la complejidad del modelo y se define de la siguiente forma:

$$BIC(M) = \log P(x_1, x_2, x_3 \dots, x_N | M) - \frac{1}{2} d \log N \quad (2.6)$$

Donde d es la dimensión de los vectores de observaciones.

Para utilizar este criterio en la tarea de segmentación, se toma una ventana de análisis de un tamaño inicial suficientemente pequeño como para suponer que existen como máximo dos segmentos acústicamente homogéneos en dicha ventana. Dicha ventana de análisis puede tener, por ejemplo, una duración de 2 segundos. A continuación, se plantean dos hipótesis:

1. Toda la ventana constituye un segmento acústicamente homogéneo H_0 .
2. Dicha ventana contiene dos segmentos acústicamente distintos H_1 , y por tanto existe un punto de cambio acústico en dicho segmento.

Si asumimos que cada segmento acústicamente homogéneo puede modelarse con una gaussiana de media μ y matriz de covarianza completa Σ $N(\mu, \Sigma)$, lo anterior se puede expresar de la siguiente manera:

1. $H_0 : (x_1, x_2, x_3 \dots, x_N) \sim N(\mu_0, \Sigma_0)$. El segmento en la ventana de estudio es acústicamente homogéneo.
2. $H_1 : (x_1, x_2, x_3 \dots, x_b) \sim N(\mu_1, \Sigma_1)$ y $(x_{b+1}, x_{b+2}, x_3 \dots, x_N) \sim N(\mu_2, \Sigma_2)$. El segmento en la ventana de estudio contiene dos segmentos acústicamente homogéneos y distintos, y el punto de cambio de un segmento a otro es la trama b .

Con lo que el BIC para ambas hipótesis se calcula así:

$$BIC_0 = -\frac{d}{2} N \log 2\pi - \frac{N}{2} \log |\Sigma_0| - \frac{N}{2} d - \frac{\lambda}{2} \left(d + \frac{1}{2} d(d+1) \right) \log N \quad (2.7)$$

$$BIC_1 = -\frac{d}{2} N \log 2\pi - \frac{b}{2} \log |\Sigma_1| - \frac{N-b}{2} \log |\Sigma_2| - \frac{N}{2} d - \lambda \left(d + \frac{1}{2} d(d+1) \right) \log N \quad (2.8)$$

Donde λ es la penalización por complejidad del modelo, de forma que cuanto mayor sea, mayor es la penalización sobre la hipótesis H_1 frente a la hipótesis H_0 .

El sistema de decisión se decanta por la hipótesis que mayor BIC ofrece, de forma que se calcula:

$$\Delta\text{BIC}(b) = \text{BIC}_1 - \text{BIC}_0 > 0 \Rightarrow \text{Punto de corte en la trama } b \quad (2.9)$$

Este valor se calcula para distintos puntos de corte b en el segmento, Por ejemplo, se analizan tramas equiespaciadas, tomando únicamente una de cada diez, para aumentar la velocidad del algoritmo. Si el resultado nos indica que no existe punto de corte en la ventana de análisis, se amplía su tamaño y se repite el proceso. En caso contrario, se marca como “cambio” y la ventana de análisis se desplaza hasta la trama siguiente a la detectada ($b + 1$).

2.4.2. Segmentación basada en T^2

Otro método de segmentación que permite segregar segmentos muy cortos, sin necesidad de estimar una matriz de covarianzas en segmentos pequeños, es la estadística Hotelling T^2 . Este método de segmentación realiza las mismas hipótesis que el BIC, utilizando el procedimiento de ventana de análisis, pero al contrario que el BIC, los dos segmentos de audio que resultan al suponer cada posible punto de corte b , se modelan con gaussianas con idéntica matriz de covarianzas, pero distinta media, de forma que se estima la matriz de covarianzas sobre todo el segmento bajo test y no sobre los segmentos a ambos lados del punto de corte, de los que únicamente hay que estimar su media.

La medida sobre la que se decide si el posible punto de corte b separa dos segmentos distintos es:

$$T^2 = \frac{b(N-b)}{N} (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) \quad (2.10)$$

Con μ_i la media de cada modelo y Σ la matriz de covarianzas sobre todo el segmento.

Y el criterio de decisión será establecer un umbral γ tal que:

$$\text{Punto de corte}(b) = \begin{cases} \text{Sí} & \text{si } T^2 > \gamma \\ \text{No} & \text{si } T^2 \leq \gamma \end{cases} \quad (2.11)$$

Es decir, si T^2 supera dicho umbral, se decidirá que existe un punto de corte en b .

2.4.3. Ventajas y desventajas de BIC y T^2

A pesar de que el BIC sea más robusto generalmente, el método de T^2 se utiliza principalmente cuando los segmentos, o la ventana de análisis es suficientemente pequeña para que la estimación de la matriz de covarianzas no se pueda realizar.

El BIC tiene como ventaja la robustez, y es que será más preciso ya que tenemos más parámetros de control, es decir, en cada ventana a analizar cada fragmento se representará con dos modelos distintos, covarianzas completas para cada fragmento por separado, al contrario que T^2 , en el que lo único variable entre hipótesis es la media del modelo tal y como se ha visto en apartados anteriores.

Una de los inconvenientes del T^2 es también su sensibilidad frente al umbral. Será complicado ajustar bien este parámetro y variará según los archivos a analizar.

2.5. Clustering

La última etapa en el diagrama de bloques de la figura 2.1 tiene como objetivo agrupar los segmentos obtenidos en la fase de segmentación en clases acústicamente homogéneas. Con esto se consigue tener todos los segmentos bien sean de un mismo locutor, música de una misma canción, un ruido determinado, etc, identificados y agrupados. Para ello se utilizan técnicas de *clustering* jerárquico, de forma que se irán agrupando iterativamente sólo los segmentos más próximos, hasta que el criterio de *clustering* decida que no se deben agrupar más segmentos.

2.5.1. Clustering basado en BIC

El BIC puede utilizarse para realizar *clustering* de segmentos de audio de forma muy similar a como se utiliza para la tarea de segmentación. El método de decisión para plantear que dos segmentos son homogéneos es evaluar el BIC de ambos segmentos conjuntamente y compararlo con el BIC de los mismos segmentos por separado. Si dicho valor es mayor que el obtenido al modelar cada segmento por separado, se consideran segmentos homogéneos.

Por supuesto, únicamente tiene sentido aplicar técnicas de *clustering* en aquellos archivos en los que se hayan detectado segmentos acústicamente distintos. Por tanto, se analizarán dichos archivos, en base a la segmentación obtenida por el BIC.

2.5.2. Clustering basado en T^2

Para decidir si dos segmentos son homogéneos, utilizando el método de T^2 , se evalúa dicho método para los dos segmentos conjuntamente y se compara con el umbral. Si el valor de T^2 está por debajo de ese umbral, se considerarán segmentos homogéneos.

Al igual que en el caso anterior, solo se aplicarán métodos de *clustering* en aquellos archivos en los que se hayan detectado segmentos acústicamente distintos.

2.6. Evaluación del sistema

Para un correcto análisis y mejora de un sistema de segmentación y clustering se deben establecer unos métodos para la evaluación de las prestaciones del mismo. Dichos métodos serán, en general, unas medidas del error cometido y dependerán de las características utilizadas en mayor medida.

2.6.1. Medida de prestaciones: DER

La métrica que nos va a servir para medir las prestaciones del sistema va a ser la tasa de error de diarización Diarization Error Rate (DER). Esta medida se ha definido en las evaluaciones del NIST (NIST Fall Rich Transcription on meetings 2006 Evaluation Plan, 2006 [NIS06]). Se mide como la fracción de tiempo que no se asigna correctamente a un locutor o a segmentos que no son habla.

Para la definición de la tarea, el sistema que realiza las hipótesis de diarización no necesita identificar a los locutores por ID o nombre definido, por lo tanto, las etiquetas de los locutores asignados en las hipótesis y en las referencias de segmentación no tienen que ser necesariamente las mismas. Al contrario que las etiquetas de “no-habla”, las cuales se marcan como saltos no etiquetados entre dos segmentos de locutor, con lo que tanto las etiquetas de referencia como las de hipótesis deben coincidir.

El script de evaluación calcula el emparejamiento óptimo para las etiquetas de los locutores entre las hipótesis y los archivos de referencia. Esto nos permite puntuar diferentes IDs entre dos archivos. Finalmente, la tasa de error de diarización se calcula de la siguiente forma [Mir06]:

$$DER = \frac{\sum_{s=1}^S dur(s) \cdot (\max(N_{ref}, N_{hyp}(s)) - N_{correct}(s))}{\sum_{s=1}^S dur(s) \cdot N_{ref}} \quad (2.12)$$

Donde S es el número total de segmentos de locutor donde tanto la referencia y las hipótesis contienen las mismas parejas de locutores. Se calcula agrupando de forma alterna las hipótesis y referencias de los locutores.

Los términos $N_{ref}(s)$ y $N_{sys}(s)$ indican el número de locutores hablando en el segmento s , y $N_{correct}(s)$ indica el número de locutores que hablan en el segmento s y que han sido correctamente emparejados en la hipótesis con la referencia. Los segmentos etiquetados como “no-habla” se consideran “hablante 0”. Cuando todos los locutores/no-habla en un segmento son aciertos, el error para ese segmento es 0.

2.7. Procedimiento del estudio

El procedimiento seguido para la implementación de las diversas técnicas que contribuyan a realizar la segmentación y clustering ha consistido en la definición del material a utilizar para la fase de entrenamiento y testeo del sistema, la definición de las medidas de evaluación que representarán las prestaciones del mismo, la definición del sistema de partida o baseline, y finalmente la implementación y comprobación de las prestaciones del sistema con las características propuestas.

2.7.1. Base de datos para el test

La base de datos utilizada es la *Switchboard* [TI06]. Esta base de datos es un corpus de conversaciones espontáneas que abarcan la creciente necesidad de bases de datos multi-locutor con las características de ancho de banda telefónico. Switchboard ha sido grabada en la sede de Texas Instruments con financiación a cargo de DARPA. El set completo consta de 2430 conversaciones de 6 minutos cada una en media. En total, alrededor de 240 horas de habla grabada y unas 3 millones de palabras de texto, hablado por 500 locutores de ambos sexos abarcando los dialectos más extendidos de Inglés Americano. En cada conversación participan únicamente dos locutores.

Además esta base de datos tiene unas características únicas, diseñadas para apoyar el desarrollo de tecnología focalizada en habla telefónica, así como investigación básica sobre habla y lenguaje conversacional espontáneos.

Primeramente, la base de datos fue grabada sin intervención humana, bajo control automático. La interacción con los sistemas fue vía tonos e instrucciones de grabación, pero los hablantes, una vez conectados, podrían “calentar” antes de que comenzaran las grabaciones.

Desde el punto de vista del factor humano, la automatización nos protege contra la posible intrusión del experimentador y su sesgo, y nos garantiza un grado de uniformidad a lo largo del periodo de adquisición de datos. La intención fue favorecer el lenguaje natural y espontáneo a los participantes. Los transcripores valoraron las grabaciones como conversaciones con un grado altísimo de naturalidad.

2.7.2. Baseline

Nuestro punto de partida, ha sido la realización del análisis de la base de datos Switchboard, a partir de la parametrización MFCC de las señales. El mejor resultado calculado y con el que se va a comparar el rendimiento de los parámetros derivados del algoritmo de separación de energía discreto (DESA) es un error DER igual al 19%. Este resultado ha sido obtenido con los 13 primeros coeficientes MFCC, el valor típico para las tareas de procesado de señal de voz [ER96], [Vaq08].

Capítulo 3

Extracción de características

En este capítulo se analiza y describe la forma de extracción de parámetros a partir de la demodulación AM-FM de la señal de voz.

3.1. Modelo AM-FM de la señal de voz

El método propuesto para el estudio de segmentación y clusterización de señales de audio, concretamente de voz, es el modelado AM-FM de la señal de voz mediante la demodulación por sub-bandas. La utilización de este modelo para señales de voz es adecuado ya que con él se representan las partes no lineales así como sus variaciones.

La información que se va a extraer de la señal va a ser la modulación en amplitud (AM) y la modulación en fase (FM) de la forma:

$$s(t) = a(t) \cos[\phi(t)] \quad (3.1)$$

Para una extracción eficiente de la información AM-FM se va a utilizar el operador energía desarrollado por Teager [Tea80] e introducido por Kaiser [Kai90], descrito por la siguiente expresión:

$$\Psi(s) = (\dot{s})^2 - s\ddot{s} \quad (3.2)$$

Donde \dot{s} es la primera derivada de la señal y \ddot{s} es la segunda derivada de la señal s .

Además, para señales AM-FM del tipo (3.1) tenemos que la relación del operador energía con la amplitud y la frecuencia se corresponde con las siguientes ecuaciones,

$$\Psi(s) \approx a^2(t)\omega_i^2(t) \quad (3.3a)$$

$$\Psi(\dot{s}) \approx a^2(t)\omega_i^4(t) \quad (3.3b)$$

Esto motivó el algoritmo de separación de energía ESA (energy separation algorithm), el cual da una estimación de la amplitud y la frecuencia instantáneas en función del operador energía Teager aplicado a la señal. La formulación del algoritmo se puede escribir así[BMQ93],

$$\hat{a}^2(t) = \Psi^2(s)/\Psi(\dot{s}) \quad (3.4a)$$

$$\hat{\omega}_i^2(t) = \Psi(\dot{s})/\Psi(s) \quad (3.4b)$$

3.2. Algoritmo de separación de energía en tiempo discreto (DESA-1)

El algoritmo implementado en esta tesis para realizar la demodulación multibanda de la señal de voz es el DESA-1. Este algoritmo propuesto por Maragos en [MKQ93], está inspirado en el caso del coseno con amplitud y frecuencia constante. El operador energía Teager-Kaiser en tiempo discreto es:

$$\Psi[x(n)] = x^2(n) - x(n-1)x(n+1) \quad (3.5)$$

Para señales AM-FM, la amplitud y frecuencia instantánea se calculan de acuerdo a las siguientes expresiones [MKQ93]:

$$|a(n)| \approx \sqrt{\frac{\Psi[x(n)]}{1 - \left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}\right)^2}} \quad (3.6)$$

$$\Omega_i(n) \approx \arccos\left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}\right) \quad (3.7)$$

Donde $y(n)$ es la derivada en tiempo discreto:

$$y(n) = x(n) - x(n-1) \quad (3.8)$$

3.3. Parametrizador

El parametrizador es el bloque fundamental y en nuestro caso tenemos dos parametrizadores que generan características diferentes. El primero, como se ha visto en la sección 2.1, obtiene los coeficientes MFCC y la FFT.

El segundo, y el principal en nuestro estudio, es el que genera las características a partir de la demodulación AM-FM de la señal, que luego serán estudiados por los bloques de segmentación y clusterización. Además, se ha de llevar a cabo un buen ajuste del sistema y una regularización de los parámetros, para conseguir una entrada comparable a los parámetros del “baseline”.

El esquema básico para el cálculo de los parámetros es el de la figura 3.1 y consta principalmente de tres etapas:

1. Banco de filtros (4 filtros)
2. Cálculo de amplitud y frecuencia instantánea (DESA1)
3. Cálculo de 5 parámetros, derivados de la amplitud y frecuencia instantánea.

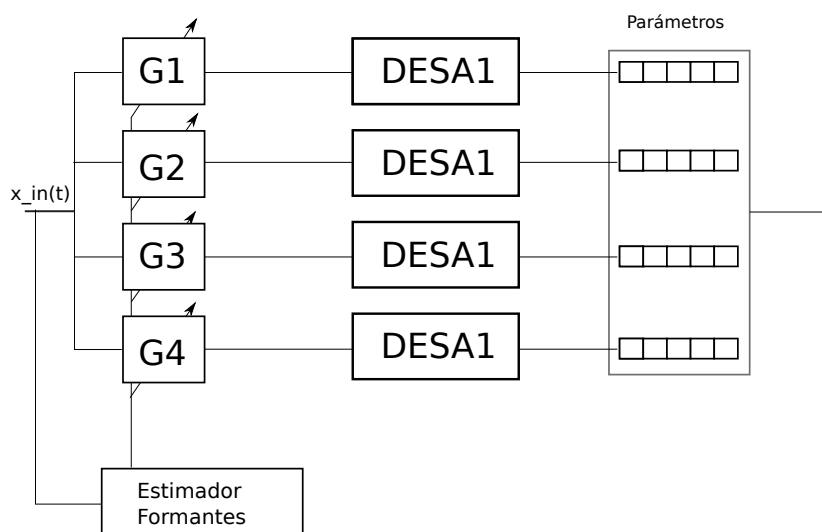


Figura 3.1: Esquema básico del parametrizador

A continuación se describe cada bloque por separado.

3.3.1. Banco de filtros

Nuestro desarrollo se centra en señales de voz con un ancho de banda limitado, como máximo, a 16 kHz. Maragos [MKQ93] introdujo el análisis mediante banco de

filtros de Gabor. El banco de filtros utilizado consta de cuatro filtros G_{1-4} . Dichos filtros tienen Q constante y un ancho de banda restringido a la octava parte de la frecuencia de muestreo para que las aproximaciones realizadas en 3.2 sigan siendo válidas.

La expresión del filtro es la de la ecuación 3.9 y en la figura 3.2 se puede ver un ejemplo de la respuesta frecuencial.

$$h(x) = e^{-\frac{x^2}{\sigma^2}} \cos(\omega x) \quad (3.9)$$

En la expresión del filtro σ representa la desviación estándar de la máscara gaussiana. El ancho de banda será $1/\sigma$ y ω es la frecuencia central del filtro. Tal como se ha dicho, el banco de filtros es de Q constante, con lo que σ va a ser elegida tal que $\sigma = Q/f$, siendo $f = \omega/2\pi$. Para nuestra aplicación se ha considerado una $Q = 1,5$, la cual nos permite restringir el ancho de banda al máximo permitido para que las aproximaciones del operador energía sean válidas. En las situaciones en las que el ancho de banda excede el máximo, se limitará, perdiendo la característica de conservar constante el factor de calidad. Esto ocurre ocasionalmente cuando la frecuencia estimada para el último filtro (G_4) es demasiado elevada, cercana a los $4kHz$, y la frecuencia de muestreo son $8kHz$.

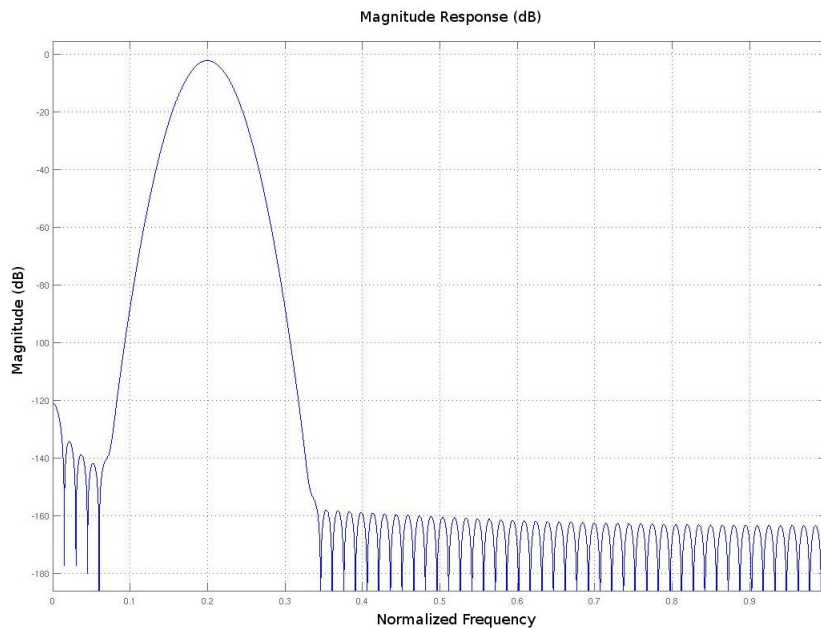


Figura 3.2: Espectro filtro de gabor

Como una segunda aproximación, los filtros fueron diseñados con un ancho de banda constante, igual a $400Hz$, más que suficiente para satisfacer las suposiciones realizadas en las aproximaciones de la formulación del DESA-1.

Estimación de la frecuencia central

La frecuencia central de los filtros se elige a partir de la frecuencia estimada de cada formante por trama. Dicha estimación es realizada a partir de las raíces del filtro predictor (LPC).

La forma de representar el predictor lineal es la siguiente:

$$\hat{x}(n) = - \sum_{i=1}^P a_i x(n-i) \quad (3.10)$$

Donde $\hat{x}(n)$ es el valor predicho de la señal, $x(n-i)$ son los valores observados anteriormente y a_i son los coeficientes del predictor. P es el orden del predictor.

El error generado por el estimador es:

$$e(n) = x(n) - \hat{x}(n) \quad (3.11)$$

Donde $x(n)$ es la señal real.

Los coeficientes del predictor lineal se calculan iterativamente mediante la implementación del algoritmo Levinson-Durbin. A partir de los coeficientes a_i se calcularán sus raíces para conseguir la frecuencia estimada de los formantes.

Puesto que la estimación es mejor para los dos primeros formantes, la frecuencia instantánea f_i y la amplitud instantánea a_i tendrán una mayor exactitud para estos dos casos.

Ejemplo de banco de filtros

En la figura 3.3 se puede ver un banco de filtros para una trama concreta, que a modo de ejemplo se han centrado las frecuencias en 400, 1000, 2000 y 3200 Hz respectivamente para cada uno de los cuatro formantes estimados.

3.3.2. Cálculo de DESA1 y características

Una vez tenemos las cuatro señales de cada formante como resultado del filtrado de la señal original por el banco de filtros, aplicamos el algoritmo DESA1 para calcular la amplitud y frecuencia instantáneas.

Los parámetros que van a obtener para configurar los vectores de características serán calculados por trama, de tamaño de 10 milisegundos:

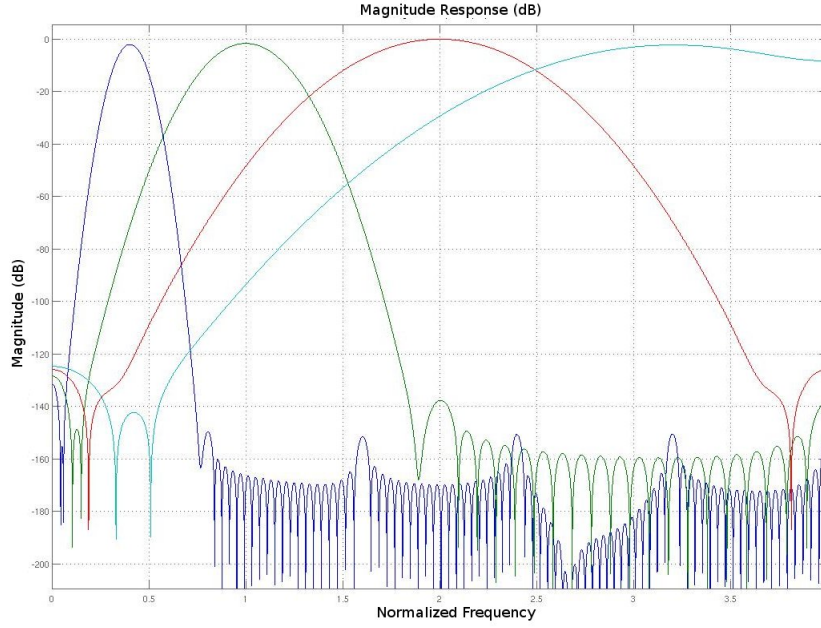


Figura 3.3: Banco de filtros. $Q = 1,5$, $G_1 : f_c = 400Hz$; $G_2 : f_c = 1000Hz$; $G_3 : f_c = 2000Hz$; $G_4 : f_c = 3200Hz$

- Amplitud y la frecuencia instantánea media por trama (IA_{mean} y IF_{mean})
- Porcentaje de modulación FM (FM-modulation percentages (FMP))
- Frecuencia ponderada por la media de la amplitud (F_i)
- Ancho de banda medio ponderado por la frecuencia instantánea (B_i)

Las ecuaciones para los parámetros anteriores son las que aparecen en (3.12)

$$IA_{mean} = \frac{\sum_N a_i}{N} \quad (3.12a)$$

$$IF_{mean} = \frac{\sum_N f_i}{N} \quad (3.12b)$$

$$F_i = \frac{\sum_{k=0}^N f_i[k] a_i^2[k]}{\sum_{k=0}^N a_i^2[k]} \quad (3.12c)$$

$$B_i = \frac{\sum_{k=0}^N [\dot{a}_i^2[k] + (f_i[k] - F_i)^2 a_i^2[k]]}{\sum_{k=0}^N a_i^2[k]} \quad (3.12d)$$

$$FMP_i = B_i / F_i \quad (3.12e)$$

3.3.3. Distribución estadística de las características

Como primer análisis y para observar la apariencia que poseen cada uno de los parámetros, se realizaron histogramas de cada uno de ellos, para archivos arbitrarios.

En la figura 3.4 se representan los histogramas de los distintos parámetros para los dos primeros formantes. Se puede observar que las distribuciones presentan una forma unimodal como puede ser el caso de la frecuencia instantánea media, F_i y B_i en las figuras 3.4(b), 3.4(c) y 3.4(d) respectivamente, y en otros casos como en la amplitud media instantánea nos encontramos con lo que parecen dos modos (fig. 3.4(a)).

Además, se puede ver, aunque con dificultad, que estas distribuciones presentan valores anómalos, alejados del rango en el que caen la mayoría de los valores. Por ejemplo, en torno a -10 en la amplitud 3.4(a), al cero en IF_{mean} , fig. 3.4(b), -6 y 12 en F_i 3.4(c) y cerca de 24 en B_i 3.4(d).

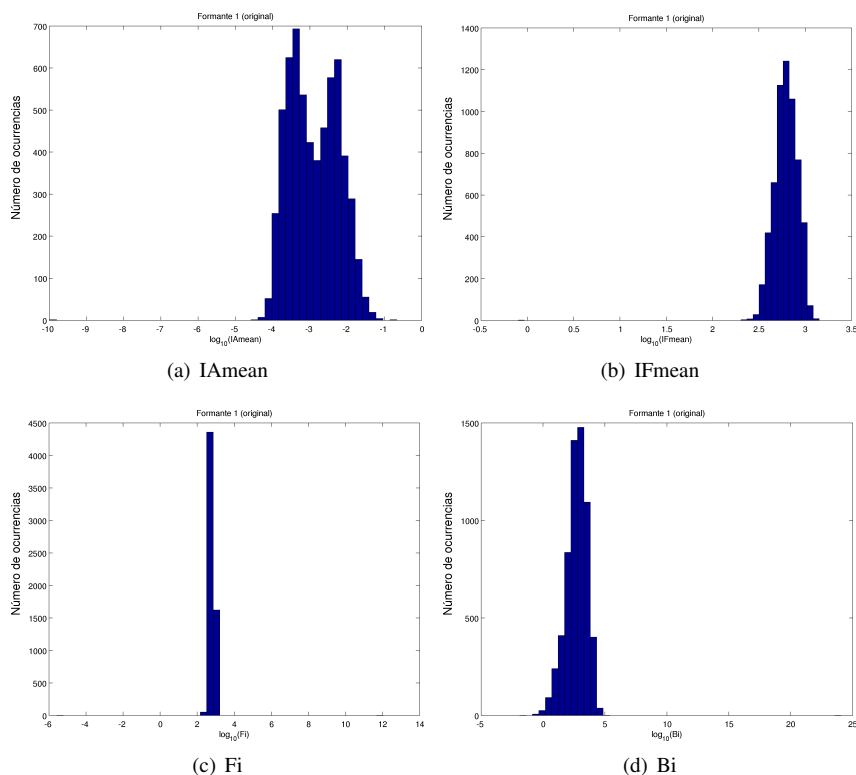


Figura 3.4: Histograma de los parámetros en fl

Visto esto, observamos que deberíamos encontrar un método para robustecer nuestros parámetros frente a valores extraños a la distribución real.

3.3.4. Regularización y normalización

Las aproximaciones realizadas para llegar a las definiciones de frecuencia y amplitud instantáneas dejan varios cabos sueltos que pueden dar lugar a un resultado numérico sin aparente sentido, como puede ser obtener un valor de energía negativo, divisiones por cero, etc. Estos valores son esporádicos, pero dependiendo de los ficheros analizados, nos encontrábamos estos valores con una mayor frecuencia de la esperada. Para resolver estos problemas, se han llevado a cabo diversas técnicas de regularización y suavizado.

Además, como se ha comentado anteriormente, uno de nuestros objetivos es encontrar una forma de combinar parámetros MFCC y los propuestos a partir del DESA1. Dicho esto, deberemos buscar la forma de tener una determinada normalización para que el resultado no se vea afectado por una diferencia excesiva en cuanto al rango de valores, varianza, etc.

Respecto a la regularización, se ha escogido un umbral empíricamente para limitar el valor mínimo de la energía. Con este valor nos protegemos ante divisiones por cero, así como cambios de fase y valores negativos (sin sentido físico).

Además, para eliminar valores espúreos de los parámetros finales, se calcula la desviación típica y la media y se descartan los valores que superan el valor de confianza del 99 %.

El resultado puede verse en los histogramas de la figura 3.5 para el primer formante. Se puede observar, comparándolos con los histogramas del apartado anterior que los valores extraños han sido eliminados.

Además, ahora podemos ver que las distribuciones tienen una forma más suave, en el caso de la amplitud 3.5(a) podríamos decir que existen dos modos, y el resto de parámetros poseen una forma más acampanada.

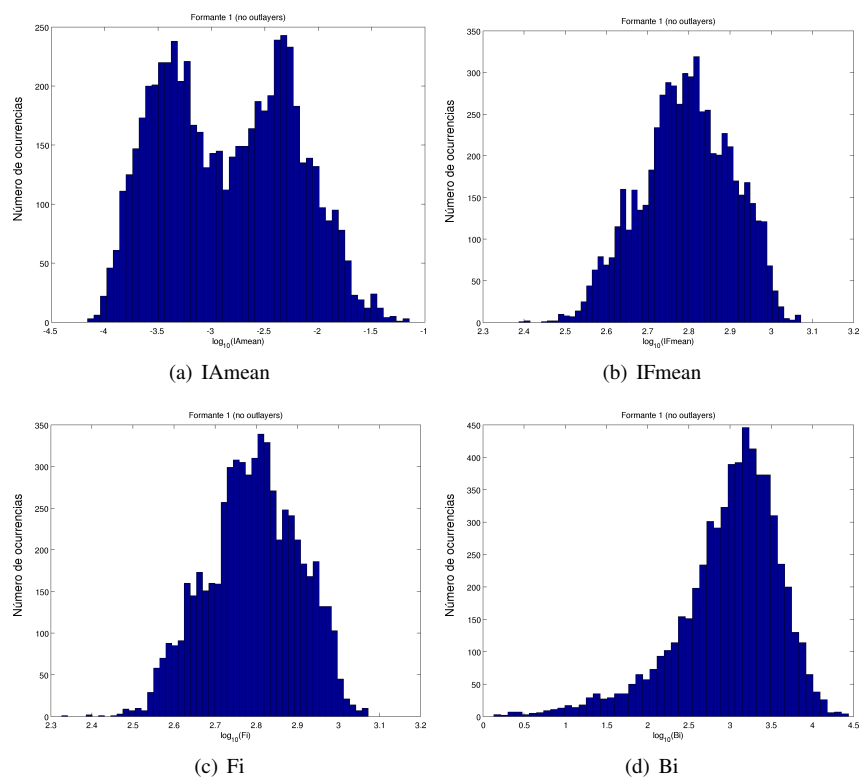


Figura 3.5: Histograma de los parámetros en f1, regularización

3.4. Selección de parámetros

Una de las tareas a realizar, es la selección de parámetros con un doble objetivo: observar cuales de ellos tienen una mayor influencia sobre la segmentación (o cuales solo nos aportan ruido), y si es posible, reducir dimensionalidad del vector de características.

Además, como hemos observado en las secciones anteriores, se obtienen estimaciones menos ruidosas en el primer y segundo formante, debido a que estimar la frecuencia central de los filtros en formantes superiores es menos preciso. Por lo tanto, a continuación se mostrarán datos de los parámetros MFCC y DESA para el primer formante.

Se han utilizado principalmente dos métodos para realizar estas tareas: cálculo del Bayesian Information Criteria (BIC) con matriz de covarianzas diagonal y Principal component analysis (PCA).

3.4.1. Matriz de covarianzas diagonal

Para realizar observaciones, se diseñó un método gráfico obteniendo el clustering mediante BIC, con matriz de covarianzas diagonal (vector de varianzas). Este método no es muy preciso, pero nos ayuda a tener una primera idea de cuáles son los parámetros que podrían discriminar mejor entre locutores. Dicho método es bastante claro ya que nuestra base de datos cuenta con conversaciones en las que solo aparecen dos locutores (ver la sección 2.7.1, titulada "Base de datos para el test" en la página 13)

La propuesta es la siguiente:

1. Se elige un subconjunto de ficheros de la base de datos.
2. Se realiza una cuadrícula, la cual es la representación ideal, cada cuadrado representa la secuencia alterna de locutores.
3. Se calcula el valor del ΔBIC tomando todas las parejas de segmentos posible. Como resultado, obtenemos una matriz triangular superior (3.13) donde $\Delta BIC_{m,n}$ representa el valor ΔBIC entre el segmento m y n , y c es una constante arbitraria, cuya finalidad es la de obtener una representación adecuada.

$$\Delta BIC = \begin{pmatrix} \Delta BIC_{1,1} & \Delta BIC_{1,2} & \cdots & \Delta BIC_{1,n} \\ c & \Delta BIC_{2,2} & \cdots & \Delta BIC_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ c & c & \cdots & \Delta BIC_{n,n} \end{pmatrix} \quad (3.13)$$

La representación a utilizar va a ser un tablero donde se aprecia una cuadrícula en la diagonal y el triángulo superior. Cada celda de la cuadrícula representará el valor de ΔBIC en dicha posición. Un valor claro indica una mayor posibilidad de cambio. El valor oscuro indica que ambos segmentos son más homogéneos. En el caso ideal, se obtendrá un “tablero de ajedrez” triangular superior como el de la figura 3.6.

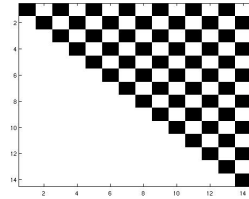


Figura 3.6: Separación ideal, transiciones alternas entre locutores

MFCC

En esta sección se muestra la aproximación de un subconjunto de los MFCC al caso ideal. El rango de intensidad es el valor ΔBIC .

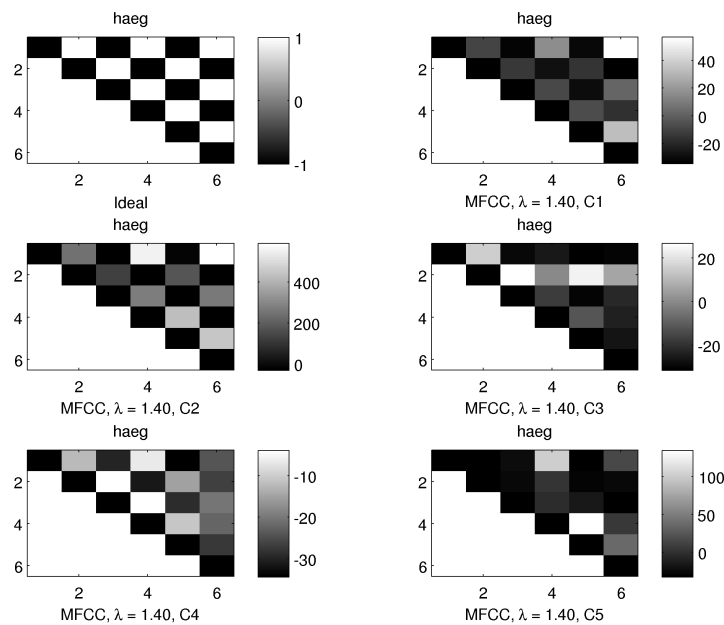


Figura 3.7: Separación con MFCC

En la figura 3.7 podemos ver la contribución que cada parámetro de los MFCCs

aporta al BIC, en este caso como primera aproximación utilizando una matriz de covarianza diagonal, en lugar de covarianzas completas. En la subfigura superior izquierda tenemos el caso ideal, y en el resto los cinco primeros coeficientes.

En esta figura se observa la aproximación que poseen dichos parámetros al caso ideal, imitando el tablero de ajedrez. Cuanto mayor es el contraste, mejor será la separación.

DESA

A continuación representamos en la figura 3.8 distintos parámetros del DESA, del primer formante. En este caso, la separación no es clara, y se producen distintos patrones (ya no es una cuadrícula “oscuro-claro-oscuro”) sino que puede ocurrir el patrón “oscuro-oscuro-claro”, cosa que no es deseable.

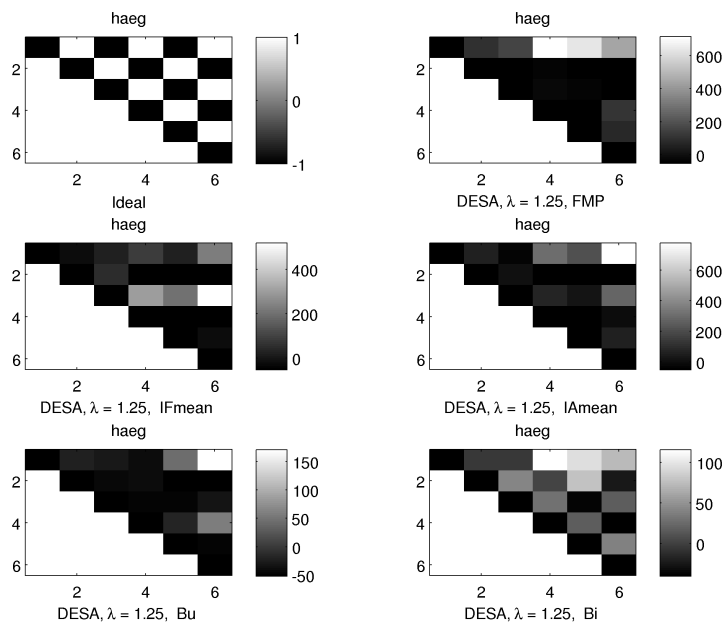


Figura 3.8: Separación con DESA

3.4.2. PCA

El segundo método utilizado fue realizar la transformación de los parámetros mediante análisis de componentes principales (Principal component analysis (PCA)). El objetivo es conseguir ordenar los parámetros y reducir dimensionalidad.

PCA se define como una transformación lineal que traslada los datos a un nuevo sistema de coordenadas tal que la proyección de la primera coordenada posee la mayor varianza, la segunda poseerá la segunda mayor varianza y así sucesivamente. PCA es teóricamente la transformación óptima para un conjunto de datos dado en términos de mínimos cuadrados.

Para una matriz \mathbf{X}^T , de media cero empírica, es decir, se ha sustraído la media del conjunto de datos, donde cada fila representa una repetición diferente del experimento, y cada columna es el resultado de un parámetro en particular, la transformación PCA viene dada por:

$$\mathbf{Y}^T = \mathbf{X}^T \mathbf{W} = \mathbf{V} \mathbf{\Sigma} \quad (3.14)$$

donde $\mathbf{V} \mathbf{\Sigma} \mathbf{W}^T$ es la descomposición en valores singulares de \mathbf{X}^T .

Este método se desechó al obtener peores resultados para un mismo número de coeficientes. El resultado original era de un 24 % de error frente al 28 % que se obtenía al aplicarle PCA.

Capítulo 4

Resultados de Segmentación y Clustering

En este capítulo se van a exponer los resultados más representativos obtenidos de la segmentación y clustering, así como los esquemas seguidos para obtener dichos valores. Además se van a presentar los esquemas básicos para calcular dichos resultados.

Inicialmente se obtuvieron los parámetros de forma continua y sin solapamiento entre tramas. Con este esquema, se aplicaron los métodos de selección de parámetros y se obtuvieron tasas de error para distintas elecciones. El entramado con solapamiento se utilizó en dos esquemas ligeramente distintos, realizándolo antes y después del filtrado multibanda.

Para todos los casos, se van a obtener finalmente los parámetros IA_{mean} , IF_{mean} , FMP , F_i y B_i (detallados en la sección 3.3.2).

4.1. Entramado sin solapamiento

Este primer esquema se propuso de inicio de tal forma que para fragmentos de la señal de entrada de 10 milisegundos, se realizaba un filtrado multibanda, donde cada filtro se centraba en la estimación de cada uno de los cuatro formantes, para a continuación aplicar DESA. Finalmente, para cada uno de los fragmentos anteriores, se saca un vector de características compuesto por los parámetros calculados en la ecuación (3.12) de la página 20.

En la tabla 4.1 tenemos la comparación entre el mejor resultado obtenido a partir de los parámetros del DESA (15 parámetros, información de los tres primeros formantes) y el valor del que partíamos que era un error del 19 % en el caso de los MFCC.

Experimento	DER	Tiempo Total (s)	Tiempo Correcto (s)	Tiempo Erróneo (s)
Baseline, MFCC 13 coefs.	19 %	26176,82	21083,92	5092,90
DESA 15 coefs. (3 formantes)	23 %	26176,82	20168,61	6008,21

Tabla 4.1: Resultados: Comparación MFCC - DESA

Para llegar al resultado del 23 % se han utilizado datos de otros experimentos: comparación para distintos valores de λ , intento de fusión tardía de parámetros (MFCC + DESA), variación del umbral del método T^2 .

El primer paso fue combinar los parámetros calculados con los MFCC. No solo esta combinación no ayuda sino que perjudica la decisión provocando como mínimo un error del 25 %, siendo este cuando solo se añade información del primer formante. (Tabla 4.2)

MFCC	DESA	DER	Tiempo Total (s)	Tiempo Correcto (s)	Tiempo Erróneo (s)
13	0	19 %	26176,82	21083,92	5092,90
13	5	25 %	26176,82	19538,69	6638,13
13	10	27 %	26176,82	19088,93	7087,89
13	15	27 %	26176,82	19091,44	7085,38
13	20	28 %	26176,82	18957,63	7219,19

Tabla 4.2: Resultados: Combinación MFCC-DESA ($\lambda = 2$)

El siguiente experimento consta de un barrido del valor de la penalización del modelo λ , para distintas situaciones: cuando los datos no son post-procesados, cuando los datos han sido limpiados de valores extraños (outlayers) y cuando además han sido normalizados. El hecho de normalizar no mejora los resultados, y quitar los elementos extraños ayuda en una pequeña cantidad, bajando como mucho un 1 % el error. Estos valores se han calculado para el mejor caso, considerando 15 parámetros, es decir, los 3 primeros formantes, y se obtiene que con un valor $\lambda = 1,5$ el valor del DER es de un 25 %. (Tabla 4.3)

λ	Sin Outlayers	Normalizado	DER
1	✓	✓	27 %
1	✓	×	26 %
1	×	×	26 %
1,25	✓	✓	25 %
1,25	✓	×	25 %
1,25	×	×	26 %
1,5	✓	✓	25 %
1,5	✓	×	25 %
1,5	×	×	26 %
1,75	✓	✓	27 %
1,75	✓	×	27 %
1,75	×	×	26 %

Tabla 4.3: Resultados: Test λ

Por último, otro experimento para la optimización de los parámetros del segmentador es la prueba de utilizar BIC y T^2 con distintos umbrales. La tabla es representativa de los resultados, consiguiendo el mejor error (23 %) para un umbral de 400. Estos resultados se pueden ver en la tabla 4.4.

BIC	T^2 /umbral	DER	Tiempo Total (s)	Tiempo Correcto (s)	Tiempo Erróneo (s)
✓	×	29 %	26176,82	19795,73	6381,09
✓	✓/ 100	27 %	26176,82	19096,14	7080,68
✓	✓/ 400	23 %	26176,82	20168,61	6008,21
✓	✓/ 600	26 %	26176,82	19242,21	6934,61

Tabla 4.4: Resultados: Comparación DESA: BIC- T^2

4.2. Entramado con solapamiento

Puesto que el análisis localizado “estándar” se basa en un entramado con solapamiento entre tramas, se realizaron distintos experimentos con solapamiento de dos maneras, realizando el filtrado antes y después del entramado. Puesto que añadir solapamiento no mejora los resultados anteriores, en esta sección se comentará brevemente el análisis realizado y se expondrán los dos mejores resultados para este caso, que pueden verse en la tabla 4.5.

Tipo	DER	Tiempo Total (s)	Tiempo Correcto (s)	Tiempo Erróneo (s)
Prefiltrado	28 %	26176,82	18768,69	7408,13
Postfiltrado	32 %	26176,82	17669,48	8507,34

Tabla 4.5: Resultados: Comparación Prefiltrado-Postfiltrado

4.2.1. Prefiltrado multibanda

En este caso, la señal es pasada por el banco de filtros, a continuación se aplica el DESA, y con la información de la amplitud y frecuencia instantánea se realiza el entramado para calcular los parámetros que conformarán nuestro vector de características (fig. 4.1).

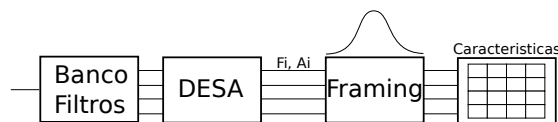


Figura 4.1: Esquema entramado. Prefiltrado

4.2.2. Postfiltrado multibanda

El filtrado en este caso se realiza después de haber hecho el entramado (figura 4.2). Este método sólo tiene sentido si se mira desde el punto de vista de la estimación de los formantes. Dicha estimación utiliza un entramado con solapamiento, por lo tanto, se tratará de utilizar el mismo número de datos tanto para dicha estimación como para el cálculo del DESA y sus características. Este método implica que el DESA ya no es continuo para todo el fichero, lo cual cabe esperar resultados peores.

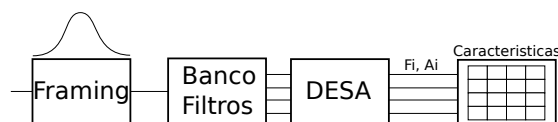


Figura 4.2: Esquema entramado. Postfiltrado

Capítulo 5

Conclusiones y líneas futuras

5.1. Conclusiones

En esta tesis se ha llevado a cabo la implementación de un sistema de evaluación para realizar un estudio de segmentación y clusterización mediante la extracción de parámetros basados en el algoritmo discreto de separación de energía DESA. A continuación se llevó a cabo el estudio del comportamiento de dicha parametrización a la tarea de segmentación.

Para ello se realizaron las tareas de extracción de parámetros a partir del modelo AM-FM, utilizando un análisis multibanda de la señal, se analizó cada parámetro por separado y se calculó el error de diarización.

Además, se planificaron distintos experimentos con distinto número de parámetros, se realizó una selección de los mismos, y se aplicó una fusión entre los parámetros de referencia (MFCC) y los DESA. También se comparó el rendimiento del sistema con y sin solapamiento de información para calcular los parámetros que servirían como entrada al BIC.

Rápidamente se desecharon los resultados con fusión de parámetros puesto que no solo mejoraba sino que el DER empeoraba. Además, para reducir dimensionalidad, tanto aplicando PCA como una selección manual no mejoró tampoco los resultados con lo que se descartó. Añadir el cuarto formante tampoco mejoraba los resultados con lo que finalmente, los mejores resultados se consiguieron con información de los tres primeros formantes.

A la vista de todos los experimentos realizados, no se puede superar el rendimiento de nuestro sistema baseline basado en los MFCC y observamos que el DESA y los parámetros derivados de él no nos aportan datos con los que el BIC pueda realizar una segmentación y clusterización precisa.

5.2. Líneas futuras

El trabajo presentado se centra en el estudio de segmentación y clustering basado en BIC y T^2 . Para continuar con el estudio de cómo influye la parametrización basado en un modelo AM-FM en la tarea de segmentación se proponen utilizar nuevas técnicas basadas en modelos de Markov y programación dinámica.

Además, nuestros experimentos se han realizado con señales limpias, sin ruido, y puesto que esta parametrización ha sido utilizada previamente en reconocimiento, buscando un sistema más robusto, se propone también realizar experimentos donde se intente segmentar y realizar clustering en entornos ruidosos, donde los MFCC podrían bajar el rendimiento.

Bibliografía

- [BMQ93] Alan C. Bovik, Petros Maragos, and Thomas F Quatieri. Am-fm energy detection and separation in noise using multiband energy operators. *IEEE Transactions on signal processing*, 41(12), December 1993.
- [CG] S. S. Chen and P. Gopalakrishnan. Speaker , environment and channel change detection and clustering via the bayesian information criterion. In *Broadcast News Trans.*
- [DM06] Dimitrios Dimitriadis and Petros Maragos. Continuous energy demodulation methods and application to speech analysis. *Speech communication*, (48):819–837, 2006.
- [ER96] Hassan Ezzaidi and Jean Rouat. Comparison of mfcc and pitch synchronous am, fm parameters for speaker identification. *ERMETIS, DSA*, 1996.
- [JQR95] C. R. Jankowsky Jr., T. F. Quatieri, and D. A. Reynolds. Measuring fine structure in speech: application to speaker identification. *IEEE*, 1995.
- [Kai90] J. F. Kaiser. On a simple algorithm to calculate the 'energy' of a signal. *IEEE ICASSP '90*, April 1990.
- [Mir06] Xavier Anguera Miró. *ROBUST SPEAKER DIARIZATION FOR MEETINGS*. PhD thesis, Universitat Politècnica de Catalunya, October 2006.
- [MKQ93] Petros Maragos, James F Kaiser, and Thomas F Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Transactions on signal processing*, 41(10), October 1993.
- [NIS06] NIST. Nist fall rich transcription on meetings 2006 evaluation plan. *NIST*, 2006.
- [RSB⁺04] J. Ramírez, J. C. Segura, C. Benítez, A. de la Torre, and A. Rubio. Efficient voice activity detector algorithms using long-term speech information. *Speech Communication*, 42:271–287, 2004.

- [Sta00] ETSI Standard. Res/stq-00018, etsi es 201 108 v1.1.2 (2000-04) speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms, April 2000.
- [Tea80] H.M. Teager. Some observations on oral air flow during phonation. *IEEE Transactions on signal processing*, ASSP 28:599–601, October 1980.
- [TI06] TI. Switchboard: A user’s manual. *UPENN*, 2006.
- [Vaq08] Carlos Vaquero Avilés-Casco. Técnicas de verificación de locutor. Master’s thesis, Universidad de Zaragoza, June 2008.
- [ZH05] B. Zhou and J. H Hansen. Efficient audio stream segmentation via t2 statistic based bayesian information criterion (t2-bic). *IEEE Trans. Speech Audio Processing*, 13(4), 2005.

Acrónimos

- BIC** Bayesian Information Criteria
- DCT** Discrete Cosine Transform
- DER** Diarization Error Rate
- DESA** Discrete Energy Separation Algorithm
- FFT** Fast Fourier Transform
- FMP** FM-modulation percentages
- LTSD** Long-Term Spectral Divergence estimation
- LTSE** Long-Term Spectral Envelope
- MFCC** Mel-frequency cepstral coefficients
- PCA** Principal component analysis
- VAD** Voice Activity Detector