Desarrollo de un Clasificador Bayesiano de Ruido y Voz Mediante Estimación por Máxima Verosimilitud.

Luis Fernando D´Haro
Ing. Electrónico
Docente Universidad Autónoma de Occidente
Cali, Colombia
lfdharo@cuao.edu.co

Resumen

Una de las tareas básicas del preprocesado de voz en los sistemas de reconocimiento automáticos del habla o en los sistemas de diálogo, es la detección de las tramas de voz con el fin de facilitar los procesos de entrenamiento, evaluación o asignación de turnos. El presente trabajo describe el desarrollo de un clasificador de Bayes para identificar tramas de voz y ruido, habiéndose entrenado un modelo de una sola Gaussiana para cada clase. La identificación se hace trama por trama, asignándole la clase con la que presente la máxima verosimilitud según el modelo. Posteriormente se aplica un filtro de mediana configurable, desde un orden 1 hasta 9, analizando con cual se obtienen los mejores resultados. Se prueba además el efecto de aplicar a los vectores de entrada normalización cepstral media (CMN) y de variar su número de dimensiones o parámetros (desde 1 hasta 33). La base de datos utilizada fue SpeechDat para dígitos aislados, escogiéndose un total de 1200 locutores. Los resultados permiten concluir los beneficios de CMN y la notable reducción del número de dimensiones basándose en el criterio de mínima varianza global e intra-clase de los datos. Finalmente se muestran algunas líneas futuras de desarrollo..

1. Introducción

Uno de las etapas previas en cualquier sistema de reconocimiento de voz, sistemas de videoconferencia, de segmentación y etiquetado de voz, es la determinación de las tramas en que inicia y termina la voz, considerándose todo lo demás como ruido. El objetivo de este trabajo fue obtener las tasas de reconocimiento de las tramas de voz y ruido. Segundo, poder determinar los parámetros de codificación de la voz más relevantes para aumentar dichas tasas. Tercero, probar distintas técnicas de preprocesado y postprocesado de los parámetros de entrada y de salida, y su efecto en la salida del reconocedor. Este proyecto está enmarcado en un proyecto de cooperación con la Universidad Politécnica de Madrid, denominado Isaías, que servirá para asignar turnos de intervención en un sistema de educación virtual. Conviene mencionar que el objetivo del reconocedor no es obtener un 100% de fiabilidad, por el contrario se busca obtener un porcentaje significativo de "falsas alarmas" que permitan entrenar modelos de reconocimiento de voz robustos, además se espera que sea de rápida ejecución dada las características de tiempo real esperadas.

Este artículo está divido en cinco secciones básicas; en la primera se mencionan las características de la base de datos seleccionada para hacer el entrenamiento, en la segunda sección se menciona el proceso de extracción y características del vector de parámetros de la voz. En la tercera sección se menciona el método de clasificación propuesto, así como el preprocesado de la voz y la aplicación de un filtro de mediana a la salida del reconocedor. En la cuarta sección se presentan los distintos experimentos realizados, sus resultados y observaciones. Finalmente, se presentan las conclusiones y líneas futuras del proyecto.

2. Características de la base de datos

Se utilizó la base de datos SpeechDat en Español para 4000 locutores, de los cuales fueron utilizados 1200. Esta base de datos fue grabada por la Universidad Politécnica de Cataluña, para aplicaciones de teleservicios de voz, entre los que se destacan: entidades bancarias, ciudades, apellidos, cantidades de dinero, fechas, dígitos aislados, etc. Para ello se empleó una interfaz telefónica ISDN a una frecuencia de 8 KHz, 8 Bits/muestra y codificación en ley A. La mitad de los locutores son hombres y la otra mitad mujeres, todos ellos con edades entre los 16 y 30 años; dichos locutores provienen de regiones diversas de España, por lo que se cuenta con una gran variedad dialectal.

Las voces han sido cuidadosamente transcritas al nivel ortográfico y segmentadas de forma automática. Se dispone de una trascripción fonética en formato SAMPA. Cada grabación cuenta con un fichero de extensión ESA en que se guarda las muestras de la voz, y de un fichero con extensión ESO en el que se guardan las etiquetas en formato ASCII SAM. Este fichero ESO, o de trascripción, incluye información sobre la palabra, o palabras,

pronunciada, nombre del locutor, sexo, edad, datos del fichero de audio, malas pronunciaciones y ruidos adicionales, entre otros. Lo más relevante de estos ficheros para este trabajo, son las etiquetas LBO y LBR. La primera contiene el valor de las muestras en que inicia y termina la voz.; la segunda, contiene la cantidad de muestras totales del fichero. Con esta información se puede determinar la cantidad de muestras a procesar, las posiciones y cantidad muestras en que se encuentra la voz y el ruido, así como facilitar el proceso de cálculo de los scores.

3. Extracción del vector de parámetros y sus características

La parametrización de la voz cumple un doble objetivo: primero, busca reducir la cantidad de información redundante de la señal de voz y, segundo, facilitar la tarea de reconocimiento y tratamiento de la misma. En este proyecto se empleó la parametrización cepstral en la escala Mel (MFCC Mel Frequency Cepstral Coefficients), adicionándole los parámetros transicionales: Delta (velocidad) y Delta-Delta (aceleración). Adicionando también la energía de la trama y sus derivadas primera y segunda. A continuación se explica con mayor detalle este proceso.

3.1 Enventanado

Dado que la voz presenta características de cuasiestacionaridad en periodos cortos de tiempo, se puede realizar un análisis espectral en tramas cortas, típicamente de 10 – 15 ms. Para hacer el enventanado se emplea, generalmente, la ventana de Hamming, pues sus características espectrales permiten mejorar los cálculos de la FFT. La ventana de Hamming se formula según (1).

$$w_n = \begin{Bmatrix} 0.54 - 0.46 \cdot \cos(\frac{2 \cdot p \cdot n}{N-1}) & 0 \le n \le N \\ 0 \qquad \textit{En los demás puntos} \end{Bmatrix} \quad (1)$$

Una vez se realiza el enventanado se procede a calcular la transforma discreta de Fourier y a partir de ella, sólo el módulo; ya que la fase no aporta información relevante para la voz, esta no se tiene en cuenta.

3.2 Filtros en la escala Mel

Estos filtros permiten calcular la energía de la señal en cada banda de frecuencia comprendida entre los 200 Hz hasta los 4KHz. Se emplearon 20 filtros o bandas, que intentan emular el comportamiento del oído humano, dándose mayor énfasis en las bajas frecuencias y menos en las altas. Los primeros 10 filtros están espaciados linealmente hasta 1 KHz y los siguientes 10, están espaciados logarítmicamente. La forma escogida de estos filtros fue triangular, siendo el máximo valor 1 en la frecuencia central y decrece linealmente hasta cero en la frecuencia central de los dos filtros adyacentes. (Figura 1).

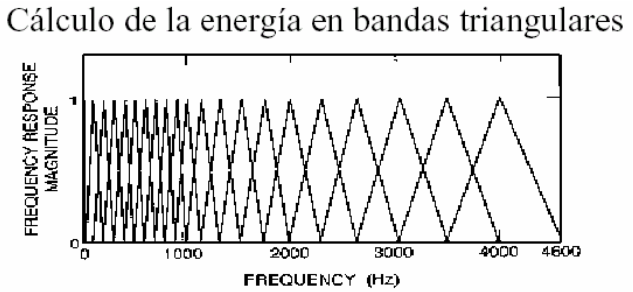

Figura 1. Forma de los filtros utilizados en el cálculo de los parámetros cepstrales

El paso siguiente es calcular el logaritmo de las energías halladas anteriormente, con lo que se consigue separar las altas componentes frecuenciales, que están relacionadas con las variaciones rápidas del espectro y que se corresponden con la frecuencia de la señal de excitación, de las bajas componentes que se corresponden con las variaciones lentas del espectro, relacionadas con la respuesta en frecuencia del filtro que modela el tracto vocal, que es el que aporta mayor información a la hora de hacer reconocimiento de voz.

Una vez obtenido el logaritmo, se procede a calcular los coeficientes cepstrales mediante la transformada inversa coseno (DCT) dada por (2).

$$MFCC_j(i) = \sum_{k=1}^{L} e(j,k) \cdot \cos\left[i \cdot \left(k - \frac{1}{2} \right) \cdot \frac{p}{L} \right] \quad (2)$$

k: La banda de frecuencias.
j: La trama en curso.
e(j,k): El logaritmo de la suma de los módulos de la FFT en la banda k de la trama j.
L: El número de bandas o filtros (20 en este caso).
M: El número total de coeficientes MEL (10 en este caso).

A los anteriores10 parámetros se le añade uno más que corresponde con el de la energía dela señal en dB, que se calcula mediante (3):

$$E(t) = 10 \cdot \log \sum_{i=1}^{L} |X(i)| \quad (3)$$

donde:

X(i) es la transformada discreta de Fourier
t es la trama objeto del cálculo
L es el número de puntos de la FFT dividido por 2

3.3 Parámetros transicionales

Dado que los MFCC reflejan las propiedades instantáneas de la voz, pues sólo dependen de la trama tratada, es conveniente obtener también las variaciones dinámicas del espectro ya que estas aportan información útil en los sistemas de reconocimiento. Inicialmente se pensó que para este trabajo serían útiles, sin embargo luego se comprobó que no lo son tanto para la detección de voz/ruido.

La fórmula utilizada para realizar este cálculo está dada por la ecuación (4):

$$\Delta\ MFCC(t) = \frac{\sum_{i=-K}^{K} i \cdot MFCC(t+i)}{\sum_{i=-K}^{K} i^2} \qquad (4)$$

En general se utiliza una ventana de cálculo de 5 puntos, por lo que K sería igual a 2, siendo t la trama que se está analizando.

Si se realiza este mismo proceso a partir de los parámetros transicionales (Delta) se obtienen los parámetros denominados Delta-Delta o de aceleración. Adicionalmente calcula la energía diferencial a partir de la energía local usando los coeficientes de regresión con una ventana de análisis de 5 puntos.

De esta forma se obtuvieron un total de 33 parámetros (10 Mel + 1 Energía, 10 Delta Mel + 1Delta Energía, 10 Delta Delta Mel + 1 Delta Delta de Energía).

3. Clasificación mediante MLE

3.1 Cálculo de la Verosimilitud mediante el teorema de Bayes

El teorema de Bayes establece que la probabilidad de la clase w_i dado el vector de características x es igual a la probabilidad a priori de la clase por la función de densidad de probabilidad $p(x / w_i)$, sobre la probabilidad total de las muestras, según se observa en la ecuación (5).

$$P(w_i / x) = \frac{p(x/w_i)P(w_i)}{p(x)} \qquad (5)$$

El término del denominador, al ser común a todas las clases no se ha considerado, en tanto que la probabilidad a priori de cada clase se ha asumido como igual para todas (0.5). La función de densidad de probabilidad se considera que es de una sola Gaussiana, donde la varianza y la media se calculan según se mencionó anteriormente.

Para realizar la estimación de los parámetros y la probabilidad de la clase según el vector de características, se ha empleado el método de estimación por máxima verosimilitud (MLE Maximum Likelihood Estimation), y para facilitar los cálculos y el procesado computacional, se ha recurrido a la función monótona creciente del logaritmo, obteniéndose así la log-verosimilitud. La cuál, para un modelo de una sola Gaussiana, se calcula según (6)

$$-\ln p(x/wi) = \sum_{C=1}^{C=33} \frac{\left(x^C - m_i^C\right)^2}{2*(s_i^C)^2} - \sum_{C=1}^{C=33} \ln\left(\frac{1}{2*s_i^C}\right) \qquad (6)$$

Para determinar cuál clase era ganadora se escoge la que presente una menor log-verosimilitud para el vector de características de entrada.

3.2 Cálculo de las medias y las varianzas

La varianza y la media de los datos globales se calculan para determinar los parámetros a emplear, y sus valores por clase para realizar el entrenamiento y el reconocimiento. Las ecuaciones empleadas para ellos fueron (7)y (8).

$$\overline{X} = \frac{1}{n}\sum_{i=0}^{N-1} X_i \qquad (7)$$

$$s^2 = \frac{1}{n}\sum_{i=0}^{N-1} X_i^2 - \overline{X}^2 \qquad (8)$$

Las covarianzas no se calcularon ya que, de experiencias anteriores, se sabe que estos parámetros tienden a presentar una baja correlación entre ellos por lo que los valores que no están sobre la diagonal principal son cercanos a 0 y por ende pueden despreciarse. Mediante el entrenamiento por clase se desarrollan dos Gaussianas diferentes, una para cada clase.

Con el fin de determinar cuáles parámetros son más importantes para el clasificador, se obtuvieron tanto la varianza y media global, como los dependientes de cada clase. Para el cálculo global, no se tienen en cuenta la distinción entre clases. Una vez obtenidos los valores, se procedió a ordenarlos de menor a mayor varianza, seleccionando siempre aquellos que presentan una menor varianza en ambos grupos.

3.3 Filtro de mediana

Con el fin de obtener un suavizado en la respuesta obtenida por el predictor, se implementó un filtro de mediana que consisten en una ventana que toma n/2 muestras anteriores, la actual y n/2 muestras futuras del resultado obtenido por el predictor; se suma la cantidad de apariciones de cada clase y se elige como resultado para la muestra actual la que tenga mayor número de apariciones. Así se eliminan glitches o cambios bruscos de la señal.

Para solucionar el problema de los extremos se planteó un filtro de anchura variable. El orden del filtro siempre es impar y hace alusión a la cantidad de muestras que toma para realizar el promediado. La figura 1 permite comprender con mayor claridad el funcionamiento de este filtro. Además permite observar la eliminación de los glitches indeseados.

3.4 CMN (Cepstral Mean Normalization)

Es una técnica clásica que permite normalizar los parámetros cepstrales, buscando robustecer el reconocedor frente al canal de grabación. Para ello se busca que los parámetros MFCC tengan una media igual a 0, que se consigue substrayendo el promedio de todos los vectores cepstrales de un mismo fichero con cada uno de los vectores de parámetros entrantes de la trama a procesar. El inconveniente que tiene es que retrasa el procesamiento en sistema en tiempo real, aunque hay varias propuestas para solventar este inconveniente. Un análisis más profundo del tema se encuentra en [2].

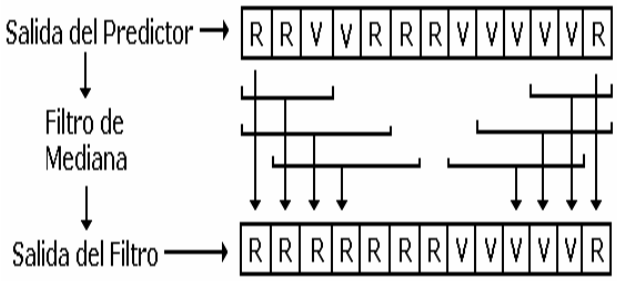

Figura 1. Resultado de un filtro de mediana con ventana igual a 5

3.5 Leave One Out (LOO)

Con elfin de entrenar con una mayor cantidad de datos y así obtener banda más estrechas de incertidumbre, se empleó este método, que consiste en dividir todo el conjunto de datos en varios subconjuntos dejando siempre por fuera uno para la fase de prueba y otro para la fase de evaluación. En este trabajo se dividieron los datos en 10 partes y sólo se dejaba 1 para la prueba. El procedimiento se repite n veces según el número de subconjuntos hechos. El inconveniente es que es computacionalmente costoso a medida que aumenta el número de datos.

3.6 Cálculo de los scores

Para el cálculo de los resultados se midió tanto la tasa de acierto general, como la de aciertos por clase. Empleando un margen de confianza del 95%. Las fórmulas empleadas tanto para la tasa, como banda están dadas por (9) y (10). Dado el gran número de tramas empleadas las bandas son muy estrechas con lo que los resultados son más confiables.

$$Tasa = 100 - 100 * (\frac{NumeroDeErrores}{TotalTramas}) \quad (9)$$

$$Banda = \quad 1.96 * \sqrt{\frac{Tasa * (100 - Tasa)}{TotalTramas}} \quad (10)$$

Total Tramas de Voz	Total Tramas de Ruido	Total de Tramas
104550	292369	396919

Tabla 1. Número total de tramas de voz y ruido empleadas en el proyecto.

4. Experimentos y resultados

Se realizaron varios experimentos en los que se variaban 3 parámetros básicos: el número de coeficientes utilizados (33, 22, 11, 6 hasta sólo 1), donde al emplear un solo parámetro se probó con el que presentara mínima dispersión intraclase, que coincidió con el de mínima varianza interclase y global, y probando con el parámetro de Energía. El segundo parámetro variable es el empleo de la normalización cepstral o no, y el último el orden del filtro de mediana (9, 7, 5, 3 y 1). Los resultados se detallan a continuación. Los mejores resultados se resaltan en negrilla, y al mejor de todos se le subraya para mayor claridad.

ENERGIA	FILTRO				
CMN	1	3	5	7	9
No	77,8	78	78,1	78,2	**78,2**
Si	85,5	85,9	86,2	86,4	**86,5**

Tabla 1. Resultados de los experimentos empleando la energía.

6 MEJORES	FILTRO				
CMN	1	3	5	7	9
No	82,2	83,4	84,1	84,4	**84,6**
Si	84,8	86,4	87	87,4	**87,6**

Tabla 3. Resultados de los experimentos empleando 6 parámetros

TODOS	FILTRO				
CMN	1	3	5	7	9
No	84,8	85,3	85,5	85,9	**86,2**
Si	85,1	85,6	85,9	86,2	**86,5**

Tabla 4. Resultados de los experimentos empleando 33 parámetros

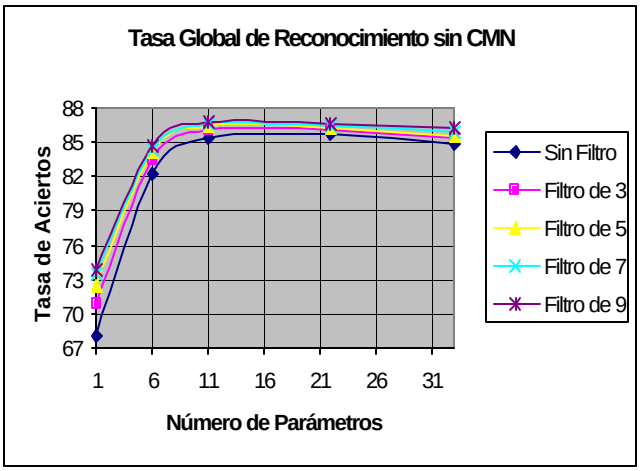

Figura 2. Tasa Global de reconocimiento sin CMN

La figura (2) permite observar como a medida que aumenta el tamaño del filtro de mediana se obtiene una mejora significativa en la tasa de aciertos, sin embargo, conforme esta crece tiende a estabilizarse. La razón de este comportamiento puede deberse a que los glitches no son de mucha duración, presentándose la mayoría en las transiciones entre voz y ruido, y ya que esta transición no suele durar más de unos 15 ms, el empleo de ventanas más grandes aporta demasiado. La gráfica también permite observar como al ir aumentando el número de parámetros, la tasa de acierto sube a un máximo ubicado en los 11 parámetros y con un filtro de orden 9 (86,8%). El hecho de que un menor número de parámetros sea necesario puede deberse a la gran variabilidad de los últimos 22 parámetros (transicionales), que si bien son útiles en reconocimiento o identificación de locutor, puede tener sentido, más para esta aplicación no lo es tanto. Otra razón, es que dada su gran variabilidad hacen más difícil hacer una discriminación interclases con ellos.

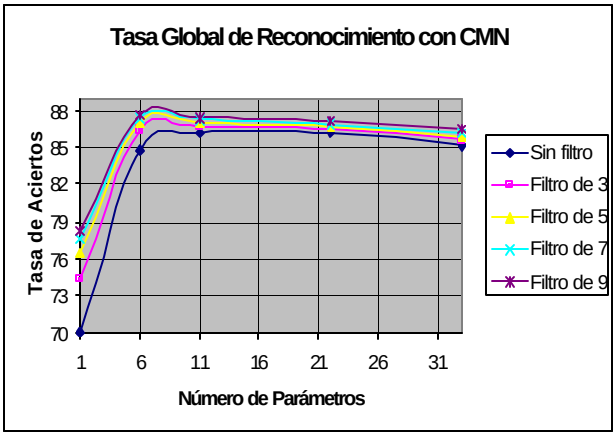

Figura 3. Tasa Global de reconocimiento aplicando CMN.

Al observar la figura (3) que muestra los resultados de aplicar CMN variando el número de parámetros y el orden del filtro, se puede observar el mismo comportamiento anterior, pero se obtiene un incremento relativo del 19.48%. De esta forma se concluye que la normalización ayuda notablemente al sistema. La razón es que al normalizar se obtienen características de canal más similares dada la gran cantidad de grabaciones distintas obtenidas.

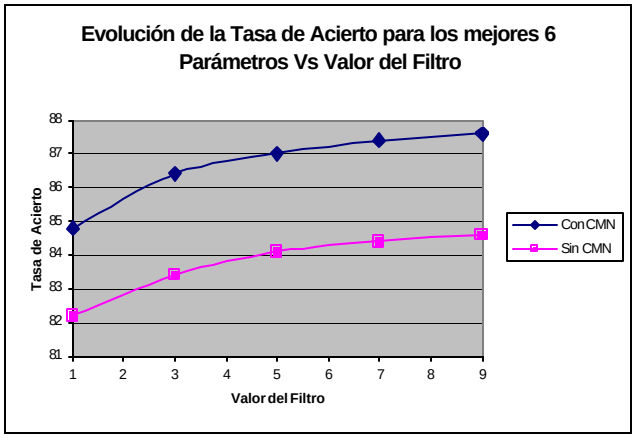

Figura 4. Tasa de acierto, con y sin CMN, para los 6 mejores parámetros con orden de filtro variable.

La figura (4) muestra en detalle el incremento logrado al usar CMN cuando el número de parámetros es 6 y el orden del filtro se varía entre 1 a 9.

5. Conclusiones y trabajos futuros

A partir de los diversos experimentos realizados, se ha podido demostrar la robustez que supone el empleo del algoritmo de normalización CMN, lográndose una tasa máxima de acierto global con un 87,6%, en comparación con el 84,6% logrado sin CMN para las mismas condiciones.

También se comprobó las mejoras que aporta el emplear un filtro de mediana a la salida del reconocedor, con el fin de eliminar cambios bruscos en el predictor. Se observa claramente el efecto que tiene el ir aumentando el ancho del filtro, en este caso se probó hasta un ancho de 9, pues se observó que a medida que se aumenta el aporte al reconocimiento tiende a ser más pequeño, dando una mejora promedio de 1,4% sin CMN y de un 2,6% con CMN.

Por otro lado, al ir agregando o quitando parámetros o características se obtuvo la máxima tasa de acierto cuando el número de parámetros es 6 al emplear CMN y de 11 cuando no se emplea CMN, lo anterior puede ser útil ya que podría agilizar los procesos de extracción de características en sistemas en tiempo real, aliviando la carga computacional. Sumado a lo anterior, se observa que de los 11 mejores parámetros, 10 pertenecen a los

parámetros MFCC mas la energía, por lo que no sería necesario calcular los parámetros transicionales, salvo el parámetro transicional de energía pues este hace parte de los mejores 11; esto se da ya que no nos interesan las particularidades del locutor, como en el reconocimiento de voz, sino únicamente la detección de las tramas de voz.

5.1 Futuros desarrollos

1.Aumentar el número de Gaussianas, permitiendo que para cada parámetro se puedan tener mas de una de ellas. El hacer esto supone un aumento en los cálculos, por lo que sería importante seleccionar bien los mejores parámetros.

2. Probar con los diferentes métodos de normalización que ofrece el CMN, pues hay algunas variaciones como el empleo de una ventana finita para calcular la media y la varianza, y no emplear todo el fichero como se hizo hasta ahora.

3. Se puede agregar un parámetro adicional que sería el de cruce por ceros, que seguramente daría mejores resultados ya que hay bastantes diferencias en el número de cruces por cero para una señal de voz que para una señal de ruido. Algunos sistemas también adicionan la medida de energías en bandas.

4. Este trabajo continuará pero implementando un reconocedor con Redes Neuronales del tipo Perceptrón Multicapa (MLP), a fin de hacer una comparación entre ambos sistemas. Igual que para este proyecto, se probará con variar la cantidad de parámetros y con los algoritmos de normalización y filtrado.

6. Bibliografía

[1] L. García P. Estudio y evaluación de algoritmo de robustez frente al ruido en un sistema de reconocimiento de voz sobre línea telefónica. UPM, Madrid. 2000. PFC ETSI Telecomunicaciones.

[2] G. F. Caminero. Estudio de técnicas de rechazo y rectificación en reconocedores de números conectados multilíngües sobre línea telefónica. Tesis doctoral, UPM, Madrid. 2000.

[3] X. HUANG, A. ACERO, et al. Spoken language processing. Prentice-Hall. 2001. ISBN: 0-13-022616-5.

[4] D. O`SHAUGHENESSY. Speech communication, Human and Machine. Adisson-Wesley, 1987.

[5] Gps-tsc.upc.es/veu/LR/LR_SPD_FDB.php3 Página de información sobre SpeechDat en español.