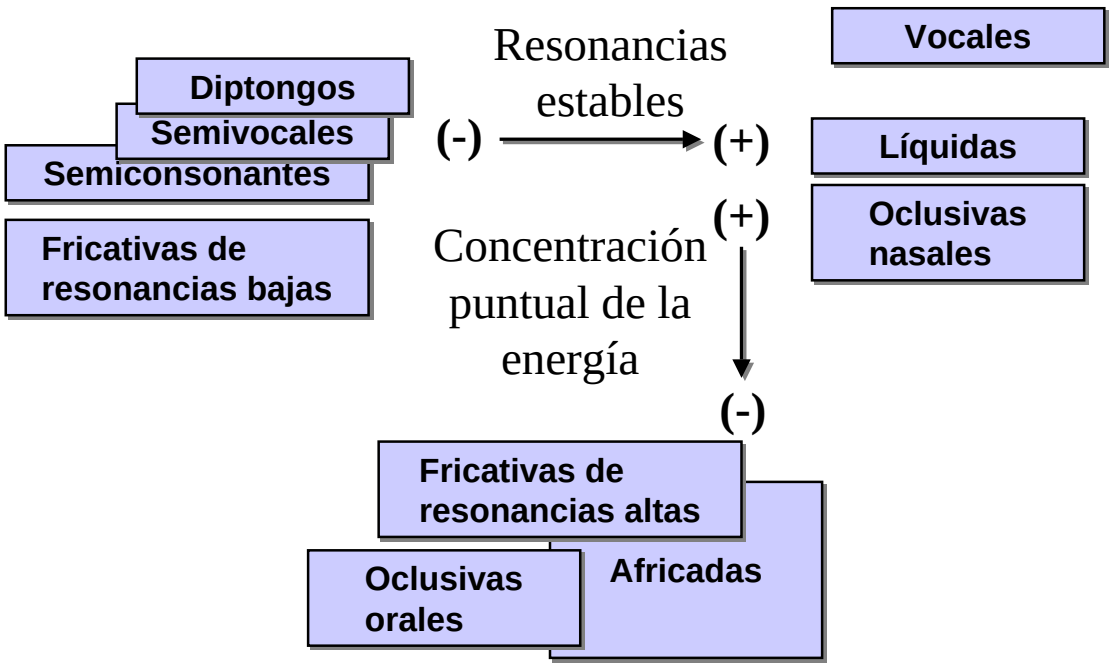
FUNDAMENTOS DEL RECONOCIMIENTO AUTOMÁTICO DE LA VOZ

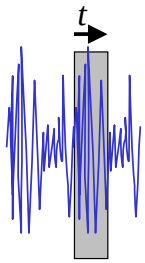
"Algoritmos de extracción de características"

Agustín Álvarez Marquina

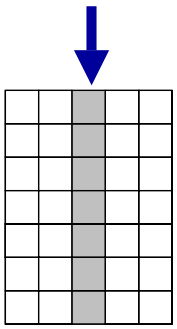
Introducción (I). Caracterización acústica de los sonidos



Introducción (II). Representación de la señal



Extractor paramétrico



Coeficientes

- **Objetivo: transformar la información presente en la traza de voz en un conjunto reducido de valores.**

- **Coeficientes estáticos**. Obtenidos a partir del análisis de pequeños fragmentos de la señal de voz (5- 20 ms).
- **Coeficientes dinámicos**. Producto de la combinación de componentes de diversos vectores.

Métodos de extracción paramétrica

- **Análisis por banco de filtros digitales.**
- **Transformada discreta de Fourier.**
- **Predicción lineal.**
- **Análisis cepstral.**
- **Predicción lineal perceptual.**

Análisis por banco de filtros digitales (I)

- **El uso de bancos de filtros digitales [PIC93] [ROB98], implementados inicialmente como filtros analógicos, ha sido históricamente la primera aproximación al procesamiento del habla.**
  - Un banco de filtros paso banda puede entenderse como un modelo sencillo de las etapas iniciales del sistema auditivo humano.
  - La señal inicial se descompone en un conjunto discreto de muestras espectrales, que contienen una información similar a la que se presenta en los niveles superiores del sistema auditivo.

Análisis por banco de filtros digitales (II)

- **Con objeto de aproximarse a la sensibilidad del oído humano, que no tiene una respuesta lineal en frecuencia existen diferentes escalas.**
- **Algunos ejemplos expresados de forma analítica, siendo *f* el valor de frecuencia en Hz, son:**
  - Escala de Bark.

$$Bark = 13 arctan\left(\frac{0.76\, f}{1000}\right) + 3.5 atan\left(\frac{f^2}{(7500)^2}\right)$$

Análisis por banco de filtros digitales (III)

- Escala de Mel.

$$m = 2595 \log_{10}\left(\frac{1+f}{700}\right)$$

- **Ésta última es la más usual en aplicaciones de tratamiento de la voz.**

Análisis por banco de filtros digitales (IV)

- **Un banco de filtros está constituido por un conjunto de filtros cada uno de los cuales retiene la información de una serie determinada de frecuencias del espectro.**

- **A su vez cada filtro puede ponderar de manera diferente las frecuencias que quedan bajo su ámbito.**

  - Un ejemplo de banco de filtros empleando escalas de Mel y 19 filtros es el de la Figura 1.

Análisis por banco de filtros digitales (V)

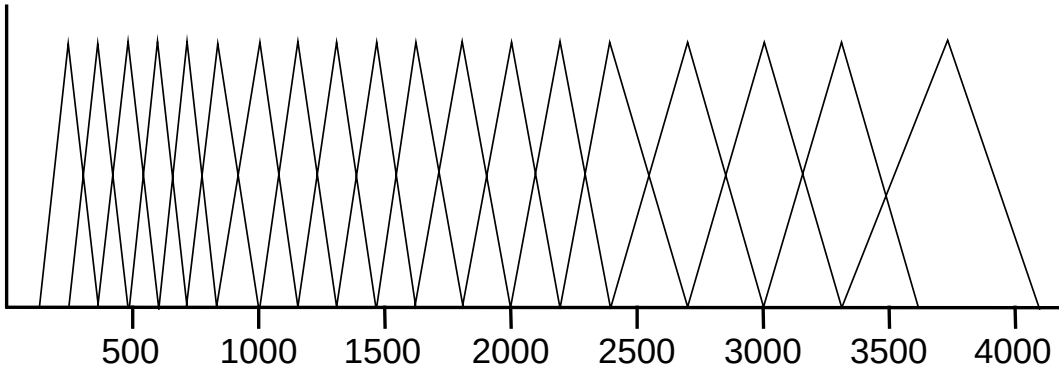


Figura 1. Banco de filtros de Holmes.

- **Este tipo de técnica, generalmente se emplea de manera conjunta con otros métodos como son el cálculo de coeficientes cepstrales.**

Transformada discreta de Fourier (I)

- **La transformada discreta de Fourier o DFT (Discrete Fourier Transform) [COO92] [DEL93] [KRA94] [DEL94], se define como:**

$$X(k) = \sum_{k=0}^{N-1} x(n) e^{-j\left(\frac{2\pi kn}{N}\right)} \qquad k = 0,1,2,\mathrm{K}, N-1$$

  - donde *N* es el número de muestras de la ventana que se va a analizar.

Transformada discreta de Fourier (II)

- **La DFT es periódica en la frecuencia con periodo f_s (frecuencia de muestreo).**
- **Presenta también como propiedades interesantes:**
  - La linealidad.

$$\alpha h(n) + \beta g(n) \Leftrightarrow \alpha H(k) + \beta G(k)$$

Transformada discreta de Fourier (III)

- La relación existente entre las operaciones de multiplicación/convolución en los dominios temporal y frecuencial.

$$h(n) \otimes g(n) \Leftrightarrow H(k)G(k)$$

$$h(n)g(n) \Leftrightarrow H(k) \otimes G(k)$$

Transformada discreta de Fourier (IV)

- **Por su parte la DFT inversa o IDFT se define como:**

$$x(n) = \sum_{k=0}^{N-1} X(k) e^{j\left(\frac{2\pi kn}{N}\right)} \qquad n = 0,1,2,\mathrm{K}, N-1$$

- **La motivación del uso de la DFT parte del hecho de la utilidad que tiene descomponer la señal de voz de partida en sus componentes en frecuencia.**

Transformada discreta de Fourier (V)

- **Un aspecto importante si queremos usar la DFT con señales de voz es que debemos asumir que al menos en periodos cortos de tiempo se cumple que la señal es estacionaria.**
  - En la realidad esto no es estrictamente así aunque podemos suponerlo.
  - La solución consiste en multiplicar la señal por una función ventana que sea 0 fuera de un determinado rango y reproducir el resultado de forma que tengamos un número de bloques iguales.

Transformada discreta de Fourier (VI)

- La ventana rectangular se define como:

$$w_n = \begin{cases} 1 & 0 \le n \le N \\ 0 & resto \end{cases}$$

- Sin embargo la utilización de esta ventana trae consigo, que en los puntos de inicio y fin exista una fuerte discontinuidad.

Transformada discreta de Fourier (VII)

- Para reducir el efecto de discontinuidad al mínimo debemos emplear tipos de ventana que tiendan a reducir a 0 los valores de las muestras en los extremos.
  - Aunque existe un buen número de tipos de ventana, la más común en el análisis de la voz es la que se conoce como ventana de Hamming (Figura 2):

$$w_n = \begin{cases} 0.54 - 0.46\cos\left(2\pi n/\left(N-1\right)\right) & 0 \le n \le N \\ 0 & resto \end{cases}$$

Transformada discreta de Fourier (VIII)

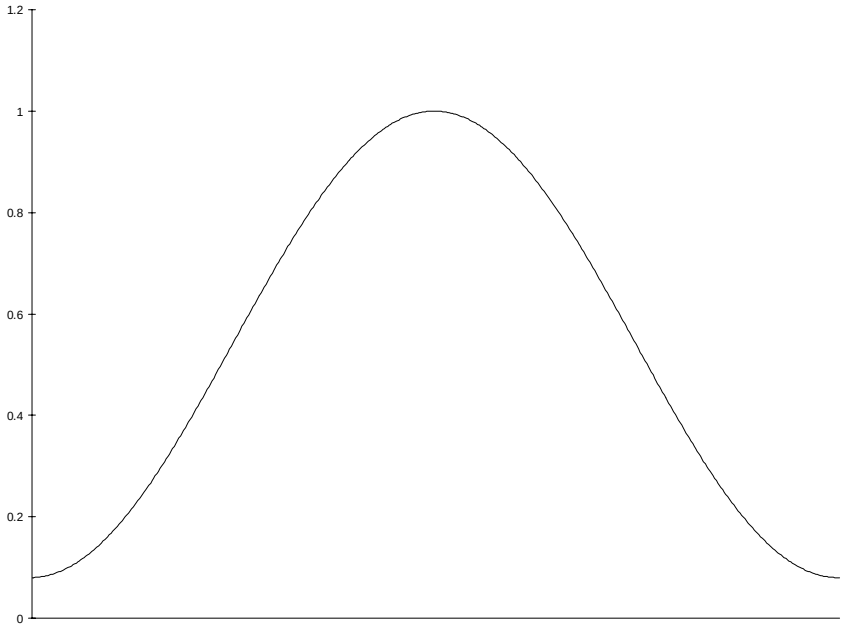


Figura 2. Ventana de Hamming.

Transformada discreta de Fourier (IX)

- **La complejidad de la DFT es de O(n2) operaciones y con objeto de acelerar el cálculo de este procedimiento, se emplea habitualmente lo que se conoce como transformada rápida de Fourier o FFT (*Fast Fourier Transform*).**
  - Simplemente es una manera eficiente de computar la DFT y su complejidad es de O(n log n), si n es una potencia de 2.

Predicción Lineal (I)

- **El método de predicción lineal o LP (*Linear Prediction*) [ITA70] [MAK75] [HAY96] [DEL93] es históricamente uno de los métodos más importantes para el análisis de la voz.**

- **Su fundamento se basa en establecer un modelo de filtro del tipo todo polo, para la fuente de sonido.**
  - La principal motivación del modelo todo polo viene dada porque permite describir la función de transferencia de un tubo, que sin pérdidas estuviese formado por diferentes secciones.

Predicción Lineal (II)

- **Constituye una aproximación razonable al habla producida por la excitación del tracto vocal causada por el conjunto de pulsos glotales.**

- **Objeciones:**
  - Los pulsos glotales no tienen una estructura espectral plana.
  - El tracto vocal no está compuesto de cilindros.
  - La cavidad nasal constituye un pasaje adicional.
  - Algunos sonidos se generan cerca de los labios como algunos sonidos fricativos sordos.

Predicción Lineal (III)

- Con un número suficiente de parámetros el modelo de predicción lineal puede constituir una aproximación adecuada a la estructura espectral de todo tipo de sonidos.

- El método de predicción lineal recibe este nombre porque pretende extrapolar el valor de la siguiente muestra de voz *x(n)* como la suma ponderada de muestras pasadas *x(n-1), x(n-2), ..., x(n-K)*:

$$\tilde{x}(n) = \sum_{i=1}^{K} a_i x(n-i)$$

Predicción Lineal (IV)

- Para ello se debe realizar el cálculo de los coeficientes ai minimizando alguna función de error *E*, concretamente de mínimos cuadrados, sobre una ventana de tamaño *N*.

$$\mathrm{E} = \sum_{\mathrm{n}=0}^{\mathrm{N-1}} \mathrm{e}^2(\mathrm{n}) = \sum_{n=0}^{N-1} \left(x(n) - \sum_{i=1}^{K} a_i x(n-i) \right)^2 \qquad 0 \le n \le N-1$$

Cepstrum (I)

- **Desde la introducción en los primeros años de la década de los 70, de las técnicas homomórficas de procesado de señal [OPP75] [PIC93], su importancia dentro del campo del reconocimiento de voz ha sido muy grande.**

- **Los sistemas homomórficos son una clase de sistemas no lineales que obedecen a un principio de superposición.**
  - De éstos, los sistemas lineales constituyen un caso especial.

Cepstrum (II)

- **La motivación para realizar un procesado homomórfico del habla viene resumida en la Figura 3:**

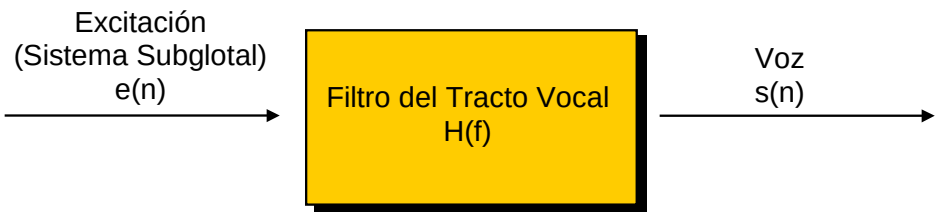


Figura 3. Las técnicas homomórficas pueden servir para separar la acción del tracto vocal (filtro lineal variable en el tiempo) de la señal de excitación.

Cepstrum (III)

- La señal de voz *s(n)* se descompone en una parte de excitación *e(n)* y en un filtro lineal $H(e^{i\theta})$.
- Así en el dominio frecuencia tenemos:

$$S(e^{i\theta}) = H(e^{i\theta})E(e^{i\theta})$$

- Para mayoría de las aplicaciones de voz sólo necesitamos la amplitud espectral.

$$\log(|S(e^{i\theta})|) = \log(|H(e^{i\theta})|) + \log(|E(e^{i\theta})|)$$

Cepstrum (IV)

- En el dominio logarítmico, las dos componentes anteriores pueden separarse empleando técnicas convencionales de procesamiento de señal.

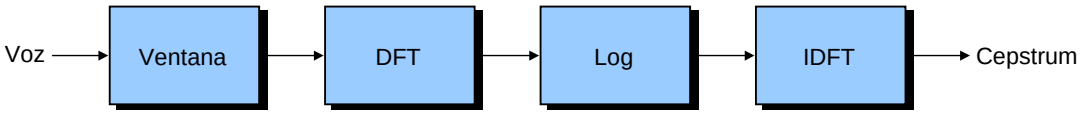


Figura 4. Análisis cepstral partiendo de la transformada discreta de Fourier.

$$c(n) = \frac{1}{N_s}\sum_{k=0}^{N_s-1}\log_{10}|S_{med}(k)|e^{j\frac{2\pi}{N_s}kn} \qquad 0 \le n \le N_s - 1$$

Cepstrum (V)

- **Es este caso, el valor *c(n)* se conoce como coeficientes cepstrales derivados de la transformada de Fourier. N_s es el número de puntos con los que se calculó la *DFT*.**
  - Esta ecuación también se conoce como la inversa de la DFT del espectro logarítmico.
    - Puede ser convenientemente simplificada teniendo en cuenta que el espectro logaritmo es una función real simétrica:

$$c(n) = \frac{2}{N_s} \sum_{k=1}^{N_s} S_{med}\left(I(k)\right) \cos\left(\frac{2\pi}{N_s} kn\right)$$

Cepstrum (VI)

- Lo habitual es usar solamente los primeros términos ($n \leq 20$). *I(k)* representa una función que traduce la posición de un valor en frecuencia al intervalo donde esté contenido.

- **Es posible a la hora de calcular un coeficiente cepstral, emplear bandas definidas según escalas de Mel.**

- **Este tipo de parámetros se conoce como coeficientes cepstrales con frecuencia en escalas de Mel o MFCC (*Mel Frequency Cepstral Coefficients*) [DAV80].**

Cepstrum (VII)

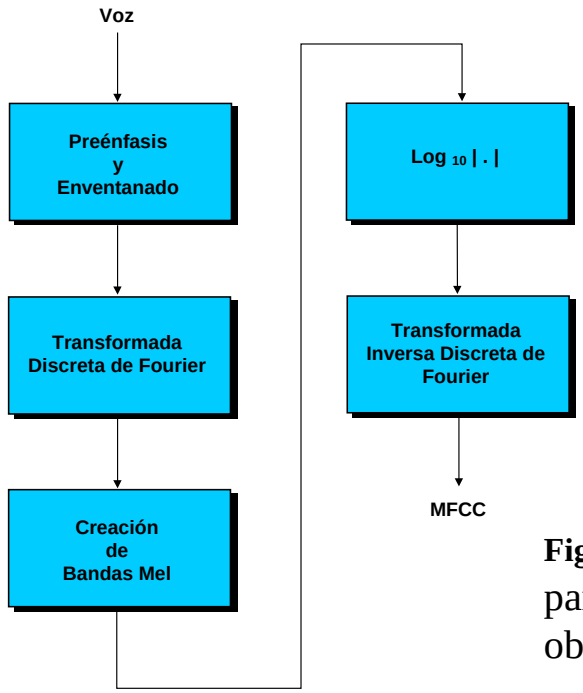


Figura 5. Esquema de parametrización para la obtención de MFCC.

Cepstrum (VIII)

- **Partiendo del análisis de predicción lineal también es posible obtener la expresión de los coeficientes cepstrales asociados:**

$$c(0) = \log(1) = 0$$

$$c(i) = -a(i) - \sum_{j=1}^{i-1} \left(1 - \frac{j}{i}\right) a(j) c(i-j) \quad 1 \le i \le N_c$$

Cepstrum (IX)

- Una transformación usual sobre este tipo de coeficientes es lo que se conoce como coeficientes cepstrales delta o coeficientes delta cepstrum.

- La expresión que permite obtener estos últimos es:

$$\Delta c_j(i) = \frac{1}{2T+1} \sum_{k=-T}^{T} k \cdot c_{j+k}(i)$$

Predicción Lineal Perceptual (I)

- La técnica de predicción lineal perceptual o PLP (Perceptual Linear Prediction) [HER90], es en esencia una combinación de las técnicas de la transformada discreta de Fourier y de predicción lineal como puede verse en la Figura 6.
  - Para obtener el análisis de banda crítica se utiliza primeramente la transformada discreta de Fourier con una ventana de Hamming de 20 ms. Posteriormente se calcula el espectro de potencia y se transfiere a una escala de Bark.

Predicción Lineal Perceptual (II)

- El segundo paso consiste en la igualación de las alturas perceptuales tiene su origen en la necesidad de compensar la diferente percepción de alturas sonoras para diferentes frecuencias.
- Tras la IDFT se calculan los coeficientes de autorregresión de un modelo todo polo.
- Adicionalmente se pueden calcular a partir de éstos los coeficientes cepstrales.

Predicción Lineal Perceptual (III)

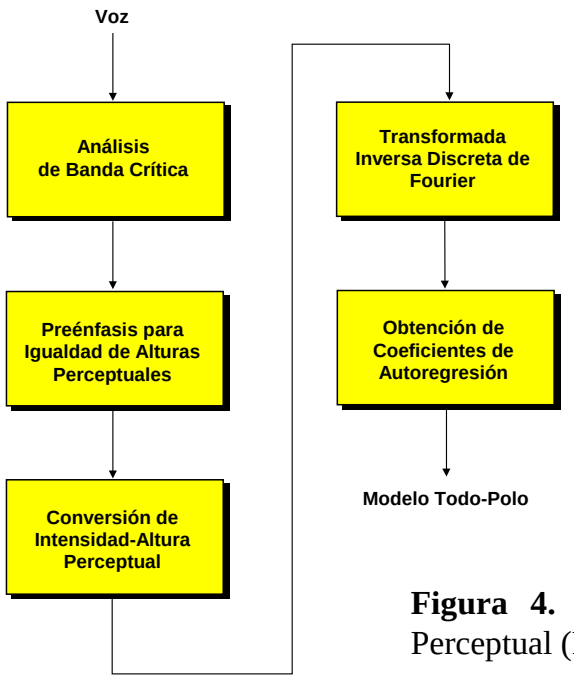


Figura 4. Predicción Lineal Perceptual (PLP).

Predicción Lineal Perceptual (IV)

- **Como extensión de la técnica anteriormente descrita, encontramos el método RASTA-PLP (RelAtive SpecTrAl) [HER92] [HER94].**
  - La motivación de este complemento viene dada por el intento de robustecer el mecanismo del algoritmo frente a distorsiones lineales en el espectro, por ejemplo debidas al canal de comunicación.
  - Una extensión del algoritmo RASTA es la conocida como J-RASTA [KOE94], que puede también compensar el ruido cuando la relación señal/ruido es baja.

Bibliografía (I)

[COO92] J. W. Cooley, "How the FFT Gained Acceptance", *IEEE SP Magazine*, enero 1992, pp. 10-13.

[DAV80] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosillabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. ASSP-28, Nº 4, agosto 1980, pp. 357-366.

[DEL93] J. R. Deller, J. G. Proakis and J. H. L. Hansen, Discrete-Time Processing of Speech Signals, Mac Millan, N. Y., 1993.

[DEL94] J. R. Deller, JR., "Tom, Dick and Mary Discover the DFT", *IEEE Signal Processing Magazine*, abril 1994, pp. 36-50.

[HAY96] S. Haykin, *Adaptive Filter Theory*, 2nd Ed., Prentice Hall, Englewood Cliffs, N. J., 1996

[HER90] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *Journal of Acoustic Society of America*, Vol. 87, Nº 4, abril 1990, pp. 1738-1752.

[HER92] H. Hermansky et al., "RASTA-PLP speech analysis technique", *Proc. of ICASSP'92*, San Francisco, Estados Unidos, 23-26 marzo 1992, pp. 121-124.

Bibliografía (II)

[HER94] H. Hermansky and N. Morgan, "RASTA Processing of Speech", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, Nº. 4, octubre 1994, pp. 578-589.

[ITA70] F. Itakura and S. Saito, "A Statical Method for Estimation of Speech Spectral Density and Formant Frequencies", *Electron. Communication*, Vol. 53, pp. 36-43

[KOE94] J. Koehler et al., "Integrating RASTA-PLP Into Speech Recognition", *Proc. of ICASSP'87*, Dallas, Estados Unidos, 6-9 abril 1987, Vol. I, pp. 421-424.

[KRA94] P. Kraniauskas, "A Plain Man´s Guide to the FFT", *IEEE Signal Processing Magazine*, abril 1994, pp. 24-35.

[MAK75] J. Makhoul, "Linear Prediction: A tutorial Review", *Proc. of the IEEE*, Vol. 63, abril 1975, pp. 124-143.

[OPP75] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice Hall, Englewood Cliffs, N. J., 1975.

[PIC93] J. W. Picone, "Signal Modeling Techniques in Speech Recognition", Proc. of the IEEE, Vol. 81, Nº 9, septiembre 1993, pp. 1215-1247.

[ROB98] A. Robinson, "Speech Analysis", ftp://svr-ftp.eng.cam.ac.uk/pub/com.sppech/info, Lent Term 1998.