

Introducción al Procesamiento de Lenguaje Natural (NLP)

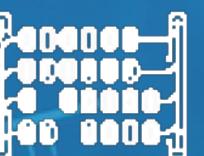
Clase II: Sistemas de Reconocimiento
Automático de Voz (ASR) II. Mel-Cepstrum y HMM

Alexander Caicedo



Universidad del
Rosario

Escuela de Ingeniería,
Ciencia y Tecnología

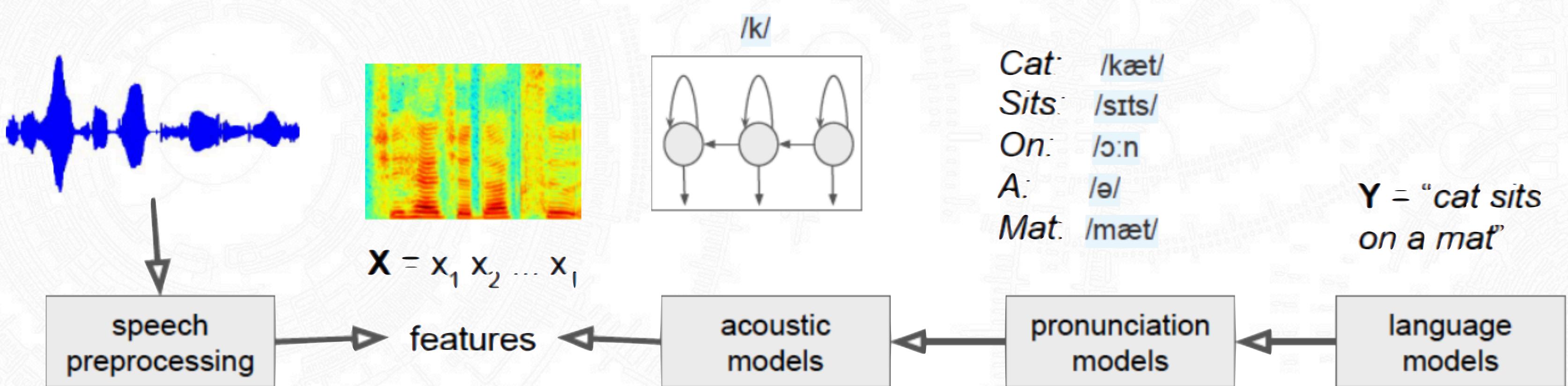


MACC
Matemáticas Aplicadas y
Ciencias de la Computación

Contenido

Sistemas ASR.
Caracterización de la señal de voz.
HMM.

ASR Systems



ASR Systems: Modelo de Lenguaje

N-grams: Establecer las probabilidades de que la palabra que siga, dadas N palabras anteriores. Esta información se saca de un corpus (puede ser la web).

$$P(\text{the}|\text{its water is so transparent that}).$$



$$\begin{aligned} P(\text{the}|\text{its water is so transparent that}) &= \\ \frac{C(\text{its water is so transparent that the})}{C(\text{its water is so transparent that})} \end{aligned}$$

i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025
eat	0	0	0.0027	0	0.021	0.0027	0.056
chinese	0.0063	0	0	0	0	0.52	0.0063
food	0.014	0	0.014	0	0.00092	0.0037	0
lunch	0.0059	0	0	0	0	0.0029	0
spend	0.0036	0	0.0036	0	0	0	0

Figure 3.2 Bigram probabilities for eight words in the Berkeley Restaurant Project corpus of 9332 sentences. Zero probabilities are in gray.

$$P(i|<\text{s}>) = 0.25$$

$$P(\text{english}|\text{want}) = 0.0011$$

$$P(\text{food}|\text{english}) = 0.5$$

$$P(</\text{s}>|\text{food}) = 0.68$$



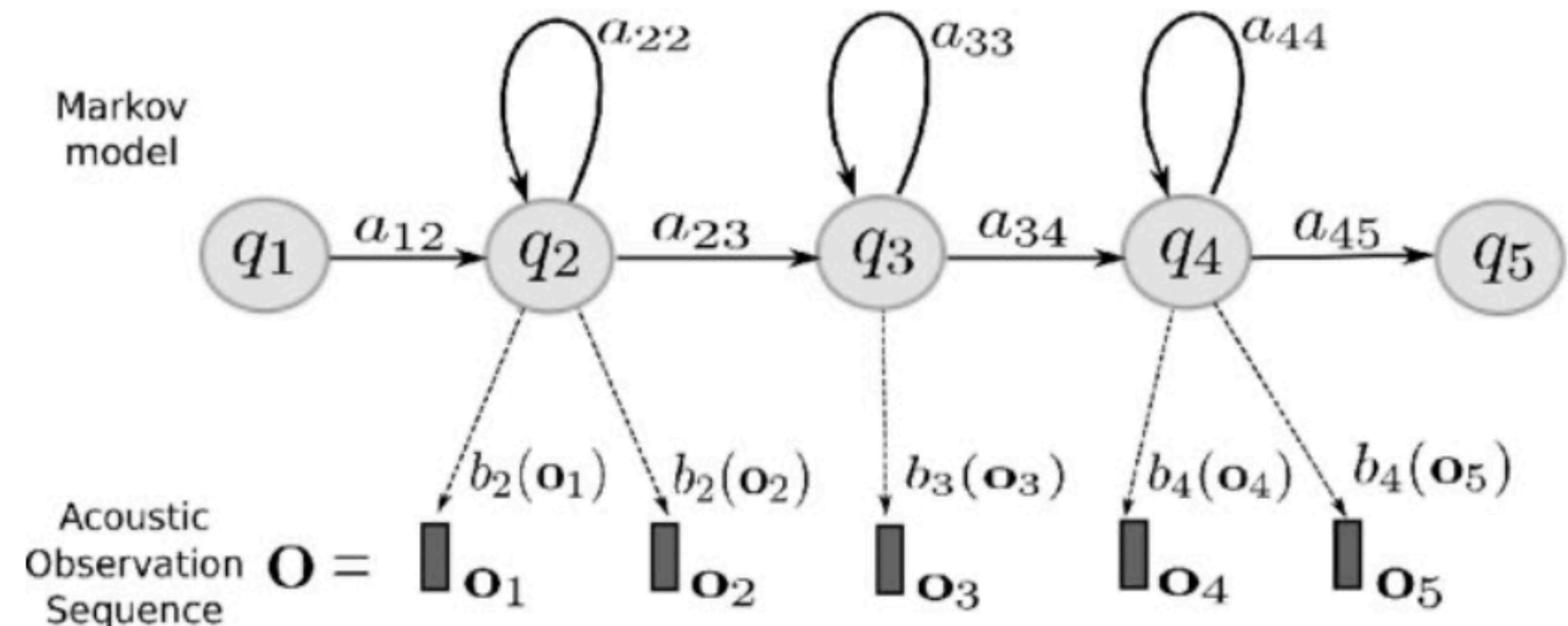
$$\begin{aligned} P(<\text{s}> i \text{ want english food } </\text{s}>) &= \\ P(i|<\text{s}>)P(\text{want}|i)P(\text{english}|want) \\ &\quad P(\text{food}|\text{english})P(</\text{s}>|\text{food}) \\ &= .25 \times .33 \times .0011 \times 0.5 \times 0.68 \\ &= .000031 \end{aligned}$$

ASR Systems: Modelo de Pronunciación

Tablas de Pronunciación: Estas son dadas por lingüistas para cada palabra del diccionario. Dividen las palabras en los sonidos que la componen (phonems o graphems). Tenga en cuenta que algunas palabras pueden tener la misma forma de ser pronunciadas.

ASR Systems: Modelo Acustico

HMM y GMM:



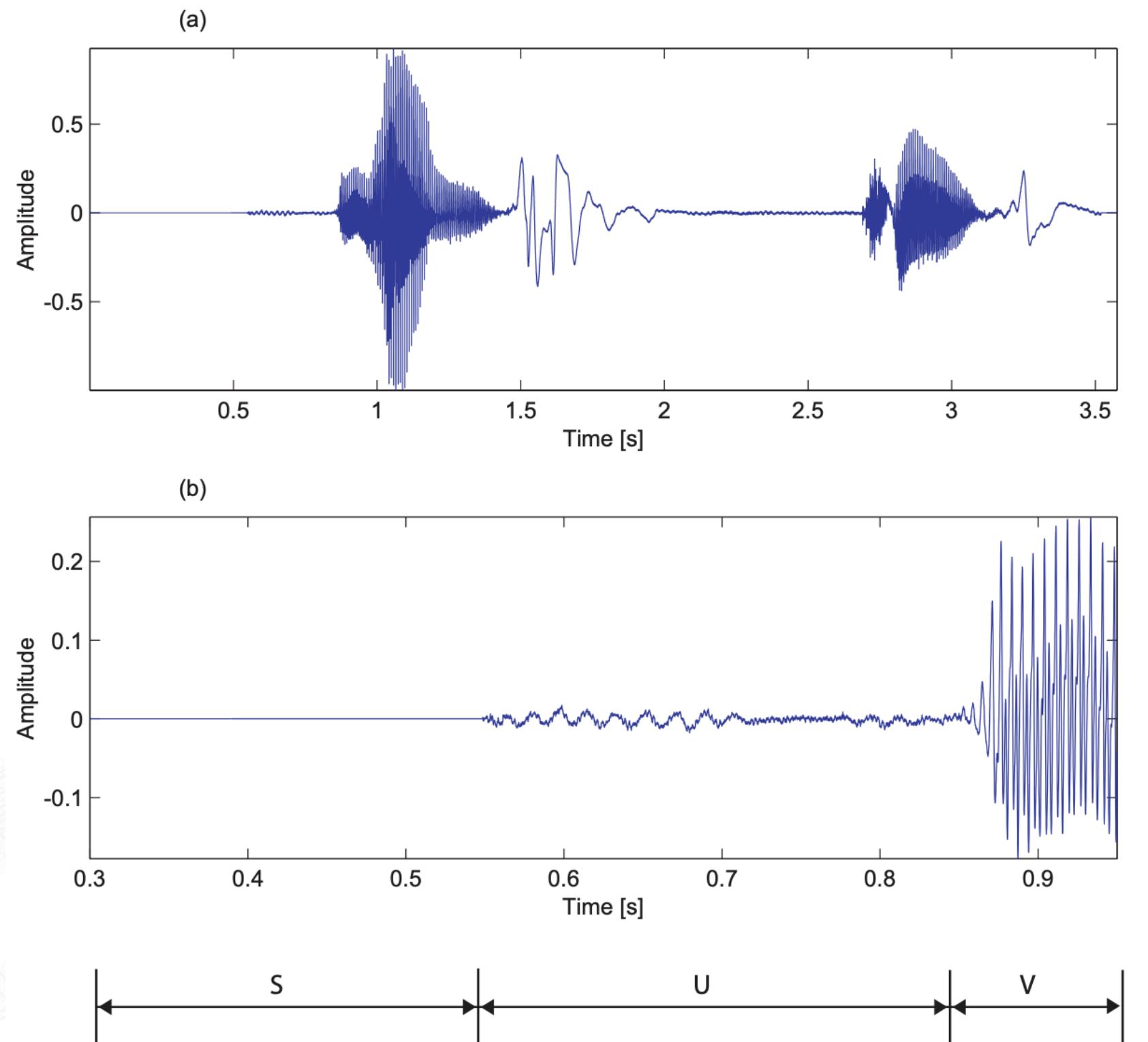
Problemas del ASR Clásico

- Requiere de tablas que son muy grandes y consumen mucha memoria.
- Sistemas deben conectarse con la nube (donde están almacenadas las tablas) para poder funcionar de forma correcta.
- El performance de estos sistemas es limitado.
- Su performance depende de la calidad del modelo de pronunciación, y del modelo de lenguaje.
- Se debe entrenar cada modelo por separado y al final concatenarlos, lo cual hace que se propaguen los errores de un modelo a otro.

Contenido

Sistemas ASR.
Caracterización de la señal de voz.
HMM.

Caracterización de la señal de Voz



La señal de voz puede ser caracterizada por tres estados:

- Silencio
- No vocal
- Vocal

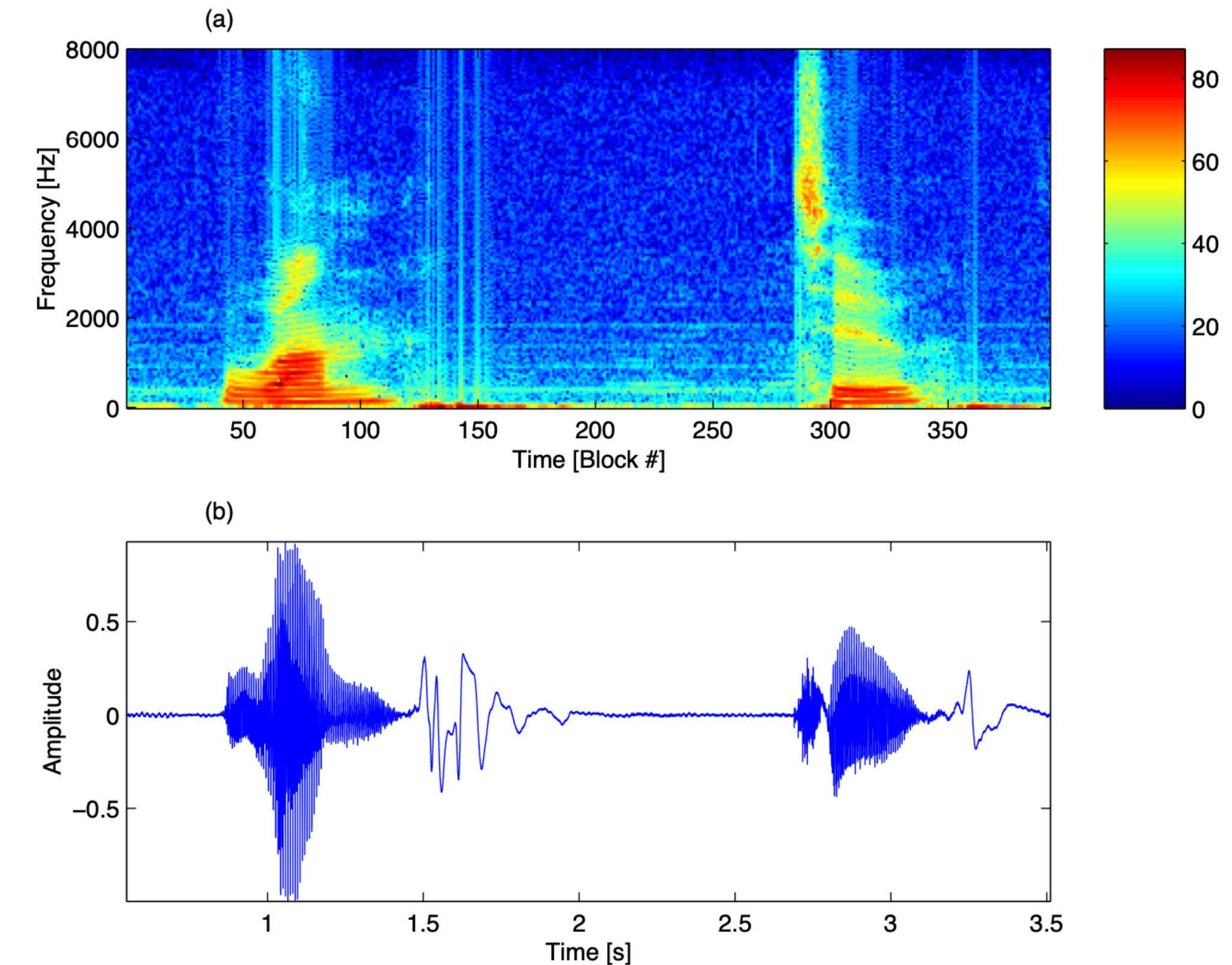
El objetivo es poder caracterizar estos estados, y a su vez relacionarlos con fonemas

Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Caracterización de la señal de Voz

Representación en tiempo-frecuencia permite reunir la dimensionalidad de la señal de voz, y obtener descriptores mas apropiados para su procesamiento.

Segmentos de 20ms, tradicionalmente con un overlap de 10ms.



Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Caracterización de la señal de Voz

¿Porqué la representación en frecuencia es apropiada para la caracterización de las señales de voz?

Caracterización de la señal de Voz

¿Porqué la representación en frecuencia es apropiada para la caracterización de las señales de voz?

Porque el lenguaje hablado se puede dividir en unidades fundamentales de sonido llamadas Fonemas,. La concatenación de estos fonemas es lo que forma las palabras que pronunciamos. Existe una relación (casi) única entre fonemas y la forma cómo se escriben las palabras.

En Español hay cerca de 24 Fonemas, en Ingles son 46

Caracterización de la señal de Voz

¿Cómo procesar la señal de voz?

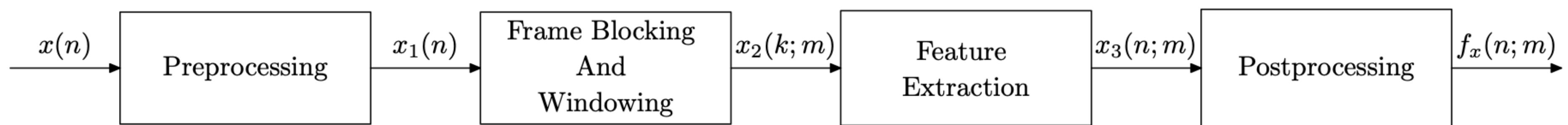
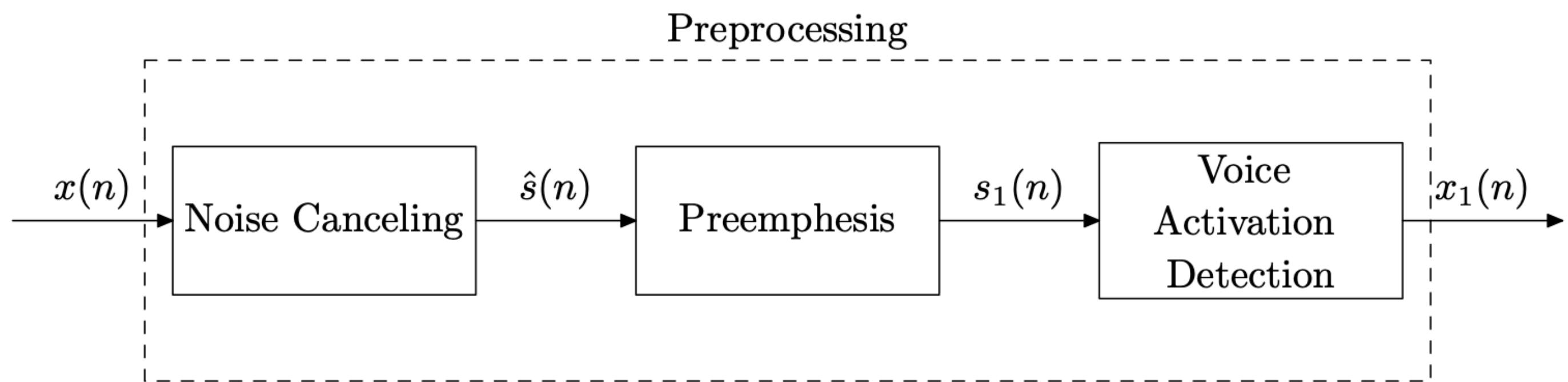


Figure 3.1: Main steps in Feature Extraction

Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Caracterización de la señal de Voz

Pre-procesamiento



- Mejorar el SNR
- Detectar cuándo se produce la señal.

Figure 3.2: Steps in Preprocessing

Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Caracterización de la señal de Voz

Voice Activation Detection

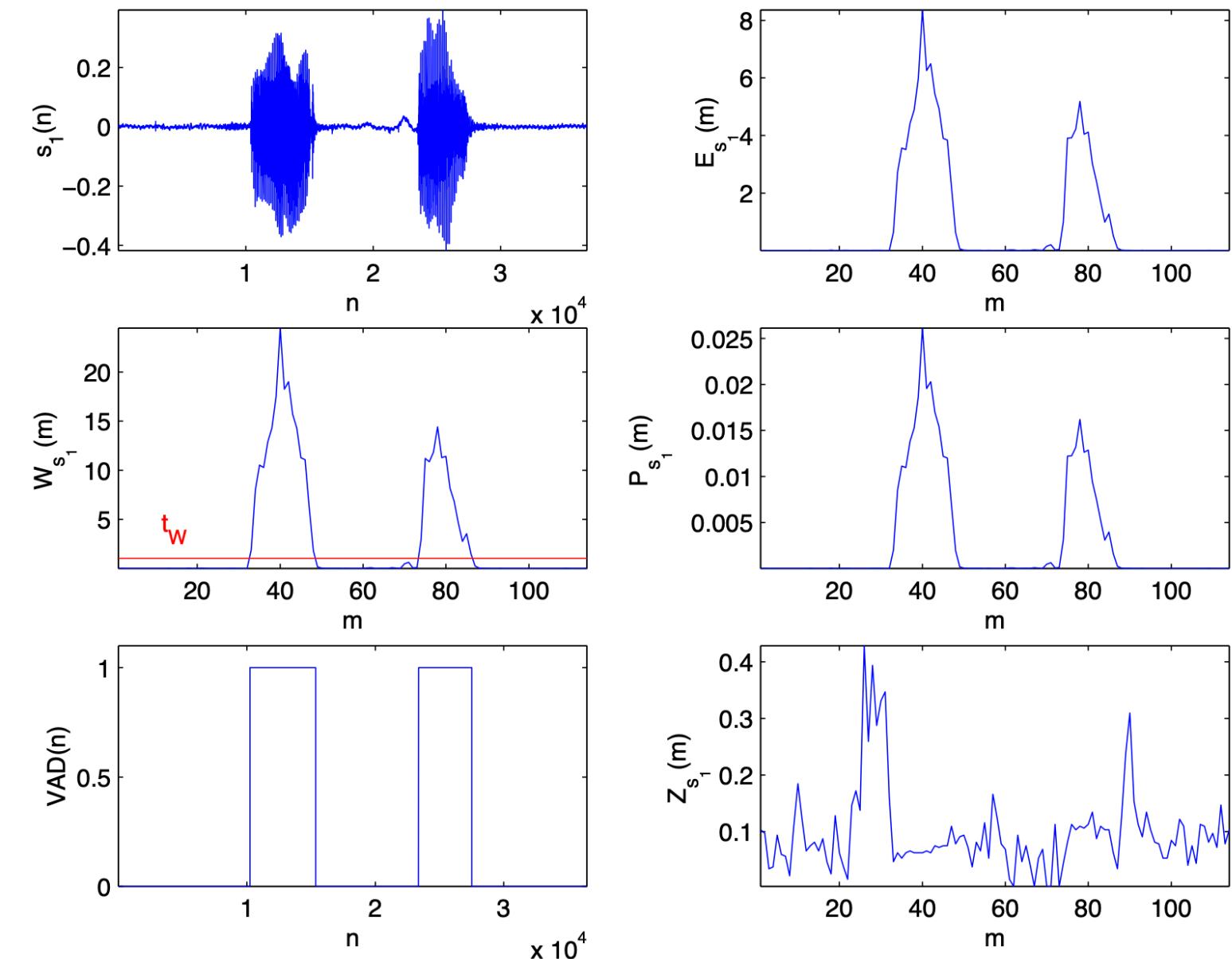


Figure 3.4: Different measures used to find speech content, $VAD(n)$

$$E_{s_1}(m) = \sum_{n=m-L+1}^m s_1^2(n)$$

$$P_{s_1}(m) = \frac{1}{L} \sum_{n=m-L+1}^m s_1^2(n)$$

$$Z_{s_1}(m) = \frac{1}{L} \sum_{n=m-L+1}^m \frac{|\operatorname{sgn}(s_1(n)) - \operatorname{sgn}(s_1(n-1)|}{2}$$

Energía

Potencia

Cruces por cero

Estimación de la variable de activación

$$W_{s_1}(m) = P_{s_1}(m) \cdot (1 - Z_{s_1}(m)) \cdot S_c$$

$$VAD(m) = \begin{cases} 1, & W_{s_1}(m) \geq t_W \\ 0, & W_{s_1}(m) < t_W \end{cases}$$

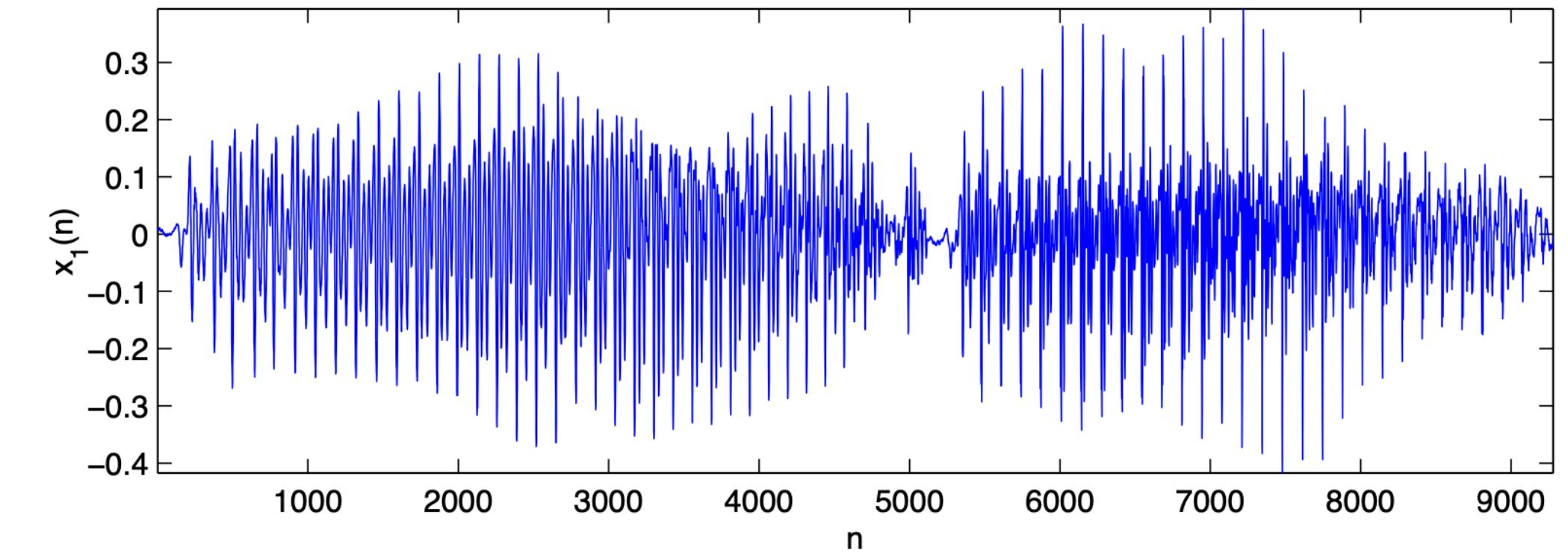
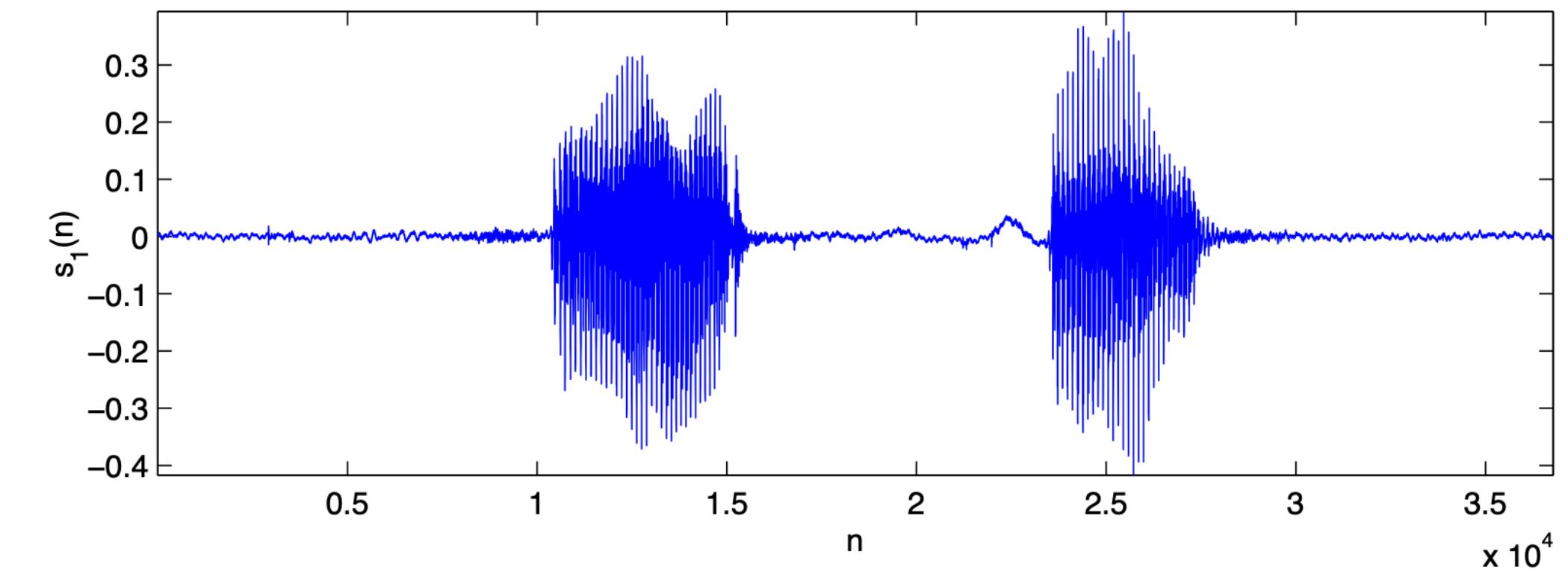
El valor de t_W se define usando segmentos de silencio.

Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Caracterización de la señal de Voz

Voice Activation Detection

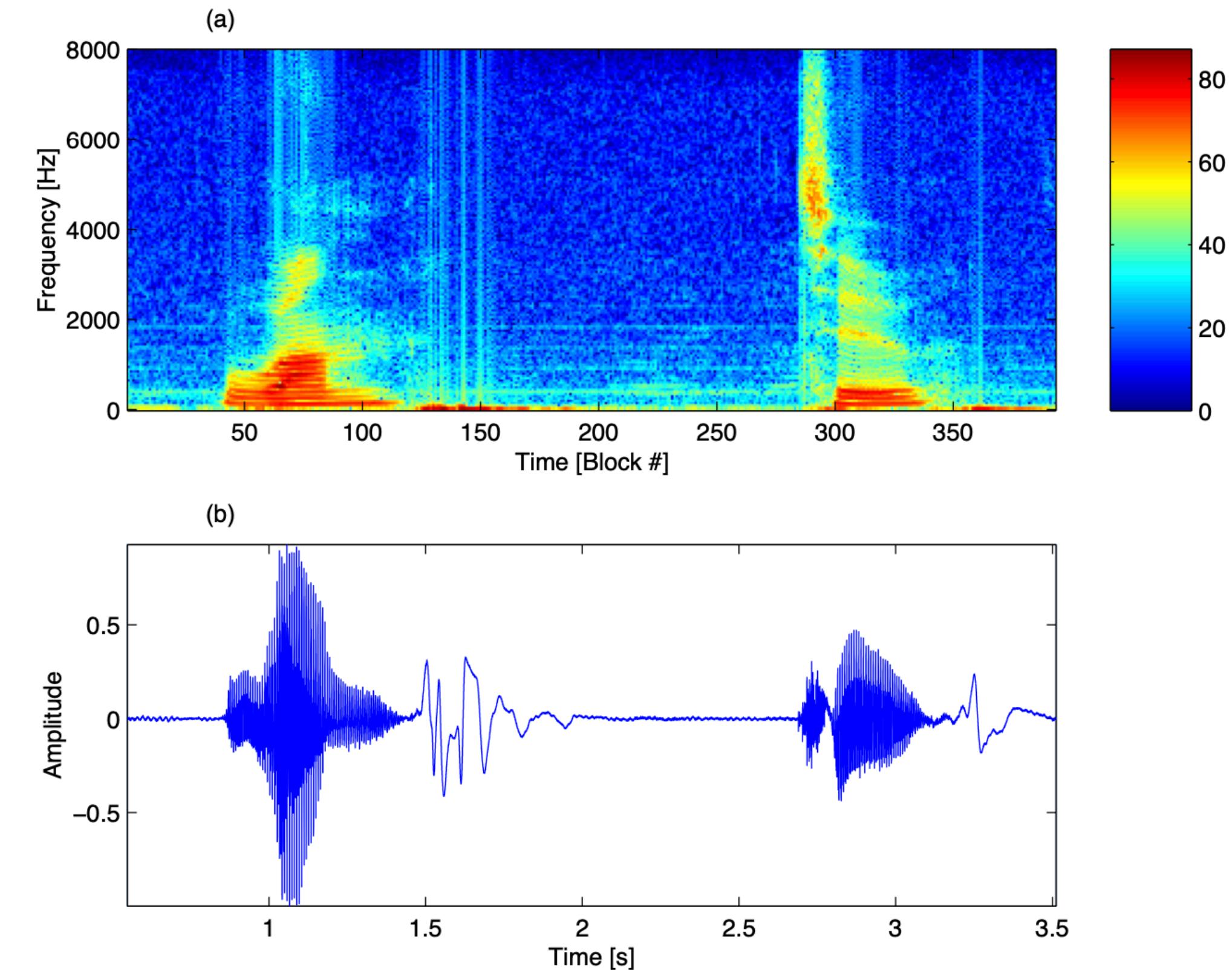
Resultado del sistema de
detección automática de voz



Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Caracterización de la señal de Voz

¿Cómo se puede detectar a partir del espectrograma segmentos de voz y de silencio?



Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Caracterización de la señal de Voz

Cepstrum

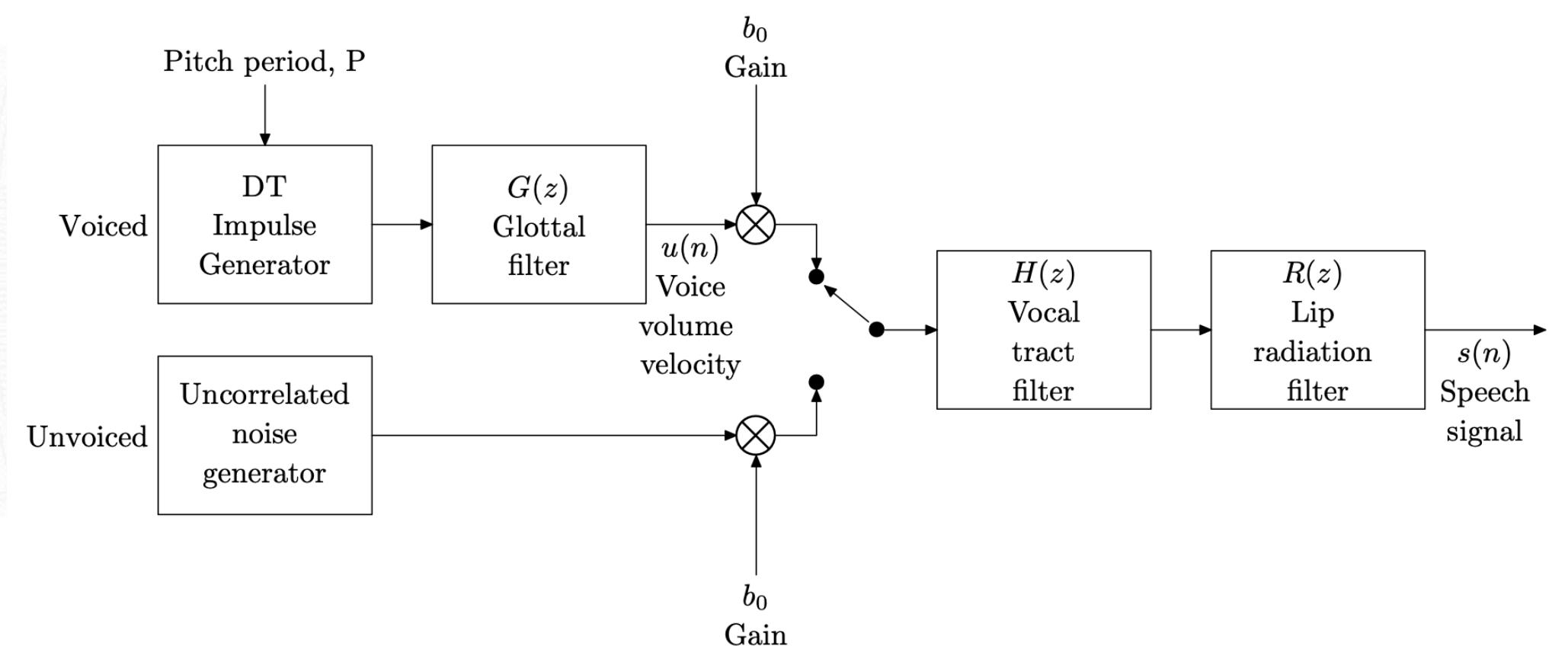


Figure 2.3: Discrete-Time Speech Production Model

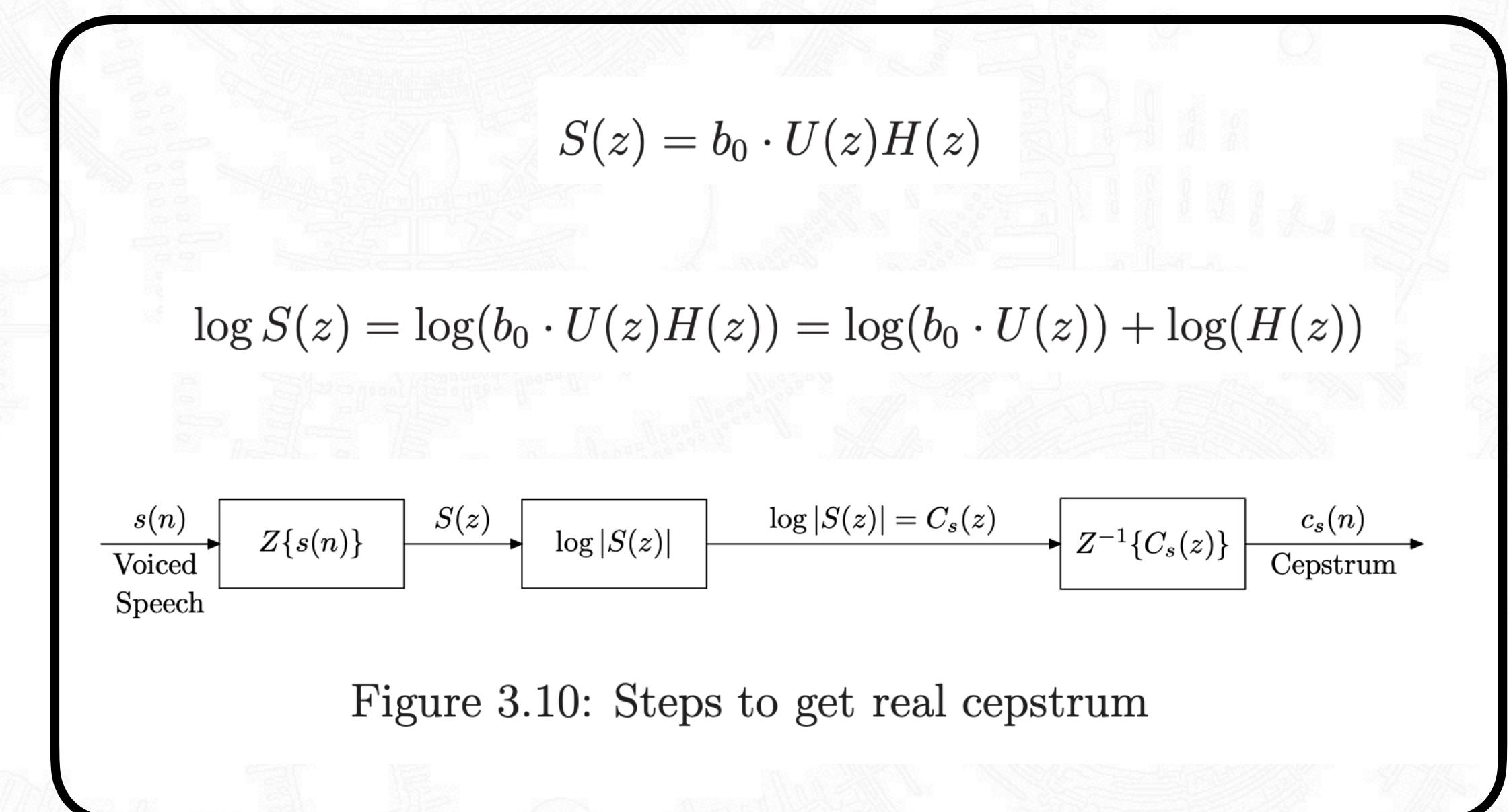


Figure 3.10: Steps to get real cepstrum

Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Caracterización de la señal de Voz

Cepstrum

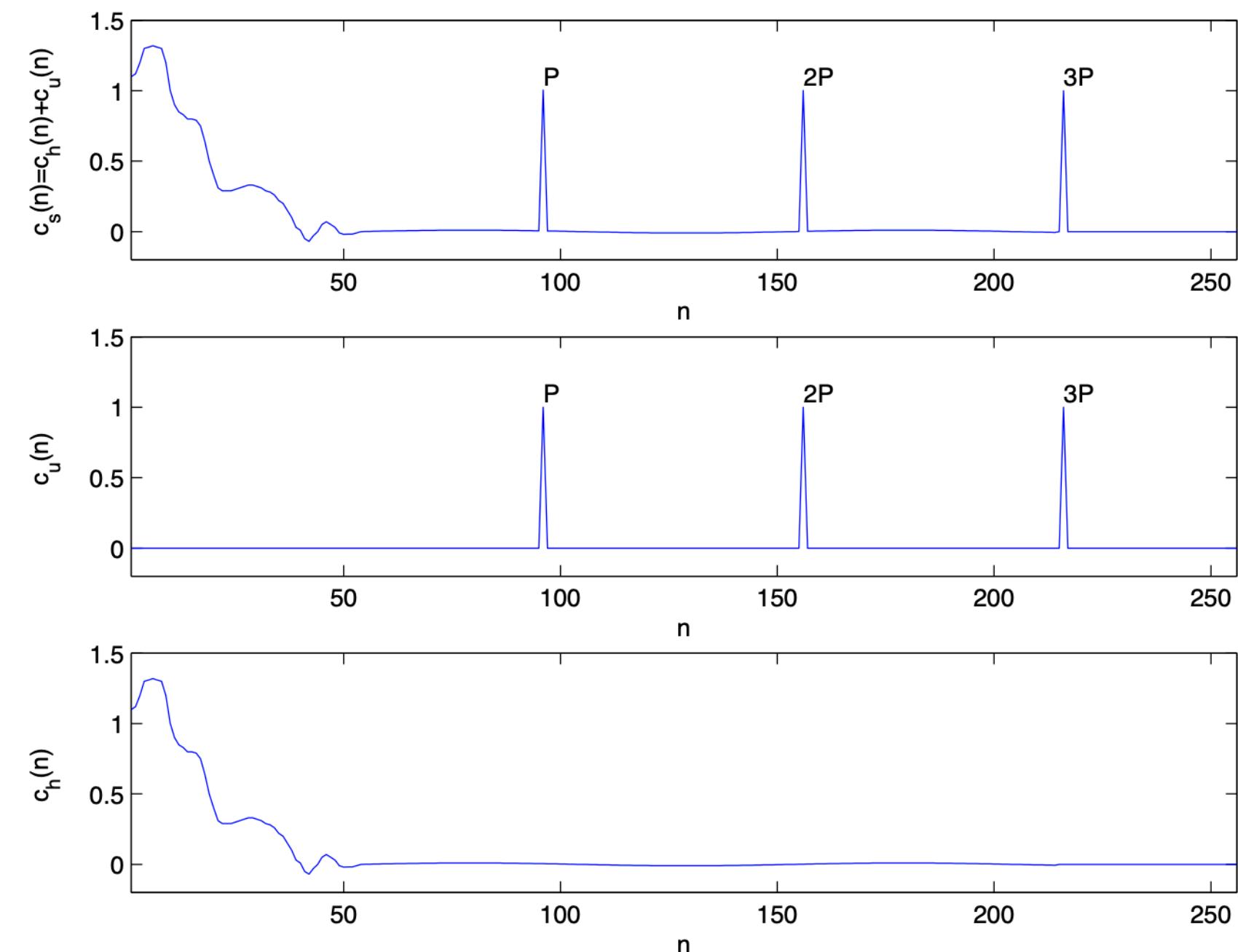


Figure 3.11: Quefrency functions for vocal tract model

Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

El tracto vocal contiene información del formante pronunciado

Caracterización de la señal de Voz

Mel-Scale

El sistema auditivo no percibe las frecuencias en una escala lineal

- Más sensible a cambio en bajas frecuencias.
- Menos sensible a cambios en altas frecuencias.

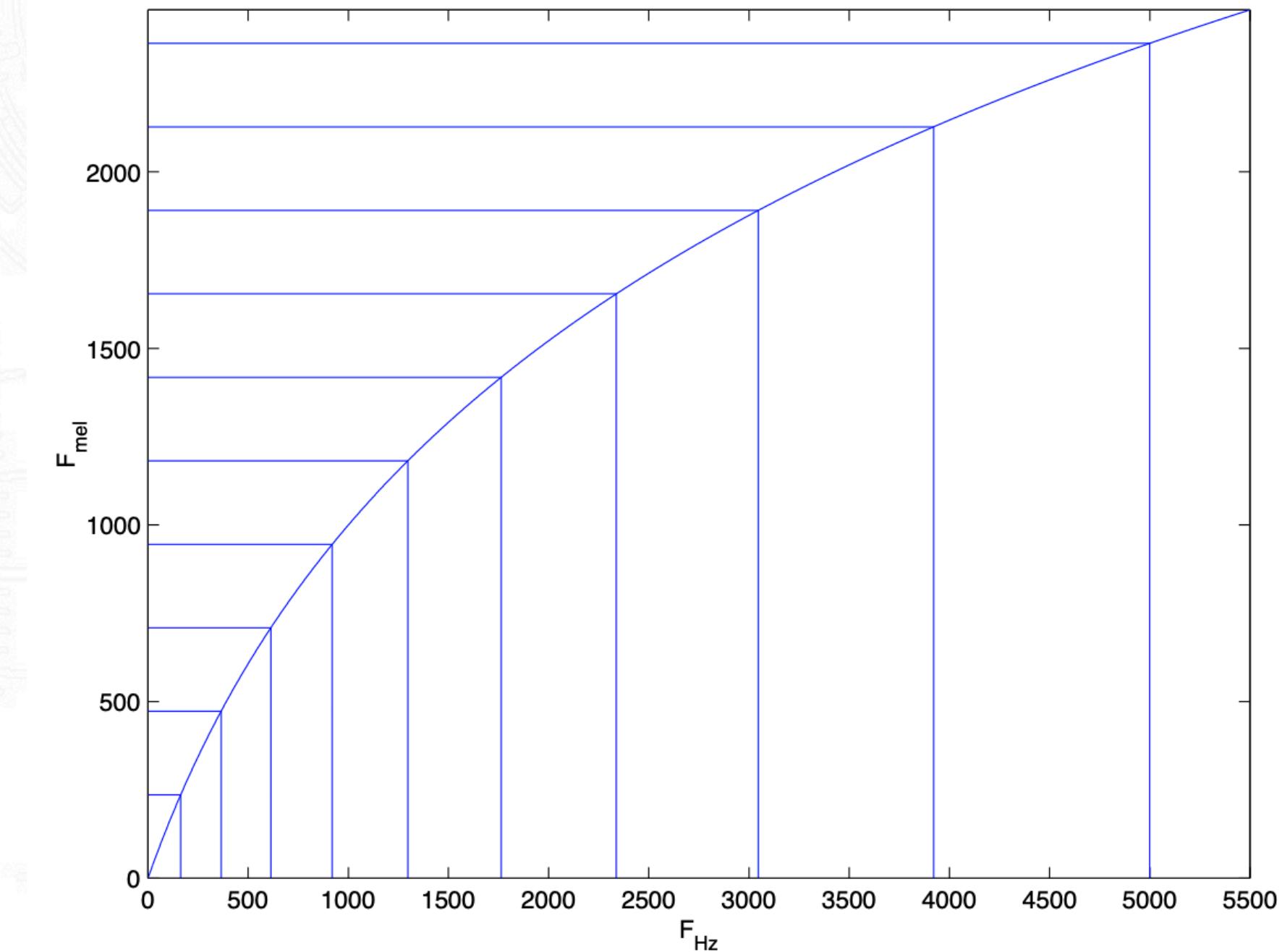


Figure 3.14: Equally spaced mel values

$$F_{\text{mel}} = 2595 \cdot \log_{10}\left(1 + \frac{F_{\text{Hz}}}{700}\right)$$

Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Caracterización de la señal de Voz

Mel-Scale

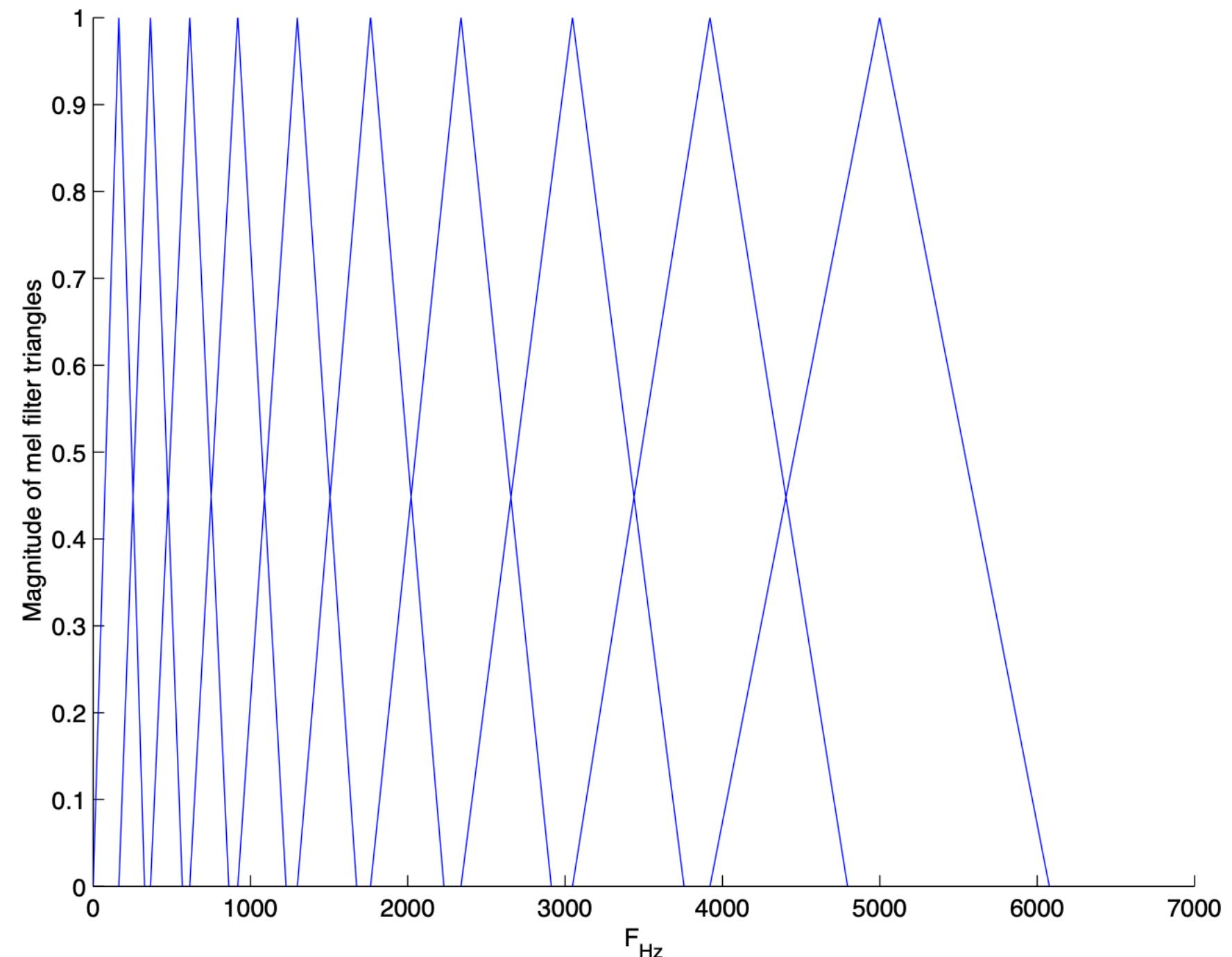


Figure 3.15: Mel scale filter bank

Para cada mel banda se calcula un coeficiente (el área en esa banda) y a ese resultado se le calcula el cepstrum.

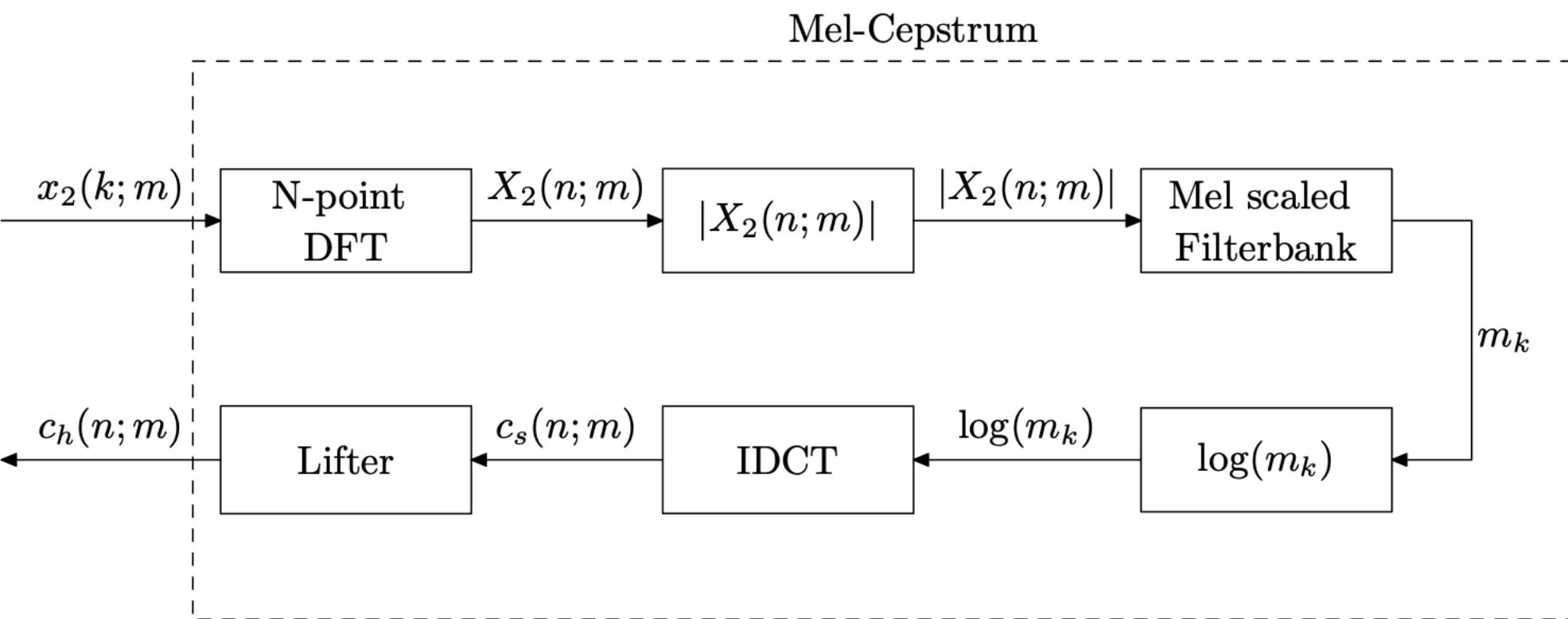


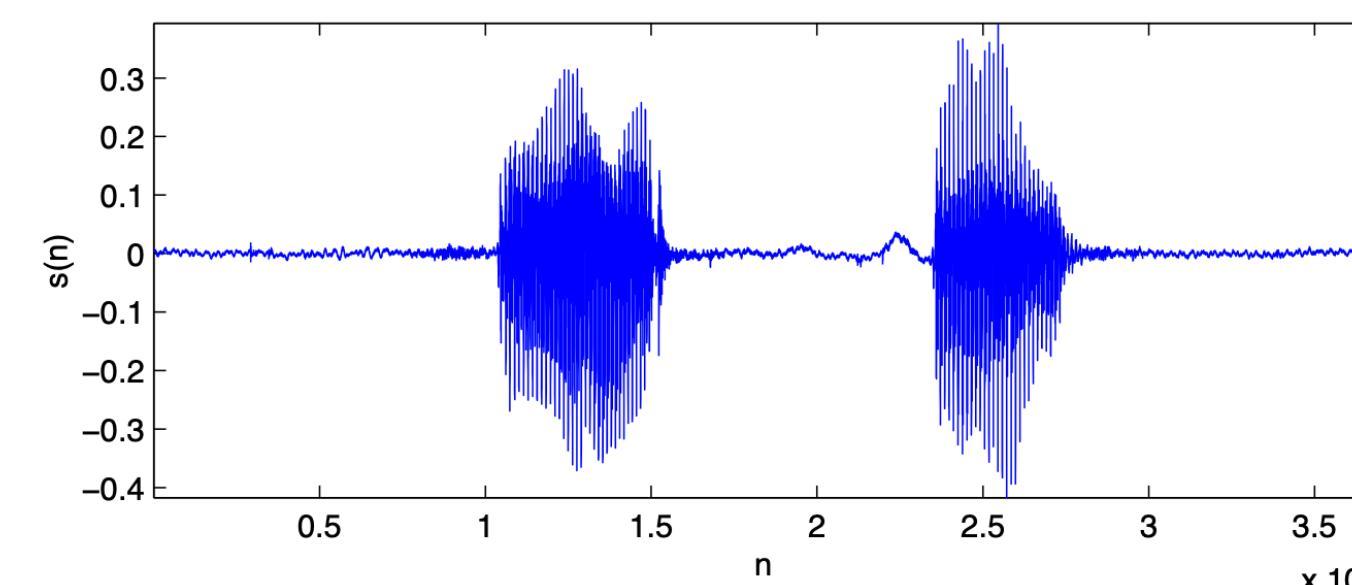
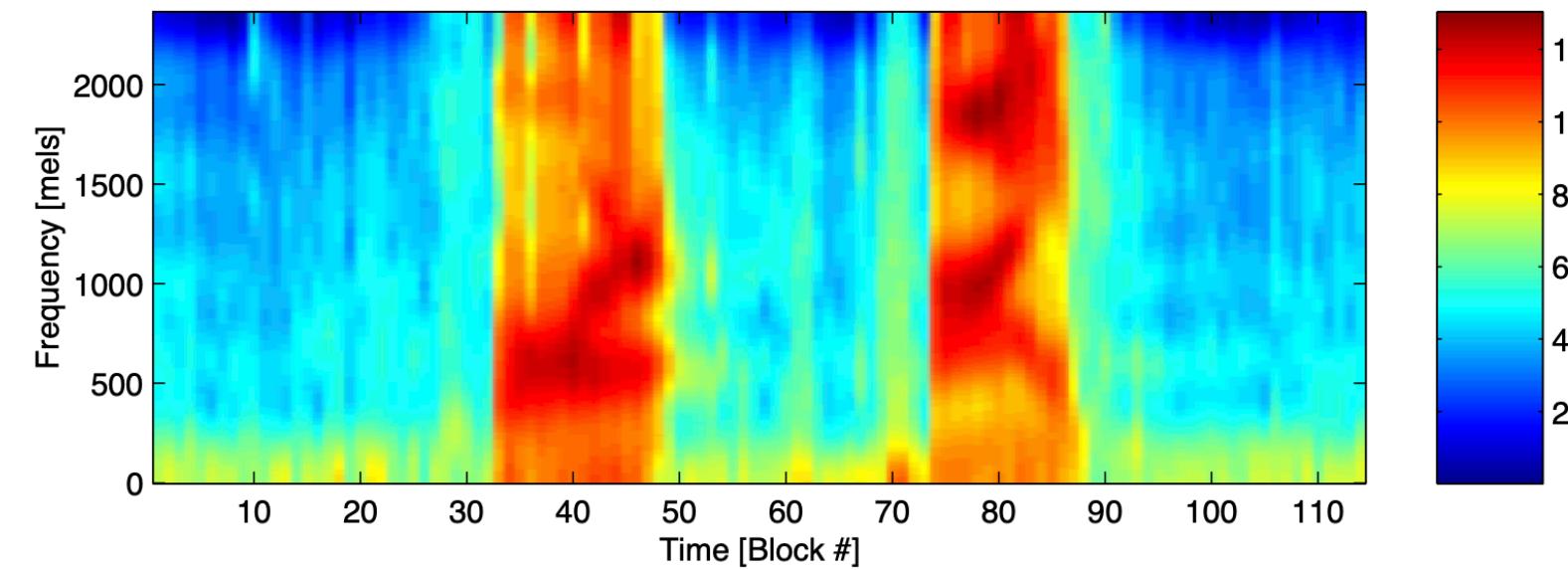
Figure 3.16: The steps in creation of Mel-Cepstrum

Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Caracterización de la señal de Voz

Mel-Scale

Espectrograma basado en Mel-Cepstrum



$$|H(n; m)| = |e^{FFT_N\{c_h(n; m)\}}|$$

Espectrograma Normal

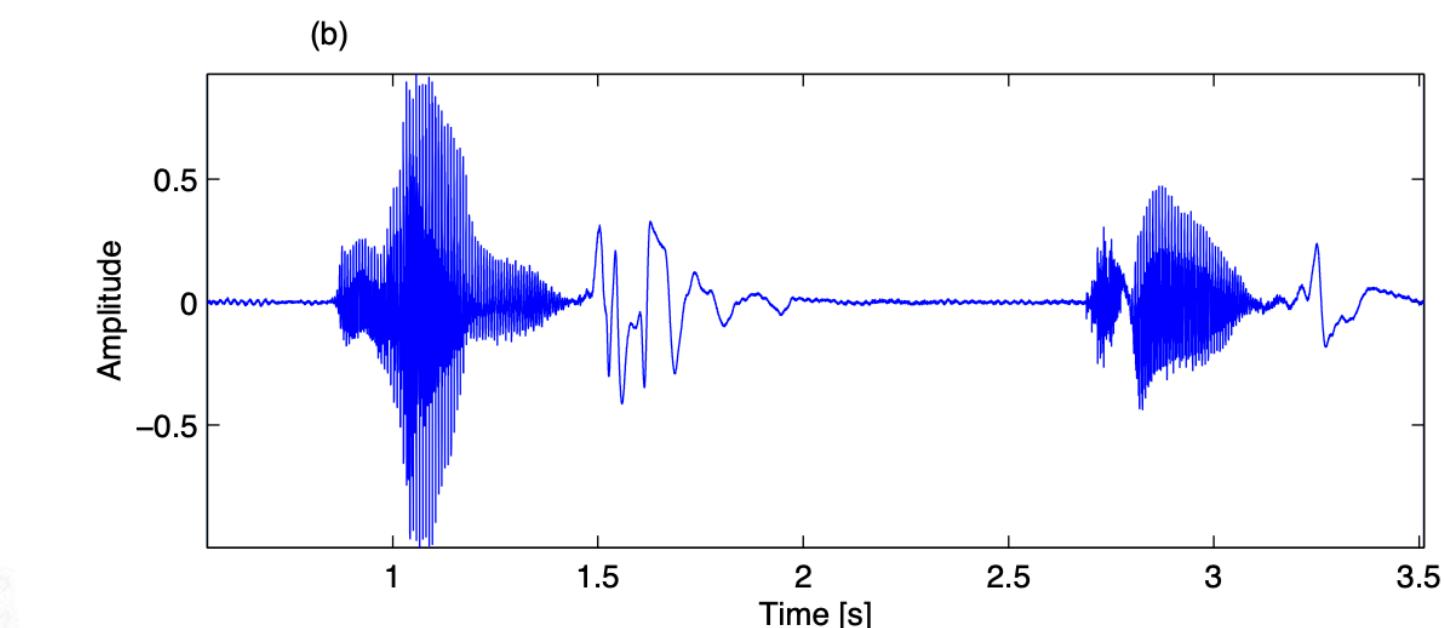
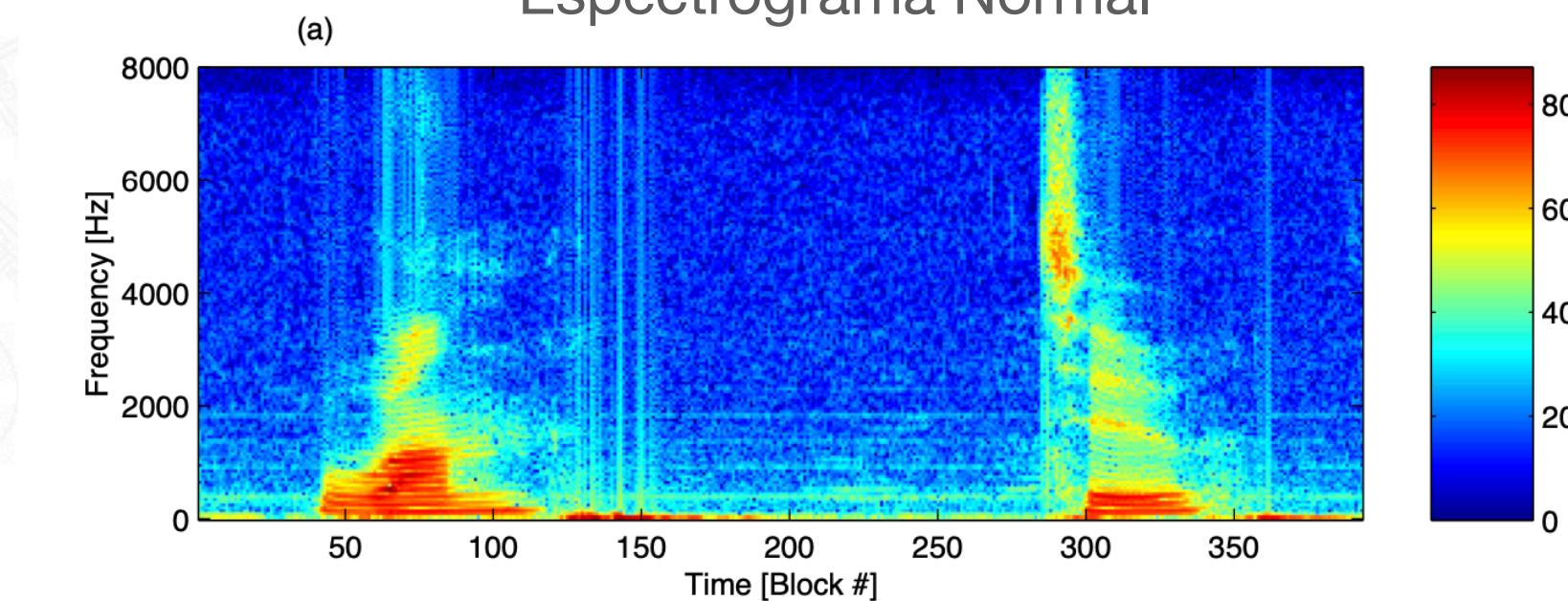


Figure 3.18: Spectral representation of cepstrum coefficients

También se puede usar el Escalograma (Wavelet). Su versión discreta incluye el Mel-scaling

Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Contenido

Sistemas ASR.
Caracterización de la señal de voz.
HMM.

Hidden Markov Models

HMM Clásico con observaciones deterministas

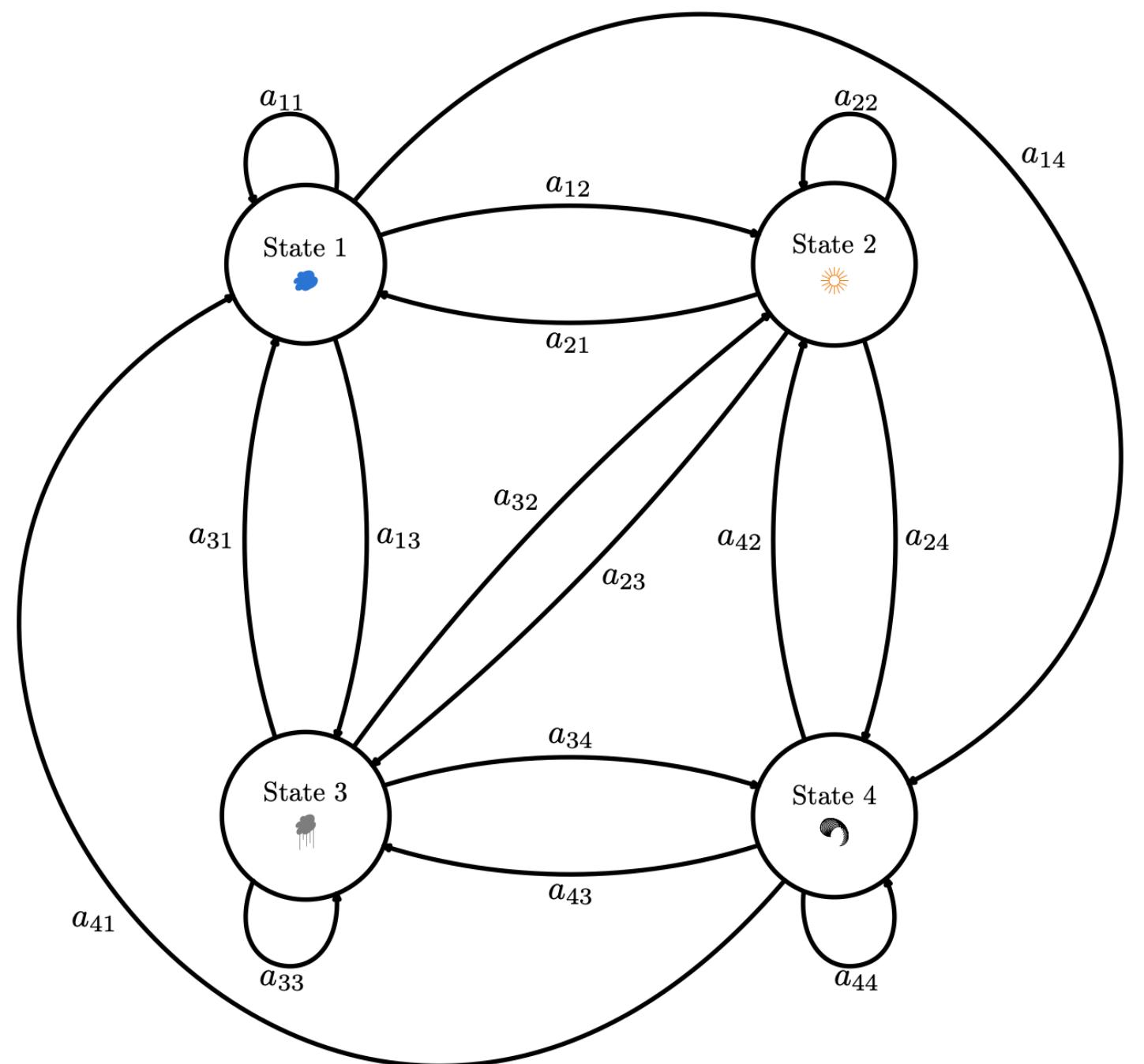


Figure 4.1: Markov model of the weather

- **State 1:** cloudy
- **State 2:** sunny
- **State 3:** rainy
- **State 4:** windy

Probabilidades de Transición

$$\begin{aligned} a_{ij} &\geq 0 & \forall j, i \\ \sum_{j=1}^N a_{ij} &= 1 & \forall i \end{aligned}$$

Matrix de Transición de Probabilidades

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \dots & a_{NN} \end{bmatrix}_{N \times N}$$

Distribución Inicial de los estados

$$\boldsymbol{\pi} = \begin{bmatrix} \pi_1 = P(q_1 = 1) \\ \pi_2 = P(q_1 = 2) \\ \vdots \\ \pi_N = P(q_1 = N) \end{bmatrix}_{N \times 1}$$
$$\sum_{i=1}^N \pi_i = 1$$

Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Hidden Markov Models

HMM Clásico con observaciones deterministas

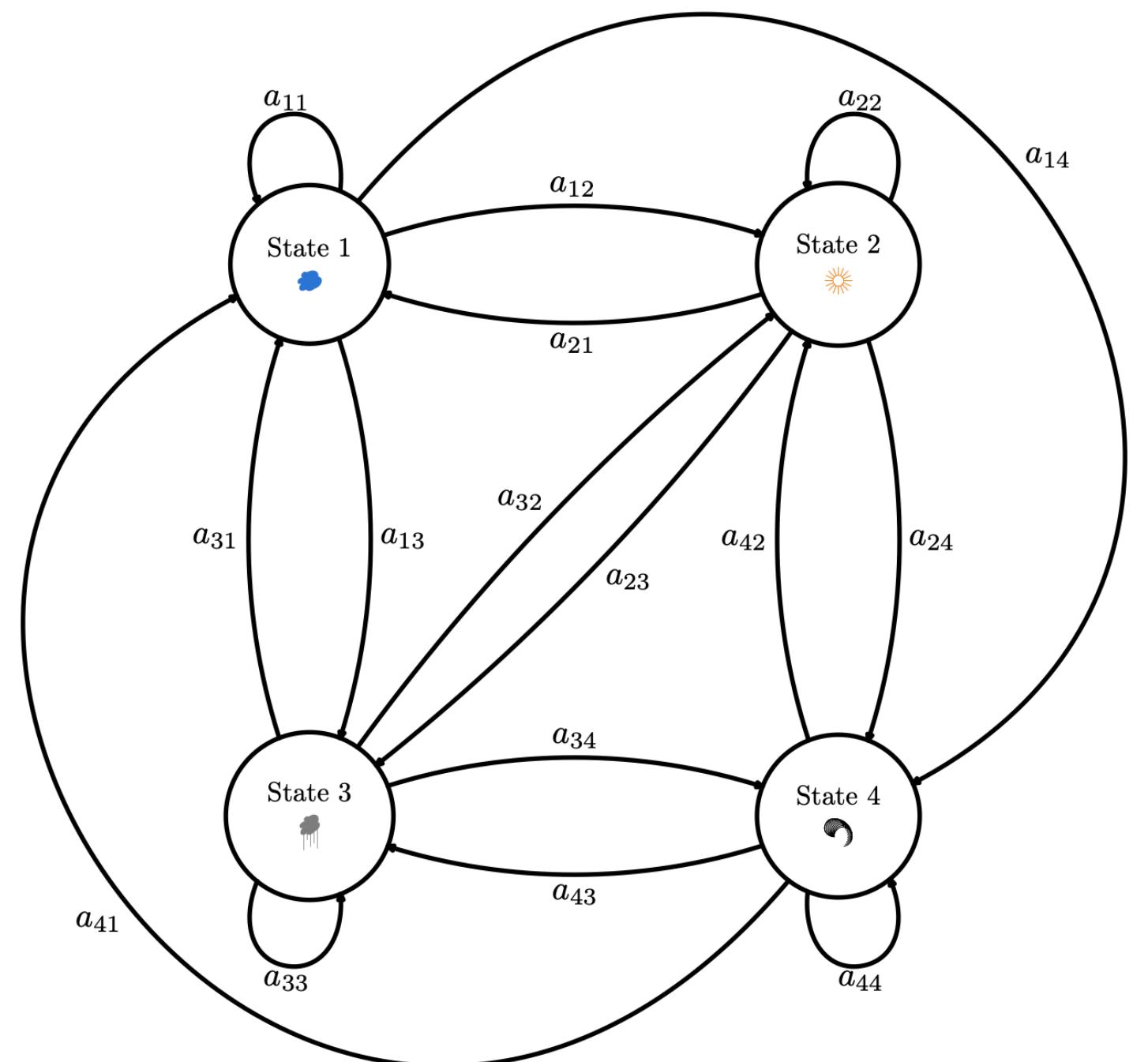


Figure 4.1: Markov model of the weather

En el modelo HMM propuesto, ¿cuál es la probabilidad de que la secuencia de eventos O ocurra?

$$\mathbf{O} = (\text{sunny, rainy, sunny, windy, cloudy, cloudy}) = (2, 3, 2, 4, 1, 1)$$

- **State 1:** cloudy
- **State 2:** sunny
- **State 3:** rainy
- **State 4:** windy

$$\begin{aligned} P(\mathbf{O}|\mathbf{A}, \boldsymbol{\pi}) &= P(2, 3, 2, 4, 1, 1|\mathbf{A}, \boldsymbol{\pi}) \\ &= P(2)P(3|2)P(2|3)P(4|2)P(1|4)P(1|1) \\ &= \pi_2 \cdot a_{23} \cdot a_{32} \cdot a_{24} \cdot a_{41} \cdot a_{11} \end{aligned}$$

Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Hidden Markov Models

¿Qué pasa si los estados ahora no son deterministicos sino probabilisticos?

Caracterización de la distribución de probabilidad de las observaciones

$$b_j(\mathbf{o}_t) = P(\mathbf{o}_t | q_t = j)$$

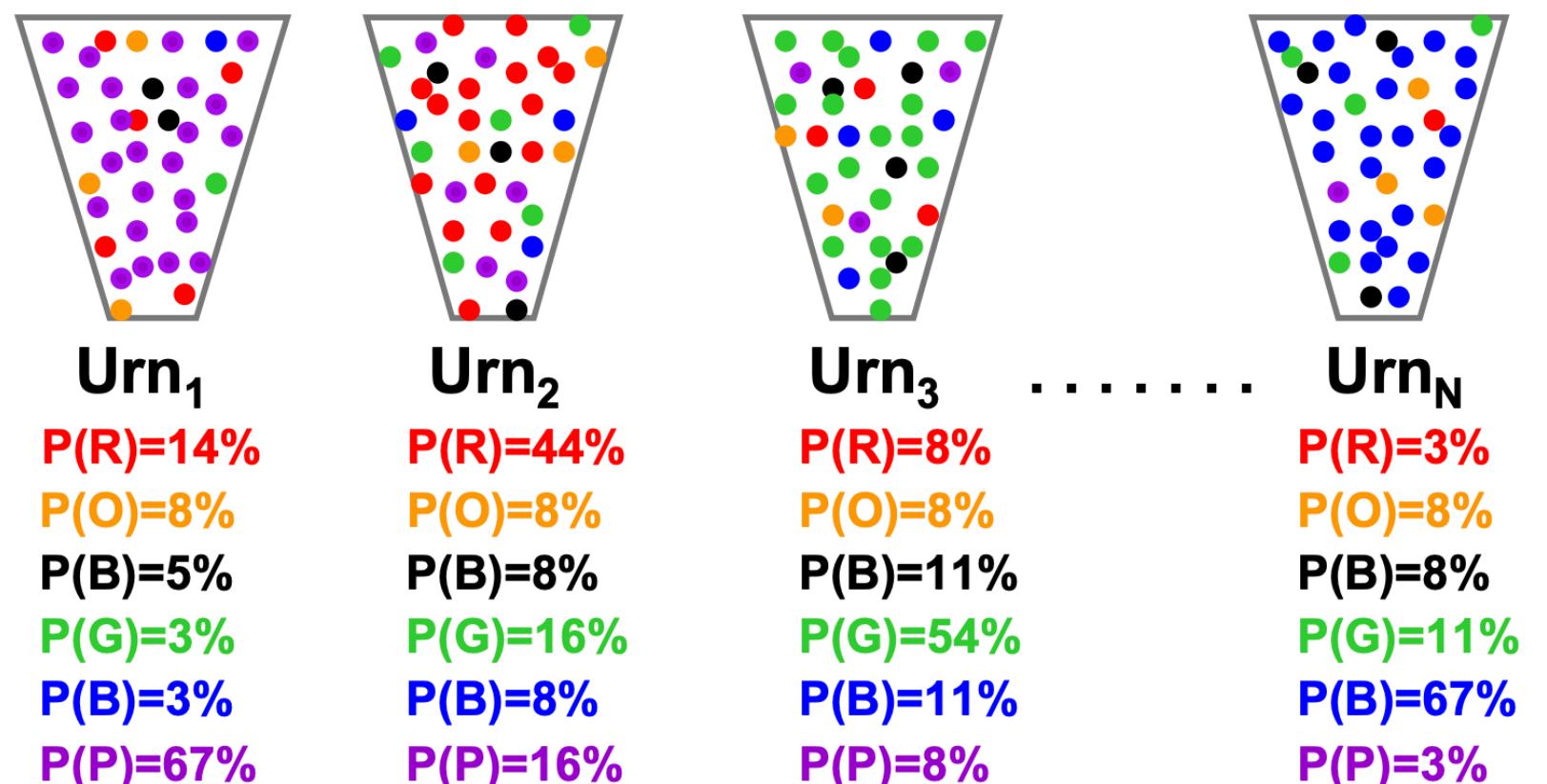


Figure 4.2: Urn and Ball example

¿Cómo el HMM genera observaciones?

1. Escojo j un estado de acuerdo a la distribución a priori.
2. Defino $t = 1; 1,2,3,\dots,T$
3. A partir de este estado escojo una observación a partir de su $b_j(o_t)$
4. Me desplazo a otro estado , según la probabilidad de transición.
5. Defino $t = t + 1$
6. Repito desde el paso 3, hasta terminar.

Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>

Hidden Markov Models

¿Pero cómo se aplica esto para Speech Recognition?

1. Suponga que N es el número de fonemas en el lenguaje.
2. Se modela las transiciones entre los fonemas (incluyendo el silencio) basado en datos escritos de entrenamiento.
3. Se tienen datos de pronunciación de los textos escritos, se realiza análisis mel-cepstral.
4. Se calcula la probabilidad que para un estado determinado (un fonema escrito) corresponda un vector mel-cepstral determinado.
5. Para un espectrograma determinado, se determina cual es la secuencia más probable que produzca este conjunto de observaciones. (Esto puede ser computacionalmente muy costoso).
6. Una vez se tienen las secuencias de fonemas, se busca en un diccionario las palabras correspondientes, y finalmente se pasa por un modelo de lenguaje.

Hidden Markov Models

Algoritmo Viterbi

1. Initialization

Set $t = 2$;
 $\delta_1(i) = \pi_i b_i(\mathbf{o}_1)$, $1 \leq i \leq N$
 $\psi_1(i) = 0$, $1 \leq i \leq N$

2. Induction

$$\begin{aligned}\delta_t(j) &= b_j(\mathbf{o}_t) \max_{1 \leq i \leq N} \delta_{t-1}(i) a_{ij}, \quad 1 \leq j \leq N \\ \psi_t(j) &= \arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}], \quad 1 \leq j \leq N\end{aligned}$$

3. Update time

Set $t = t + 1$;
Return to step 2 if $t \leq T$;
Otherwise, terminate the algorithm (goto step 4).

4. Termination

$$\begin{aligned}P^* &= \max_{1 \leq i \leq N} [\delta_T(i)] \\ q_T^* &= \arg \max_{1 \leq i \leq N} [\delta_T(i)]\end{aligned}$$

5. Path (state sequence) backtracking

(a) Initialization

Set $t = T - 1$

(b) Backtracking

$$q_t^* = \psi_{t+1}(q_{t+1}^*)$$

(c) Update time

Set $t = t - 1$;
Return to step (b) if $t \geq 1$;
Otherwise, terminate the algorithm.

Un problema es que se trabaja con producto de probabilidades, para evitar errores numéricos se puede trabajar con el logaritmo de estas probabilidades.

Source: <https://www.diva-portal.org/smash/get/diva2:831263/FULLTEXT01.pdf>



Gracias
Preguntas?