Coeficientes Cepstrais de Frequência Mel (MFCCs)

Universidade Federal do Ceará
Campus Sobral
Engenharia Elétrica e Engenharia de Computação

Processamento Digital de Sinais (SBL0085)

Prof. C. Alexandre R. Fernandes

Conteúdo

0. O que são os MFCCs?

- Os coeficientes cepstrais de frequência mel (Mel-Frequency Cepstral Coefficients - MFCCs) são uma representação comum do timbre (ou qualidade tonal) do som

- MFCC são compostos de 10 a 20 coeficientes (em geral 16).

- MFCCs são a representação de áudio mais utilizada em reconhecimento de fala e processamento de música.

- O objetivo principal do MFCC é representar o áudio de uma forma que imite a percepção auditiva humana, focando nas características que realmente importam para distinguir fonemas.

0. O que são os MFCCs?

- Os MFCCs são uma representação compacta, discriminativa e perceptualmente relevante de sinais de áudio, especialmente de fala.

- Transformam um sinal acústico bruto em um conjunto reduzido de características que preserva a informação mais importante para tarefas de classificação, reconhecimento e modelagem estatística.

- Aproximação da percepção auditiva humana - modela a resolução em frequência do sistema auditivo humano:
  - mais sensível a variações em baixas frequências;
  - progressivamente menos sensível em altas frequências.

0. O que são os MFCCs?

- Os MFCCs enfatizam componentes espectrais que são perceptualmente mais relevantes, tornando a representação mais alinhada com a forma como humanos distinguem sons e fonemas.

- Computacionalmente eficiente → adequado para aplicações em tempo real e sistemas embarcados.

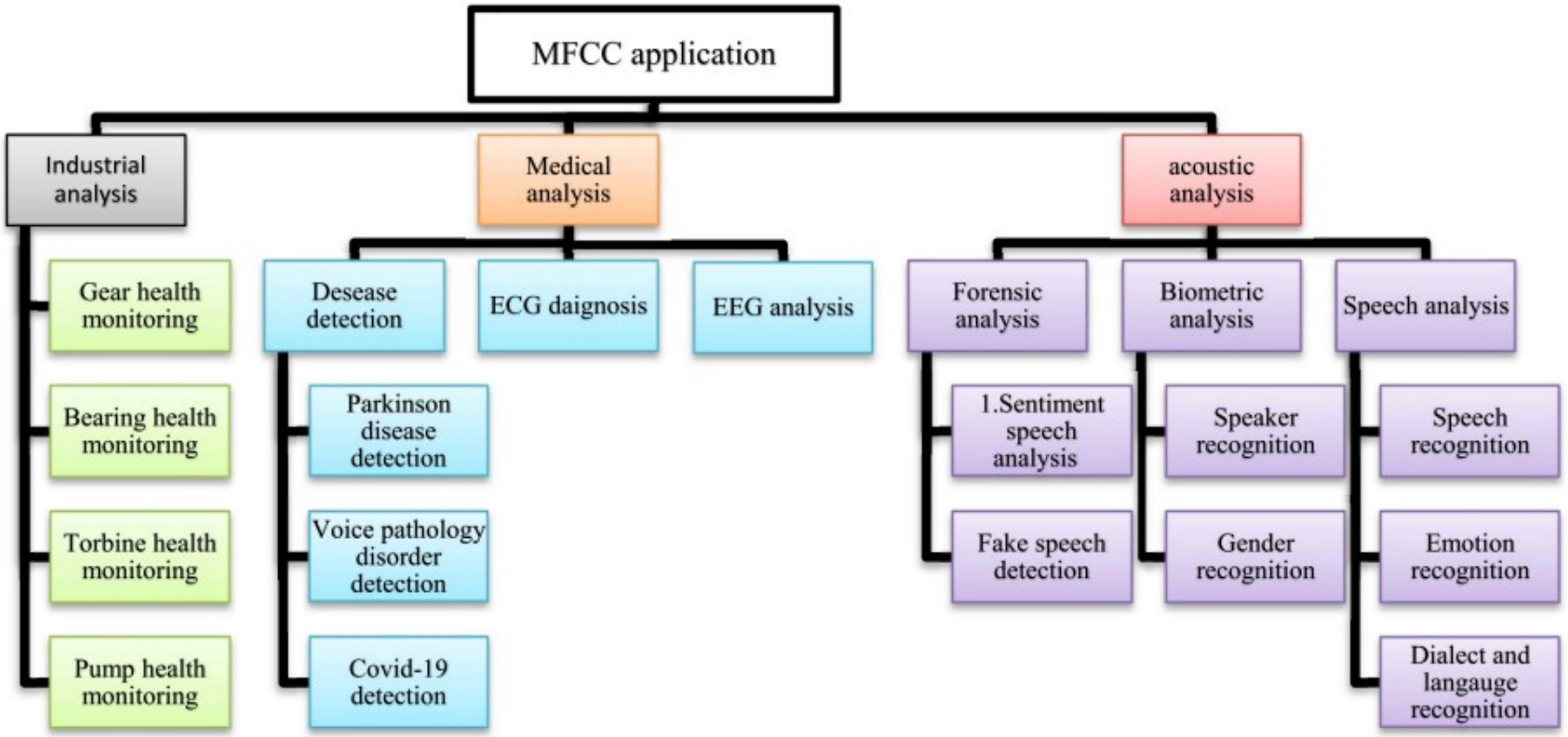
- Os MFCCs tornaram-se muito populares em reconhecimento automático de fala e áudio por décadas.

0. O que são os MFCCs?

- Os MFCCs demonstraram excelente desempenho em:
  - reconhecimento automático de fala;
  - identificação de locutor;
  - classificação de sons ambientais;
  - indexação e recuperação de áudio.

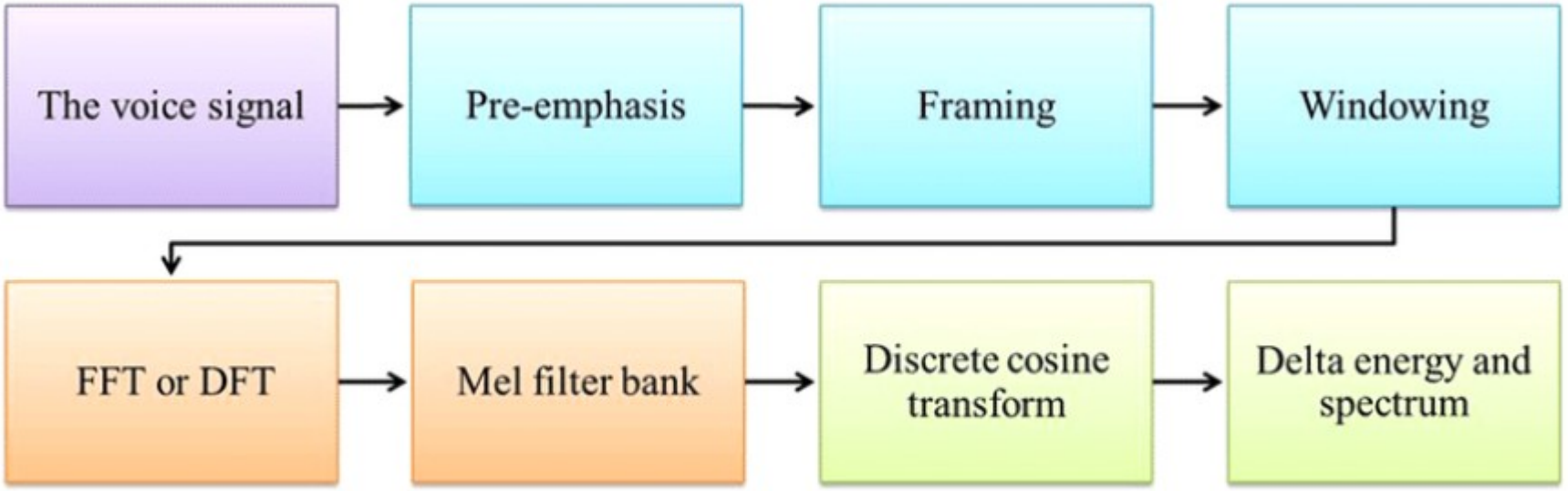
- Em resumo, os MFCCs:
  - incorporam conhecimento perceptual humano;
  - fornecem uma representação compacta e pouco correlacionada;
  - capturam características espectrais relevantes do sinal;
  - são robustos, eficientes e amplamente validados na prática.

0. O que são os MFCCs?



0. O que são os MFCCs?

- Principais etapas do cálculo dos MFCCs



I. Pré-ênfase (opcional)

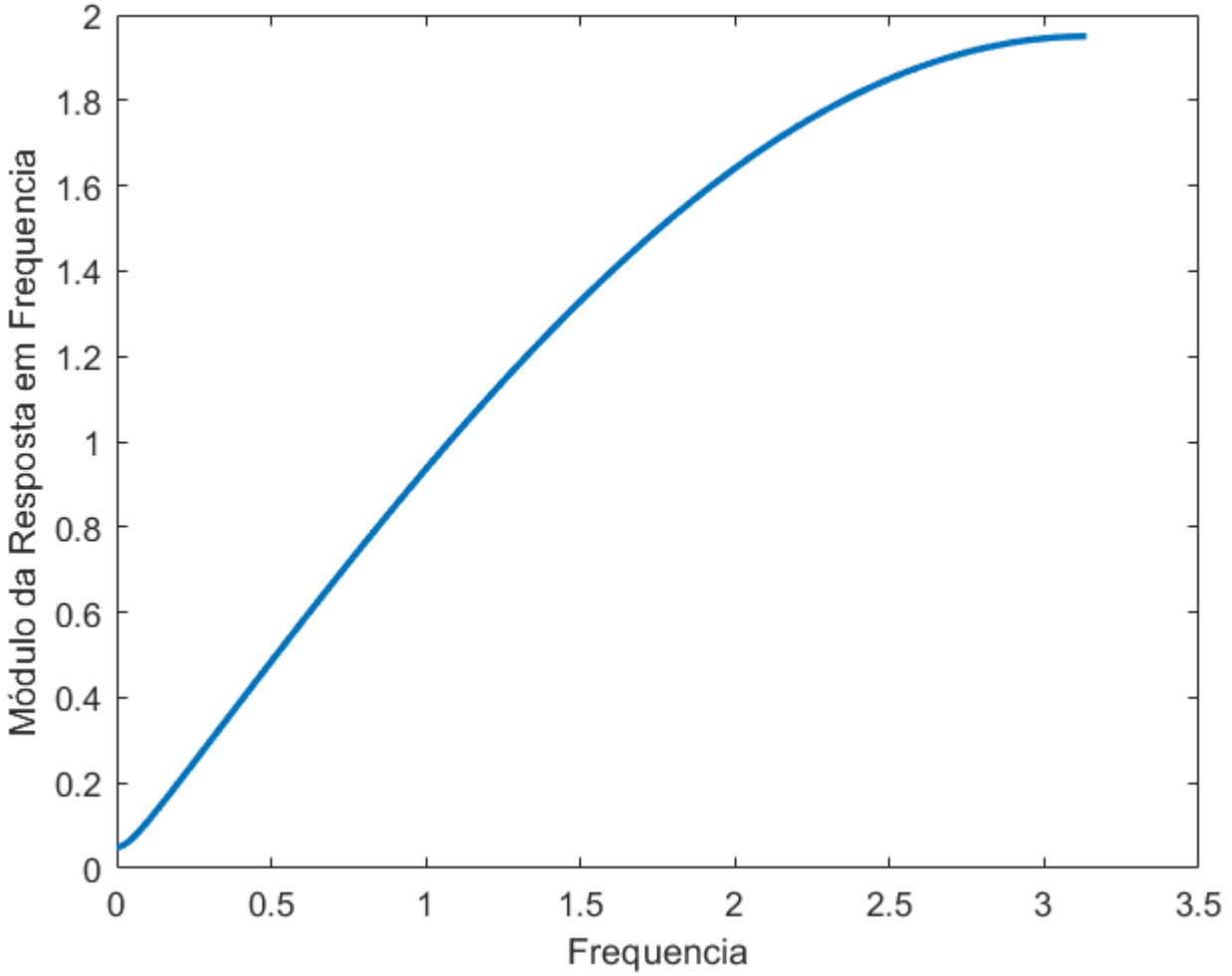
- Filtro para amplificar componentes de alta frequência, compensando a atenuação natural do trato vocal.
- O trato vocal atua naturalmente como um filtro passa-baixa. A pré-ênfase suaviza o espectro e aumenta a discriminação das formantes.
- A pré-ênfase equilibra o espectro, ajudando o modelo a capturar informações importantes contidas nas altas frequências.

$$y[n] = x[n] - \alpha x[n-1],$$

em geral, $0.9 \leq \alpha \leq 1$.

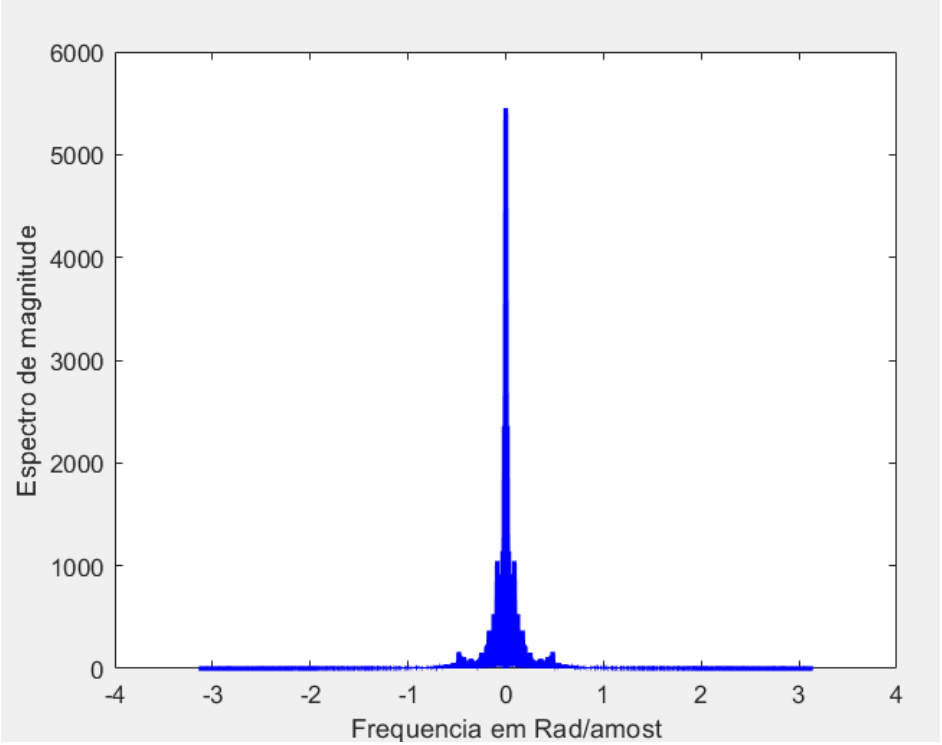
I. Pré-ênfase (opcional)

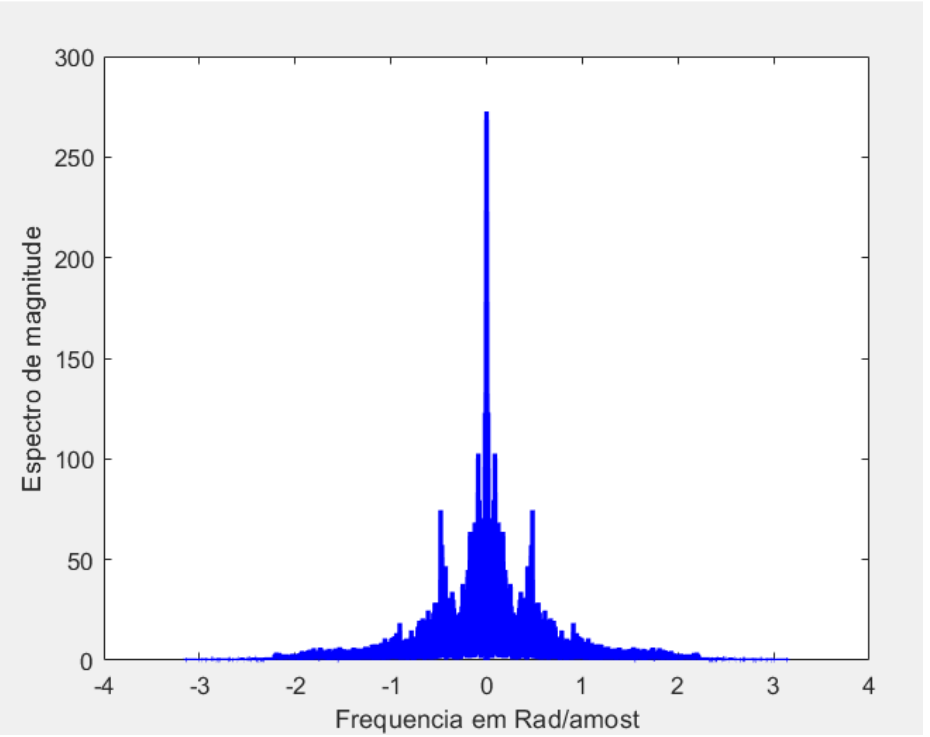
- Resposta em frequência do filtro pré-ênfase para α = 0,95



I. Pré-ênfase (opcional)

- Espectro do sinal original x sinal filtrado (áudio botão)





II. Segmentação (*Framing*)

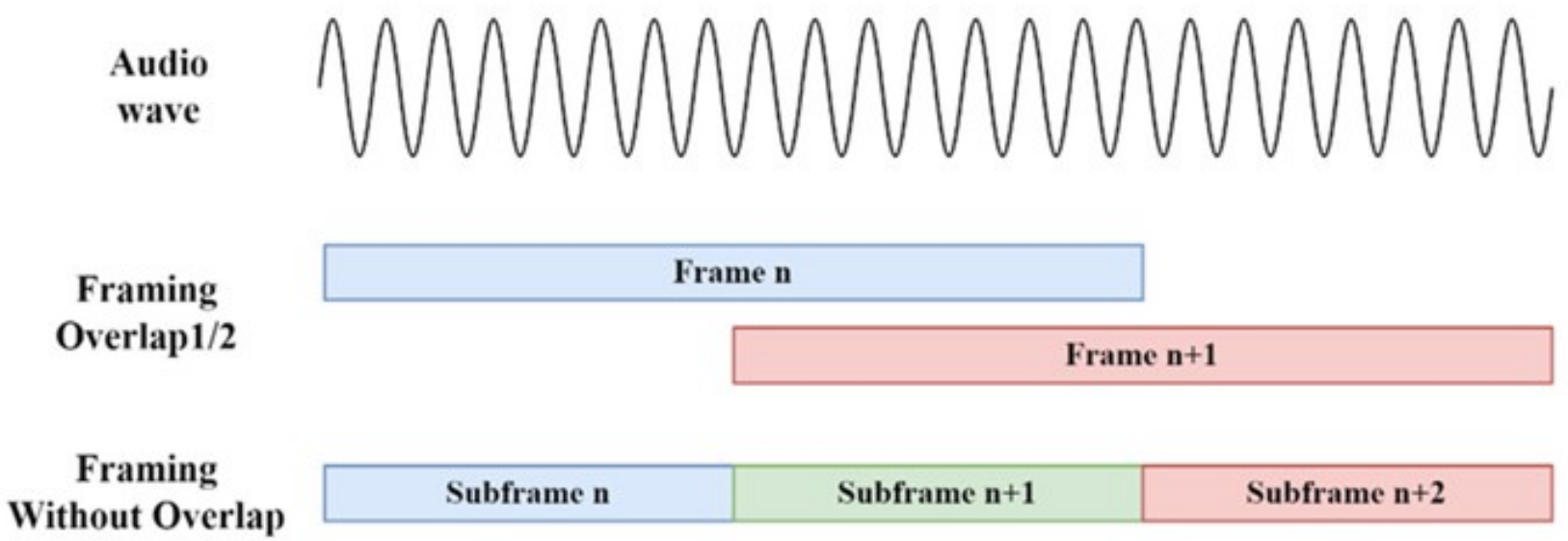
- Divide-se o sinal em janelas curtas de 10–40 ms.

- A fala humana é não estacionária, mas pode ser considerada estacionária em intervalos curtos.

- Estacionariedade: características estatísticas não variam com o tempo → Espectro de Fourier não varia com o tempo.

- O processamento é feito dentro de cada janela, onde o sinal pode ser considerado estacionário.

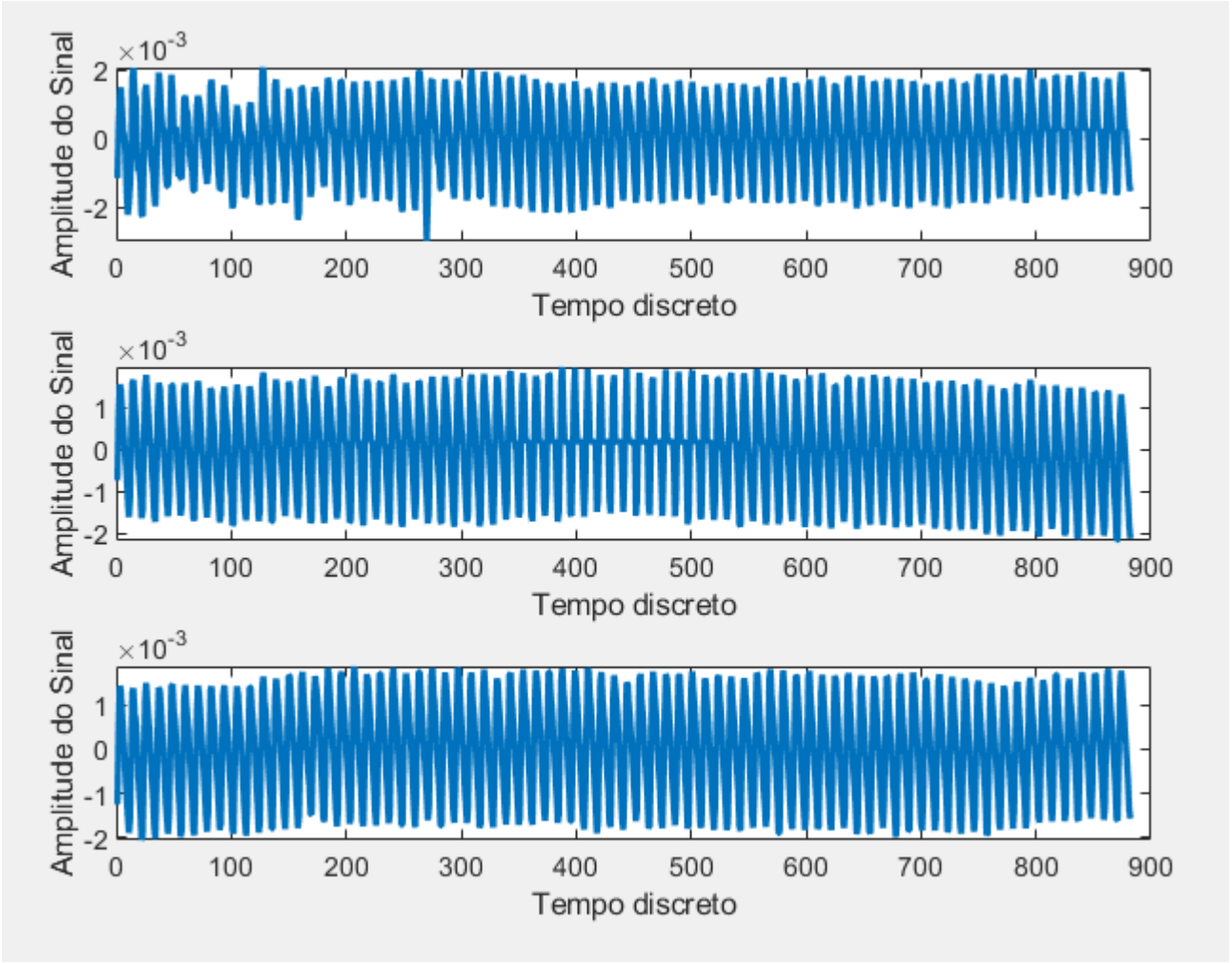
II. Segmentação (*Framing*)



- Nesta etapa, é comum remover a media dos sinais (opcional).

II. Segmentação (*Framing*)

- 3 exemplos de *frames* (áudio botão)



III. Aplicação de Janela

- Multiplicação do quadro (*frame*) por uma janela suavizante, geralmente Hamming, para reduzir descontinuidades nos limites.

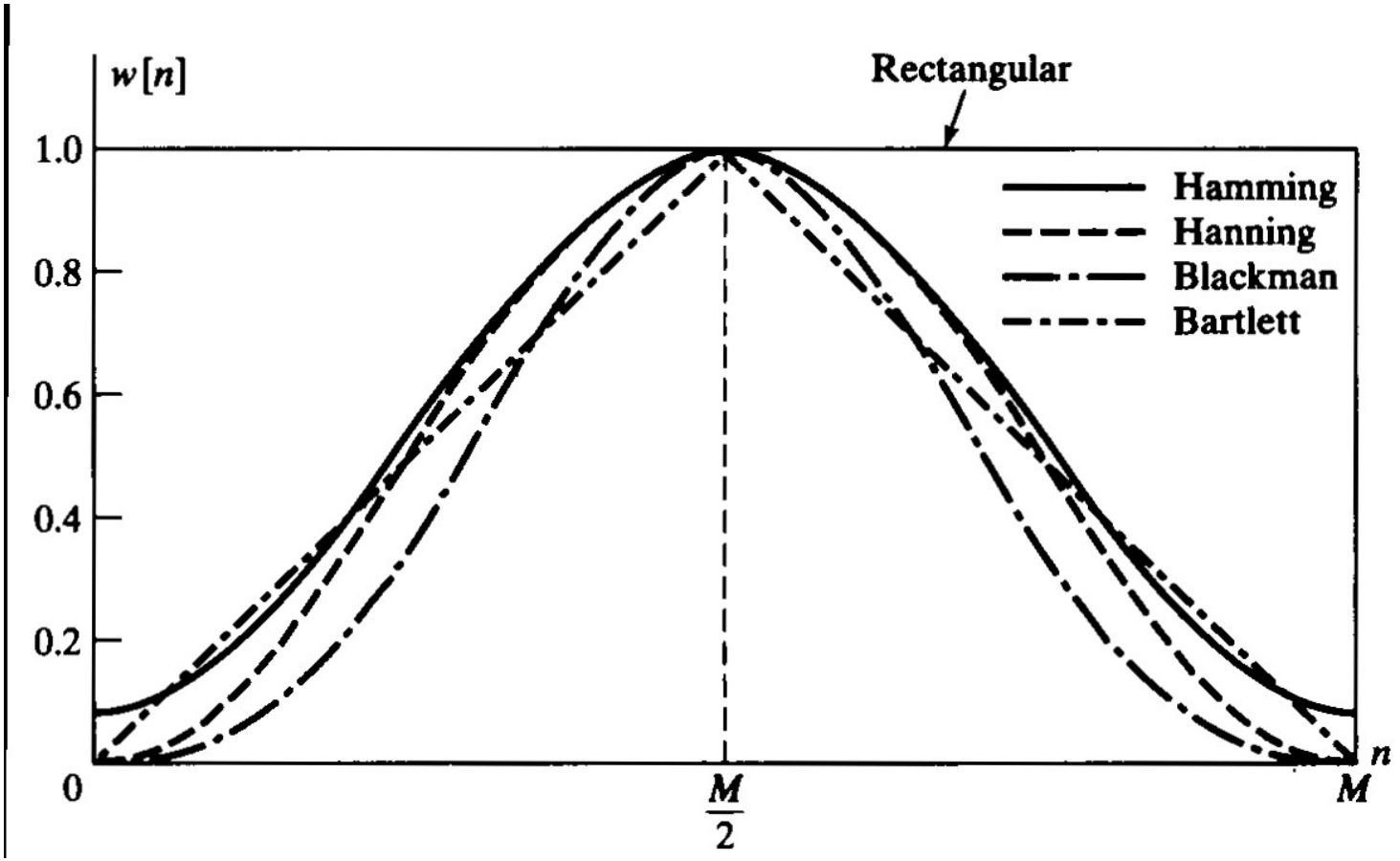
- A janela de Hamming evita descontinuidades nas bordas dos quadros, que causariam distorções no domínio da frequência.

- Janela de Hamming:

$$x_{w,k}[n] = x_k[n]\, w[n] \qquad w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right)$$

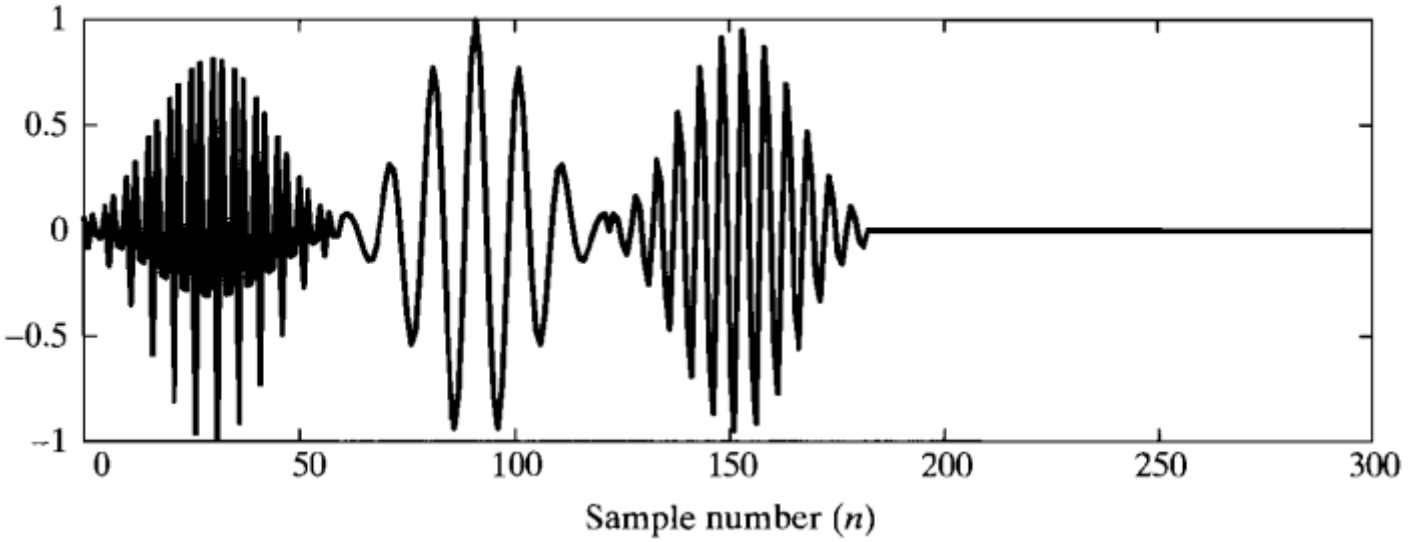
III. Aplicação de Janela

- Janelas no domínio do tempo → truncamento não abrupto

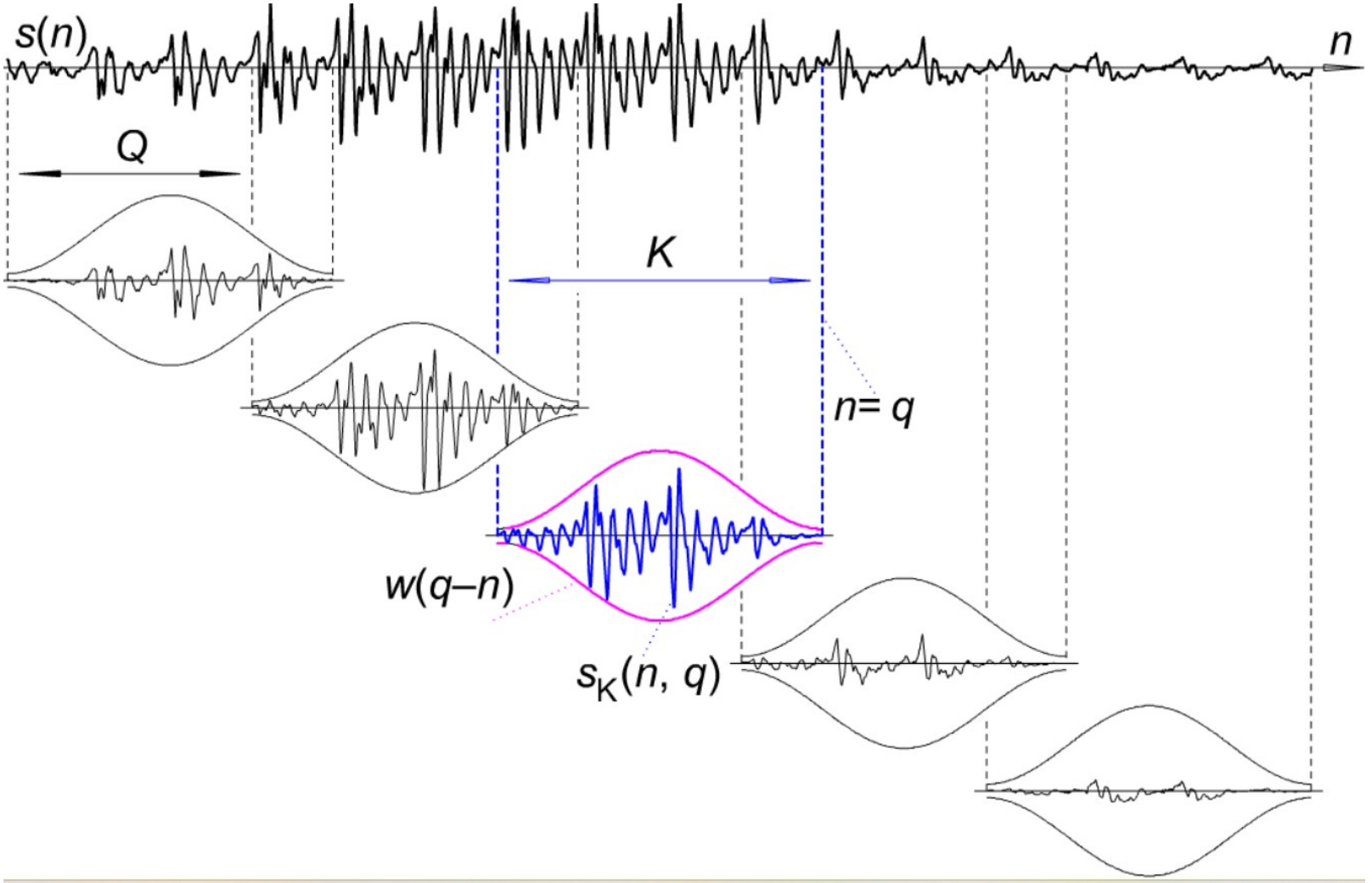


III. Aplicação de Janela

- As janelas fazem um truncamento mais suave do sinal → sinal menos distorcido na frequência .
- Cossenos trucados por janelas:

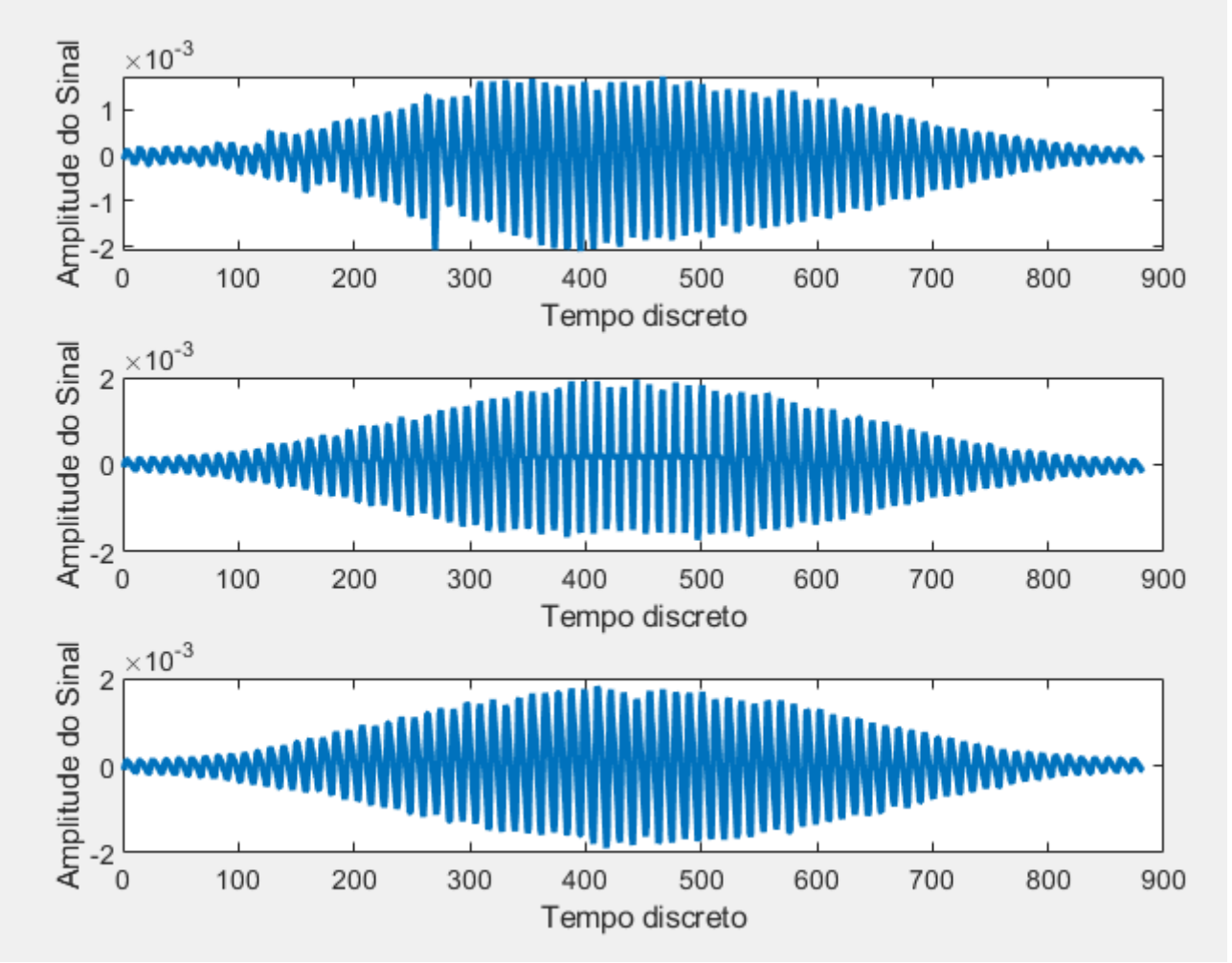


III. Aplicação de Janela



III. Aplicação de Janela

- 3 exemplos de *frames* com janelamento (áudio botão)



IV. Transformada Discreta de Fourier (DFT)

- Computa-se o espectro de magnitude para obter a distribuição de energia por frequência.

- Geralmente utiliza-se apenas o espectro positivo (devido à simetria).

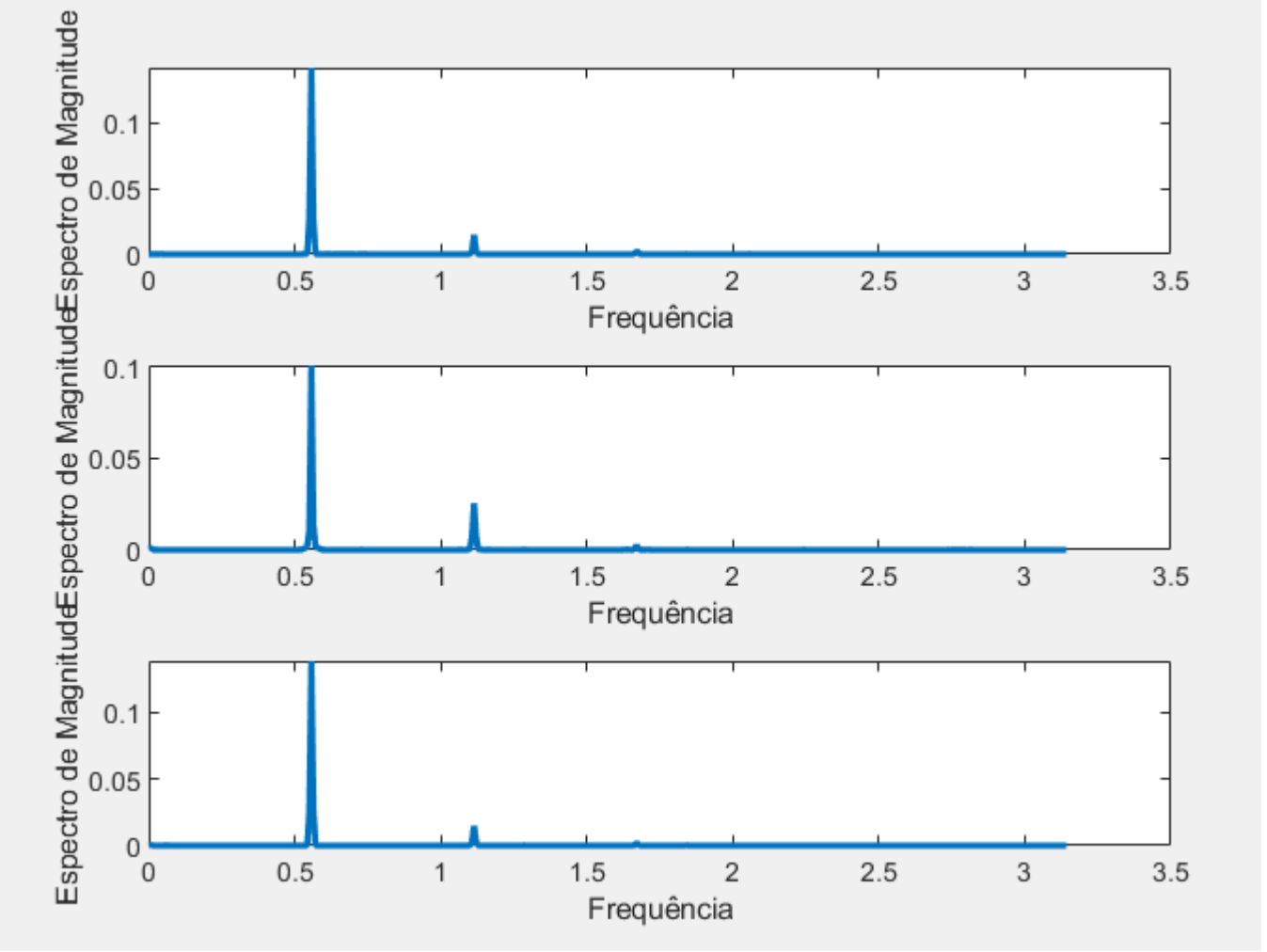
- Permite analisar o conteúdo espectral que compõe o som.

- A percepção auditiva é altamente dependente da distribuição espectral, não do sinal no tempo.

$$X_k[m] = \sum_{n=0}^{N-1} x_{w,k}[n]\, e^{-j2\pi nm/N}, \qquad m = 0, \ldots, N-1$$

IV. Transformada Discreta de Fourier (DFT)

- 3 exemplos de DFTs (áudio botão)



V. Energia Espectral

- Converte-se a DFT complexa em energia (ou potência).
- A audição humana responde mais diretamente à magnitude (ou energia) do que à fase do sinal.
- A fase é descartada.

$$|X_k[m]|^2 = X_k[m]\, X_k^*[m]$$

VI. Banco de Filtros Mel

6.1 Escala Mel → alteração no eixo das frequências

- Mapeamento da energia da DFT para o eixo Mel (escala perceptual).
- A percepção humana de frequência é aproximadamente linear abaixo de 1 kHz e logarítmica acima disso.
- A escala Mel lineariza essa percepção.

$$\text{mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad f(\text{mel}) = 700\left(10^{\text{mel}/2595} - 1\right)$$

VI. Banco de Filtros Mel

6.1 Escala Mel → alteração no eixo das frequências

- O ouvido humano é muito sensível a pequenas mudanças em baixas frequências, mas tem dificuldade em distinguir mudanças em frequências altas.

- A Escala Mel "achata" o espectro para que ele corresponda à percepção não linear humana.

VI. Banco de Filtros Mel

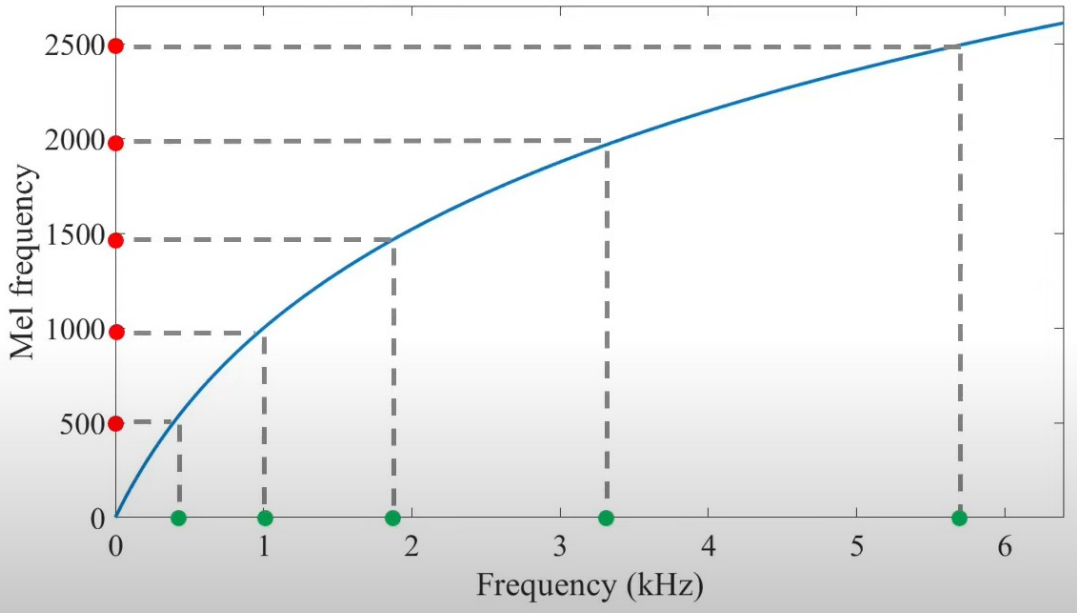
6.1 Escala Mel → alteração no eixo das frequências

- A audição humana processa as frequências em uma escala logarítmica

Nota musical	*Intervalo com a nota fundamental*	*Afinação natural*	*Frequência (Hz)* [Ocultar]
Dó	Dó uníssono	1/1=1,000	132,000
Dó #	Semitom	25/24=1,042	137,544
Ré b	Segunda diminuta	27/25=1,080	142,560
Ré	Segunda maior	9/8=1,125	148,500
Ré #	Segunda aumentada	76/74=1,172	154,704
Mi b	Terça menor	6/5=1,200	158,400
Mi	Terça maior	5/4=1,250	165,000
Mi #	Terça aumentada	125/96=1,302	171,864
Fá b	Quarta diminuta	32/25=1,280	168,960
Fá	Quarta perfeita	4/3=1,333	175,956
Fá #	Quarta aumentada	25/18=1,389	183,348
Sol b	Quinta diminuta	36/25=1,440	190,080
Sol	Quinta perfeita	3/2=1,500	198,000
Sol #	Quinta aumentada	25/16=1,563	206,316
La b	Sexta menor	8/5=1,6	211,200
Lá	Sexta maior	5/3=1,667	220,044
Lá #	Sexta aumentada	152/72=1,737	229,284
Si b	Sétima menor	9/5=1,800	237,600
Si	Sétima maior	15/8=1,875	247,500
Si #	Sétima aumentada	125/64=1,953	257,796
Dó b	Oitava diminuta	48/25=1,920	253,440
Dó	Oitava perfeita	2/1=2,000	264,000

VI. Banco de Filtros Mel

6.1 Escala Mel → alteração no eixo das frequências



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

Take m_k equidistant points on the mel frequency axis

Compute the f_k corresponding frequencies

VI. Banco de Filtros Mel

6.2 Aplicação dos filtros triangulares Mel

- O espectro de potência é passado por um conjunto de filtros triangulares na frequência (geralmente entre 26 e 40 filtros) espaçados de acordo com a Escala Mel.
- O espectro DFT possui centenas ou milhares de amostras altamente correlacionadas.
- Os filtros triangulares são usados para dividir os espectro DFT em poucas dezenas de bandas.
- Tal como na Etapa 2 (aplicação de Janela), o filtro triangular evita um truncamento abrupto.

VI. Banco de Filtros Mel

6.2 Aplicação dos filtros triangulares Mel

- Por que filtros triangulares (e não retangulares)?
  - Filtros triangulares possuem transições suaves, evitando efeitos de Gibbs → Introduzem menos distorção espectrais;

- Os filtros triangulares → reduzir a dimensionalidade espectral com truncamento suave.

- Agrupam a energia espectral em algumas bandas.

VI. Banco de Filtros Mel

6.2 Aplicação dos filtros triangulares Mel

- Determinam-se M pontos de corte igualmente espaçados na escala Mel, convertendo-os de volta para Hz e depois para os índices da DFT (ver slide anterior).

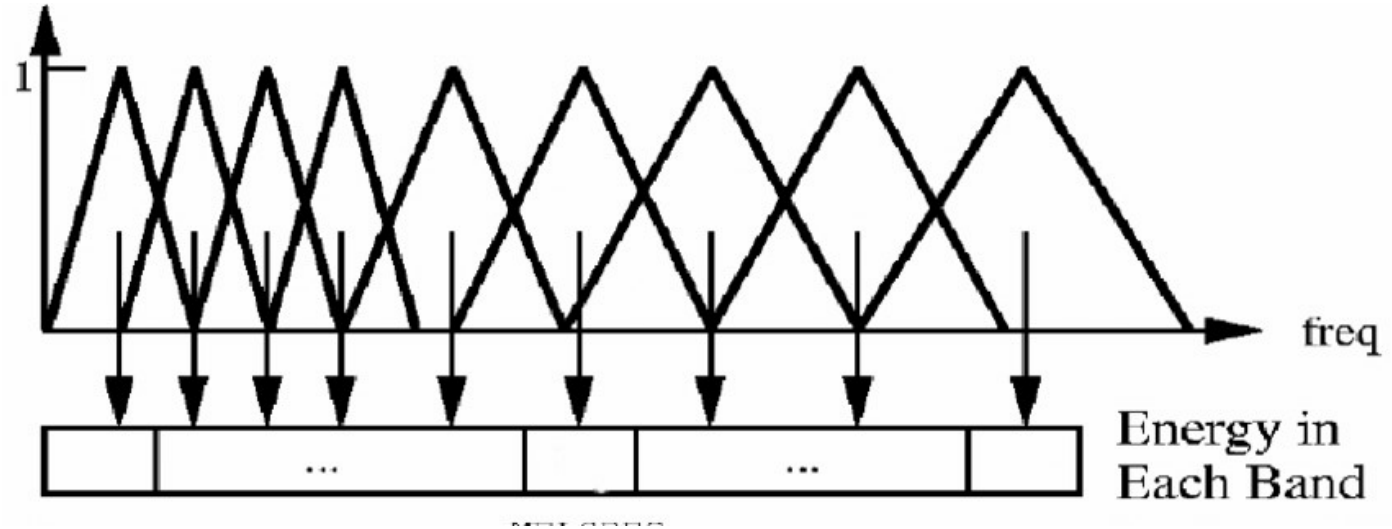
- Aplicação de filtro triangular:

$$H_m(f) = \begin{cases} 0, & f < f_{m-1} \\ \dfrac{f - f_{m-1}}{f_m - f_{m-1}}, & f_{m-1} \leq f \leq f_m \\ \dfrac{f_{m+1} - f}{f_{m+1} - f_m}, & f_m \leq f \leq f_{m+1} \\ 0, & f > f_{m+1} \end{cases}$$

VI. Banco de Filtros Mel

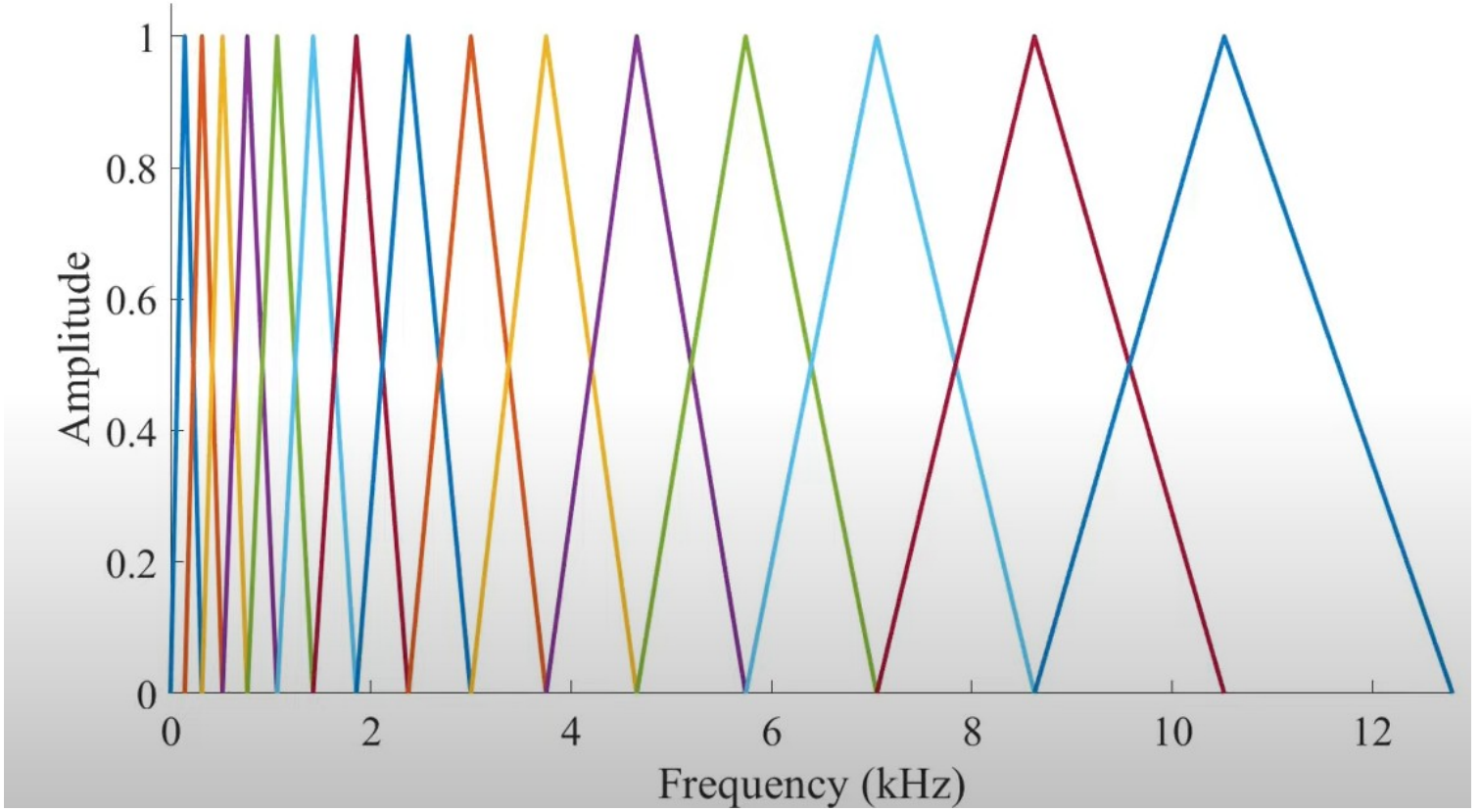
6.2 Aplicação dos filtros triangulares Mel

- Aplicação de filtro triangular na frequência → Quem estiver longe de certas frequências, terá peso menor
- Bandas de baixa frequência são estreitas;
- Bandas de alta frequência são largas.



VI. Banco de Filtros Mel

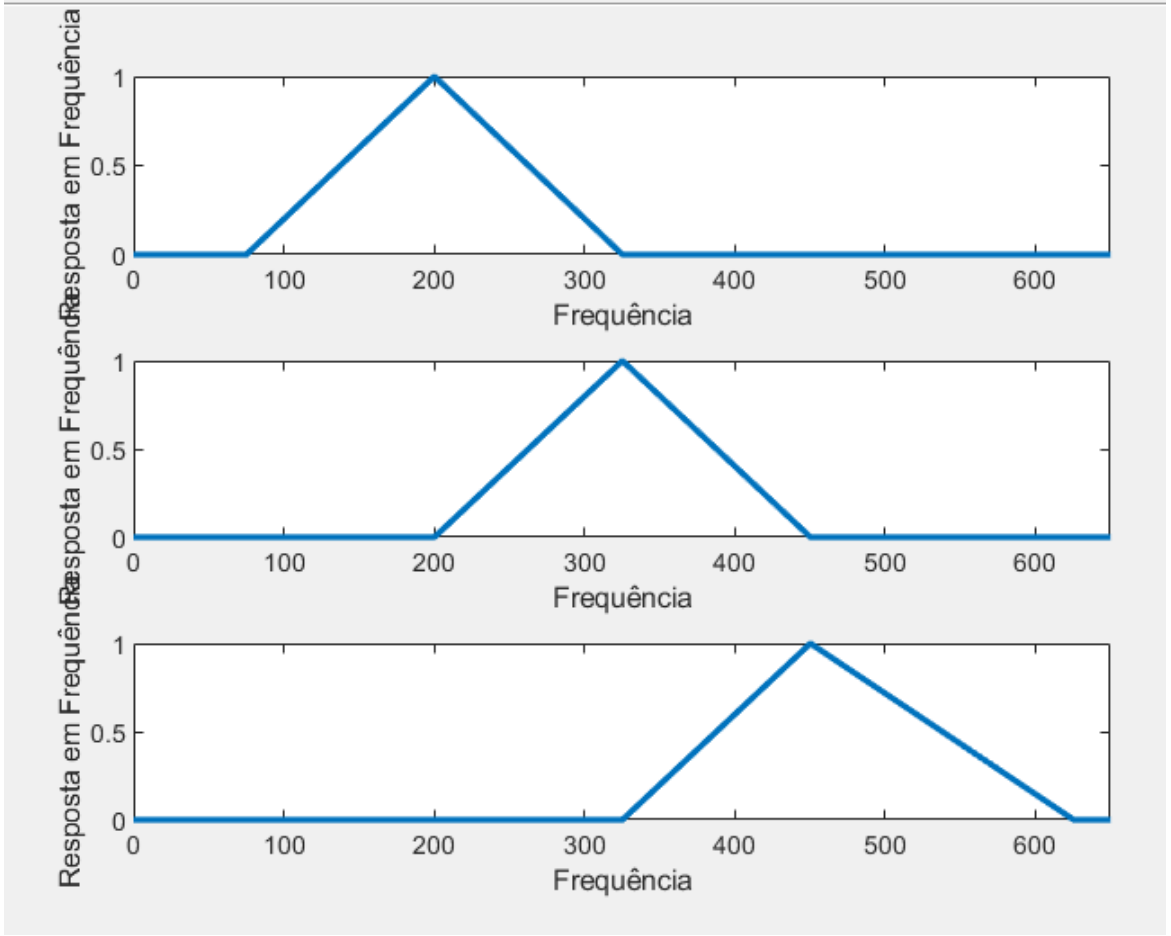
6.2 Aplicação dos filtros triangulares Mel



VI. Banco de Filtros Mel

6.2 Aplicação dos filtros triangulares Mel

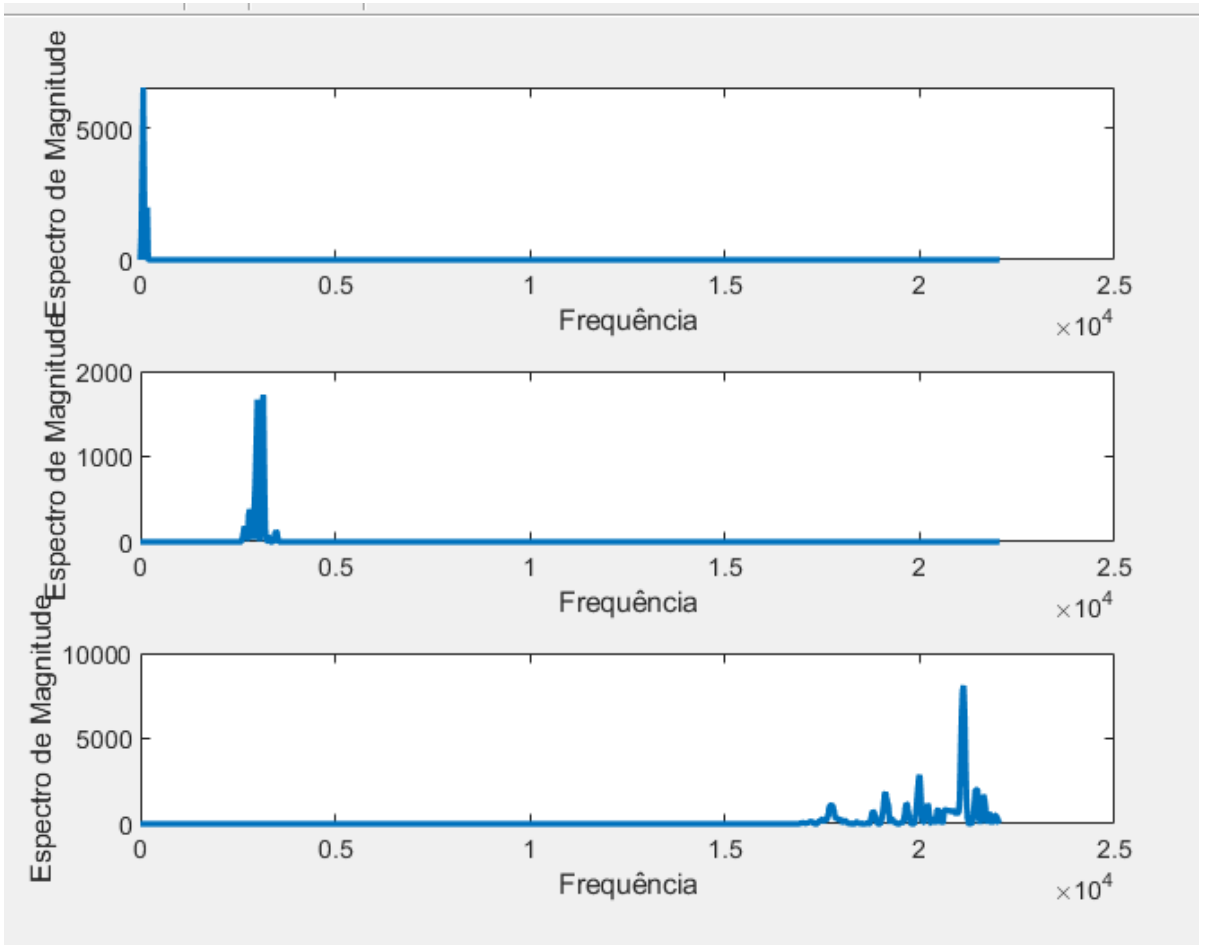
3 exemplos de filtros triangulares



VI. Banco de Filtros Mel

6.2 Aplicação dos filtros triangulares Mel

3 exemplos de sinais filtrados (áudio botão)



VI. Banco de Filtros Mel

6.3 Energia filtrada em cada banda Mel

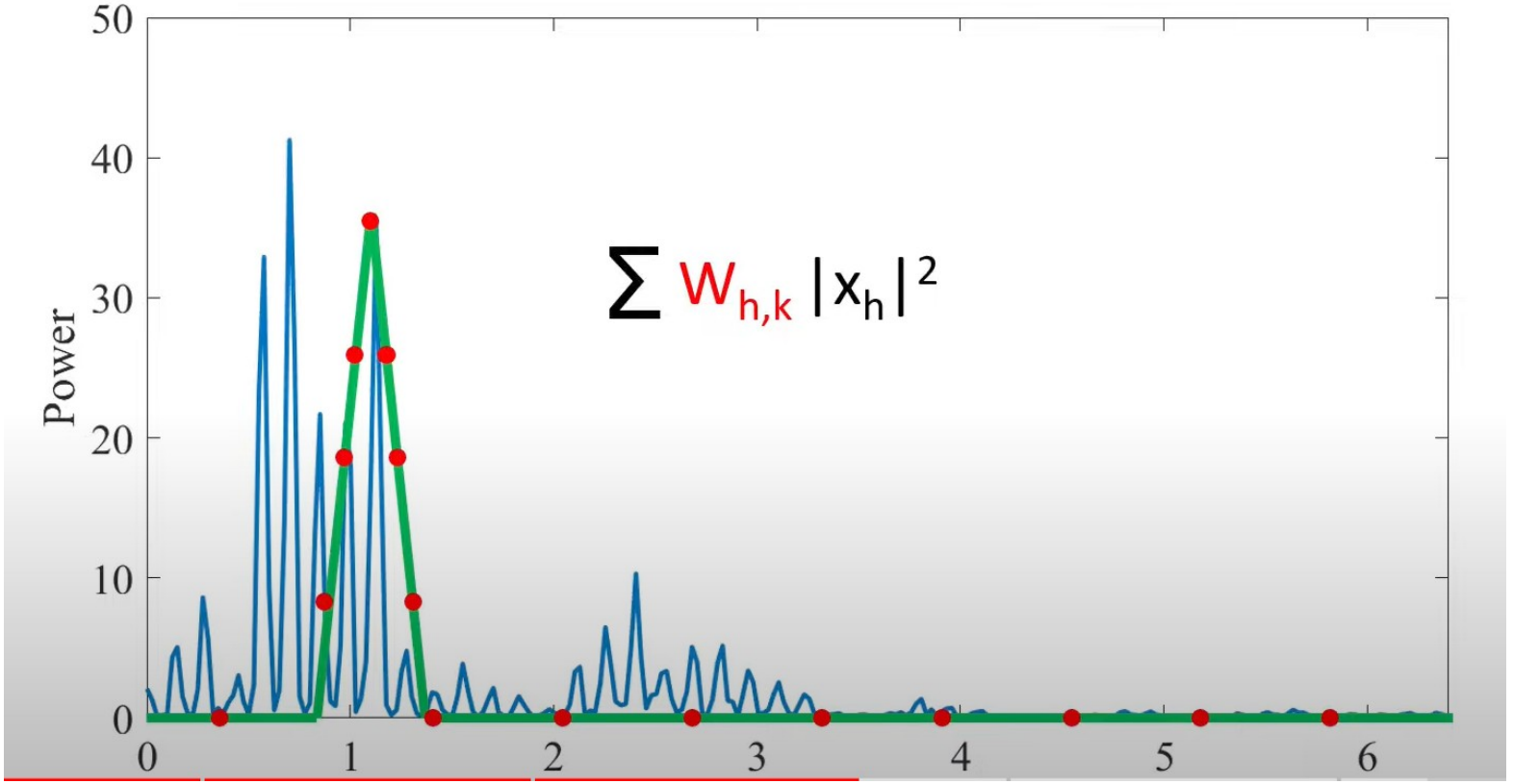
- Cálculo da energia de cada banda (bloco na frequência)

- Energia da banda m

$$E_m = \sum_{m_f=0}^{N/2} |X_k[m_f]|^2 \, H_m(m_f)$$

VI. Banco de Filtros Mel

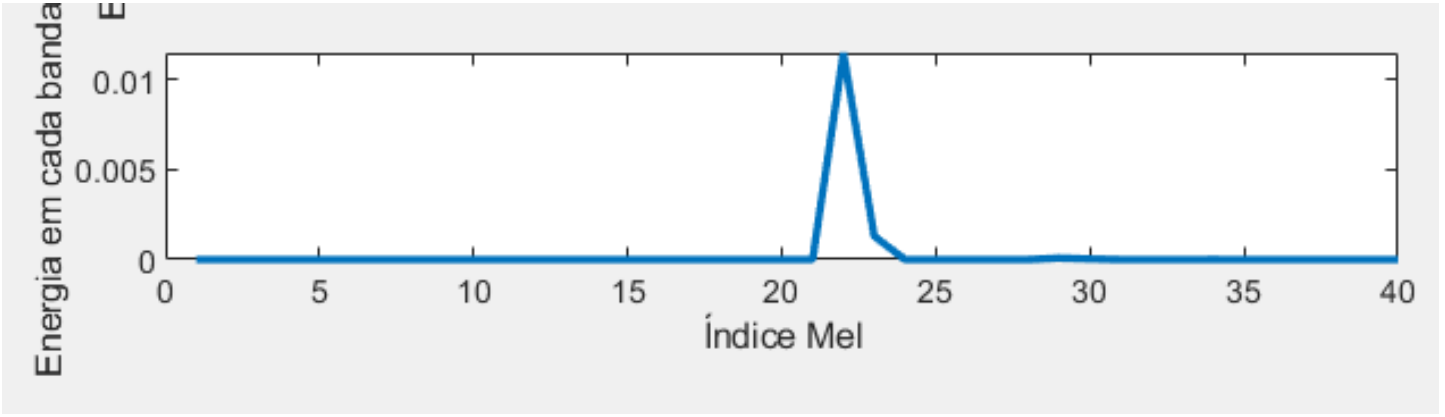
6.3 Energia filtrada em cada banda Mel



VI. Banco de Filtros Mel

6.3 Energia filtrada em cada banda Mel

Exemplo energia em cada banda Mel (áudio botão)



VI. Banco de Filtros Mel

- As etapas 6.1, 6.2 e 6.3 podem ser consideradas uma etapa só.

- Aqui nos slides, nós separamos a etapa 6 em 3 sub-etapas para efeitos didáticos:
    - Escala Mel
    - Banco de filtros triangulares
    - Cálculo da energia

VII. Logaritmo das energias Mel

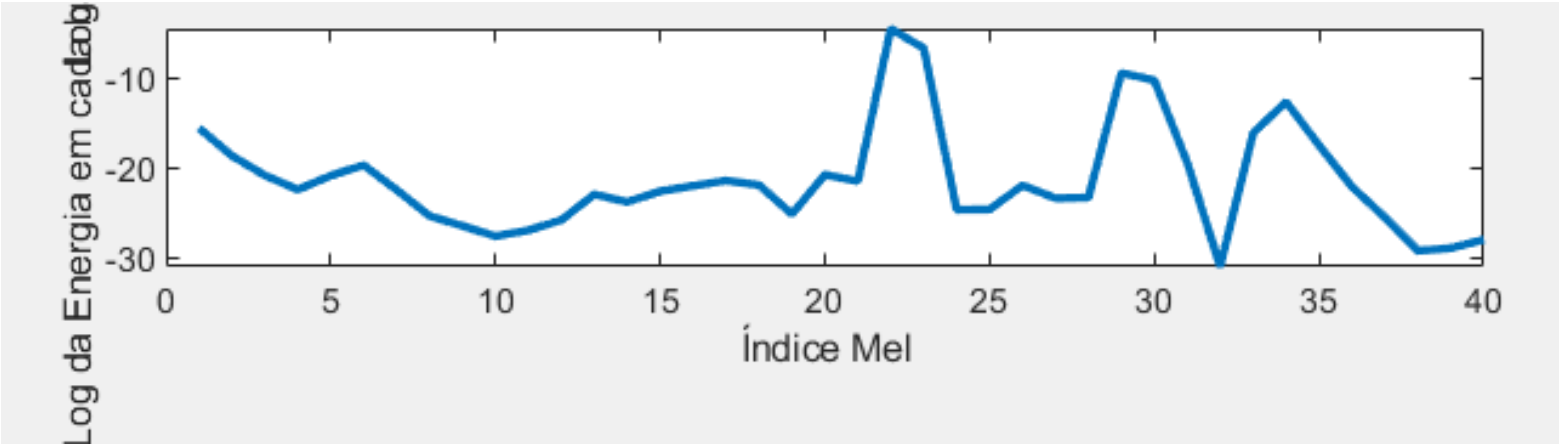
- O sistema auditivo tem resposta aproximadamente logarítmica à intensidade sonora (medida em Decibéis):

$$\tilde{E}_m = \log(E_m)$$

- Ou seja, o ouvi humana interpreta o som em escala logarítmica tanto na frequência quanto na intensidade.

VII. Logaritmo das energias Mel

- Exemplo log da energia em cada banda Mel (áudio botão)



VIII. Transformada Discreta do Cosseno (DCT)

- Aplicação da DCT no log das energias mel:

$$c_n = \sum_{m=1}^{M} \tilde{E}_m \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right], \qquad n = 0, \ldots, K-1,$$

- A DCT descorrelaciona as componentes e concentra a maior parte da energia informacional nos primeiros coeficientes.

- Isso permite reduzir dimensionalidade sem perda significativa de poder discriminativo;

VIII. Transformada Discreta do Cosseno (DCT)

- Uso de poucos parâmetros (tipicamente 12–13 coeficientes).
- Facilita a modelagem por classificadores estatísticos (GMMs, HMMs, SVMs, redes neurais).
- Isso separa grosseiramente as fontes espectrais (trato vocal) da excitação (pregas vocais).

VIII. Transformada Discreta do Cosseno (DCT)

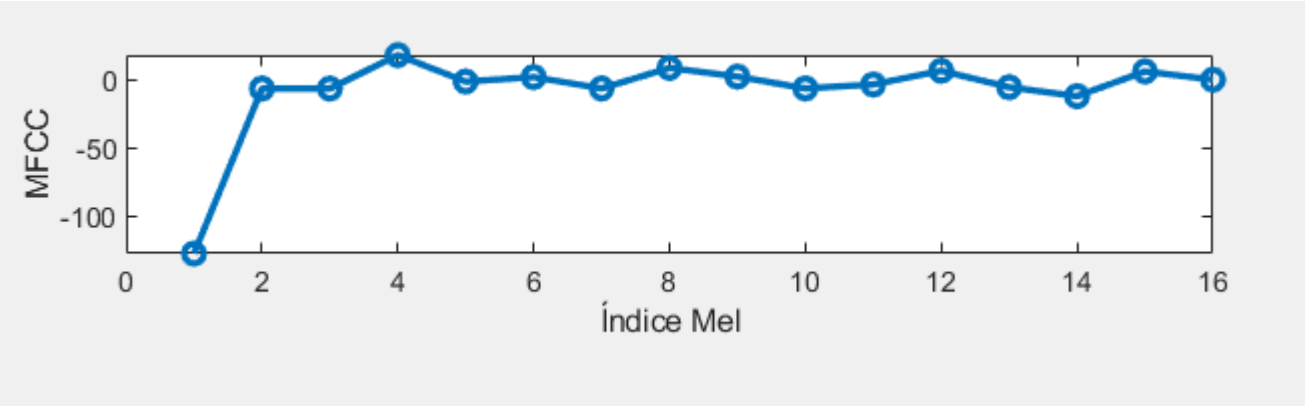
- Isso é conhecido como o Cepstro do Sinal.

- O nome "cepstro" (em inglês, cepstrum) é um anagrama da palavra "espectro" (spectrum), criado intencionalmente por seus inventores.

- Cepstro → espectro de um espectro logarítmico ou análise de frequência de uma análise de frequência.

VIII. Transformada Discreta do Cosseno (DCT)

- Exemplo de MFCC (áudio botão)



Resumo

Resumo da pipeline com equações principais

1. Pré-ênfase:

   $$y[n] = x[n] - \alpha x[n-1]$$

2. Framing e janela:

   $$x_{w,k}[n] = x[n + kH]\, w[n]$$

3. FFT:

   $$X_k[m] = \sum x_{w,k}[n] e^{-j2\pi nm/N}$$

4. Energia:

   $$|X_k[m]|^2$$

5. Filtros Mel:

   $$E_m = \sum |X_k[m_f]|^2 H_m(m_f)$$

6. Log:

   $$\tilde{E}_m = \log(E_m)$$

7. DCT:

   $$c_n = \sum \tilde{E}_m \cos\left(\frac{\pi n}{M}(m - \frac{1}{2})\right)$$