PPGEE
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA PARAÍBA

Processamento Digital de Sinais de Voz

Aula 05 – *Análise Espectral e Temporal de Sinais de Voz*

Profa. Silvana Luciene do N. Cunha Costa, D.Sc.

Transformada de Fourier a Curto Intervalo de Tempo (*Short Time Fourier Transform - STFT*)

- sons em estado estacionário, como vogais, são produzidos por excitação periódica de um sistema linear

- espectro do sinal de voz é o produto da excitação e a resposta em frequência do trato vocal

- Sinal de voz é variante no tempo

- A STFT pode capturar as mudanças na voz que ocorrem no tempo

DTFT e DFT

- DTFT de um sinal de voz $x(n)$

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} = DTFT\{x(n)\}$$

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n}d\omega = DTFT^{-1}\{X(e^{j\omega})\}$$

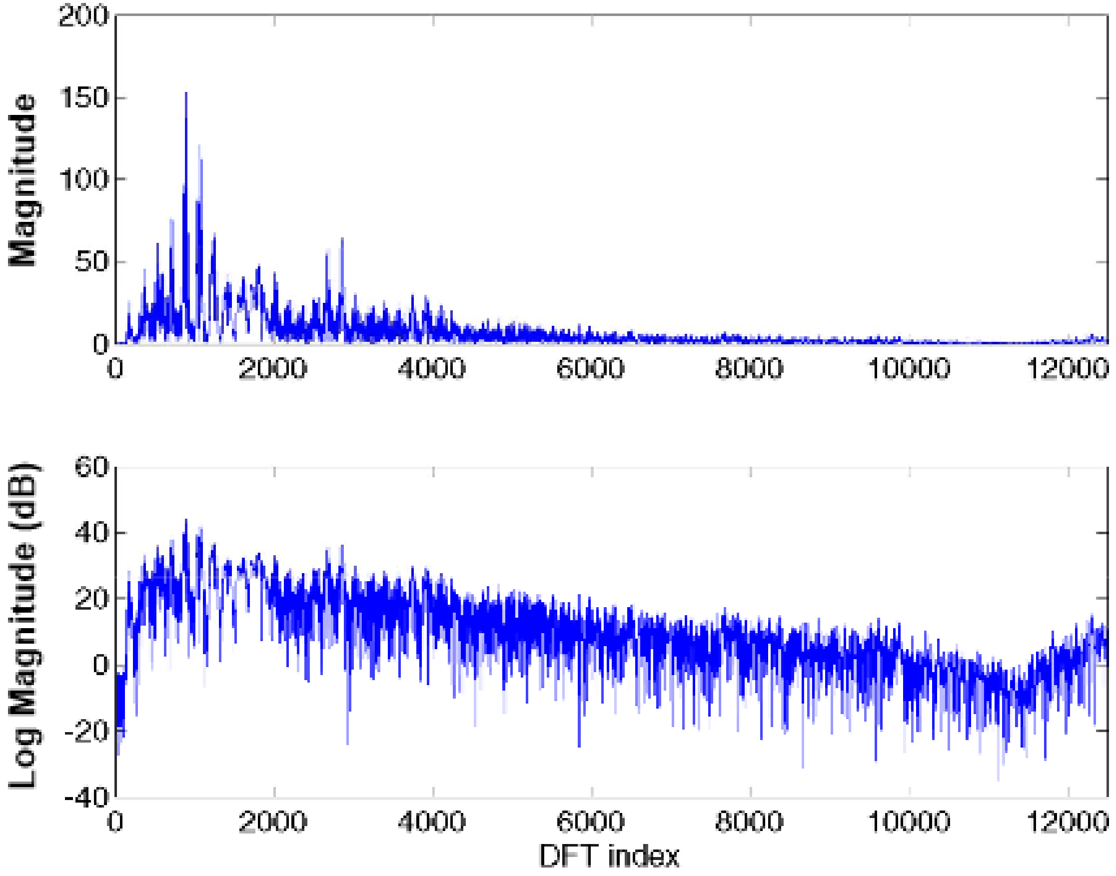
- A DTFT e DFT para a duração infinita do sinal pode ser calculada (a DTFT) e aproximadas (a DFT) pelo seguinte:

$$X(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\omega m} \quad (DTFT)$$

$$X(k) = \sum_{m=0}^{L-1} x(m)w(m)e^{-j(2\pi/L)km}, \quad k = 0,1,\ldots,L-1$$

$$= X(e^{j\omega})\Big|_{\omega=(2\pi k/L)} \qquad (DFT)$$

DTFT para 25000 amostras de voz



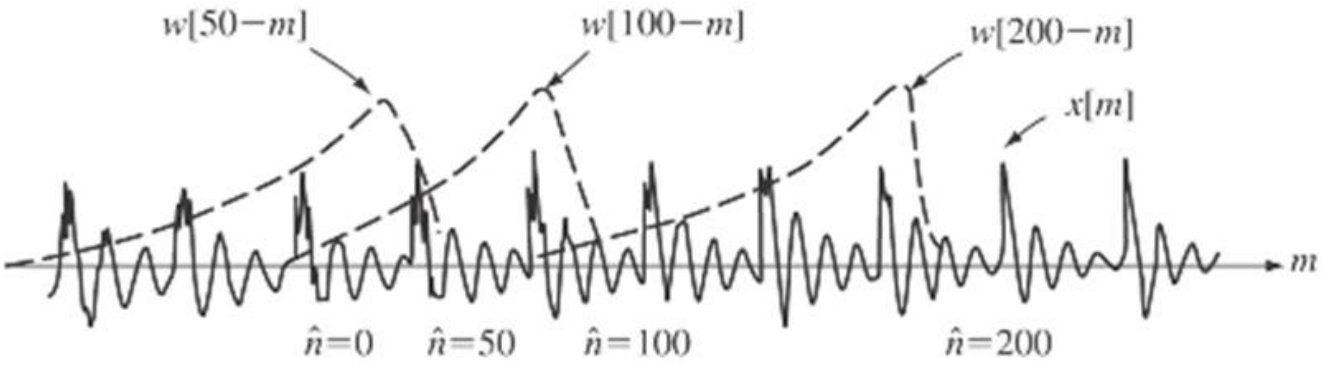
Análise de Fourier a curto intervalo de tempo (STFT)

- Definição de STFT

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)w(\hat{n}-m)e^{-j\hat{\omega}m}$$

$\hat{n}$ e $\hat{m}$ variáveis.

- $w(\hat{n}-m)$ é uma janela real que determina a porção de $x(n)$ que é usada no cálculo de $X_{\hat{n}} = (e^{j\hat{w}})$.

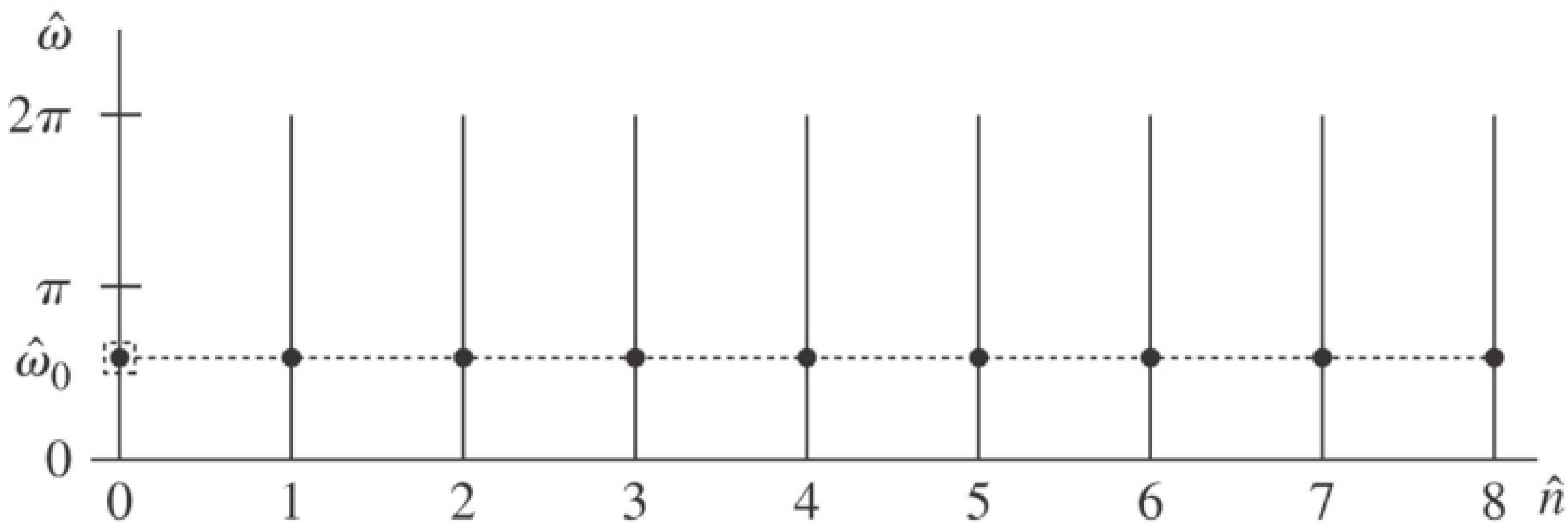


STFT

- STFT is a function of two variables, the time index, $\hat{n}$, which is discrete, and the frequency variable, $\hat{\omega}$, which is continuous

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)w(\hat{n}-m)e^{-j\hat{\omega}m}$$

$$= DTFT\ (x(m)w(\hat{n}-m)) \Rightarrow \hat{n} \text{ fixed, } \hat{\omega} \text{ variable}$$



Frequências da STFT

- the STFT is periodic in ω with period 2π, i.e.,

  $X_{\hat{n}}(e^{j\hat{\omega}}) = X_{\hat{n}}(e^{j(\hat{\omega}+2\pi k)}), \forall k$

- can use any of several frequency variables

  to express STFT, including

  --$\hat{\omega} = \hat{\Omega}T$ (where T is the sampling period for $x(m)$) to represent analog radian frequency, giving $X_{\hat{n}}(e^{j\hat{\Omega}T})$

  --$\hat{\omega} = 2\pi\hat{f}$ or $\hat{\omega} = 2\pi\hat{F}T$ to represent normalized frequency $(0 \le \hat{f} \le 1)$ or analog frequency $(0 \le \hat{F} \le F_s = 1/T)$, giving $X_{\hat{n}}(e^{j2\pi\hat{f}})$ or $X_{\hat{n}}(e^{j2\pi\hat{F}T})$

Sinal Recuperado da STFT

- since for a given value of $\hat{n}, X_{\hat{n}}(e^{j\hat{\omega}})$ has the same properties as a normal Fourier transform, we can recover the input sequence <u>exactly</u>
- since $X_{\hat{n}}(e^{j\hat{\omega}})$ is the normal Fourier transform of the windowed sequence $w(\hat{n}-m)x(m)$, then

$$w(\hat{n}-m)x(m)=\frac{1}{2\pi}\int_{-\pi}^{\pi}X_{\hat{n}}(e^{j\hat{\omega}})e^{j\hat{\omega}m}d\hat{\omega}$$

- assuming the window satisfies the property that $w(0)\neq 0$ (a trivial requirement), then by evaluating the inverse Fourier transform when $m=\hat{n}$, we obtain

$$x(\hat{n})=\frac{1}{2\pi w(0)}\int_{-\pi}^{\pi}X_{\hat{n}}(e^{j\hat{\omega}})e^{j\omega\hat{n}}d\hat{\omega}$$

Sinal Recuperado da STFT

$$x(\hat{n}) = \frac{1}{2\pi w(0)} \int_{-\pi}^{\pi} X_{\hat{n}}(e^{j\hat{\omega}}) e^{j\omega\hat{n}} d\hat{\omega}$$

- with the requirement that $w(0) \neq 0$, the sequence $x(\hat{n})$ can be recovered exactly from $X_{\hat{n}}(e^{j\hat{\omega}})$, if $X_{\hat{n}}(e^{j\hat{\omega}})$ is known for all values of $\hat{\omega}$ over one complete period
  - sample-by-sample recovery process
  - $X_{\hat{n}}(e^{j\hat{\omega}})$ must be known for every value of $\hat{n}$ and for all $\hat{\omega}$

can also recover sequence $w(\hat{n}-m)x(m)$ but can't guarantee that $x(m)$ can be recovered since $w(\hat{n}-m)$ can equal 0

Propriedades da STFT

$$X_{\hat{n}}(e^{j\hat{\omega}}) = DTFT[w(\hat{n}-m)x(m)] \quad \hat{n} \text{ fixed, } \hat{\omega} \text{ variable}$$

- relation to short-time power density function

$$S_{\hat{n}}(e^{j\hat{\omega}}) = |X_{\hat{n}}(e^{j\hat{\omega}})|^2 = X_{\hat{n}}(e^{j\hat{\omega}}) \cdot X_{\hat{n}}^*(e^{j\hat{\omega}}) = DTFT\left[R_{\hat{n}}(k)\right] \quad \hat{n} \text{ fixed}$$

$$R_{\hat{n}}(k) = \sum_{m=-\infty}^{\infty} w(\hat{n}-m)x(m)w(\hat{n}-m-k)x(m+k) \Leftrightarrow S_{\hat{n}}(e^{j\hat{\omega}})$$

- Relation to regular $X(e^{j\hat{\omega}})$ (assuming it exists)

$$X(e^{j\hat{\omega}}) = DTFT\left[x(m)\right] = \sum_{m=-\infty}^{\infty} x(m)e^{-j\hat{\omega}m}$$

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{-j\theta})X(e^{j(\hat{\omega}-\theta)})e^{-j\theta\hat{n}}\,d\theta$$

$$\left[w(\hat{n}-m)\square x(m) \leftrightarrow W(e^{-j\theta})e^{-j\theta\hat{n}} * X(e^{j\theta})\right]$$

Propriedades da STFT

- assume $X(e^{j\hat{\omega}})$ exists

$$X(e^{j\hat{\omega}}) = DTFT\left[x(m)\right] = \sum_{m=-\infty}^{\infty} x(m)e^{-j\hat{\omega}m}$$

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} W(e^{-j\theta})X(e^{j(\hat{\omega}-\theta)})e^{-j\theta\hat{n}}d\theta$$

- limiting case

$$w(\hat{n}) = 1 \quad -\infty < \hat{n} < \infty \Leftrightarrow W(e^{j\hat{\omega}}) = 2\pi\delta(\hat{\omega})$$

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} 2\pi\delta(-\theta)X(e^{j(\hat{\omega}-\theta)})e^{-j\theta\hat{n}}d\theta = X(e^{j\hat{\omega}})$$

i.e., we get the same thing no matter where the window is shifted

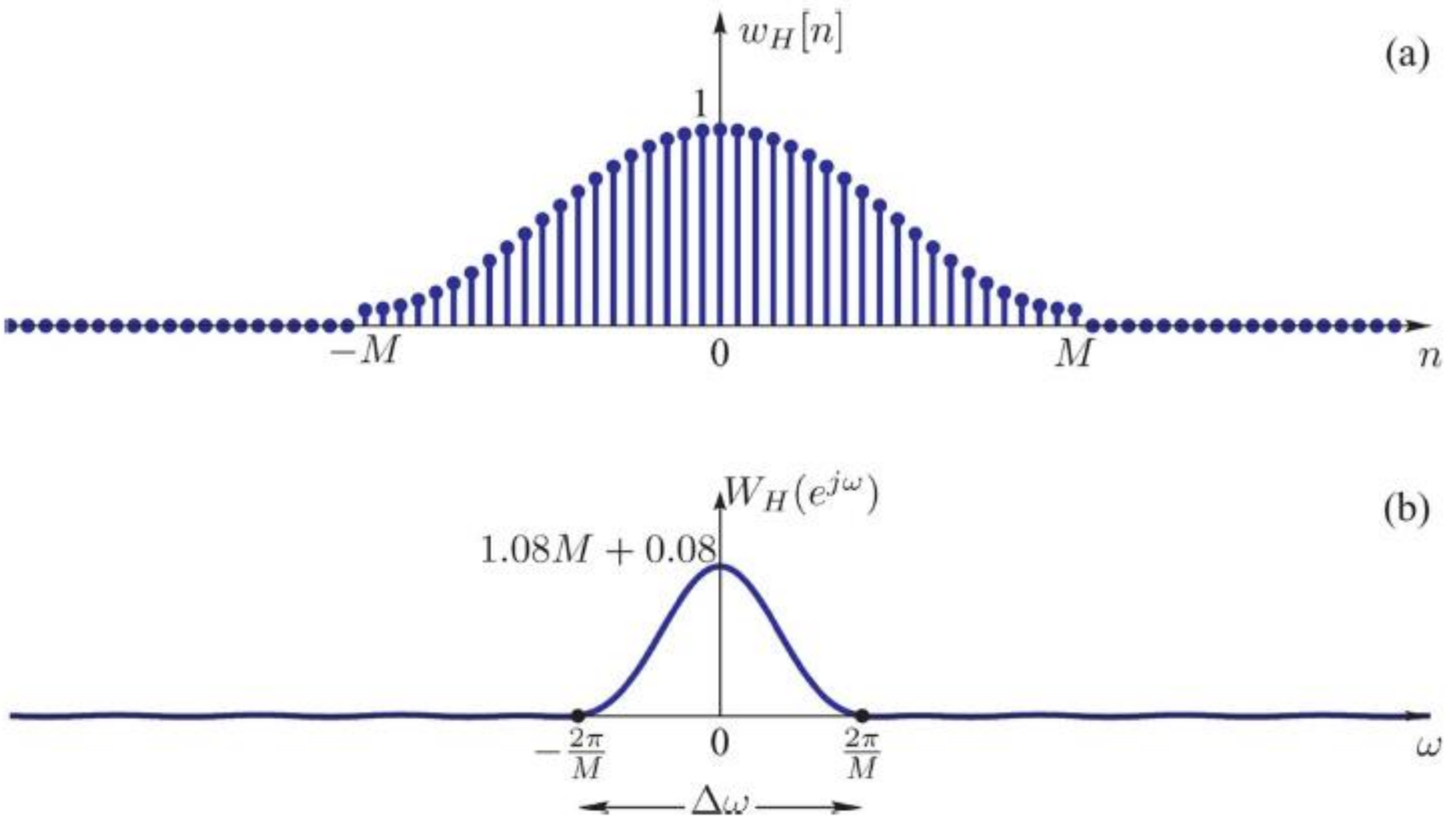
Efeitos do comprimento das janelas

- for $X_{\hat{n}}(e^{j\hat{\omega}})$ to represent the short-time spectral properties of $x(\hat{n})$ inside the window $\Rightarrow$ $W(e^{j\theta})$ should be much narrower in frequency than significant spectral regions of $X(e^{j\hat{\omega}})$--i.e., almost an impulse in frequency
- consider rectangular and Hamming windows, where width of the main spectral lobe is inversely proportional to window length, and side lobe levels are essentially independent of window length

Rectangular Window: flat window of length L samples; first zero in frequency response occurs at F_S/L, with sidelobe levels of -14 dB or lower

Hamming Window: raised cosine window of length L samples; first zero in frequency response occurs at $2F_S/L$, with sidelobe levels of -40 dB or lower

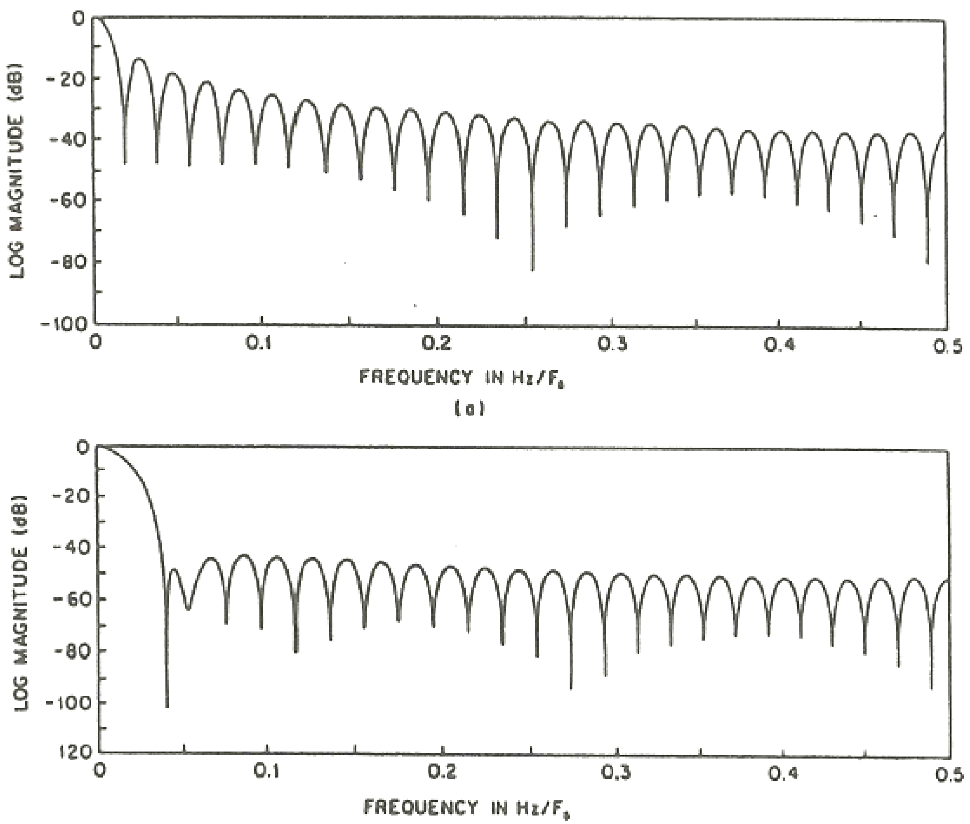
Janelas



L=2M+1-point Hamming window and its corresponding DTFT

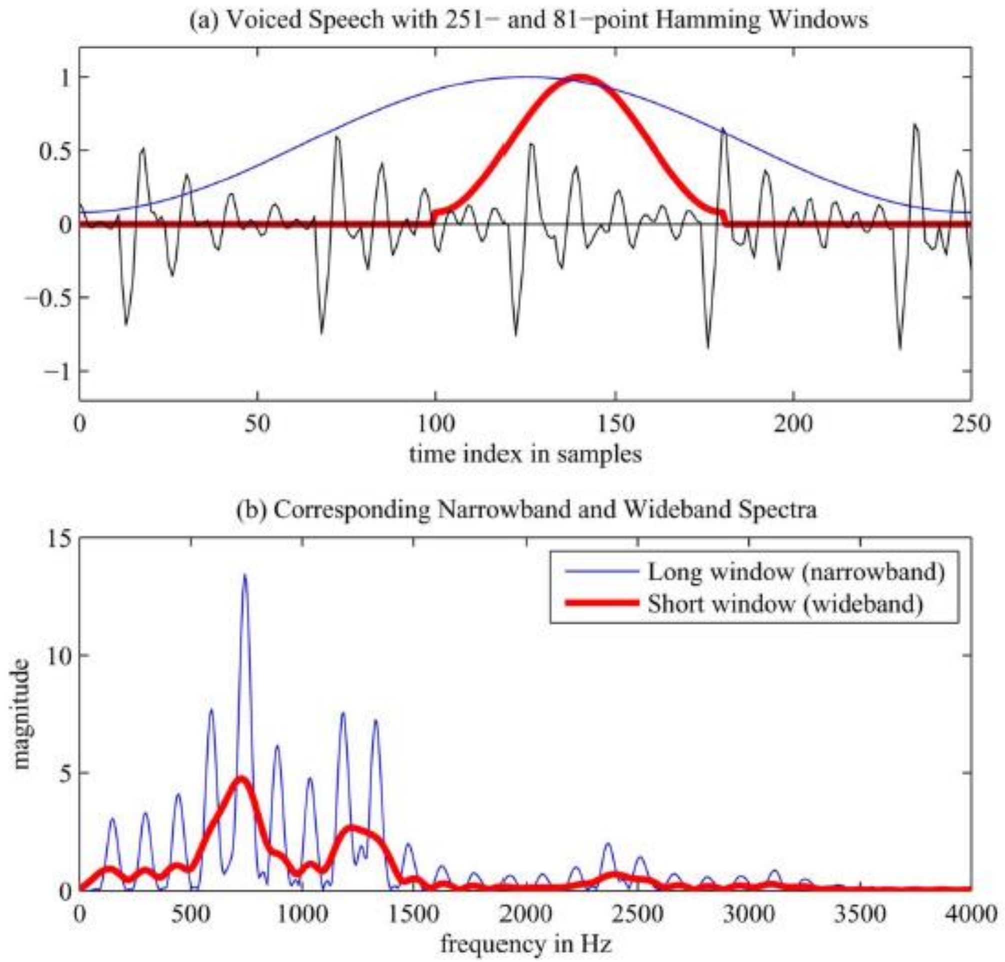
Efeitos do comprimento das janelas

Resposta em frequência das janelas



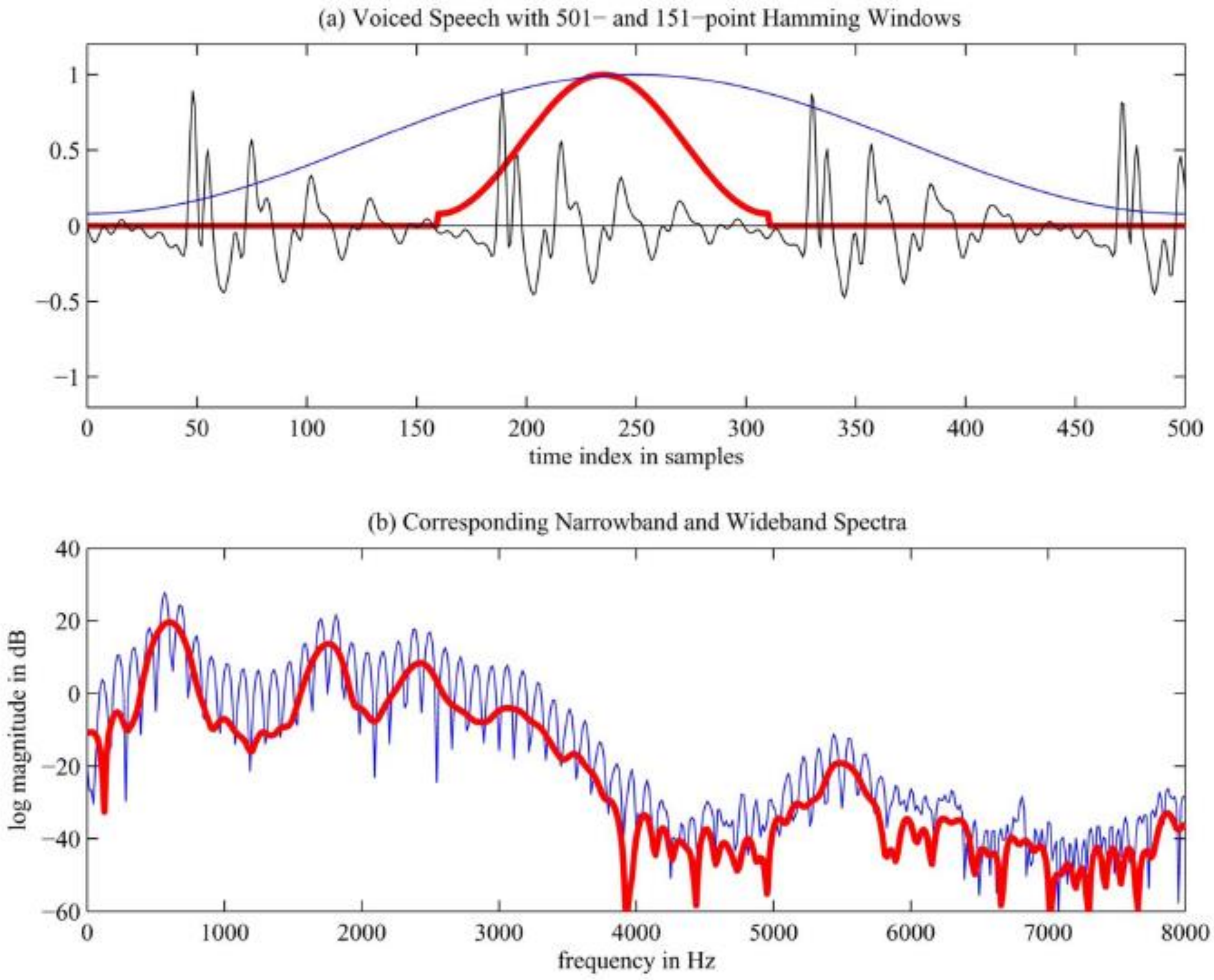
Efeitos do comprimento das janelas -

Janela de Hamming – segmento sonoro



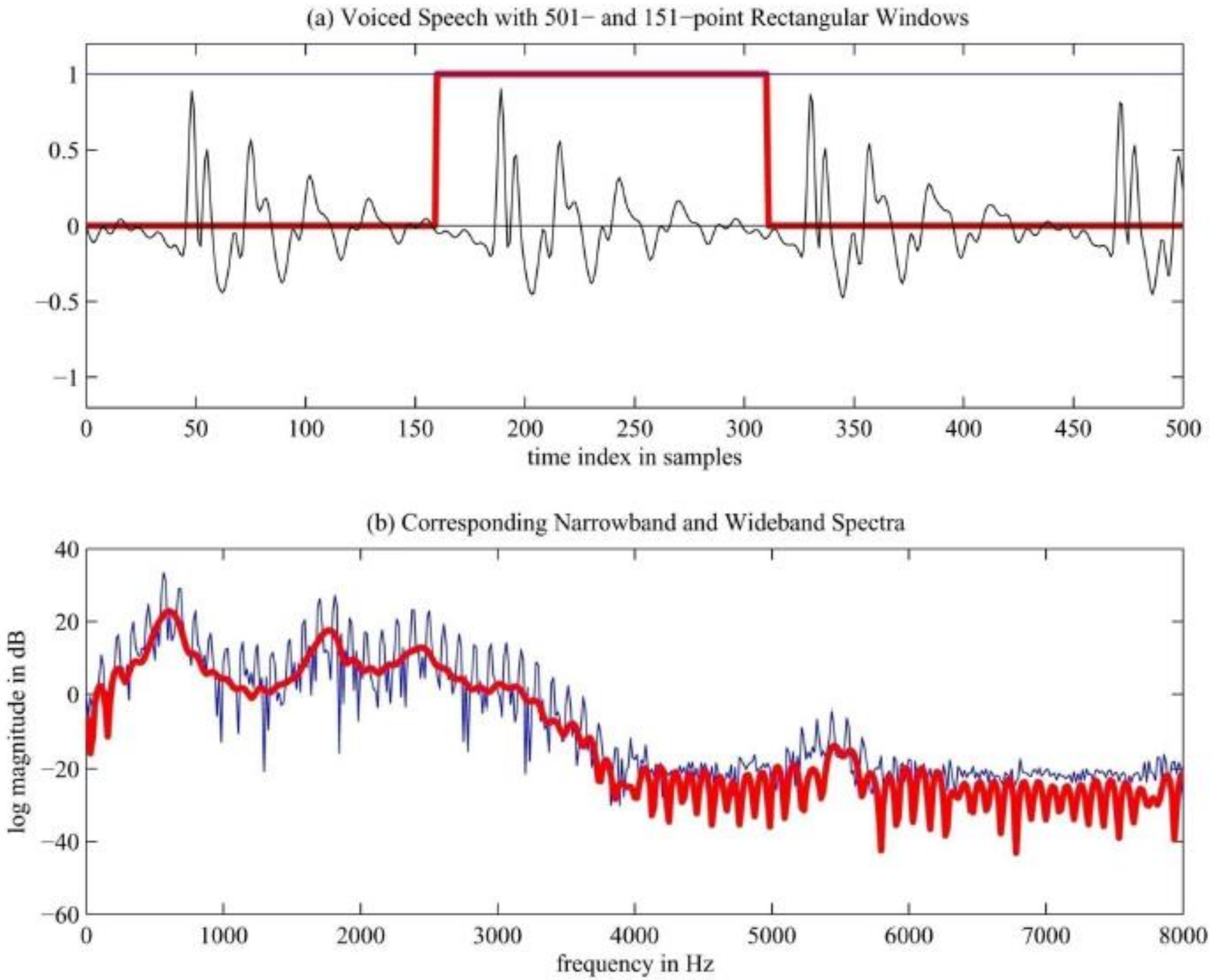
Efeitos do comprimento das janelas -

Janela de Hamming – segmento sonoro



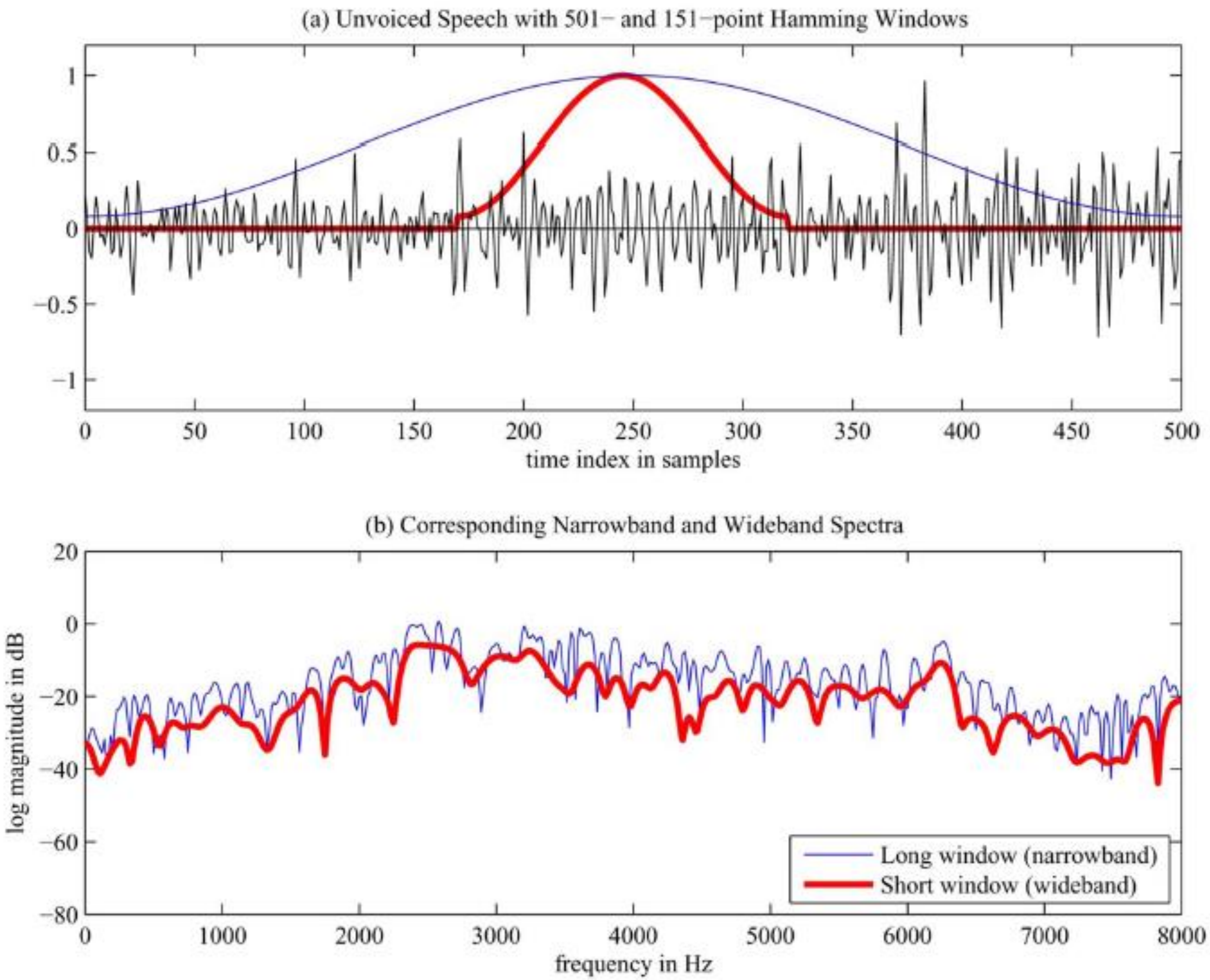
Efeitos do comprimento das janelas -

Janela Retangular - segmento sonoro



Efeitos do comprimento das janelas -

Janela Hamming – segmento surdo



Relação com a Autocorrelação a curto intervalo de tempo

 $X_{\hat{n}}(e^{j\hat{\omega}})$ is the discrete-time Fourier transform of $w[\hat{n}-m]x[m]$ for each value of $\hat{n}$, then it is seen that

$$S_{\hat{n}}(e^{j\hat{\omega}}) = |X_{\hat{n}}(e^{j\hat{\omega}})|^2 = X_{\hat{n}}(e^{j\hat{\omega}})\, X_{\hat{n}}^*(e^{j\hat{\omega}})$$

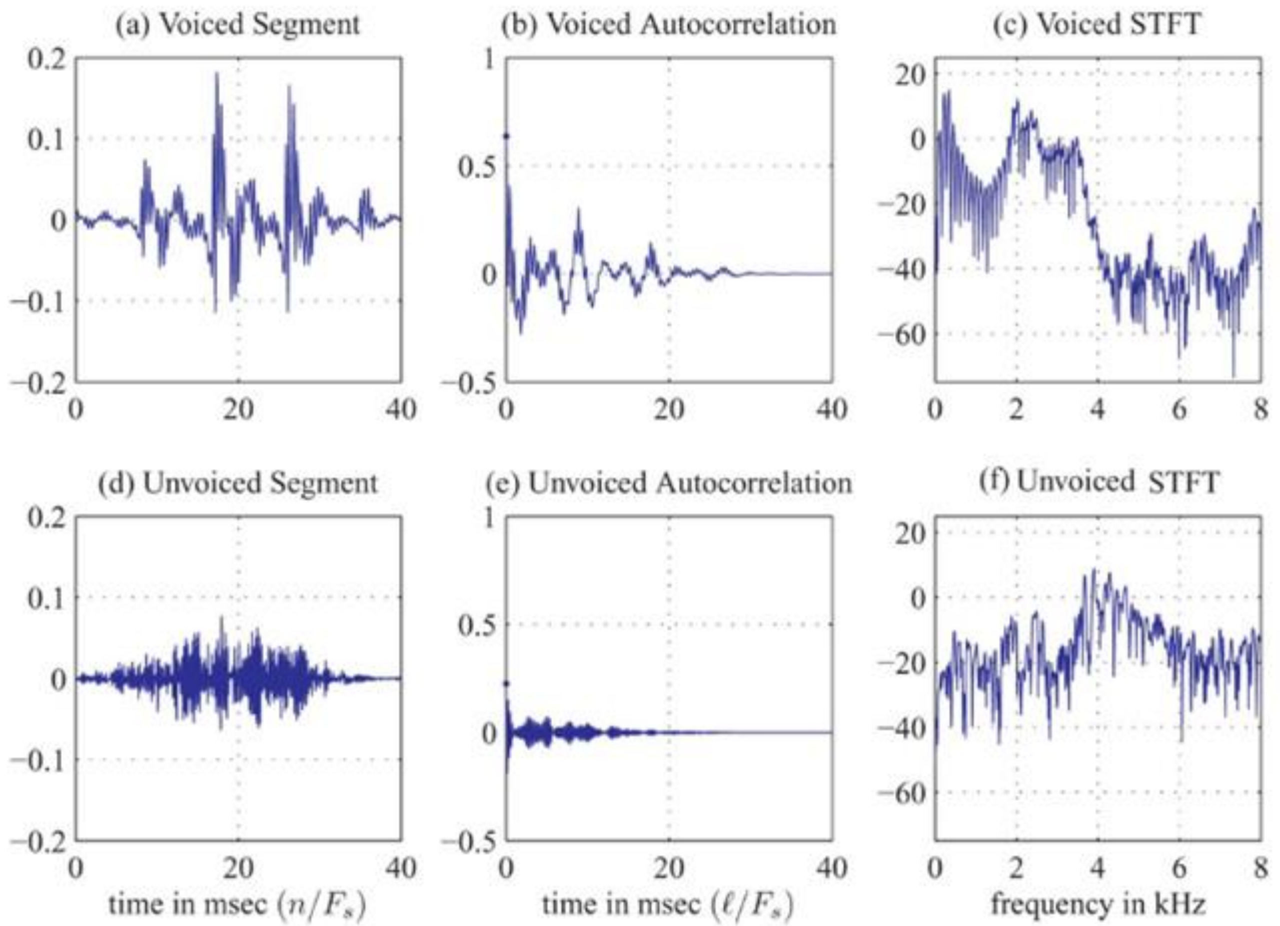
is the Fourier transform of

$$R_{\hat{n}}(l) = \sum_{m=-\infty}^{\infty} w[\hat{n}-m]x[m]w[\hat{n}-l-m]x[m+l]$$

which is the short-time autocorrelation function.

Thus the above equations relate the short-time spectrum to the short-time autocorrelation,

Autocorrelação a curto intervalo de tempo e STFT



Resumo da TF vista da STFT

- interpret $X_{\hat{n}}(e^{j\omega})$ as the normal Fourier transform of the sequence $w(\hat{n}-m)x(m), -\infty < m < \infty$
- properties of this Fourier transform depend on the window
  - frequency resolution of $X_{\hat{n}}(e^{j\omega})$ varies inversely with the length of the window => want long windows for high resolution
  - want $x(n)$ to be relatively stationary (non-time-varying) during duration of window for most stable spectrum => want short windows

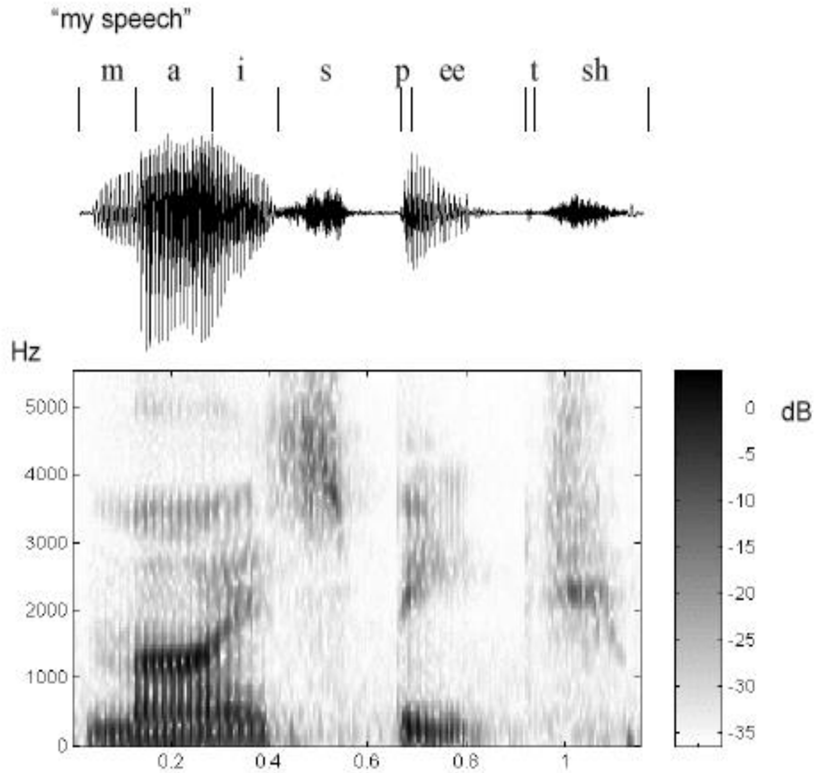
⇔ as usual in speech processing, there needs to be a compromise between good temporal resolution (short windows) and good frequency resolution (long windows)

Análise Espectrográfica

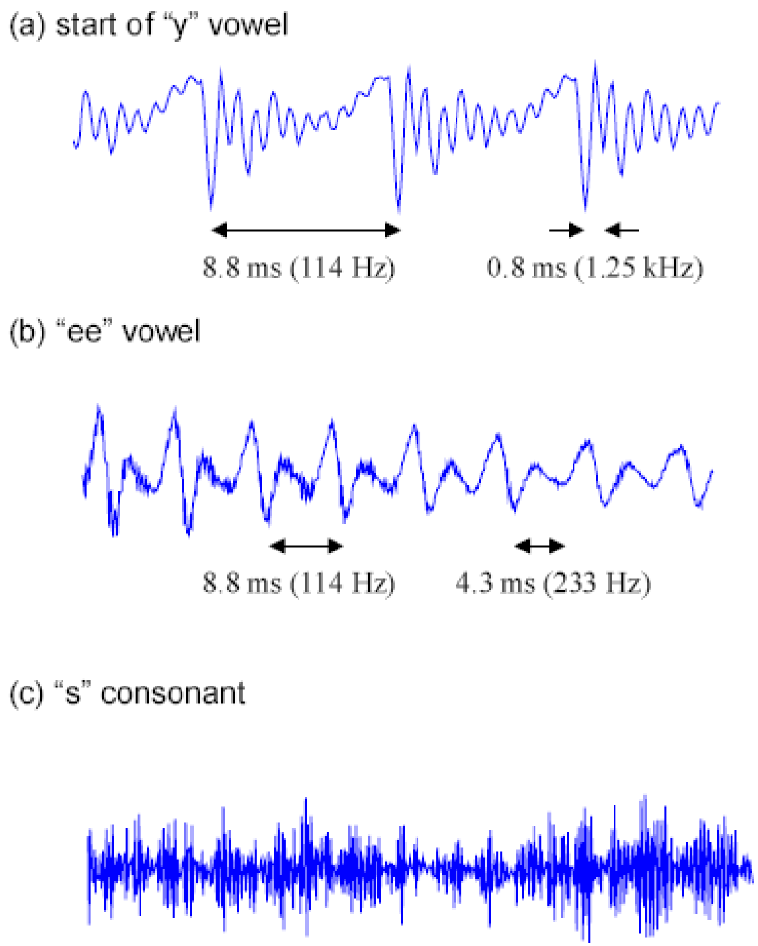
- Permite examinar as propriedades tempo-frequência do sinal de voz e identificar as unidades linguísticas da fala.
- A maior parte das propriedades acústico-fonéticas da fala podem ser identificadas diretamente dos espectrogramas.
- Fala contém mais informações que a escrita, já que contém informações sobre o estado geral do indivíduo (gênero, emoções, idade, sotaque regional, entre outros).
- O espectrograma traduz o que é normalmente percebido pelos nossos ouvidos num domínio visual, numa representação tempo-frequência.
- Permite decodificar o sinal de voz

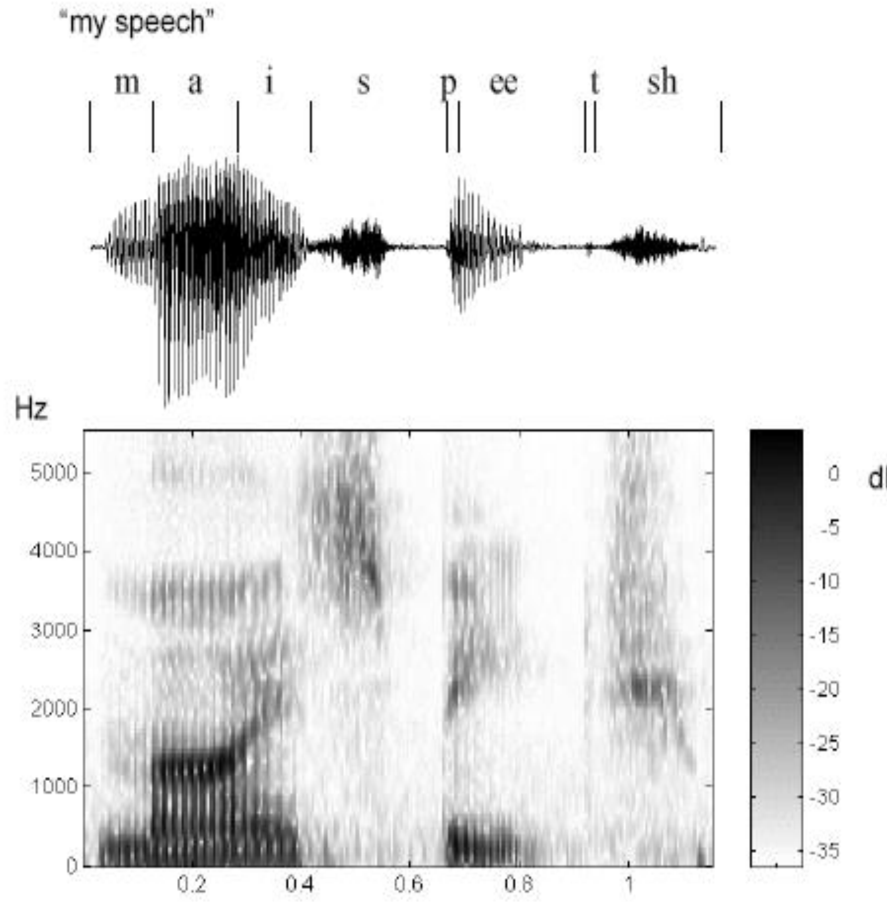
Exemplo - Espectrograma

- Áreas escuras do espectrograma mostram alta intensidade;
- Segmentos sonoros são mais fortes que segmentos surdos;
- Faixas escuras na horizontal são os picos dos formantes;
- “s” tem alta frequência (cerca de 4,5 kHz);
- “sh” é de frequência mais baixa por que a língua está mais para trás;
- Faixas verticais em “my” são os fechamentos laríngeos individuais;
- O “y” de “my” é um ditongo com duas vogais sucessivas.

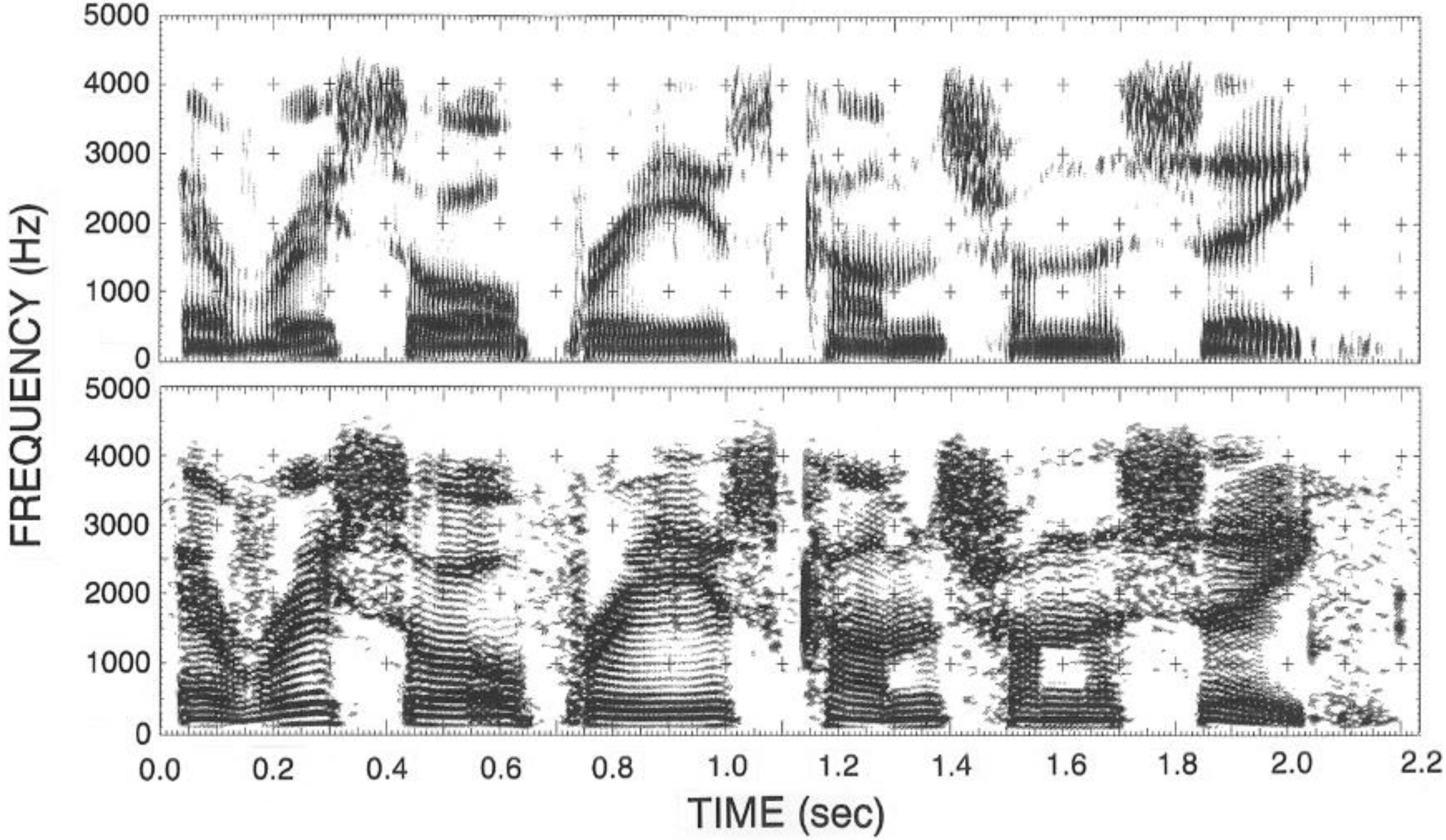


Formas de onda e espectrograma

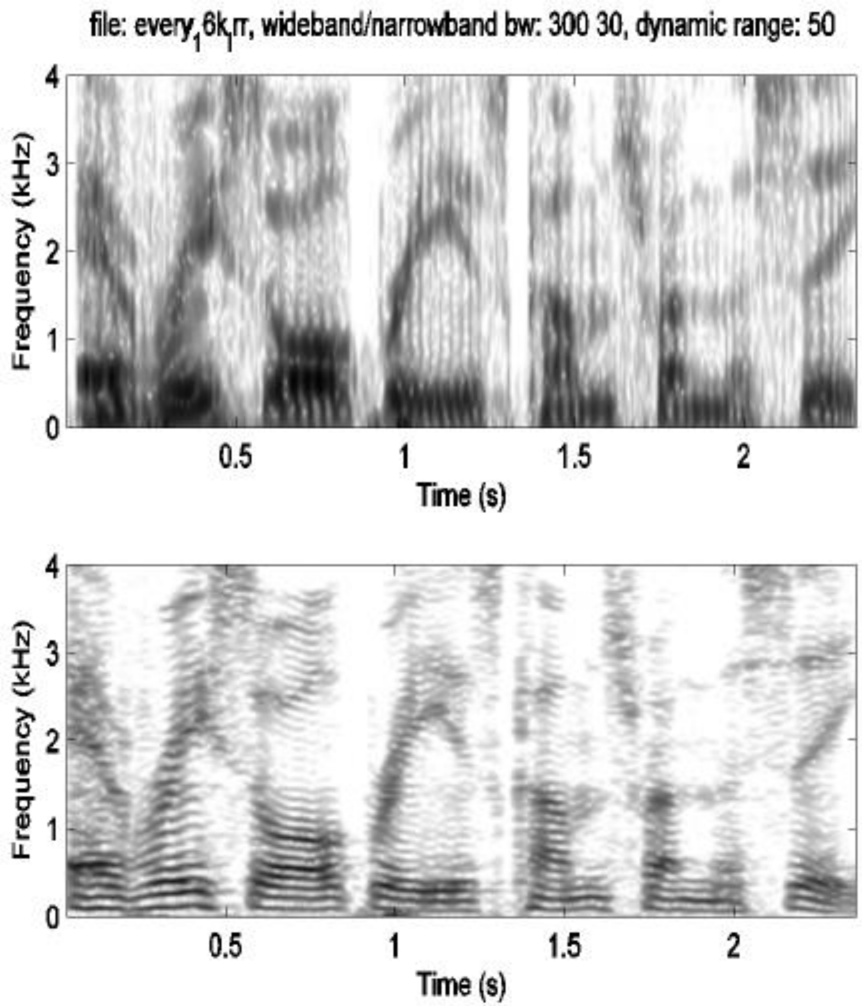




Exemplo - Espectrograma



Espectrograma - Tipos

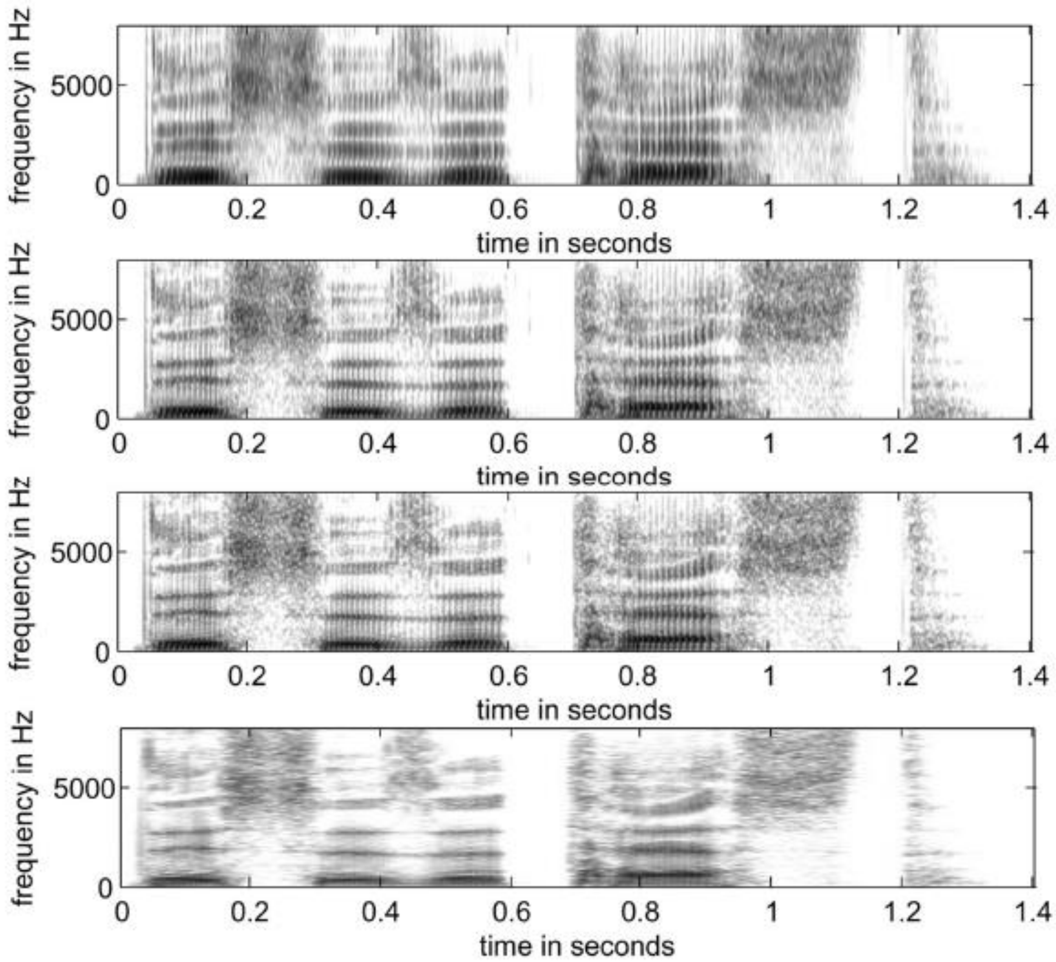


- **wideband spectrogram**
  - follows broad spectral peaks (formants) over time
  - resolves most individual pitch periods as vertical striations since the IR of the analyzing filter is comparable in duration to a pitch period
  - what happens for low pitch males—high pitch females
  - for unvoiced speech there are no vertical pitch striations
- **narrowband spectrogram**
  - individual harmonics are resolved in voiced regions
  - formant frequencies are still in evidence
  - usually can see fundamental frequency
  - unvoiced regions show no strong structure

Espectrogama - análise

- Speech Parameters ("This is a test"):
  - sampling rate: 16 kHz
  - speech duration: 1.406 seconds
  - speaker: male
- Wideband Spectrogram Parameters:
  - analysis window: Hamming window
  - analysis window duration: 6 msec (96 samples)
  - analysis window shift: 0.625 msec (10 samples)
  - FFT size: 512
  - dynamic range of spectral log magnitudes: 40 dB
- Narrowband Spectrogram Parameters:
  - analysis window: Hamming window
  - analysis window duration: 60 msec (960 samples)
  - analysis window shift: 6 msec (96 samples)
  - FFT size: 1024
  - dynamic range of spectral log magnitudes: 40 dB

Espectrograma - Tipos



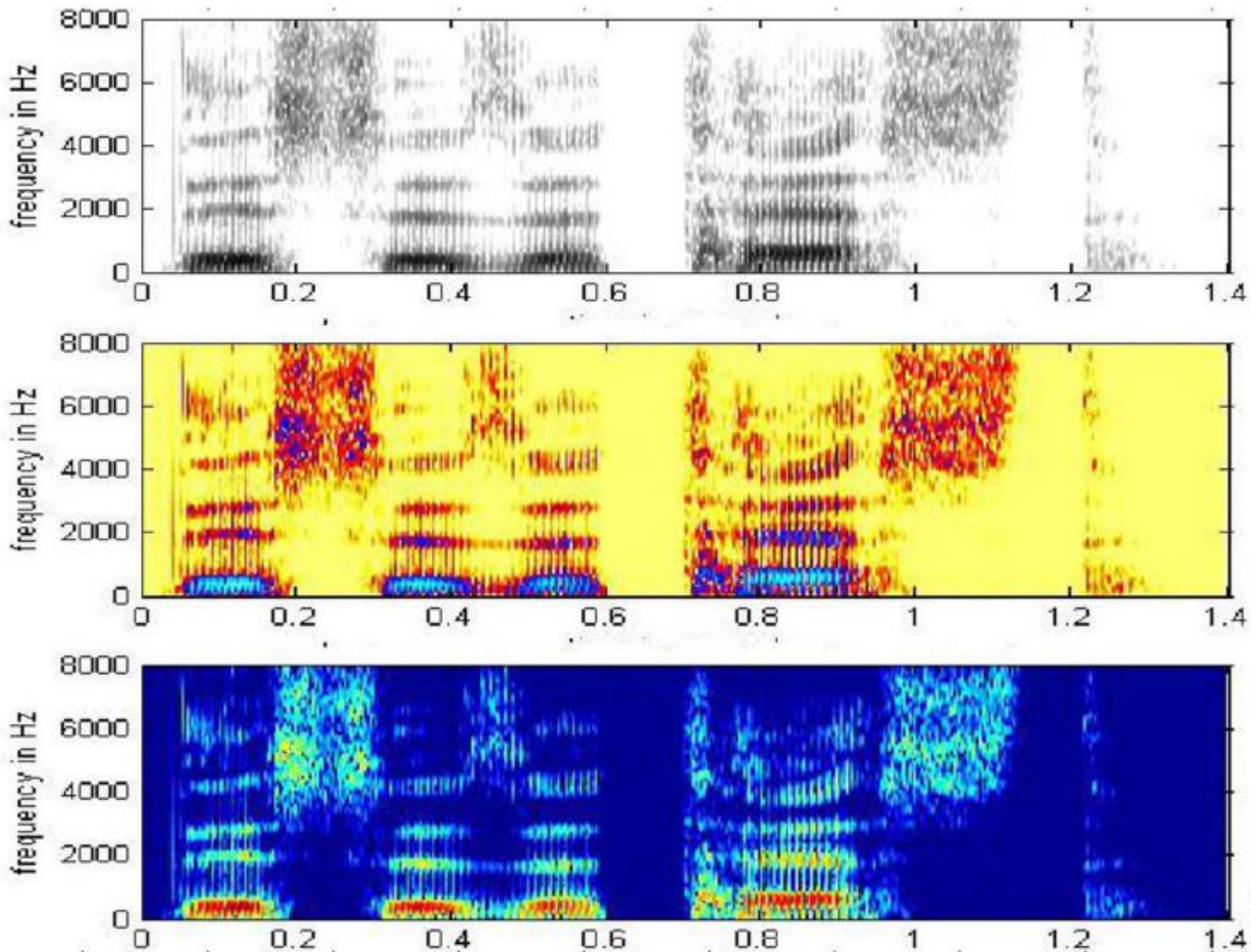
Top Panel:
3 msec (48 samples) window

Second Panel:
6 msec (96 samples) window

Third Panel:
9 msec (144 sample) window

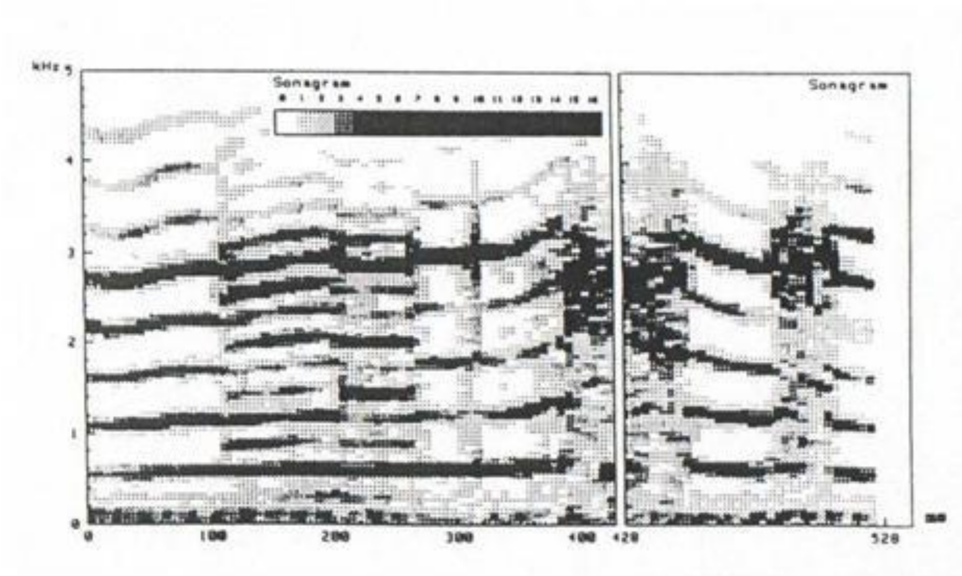
Fourth Panel:
30 msec (480 sample) window

Espectrograma - Tipos



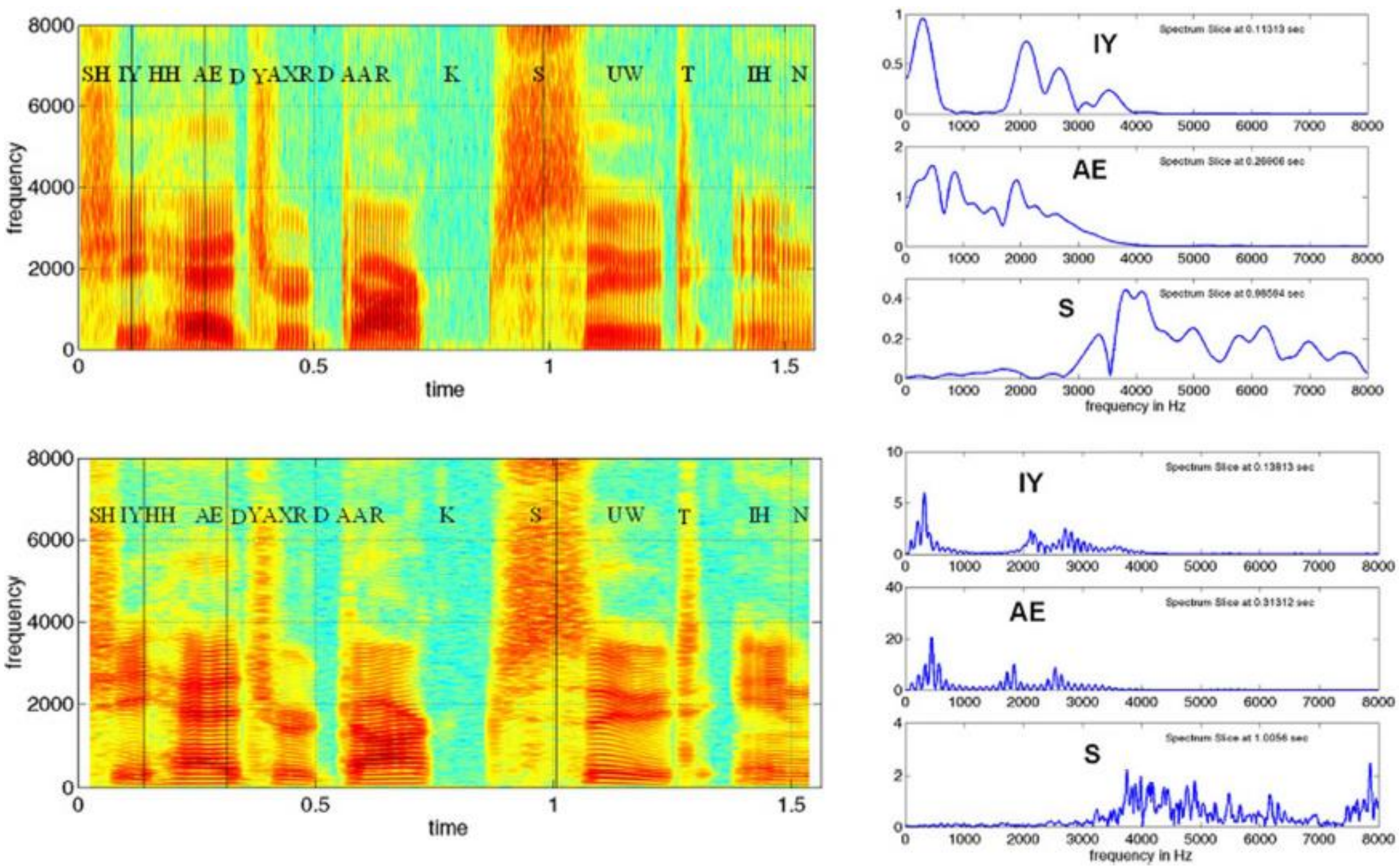
Espectrogama - exemplo

Espectrograma de um choro recém-nascido



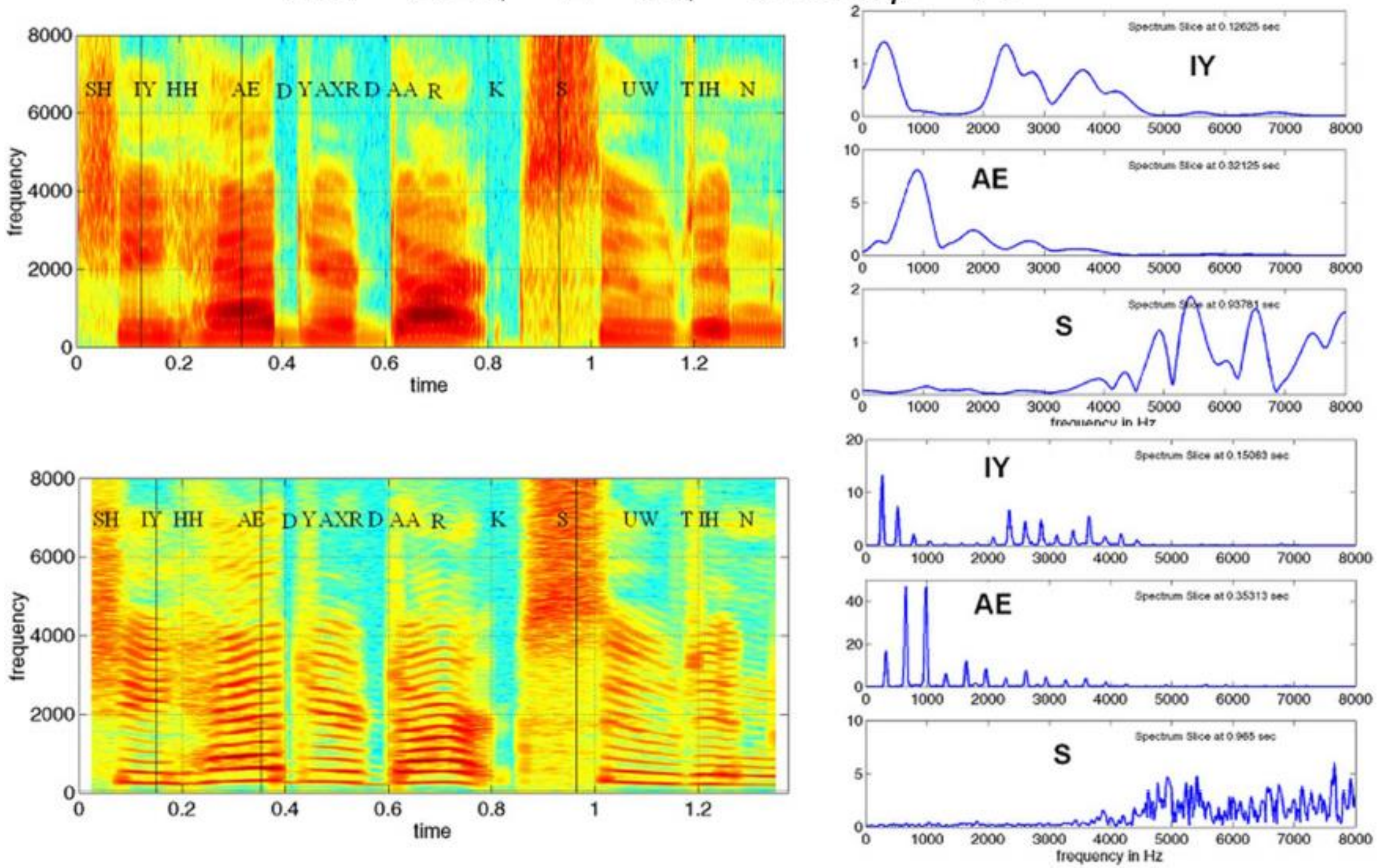
Espectrograma – Tipos - male

$\text{nfft} = 1024, \quad L = 80, \quad Overlap = 75$



Espectrograma – Tipos - female

$$\text{nfft} = 1024, \quad L = 80, \quad Overlap = 75$$



Tarefas pós-aula

- Encontrar a autocorrelação a curto intervalo de tempo para as palavras aplausos e seu primeiro nome;
- Achar a Transformada de Fourier para a função de autocorrelação.
- Plotar os gráficos de autocorrelação e sua transformada (Densidade espectral);
- Construir espectro e espectrograma das palavras, caracterizando os fonemas no gráfico.
- Adicionar ruído à palavra e verificar os efeitos pelo espectrograma;
- Variar o tamanho das janelas, observar e relatar os efeitos no espectrograma resultante;
- Prazo de entrega: três semanas após a aula.

Referências

- Rabiner, Lawrence. Lecture 9: Short-Time Fourier Transform (STFT) Concepts. Digital Speech Processing Course (Winter 2013). Disponível em: http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/digital%20speech%20processing%20course/lectures_new/Lecture%209_winter_2012.pdf. Acesso em 30/03/2013.

- Naylor, Patrick A. SPEECH PROCESSING. Overview. Imperial College London, Spring Term 2008/9. Disponível em: http://www.ee.ic.ac.uk/hp/staff/pnaylor/notes/Overview.pdf. Acesso em 01/04/2013.

- Juang, Gina. Chaos in Vocal Cord Vibration – A Look at the Evidences and Promises it Provid.e