

# *Processamento Digital de Sinais de Voz*

Aula 02- *Fundamentos de produção da voz*

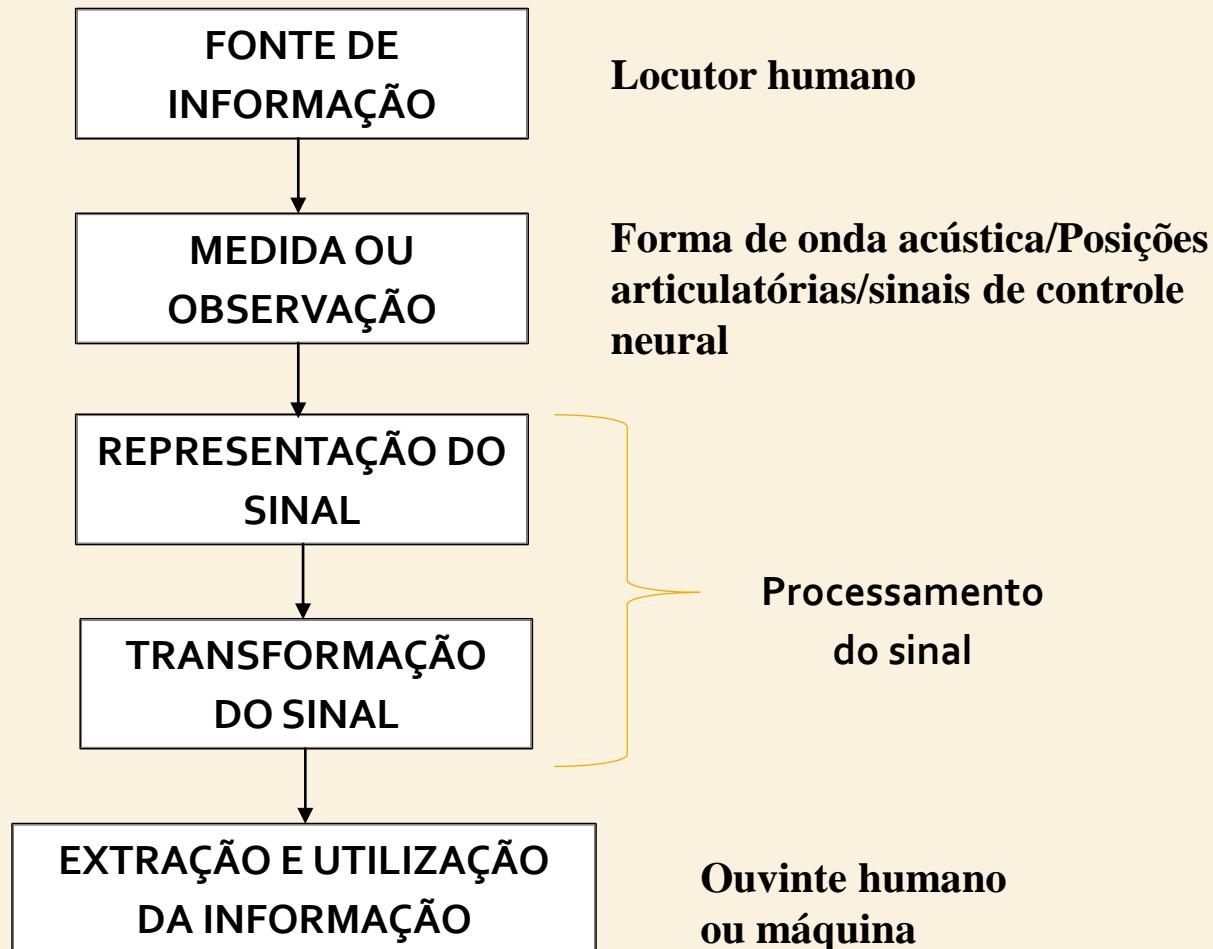
---

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

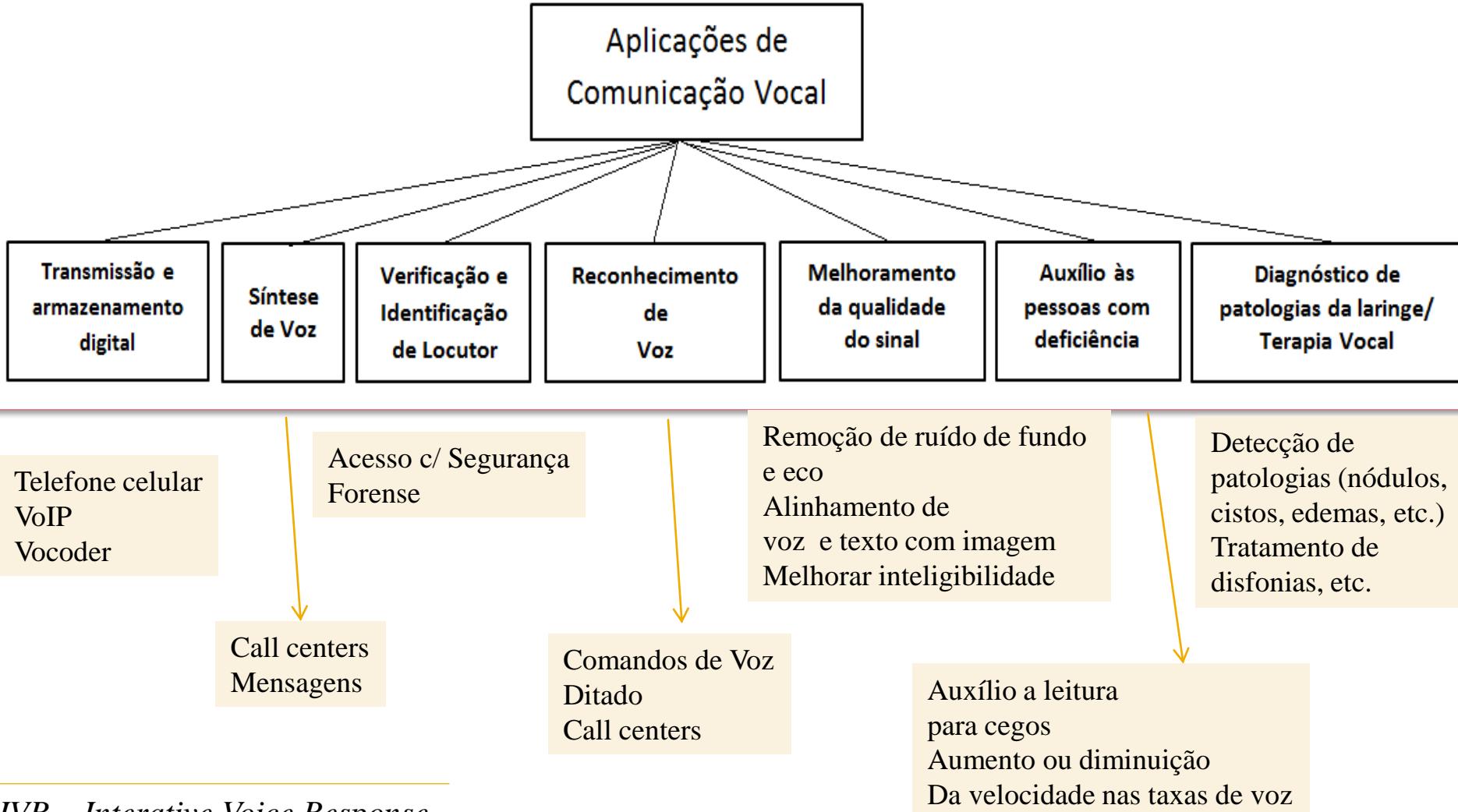


# Modelo do Processamento de Voz

Visão Geral da manipulação e processamento da informação



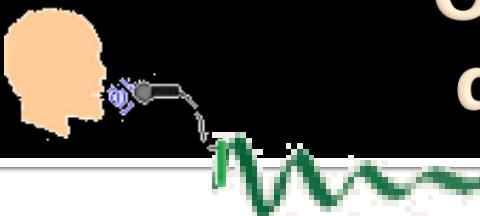
# Aplicações Típicas





# Conceitos Básicos

- A voz é o meio mais natural de comunicação humana.
- A voz pode ser usada mesmo sem contato visual;
- A voz é uma das extensões mais fortes da personalidade humana.
- A partir, unicamente da voz, é possível identificar várias características de quem fala (idade, sexo, lugar onde mora, estado emocional, estado de saúde, grupo sócio-econômico-cultural, etc.).<sup>1</sup>
- <sup>1</sup> <http://www.dsc.ufcg.edu.br/~joseana/PDSV.html/Notas> de aula/[Notas de Aula 02 \(.zip\)](#) (Conceitos Fundamentais)

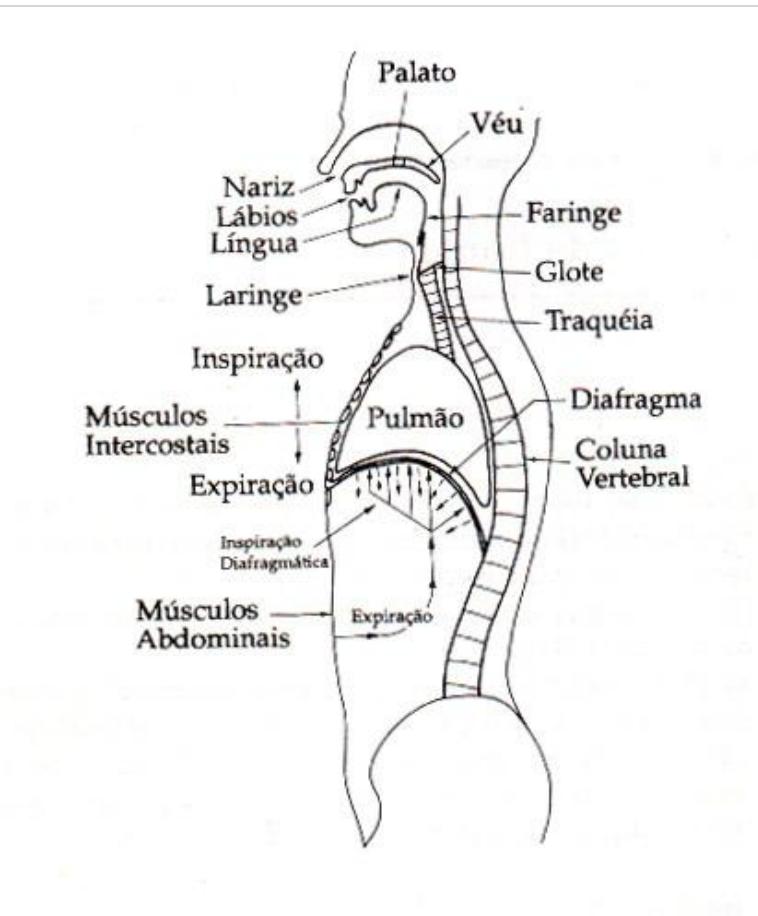


# O Mecanismo Natural de Produção da Fala

- Fisiologia do processo de produção da fala
  - Modelamento físico e matemático para a construção do sistema desejado.
- A voz humana
  - Movimento sonoro audível
  - Resultado da ação de um conjunto de estruturas que formam um sistema para produção dos sons → pulmões, traqueia, laringe, faringe, cavidades nasais e a cavidade oral.



# Anatomia do Aparelho Fonador

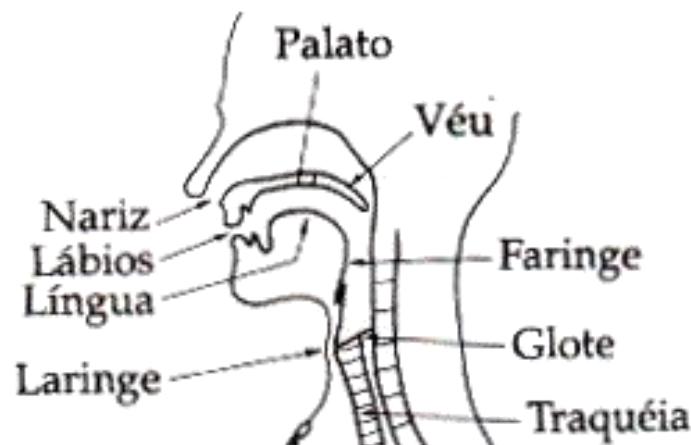


É o movimento respiratório que proporciona o suprimento de energia necessário à fonação.

A fala é produzida através da liberação de ar dos pulmões para o trato vocal.

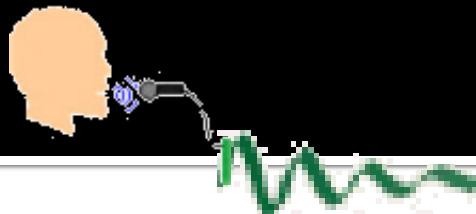


# O Trato Vocal



- Região compreendida entre a glote e os lábios, da qual participam várias cavidades: laringe, faringe, cavidades oral e nasal.

# O Trato Vocal



- Tubo sonoro de 17cm de comprimento e 4cm de diâmetro, em posição de repouso
- Ressonância: 500, 1500 e 2500Hz.

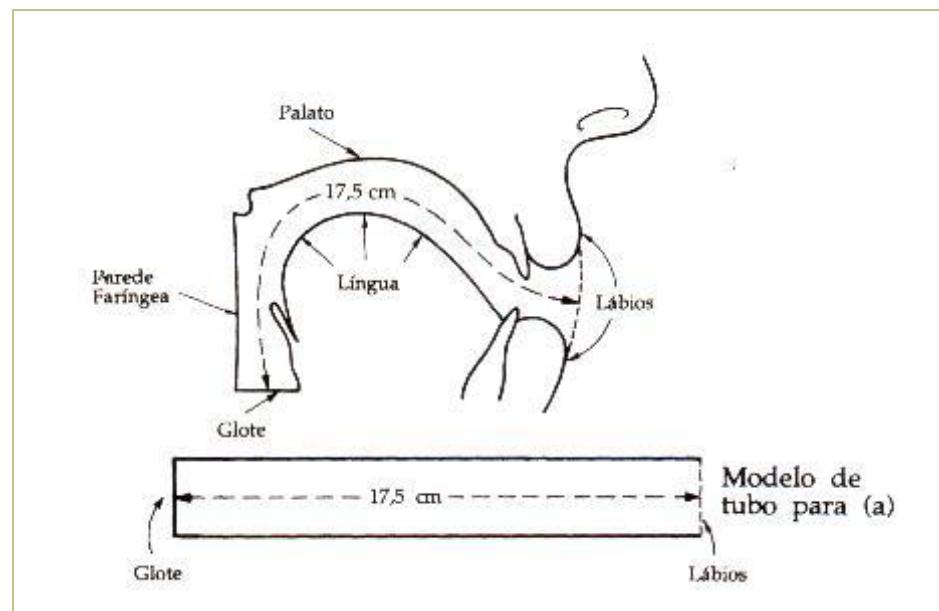
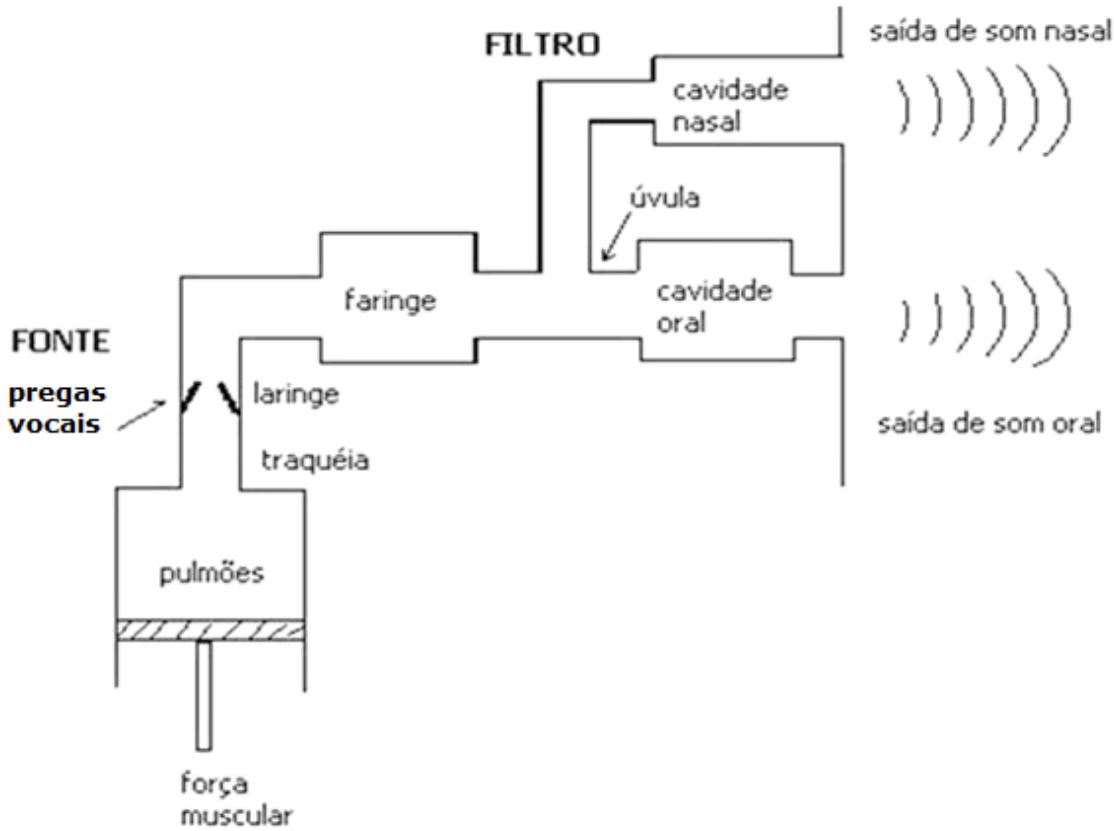


Diagrama da forma do trato oro-faríngeo e um modelo de tubo do trato para a vogal /a/ (Fonte: RUSSO, 1993).



# A produção da voz humana



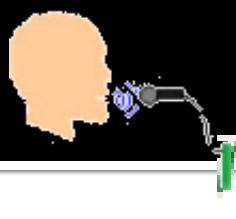
O ar é conduzido para fora dos pulmões pela traqueia, passando pela laringe, onde estão as pregas vocais.

No espaço entre as dobras vocais (globo), o fluxo contínuo de ar dos pulmões é geralmente transformado em vibrações rápidas e audíveis quando falamos.

É produzida uma sequência de pulsos cuja frequência é controlada pela pressão do ar e pela tensão e comprimento das pregas vocais.

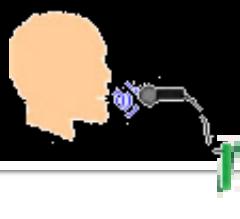
**Diagrama de blocos para a produção da voz humana – Sistema Fonte-filtro.**

Fonte: Deller, Proakis and Hansen, 2010. Adaptação.



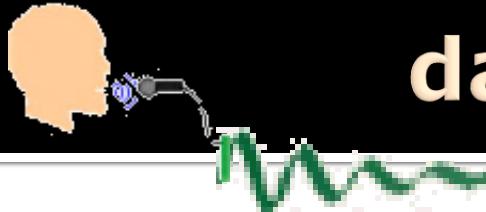
# Classificação dos sons da fala

- **vozeados ou sonoros** → Sons produzidos pela vibração das pregas vocais. Ex: vogais.
- **não-vozeados ou surdos** → Sons nos quais não há vibração das pregas vocais. Ex:[s].
- **Fricativos** → Sons nos quais o fluxo de ar é constringido em algum ponto do trato vocal, elevando-se a língua em direção ao palato, tornando-se turbulento e produzindo um ruído de amplo espectro. Ex: [f].



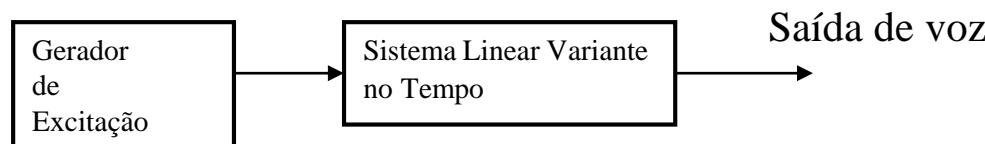
# Classificação dos sons da fala

- **Plosivos** → Sons nos quais o fluxo de ar é interrompido totalmente em algum ponto do trato vocal, e então a pressão formada liberada de uma só vez. Ex: [p] e [t].
  - Sons fricativos ou plosivos também podem ser surdos ou sonoros. Ex.: [f] (fricativo surdo) e [v] (fricativo sonoro).
- **Misto** → Um som pode ser simultaneamente sonoro e surdo. Ex: fonema /z/ na frase “três zebras”.
  - Alguns sons da voz são de uma região curta de silêncio, seguida por um região de voz sonora, voz surda, ou ambas.
  - Também denominados fricativos surdos ou fricativos sonoros.



# Teoria Acústica da Produção da Fala

- Constituída de representações matemáticas do processo de produção da fala.
- Base para toda a análise e síntese realizada com os sinais da fala.



Modelo simplificado de produção de fala

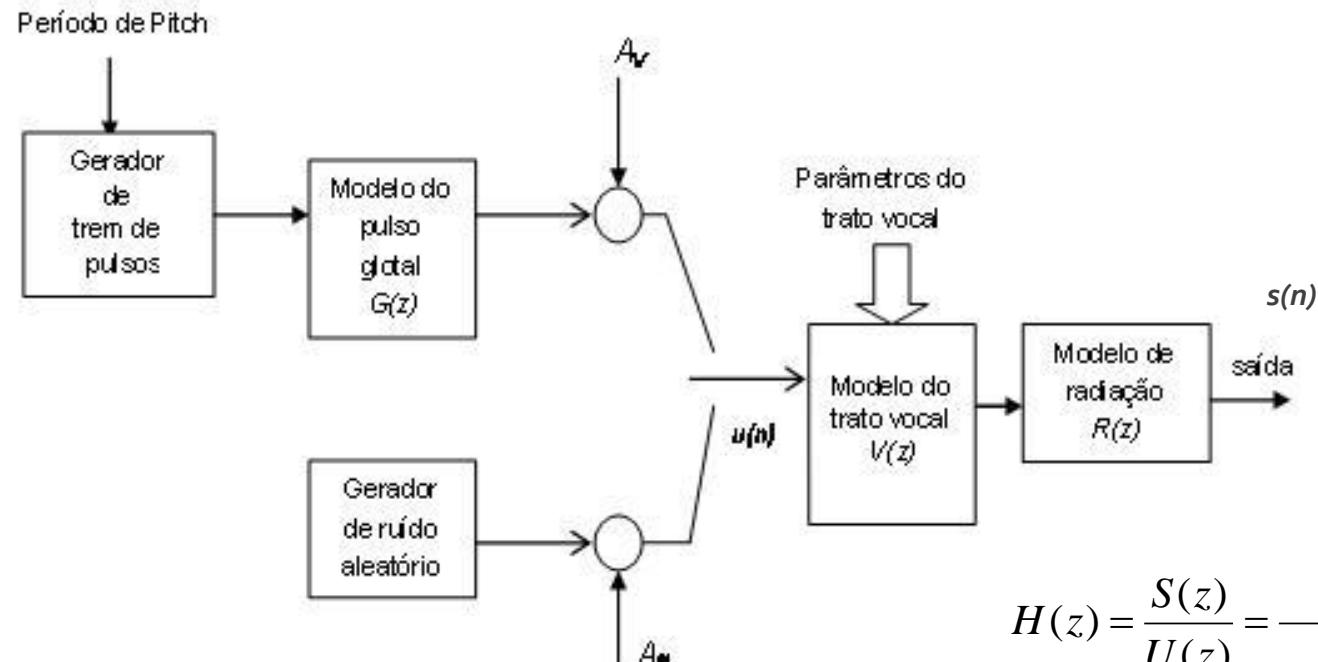
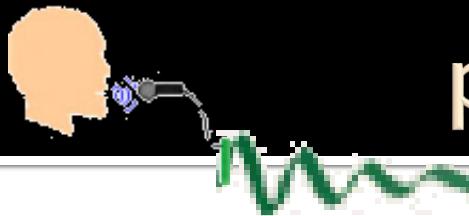
Gerador de excitação → trem de pulsos glóticos/ruído aleatório

Sinais sonoros

Sinais surdos

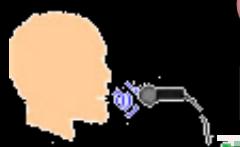


# Modelo geral discreto no tempo para produção de fala



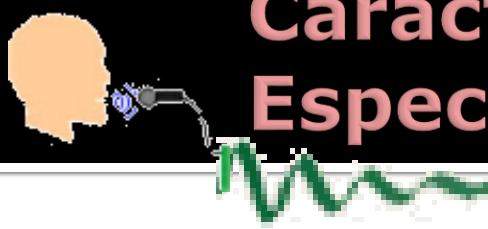
$$H(z) = \frac{S(z)}{U(z)} = \frac{G(z)}{1 - \sum_{k=1}^p \alpha_k z^{-k}}$$

$\alpha_k$  - coeficientes de predição linear (coeficientes LPC).  
 $p$  - ordem do preditor.



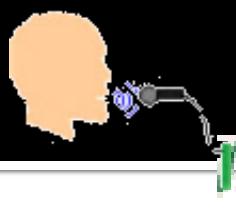
# Características Temporais e Espectrais do sinal de Voz

- O sinal de voz é um sinal cujas características estatísticas variam fortemente com o tempo.
- Pode ser considerado estacionário em trechos muito pequenos - da ordem de dezenas de milisegundos (16 a 32 ms).
- O sinal de voz pode ser dividido em segmentos curtos, nos quais possui propriedades acústicas semelhantes, podendo ser analisado e processado nesses intervalos curtos de tempo.



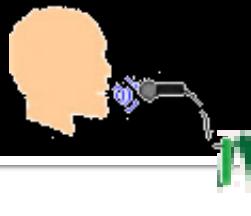
# Características Temporais e Espectrais do sinal de Voz

- Uma das mais importantes características da voz é que a voz não é constituída de sons discretos bem definidos. Existe uma certa intercalação entre os sons em uma elocução → co-articulação.
- As variações evidentes na forma de onda da voz, são uma consequência direta dos movimentos do sistema articulatório da voz, o qual raramente permanece fixo por um considerável período de tempo.
- Estes articulatórios são tecidos humanos e/ou músculos, os quais são movidos de uma posição para outra de forma a produzir os sons de voz desejados



# Características Temporais e Espectrais do sinal de Voz

- Os órgãos usados para produção da voz são compartilhados com outras funções tais como, respiração, alimentação e olfato.
- Expressões faciais ou gestos, podem ser usadas para fornecer entradas adicionais para o ouvinte.
- Para portadores de deficiência auditiva, por exemplo, esses gestos e expressões são de extrema importância na compreensão da mensagem.



# Características Temporais e Espectrais do sinal de Voz

- Amostragem dos sinais de voz
  - A comunicação humana típica está limitada na faixa de 7-8kHz.
  - Em sons sonoros, as altas frequências estão mais do que 40dB abaixo do pico do espectro.
  - Por outro lado, para sons surdos, o espectro não cai consideravelmente mesmo acima de 8kHz.
  - Para o processo de estimação das três primeiras frequências formantes da voz sonora → porção do espectro até 3,5kHz.
  - Pode-se usar uma taxa de amostragem, segundo o critério de Nyquist, de 8kHz.

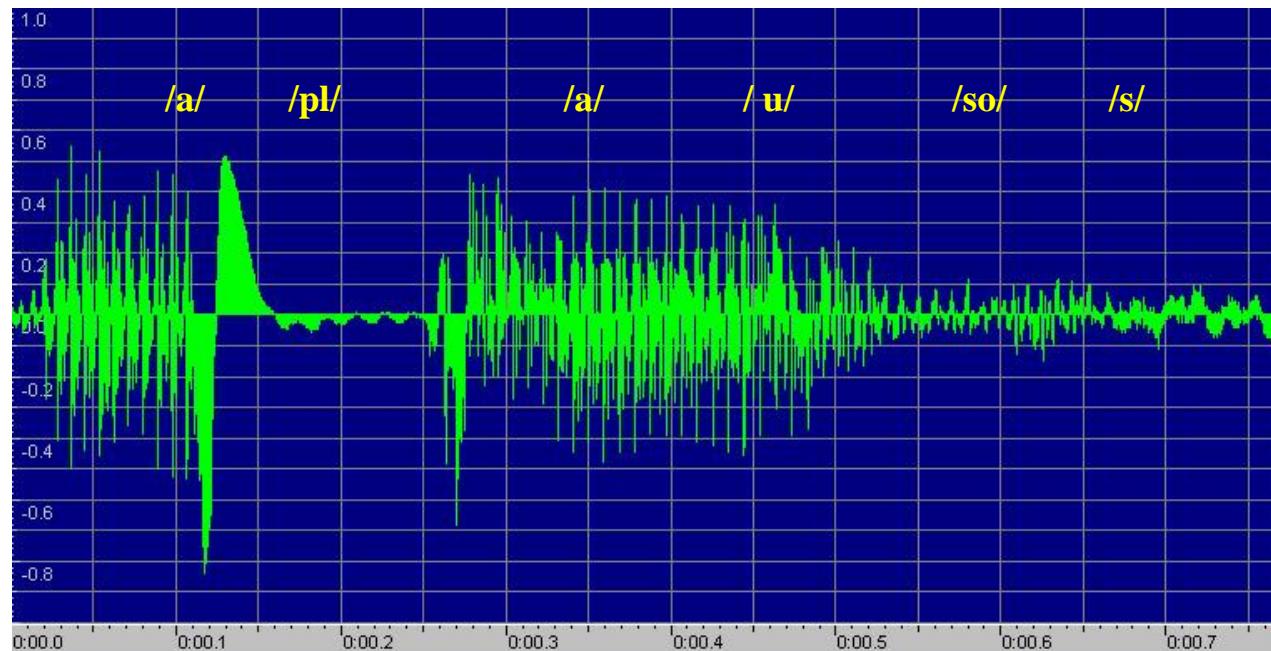


# Representação dos Sinais de Voz

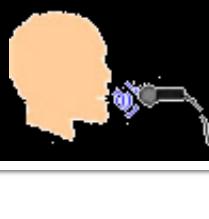
## ■ Forma de Onda:

- Aspectos acústicos e psicológicos da voz;

- Referência para realimentação visual → melhoramento da fala (periodicidade, intensidade, duração, etc.);

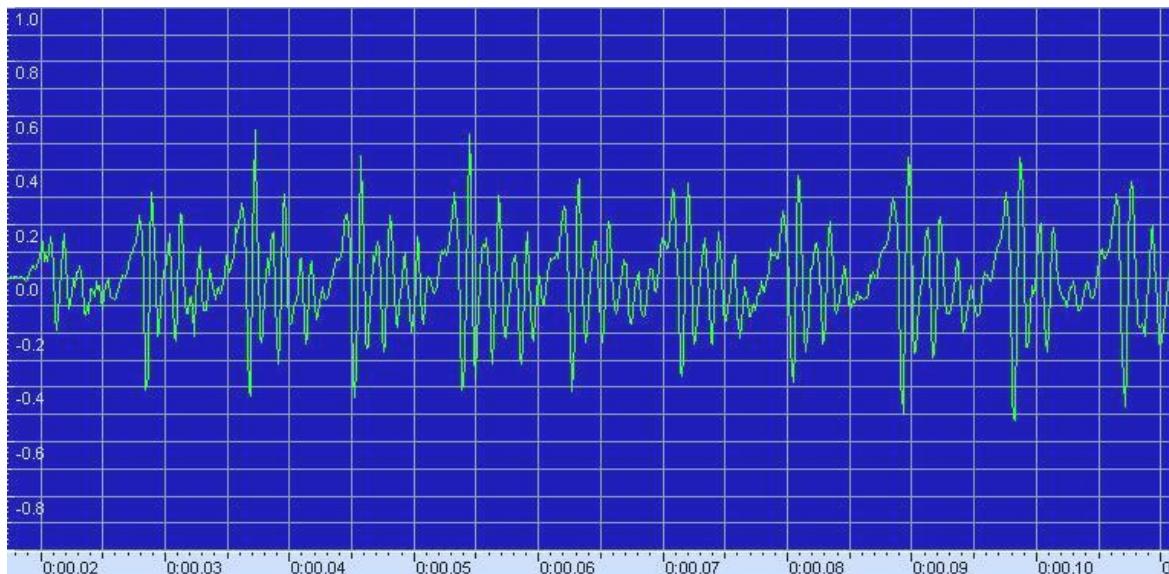


Forma de onda da palavra *aplausos*



# Representação dos Sinais de Voz

## ■ Sons Sonoros



Na produção dos sons sonoros, são obtidas ondas de pressão quase periódicas, excitando o trato vocal, atuando como um ressonador e modificando o sinal de excitação, produzindo frequências de ressonância – as formantes – que caracterizarão os diferentes sons sonoros.

Forma de onda da vogal /a/ na palavra aplausos

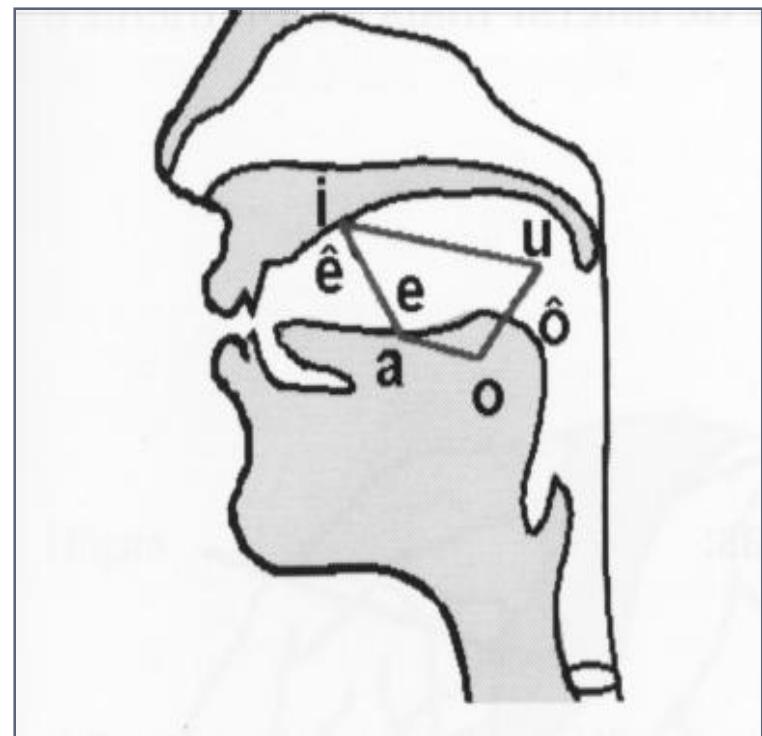


# Classificação dos sons da fala

As vogais:

## 1. Zona de articulação:

- média: a (amor)
- anteriores: é, ê, i (pé, crê, vi)
- posteriores: ó, ô, u (pó, avô, caju)



# Classificação dos sons da fala

## As vogais:

- Vogais anteriores (é, ê, i) - são emitidas abaixando-se a ponta da língua e elevando-se progressivamente a sua parte anterior em direção ao palato duro.
- Vogais posteriores (ó, ô, u) - exigem que se eleve cada vez mais a base da língua em direção ao palato, enquanto os lábios vão tomado uma forma arredondada e fechada.
- A vogal média a é produzida com a língua em posição de descanso e a boca entreaberta.

# Classificação dos sons da fala

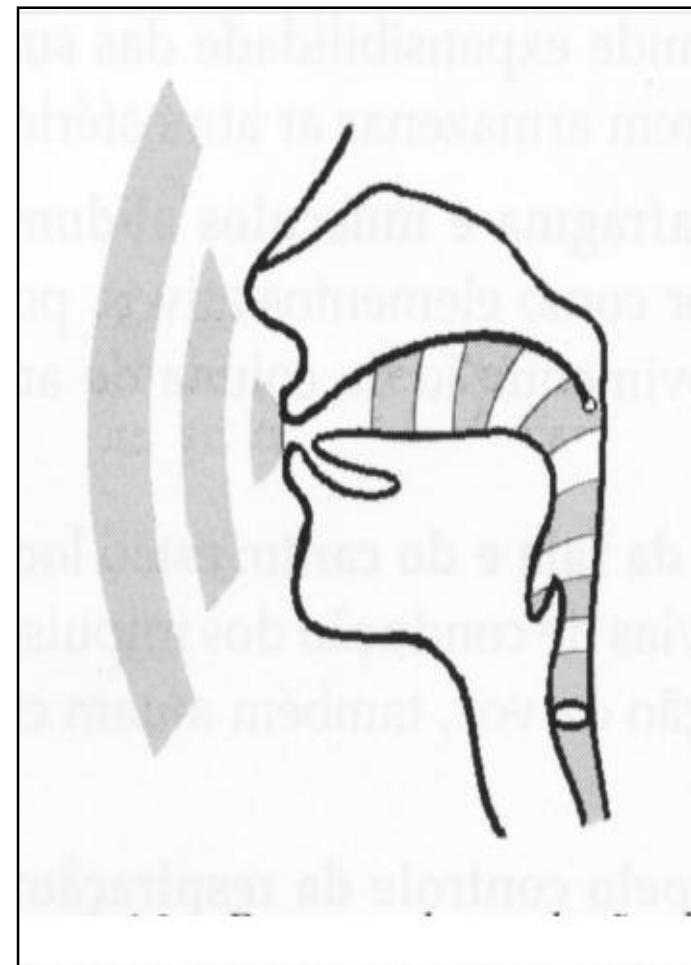
## As vogais:

### 2. Ressoador principal:

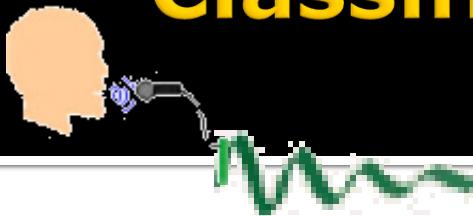
- **Oral** - a, é, ê, i, ó, ô, u como nas palavras *pato, ré, dê, vi, nó, fogo, luva*
- **Nasal** - a, e, i, o , u quando acentuadas pelo til ou quando antecedem o n ou o m, como nas palavras *vã, vento, fim, bom, fundo*

# Classificação dos sons da fala

- A úvula e palato mole exercem uma função de grande importância na fonação, pois podem modificar o timbre do fonema.
- Pelo movimento dessas estruturas, a coluna de ar pode ser expelida pela boca dando origem aos fonemas orais, ou pelo nariz, quando então se formam os fonemas nasais.

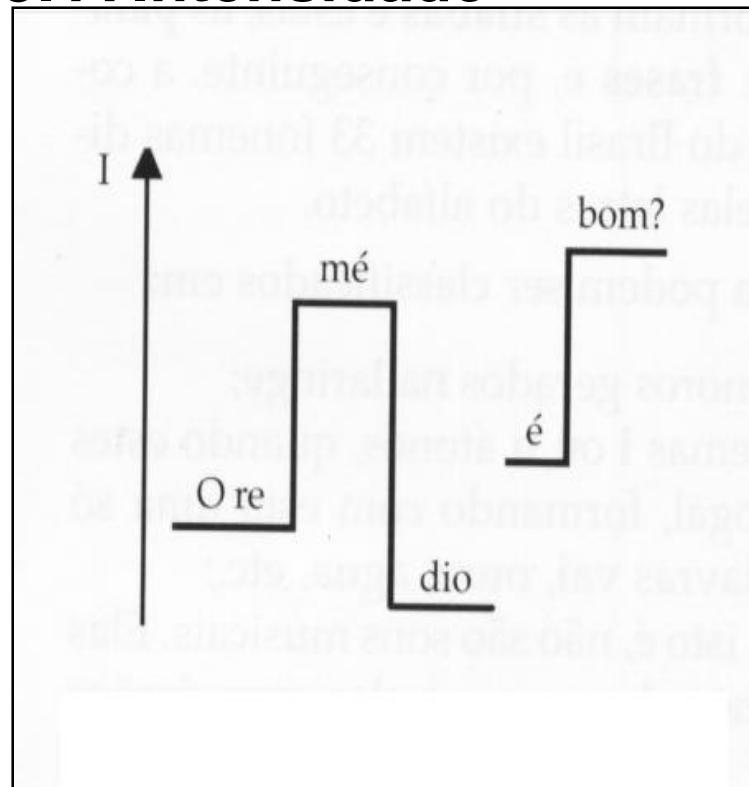


# Classificação dos sons da fala



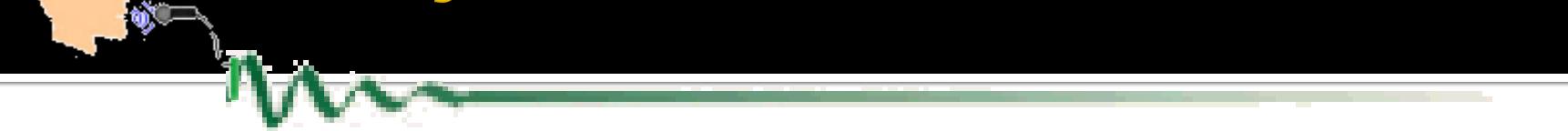
## As vogais:

### 3. A Intensidade



- tônicas: já, acarajé, pelo, aqui, poldro, pus
- subtônicas: armistício, cafezinho, compadre
- átonas: vela, bole, tição, dado, julgar
- **sílaba tônica** - maior intensidade
- **sílabas subtônicas** - média intensidade
- **sílabas átonas** - baixa intensidade

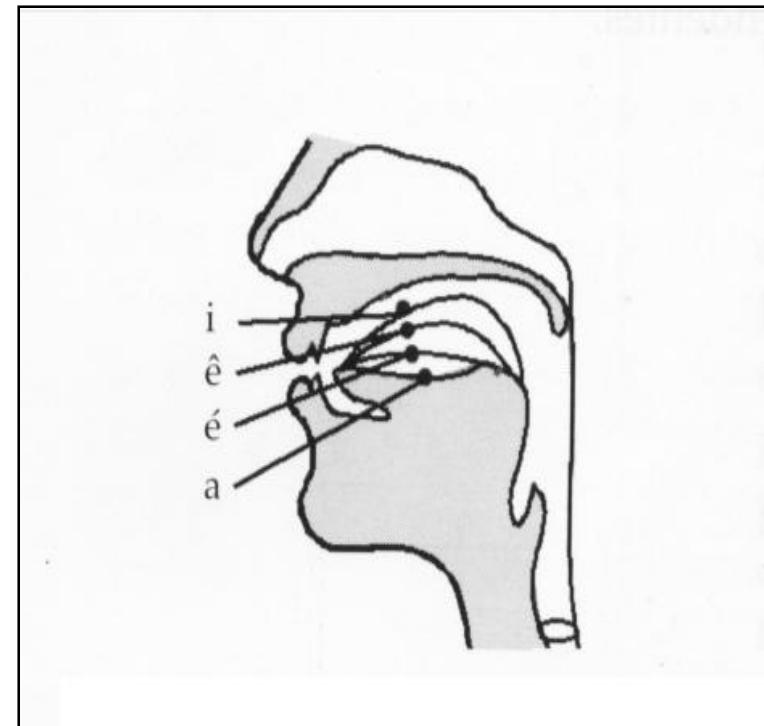
# Classificação dos sons da fala



## As vogais:

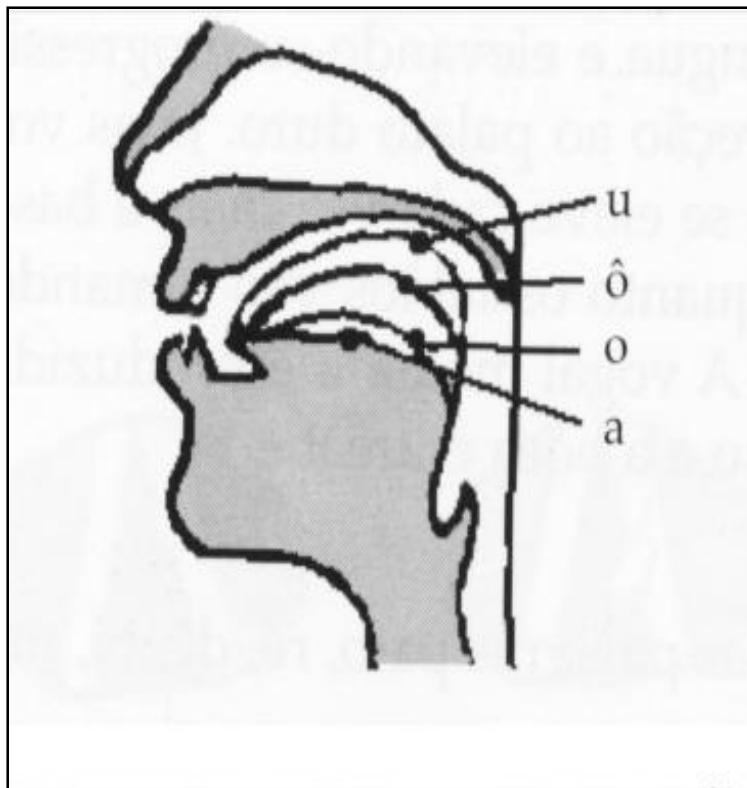
### O timbre

- abertas: a, é, ó (vá, fé, jiló)
- fechadas: ê, ô, i, u e todas as nasais (ipê, dor, vi, itu, vã, fenda)
- reduzidas: as vogais átonas orais ou nasais (revela, bule, sisal, rato, unção, amei, então)



# Classificação dos sons da fala

## As consoantes:



## Classificação:

### 1. Modo de articulação:

- **Oclusivas:** quando ocorre o impedimento completo ao fluxo de saída do ar pela boca
- **Constritivas:** quando ocorre impedimento parcial à expulsão do ar pela boca.

# Classificação dos sons da fala



**Neste grupo estão as consoantes:**

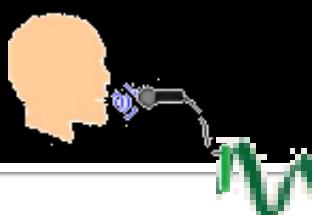
- **fricativas**: quando o ar é expulso por um conduto oral estreitado ou pelos lábios quase fechados (f, v, x, ç, s, z, j)
- **vibrantes**: quando, ao sair, o ar vibra de modo áspero ( r )
- **laterais**: quando a língua fica em contato com o palato e obstrui o canal central da boca, deixando que o ar se escoe por canais situados próximos às bochechas (l, lh)

# Classificação dos sons da fala



## 2. Ponto de articulação

- **bilabiais**: (p, b, m)
- **labiodentais**: (f, v)
- **linguodentais**: (t, d)
- **alveolares**: língua e alvélos (s, z, l, n)
- **palatais**: dorso da língua junto ao palato duro ( j, g com som de j, e o x, lh, nh)
- **velares**: região posterior da língua junto ao palato mole ( c com som de k, q, e o g com som de guê)

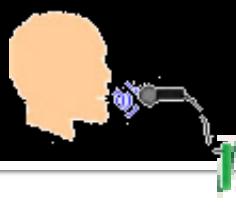


# Frequência fundamental

## Sons Sonoros

- O fluxo de ar vindo dos pulmões é controlado pela abertura e fechamento das pregas vocais.
- Para sons vozeados (sonoros), a frequência fundamental,  $F_0$ , ou "*pitch*", corresponde à frequência do sinal excitatório proveniente da glote.
- $F_0 \rightarrow$  Frequência de vibração das pregas vocais.





# Representação dos Sinais de Voz

- A frequência média dos pulsos é denominada frequência fundamental de excitação,  $F_0$  e o período fundamental (*Pitch*),  $P$ , é dado por:

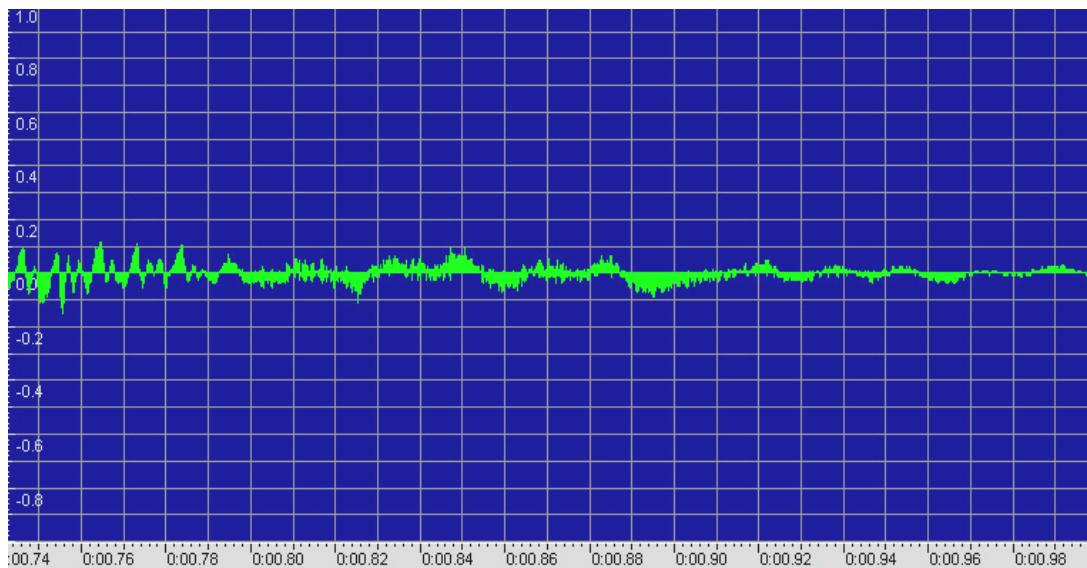
$$P = 1/F_0$$

- A frequência fundamental média no Português brasileiro está em torno de 105 Hz para os homens, 213 Hz para as mulheres e 290 Hz para as crianças.
- Oscilações em torno desta frequência fundamental são chamadas “*jitter*”.



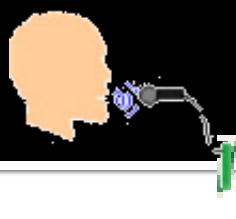
# Representação dos Sinais de Voz

## Sons Surdos:



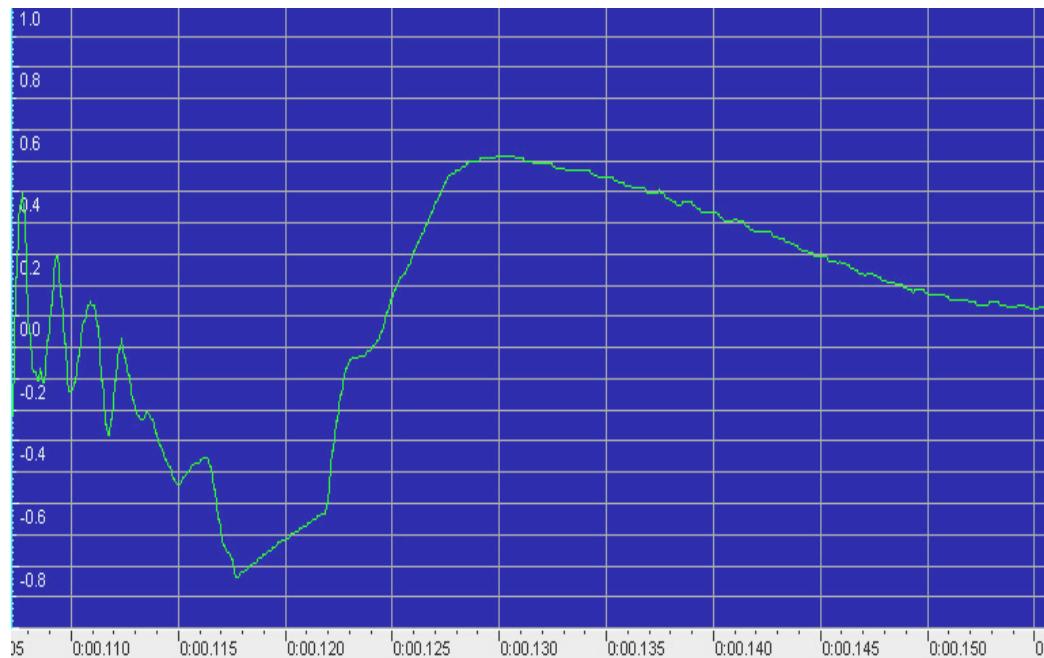
É produzida uma constrição em algum ponto do trato vocal → o ar adquire velocidade suficientemente alta para produzir turbulência gerando uma fonte de ruído (ruído branco) para excitar o trato vocal. A glote permanece aberta, não havendo vibração das dobras vocais.

Forma de onda do fonema /s/ na palavra aplausos.



# Representação dos Sinais de Voz

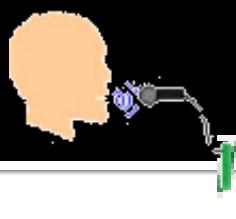
## ■ Sons Plosivos



Forma de onda do fonema /p/ na palavra aplausos.

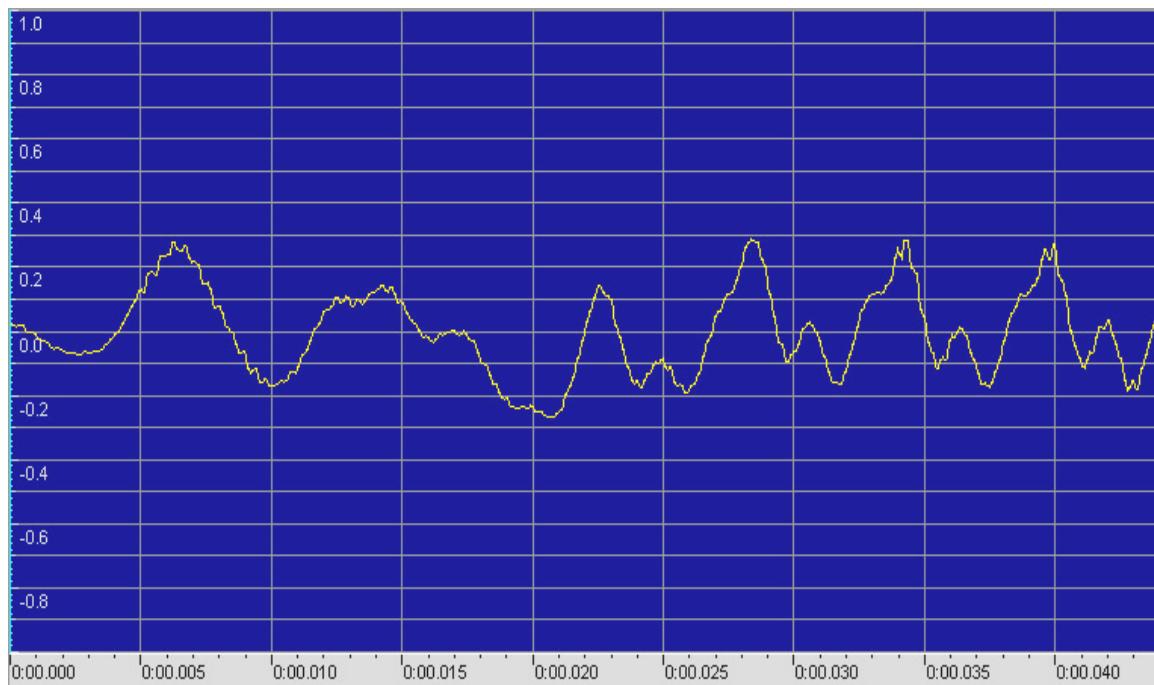
O ar é totalmente dirigido à boca, estando esta completamente fechada.

Com o aumento da pressão, a oclusão é rompida bruscamente, gerando um pulso que excita o aparelho fonador. Com a excitação ocorre um movimento rápido dos articuladores em direção à configuração do próximo som.



# Representação dos Sinais de Voz

- Sons com excitação mista:



Combinação de vibração das pregas vocais e de excitação turbulenta.

Forma de onda do fonema /v/ na palavra viajar.

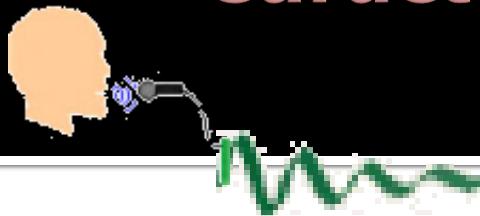


# Representação dos Sinais de Voz



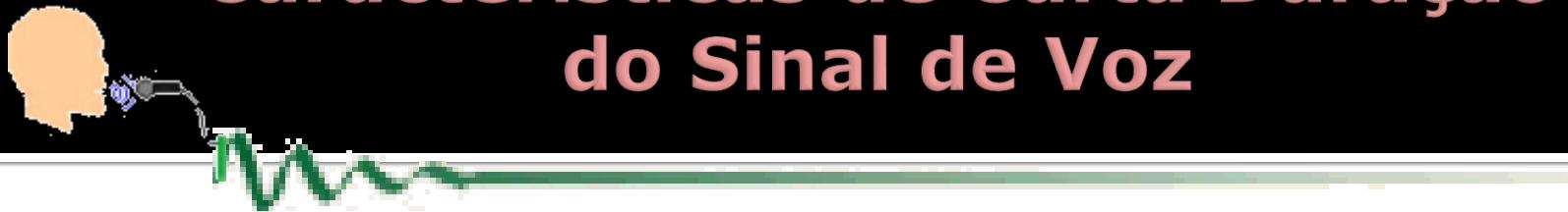
- Fricativos sonoros: /j/, /v/ e /z/
- Fricativos surdos: /s/
- Os fonemas fricativos que compõem o sistema fonológico do português brasileiro distinguem-se quanto aos seguintes pontos de articulação: lábiodentais (/f/, /v/), alveolares (/s/, /z/), e palatais ( /ʃ/ e /ʒ/ ) .

# Características de Curta Duração do Sinal de Voz



- Segmentos de voz localmente estacionários, de período em torno de 32 ms, são denominados segmentos de curta duração ("*short time*");
- Segmentos de voz com duração bem maior são denominados segmentos de longa duração ("*long time*").
- Ergodicidade para curtos intervalos de tempo (10 a 20ms).
- Informações significantes sobre o fenômeno físico da produção da voz podem ser obtidos através da análise espectral de curta duração.

# Características de Curta Duração do Sinal de Voz



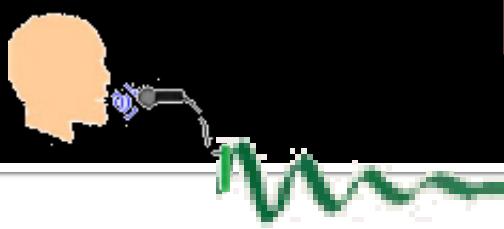
## ■ Energia Segmental

$$E_{seg} = N_A \cdot E\{[s(n) - \mu_s(n)]^2\}$$

$s(n)$  → sinal de voz,

$\mu_s(n)$  → média de  $s(n)$  e

$N_A$  → número de amostras do segmento em análise.



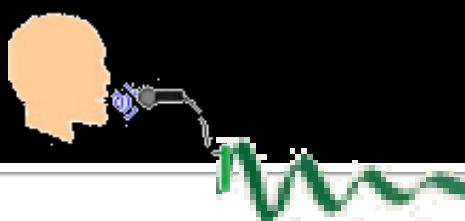
# Energia Segmental

- Considerando-se, ainda, ergodicidade, estacionariedade no sentido amplo e média nula, para o sinal de voz no intervalo citado, a  $E_{seg}$  é definida por:

$$E_{seg} = N_A \cdot E\{[s(n)]^2\} = \sum_{n=0}^{N_A-1} [s(n)]^2 \quad \text{e}$$

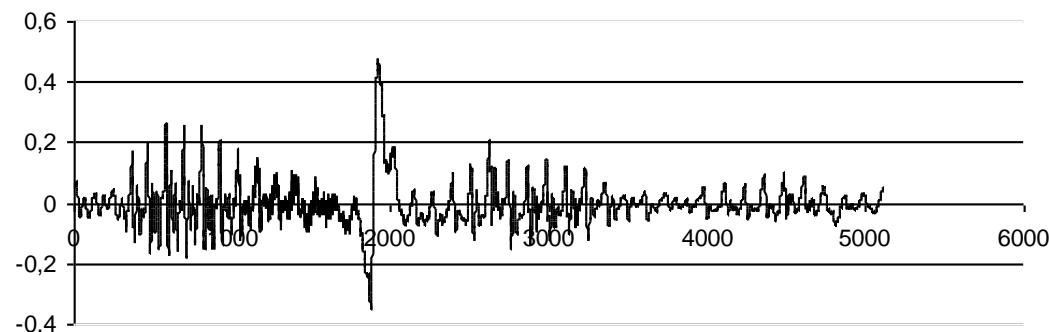
$$E_{seg} (dB) = 10 \log [E_{seg}]$$

- *Parâmetro útil na diferenciação entre segmentos surdos e sonoros do sinal de voz, já que amplitude nos segmentos surdos é bem mais baixa que nos segmentos sonoros.*

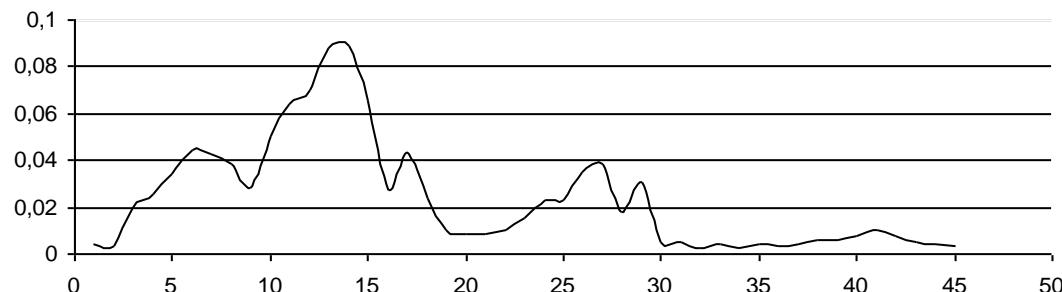


# Energia Segmental

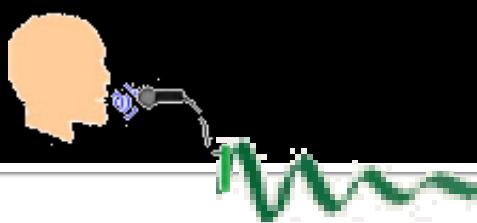
- Exemplo:



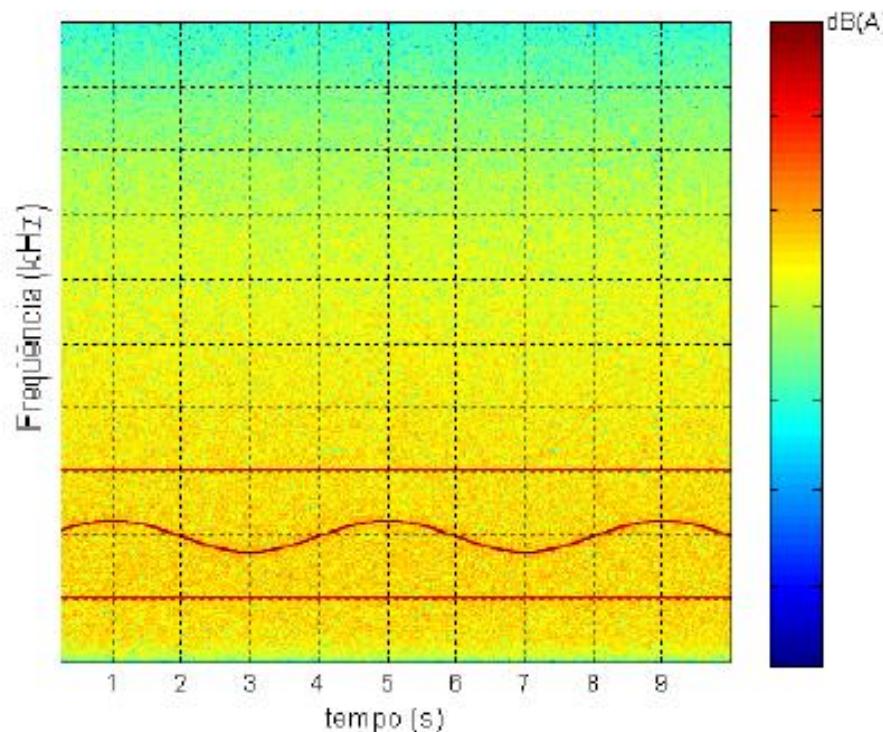
Forma de Onda da Palavra "desliga"



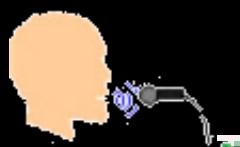
Energia Segmental da Palavra "desliga"



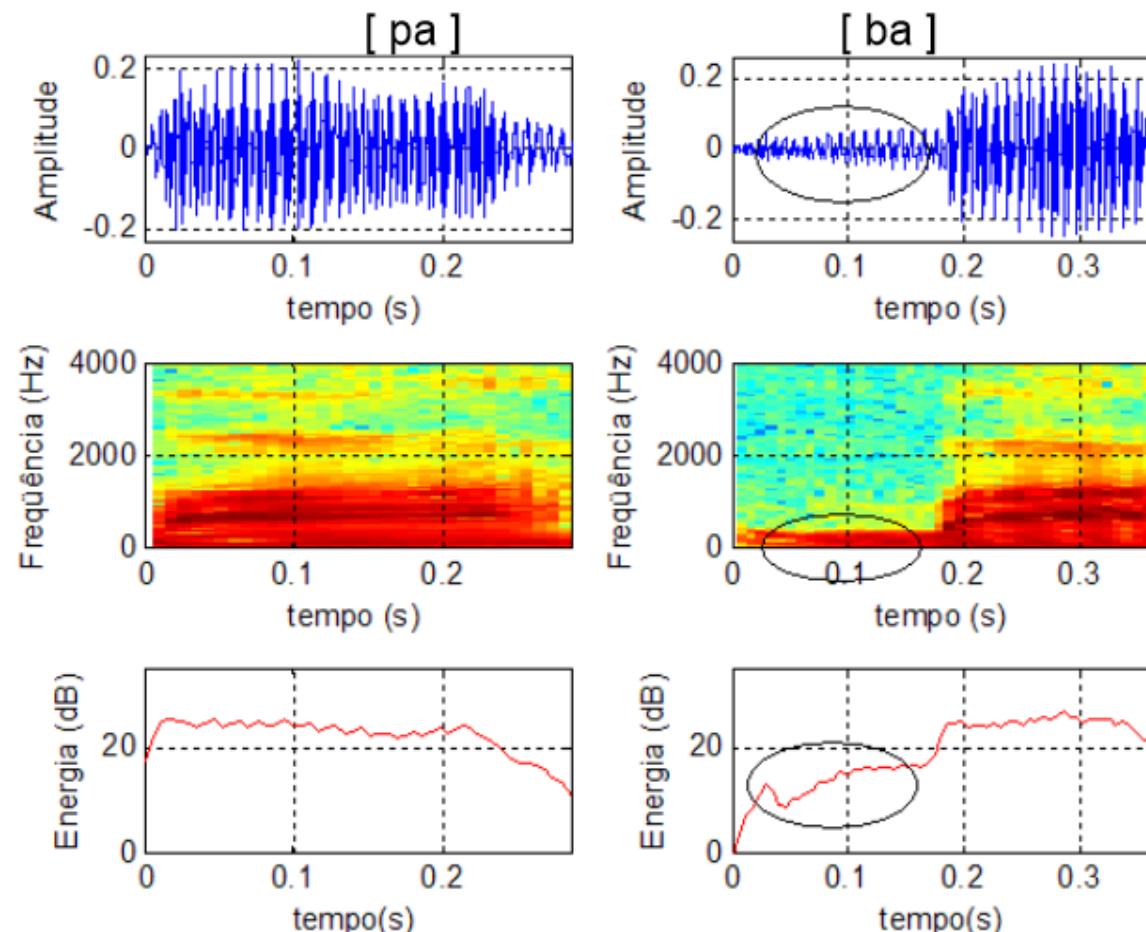
# Espectrograma



Espectrograma A-ponderado para sinais tonais e ruído branco.  
RSR = 0 dB.

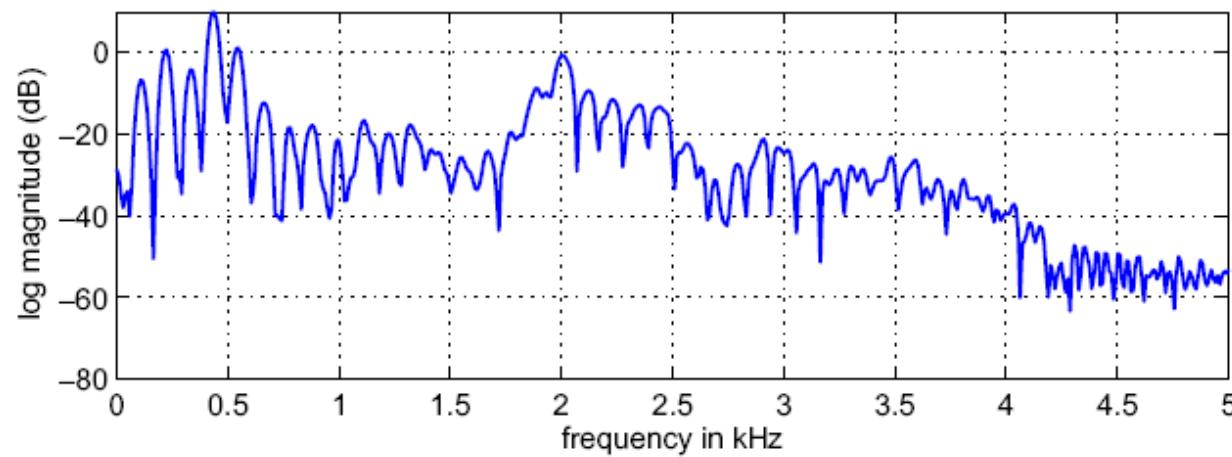
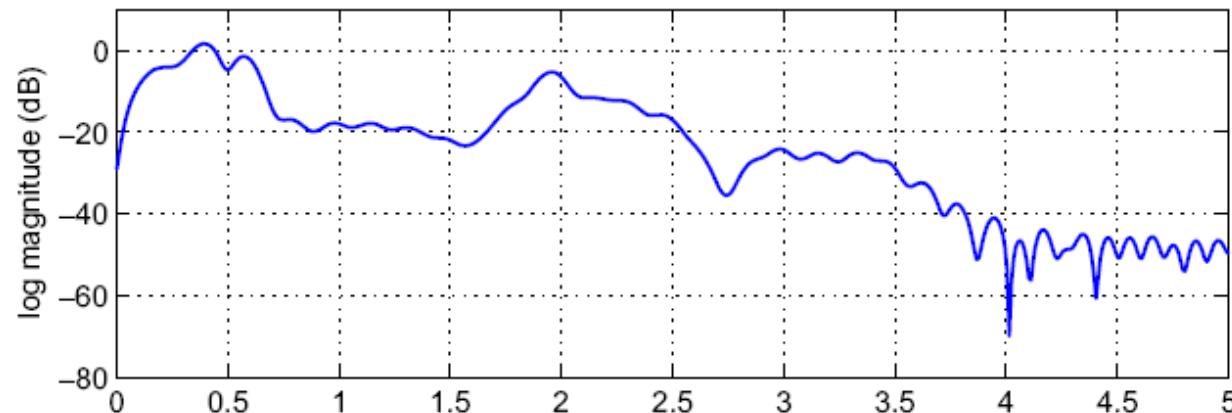
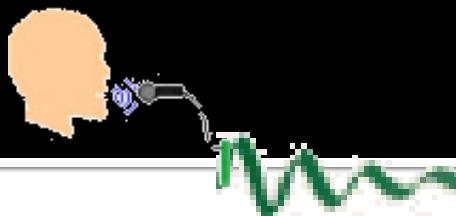


# Espectrograma



forma de onda, espectrograma e registro sonoro para [pa], [ba]

# Espectro

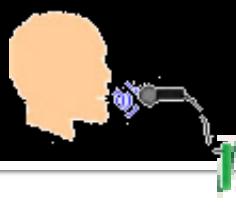




# Taxa de Cruzamentos por Zero -TCZ



- Indica o número de vezes em que as amostras de um sinal mudam de sinal (polaridade), ou sejam, cruzam o zero, limiar tomado como referência.
- A TCZ dá uma ideia do conteúdo de frequências do sinal em análise.
- Altas taxas de TCZ caracterizam os sons surdos, enquanto baixas taxas indicam a presença de sons sonoros.
- Parâmetro bastante eficaz na identificação de consoantes fricativas surdas



# Taxa de Cruzamentos por Zero

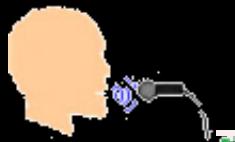
- Esse parâmetro, a curto intervalo de tempo, é definido como (DELLER, PROAKIS & HANSEN, 1993):

$$TCZ(m) = \frac{1}{N_A} \sum_{n=m-N_A+1}^m \frac{|\operatorname{sgn}[s(n)] - \operatorname{sgn}[s(n-1)]|}{2} \cdot w(m-n)$$

onde

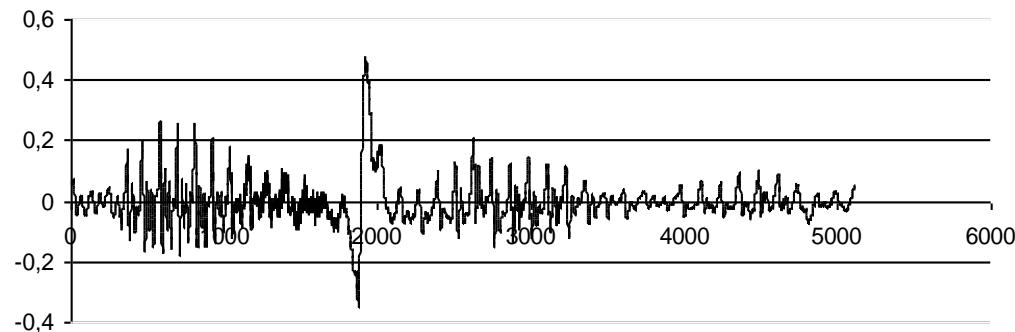
$$\operatorname{sgn}[s(n)] = \begin{cases} +1, & s(n) \geq 0 \\ -1, & s(n) < 0 \end{cases}$$

e  $w(m)$  é a janela, frame ou segmento do sinal em análise e  $N_A$  é o tamanho da janela, ou o número de amostras do segmento.

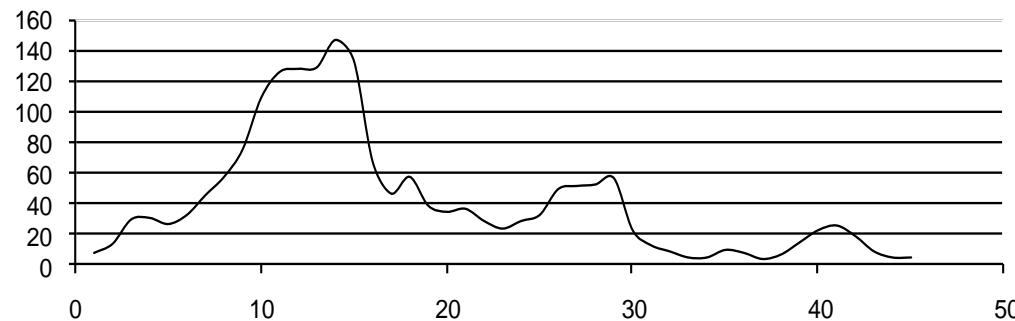


# Taxa de Cruzamentos por Zero

## ■ Exemplo:

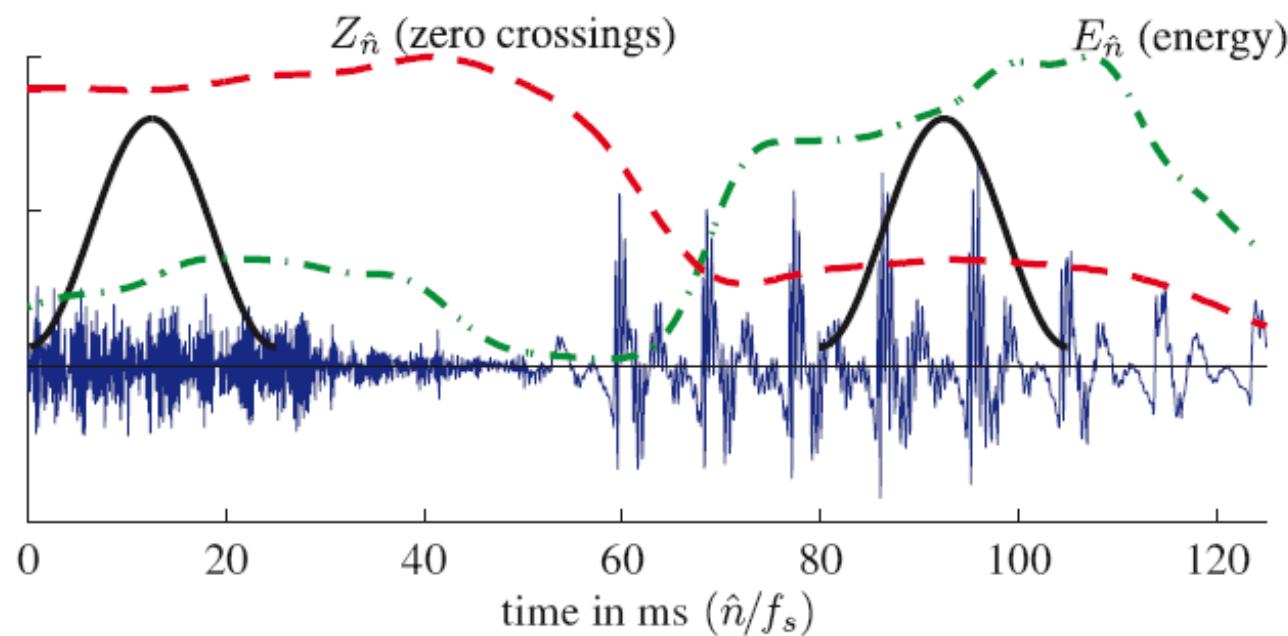
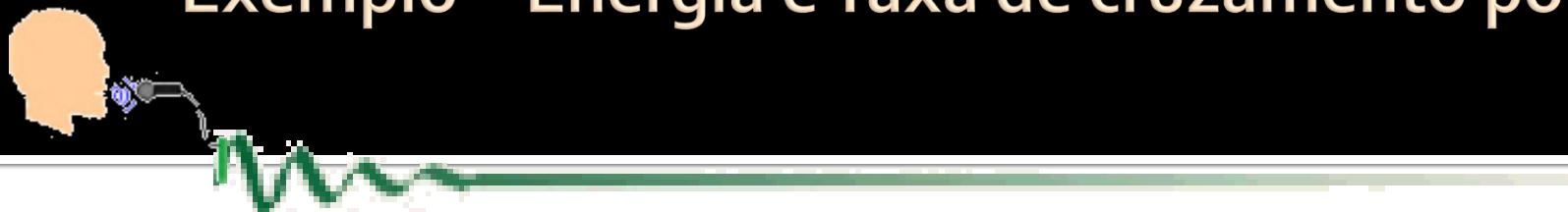


Forma de Onda da Palavra "desliga"

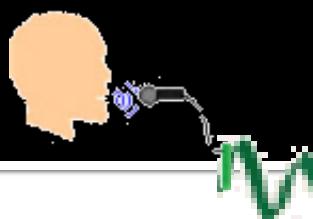


TCZ da Palavra "desliga"

# Exemplo – Energia e Taxa de cruzamento por zeros



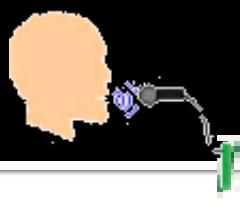
[http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/digital%20speech%20processing%20course/final\\_speech\\_paper\\_1\\_2008.pdf](http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/digital%20speech%20processing%20course/final_speech_paper_1_2008.pdf)



# Função de AutoCorrelação

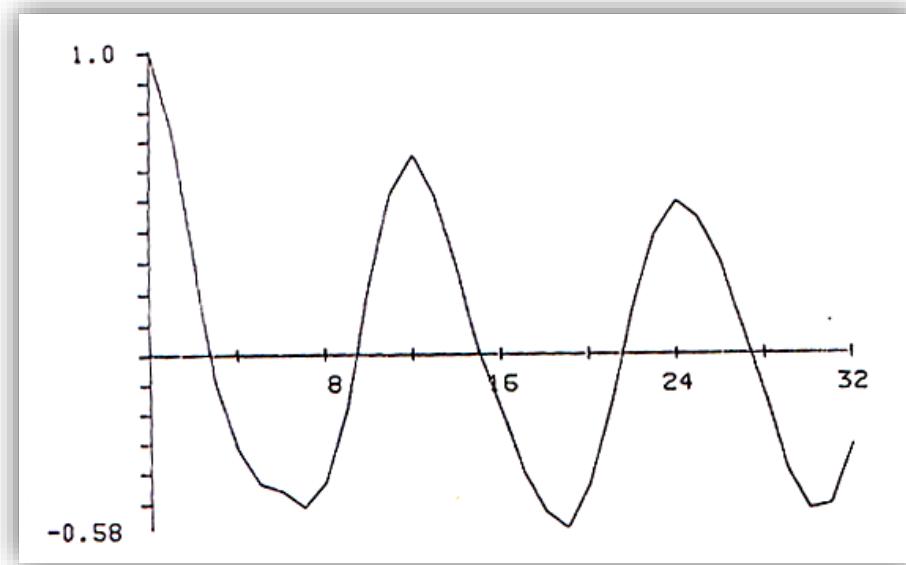
- Uma estimativa da função de autocorrelação de um processo aleatório ergódico pode ser obtida pela estimativa da função de autocorrelação medida no tempo para um segmento longo (porém finito) do sinal.
- A expressão geral que define a função de autocorrelação de um sinal  $x(n)$  é dada por (Rabiner & Schafer, 1978):

$$R_{xx}(k) = [x(n) \cdot x(n+k)] = \frac{1}{N} \sum_{n=0}^{N-k-1} x(n) \cdot x(n+k); \quad k \geq 0$$



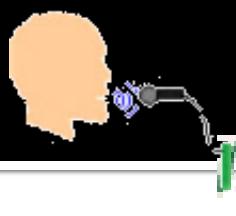
# Função de AutoCorrelação

- Mostra a dependência estatística entre as amostras de um sinal de voz, tomadas em instantes de tempo distintos com relação a um instante de referência.



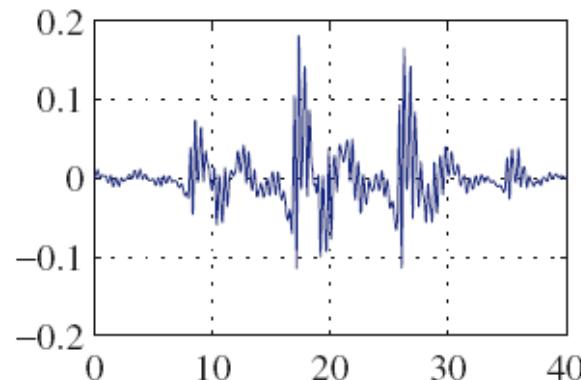
**Função de autocorrelação de um sinal de voz.**

*Obs: também pode ser obtida para segmentos curtos do sinal de voz.*

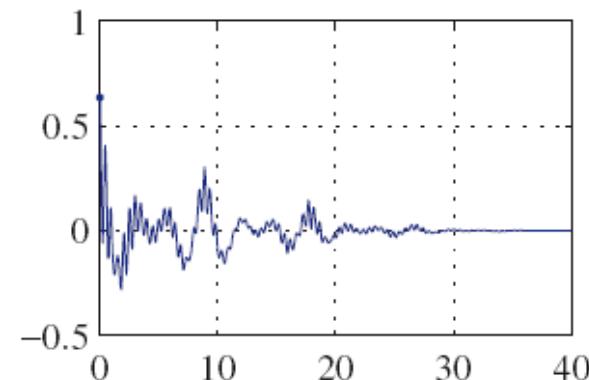


# Autocorrelação

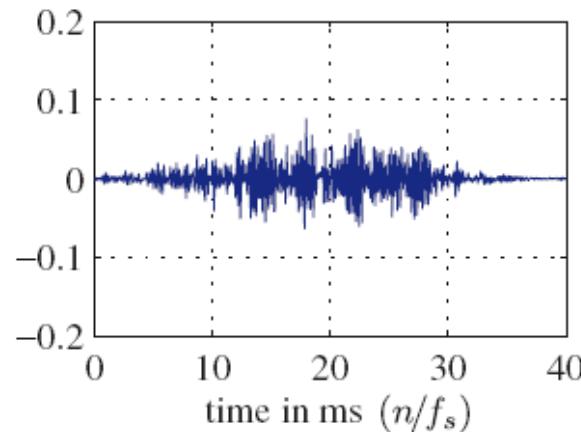
(a) Voiced Segment



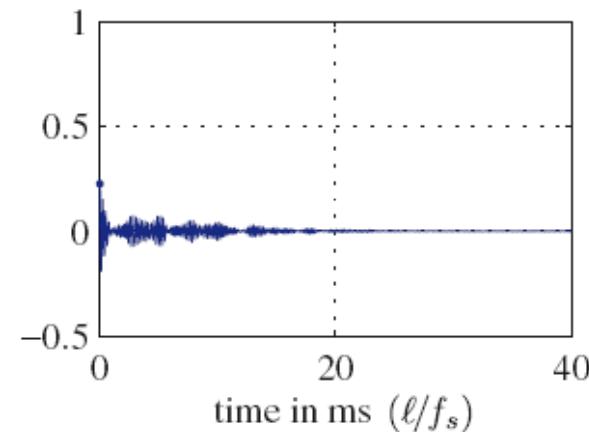
(b) Voiced Autocorrelation



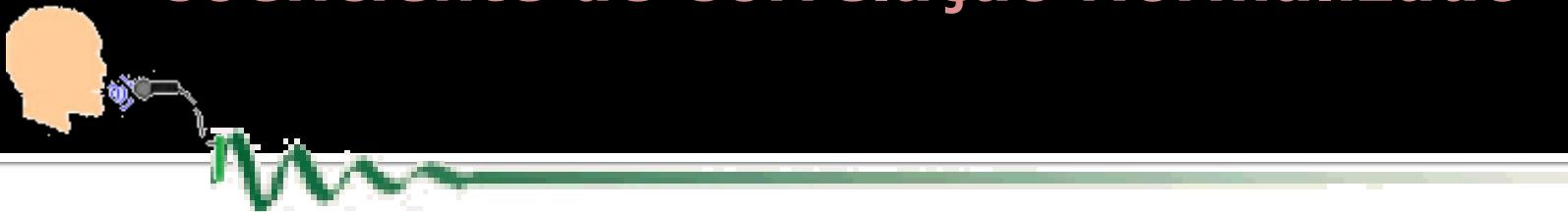
(c) Unvoiced Segment



(d) Unvoiced Autocorrelation

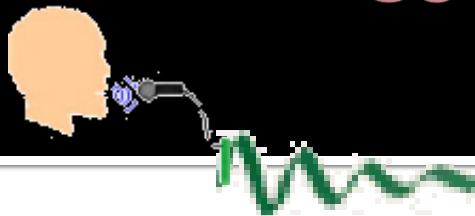


# Coeficiente de Correlação Normalizado



- Bastante útil na classificação dos sons da fala, sendo bastante utilizado na distinção dos sons sonoros e surdos.
- Para sons sonoros → próximos da unidade → alta correlação, pois há uma concentração de energia, nesses sons, nas baixas frequências do espectro.
- Para sons surdos → valores próximos de zero.
- Para intervalos de silêncio → valores variam com o ambiente (entre os valores obtidos para os sons sonoros e surdos).

# Coeficiente de Correlação Normalizado



- Utilizando apenas o primeiro coeficiente de correlação  $\rho_1$  é possível obter as informações necessárias para o auxílio na identificação dos diversos sons da fala.

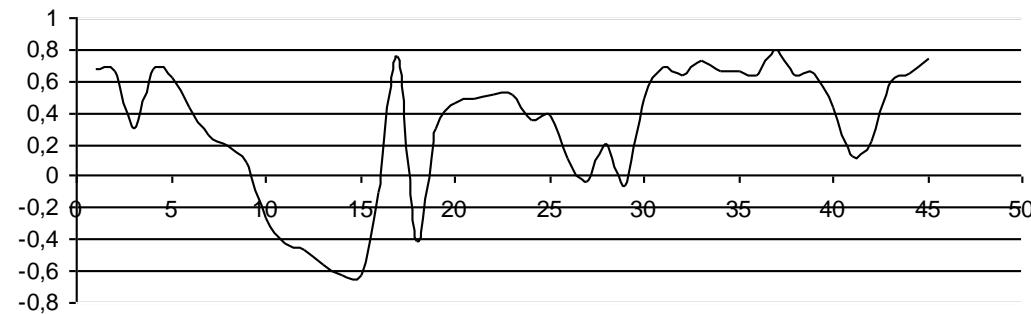
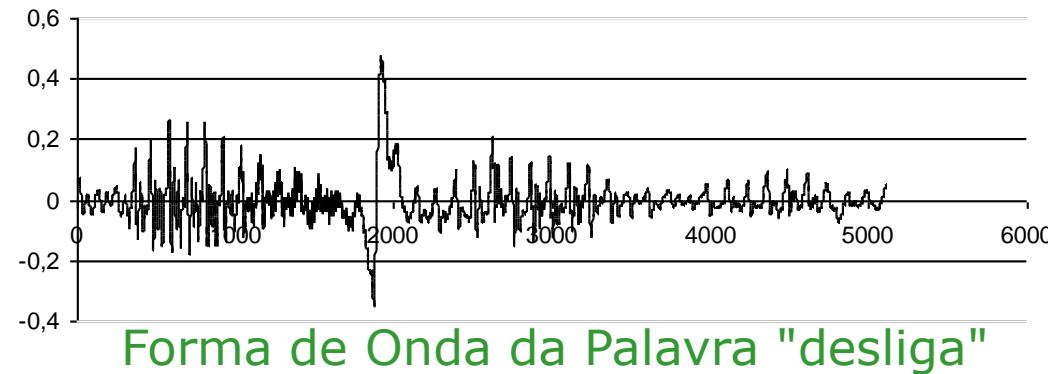
$$\rho_1 = \frac{\sum_{n=1}^{N_A} [s(n) \cdot s(n-1)]}{\sqrt{[\sum_{n=1}^{N_A} s^2(n)][\sum_{n=0}^{N_A-1} s^2(n)]}}$$

Em que:  $s(n)$  → sinal de voz e  
 $N_A$  → Tamanho do segmento em processamento.



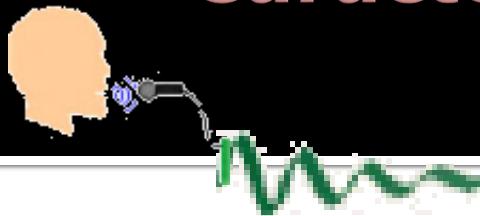
# Coeficiente de Correlação Normalizado

- Exemplo:



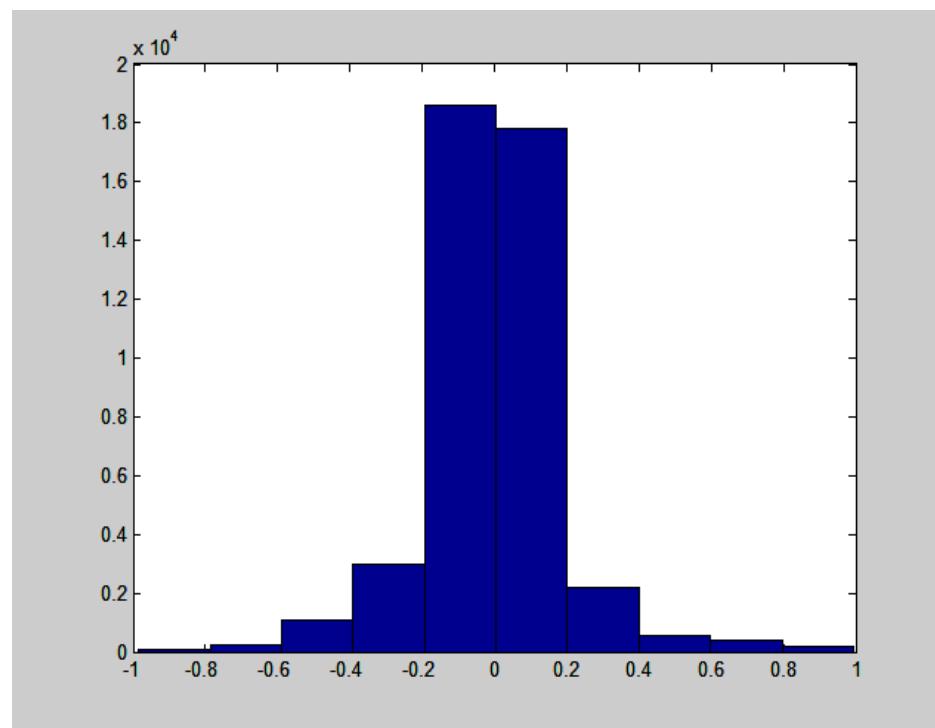
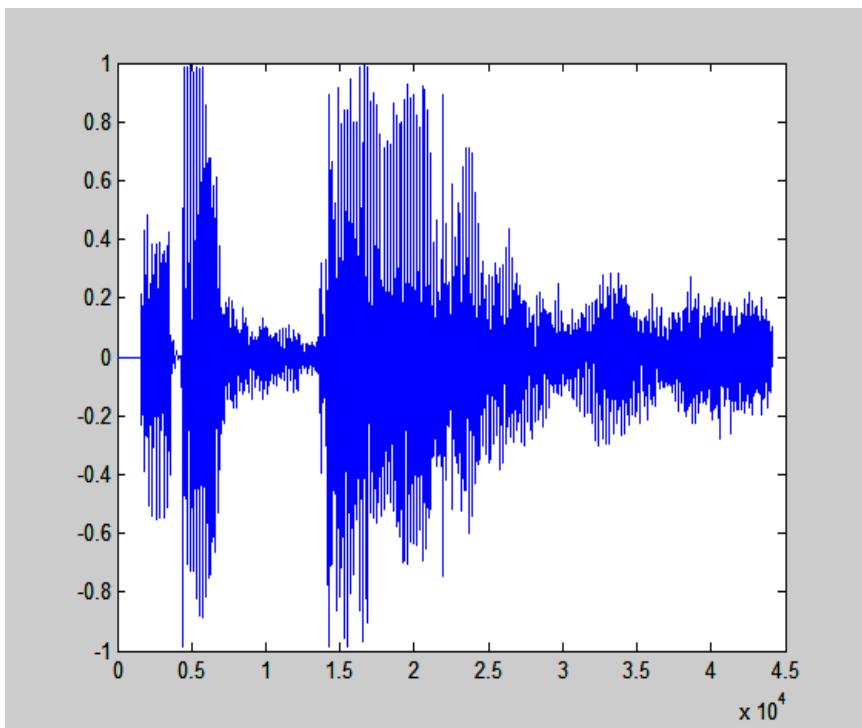
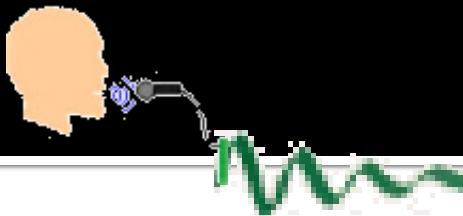
Coeficiente de Correlação Normalizado da Palavra "desliga"

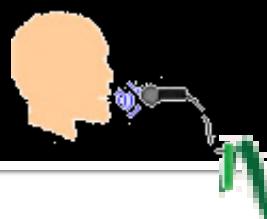
# Características de Longa Duração do Sinal de Voz



- Função densidade de probabilidade (f.d.p.)
  - A f.d.p. especifica não somente os valores médios do sinal, como também indica outras características importantes do sinal de voz, tais como faixa de amplitude e valor de pico.
  - A função densidade de amplitudes para sinais de voz depende da largura de faixa utilizada e das condições de gravação.
  - Uma primeira aproximação é o modelo “*two-sided*” exponencial ou Laplaciano. Entretanto, a f.d.p. Gamma é uma aproximação ainda melhor nesse caso.

# Histogramma





# fdp de um sinal de voz

Gaussiana

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

Laplaciana

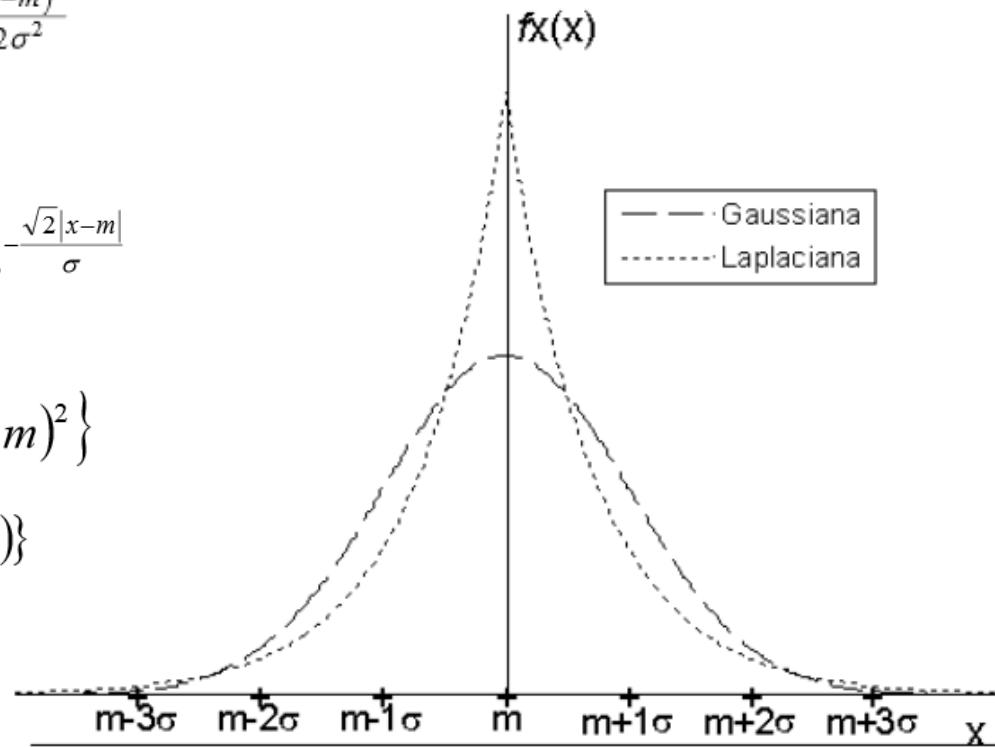
$$f_x(x) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{\sqrt{2}|x-m|}{\sigma}}$$

sendo

$$\sigma^2 = E\{(x - m)^2\}$$

e

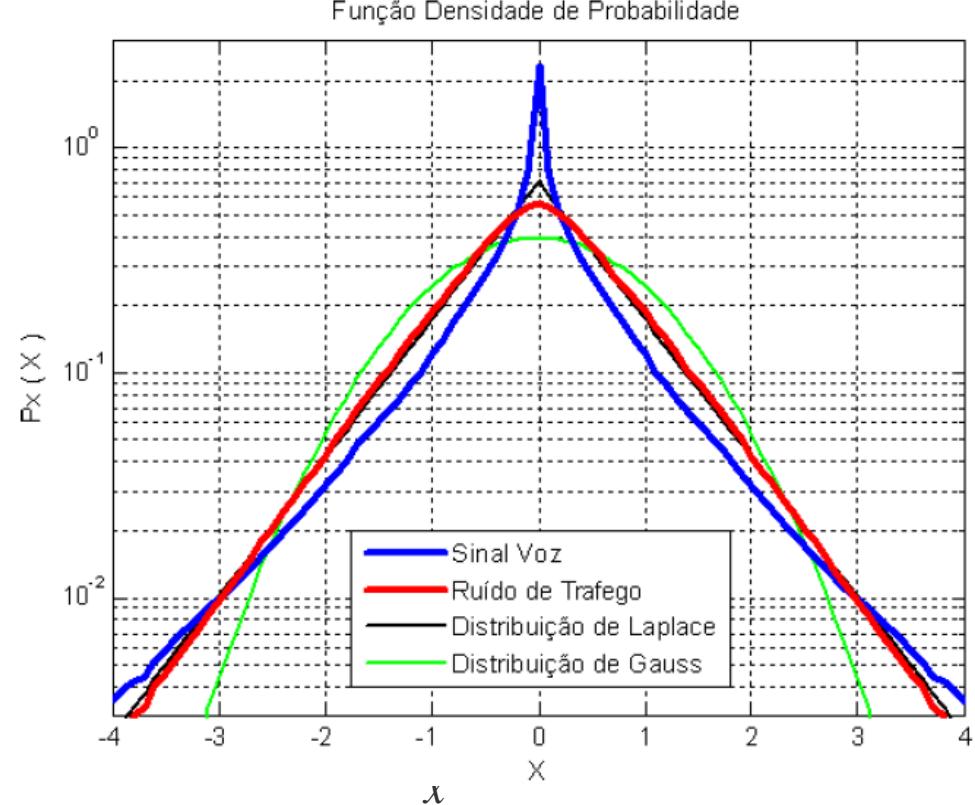
$$m = E\{(x)\}$$



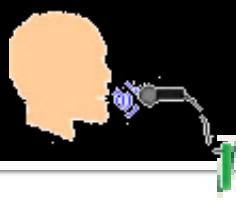
# Função densidade de probabilidade (f.d.p.)



Para sinais de voz de curta duração, a melhor aproximação é a f.d.p. Gaussiana, independentemente se os segmentos são de alta ou baixa energia.



Comparação da f.d.p. para um sinal de voz com valores teóricos de f.d.p.'s Gamma e Laplace.



# Densidade espectral de potência

- A densidade espectral de potência de sinais de voz pode ser estimada pela Transformada de Fourier da função de autocorrelação

$$S_{xx}(\Omega) = \Im\{R_{xx}(k)\}; \quad k \rightarrow \infty$$

- O espectro de sinais de voz pode ser obtido pela média ao longo de L segmentos do sinal, a partir das estimativas da densidade espectral de potência de curta duração, com periodogramas de N componentes, ou seja, para um l-ésimo segmento:

$$S_{xx}(\Omega, l) = \frac{1}{N} |X(j\Omega, l)|^2$$



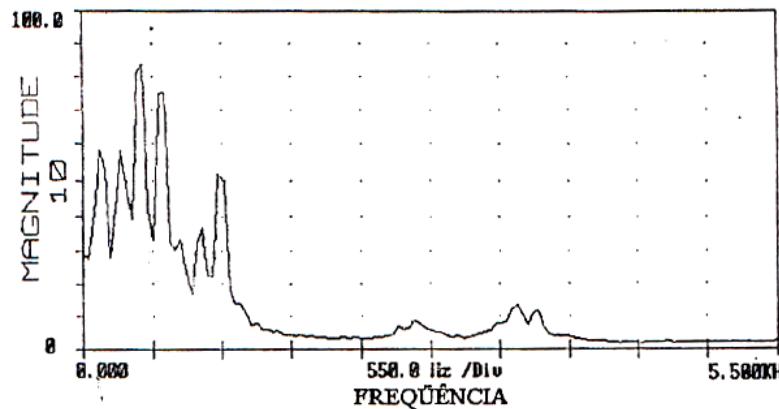
# Densidade espectral de potência

onde

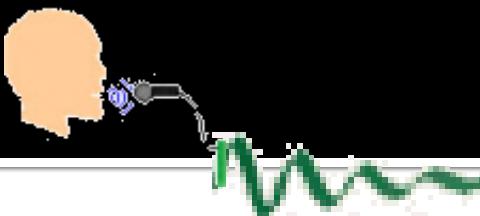
$$X(j\Omega, l) = \Im\{S(n, l)\} = \sum_{n=0}^{N-1} S(n, l) e^{-jn\Omega}$$

cuja média ao longo de L segmentos é dada por:

$$S_{ss}(\Omega) = \frac{1}{L} \sum_{l=1}^L S_{ss}(\Omega, l)$$



**Densidade Espectral de Potência da palavra aplausos.**



# Referências

Lawrence R. Rabiner and Ronald W. Schafer. Introduction to Digital Speech Processing. Foundations and Trends R in Signal Processing., Vol. 1, Nos. 1–2 (2007) 1–194. DOI: [10.1561/2000000001](https://doi.org/10.1561/2000000001).

Abraham Alcaim e Carlos Alexandre dos Santos Oliveira. Fundamentos do Processamento de sinais de voz e imagem. Editora PUCRio – 2011.