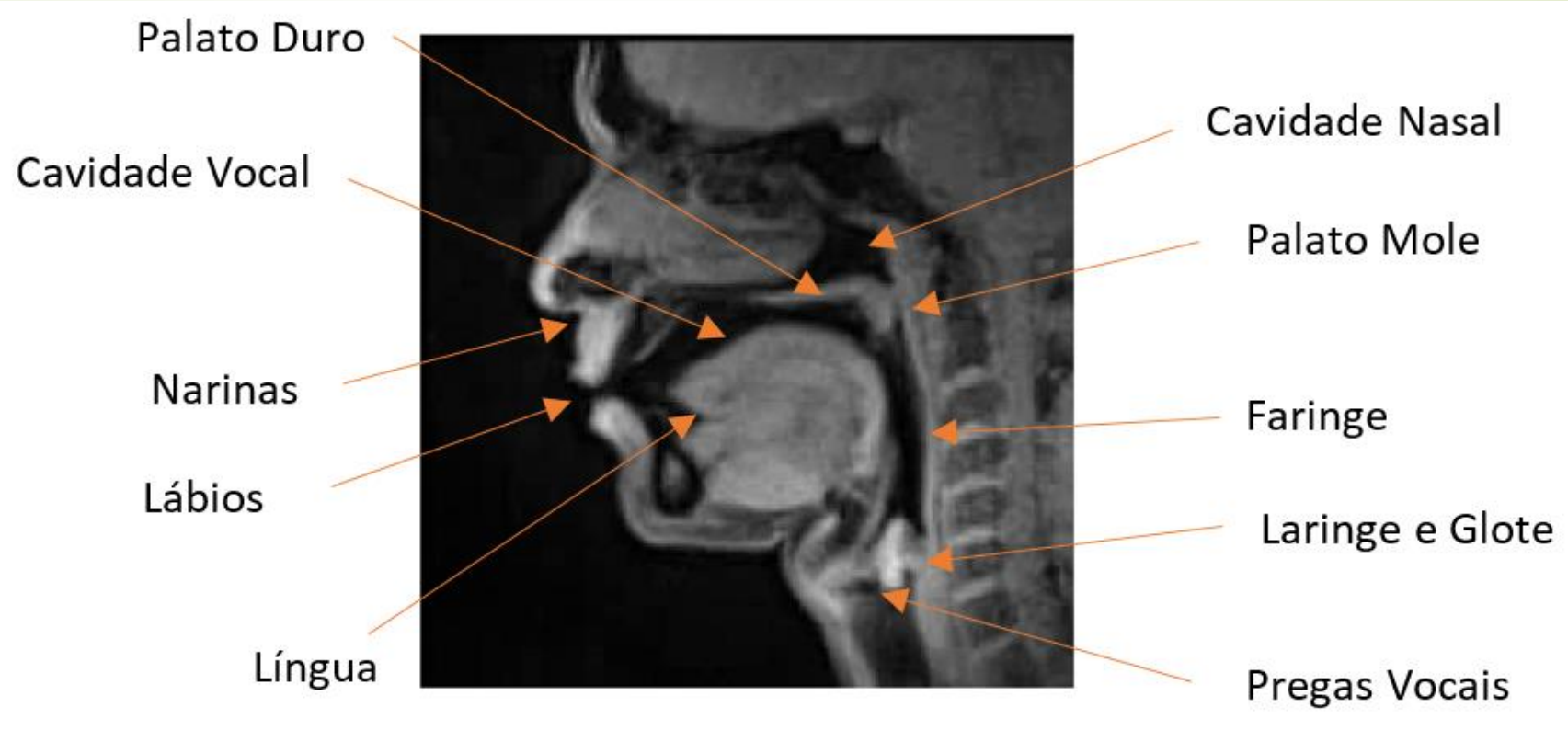
Aparelho Fonador Humano e Processamento Digital do Sinal de Voz

ESTI019 – Comunicações Multimídia

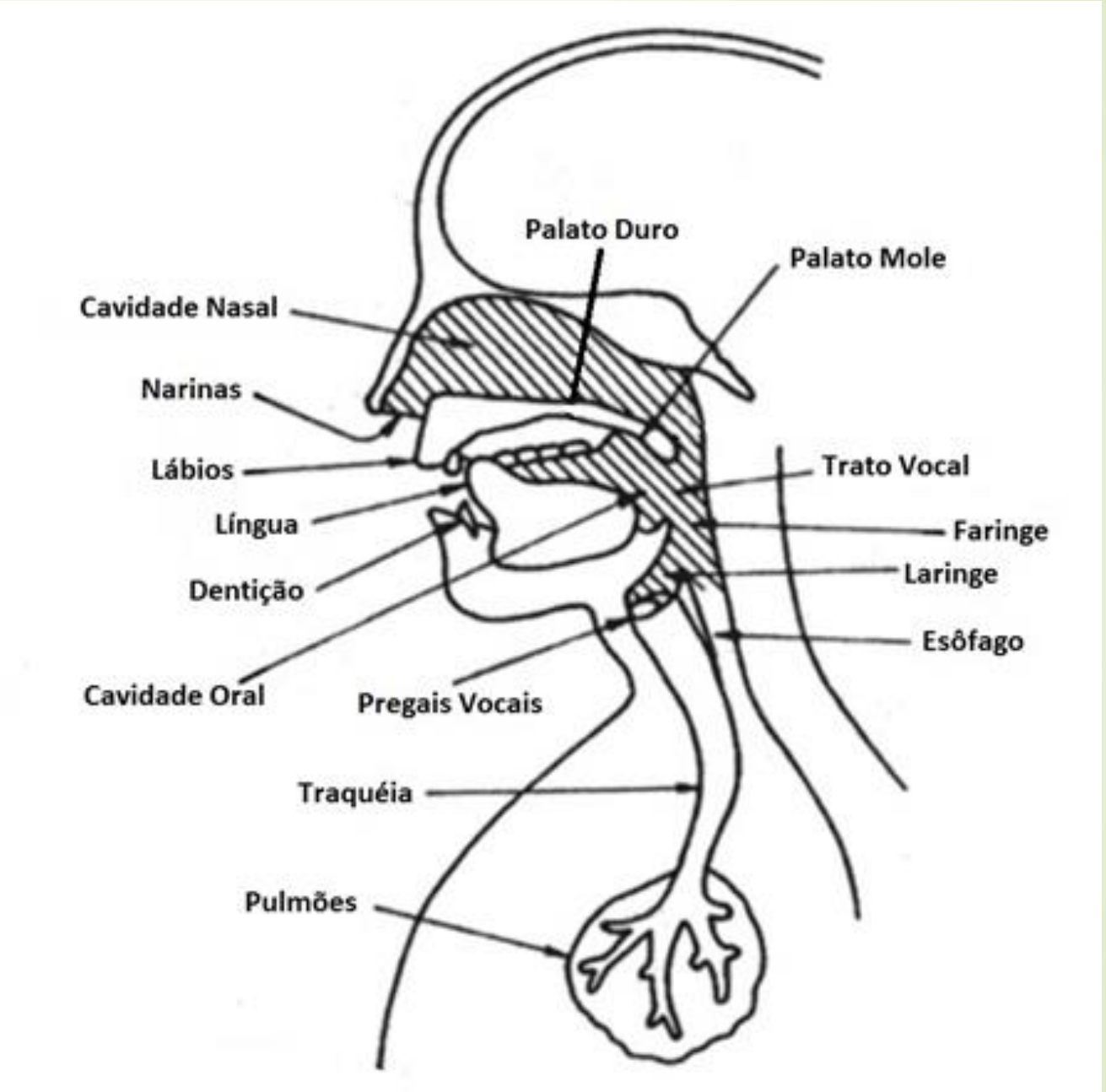
Profs. Celso S. Kurashima, Kenji Nose e Mário Minami

UFABC

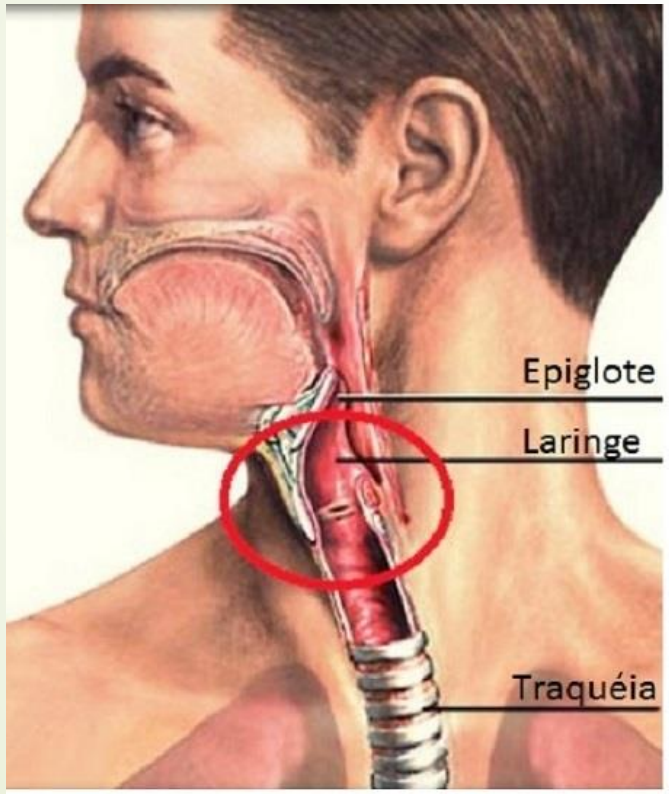
Raio-X do Aparelho Fonador Humano (AFH)

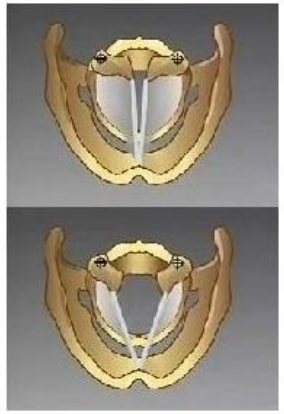


Elementos Anatômicos AFH



Laringe e Pregas Vocais





Articulações entre cartilagens Aritenóide e Cricóide

Tensão variável pelas articulações entre cartilagens Tireóide e Cricóide

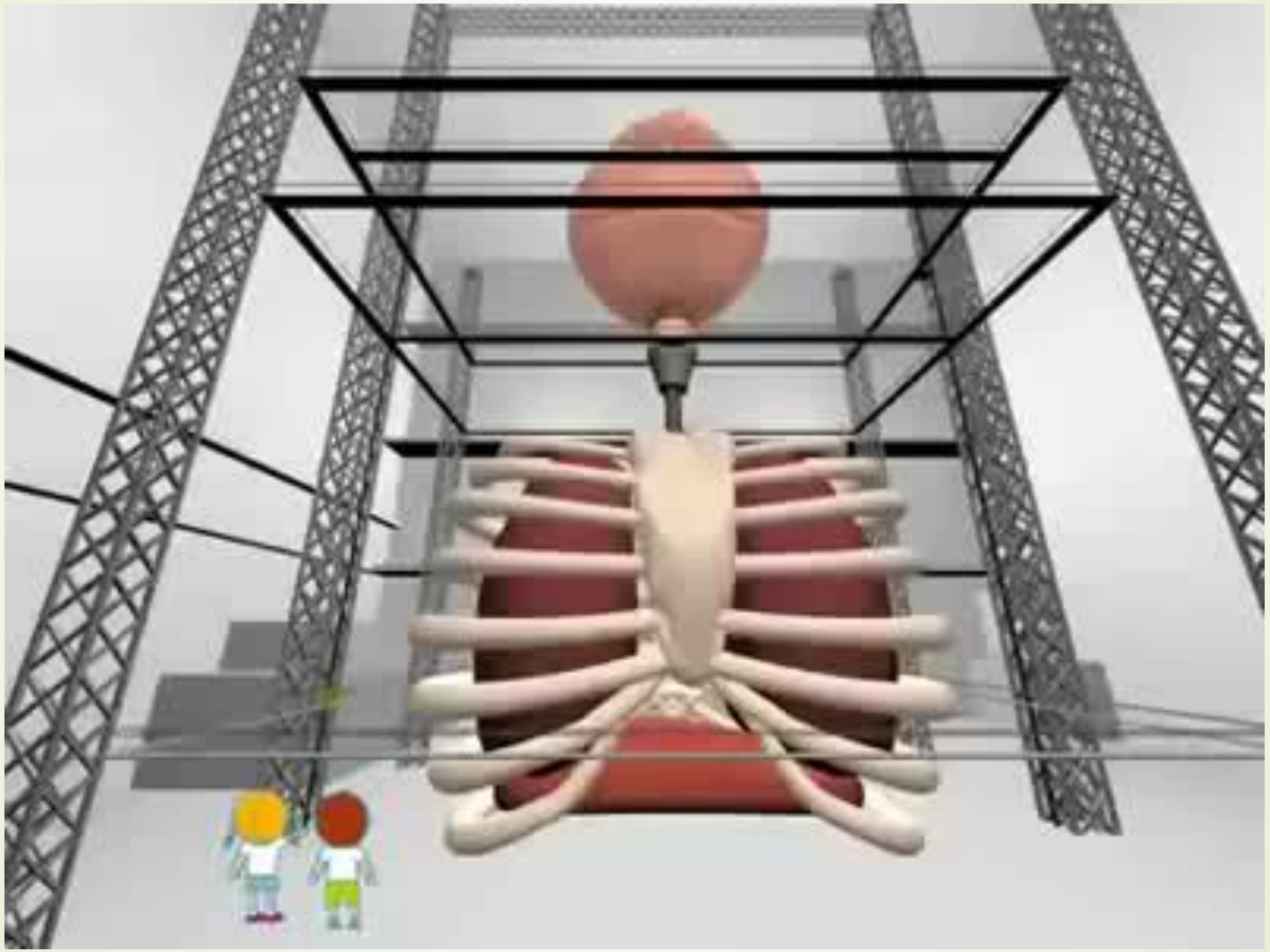
Cordas (Pregas) Vocais

- Laringe e Cordas Vocais

https://www.youtube.com/watch?v=0H5WKQ--q4c

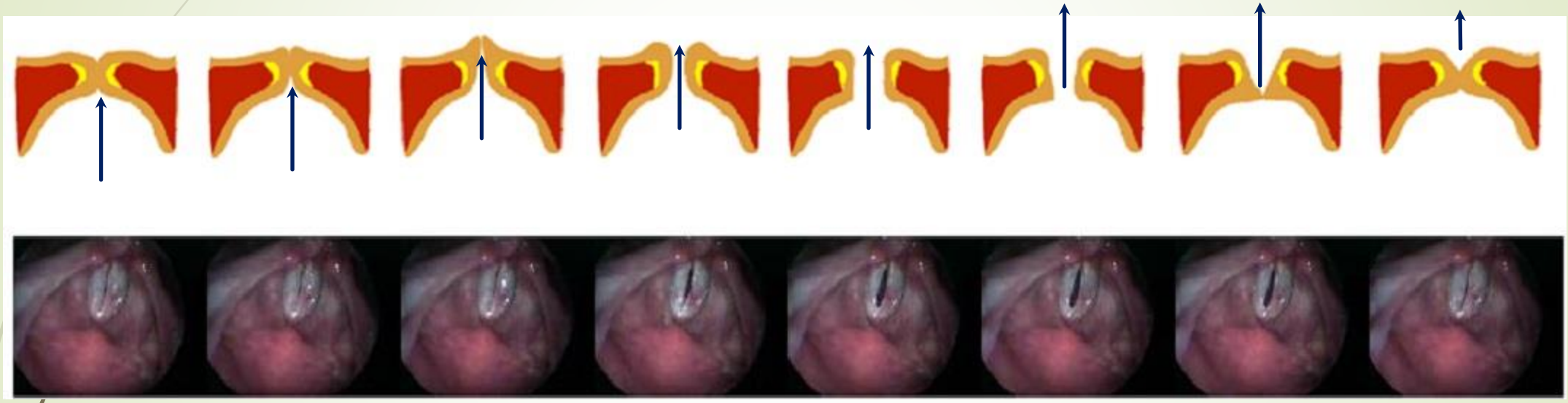
Vários são os músculos que atuam nas Cordas (Pregas) Vocais.

Fábrica da Voz: Sistema Completo



https://www.youtube.com/watch?v=SfhGbCjHA3w

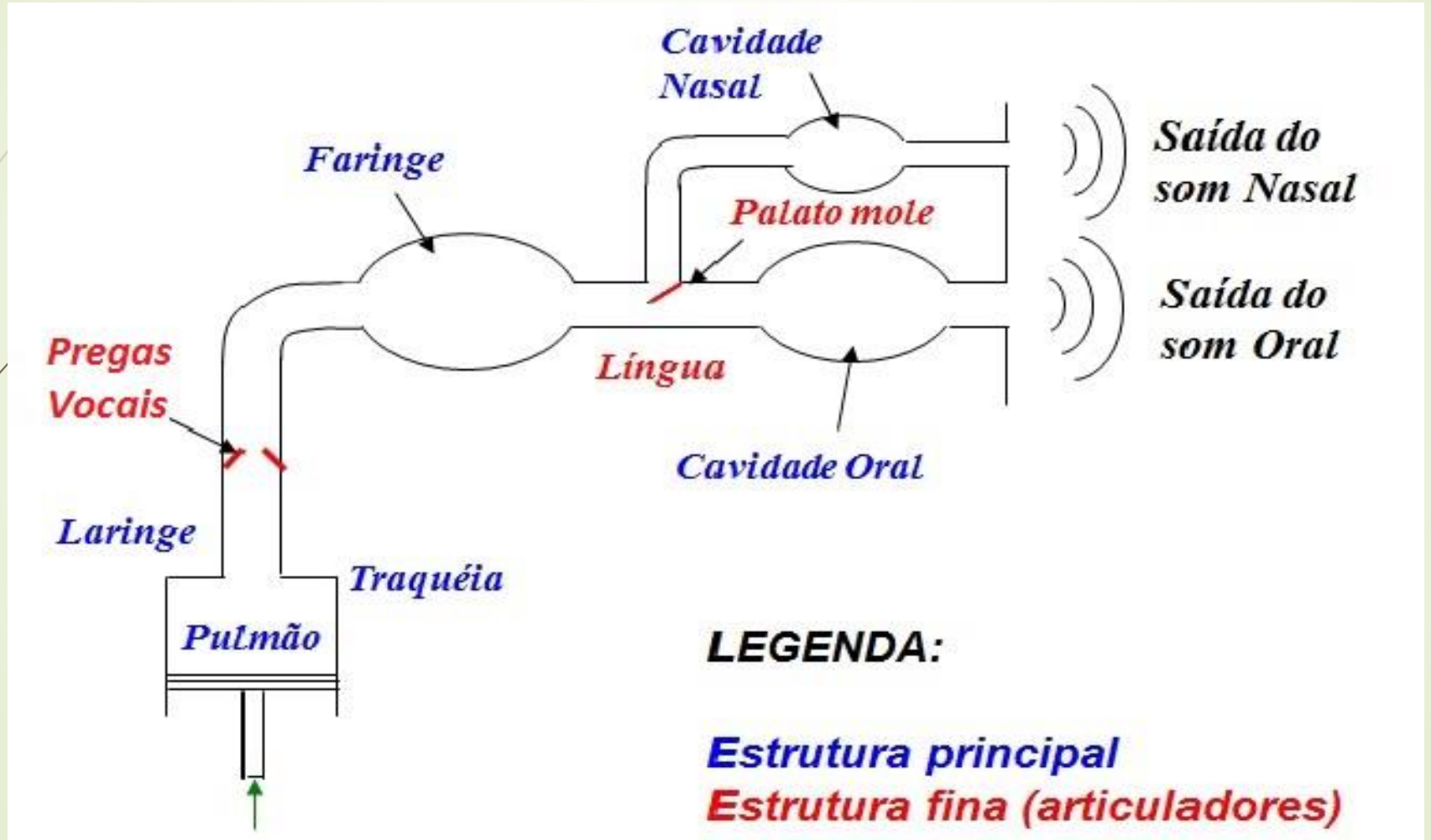
Ciclo de Vibração das Pregas Vocais



Adução
Fechamento

Abdução
Abertura

Estruturas e Articuladores Trato Vocal e Nasal

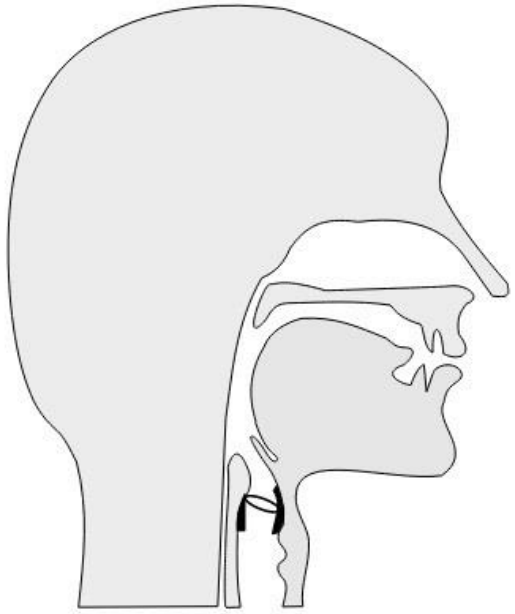


Comentários sobre a Estrutura do Trato Vocal

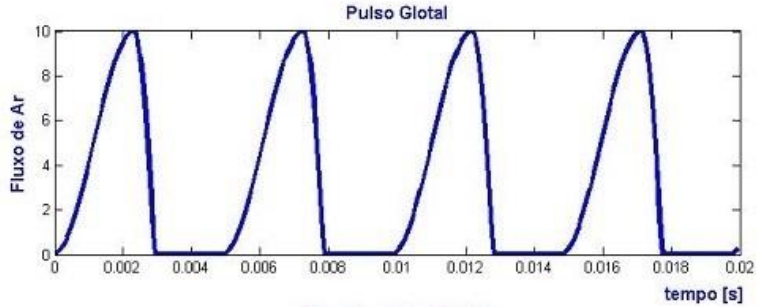
- cavidades → filtro acústico → estrutura ressonante da voz humana
- homem adulto → tamanho do trato vocal, em média, é de 17 cm
  - mulher adulta, cerca de 14 cm; na criança em idade escolar, 10 cm
- Com os articuladores, a seção transversal varia de 0 a 20 cm^2
- Comprimento médio do trato nasal num homem adulto é de 12 cm
  - Utilização controlada através do palato mole
- A faixa de abertura do palato mole num homem adulto vai de 0 a 5 cm^2
  - sons nasais (palato mole aberto) ou
  - sons não nasais (palato mole fechado).
- ***A excitação para o filtro acústico*** será periódica ou não, dependendo do que ocorrer na laringe, pela atuação das cordas vocais
  - cordas vocais vibrando na passagem do ar temos a excitação (periódica) dos chamados sons ***sonoros*** (“**voiced’s**”)
  - quando as cordas vocais não vibram a excitação é dita ***surda*** (“**unvoiced**”).

Modelo Fonte-Filtro

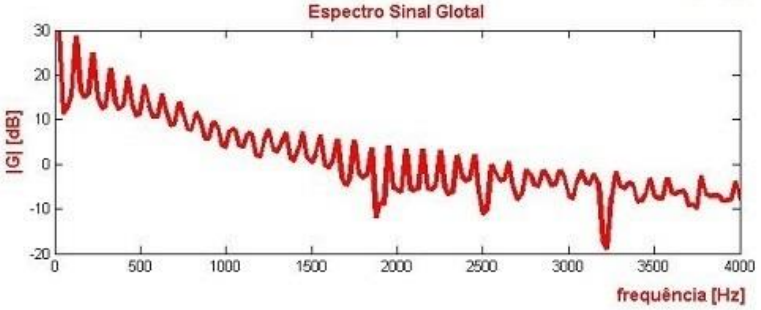
Ressonâncias Filtro Acústico



Especto da Voz



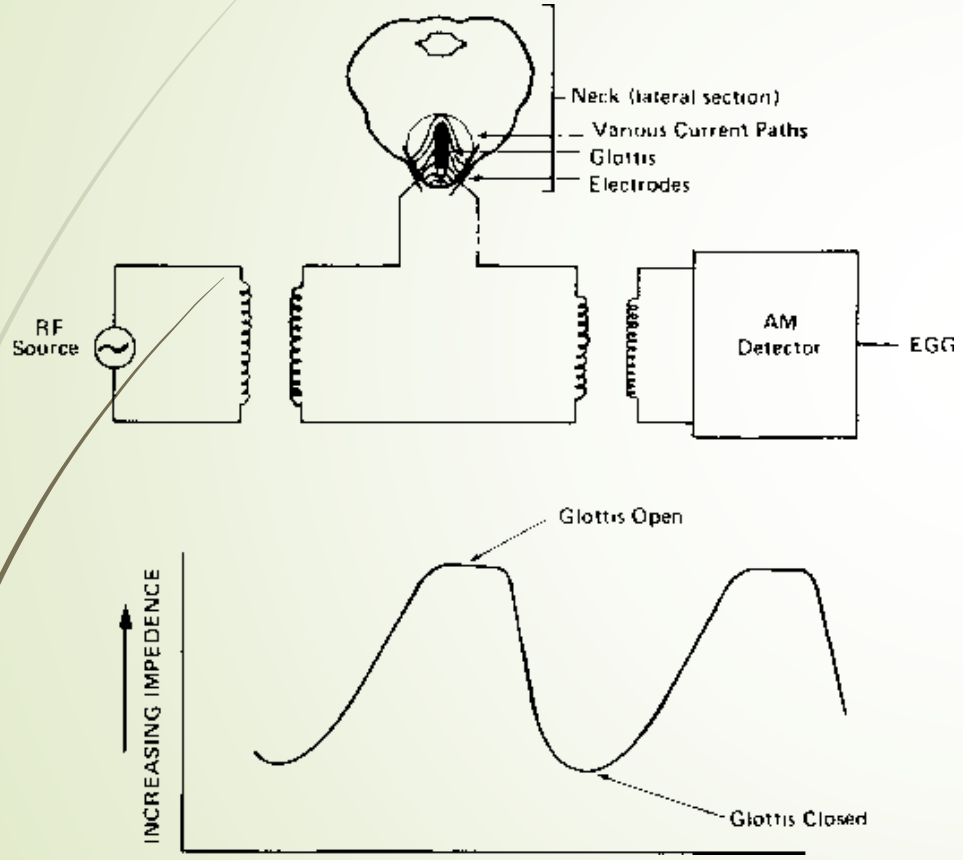
Pulsos Glotais



Espectro

Eletroglotografia (EGG)

É uma técnica não invasiva usada para registrar comportamento na laringe indiretamente pela medição da impedância eléctrica na garganta durante a fonação



Um sinal de alta frequência (300kHz a 5Mz) com minima potência (µW) e baixa corrente (≈ 1mA) e baixa tensão (<0.5V), para segurança fisiológica, atravessa dois eletrodos,

instalados na superfície da garganta próximo da cartilagem tireóide.

Electroglottography, in: https://www2.ims.uni-stuttgart.de/EGG/frmst2.htm

Análise Temporal da Voz

- Segmentos
- Janelas: Processamento de Tempo Curto
- Energia de Tempo Curto
- Pitch, T_0
- Frequência Fundamental f_0

Segmentos (Quadros)

- Divisão do sinal em trechos “quase-estacionários”
- freqüência de amostragem f_a, segue o teorema de Nyquist (dobro da frequência máxima do sinal)
- duração do segmento, T_s , é fixa:

  Maior que uma transição da articulação

  Menor que a duração de uma vogal “rápida”

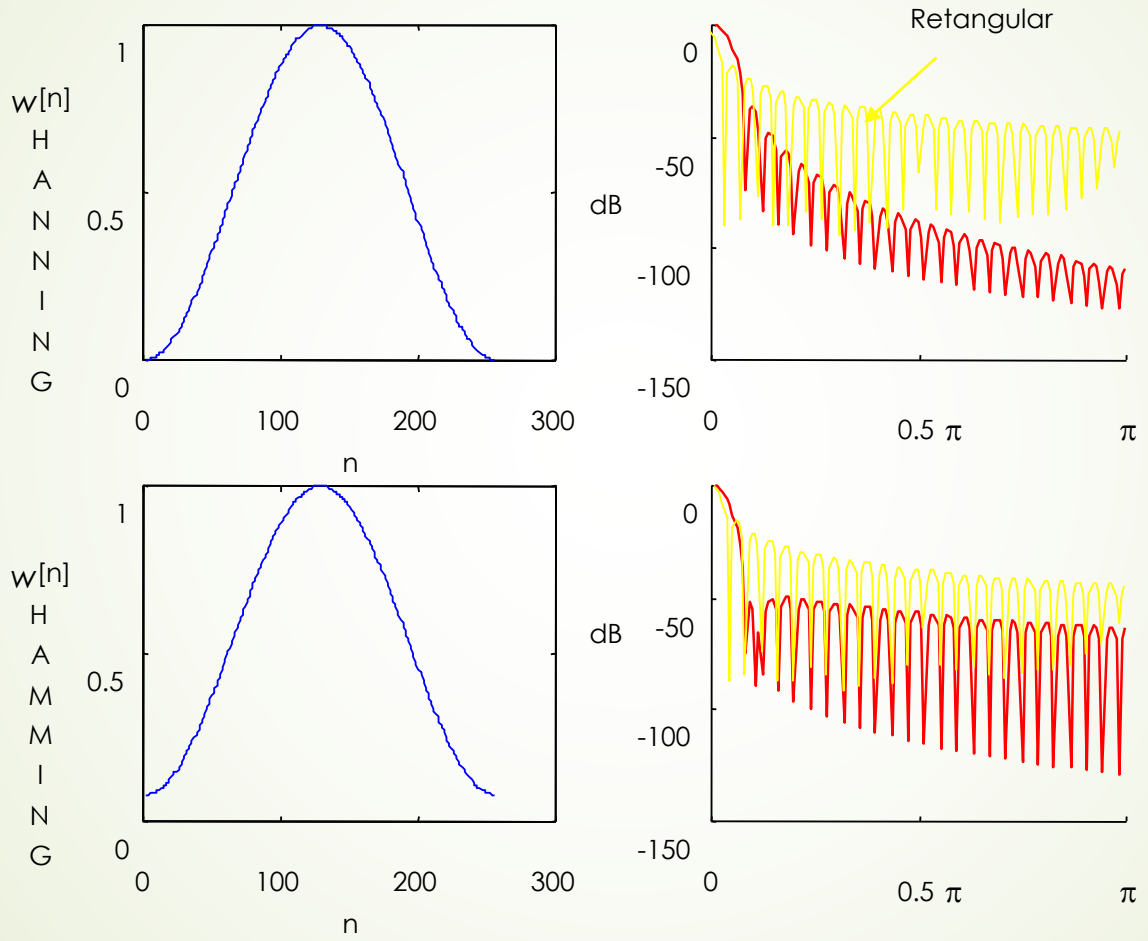
- *Tamanho da Janela*, *ou* número de amostras por segmento, N_J, é dado por:

$$N_J = f_a T_s = \frac{T_s}{T_a}$$

Pois, $f_a = \frac{1}{T_a}$

onde T_a é o período de amostragem.

Janelas e Espectro das Janelas



Energia de Curto Prazo

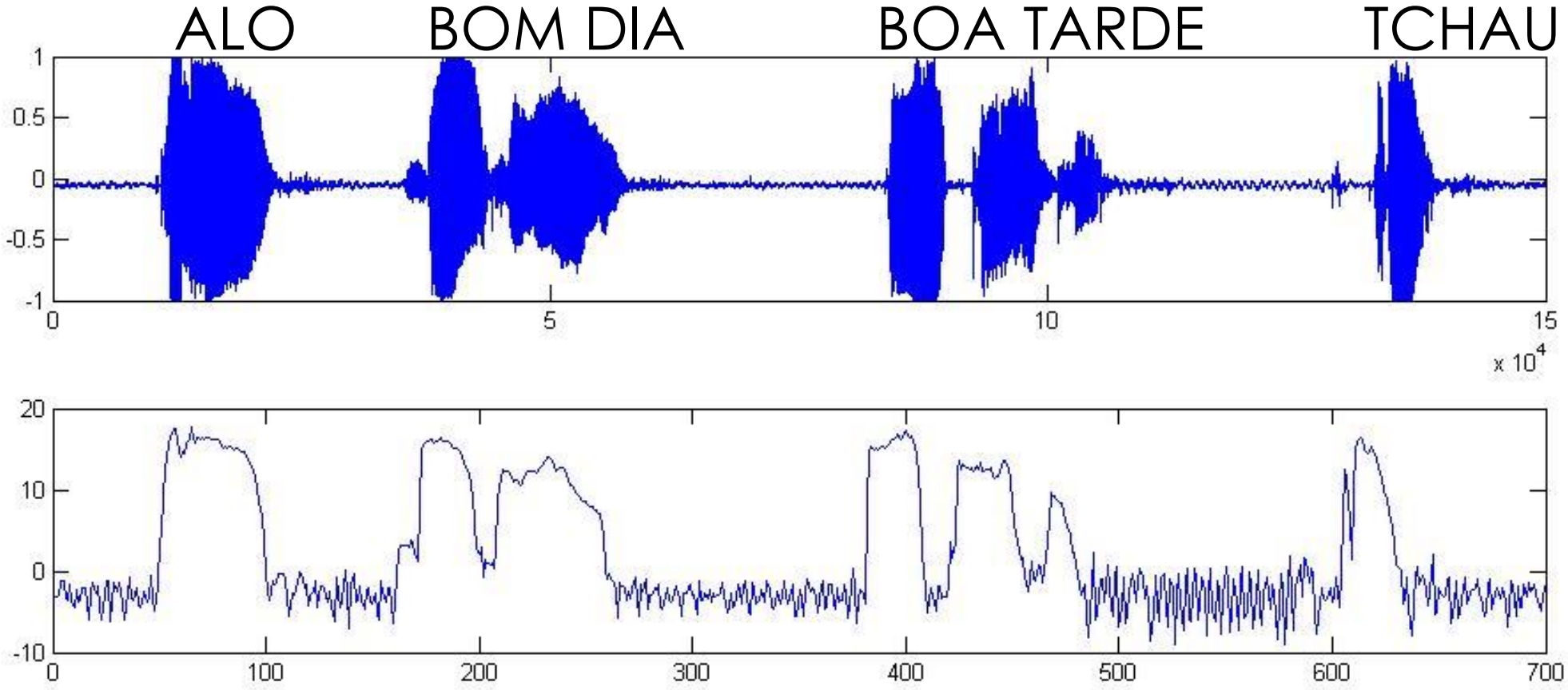
Num l-ésimo Segmento, de tamanho N_J :

$$E(l) = \frac{1}{N_J} \sum_{m=0}^{N_J - 1} \left(x_l(m) \right)^2 \qquad 0 \leq l \leq N_t$$

$$E_{dB}(l) = 10 \log E(l)$$

Sendo N_t o número total de segmentos do sinal
N_J o tamanho de cada segmento.

Contorno de Energia



Sons Vocálicos e Consonantais

- ***Sons vocálicos***, quando o fluxo de ar praticamente não sofre restrições à sua passagem pelo trato vocal. Sons Vocálicos possuem ***maior Energia***.
- ***Sons consonantais***, quando as restrições (constrições) são significativas, diminuindo assim significativamente sua amplitude, e assim possuem ***menor Energia***.

Densidade Espectral de Potência

$$S_i(f) = 10\log\left|\mathbf{S}_i[f]\right|^2$$

$$E(f) = 10\log_{10}\left|\mathbf{E}[f]\right|^2$$

$$I(f) = 10\log_{10}\left|\mathbf{I}[f]\right|^2$$

$$S_f(f) = 10\log_{10}\left|\mathbf{S}_f[f]\right|^2$$

Transformada Discreta de Fourier (TDF)

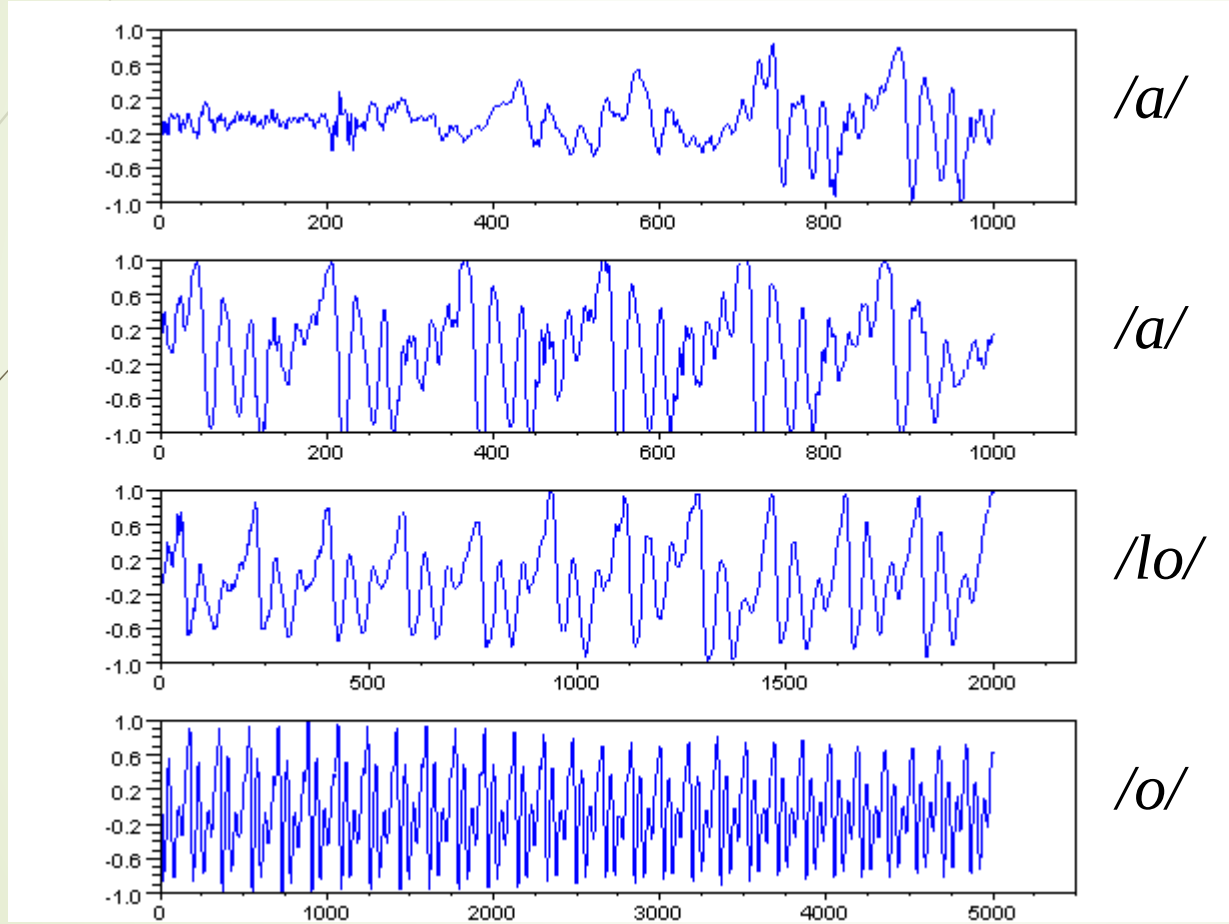
$$\mathbf{S}_i[f] \xleftrightarrow{TDF} s_i(t)$$

$$\mathbf{E}[f] \xleftrightarrow{TDF} e(t)$$

$$\mathbf{I}[f] \xleftrightarrow{TDF} i(t)$$

$$\mathbf{S}_f[f] \xleftrightarrow{TDF} s_f(t)$$

Período Fundamental T_0, Pitch, Frequência Fundamental f_0



$$f_0 = \frac{1}{T_0}$$

Palavra “ALÔ”

f_0, T_0 e Pitch

- O tempo decorrido entre duas aberturas sucessivas das cordas vocais é chamado de *período fundamental* T_0 e a freqüência de vibração é chamada de freqüência fundamental f_0 da fonação:

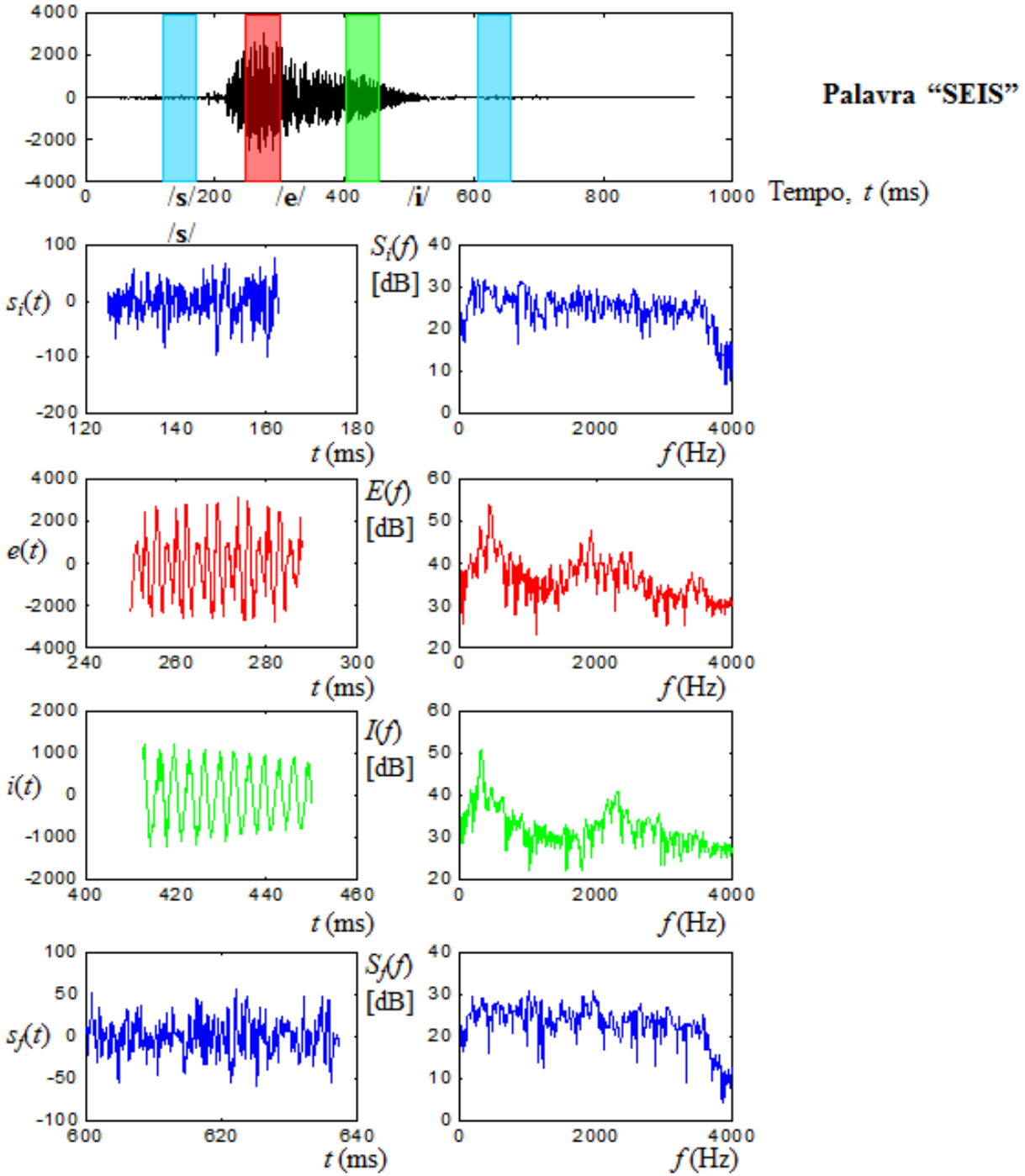
$$f_0 = \frac{1}{T_0}$$

- O "pitch" significa na terminologia da psicologia cognitiva (e também em psicolingüística e psicofisologia) a *percepção humana da freqüência fundamental*
- Pitch pode ser tanto "percebido" pela f_0 quanto por T_0

FONEMAS SONOROS e SURDOS

- **Fonemas sonoros** (“voiced”), ou Excitação Sonora, são aqueles que ocorrem quando a força muscular empurra o ar dos pulmões, saindo pela traquéia e atravessando a glote, onde periodicamente o fluxo pode ser interrompido pelo movimento das cordas vocais. Os movimentos periódicos de abertura e o fechamento da glote ocorrem em resposta à pressão sub-glotal do ar da traquéia, sendo que este ciclo periódico é o responsável pelo formato da onda de ar emergente da glote.
- **Fonemas surdos**, ou Excitação Surda (“unvoiced”) ocorre quando forma-se uma constrição em algum lugar do trato vocal e depois força-se a passagem do ar através desta constrição.

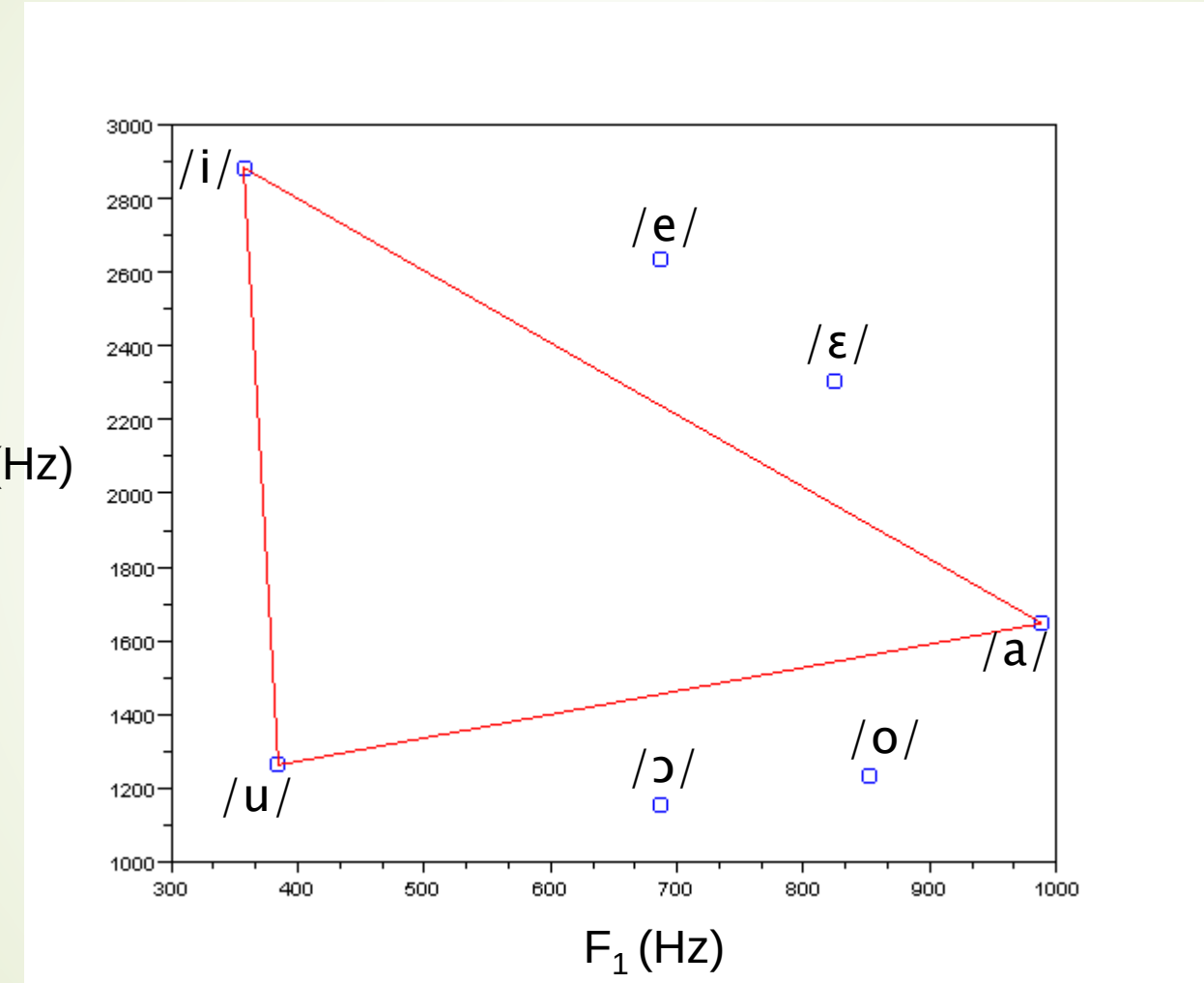
Sinais de Voz Sonoros e Surdos



Análise do Sinal

- 1000ms de amostra, f_a = 8kHz, banda telefônica (300-3400Hz)
- Segmentos de 37.5ms, J = 8000*0.0375 = 300
- Ruído de fundo 13dB, Fonema **/s/** 25dB acima. Fonema consonantal, surdo constritivo: espectro quase branco (Fonema consonantal sonoro constritivo, p.ex. /v/ de NOVE).
- Vogais **/e/** e **/i/** possuem “picos” e “vales” no espectro. Nas frequências ***Formantes*** f_1, f_2, f_3, f_4 , ou de *ressonância* nas cavidades, o sinal atinge maiores potências (em f_1, 50dB). Nas *anti-ressonâncias* a potência cai a 30 dB.

Triângulo das Vogais, português



Adaptado de Russo e Behlau, 1993

Plano de monotongos do português de São Paulo

https://pt.wikipedia.org/wiki/Fonologia_da_l%C3%ADngua_portuguesa#:~:text=O%20portugu%C3%AAs%20usa%20a%20altura,maioria%20dos%20dialetos%20do%20Brasil).

Classificação das vogais

	Front	Central	Back
Close	i	(ɨ)	u
Close-mid	e		o
Open-mid	ɛ	ɐ	ɔ
Open		a	

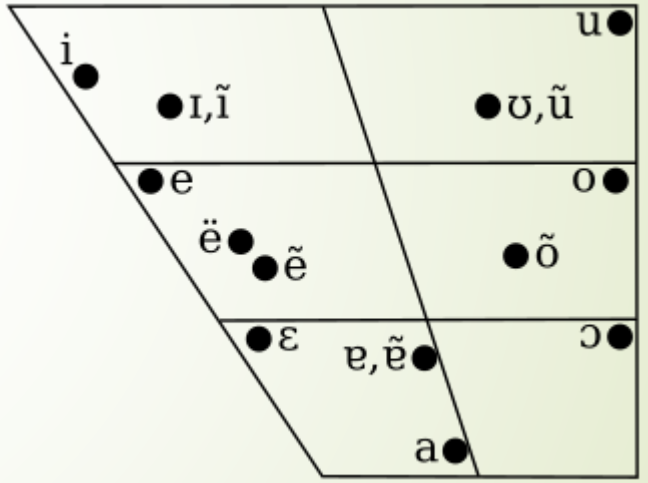
Português brasileiro

Vogais orais

/ɛ/	/'sɛ/	*sé*	/ɛ/
/e/	/'se/	*sê*	/e/
/ɪ ~ i/	/'si/	*se*	/ɨ/
/i/	/'si/	*si*	/i/
/ɔ/	/'pɔs/	*pós*	/ɔ/
/o/	/'pos/	*pôs*	/o/
/u/	/'tu/	*tu*	/u/
/ɐ/	/'ʁɐmɐ/	*rama*	/ɐ/
/a/	/'awmɐ/	*alma*	/a/

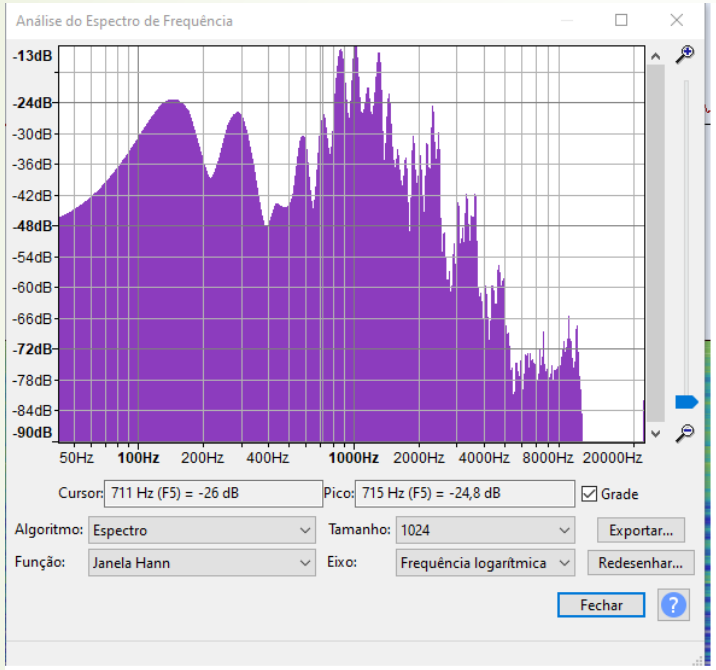
Vogais nasais

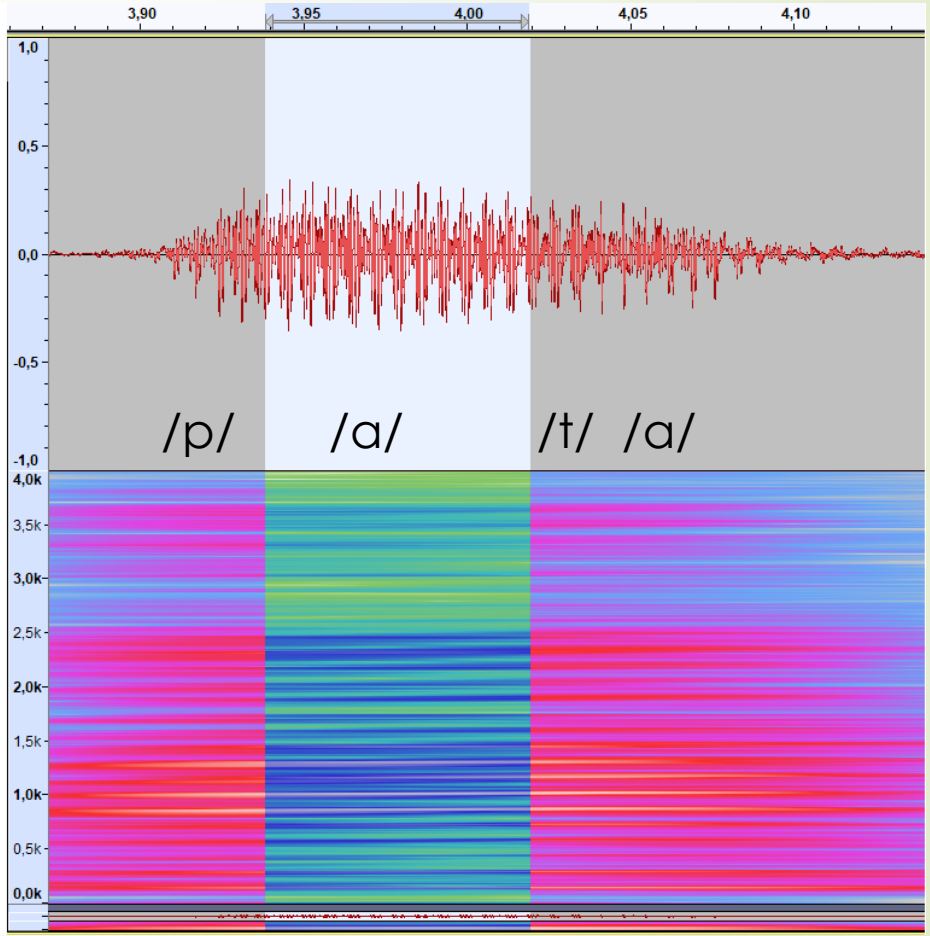
/ĩ/	/'vĩ/	*vim*	/ĩ/
/ẽ/	/'ẽtɾu/	*entro*	/ẽ/
/ɐ̃/	/'ɐ̃tɾu/	*antro*	/ã/
/õ/	/'sõ/ (in some dialects: /õw̃/)	*som*	/õ/

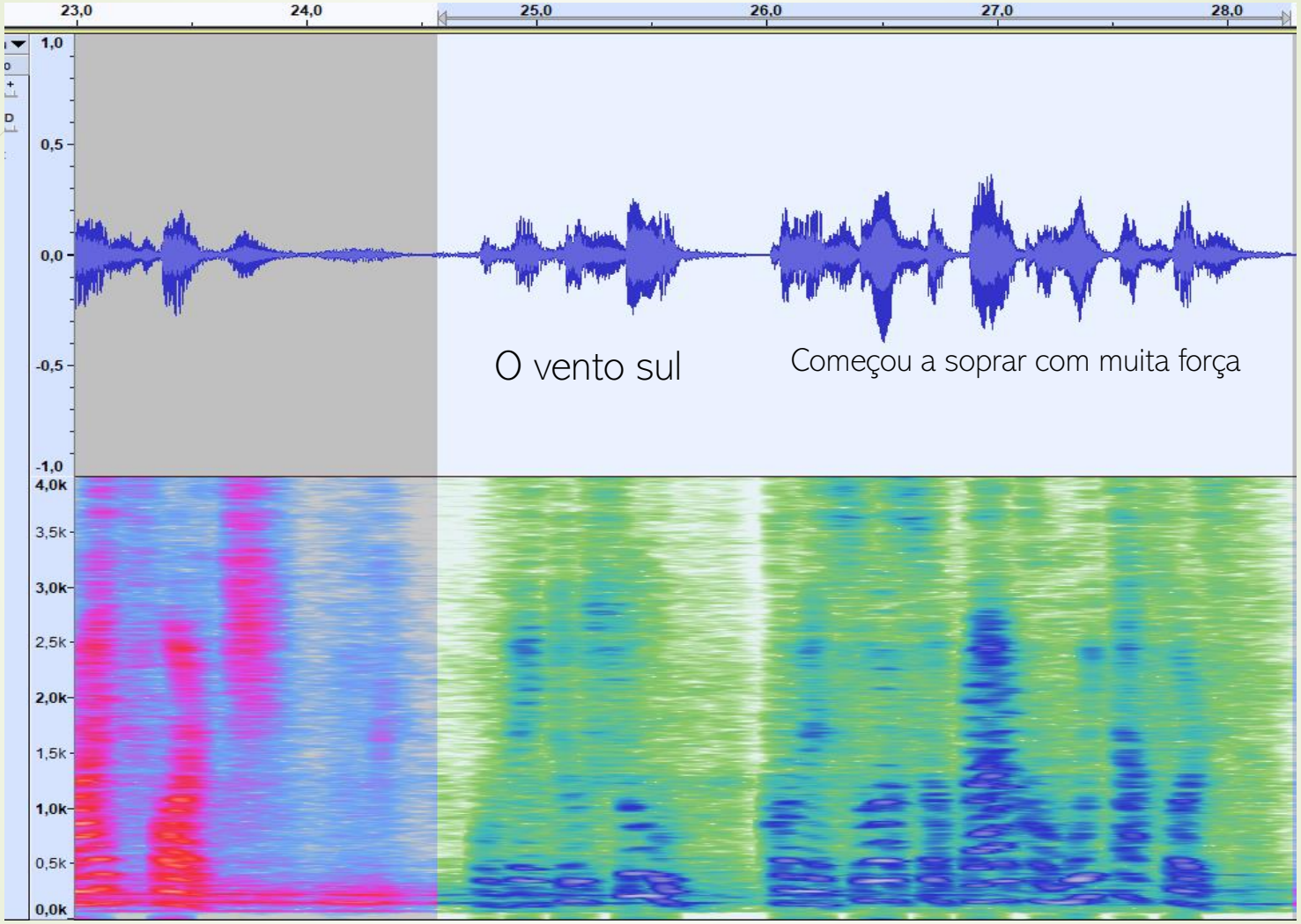


Barbosa, Plínio A. & Eleonora C. Albano (2004), "Brazilian Portuguese", *Journal of the International Phonetic Association* ***34***(2): 227-232 doi:10.1017/S0025100304001756

Espectro e Forma de Onda/ Espectrograma







Espectrograma

Exemplo para um sinal sonoro, como uma vogal

$$S(\Omega) = U(\Omega).H(\Omega).R(\Omega)$$

Modelo da Glote

Modelo do Trato Vocal

Modelo da Radiação da Fala

- Os três modelos lineares e separáveis, para simplicidade
- Propagação desde os pulmões, na traquéia, glote e trato vocal, através de uma onda de pressão plana, propagando-se progressivamente até os lábios

Modelo da Fonte de Excitação

Sonora: Cadeia quase-periódica de bolsões de ar

Surda: tipo turbulento, como ruído

Plosiva: Escape de ar após oclusão total

Sussurro (fricativo): Passagem através da glote semi-fechada

Silêncio: regiões do sinal sem som

Modelam a geração de fonemas de "mesma denominação"

Excitação SONORA ("Voiced")

- Características importantes:
  - Frequência fundamental *f0*
  - Duração de cada fase (aberta e fechada)
  - O instante da oclusão da Glote
  - O formato de cada pulso (abertura, fechamento)
- Exemplo de Modelo, no domínio-Z:

$$S(z) = \Theta_0 U(z) H(z) R(z)$$
$$= \Theta_0 E(z) G(z) H(z) R(z)$$

Comentários sobre o modelo:

- Os termos no domínio Z, correspondem exatamente aos análogos em w (contínuo)
- Coeficiente de ganho Θ_0
- $E(z)$ é a transformada Z do trem de impulsos $e(n)$, com período de pitch P
- $G(z)$ é o filtro de trato vocal (glote), $g(n)$ sua reposta impulsiva
- Logo,

$$u(n) = \sum_{i=-\infty}^{\infty} g(n - iP)$$

Excitação SURDA

- Um tipo de excitação surda são sons que friccionam com grandes constrições no trato vocal (fricativo)
- Outro tipo é um súbito escape de ar depois da abertura rapidíssima de uma oclusão (plosivo)
- O modelo para ambos é um **ruído branco *N*(*z*)**:

$$S(z) = \Theta_0 N(z) H(z) R(z)$$

Ou, no domínio da frequência (DTFT):

$$S(w) = \Theta_0 N(w) H(w) R(w)$$

Modelamento do Trato Vocal

- Comprimento de onda de uma onda plana acústica de 4kHz:

$$\lambda_{4kHz} = \frac{v_{som}}{f} = \frac{340\,\text{m/s}}{4000\,\text{ciclos/s}} = 8.5\,\text{cm}$$

- Como o diâmetro do trato vocal é de ± 2 cm, a hipótese de uma onda plana se propagando dentro dele, é razoável.
- Leis importantes: da Continuidade e de Newton

$p(x,t)$ pressão sonora

$\vec{v}_\zeta(x,y,z,t)$ Vetor velocidade no ar de uma partícula ζ

ρ Densidade do ar no tubo

$$\frac{1}{\rho v_{som}^2}\frac{\partial p(x,t)}{\partial t} = -\nabla \vec{v}_\zeta(x,y,z,t)$$

$$\rho\frac{\partial \vec{v}_\zeta(x,y,z,t)}{\partial t} = -\nabla \bullet p(x,t)$$

Aproximação onda plana propagando na direção *x* (origem na glote para os lábios):

$A(x,t)$ seção transversal variável do trato vocal, na posição x e instante t

$\vec{v}(x,t)$ velocidade de um volume de ar, na posição x e instante t:

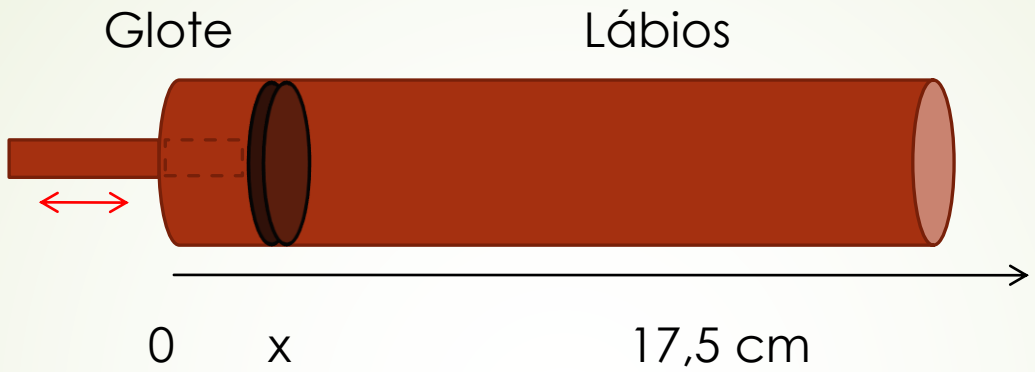
$$\vec{v}(x,t) = A(x,t)\vec{v}_{\zeta}(x,t)$$

Substituindo nas expressões tridimensionais:

$$-\frac{\partial v(x,t)}{\partial x} = \frac{1}{\rho v_{som}^2}\frac{\partial\left[p(x,t)A(x,t)\right]}{\partial t} + \frac{\partial A(x,t)}{\partial t}$$

$$-\frac{\partial p(x,t)}{\partial x} = \rho\frac{\partial\left[v(x,t)/A(x,t)\right]}{\partial t}$$

Modelo de 1 Tubo sem Perdas



Terminação ABERTA, lábios abertos, o desvio da pressão será nulo em $x=l$ (l = 17,5cm), em relação à pressão ambiente:

$$p(l,t) = p_{labios}(t) = 0$$

Para regime permanente, a fonte na glote pode ser modelada por exponencial complexa:

$$v(0,t) = u_{glote}(t) = \overline{U}_{glote}(\Omega)e^{j\Omega t}$$

Modelo para lábios abertos:

$$v(l,t) = \frac{\overline{U}_{glote}(\Omega)}{\cos(\Omega l / v_{som})} e^{j\Omega t} \overset{def}{=} \overline{U}_{labios}(\Omega) e^{i\Omega t}$$

Onde $\overline{U}_{glote}(\Omega)$ é o fasor para o sinal $u_{glote}(t)$.

A função de transferência para o trato vocal é dada pela relação entre os fasores das velocidades nos lábios e na glote:

$$H(\Omega) = \frac{\overline{U}_{labios}(\Omega)}{\overline{U}_{glote}(\Omega)} = \frac{u_{labios}(t)}{u_{glote}(t)} = \frac{1}{\cos(\Omega l / v_{som})}$$

Resultados para lábios abertos:

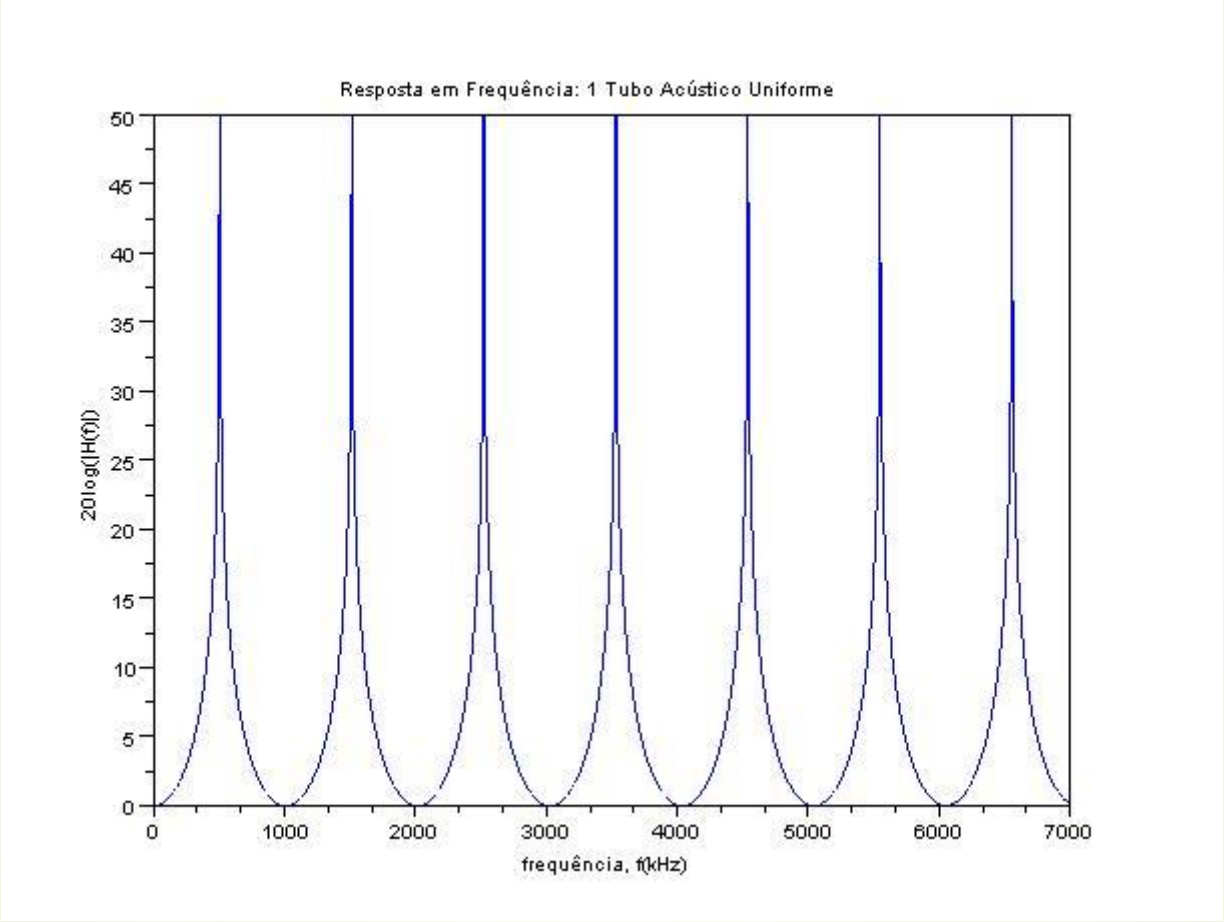
- As frequências de ressonância f_i para este modelo são obtidas igualando o denominador a zero:

$$\frac{\Omega_i l}{v_{som}} = \frac{\pi}{2}(2i-1) \qquad \text{para} \qquad i = 1,2,3,4,...$$

- Como $\Omega_i = 2\pi f_i$, as ressonâncias ocorrerão nas frequências:

$$f_i = \frac{v_{som}}{4l}(2i-l) \qquad \text{para} \qquad i = 1,2,3,4,...$$

Resposta em Frequência, modelo 1 Tubo



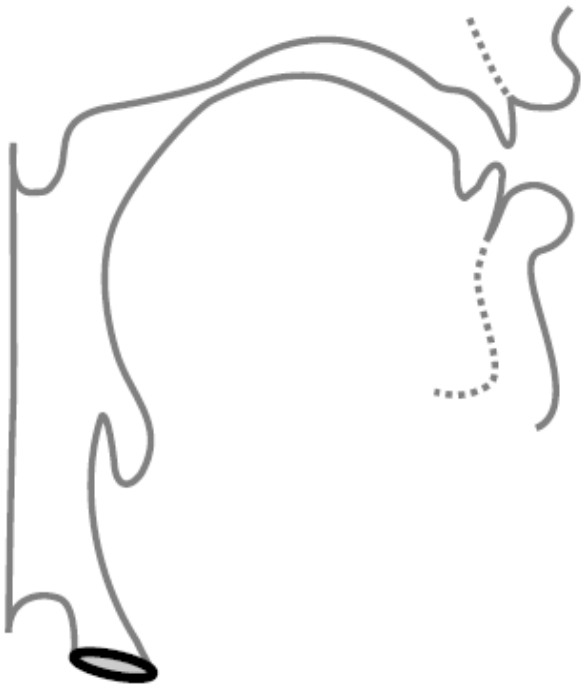
$v_{som} = c = 353.027$ m/s, $\Theta = 37^{o}$C, $l = 17.5$ cm

fonte: http://www.sengpielaudio.com/calculator-speedsound.htm

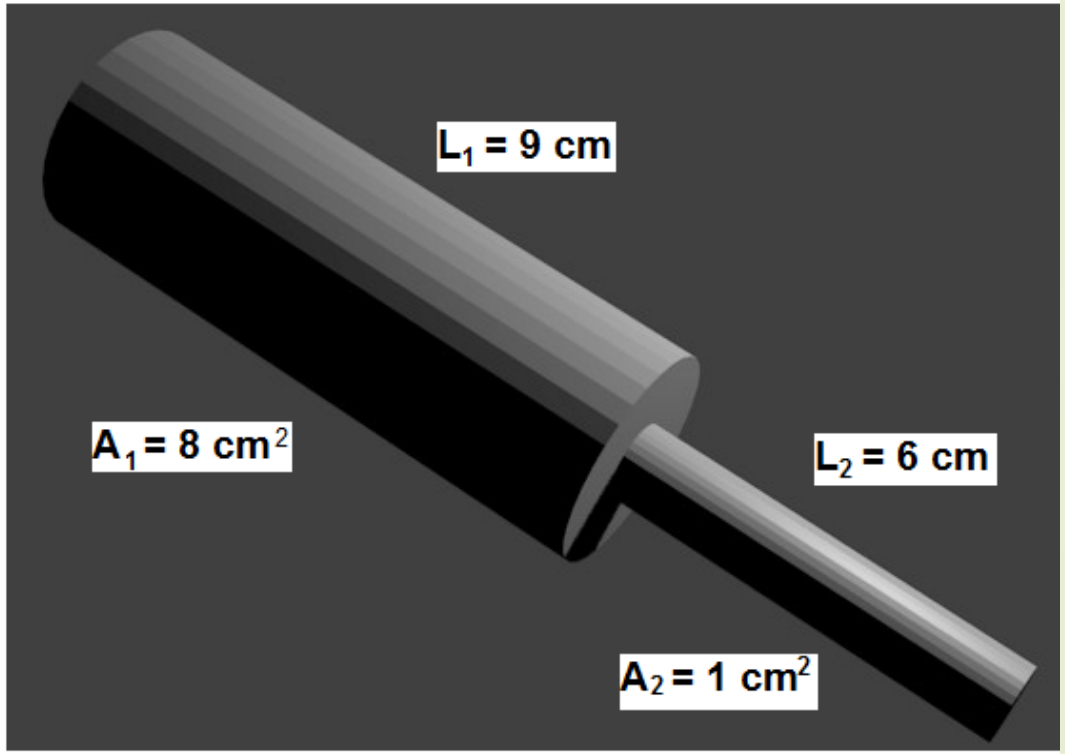
Script Matlab modelo de 1 tubo

```
%  modelamento acustico do trato vocal
%  Minami - 29 agosto 2013
maxgain = 50;  % ganho maximo = 50dB
l = 17.5e-2;  % comprimento do trato vocal em m
v = 350;  % velocidade do som, m/s, no ar na temperatura de 37 celsius
fat = l/v;
omega = 0:6000;
w = 2*pi*omega*fat;
den = cos(w);
ntot = prod(size(den));
for i=1:ntot
  H(i) = 20*log10(abs(1/den(i)));
  if H(i)>maxgain
    H(i) = maxgain;
  end
end
clf
plot(omega,H);
xtitle( 'Resposta em Frequência de um Tubo Acústico Uniforme',
'frequência, f(kHz)', '20log(|H(f)|)');
```

Modelo 2 Tubos

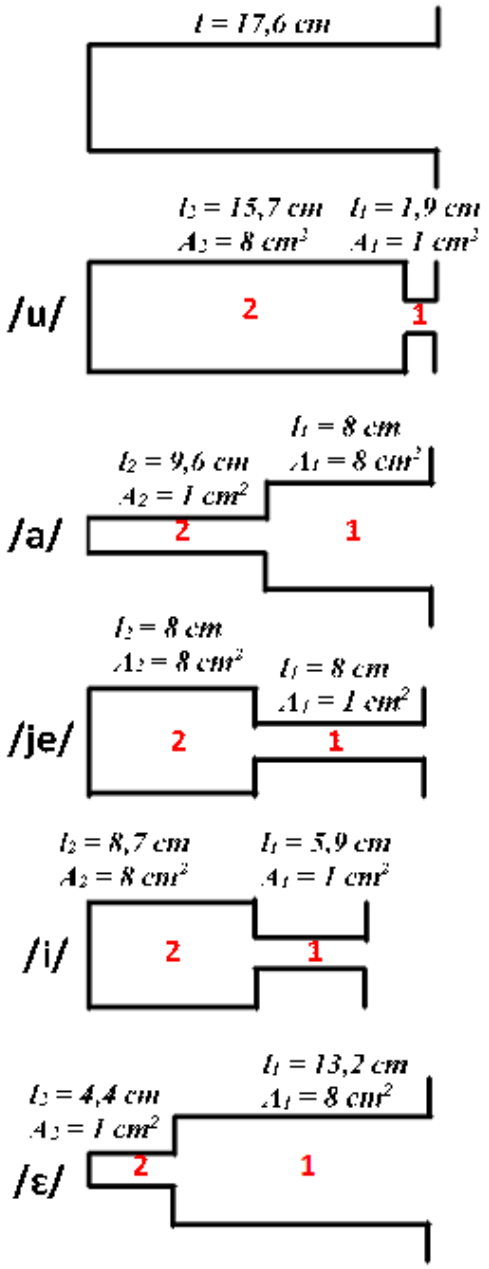


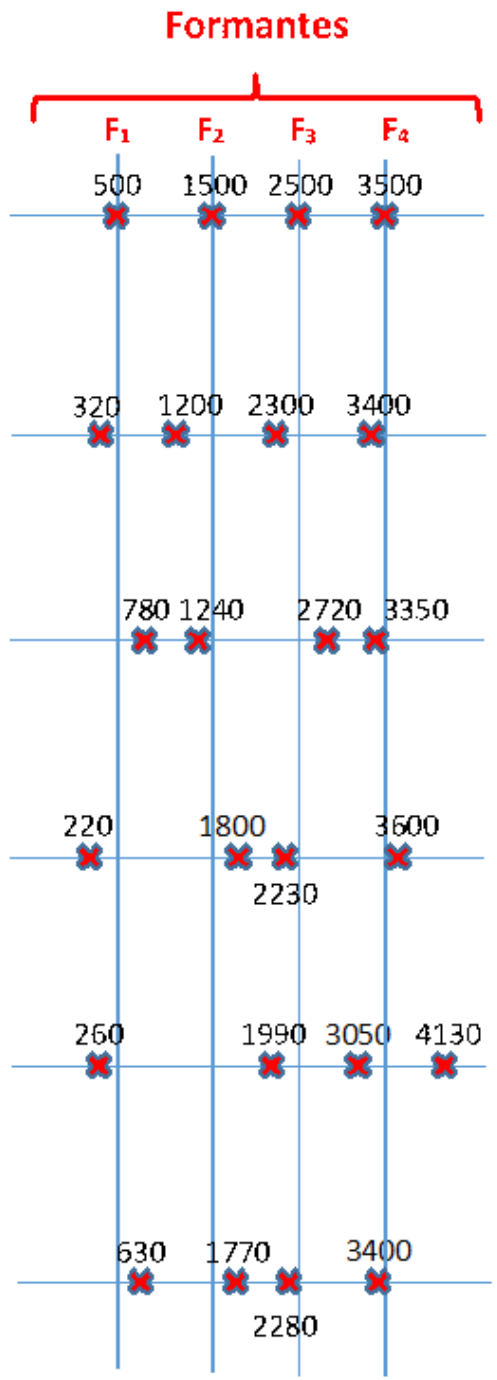
Articulações Trato Vocal
para fonema sonoro /i/



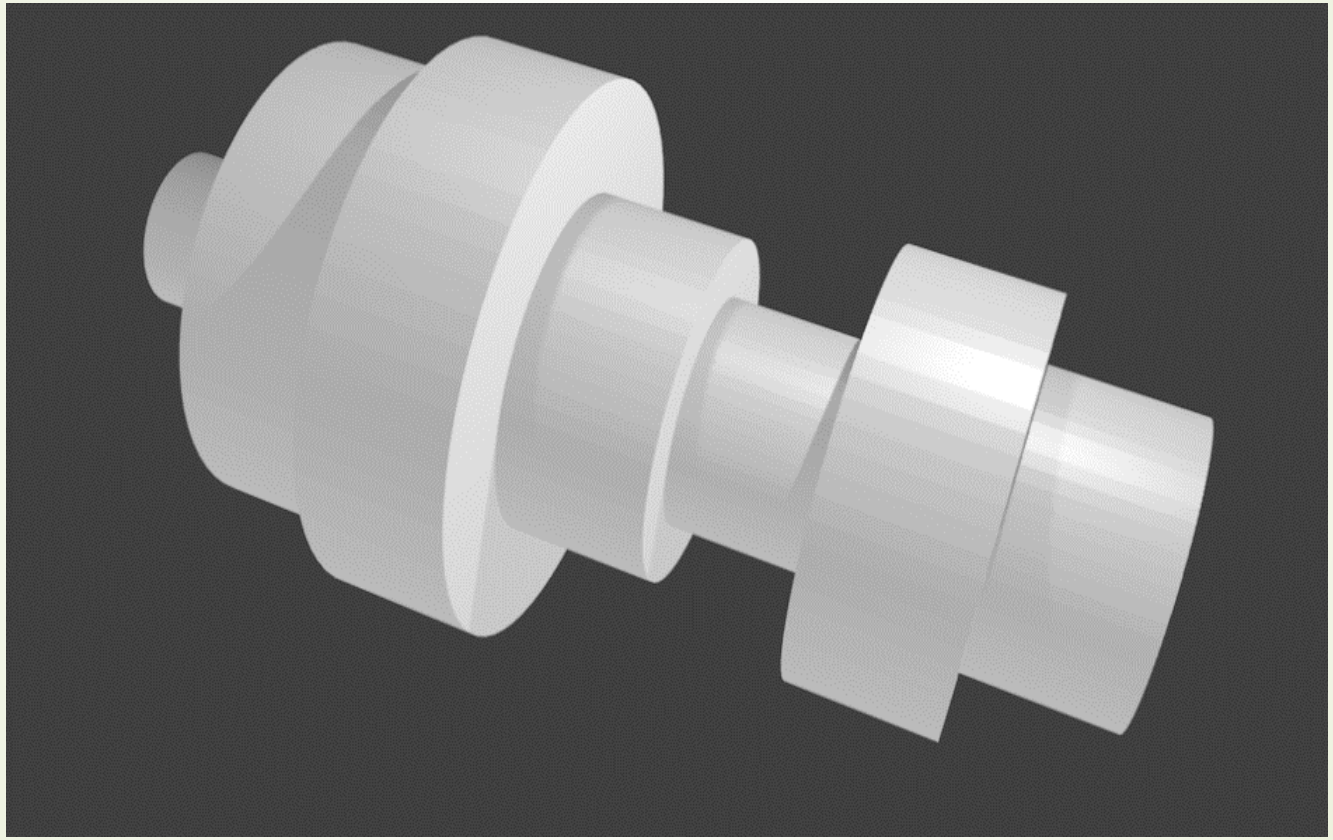
Modelo de Dois Tubos para o Trato Vocal /i/

Frequências de Ressonância do Modelo 2 tubos

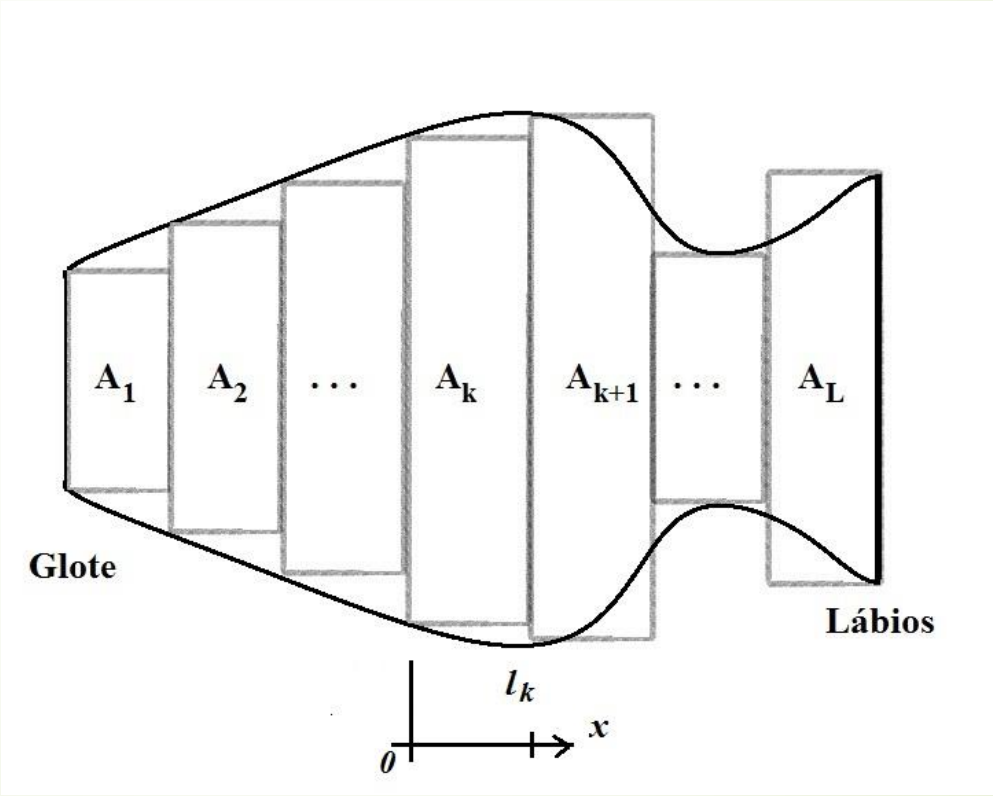




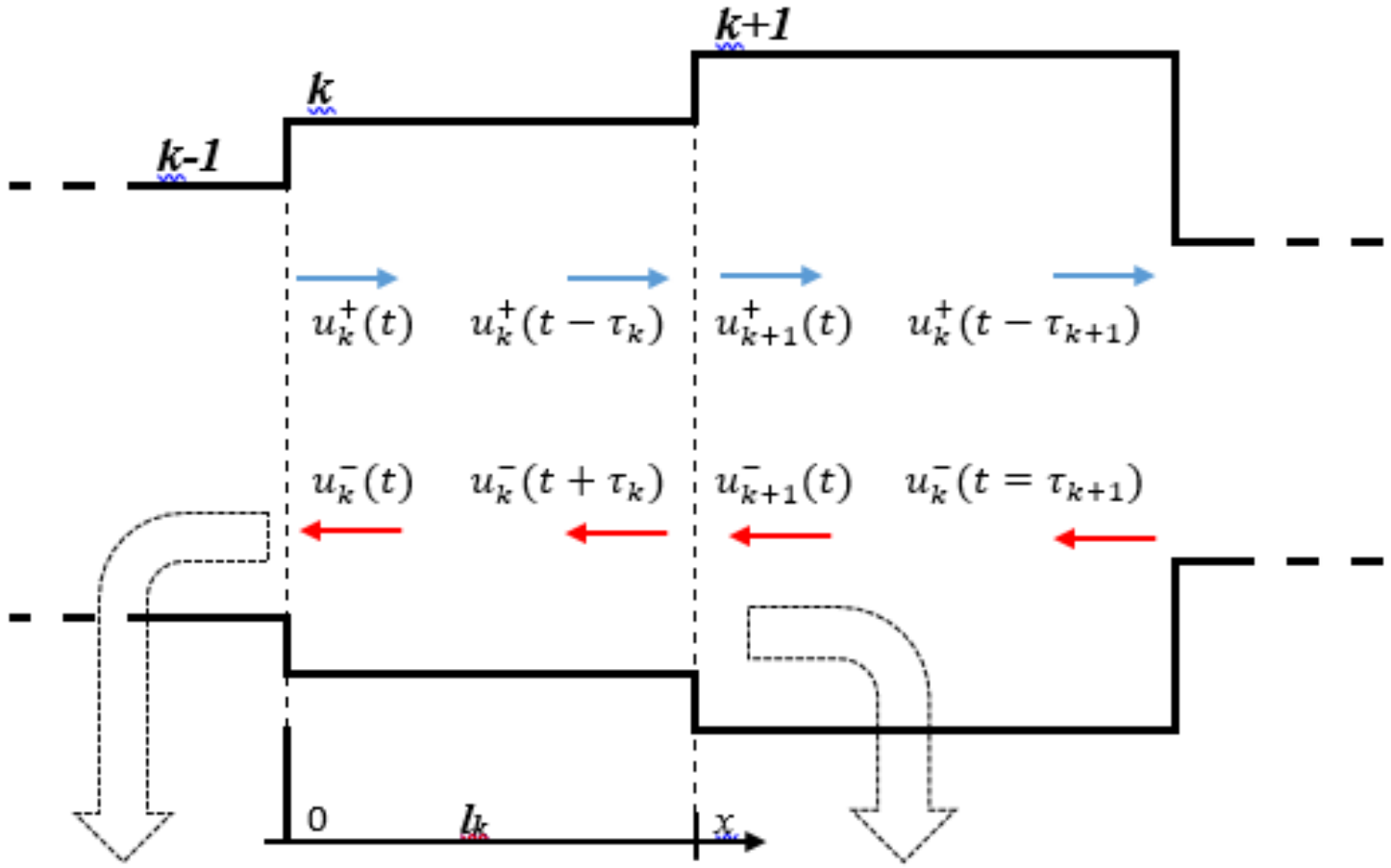
Multi-Tubos



Aproximação para qualquer conformação



Condições de Contorno



$$u_k^+(t) = u_k^+(t - 0/c)$$
$$u_k^-(t) = u_k^-(t + 0/c)$$
$$u_k(0,t) = u_k^+(t) - u_k^-(t)$$

$$u_k^+(t-\tau_k) = u_k^+(t - l_k/c)$$
$$u_k^-(t+\tau_k) = u_k^-(t + l_k/c)$$
$$u_k(l_k,t) = u_k^+(t-\tau_k) - u_k^-(t+\tau_k)$$

Ondas Progressivas e Regressivas

$$u_{k+1}^{+}(t) = u_{k}^{+}(t-\tau_k)\left[\frac{2A_{k+1}}{A_{k+1}+A_k}\right] + u_{k+1}^{-}(t)\left[\frac{A_{k+1}-A_k}{A_{k+1}+A_k}\right]$$

$$u_{k}^{-}(t+\tau_k) = u_{k+1}^{-}(t)\left[\frac{2A_{k+1}}{A_{k+1}+A_k}\right] - u_{k}^{+}(t-\tau_k)\left[\frac{A_{k+1}-A_k}{A_{k+1}+A_k}\right]$$

Coeficientes de Transmissão e Reflexão

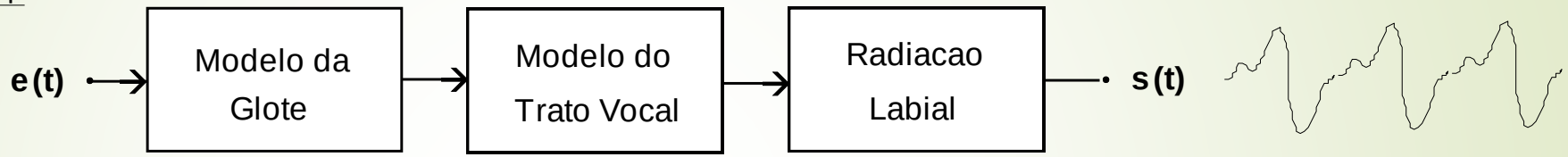
Transmissão: $r_k^+ = \frac{2A_{k+1}}{A_{k+1}+A_k}$

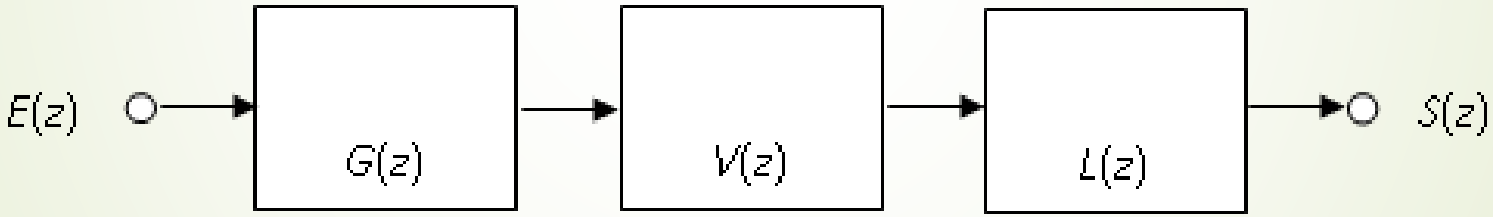
Reflexão: $r_k^- = \frac{A_{k+1}-A_k}{A_{k+1}+A_k}$

$$r_k \stackrel{\text{def}}{=} r_k^-$$

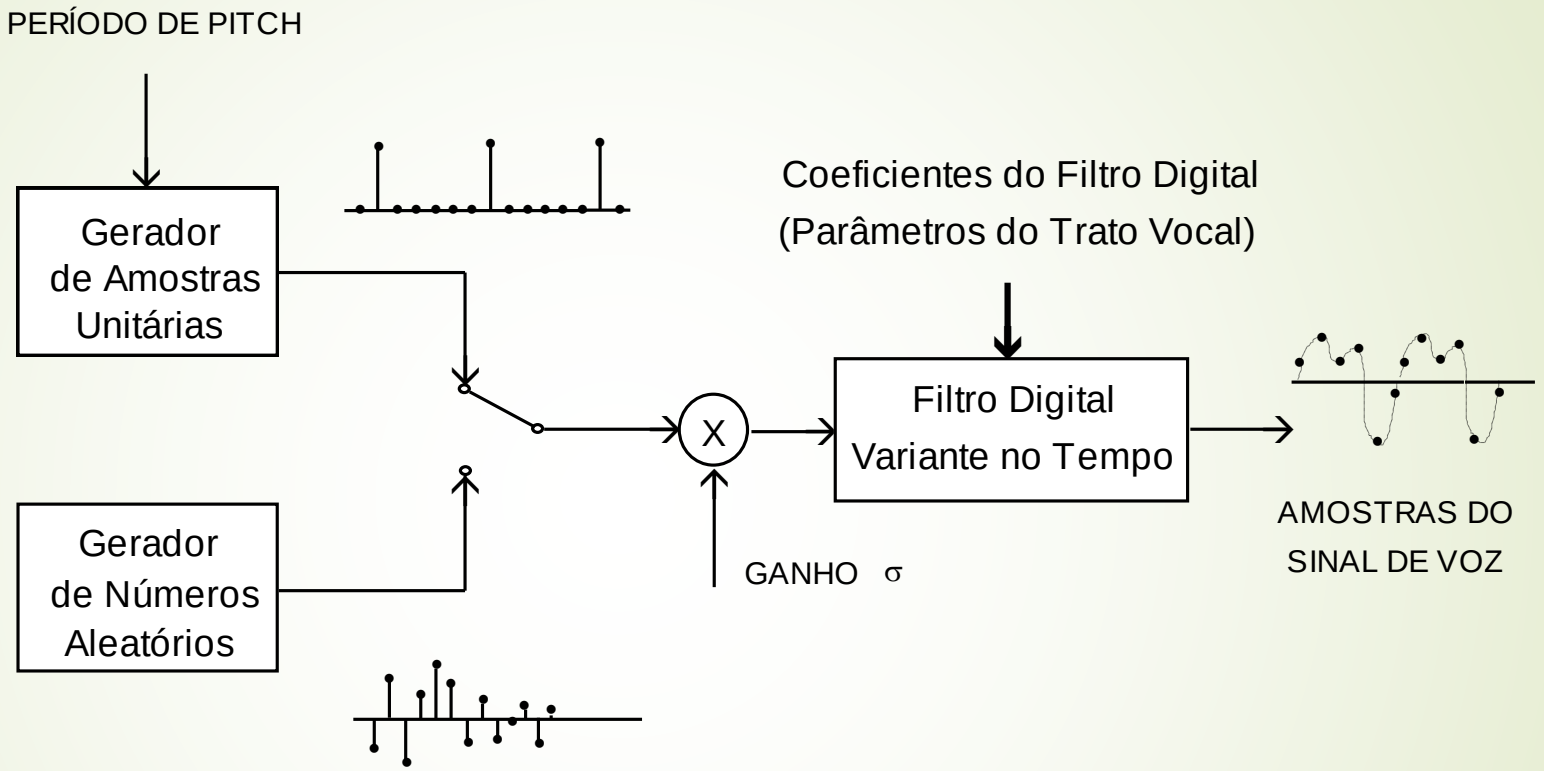
$$-1 \leq r_k \leq 1.$$

Modelo do Trato Vocal



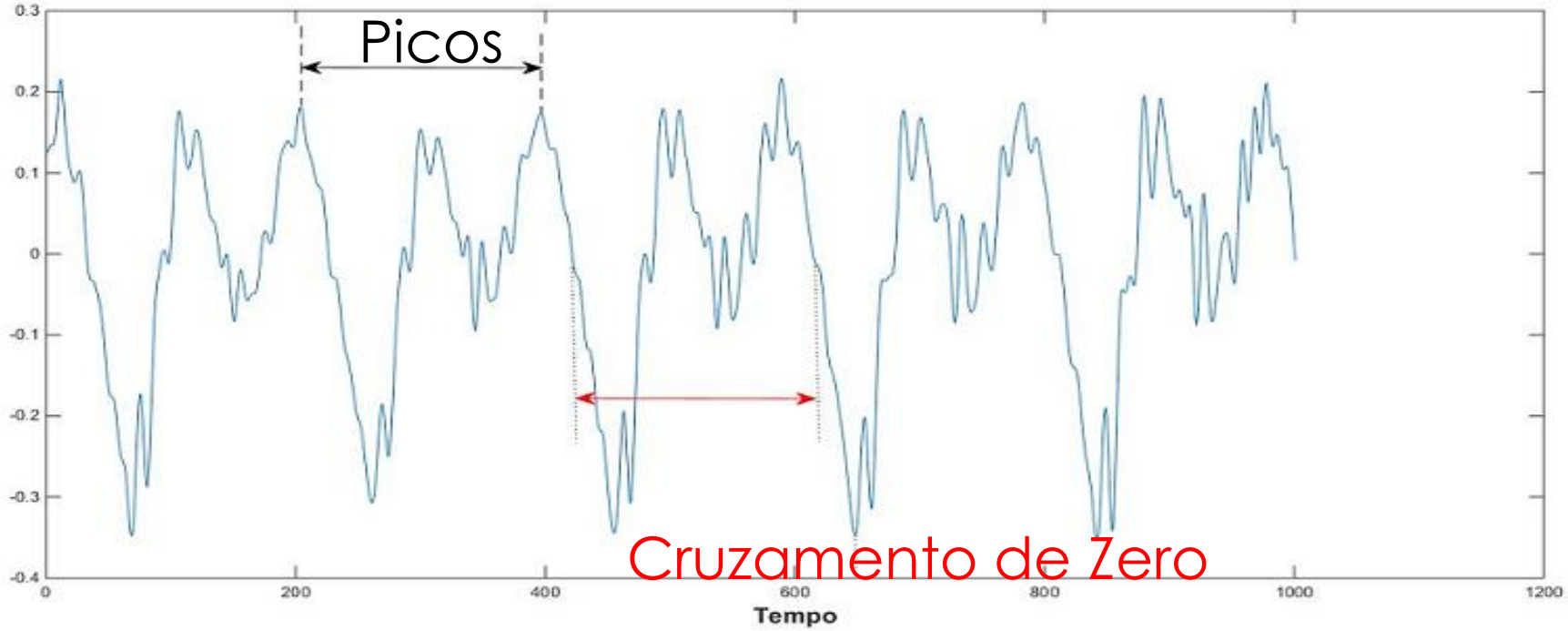


Modelo de Tempo Discreto

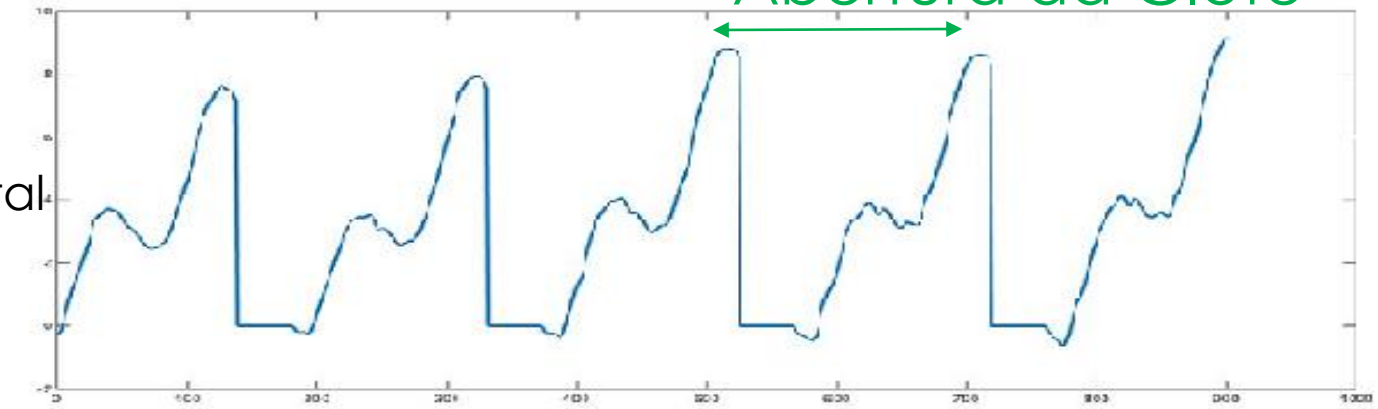


$$H(z) = \sigma_{LPC} \frac{1}{A(z)}$$

Pitch Diferentes Definições



Pulso Glotal



Determinar o Pitch pela Autocorrelação

$$r(d,q) = \frac{1}{K} \sum_{n=q}^{q+K+d} s(n)s(n+d)$$

Janela K, ponto q

RABINER, L. R. (1977, February). **On the Use of Autocorrelation Analysis for Pitch Detection**. *IEEE Transactions On Acoustics, Speech, And Signal Processing*, pp. 24-33.

Questões (P2 - 0,5 pts) para entregar:

1. (**Questão Obrigatória, Já fazendo parte da P2, isto é, quem não entregar esta, não terá as outras corrigidas nem incluídas na nota**) Grave com o Audacity (ou outro programa) com sua própria voz, os seguintes CINCO (05) roteiros de arquivos, salvos no formato .wav:

   a) Seu nome completo, pronunciado com clareza, salve como "NomesSobrenomes.wav", pex. "JoaoCarlosMartinsDeSouza.wav"

   b) Os números do seu RA completo, pronunciado dígito a dígito, salve como "RAxxxxxxxx.wav", pex. a aluno com RA11062416 grava os dígitos "um", "um", "zero", "seis", "dois", "quatro", "um", "seis" e nomeia o arquivo como "RA11062416.wav".

   c) Os pares de seu RA completo, p.ex. o aluno com RA11062416, grava os pares "onze", "seis", "vinte e quatro" e "dezesseis", e salva como "RA11062416_pares.wav".

   d) Grave o trecho de O burrinho pedrês em Sagarana, salve como "Burrinho_NomeSobrenome.wav":

   "*Folgado, Sete-de-Ouros endireitou para a coberta. Farejou o cocho. Achou milho. Comeu. Então, rebolcou-se, com as espojadelas obrigatórias, dançando de patas no ar e esfregando as costas no chão. Comeu mais*"

   e) Grave o trecho da Lírica de Camões e salve como Camoes_NomeSobrenome::

   "*Campos bem aventurados,*
   Tornai-vos agora tristes,
   Que os dias em que me vistes
   Alegre, já são passados".

Questões para entregar/PROVA P2 (2,0 pts)

2. O que são o **pitch** e fundamental f_0? Usando a forma de onda e o espectrograma e do Audacity, encontre um **pitch** no seu Nome e uma f_0 no seu sobrenome, usando o arquivo **1.a** anterior, imprima a imagem usada e destaque suas respostas na imagem, indicando de qual fonema está efetuando cada medição (*).
3. O que é o modelo Fonte-Filtro para o trato vocal? Quais são os tipos de Fonte e os articuladores no Filtro?
4. O que são as formantes f_1 a f_4? Usando dois números diferentes dentro do arquivo **1.b** anterior, usando o espectrograma ou o espectro, determine as formantes de duas vogais diferentes, imprimindo e destacando na imagem estes valores (*).
5. No que consiste o modelo de tubos acústicos para as formantes? O que é o coeficiente de reflexão ($\boldsymbol{k_i}$) neste modelo?
6. Desenhe o modelo de tubos acústicos do trato vocal, sendo o vetor dos coeficientes de reflexão, considerando $A_1 = \pi$ cm^2 para **k** = `[0.5 -0.7 -0.3 0.2 -0.4 0.9]`
7. Quais as diferenças entre as excitações surda e sonora? E o que são as consoantes? Registre trechos no arquivo **1.d** anterior para trechos sonoros, surdos e consoantes, usando ou forma de onda ou espectrograma, indicando qual fonema está destacando (*).
8. O que são os fonemas plosivos? No arquivo **1.e** anterior, encontre-os através do espectrograma e da forma de onda e registre nas figuras qual fonema destacou (*).
9. O que é a Energia de Tempo-Curto? Que informações ela revela?
10. Pesquise pelo menos dois métodos (um temporal e outro na frequência) para determinação do **Pitch**.

(*) Se desejar, envie TAMBÉM o trecho específico de voz .wav analisado.