

Geração de prosódia para o português brasileiro em sistemas text-to-speech

Felipe Cortez de Sá

Universidade Federal do Rio Grande do Norte

Junho de 2018

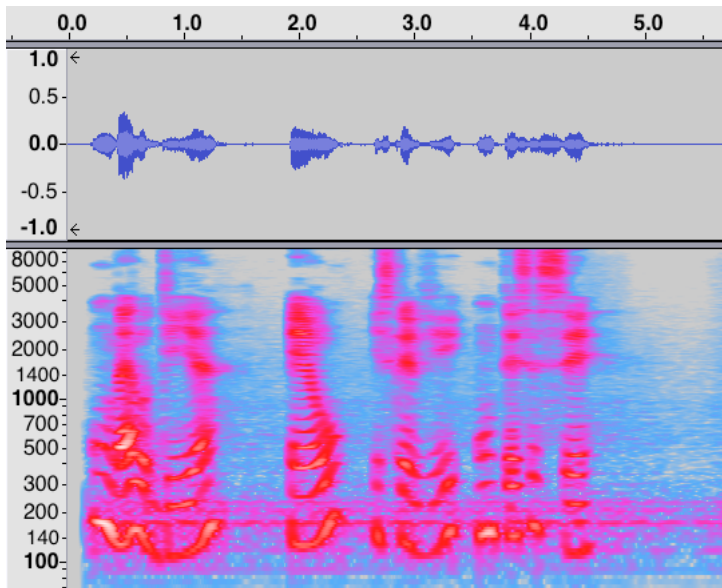
- 1 Introdução
- 2 Fundamentação teórica
 - Prosódia
 - Sistemas text-to-speech
- 3 Implementação
- 4 Perguntas

It would be a considerable invention indeed, that of a machine able to mimic speech, with its sounds and articulations. I think it is not impossible.” (Leonhard Euler, 1761)

- Euler e Wolfgang von Kempelen
- *Voice User Interfaces*
 - *Apple - Siri*
 - *Google Assistant*
 - *Microsoft - Cortana*
 - *Amazon - Alexa*
- Acessibilidade
- Ensino de linguagens
- Estudo de linguística
- Prosódia afetiva

- pros (verso) - odé (canto)
- Suprasegmental
- Frequência
- Duração
- Intensidade

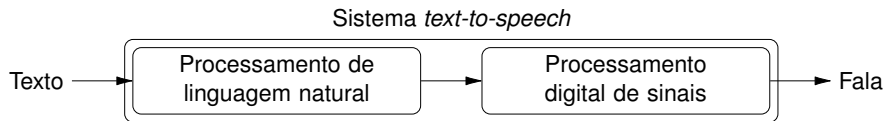
Prosódia



Função prosódica

- Suprasegmental
- Afetiva
- Aumentativa

Sistemas text-to-speech



■ Front end

- Normalização de texto
- Conversão grafema-fone
- Geração de prosódia

■ Back end

- Síntese articulatória
- Síntese por formantes
- Síntese concatenativa
- Síntese por Hidden Markov Models e Deep Neural Networks

- Normalização de texto
 - A conta deu R\$ 20, V. Exa. Pode conferir?
- Conversão grafema-fone
 - Gosto de pão
 - [ɡɔstu] (gósto)
 - [ɡostu] (gôsto)
- Geração de prosódia

O desafio da geração de prosódia

O desafio da geração de prosódia

Capítulo LV - O Velho Diálogo de Adão e Eva

Brás Cubas: ?

Virgília:

Brás Cubas:

Virgília: !

Brás Cubas:

Virgília:

. ?

.

Brás Cubas:

Virgília:

Brás Cubas: !

Virgília. ?

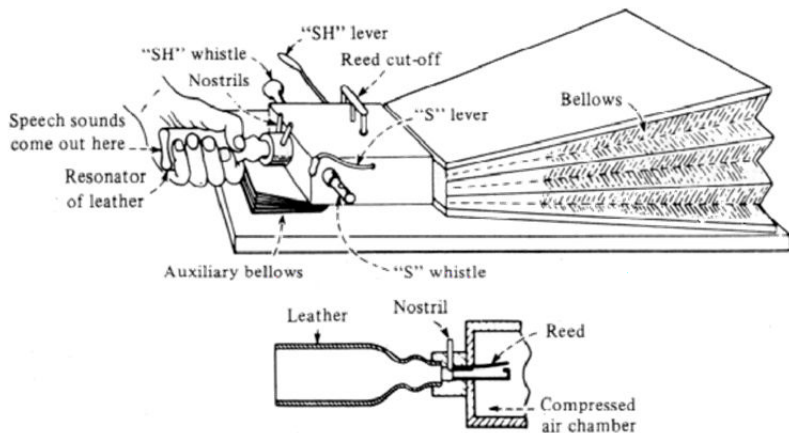
Brás Cubas. !

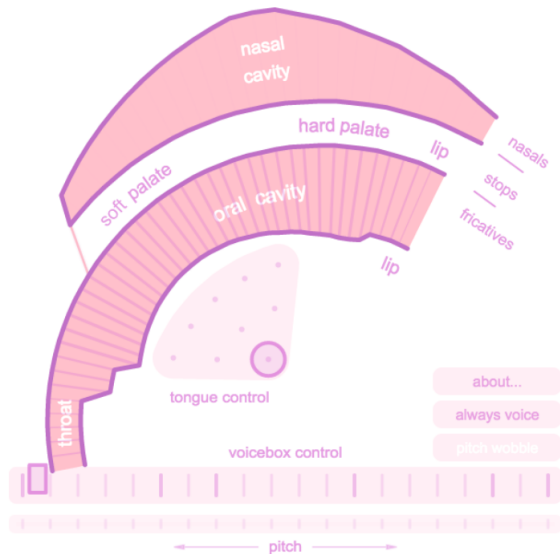
Virgília. !

- Heurísticas derivadas a mão
- Sistemas baseados em análise gramatical
- Métodos baseados em corpus

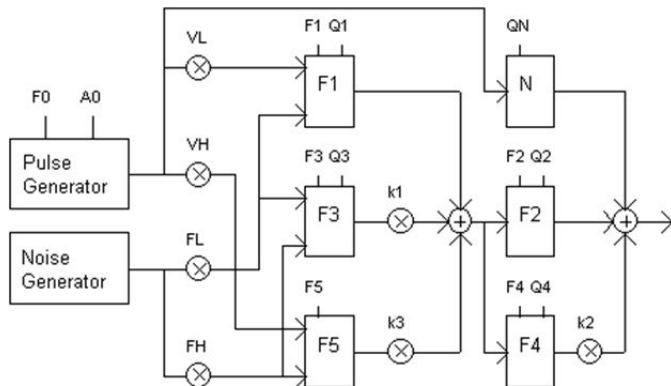
- Síntese articulatória
- Síntese por formantes
- Síntese concatenativa
- Síntese por Hidden Markov Models e Deep Neural Networks

Máquina de Wolfgang von Kempelen (1778)



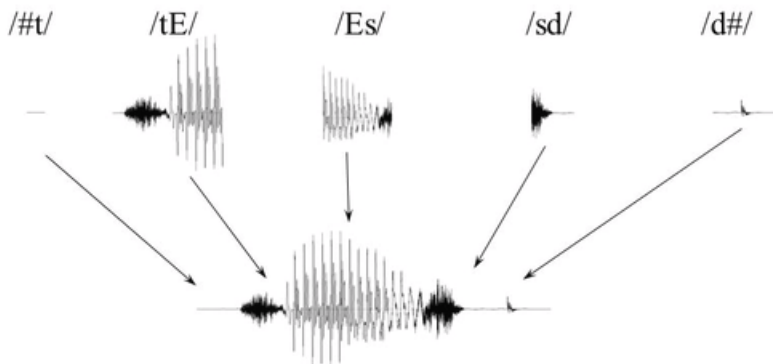


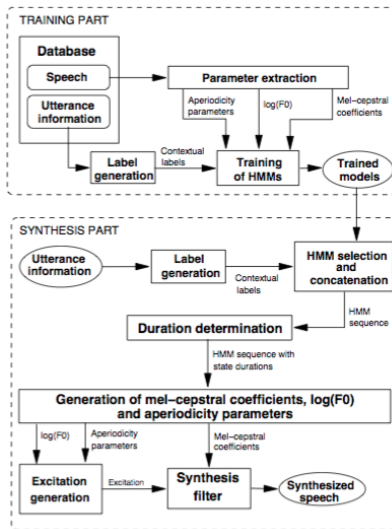
Formantes



Concatenativa

test = /tEsd/ = /#t/ + /tE/ + /Es/ + /sd/ + /d#/

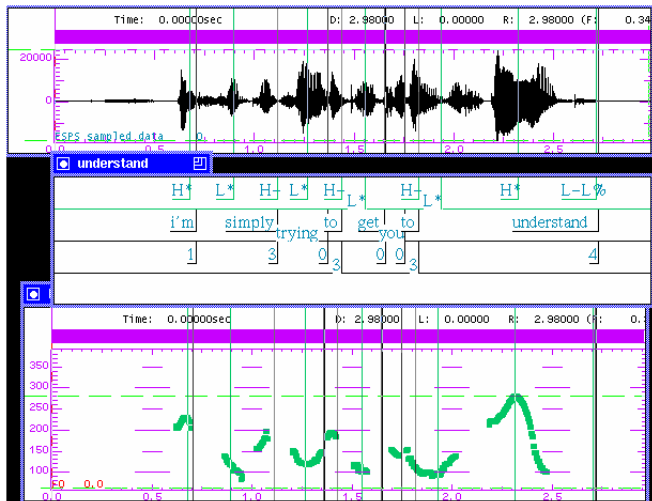


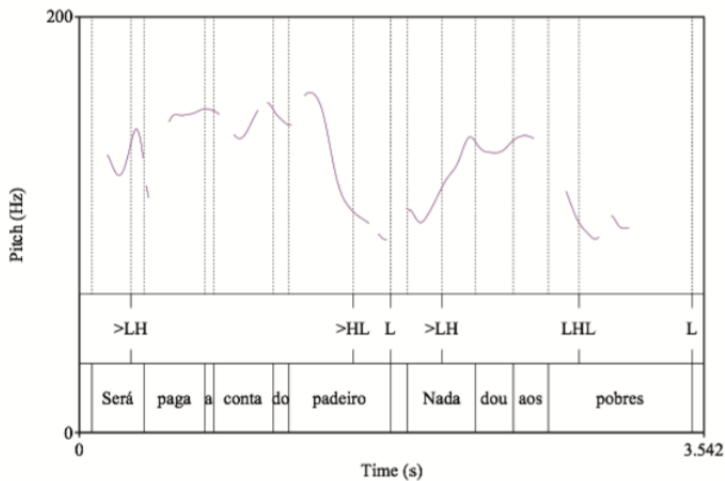


Sistemas TTS para o português brasileiro

- Aiuruetê (1997) – curvas de frequência pré-definidas
- eSpeakNG (2006) – *pre-head, head, nucleus, tail*
- Couto et al (2010) – probabilístico
- LianeTTS (2011) – partes do discurso

- ToBI – Tone Breaks and Indices
- DaTo – Dynamic Tones
- INTSINT – International Transcription System for Intonation





Ele FOI lá HOje?

[↑ > ↑↑ ↓]

- SSML – Speech Synthesis Markup Language
- EmotionML
- Anotações entoacionais

```
< speak >
```

```
  Siga < emphasis level = "strong" >aquele< / emphasis > carro.
```

```
< / speak >
```

```
<emotionml version="1.0" xmlns="http://www.w3.org/2009/10/emotionml">  
  <emotion category-set="http://www.w3.org/TR/emotion-voc/xml#everyday-categories">  
    <emotion>  
      <category name="happy" />  
      Que bom te ver!  
    </emotion>  
  </emotionml>
```

(Utterance Words

(The

(boy ((accent L*))

saw

the

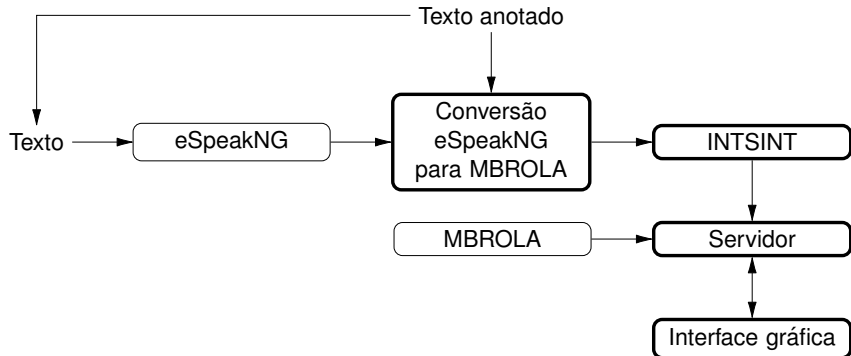
(girl ((accent H*) (tone L-)))

with

the

(telescope ((accent H*) (tone H-H%))))))

Arquitetura



```
$ espeak-ng -v pt-br 'Bom dia' -x -q --sep=  
$ b/'o/N dZ/'i/&
```

Conversão de fones

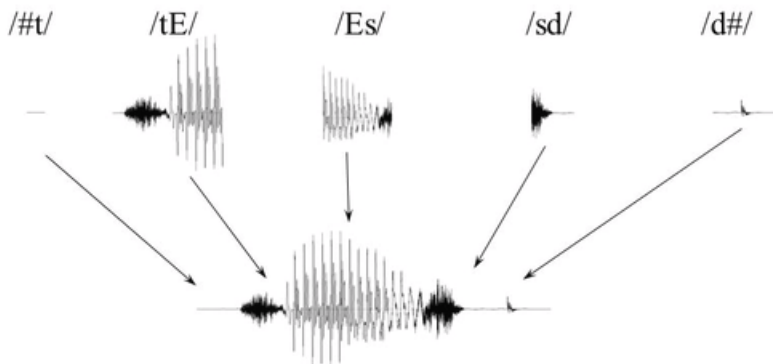
b | b | _b_arco
k | k | _c_om
d | d | _d_ose
& | a | v_a_le
6 | @ | tam_a_nho
n | n | _n_unca

Regra	Cálculo
Top	$\text{key} \times \sqrt{2^{range}}$
Middle	key
Bottom	$\text{key} / \sqrt{2^{range}}$
Higher	$\sqrt{P_{i-1} \times T}$
Same	P_{i-1}
Lower	$\sqrt{P_{i-1} \times B}$
Upstepped	$\sqrt{P_{i-1} \times \sqrt{P_{i-1} \times T}}$
Downstepped	$\sqrt{P_{i-1} \times \sqrt{P_{i-1} \times B}}$

_ 150 50 150
t 70 50 125
e 125 50 75
c 70 50 125
e 125 50 75
c 70 50 125
e 116 20 232 80 300
_ 150 50 150

Concatenando dífonos

test = /tEsd/ = /#t/ + /tE/ + /Es/ + /sd/ + /d#/



Demonstração!

- Adicionar suporte a outros modelos de análise entoacional
- Usar *Natural Language Understanding* para estimar prosódia
- Gerar prosódia a partir de marcação SSML
- Criação de corpus anotado com prosódia para o português brasileiro

*If you understood everything I said, you'd be
me* (Miles Davis)