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

Resumo

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)

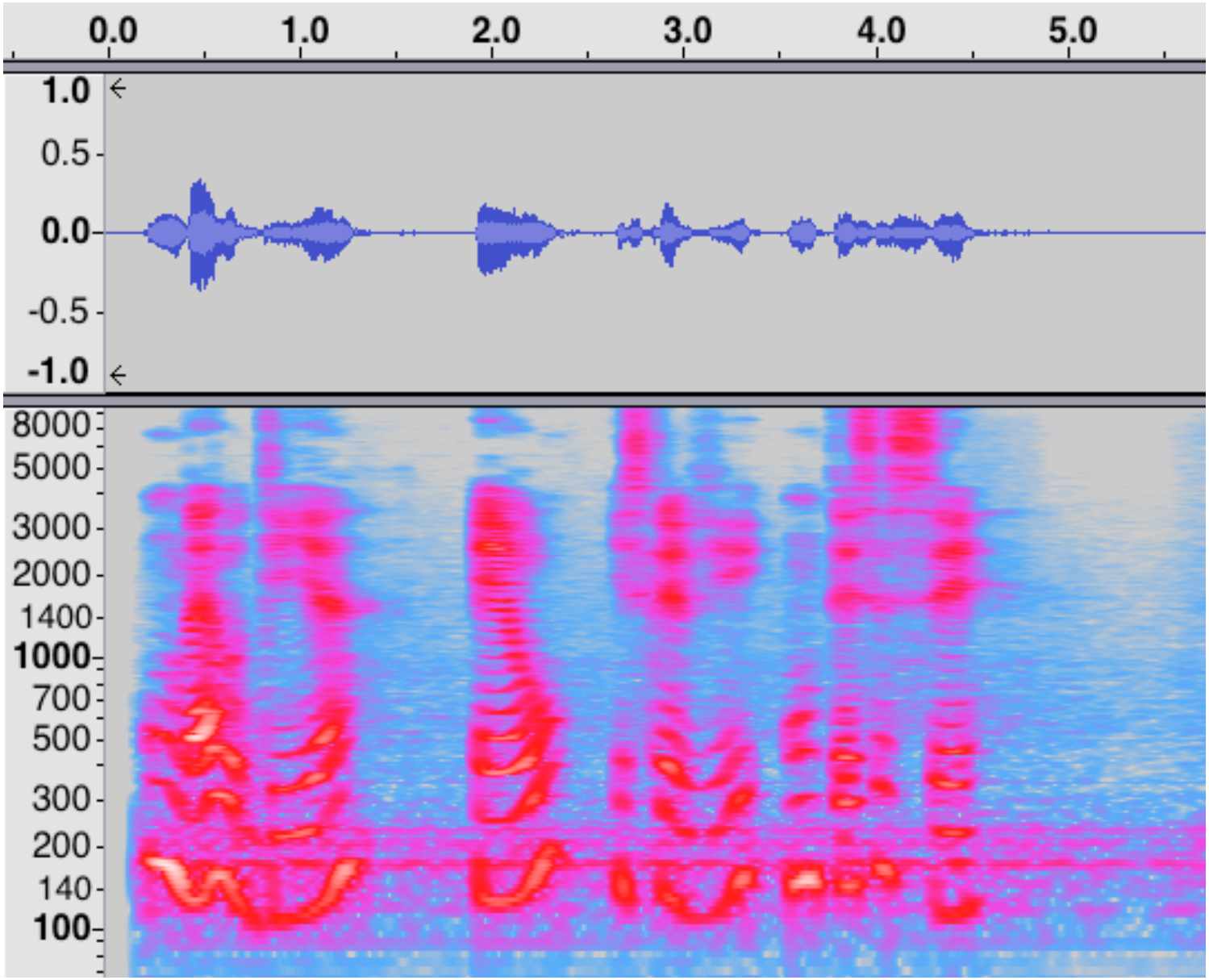
Motivação

- 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

Prosódia

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

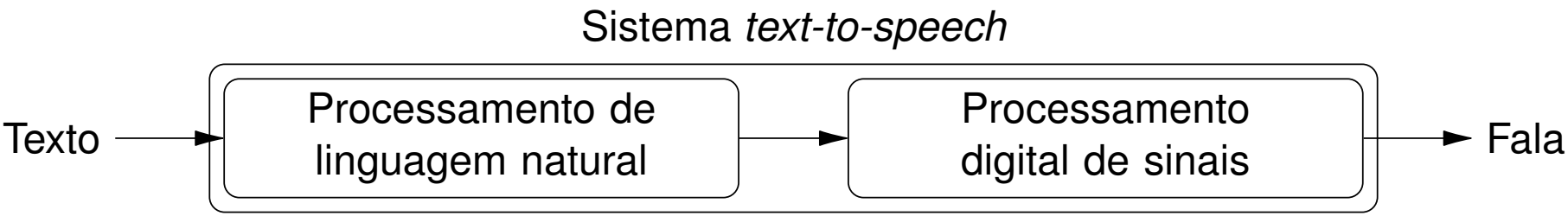
Prosódia


0.0
1.0
2.0
3.0
4.0
5.0
1.0
0.5
0.0
-0.5
-1.0
8000
5000
3000
2000
1400
1000
700
500
300
200
140
100

Função prosódica

- Suprassegmental
- Afetiva
- Aumentativa

Sistemas text-to-speech



Arquitetura

- 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

Front end

- Normalização de texto
  - A conta deu R$ 20, V. Exa. Pode conferir?
- Conversão grafema-fone
  - Gosto de pão
  - [gɔstu] (gósto)
  - [gostu] (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. . . . . !

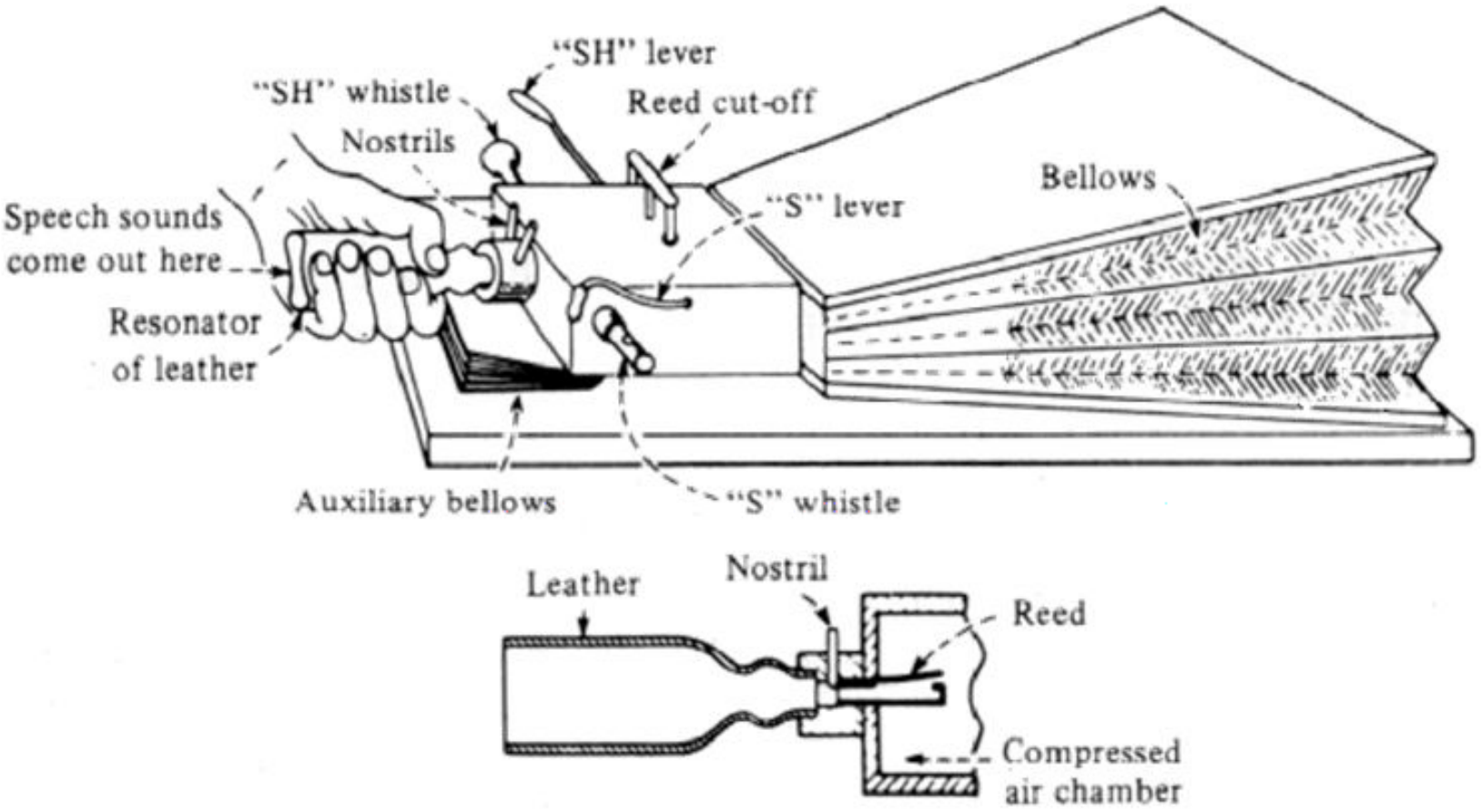
Geração de prosódia

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

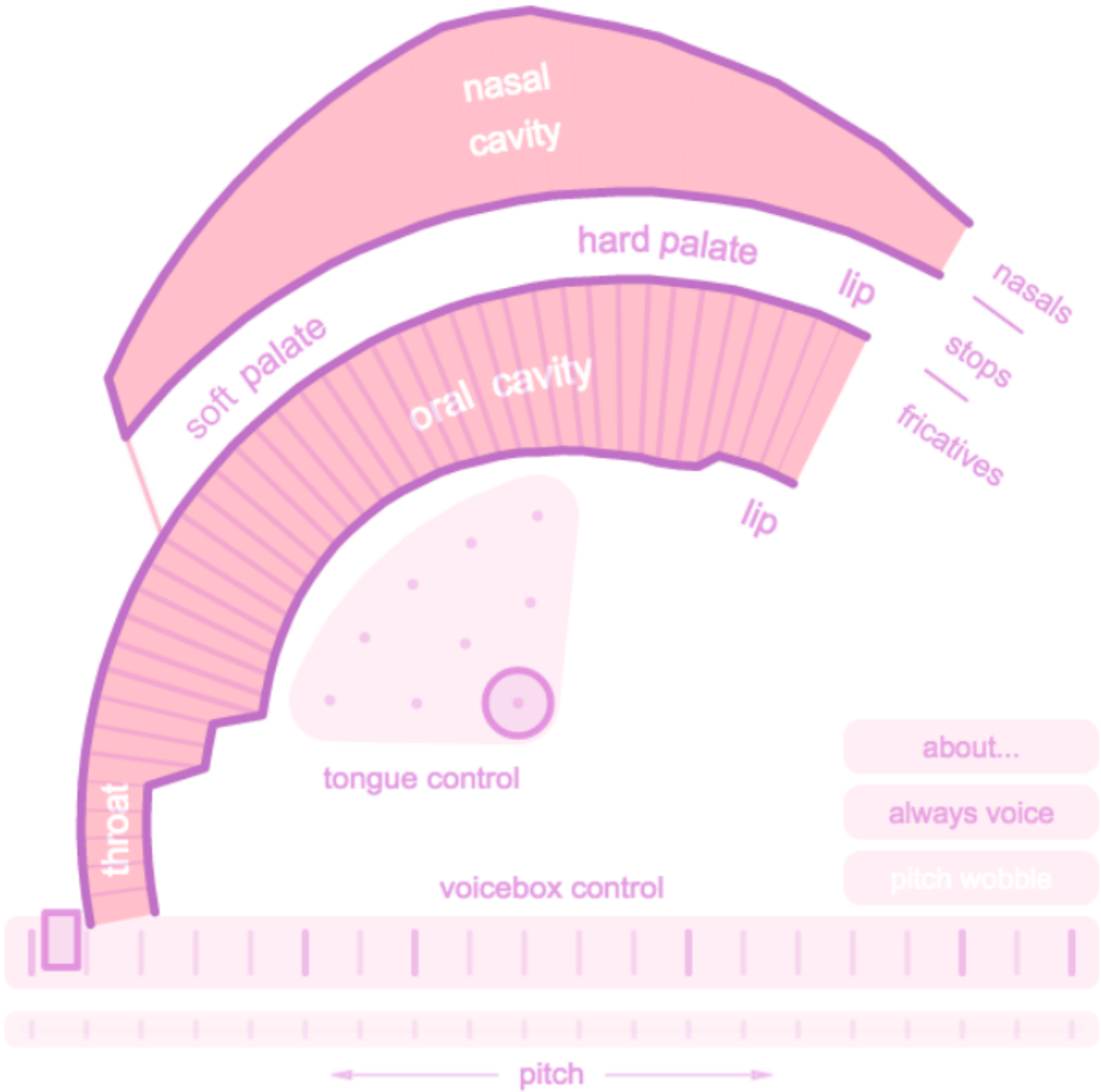
Back end

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

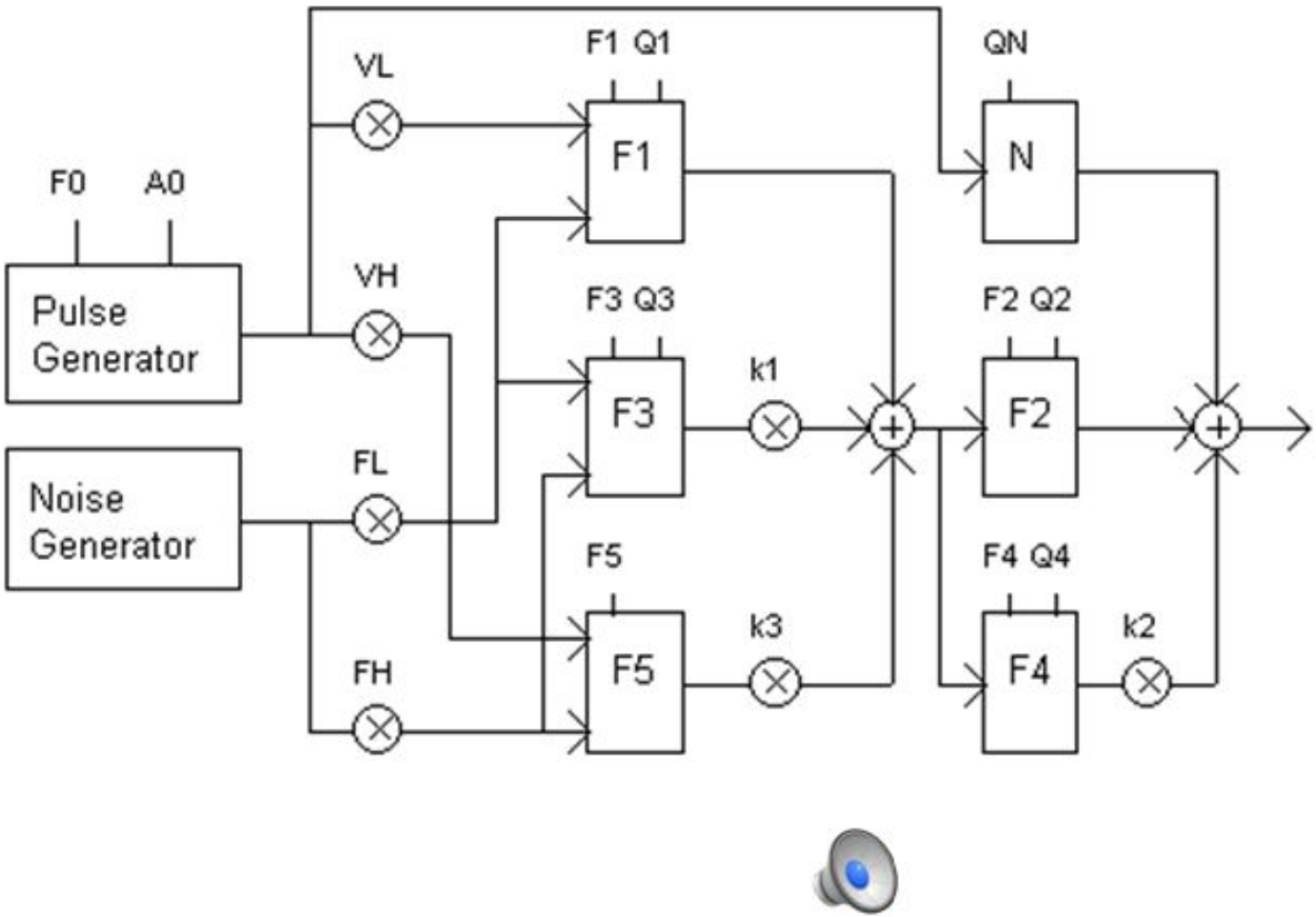
Máquina de Wolfgang von Kempelen (1778)



Articulatória

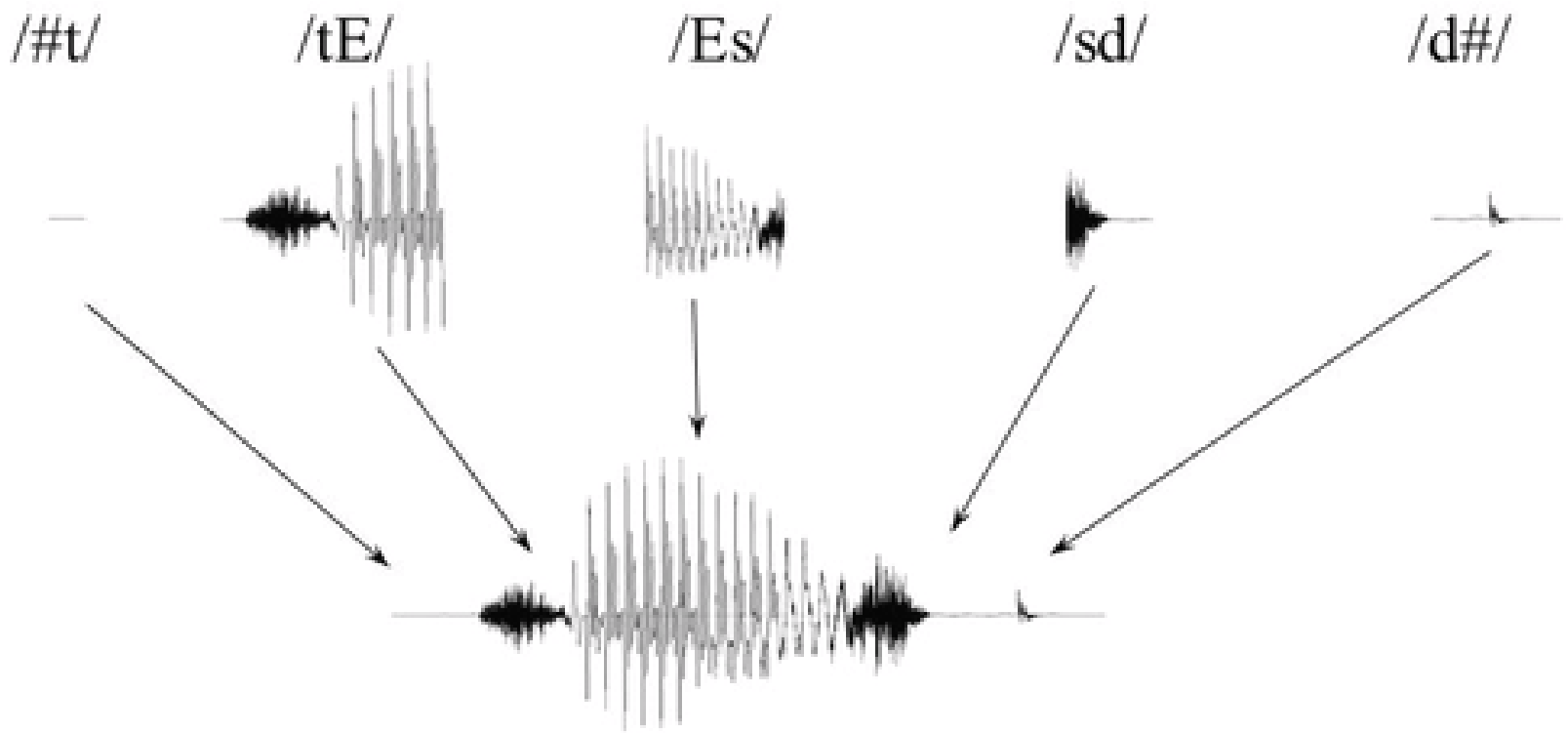


Formantes

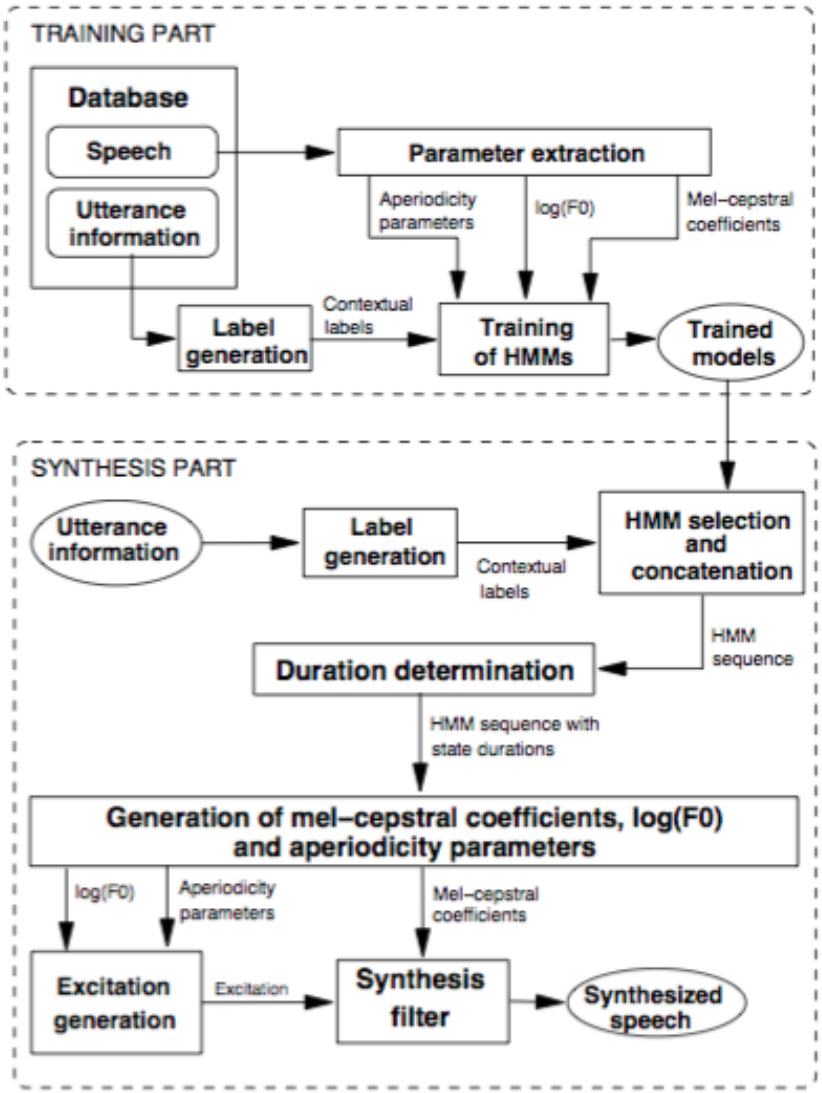

F1 Q1
QN
VL
F1
N
F0
A0
Pulse Generator
VH
F3 Q3
F2 Q2
k1
F3
F2
Noise Generator
FL
F5
F4 Q4
k3
k2
F5
F4
FH

Concatenativa

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



HMM


TRAINING PART
Database
Speech
Utterance information
Parameter extraction
Aperiodicity parameters
log(F0)
Mel–cepstral coefficients
Label generation
Contextual labels
Training of HMMs
Trained models
SYNTHESIS PART
Utterance information
Label generation
Contextual labels
HMM selection and concatenation
HMM sequence
Duration determination
HMM sequence with state durations
Generation of mel–cepstral coefficients, log(F0) and aperiodicity parameters
log(F0)
Aperiodicity parameters
Mel–cepstral coefficients
Excitation generation
Excitation
Synthesis filter
Synthesized speech

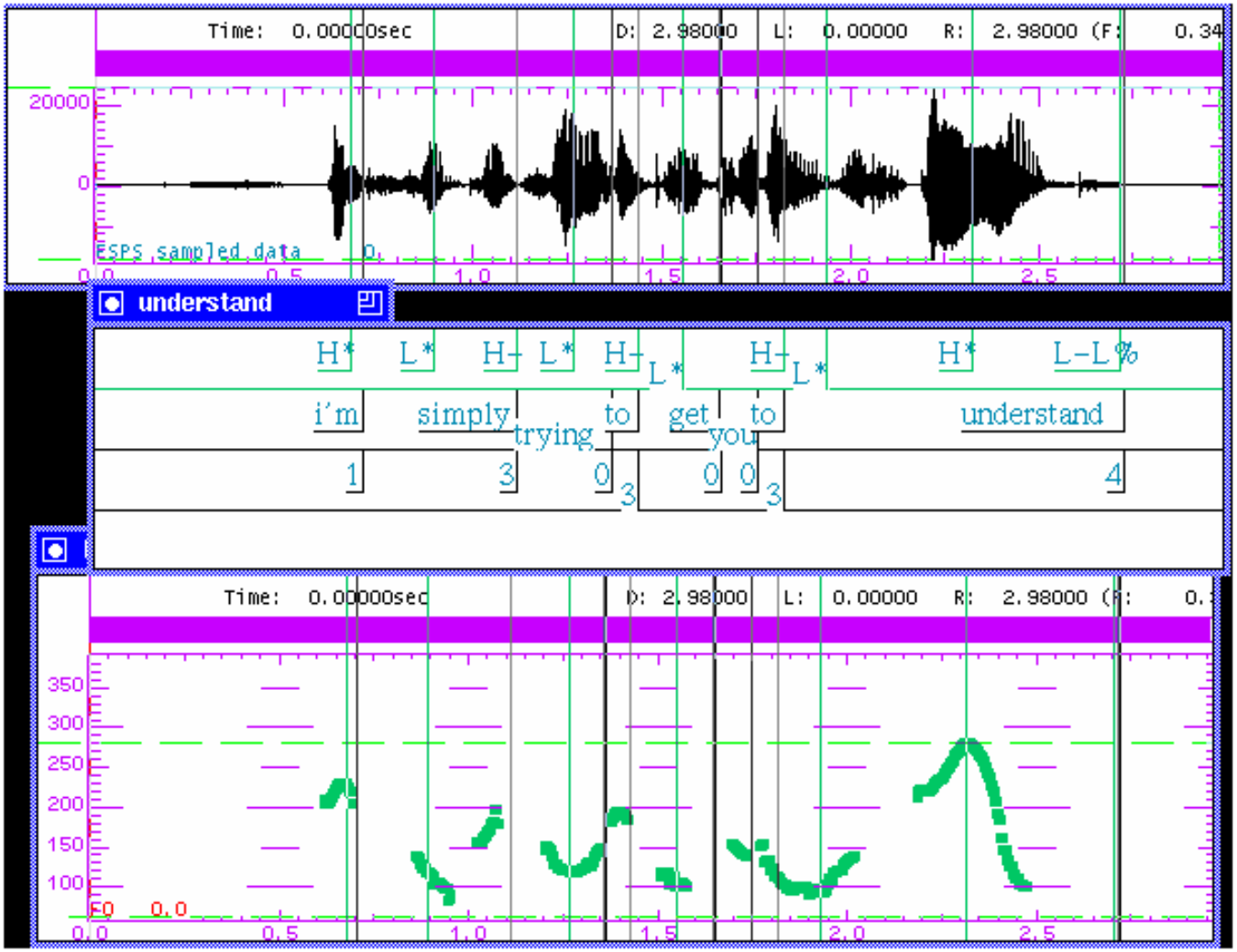
Sistemas TTS para o português brasileiro

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

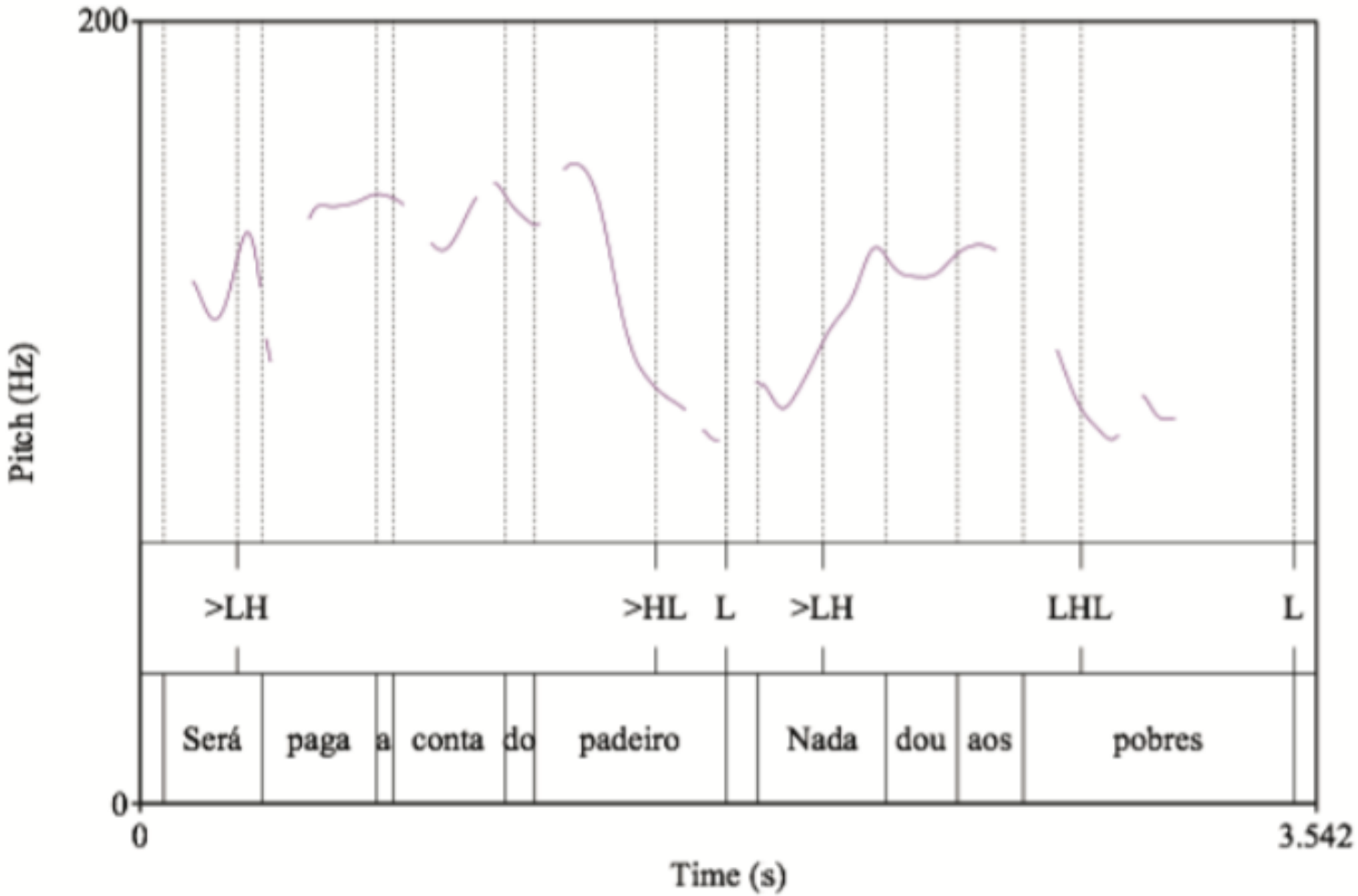
Analisando prosódia

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

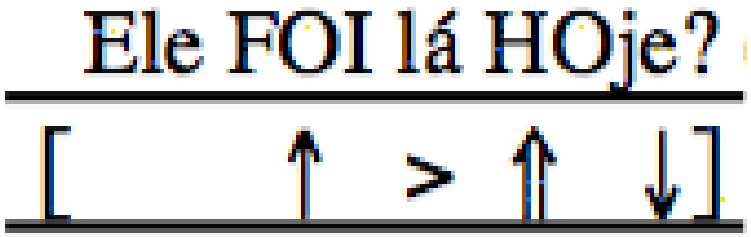
ToBI


Time: 0.00000sec
D: 2.98000
L: 0.00000
R: 2.98000
20000
0
ESPS sampled data
understand
H* L* H- L* H- L* H- L*
i'm simply trying to get to understand
1 3 0 3 0 0 3 4
350
300
250
200
150
100
F0

DaTo


200
Pitch (Hz)
0
>LH
>HL
L
>LH
LHL
L
Será
paga
a
conta
do
padeiro
Nada
dou
aos
pobres
0
3.542
Time (s)

INTSINT



Complementando texto

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

SSML

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

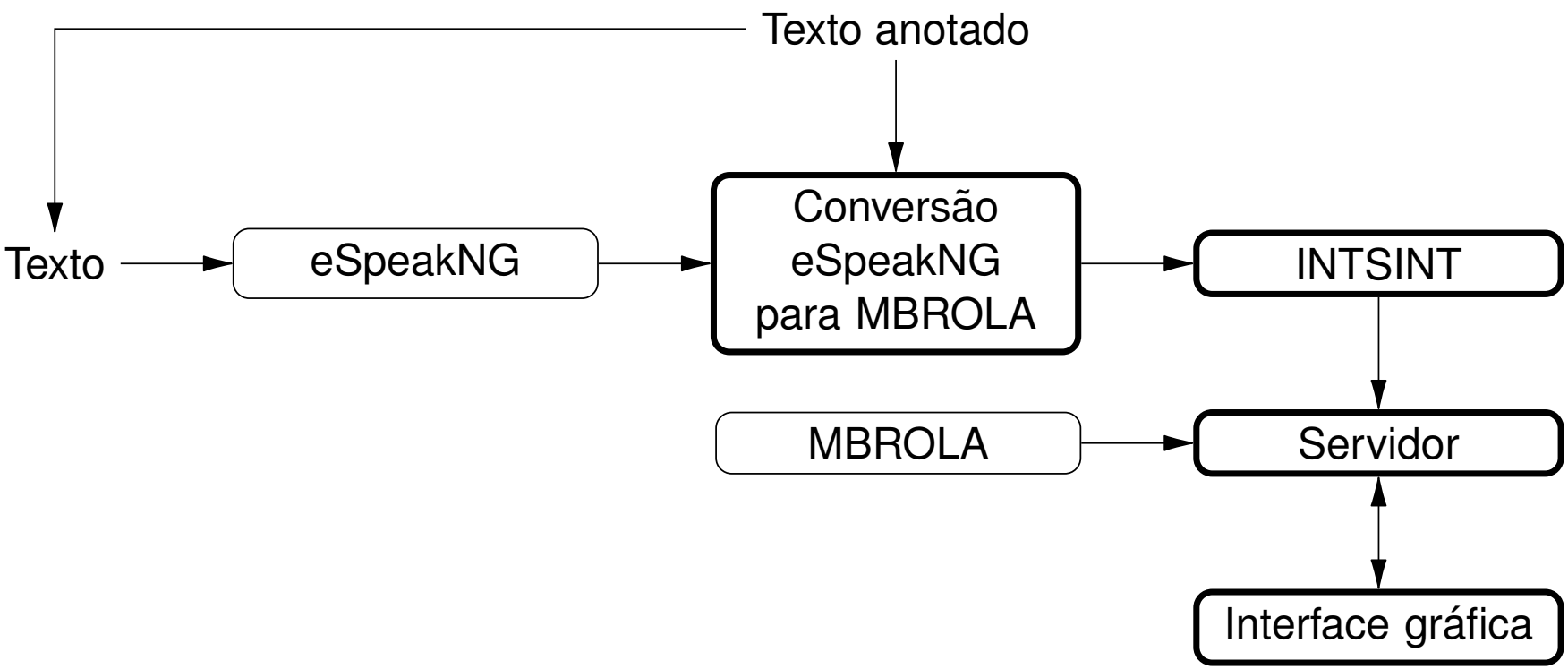
EmotionML

```
<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>
```

Anotações manuais

```
(Utterance Words
 (The
  (boy ((accent L*)))
  saw
  the
  (girl ((accent H*) (tone L-)))
  with
  the
  (telescope ((accent H*) (tone H-H%))))))
```

Arquitetura



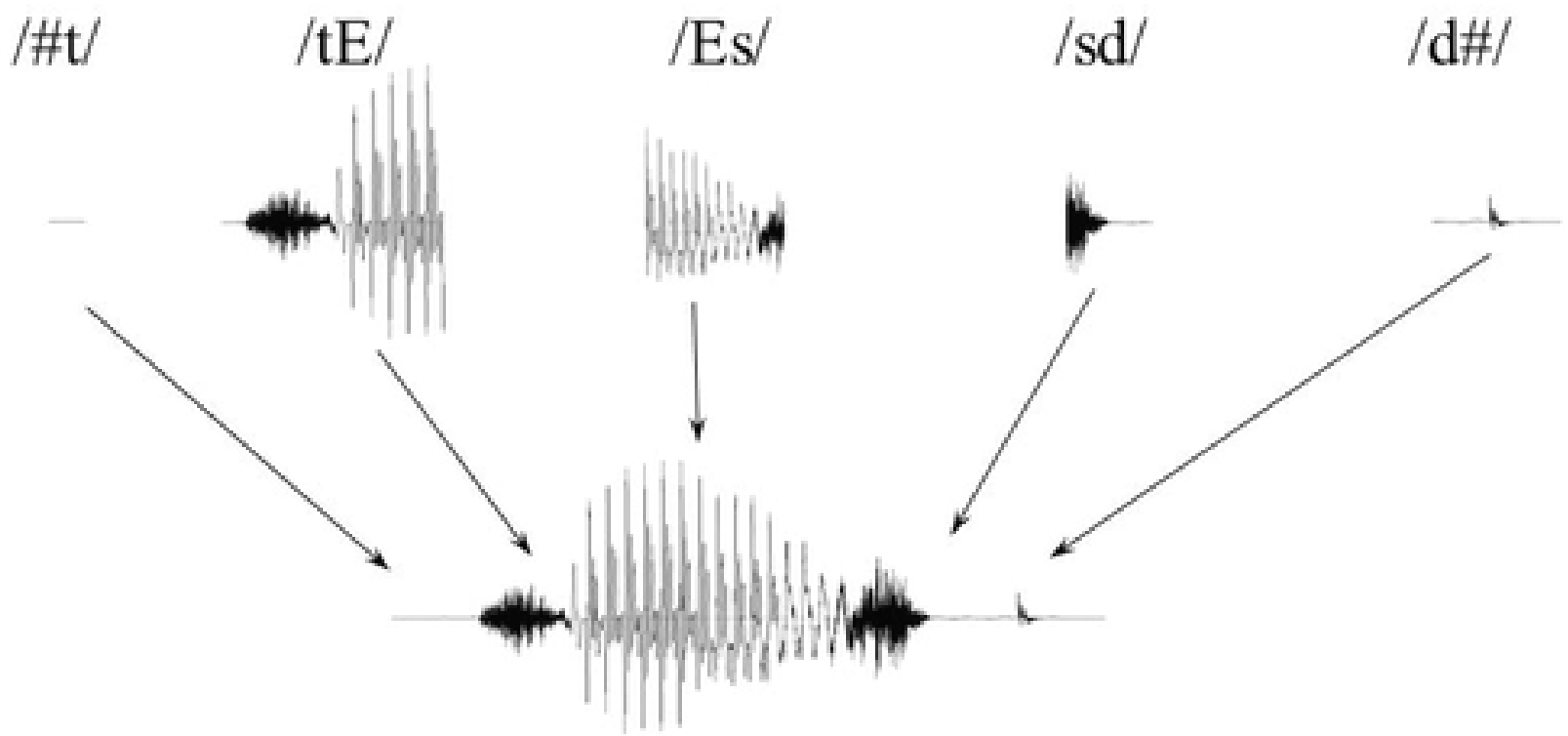
Regra	Cálculo
Top	key $\times \sqrt{2^{range}}$
Middle	key
Bottom	key$/\sqrt{2^{range}}$
Higher	$\sqrt{P_{i-1} \times T}$
Same	P_{i-1}
Lower	$\sqrt{P_{i-1} \times B}$
Upstepped	$\sqrt{P_{i-1} \times \sqrt{P_{i-1} \times T}}$
Downstepped	$\sqrt{P_{i-1} \times \sqrt{P_{i-1} \times B}}$

MBROLA

```
_ 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!

Trabalhos futuros

- 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

Perguntas

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