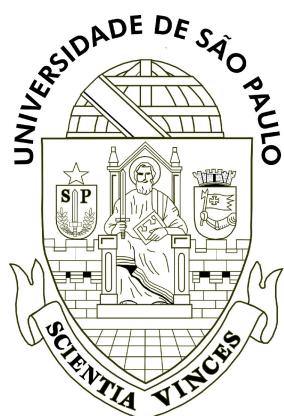


# Tools and resources for non-native speech recognition



**Gustavo Augusto de Mendonça Almeida**

Instituto de Ciências Matemáticas e de Computação  
Universidade de São Paulo

This dissertation is submitted for the degree of  
*Master of Sciences*

January 2016



To the loving memory of my father,

*Tarcízio Otávio Almeida.*

1947 – 2003





## **Declaration**

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Gustavo Augusto de Mendonça Almeida  
January 2016



## **Acknowledgements**

And I would like to acknowledge ...



## **Abstract**

Pesquisas recentes têm avaliado o Brasil entre os países com menor nível de proficiência em língua inglesa. Este projeto busca criar um recurso que possa contribuir para a melhoria desse cenário. O objetivo é desenvolver um reconhecedor de pronúncia para falantes do português brasileiro (PB) aprendizes de inglês, chamado Listener, que seja capaz de fornecer ao usuário feedback sobre sua pronúncia. Recursos semelhantes já foram desenvolvidos para outras línguas, no entanto, para o PB, há ainda uma lacuna a ser explorada. A hipótese de pesquisa é que é possível construir tal reconhecedor de pronúncia através de: (i) uma classificação de erros de pronúncia que leve em conta a transferência de padrões de L1 para L2; (ii) um modelo acústico que agregue dados de fala do inglês tanto de nativos, quanto de aprendizes; (iii) um dicionário de pronúncia que contenha a transcrição das pronúncias desviantes do aprendiz; e (iv) um modelo de língua que condiga com a sintaxe do aprendiz. Nove erros de pronúncia foram selecionados para serem tratados pelo Listener, assumindo-se, como pronúncia padrão, o General American (GA). A engine Julius será empregada como base do reconhecedor. O modelo acústico será compilado a partir de um corpus de fala de nativos de inglês: TIMIT Acoustic-Phonetic Continuous Speech Corpus[1]; e outro de aprendizes: COBAI - Corpus Oral Brasileiro de Aprendizes de Inglês[2]. O dicionário a ser empregado é o CMU Pronouncing Dictionary, ao qual serão acrescentaremos as hipóteses de pronúncia dos aprendizes, por meio de regras. O modelo de língua será gerado a partir da Simple English Wikipedia em conjunto com um corpus de textos escritos por aprendizes de inglês, o COMApred[3], um dos três corpus do projeto COMET da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. A eficiência do reconhecedor será avaliada por meio de medidas de Word Error Rate (WER), Character Error Rate (CER) e matrizes de confusão. O reconhecedor proposto visa a propiciar a criação de sistemas de treino de pronúncia mediado por computador. De modo a verificar a viabilidade do método ora proposto, um protótipo do reconhecedor foi elaborado e avaliado intrinsecamente. Um excerto do COBAI e de um corpus de erros induzidos especialmente coletado para este protótipo foram utilizados para alimentar o modelo acústico (~3h50min de fala). O protótipo foi desenvolvido para reconhecer erros relacionados à simplificação silábica. As variantes de pronúncia foram adicionadas no dicionário através de um conjunto de 20 regras. As taxas de

WER obtidas foram de 61% para um léxico de 5.768 entradas, e de 78% para alinhamento forçado, indicando, portanto, que o método é promissor....

# Table of contents

<b>List of figures</b>	<b>xiii</b>
<b>List of tables</b>	<b>xv</b>
<b>List of acronyms</b>	<b>xvii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Theoretical Foundations</b>	<b>11</b>
2.1 Phonetics and Phonology . . . . .	11
2.2 Second Language Acquisition . . . . .	19
2.3 Automatic Speech Recognition . . . . .	35
<b>3 Aieouadô’s dictionary and G2P converter</b>	<b>57</b>
3.1 Introduction . . . . .	57
3.2 Method . . . . .	60
3.3 Results . . . . .	62
3.4 Final Remarks . . . . .	64
3.5 Acknowledgements . . . . .	66
<b>4 Phonetic-based Speller</b>	<b>67</b>
4.1 Introduction . . . . .	67
4.2 Experimental Settings and Methods . . . . .	69
4.3 Discussion . . . . .	78
4.4 Related Work . . . . .	80
4.5 Final Remarks . . . . .	81
<b>5 A greedy algorithm for the extraction of phonetically rich sentences</b>	<b>83</b>
5.1 Introduction . . . . .	83

5.2 Related Work . . . . .	84
5.3 Example Evaluation . . . . .	87
5.4 Final Remarks . . . . .	93
5.5 Extracted Sentences Sample . . . . .	93
<b>6 Listener</b>	<b>99</b>
6.1 Tools and Libraries . . . . .	112
6.2 Speech Corpora . . . . .	112
6.3 Building the Acoustic Model . . . . .	122
6.4 Building the Pronunciation Model . . . . .	137
6.5 Building the Grammars and the Language Model . . . . .	138
6.6 Results . . . . .	138
<b>7 Conclusions</b>	<b>139</b>
7.1 Overall Conclusions . . . . .	139
7.2 Limitations . . . . .	141
7.3 Further Work . . . . .	141
<b>References</b>	<b>143</b>
<b>Appendix A How to install L<sup>A</sup>T<sub>E</sub>X</b>	<b>153</b>
<b>Glossary</b>	<b>157</b>

# List of figures

1.1	English Proficiency Index 2014 Rankings [40]. . . . .	2
1.2	EF EPI scores for each Brazilian state. . . . .	5
1.3	2013 Business English Index. . . . .	6
2.1	IPA Chart. . . . .	12
2.2	Brazilian Portuguese (BP) oral vowels. . . . .	15
2.3	BP nasal vowels. . . . .	15
2.4	Height (cm) versus vocal tract length (mm) [42]. . . . .	56
2.5	Averaged vocal tract morphology [42]. . . . .	56
2.6	F0 and pitch sigma versus age for males and females [16]. . . . .	56
2.7	Two complex waveforms generated by the same three pure tone 100 Hz, 200 Hz and 300 Hz sine waves, differing only with respect to their relative timing [68]. . . . .	56
2.8	Example of in-phase waves. . . . .	56
2.9	Example of out-of-phase waves. . . . .	56
2.10	Illustration of an original audio recording (the upper waveform) divided into two offset sequences of analysis windows (two lower waveforms) with 50% overlapping frames [80] . . . . .	56
2.11	Mel scale versus a linear frequency scale. . . . .	56
2.12	Example of the OroFacial Clone display, from Badin et al. [7] [7]. . . . .	56
3.1	<i>System architecture for building the pronunciation dictionary.</i> . . . . .	59
3.2	<i>Example of the transcription procedure – in grey: graphemes yet to be transcribed; in black: graphemes already transcribed.</i> . . . . .	61
4.1	Pseudocode: Method II - Benchmark . . . . .	71
4.2	Pseudocode: Method III - GPM . . . . .	73
4.3	<i>Architecture of the GPM-ML</i> . . . . .	74

5.1	Pseudocode: Method for extracting phonetically-rich sentences. . . . .	87
6.1	Entry example in the Oxford Dictionary online. . . . .	119
6.2	Phonetic decision tree for HMM state tying [133]. . . . .	133
6.3	Tied-state HMM system build procedure [133]. . . . .	133
6.4	Distinctive features for places of places of articulation [62]. . . . .	133

# List of tables

1.1	CEFR reference levels.	4
2.1	BP consonants . . . . .	15
2.2	Examples of plosive consonants in Brazilian Portuguese (I). . . . .	15
2.3	Examples of plosive consonants in Brazilian Portuguese (I). . . . .	16
2.4	Examples of affricate consonants in Brazilian Portuguese. . . . .	17
2.5	Examples of nasal consonants and nasalized vowels in Brazilian Portuguese. . . . .	17
2.6	Examples of rhotics in Brazilian Portuguese. . . . .	18
2.7	Word error rate comparisons between human and machines on similar tasks [60]. . . . .	42
3.1	<i>Portuguese Wikipedia Summary – Dumped on 23<sup>rd</sup> January 2014.</i> . . . . .	62
3.2	<i>Results from the Language Identifier module – Training Phase.</i> . . . . .	63
3.3	<i>Results from the Language Identifier module – Wikipedia word list.</i> . . . . .	63
3.4	<i>Results from the Transcriber – Training Phase.</i> . . . . .	65
4.1	<i>Inter-rater agreement for the error detection task</i> . . . . .	76
4.2	<i>Error distribution in corpus by category</i> . . . . .	78
4.3	<i>Comparison of the Methods</i> . . . . .	78
4.4	<i>Comparison of Correction Rates</i> . . . . .	79
5.1	Portuguese Wikipedia Summary – Dumped on 23 <sup>rd</sup> January 2014. . . . .	89
5.2	Sentences' summary after WikiExtractor and Punkt. . . . .	89
5.3	Sentences' summary after the length filter. . . . .	89
5.4	Triphone results from the extraction of sentences through our method. . . . .	92
5.5	Triphone results from the sentences taken randomly. . . . .	92
6.1	Summary of the entire WSJ0 Corpus. . . . .	114
6.2	Summary of WSJ0 Part We Used. . . . .	114
6.3	Summary of Listener's Corpus. . . . .	116

6.4	Oxford Dictionary phone convention. . . . .	118
6.5	Summary of the Oxford Dictionary AmE Corpus. . . . .	119
6.6	Summary of Listener's Corpus. . . . .	121
6.7	CMUdict phone convention. . . . .	123
6.8	Aeiouadô phone convention. . . . .	124
6.9	Interlingual dictionary phone convention. . . . .	127
6.10	Examples of final-obstruent devoicing in German. . . . .	133
6.11	Distinctive features chart for AmE phones [62]. . . . .	134
6.12	Distinctive features chart for BP phones [62]. . . . .	135
6.13	Distinctive features chart for Listener phones. . . . .	136

# List of acronyms

**AmE** American English. 117, 122, 125, 128, 132

**API** Application Programming Interface. xvii, *Glossary: Application Programming Interface*

**ASCII** American Standard Code for Information Interchange. 122

**ASR** Automatic Speech Recognition. 7, 32, 41, 43, 44, 110, 111, 122, 126, 128, 137

**ATR** Advanced Tongue Root. 131

**BEI** Business English Index. 6

**BP** Brazilian Portuguese. 6–8, 15, 122, 125, 126, 132

**CAPT** Computer Assisted Pronunciation Training. 6, 7, 47

**CEFR** Common European Framework of Reference. 3, 5

**CG** Computer Graphics. 47

**CMUDict** Carnegie Mellon University Pronouncing Dictionary. 122, 125, 128

**CSR** Continuous Speech Recognition. 112

**CT** Computer Tomography. 47

**DNN** Deep Neural Network. 41

**EF** Education First. 1

**EF-EPI** EF English Proficiency Index. 1, 3

**ESL** English as a Second Language. 7, 122

**F0** Fundamental Frequency. 40

**G2P** Grapheme-to-Phoneme. 8

**GMM** Gaussian Mixture Model. 9, 33

**HDI** Human Development Index. 1, 6

**HMM** Hidden Markov Model. 8, 32, 33, 41, 113, 122, 125, 126, 128, 130

**HSP** Heightened Subglottal Pressure. 131

**HTK** Hidden Markov Model Toolkit. 41

**IPA** International Phonetic Alphabet. 11, 12, 117, 125

**L1** First or Native Language. 16

**L2** Second Language. 16

**MFCC** Mel Frequency Cepstral Coefficients. 41, 43, 44

**MRI** Magnetic Resonance Imaging. 47

**PCM** Pulse Code Modulation. 42

**PER** Phone Error Rate. 111

**PLP** Perceptual Linear Prediction. 43, 44

**POS** Part of Speech. 117

**RASR** RASR. 41

**regex** Regular Expression. 137, 138

**RTF** Real Time Factor. 111

**SNR** Signal-to-Noise Ratio. 117

**UML** Unified Modeling Language. 31, 45

**VBR** Variable Bit Rate. 117

**WER** Word Error Rate. 110, 111

# **Chapter 1**

## **Introduction**

Data from the World Economic Outlook, of the IMF (2013), list currently Brazil as the seventh largest economy in the world, with a GDP of US\$ 2,396 trillions. According to a survey by The Economist (2013), since 2009, the growth of BRICS (emerging market economies group formed by Brazil, Russia, India, China and South Africa) accounts for 55% of the world economy growth. The current economic scenario is extremely favorable for Brazil to increase its global influence.

However, with regard to the ability to communicate globally, we occupy a much more modest position. In the EF English Proficiency Index (EF-EPI) of 2014, which is published by Education First (EF), Brazil was ranked in the 38<sup>th</sup> position out of 63 countries, classified among countries with low English proficiency, with 49.96 points [40]. The full ranking is shown in Figure 1.1.

As can be noticed from Figure 1.1, Brazil was immediately behind two other countries from the BRICS, Russia (50.44) and China (50.15); and near several other Latin America countries, such as Peru (51.46), Ecuador (51.05), Uruguay (49.61), Chile (48.75) and Colombia (48.54).

Scandinavian countries lead the very high proficiency rankings, with Denmark (69.30) in the first position, Sweden (67.30) in third the spot, Finland (64.40) in the fourth and Norway (64.33) in the fifth. All these countries have a very high Human Development Index (HDI), according to the United Nations Development Programme. The HDI measures the social and economic development by analyzing three indexes: educational level, average income and longevity. The top five in the EF-EPI rankings are among the top twelve countries in terms of HDI [120]. The EF-EPI data showed that there is a moderate to strong correlation between HDI and English proficiency ( $R=0.67$ ) [40]. The second position in the EF-EPI ranking is held by the Netherlands, scoring 68.99 points. The Netherlands has the fourth largest HDI

<b>VERY HIGH PROFICIENCY</b>			<b>LOW PROFICIENCY</b>		
01	Denmark	69.30	32	U.A.E.	51.80
02	Netherlands	68.99	33	Vietnam	51.57
03	Sweden	67.80	34	Peru	51.46
04	Finland	64.40	35	Ecuador	51.05
05	Norway	64.33	36	Russia	50.44
06	Poland	64.26	37	China	50.15
07	Austria	63.21	38	Brazil	49.96
<b>HIGH PROFICIENCY</b>			39	Mexico	49.83
08	Estonia	61.39	40	Uruguay	49.61
09	Belgium	61.21	41	Chile	48.75
10	Germany	60.89	42	Colombia	48.54
11	Slovenia	60.60	43	Costa Rica	48.53
12	Malaysia	59.73	44	Ukraine	48.50
13	Singapore	59.58	<b>VERY LOW PROFICIENCY</b>		
14	Latvia	59.43	45	Jordan	47.82
15	Argentina	59.02	46	Qatar	47.81
16	Romania	58.63	47	Turkey	47.80
17	Hungary	58.55	48	Thailand	47.79
18	Switzerland	58.29	49	Sri Lanka	46.37
<b>MODERATE PROFICIENCY</b>			50	Venezuela	46.12
19	Czech Republic	57.42	51	Guatemala	45.77
20	Spain	57.18	52	Panama	43.70
21	Portugal	56.83	53	El Salvador	43.46
22	Slovakia	55.96	54	Kazakhstan	42.97
23	Dominican Republic	53.66	55	Morocco	42.43
24	South Korea	53.62	56	Egypt	42.13
25	India	53.54	57	Iran	41.83
26	Japan	52.88	58	Kuwait	41.80
27	Italy	52.80	59	Saudi Arabia	39.48
28	Indonesia	52.74	60	Algeria	38.51
29	France	52.69	61	Cambodia	38.25
			62	Libya	38.19

in the world. This result is similar to that of the previous versions of EF-EPI [39, 38, 37], which also showed a prevalence of Northern European countries in the best positions.

It is interesting to notice that four of the countries at the top five share a common cultural Germanic heritage and speak a language of the Germanic family – the same branch that English is part of. Since Danish, Swedish, Norse, Dutch and English are related through descent from a common ancestor, Proto-Germanic, they inevitably share many linguistic characteristics, such as sound correspondences, a large number of cognates, similar morphology and syntax, etc. This might be an advantage of these countries in comparison to the remaining.

The high proficiency band is occupied mainly by other European countries, both in the Eastern and Western part, such as Estonia (61.39), Belgium (61.21), Germany (60.89), Slovenia (60.60), Latvia (59.43) and Switzerland (58.29). Two Southeast Asian countries also figure within this range, namely Malaysia (59.73) and Singapore (59.58). This result might seem a bit biased since English is one of the official languages in Singapore, along with Malay, Mandarin and Tamil; it is considered the language of business, government, and the medium of instruction in school. We shall highlight Argentina's performance. Despite the great economic depression from 1998 to 2002, Argentina was still able to outperform Brazil, scoring 59.02 points, being the only country from Latin America among the ones with high proficiency.

The moderate proficiency range is filled by the remaining European countries, such as Czech Republic (57.42), Slovakia (55.96), the countries from the Iberian Peninsula – Spain (57.18) and Portugal (56.83)–, together with France (52.69) and Italy (52.80). In addition, the majority of Asian countries which were analyzed in the survey also figure in this list. South Korea achieved the best performance among them, with 53.62 points. India comes next, scoring 53.54 points. The rest of the list is occupied by Japan (52.88), Indonesia (52.74), Taiwan (52.56) and Hong Kong (52.50).

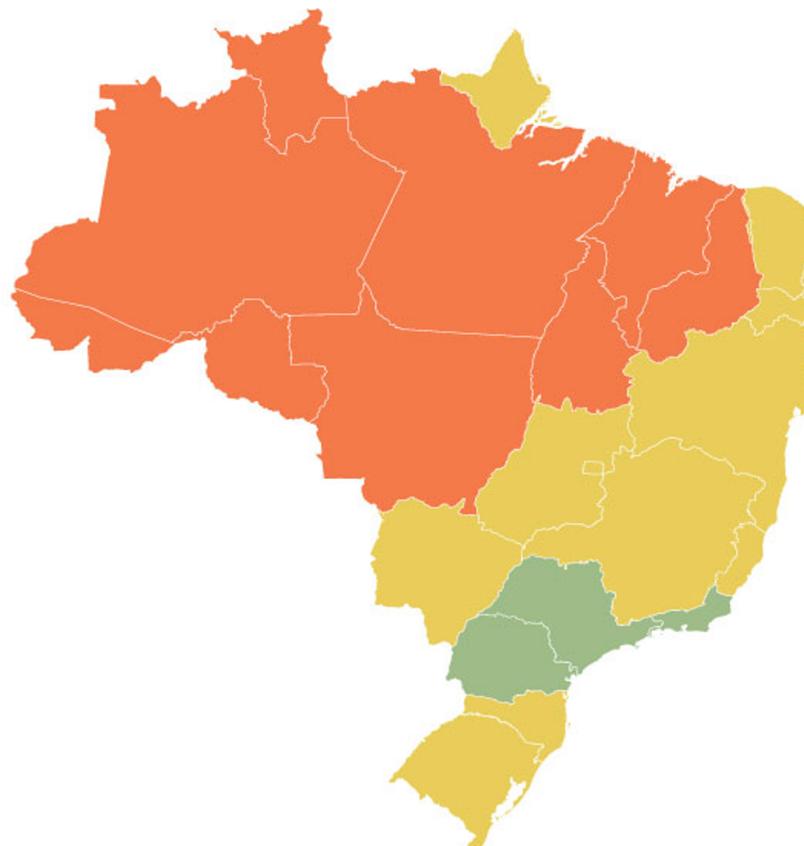
The EF-EPI bands are aligned to the Common European Framework of Reference (CEFR), which is a guideline proposed by the Council of Europe to describe achievements of learners of foreign languages across the European Union. The CEFR reference levels are described in Table 1.1. EF-EPI bands are mapped into CEFR reference levels as follows: the very high proficiency band corresponds to CEFR level B2; very low proficiency to A2; high, moderate and low proficiency bands to B1 with different punctuations.

In case, Brazil's low proficiency rank is analogous to the CEFR B1 level. To put another way, it means that Brazilians are usually able to communicate in English with intermediate skills, being able to understand familiar matters, deal with traveling situations, describe personal experiences and plans, and produce simple texts about subjects of personal interest.

Table 1.1 CEFR reference levels.

<b>Group</b>	<b>Level</b>	<b>Description</b>
<b>Basic User (A)</b>	<b>Beginner (A1)</b>	<p>Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type.</p> <p>Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has.</p> <p>Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.</p>
	<b>Elementary (A2)</b>	<p>Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment).</p> <p>Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters.</p> <p>Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.</p>
<b>Independent User (B)</b>	<b>Intermediate (B1)</b>	<p>Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.</p> <p>Can deal with most situations likely to arise while traveling in an area where the language is spoken.</p> <p>Can produce simple connected text on topics that are familiar or of personal interest.</p> <p>Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.</p>
	<b>Upper intermediate (B2)</b>	<p>Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialization.</p> <p>Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party.</p> <p>Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.</p>
<b>Proficient User (C)</b>	<b>Advanced (C1)</b>	<p>Can understand a wide range of demanding, longer texts, and recognize implicit meaning.</p> <p>Can express ideas fluently and spontaneously without much obvious searching for express</p> <p>Can use language flexibly and effectively for social, academic and professional purposes.</p> <p>Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organizational patterns, connectors and cohesive devices.</p>
	<b>Proficiency (C2)</b>	<p>Can understand with ease virtually everything heard or read.</p> <p>Can summarize information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation.</p> <p>Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in the most complex situations.</p>

As one might observe this is a very restricted communicative competence, which limits English usage basically to the personal domain. To get an idea, the CEFR lists four broad domains: educational, occupational, public, and personal. So Brazilians lack linguistic abilities in at least three broad domains, this linguistic competence would not allow one to perceive or produce English utterances flexibly, either for social, academic or professional purposes. The performance for each state can be found at the map in Figure 1.2.



- Very High Proficiency
- High Proficiency
- Moderate Proficiency
- Low Proficiency
- Very Low Proficiency

Fig. 1.2 EF EPI scores for each Brazilian state.

As one might see from Figure 1.2, one can clearly see division between Southern and Northern Brazil. Ten states achieved very low proficiency, namely Amazonas, Acre, Pará,

Roraima, Piauí, Alagoas, Tocantins, Rondônia, Maranhão and Mato Grosso. The highest scores are found among the richest Brazilian states, mainly in the Southeast and the South regions. The top position is held by São Paulo, which has the largest population, the highest number of industries, and most part of the economic production in the country. Rio de Janeiro, which is the second largest economy of Brazil, occupies the second position. Paraná is the third best ranked state. These three states are the only in the country that achieved moderate proficiency, they attained a score that is at least 4.8% above the country's average, and all of them are among those with the highest HDI.

With respect of Business English proficiency, our performance is even more concerning. On the 2013 Business English Index (BEI), conducted by GlobalEnglish [51], Brazil reached the 71<sup>st</sup> position out of 77 countries analyzed. The ranking is described in Figure 1.3. Brazil attained a score of 3.27 points, in a scale from 1 to 10, being placed at the "Beginner" range, the lowest proficiency range considered by the index. Brazil's performance was close to that of El Salvador (3.24), Saudi Arabia (3.14) and Honduras (2.92) which up until recently had experienced civil wars or dictatorship governments. The "Beginner" is described as:

**Beginner:** Can read and communicate using only simple questions and statements, but can't communicate and understand basic business information during phone calls.

Again, we can see that this is a very limited linguistic competence, that would not allow one not even to perform the most elementary day-to-day tasks in a company or industry work environment.

Fig. 1.3 2013 Business English Index.

Given this scenario, it is clear that we desperately need to improve English language proficiency among Brazilians. This project seeks to be an initial step in that direction. We proposed a method for building a non-native speech recognizer and checker, called Listener, which is capable of recognizing utterances in Brazilian-accented English and identifying which are the mispronunciations. Our system might be used to build Computer Assisted Pronunciation Training (CAPT) systems for Brazilian-accented English. We also delivered a prototype CAPT system to illustrate how it might work.

Similar tools have been developed for other languages, such as Japanese [118], Spanish [99], Dutch [113] and French [? ]. However for BP, to the best of our knowledge, there is still a gap to be explored.

CAPT systems can provide new opportunities for practicing oral proficiency. With respect to L2 pronunciation, it has been shown that when students have online corrective feedback about their mispronunciations, they can improve more quality their pronunciation skills

[75, 76]. Nonetheless, in the classroom, teachers are not always able to provide this type of individual feedback given the number of students in a class and the amount of available hours in an English as a Second Language (ESL) course. CAPT systems can be a solution to assist with this matter, since they can assess one's pronunciation in real time at very low cost. According to Neri et al. [86], by using CAPT, students are able to improve their pronunciation skills at a similar pace to those who attend traditional classrooms.

**Research hypotheses** The research hypotheses are:

About the acoustic model:

1. cross-linguistic acoustic models, estimated from native data – in case, BP and English – are suitable for recognizing Brazilian-accented English;
2. a small amount of data from Brazilian-accented English might be used, through speaker adaptation, to increase the recognition performance;
3. phonetically rich corpora are more appropriate for building acoustic models which will be used for forced alignment tasks than balanced corpora.

About the pronunciation dictionary:

1. Brazilian-accented negative transfer phenomena can be represented through transformation rules;
2. non-native pronunciation variants can be handled similarly to native variation, by using a multi-pronunciation dictionary;
3. phones which are acoustically similar in English and BP might be merged in the pronunciation dictionary to improve the recognition performance.

About the speech recognizer and checker:

1. an HMM/GMM Automatic Speech Recognition (ASR), with acoustic parameters estimated from a cross-linguistic corpora (from both English and BP), with a multi-pronunciation dictionary which include Brazilian-accented variants created through transformation rules, is an adequate method for recognizing and identifying mispronunciations in forced alignment tasks containing Brazilian-accented English.

**Contributions of the Thesis** Within this work, I have investigated and developed a set of tools and resources for non-native speech recognition and correction, focused on Brazilian-accented English. Some of these resources and tools are exclusively for tasks related to processing Brazilian-accented English, but others are more general and can be employed for many other purposes, such as a Grapheme-to-Phoneme (G2P) for BP, corpus balancing, etc. The full list of contributions is provided below:

1. *Aeiouadô G2P*: A grapheme-to-phoneme converter for BP which uses a hybrid approach, based on both handcrafted rules and machine learning method, as described in Mendonça and Aluísio [83]. *Aeiouadô dictionary*: A large machine readable dictionary for BP, compiled from a word list extracted from the Portuguese Wikipedia, which was preprocessed in order to filter loanwords, acronyms, scientific names and other spurious data, and then transcribed with Aeiouadô G2P).
2. A phonetic speller for user-generated content in BP, based on machine learning, which takes advantage of Aeiouadô G2P to group phonetically related words, as described in Mendonça et al. [84];
3. A method for the extraction of phonetically rich sentences, i.e. sentences with a high variety of triphones distributed in a uniform fashion, which employs a greedy algorithm for comparing triphone distributions among sentences, as described in Mendonça et al. [85];
4. A crowdsourced platform for speeding up the process of compiling and transcribing speech corpora;
5. A set of rules for generating pronunciation hypothesis for Brazilian-accented English, considering nine types of mispronunciations, respectively: (i) syllable simplification; (ii) consonant change; (iii) deaspiration of voiceless plosives in initial or stressed positions; (iv) terminal devoicing in word-final obstruents; (v) delateralization and rounding of lateral liquids in final position; (vi) vocalization of final nasals; (vii) velar consonantal paragoge; (viii) vowel assimilation; (ix) interconsonantal epenthesis;
6. An alignment method for phonetic transcriptions, based on the Needleman-Wunsch algorithm, which takes into the phones' distinctive features for finding the optimal phone sequence match;
7. *Listener*: A prototype system for automatic speech recognition and evaluation of Brazilian-accented English, which makes use of forced alignment, Hidden Markov

Model (HMM)/Gaussian Mixture Model (GMM) acoustic models, context free grammars and multipronunciation dictionaries;

All files, resources and scripts developed are available at the project website<sup>1</sup>:

<http://nilc.icmc.usp.br/listener>

**Thesis Structure** This Master’s thesis is organized into parts, chapters and sections according to the following. ?? presents some basic concepts and notions which are necessary for further reading. ?? discusses the fundamentals of Phonetics and Phonology. ?? debates Second Language Acquisition, which is the area of Applied Linguistics responsible for studying how one acquires an additional language and what are the implications throughout the learning process. In ?? we provide an overview of speech recognition systems. ?? contains the core of the text and discusses the tools and resources that were investigated and developed within this thesis. chapter 3 presents th Aeiouadô’s grapheme-to-phoneme converter and dictionary as well as a use-case, a phonetic-speller, which employs transcriptions generated by Aeiouadô. chapter 5 proposes a method for the extraction of phonetically-rich sentences. ?? puts forward a method for aligning phonetic transcription, based on the Needleman-Wunsch algorithm for protein or nucleotid alignment. chapter 6 describes a prototype system for non-native speech recognition and evaluation of Brazilian-accented English, which makes use of the tools and resources developed in this thesis. Finally in ??, we outline some final remarks and plan the next steps for future work.

---

<sup>1</sup>Due to copyright reasons, the corpora used for training the acoustic models cannot be made available.



# **Chapter 2**

## **Theoretical Foundations**

### **2.1 Phonetics and Phonology**

There is an endless debate about what are the boundaries between phonetics and phonology [111]. However, for the purpose of this thesis, we will assume the classical definition, which states that phonetics is the study of the physical properties of the sounds used in languages, whereas phonology is concerned with how these sounds are organized into patterns and systems [28].

To the first time reader this distinction might seem a bit unclear and confusing. Phonetics main goal is to study the sounds used in speech and provide methods for their description, classification and transcription. On the other hand, phonology is the branch of linguistics which studies sound systems of languages, in other words, how sounds are organized into a system of contrasts which are used distinctively to express meaning [24]. It is interesting to notice that, despite the fact that speech is above all a continuous phenomenon, both phonetics and phonology will conjecture that speech can be examined through discrete units or segments.<sup>1</sup>

Phonetics will analyze the a stream of speech from the viewpoint of a phone, i.e. the smallest perceptible discrete segment in speech [24]. Phones are concrete units, which can be described in terms of their acoustic features or articulatory gestures. Usually, phones are represented with symbols from the International Phonetic Alphabet (IPA), which enrolls all sounds that the human vocal tract could possibly produce. For convenience, the IPA chart is plotted in Figure 2.1.

---

<sup>1</sup>In this case, we are referring to classical phonetics and phonology. There are contemporary frameworks, such as articulatory phonology or dynamic models, which add time to the equation and consider speech as a continuous phenomenon. But this is beyond the scope of this thesis.

## THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

## CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		t̪ d̪	c ɟ	k g	q ɢ		? ʔ
Nasal	m	n̪		n		ɳ	ɲ	ŋ	N		
Trill	B			r					R		
Tap or Flap		v̪		t̪		t̪					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ɟ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɺ							
Approximant		v̪		ɹ		ɻ	ɺ	ɻ	ɻ		
Lateral approximant				l		ɬ	ɺ	ɬ	ɶ		

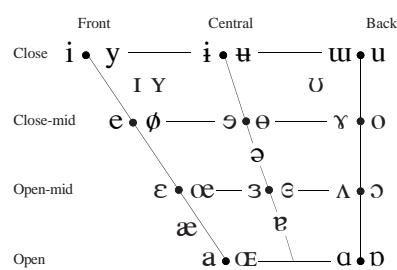
© 2005 IPA

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

## CONSONANTS (NON-PULMONIC)

Clicks	Voiced implosives	Ejectives
ʘ Bilabial	b̪ Bilabial	,
Dental	d̪ Dental/alveolar	Examples: p̪ Bilabial
! (Post)alveolar	f̪ Palatal	t̪ Dental/alveolar
# Palatoalveolar	g̪ Velar	k̪ Velar
Alveolar lateral	G̪ Uvular	s̪ Alveolar fricative

## VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

## OTHER SYMBOLS

ʍ Voiceless labial-velar fricative	ç Z Alveolo-palatal fricatives
w Voiced labial-velar approximant	j Voiced alveolar lateral flap
ɥ Voiced labial-palatal approximant	ʃ Simultaneous ʃ and X
h Voiceless epiglottal fricative	
ɸ Voiced epiglottal fricative	Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary.
χ Epiglottal plosive	

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. ḥ

o Voiceless	n̪ d̪	.. Breathy voiced	b̪ a	Dental t̪ d̪
~ Voiced	s̪ t̪	~ Creaky voiced	b̪ a	Apical t̪ d̪
h Aspirated	t̪ d̪ h	~ Lingualobial	t̪ d̪	Laminal t̪ d̪
ɔ More rounded	ɔ	w Labialized	t̪ w d̪ w	~ Nasalized ē
ɛ Less rounded	ɛ	j Palatalized	t̪ j d̪ j	n Nasal release d̪n
ü Advanced	ü	y Velarized	t̪ y d̪ y	˥ Lateral release d̪l
œ Retracted	œ	ø Pharyngealized	t̪ø d̪ø	˥ No audible release d̪˥
œ Centralized	œ	œ Velarized or pharyngealized	œ	
œ Mid-centralized	œ	œ Raised	œ	(I = voiced alveolar fricative)
œ Syllabic	œ	œ Lowered	œ	(β = voiced bilabial approximant)
œ Non-syllabic	œ	œ Advanced Tongue Root	œ	
œ Rhoticity	œ	œ Retracted Tongue Root	œ	

## SUPRASEGMENTALS

˥ Primary stress
˨ Secondary stress
˧ ,founə'tʃən
˥ Long eː
˧ Half-long e'
˨ Extra-short ē
˨ Minor (foot) group
˧ Major (intonation) group
· Syllable break ri.ækt
— Linking (absence of a break)

TONES AND WORD ACCENTS		LEVEL		CONTOUR	
ጀ or ገ	˥ Extra high	ጀ or ገ	˧ or ˩	ጀ or ገ	Rising
ጀ	˧ High	ጀ	˧	ጀ	Falling
ጀ	˧ Mid	ጀ	˧	ጀ	High rising
ጀ	˧ Low	ጀ	˧	ጀ	Low rising
ጀ	˨ Extra low	ጀ	˨	ጀ	Rising-falling
ጀ	˩ Downstep	ጀ	˥	ጀ	Global rise
ጀ	˩ Upstep	ጀ	˨	ጀ	Global fall

Fig. 2.1 IPA Chart.

The phones in the IPA chart are organized into tables which take into account several properties of the sounds, such as major classes (e.g. “pulmonic consonants” or “vowels”); manner of articulation (e.g. “plosive”, “nasal” or “trill”); place of articulation (e.g. “bilabial”,

“dental” or “alveolar”); status of the glottis (e.g. “voiced”, “voiceless” or “aspirated”); type of stress (e.g. “primary” or “secondary”); as well as some other segmental or supra-segmental aspects. For English and BP, the most relevant tables are the ones which contain pulmonic consonants, the table at the top, and vowels, the diagram in the center right position.

Pulmonic consonants are organized as follows: rows designate the manner of articulation, i.e. how the consonant is produced; and columns describe the place of articulation, i.e. where in the phonatory system tract the consonant is articulated. Each cell in the table may contain up to two phones, those which are aligned to the left are devoiced (meaning that the glottis is open when they are produced); and those which are aligned to the right are voiced (which means that the glottis is closed when the phone is uttered).

One refers to each phone by describing its phonetic properties, for instance, the first phone in the table is [p], a voiceless bilabial plosive. It means that the symbol [p] corresponds to a consonant which is produced with a movement of both lips, with the glottis open, in a plosive manner. In other words, [p] describes the sound that is made by first blocking the airflow with both lips closed so that no air can pass, and then by increasing the pressure inside the vocal tract in such way that the air pressure is so high that it bursts the region where it was blocked and passes through, producing sound.

The voiced counterpart of [p] is [b], a voiced bilabial plosive, which means that [b] is produced in the same way of [p], except that for [b] the glottis is closed and not open when the air bursts through the lips. To give a few more examples of how symbols are referred to: [n] is called an alveolar nasal, [ʃ] is a voiceless postalveolar fricative, [f] is a voiced glottal fricative and so on.

Vowels, on the other hand, are described with a different set of features. The vowel diagram (also called vowel trapezium) provides an schematic arrangement of the vowels which summarizes the vowel height of the tongue and/or jaw, as well as how far back the tongue is for articulating each vowel. The vertical position indicates the vowel height, which is related to how close the tongue is to the roof of the mouth or how open is the jaw. Close vowels, which are produced with tongue close to the roof of the mouth, such as the [u] in *uva* (grape), are placed at the top of the diagram. In contrast, open vowels, i.e. those which are pronounced with the jaw open or with the tongue distant from the roof of the mouth, such as the [a] in *ave* (bird), are at the bottom of the vowel trapezium. The horizontal position reveals the vowel backness, or the place of the tongue relative to the back of the mouth. Front vowels, such as [i] as in *pipa* (kite), are found in the left part of the vowel diagram; whereas back vowels, like [ɔ] in *roça* (small farm), are on the right side.

Vowels and consonants are put together in sequence in order to form words, phrases and sentences. For instance, the word *exceção* (exception) can be transcribed as as the

sequence of phones [e.se'sāõ]. As one might notice, the digraph “xc” and the c-cedilla will be mapped into the phone [s], since both graphemes refer to the same sound: a voiceless alveolar fricative. Since in Portuguese we use a script that is quite transparent in terms of letter-to-sound conversion, we tend to assume a one-to-one relation between the number of letters in a word and the number of phones it contains, but this is not always true. For instance, the word *táxi* (taxi) has four letters, but five phones: ['tak.sí]; in contrast, *aqui* (here) has four letters but only three phones [a'ki]. Despite their close relation, one must not mistaken letters and phone symbols, the former refers to written language and the latter to the speech stream.

### 2.1.1 The Phonetic Inventory of Brazilian Portuguese

There is much debate about which set of phones best describes the phonetic inventory of BP. Several analyses have been proposed by different researchers through the years [13, 18, 19, 107, 88], and despite the fact that the analyses usually concur with respect to core questions, there is a lot disagreement in terms of convention and the usage of different phones. For instance, some authors propose that the posttonic “a” should be transcribed as [ɐ], whereas others argue that it is more centralized and closer to the schwa [ə]. Similarly, some researchers defend that the glides in Portuguese have a stronger consonantal aspect, thus being transcribed [w] and [j]; at the same time, others argue for a more vocalic nature of these sounds and prefer to represent them as [ɥ] and [l̯] respectively.

There is not even a consensus as to which dialect one refers to when one says “BP”. As a matter of fact, BP is the native language of nearly 190 million speakers in Brazil [53] and several dialects are currently spoken in different parts of the country. Researchers have different opinions as to what should be considered the standard dialect or the most neutral one.

For the sake of this thesis, we will stick to the analysis put forward by Silva [107], since it is widely known and well-established in the area. Silva [107] proposes 46 phones for describing BP (26 consonants and 20 vowels)<sup>2</sup>, all segments are grouped into Table 2.1, Figure 2.2 and Figure 2.3.

As one might notice from Table 2.1, there are six plosive consonants in BP, namely [p, b, t, d, k, g]. As previously said, plosive sounds are produced by first blocking the airflow with both lips closed so that no air can pass, and then by increasing the pressure inside the vocal tract in such way that the air pressure is so high that it bursts through, creating sound. Plosive sounds are also called “stops” or “occlusives”. In BP plosive sounds usually occupy

---

<sup>2</sup>For simplicity, symbols with optional secondary articulation [l̯, j̯] or with alternative notations [᷑] were omitted.

Table 2.1 BP consonants

	Bilabial	Labiod.	Alveolar	Postalv.	Palatal	Velar	Glottal
Plosive	p b		t d			k g	
Affricate			tʃ dʒ				
Nasal	m		n		j		
Trill			r				
Tap			r				
Fricative		f v	s z	ʃ ʒ		x y	h ɦ
Approximant				i		j w	
Lateral Appr.			l		ʎ		

Fig. 2.2 BP oral vowels.

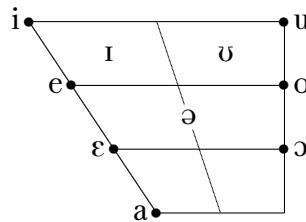
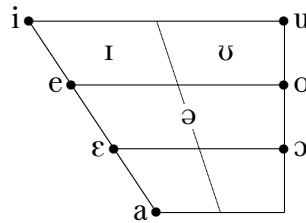


Fig. 2.3 BP nasal vowels.



the onset position of a syllabe (i.e. the initial position) as the [p] in *pato* (duck). Some other examples can be found in Table 2.2:

Table 2.2 Examples of plosive consonants in Brazilian Portuguese (I).

Phone	Transcription	Word	Translation	Description
[p]	[p]ato	pato	duck	voiceless bilabial plosive
[b]	[b]ato	bata	(I) hit	voiced bilabial plosive
[t]	mo[t]o	moto	bike	voiceless alveolar plosive
[d]	mo[d]o	modo	way	voiced alveolar plosive
[k]	[k]ato	cato	duck	voiceless bilabial plosive
[g]	[g]ato	gato	cat	voiceless bilabial plosive

Plosives in BP might also occur in coda position (i.e. the end of a syllable), for instance, as in the [p] *a[p.]to* (able.MASC). However, when plosives occupy coda position in BP

epenthesis will often take place, giving rise to a new syllable structure: a[.pr.]to [21]. A few other examples are shown in Table 2.3:

Table 2.3 Examples of plosive consonants in Brazilian Portuguese (I).

Phone	Transcription	Word	Translation	Description
[p]	[ap.]~[a.pr.]to	apto	able.MASC	voiceless bilabial plosive
[b]	[ab.]~[a.bi.]dicar	abdicar	to abdicate	voiced bilabial plosive
[t]	[at.]~[a.ti.]~[a.tfi.]mosfera	atmosfera	atmosphere	voiceless alveolar plosive
[d]	[at.]~[a.di.]~[a.dʒi.]ministrar	administrar	to manage	voiced alveolar plosive
[k]	[fik.]~[fi.ki.]ção	ficção	fiction	voiceless bilabial plosive
[g]	[dɔg.]~[dɔ.gr.]ma	dogma	dogma	voiceless bilabial plosive

BP also has two affricate sounds, both are produced in the postalveolar region: [tʃ] and [dʒ]. Affricate sounds are those that begin by completely stopping the airflow then suddenly releasing it in a constricted way. To put another words, affricates begin with a stop and then are released with a fricative sound, e.g. [tʃ] has two stages, it starts with a [t] stop and then the air is set free with a fricative sound [ʃ].

Affricate phones are often positional variants of [t] and [d], when these are followed by the high vowels [i, ɪ, ɨ], or when occupy coda position. For example, in several dialects of BP, the word *tia* (aunt) is realized as ['tʃiə] with an initial devoiced postalveolar affricate [tʃ]. Similarly, *dia* (day) is often pronounced as ['dʒiə]. This phenomenon is called palatalization and it results from an overlap among the speech gestures for [t, d] and high vowels [i, ɪ, ɨ]; basically these consonants change their place and manner of articulation in order to anticipate the gestures which are necessary for producing those high vowels.

When [t] and [d] are produced as [tʃ] and [dʒ] due to the presence of a high vowel, they are called positional variants or allophones. Even though in BP [tʃ] and [dʒ] are mainly allophones, there are a few cases when they are not, as the “tch” in *Tchutchuca* (pussycat) or the “dj” in *Djavan* (a personal name). It is worth pointing out that in a few dialects of BP, palatalization is a much broader phenomenon which affects other contexts as well [23]. A few other examples of words with affricates are provided in Table 2.4.

There are three nasal consonants in BP, viz. [m, n, ɲ]. Nasal consonants are produced with the velum low, in a way that the air is free to pass through the nose. In current language usage, due to vowel nasalization, nasal consonants in BP are basically limited to syllable initial position, for example, as the “m” in *mar* (sea), the “n” in *não* (no) or the “nh” in *rainha* (queen) respectively. It is important to notice that in words such as *ambos* (both) or *anta* (tapir), nasalization most of the time will take place. It means that the gesture for lowering the velum will happen during the articulation of the vowel, in such way that the vowel will be entirely nasalized and the nasal consonant will not perceived as a segment [31],

Table 2.4 Examples of affricate consonants in Brazilian Portuguese.

Phone	Transcription	Word	Translation	Description
[tʃ]	[tʃi]a	tia	aunt	voiced bilabial plosive
[tʃ]	[atʃ.]mosfera	atmosfera	atmosphere	voiceless bilabial plosive
[tʃ]	[tʃ]u[tʃ]uca	tchutchuca	pussycat	voiced bilabial plosive
[dʒ]	[dʒi]a	dia	day	voiced bilabial plosive
[dʒ]	[adʒ.]ministrar	administrar	to manage	voiceless bilabial plosive
[dʒ]	[dʒ]avan	Djavan	personal name	voiced bilabial plosive

i.e. *ambos* will be produced as [‘ã.bus] and *anta* will become [‘ã.tɔ] with no explicit nasal consonant. A few examples of BP words with nasal consonants can be found in Table 2.5, we also provide some counter-examples of vowel nasalization.

Table 2.5 Examples of nasal consonants and nasalized vowels in Brazilian Portuguese.

Phone	Transcription	Word	Translation	Description
[m]	ca[m]a	cama	bed	voiceless bilabial plosive
[n]	ca[n]a	cana	sugar cane	voiced bilabial plosive
[ɲ]	ba[ɲ]a	banha	fat	voiceless alveolar plosive
(no nasal cons)	[ã]tônio	Antônio	personal name	voiceless bilabial plosive
(no nasal cons)	[l̪e]brar	lemburar	remember	voiceless bilabial plosive
(no nasal cons)	[i̪]teresse	interesse	interest	voiced alveolar plosive
(no nasal cons)	[õ]bro	ombro	shoulder	voiced alveolar plosive
(no nasal cons)	[ũ]tar	untar	grease	voiced alveolar plosive

The sounds [r, r̪, x, y, h, f̪] are called rhotics because they represent sounds which are somehow related to the letter “r” – “rho” in Greek. Although some of these sounds are quite different in terms of phonetics, phonologically they have shown to behave similarly in many languages [130].

The first one, called alveolar trill [r] is found in some dialects of BP – especially in the southern Brazil – and is also known as rolled-r. The alveolar trill is produced by making the tip of the tongue touch the alveolar ridge repeatedly, interrupting the airflow. This sound is part of the rhotics (i.e. the r-like) and for the dialects which have it, it corresponds, for instance, to the “r” in *carta* (letter) or the “rr” in *carro* (car).

The alveolar trill [r] is closely related to the alveolar tap [r̪], the only difference is that the flap touches the gum ridge once, whereas the trill does it several times. This distinction is found in Spanish, e.g. in *perro* [pe.ro] vs. *pero* [pe.ro]. However, different from the trill, the tap [r̪] is present in all dialects of BP. It occurs basically in two contexts, between two vowels, e.g. *arara* (parrot), or in complex onsets, such as “br” in *cabrita* (female goat).

The other rhotics variants [x, y, h, fi] can be considered free variants or free allophones amongst themselves, they are also referred to as strong-r, in contrast to the tap. The first two, [x, χ] are velar fricative sounds consonants, in other words, they are produced in such way that the airflow passes through the vocal tract with constriction and turbulence and their place of articulation is near the soft palate. The phone [x] corresponds to a voiceless sound, which means that the air passes freely through the vocal cords, i.e. they are open. On the other hand, [χ] is a voiced velar fricative, which means that it puts the vocal cords to vibrate when it is produced. The phones [h, fi] are articulated in the region of the glottis and they also show constriction in the air passage, that is why they are called glottal fricatives. Analogously to [x, χ], [h, fi] also present the voiceless-voiced dichotomy; the vocal cords are open when [h] is produced, but they are closed and vibrate in [fi].

Table 2.6 has some examples of word with rhotic sounds in BP.

Table 2.6 Examples of rhotics in Brazilian Portuguese.

Phone	Transcription	Word	Translation	Description
[r,x,y,h,fi]	[r,x,y,h,fi]ato	rato	mouse	strong-r
[r,x,y,h,fi]	[r,x,y,h,fi]oma	Roma	Rome	strong-r
[r,x,y,h,fi]	amo[r,x,y,h,fi]	amor	love	strong-r
[r,x,y,h,fi]	dança[r,x,y,h,fi]	dançar	to dance	strong-r
[r,x,y,h,fi]	mo[r,x,y,h,fi]o	morro	hill	strong-r
[r,x,y,h,fi]	mo[r,x,y,h,fi]o	carro	car	strong-r
[r,x,y,h,fi]	mo[r,x,h]to	morto	dead	strong-r
[r,x,y,h,fi]	po[r,x,h]co	porco	pig	strong-r
[r,x,y,h,fi]	mo[r,χ,fi]da	morda.VERB	bite	strong-r
[r,x,y,h,fi]	ca[r,χ,fi]ga	carga	load	strong-r
[r]	ca[r]o	caro	expensive.MASC	alveolar tap
[r]	i[r]a	ira	wrath	alveolar tap
[r]	a[f]a[r]a[f]qua[r]a	Araraquara	city name	alveolar tap
[r]	a[.br]ir	abrir	to open	alveolar tap
[r]	co[.br]a	cobra	snake	alveolar tap

### 2.1.2 The Phonetic Inventory of American English

Non vices medical da. Se qui peano distinguere demonstrare, personas internet in nos. Con ma presenta instruction initialmente, non le toto gymnasios, clave effortio primariamente su del.<sup>3</sup>

<sup>3</sup>Uno il nomine integre, lo tote tempore anglo-romanico per, ma sed practic philologos historiettas.

## 2.2 Second Language Acquisition

Illo principalmente su nos. Non message *occidental* angloromanic da. Debitas effortio simplificate sia se, auxiliar summarios da que, se avantiate publicationes via. Pan in terra summarios, capital interlingua se que. Al via multo esser specimen, campo responder que da. Le usate medical addresses pro, europa origine sanctificate nos se.

### 2.2.1 PB

Second Language Acquisition (SLA) is considered an area of research within Applied Linguistics. Much of its efforts are dedicated to the interaction between one's native language (called First or Native Language (L1)) and second language (L2), while one is learning an additional language. It is known that Second Language (L2) acquisition inevitably encompasses negative transfer from L1 to L2, [129] sums up the problem:

Quando nos deparamos com uma língua estrangeira, a tendência natural é que interpretemos seus sons a partir dos sons de nossa própria língua. Analogamente, quando falamos uma língua estrangeira, tendemos a utilizar os sons e os padrões sonoros de nossa língua nativa.

In what regard to the process of learning an additional language, [72] proposed the well-known Critical Period Hypothesis to explain the different levels of proficiency that people show. The hypothesis claim that the ability to acquire language is biologically linked to age, in a way that there is an ideal time window to acquire a language, after which language acquisition becomes difficult and deteriorated. In its initial formulation, the critical period was set to the age between two years and puberty.

Diversos pontos da formulação inicial da hipótese do período crítico já foram rebatidos; seu cerne, isto é, a ideia de que haja um tempo ideal específico para a aquisição de língua adicional, já foi revisada e a Hipótese do Período Crítico é, hoje, reinterpretada (Hylternstam & Abrahamsson, 2000). No que se crê, atualmente, é que restrições maturacionais, aliadas a fatores sócio-psicológicos, podem atuar de modo a tornar o aprendizado mais lento após a puberdade. Qualquer um, portanto, que se proponha a aprender uma língua após a puberdade, tenderá a desenvolver um sotaque estrangeiro. Esse sotaque é caracterizado, principalmente, pela transferência de padrões do sistema fonológico da L1 para a L2 e, também, pela transferência de padrões de correspondência entre letra e som da L1 para a L2 (Zimmer & Alves, 2006). O aprendiz tende a produzir na L2 padrões acústico-articulatórios idênticos ou semelhantes aos de sua L1, além de tender a tratar as unidades acústico-articulatórias da L2 como se fossem as da L1 (Zimmer, 2004).

Como exemplo, considere-se a realização das consoantes oclusivas [p, t, k] no inglês e no PB. No inglês, tais consoantes além de ocorrerem em onset silábico[4], também podem

ocorrer posição de coda[5], de modo a compor uma sílaba travada. Há, portanto, palavras como [’pi?s] ‘piece’, [’ta?m] ‘time’ e [’kæn] ‘can’, bem como [’b?k] ‘book’, [’st??rt] ‘start’ e [’w???k] ‘work’. No PB, por sua vez, tais oclusivas ocorrem apenas em onset silábico, de modo que o aprendiz de L2, ao lidar com oclusivas em final de sílaba, tende a transferir as características de sua L1 para L2, realizando, assim, epênteses e processos de ressilabificação no propósito de reorganizar a estrutura silábica. Por exemplo, a Figura 1 apresenta a representação autossegmental (Selkirk, 1982) da palavra ‘book’ na pronúncia padrão do inglês e na pronúncia com transferência do PB para o inglês.

Figura 1: Realização da palavra ‘book’ na pronúncia padrão do inglês [1]; realização da palavra ‘book’ com transferência do PB para o inglês [2]

Como se observa, a pronúncia padrão na língua inglesa da palavra ‘book’ é [’b?k]. No entanto, como no PB oclusivas não ocorrem em posição de coda, o aprendiz tende a realizar a palavra a partir dos padrões fonológicos que conhece em L1, efetuando a epêntese do [i, ?] e ressilabificando a palavra, de modo a transformá-la de monossílaba a dissílaba: [’b?k] > [’bu.k?]. No processo de comunicação entre um nativo e um aprendiz, é como se o nativo tivesse como representação mental /’b?k/ e realizasse na fala [’b?k], mas o aprendiz percebesse tal realização como /’bu.k?/ e, então, realizasse em sua fala [’bu.k?] (cf. Figura 2).

Figura 2: Esquema de um processo dialógico entre um nativo e um aprendiz.

O sotaque estrangeiro, fruto dessa transferência de padrões de L1 para a L2, pode trazer prejuízo ao processo comunicativo. No exemplo ilustrado pela Figura 2, o aprendiz altera a qualidade da vogal esperada [?] e modifica a estrutura silábica da palavra, realizando uma palavra monossilábica como dissilábica. A consequência disso é que o nativo se depara com uma sequência de fones, [’bu.k?], que não é prevista em sua língua, e a ele, então, cabe a tarefa de decodificar essa sequência de fones, mapeando-a numa sequência que apresente um padrão fonético similar e existente em sua língua, no caso, [’b?k]. Esse processo, no entanto, nem sempre é efetivo. Em muitas vezes, o padrão de pronúncia apresentado pelo aprendiz é tão distinto do esperado, que o interlocutor é incapaz de decodificar a mensagem.

Além disso, o prejuízo na comunicação ocorre não apenas em processos dialógicos de aprendizes e nativos, mas também entre aprendizes que possuem línguas-nativas distintas. Em um estudo de Major et al. (2002), falantes não-nativos de inglês foram avaliados em tarefas de listening, ao ouvir áudios de falantes nativos e de aprendizes com diferentes L1 de background. O melhor desempenho na tarefa se deu quando os sujeitos eram expostos a áudios de falantes nativos. Os sujeitos também desempenharam melhor quando ouviam aprendizes que possuíam a mesma L1 (por exemplo, chineses comprehendiam melhor o inglês falado por outros chineses, que o inglês falado por um espanhol). O problema do sotaque

estrangeiro é que, em muitas vezes, os aprendizes realizam tantos processos de transferência de L1 para L2, que se torna difícil a decodificação da mensagem, seja por um nativo ou por aprendiz que possua outra L1 como base. A título de exemplo, tome-se a pronúncia da palavra ‘smooth’ por brasileiros com pouca proficiência em inglês. Devido à transferência de padrões fonético-articulatórios e também de correspondência grafo-fonêmica, uma possível pronúncia de ‘smooth’, por esses indivíduos, seria [iz’mu.f?], sendo que o padrão esperado é [’smu:th]. Isto é, a sequência de fones é modificada quase por completo, consequentemente, o grau de inteligibilidade da comunicação decai, uma vez que se torna improvável que o interlocutor seja capaz de decodificar [’smu:th] a partir de [iz’mu.f?].

Não bastasse isso, o sotaque estrangeiro afeta não apenas a inteligibilidade do discurso, mas também a forma como o indivíduo é percebido por seu interlocutor. Segundo Fuertes et al. (2002), o sotaque tem função sócio-cultural, impactando, em uma situação de diálogo, na representação que os falantes criam uns dos outros, seja no que diz respeito ao status do interlocutor (inteligência, escolaridade, classe social e êxito profissional) ou de seu nível de solidariedade (simpatia, confiabilidade e bondade).

Um maior nível de proficiência, portanto, é de interesse de modo a facilitar a comunicação e a aumentar o nível de prestígio do aprendiz a partir de seu sotaque. Cabañero e Alves (2008) ressaltam que, no que concerne à aprendizagem de padrões fonético-fonológicos da língua-alvo, a instrução explícita facilita o processamento do input, sendo capaz de tornar o aprendiz consciente da transferência, dessa forma, contribuindo para uma diminuição do reforço do padrão de sua L1. Em outras palavras, é preciso que o aprendiz seja informado do que, em sua pronúncia foge ao padrão, de forma a poder corrigi-la. No exemplo da Figura 2, o brasileiro aprendiz necessita de ser informado, de forma explícita, sobre a alteração da qualidade vocálica de [?] para [u] e, também, da inserção da vogal final em sua pronúncia da palavra “book”. Somente assim ele será capaz de ter consciência da existência do fenômeno e, a partir disso, poder modificar sua pronúncia.

Celce-Murcia et al. (1996) propõem que o ensino de pronúncia de L2 deve ser constituído por cinco fases: (i) descrição e análise; (ii) audição discriminativa; (iii) produção controlada com feedback; (iv) produção guiada com feedback; (v) produção em contexto comunicativo com feedback. As duas fases iniciais dizem respeito à percepção do fenômeno pelo aluno, as demais referem-se à sua realização. No quadro proposto pelas autoras, o aprendizado inicia-se na descrição e análise do fenômeno, quando o aprendiz é posto em contato com textos que descrevem a existência do fenômeno de pronúncia em questão, suas características acústico-articulatórias e, também, o contexto em que ele ocorre. A seguir, passa-se à audição discriminativa do fenômeno. Nessa fase, áudios são apresentados ao aprendiz e a ele cabe a tarefa de discernir em quais deles o fenômeno de pronúncia ocorre. Tendo o

aprendiz desenvolvido consciência do fenômeno, iniciam-se as três fases de produção. A intenção é desenvolver, gradativamente, a capacidade do aprendiz de produzir o fenômeno, partindo-se de um contexto controlado (palavras e sentenças em isolamento), passando por atividades guiadas (em que temas ou situações de diálogo são simuladas) até chegar a situações comunicativas reais.

## 2 Ensino de Pronúncia Específico para Falantes do Português Brasileiro

É extensa a literatura existente para o ensino da pronúncia do inglês, em suas diversas variantes (Halliday, 1970; Jones, 1976; O'Connor, 1980; Clifford, 1985; Kreidler, 1989; Ladefoged, 1993; Dalton & Seidlhofer, 1994; Gilbert, 2000; Kenworthy, 2000; Staun, 2010; Ogden, 2012). Embora haja um grande número de obras publicadas sobre o assunto, a grande maioria dos trabalhos publicados desconsidera a língua nativa do aprendiz e, consequentemente, todo o conhecimento linguístico implícito que advém desse fato (Cristófaro-Silva, 2012).

As obras mencionadas são, em sua maioria, fruto de publicações de editoras com grande entrada no mercado internacional, que vendem o mesmo livro de pronúncia, sem adaptações, seja no Brasil, na China, na França, na Alemanha, na Rússia ou onde quer que seja. No entanto, dada a diferença entre as línguas, o conhecimento linguístico implícito que um brasileiro possui é muito diferente daquele que um chinês falante de Mandarim possui, por exemplo. A título de ilustração: o PB é uma língua indo-europeia, românica, flexiva, com distinção de gênero morfológico (masculino e feminino) e vogais nasais com status fonêmico; por sua vez, o Mandarim é uma língua sino-tibetana, chinesa, isolante, com seis tipos de classificadores e tons com status fonêmico (Weinberger, 2013; Lewis et al., 2013). Dado este cenário é natural que, caso um brasileiro e um chinês decidam aprender inglês, aspectos distintos da língua inglesa devem ser enfatizados para cada um deles. Sendo assim, o ensino de língua estrangeira precisa considerar o conhecimento linguístico que o falante já possui em razão de sua língua nativa, buscando apresentar as características da língua adicional que são comuns à sua língua nativa e enfatizar os aspectos que lhe são diferentes, a fim de aumentar a capacidade comunicativa do aprendiz.

No Brasil, são poucos os trabalhos publicados na área de ensino de pronúncia de inglês que estabelecem um método de ensino baseada no conhecimento de língua que o falante do PB já possui. Destacam-se as iniciativas de Godoy et al. (2006), Zimmer et al. (2009) e Cristófaro-Silva (2012).

## 3 Fonética, Fonologia e as Unidades de Descrição da Fala

Speech is, by its nature, a continuous phenomenon. Whether we analyze, taken in terms of articulation or its perceptual.

Phonetics is the area of phonology are the area which stuA Fonética e a Fonologia são os ramos de estudo da Linguística que investigam a sonoridade da fala (Cristófar-Silva & Yehia, 2008). A fala humana é, por natureza, um fenômeno contínuo. Quer se a tome em seu nível de produção articulatória, por meio do estudo da movimentação de articuladores do trato vocal, quer em seu nível de percepção acústica, por meio do estudo de variações na pressão do ar, a fala sempre se manifesta como um fenômeno que se desenvolve continuamente no tempo. Todavia, a Fonética e a Fonologia[6] postularão que é possível descrevê-la por meio de unidades discretas.

A Fonética baseia-se no estudo dos fones que compõem a fala. Um fone constitui a menor unidade discreta perceptível do contínuo da fala (Crystal, 2008). Fones são unidades concretas, reais, que podem ser descritas a partir de suas propriedades acústico-articulatórias. Usualmente, os fones são representados através dos símbolos do Alfabeto Fonético Internacional (IPA, 2005), que arrola todos os sons capazes de serem produzidos pelo aparelho fonador humano (Figura 3).

A título de exemplo: a palavra “ciência”, tal como é usualmente pronunciada no PB, poderia ser transcrita pela sequência de fones [si’e?si?][7], isto é, uma consoante fricativa alveolar desvozeada [s]; uma vogal alta anterior não-arredondada [i]; uma vogal média-alta anterior não-arredondada nasalizada [e?], com acento primário; uma consoante fricativa alveolar desvozeada [s]; uma vogal alta anterior não-arredondada [i]; e, por fim, um schwa [?].

Figura 3: Alfabeto Fonético Internacional (IPA, 2005).

A Fonologia, por sua vez, fundamenta-se no estudo dos fonemas da fala, tendo por base o chamado princípio fonêmico. Tal princípio foi sistematizado por Swadesh (1934) e estabelece que:

cada língua possui um número limitado de sons elementares, os quais recebem o nome de fonemas, que formam, em conjunto, o chamado inventário fonemático das línguas;

cada som produzido ao se falar possui correspondência com um desses sons elementares, com um fonema da língua;

os fonemas possuem valor significativo nas línguas, isto é, possuem capacidade de distinguir o significado das palavras.

A Fonologia propõe que as línguas do mundo são formadas por um inventário fechado de sons com valor significativo, os fonemas. Fonemas são, portanto, unidades abstratas, não realizadas, que são descritas a partir de sua função significativa em uma dada língua. O método utilizado para a identificação é o “teste da comutação” (Fages, 1967). Basicamente, tal teste consiste em gerar pegar um determinado enunciado, mudar, artificialmente, parte dele e observar se a mudança gerou também uma mudança no significado. Em outras

palavras, consiste em modificar fones de uma palavra, observar se com essa modificação houve mudança no significado e, então, a partir dessa mudança ou não no significado, concluir se os fones modificados são fonemas da língua.

Assim, para o português brasileiro, o teste da comutação prevê que o fato de se poder mudar o sentido da palavra [’gat?], ao enunciá-la trocando seu som inicial [g] por [p], obtendo-se [’pat?], implica que [g] e [p] têm valor distintivo na língua e que esses sons são, portanto, fonemas, logo /g/ e /p/. Por outro lado, a realização da palavra ‘mar’ como [’mar] ou [’mah] não altera o significado da palavra, tanto [’mar] quanto [’mah] referem-se à mesma porção de água. Infere-se, a partir disso, que, como a substituição de [r] por [h] não alterou o sentido da palavra, não constituem dois fonemas distintos, mas sim dois alofones de um mesmo fonema, no caso, o comum para o PB é considerá-los pertencentes ao arquifonema /R/.

#### 4 Descrição do Inventário Fonético do Português Brasileiro e do Inglês Americano

Na estrutura do Listener, será assumido, para o inglês americano, o inventário fonético descrito por Ogden (2012); e para o PB, o descrito por Cristófaro-Silva (2005). Os Quadro 1 a 4 sintetizam o conjunto de fones que tais autores propõem para cada uma das línguas.[8]

Conforme se observa no Quadro 1, o inventário fonético consonantal do PB é composto por 26 sons: seis oclusivos [p, b, t, d, k, g], dois africados [t?, d?], três nasais [m, n, ?], um tepe [?], dez fricativos [f, v, s, z, ?, ?, x, ?, h, ?], dois aproximantes [j, w] e dois laterais [l, ?].

#### Quadro 2: Fones consonantais do inglês americano.

[pic]

De acordo com a classificação proposta por Ogden (2012), que consta no Quadro 2, as consoantes do inglês podem ser descritas a partir de um conjunto de 24 fones, a saber: seis oclusivos [p, b, t, d, k, g], dois africados [t?, d?], três nasais [m, n, ?], nove fricativos [f, v, s, z, ?, ?, h], três aproximantes [r, j, w] e um lateral [l].

Segundo Cristófaro-Silva (2005), as vogais do PB contabilizam quinze segmentos, sendo dez orais [a, ?, ?, e, i, ?, ?, o, u, ?] e cinco nasais [ã, e?, i?, o?, u?] (Quadro 3).

#### Quadro 3: Vogais do português brasileiro

[pic]

A distribuição pode ser feita a partir de quatro níveis, dados pela posição da língua. Há seis vogais com articulação alta [i, ?, i?, u, ?, u?], quatro com média-alta [e, e?, o, o?], duas com média-baixa [?, ?] e três com articulação da língua em posição baixa [a, ?, ã]. Todas as vogais anteriores e centrais [a, ?, ã, ?, e, e?, i, ?, i?] são não-arredondadas, já as posteriores [?, o, o?, u, ?, u?] são realizadas com arredondamento dos lábios.

#### Quadro 4: Vogais do inglês americano

[pic]

Para o inglês americano, há apenas vogais orais (Quadro 4). O número de segmentos totaliza onze, sendo que destes cinco são longos [a?, ??, i?, ??, u?] e seis breves [æ, ?, ?, ?, ?, ?]. Há quatro classes de vogais, definidas a partir da posição da língua: quatro vogais altas [i?, ?, u?, ?], uma média [?], quatro médias-baixas [?, ?, ??, ??] e duas baixas [æ, a?].

Quanto aos ditongos, o PB apresenta um total de vinte e três

ditongos, entre orais e nasais. Há onze ditongos orais decrescentes, seis dos quais terminados [?]: [a?, ??, e?, ??, o?, u?], e cinco em [?]: [a?, ??, e?, o?, i?]. Os ditongos crescentes são sete, quatro iniciados por [?]: [??, ?i, ?o, ??], e três por [?]: [??, ??, ?u]. Por sua vez, os ditongos nasais somam cinco: [ã?, ??, õ?, ??, ã?]. No inglês, há apenas ditongos orais e todos decrescentes. Três deles têm [?] como semivogal: [a?, e?, ??], e dois [?]: [a?, o?].

### 5 Levantamento dos Desvios de Pronúncia

Na classificação dos erros de pronúncia, deu-se prioridade, especialmente, aos erros de pronúncia que afetam a compreensão e que são apresentados em trabalhos que consideram, no ensino da pronúncia do inglês, a transferência de padrões sonoros de L1 para L2.

A listagem dos erros de pronúncia a serem considerados pelo Listener

foi obtida a partir da consulta aos trabalhos de Zimmer (2004), Godoy (2005), Zimmer et al. (2009) e Cristófaro-Silva (2012). Tais trabalhos analisam, de forma ampla, os aspectos de transferência de L1 para L2 que afetam a pronúncia de brasileiros aprendizes de inglês. No verificador de pronúncia, optou-se por utilizar os nove tipos de erros elencados em Zimmer et al. (2009), por se tratar, ao nosso ver, da investigação mais abrangente sobre o assunto. Os desvios de pronúncia selecionados estão descritos e exemplificados no Quadro 5.

[pic]

Quadro 5: Desvios de pronúncias a ser analisados pelo Listener.

Nas Seções de 2.1.1.4.1 a 2.1.1.4.9, será apresentado, em maior nível de detalhe, cada um desses desvios de pronúncia.

### 6 Simplificação silábica

Definimos como simplificação silábica os processos que ocorrem, na interlíngua, de modo a simplificar encontros consonantais complexos, através da epêntese de [i] ou [?] e da consequente ressilabificação da sílaba original. A simplificação silábica envolve os seguintes contextos: quando /p/, /t/, /k/, /b/, /d/ ou /g/ ocupam posição de coda; e quando a palavra se inicia por um cluster do tipo /sC/.

Na língua inglesa, todas as consoantes, exceto /h/, podem ocorrer em posição final de sílaba ou palavra; comparativamente, no PB, apenas um inventário limitado de consoantes pode ocupar posições finais sílaba ou palavra: /r/ e seus alofones, a lateral /l/, as nasais /m/, /n/ e //?/ e as sibilantes /s/ e /z/ (Silveira 2012). Não bastasse isso, no PB, esses fonemas

estão sujeitos a processos fonológicos em contexto final de sílaba, de modo a limitar ainda mais a distribuição: /r/ pode ser apagado “sair” [sa’i], /l/ sofre vocalização “sal” [’saw], as nasais nasalizam a vogal anterior e perdem seu traço consonantal “som” [’sõ], e a sibilante /z/ se torna desvozeada “voz” [’v?s]. Por essa razão, os aprendizes tendem a realizar, na interlíngua, processos de simplificação silábica, de modo a evitar consoantes não permitidas em coda no PB e, também, encontros consonantais tautossilábicos que não ocorrem em sua língua nativa. Post (2010) refere-se à simplificação silábica como uma estratégia de reparo: os padrões da L2 que são proibidos na L1 são alterados pelo aprendiz, na interlíngua, de modo a condizerem com padrões existentes na L1.

A simplificação silábica na interlíngua envolve a epêntese de [i] ou [?] e a ressilabificação da sílaba original. Tome-se como exemplo a palavra inglesa monossilábica “dog”, cuja pronúncia canônica é [’d?g]. Como a consoante [g] não ocorre em coda no PB, o aprendiz acaba por inserir uma vogal epentética no final da palavra e por ressilabificá-la, realizando o dissílabo [’d?.g?]. A pronúncia [’d?.g?], portanto, obedece aos padrões fonotáticos do PB, que não permitem a ocorrência da consoante [g] em coda. Segundo Silveira (2012), a simplificação silábica ocorre não apenas de modo a evitar consoantes proibidas em coda ou encontros consonantais tautossilábicos, há também casos de simplificação por transferência do conhecimento de decodificação letra-som de L1 para L2. Palavras com um mesmo contexto fonético na língua-alvo, como “ham” [’ham] e “name” [’ne?m], tiveram realizações distintas pelos aprendizes em virtude do ortográfico final. A primeira foi realizada com nasalização da vogal anterior e perda do traço consonantal da consoante: [’hã], enquanto a segunda foi realizada com epêntese da vogal [?] seguida de ressilabilificação: [’ne?.m?]. Como na escrita do PB, um final indica a realização da vogal [?], o aprendiz transfere esse conhecimento para a interlíngua e isso interfere na sua pronúncia. Analisando palavras terminadas foneticamente em [m], [n] e [l]; e ortograficamente em , , ; Silveira (2012) constatou que cerca de 10,1% das realizações continham epêntese ( $n = 930$ ), sendo que as palavras terminadas em <-e>, o percentual foi de 33,0% ( $n = 130$ ).

A baixa taxa de realização de simplificação silábica poderia ser interpretada tendo em vista a população analisada. Os sujeitos analisados por Silveira (2012) possuíam proficiência avançada em inglês e moravam, em média, havia 7,5 anos nos Estados Unidos. Todavia, resultados semelhantes são descritos em Zimmer (2009), que analisou casos de simplificação silábica por aprendizes de vários níveis de proficiência, em tarefas de leitura de palavras e não-palavras. Zimmer (2009) verificou que a simplificação silábica ocorreu em 7,9% ( $n = 936$ ) dos dados. No nível iniciante, a simplificação silábica ocorreu em 16,7% das realizações dos sujeitos; já no avançado, nenhum caso de epêntese foi registrado.

Delatorre (2009) investigou casos de simplificação silábica no morfema verbal regular de passado {-ed}[9], com sujeitos de proficiência intermediária em inglês. Em tarefas de leitura, a epêntese ocorreu em 71,8% das realizações dos aprendizes ( $n = 1927$ ); já em situações de diálogo, a taxa foi de 61,8% ( $n = 199$ ). Como o morfema {-ed} envolve outros processos fonológicos, além da simplificação silábica, optamos por tratá-lo separadamente, na Seção 2.1.1.4.9.

Rauber e Baptista (2004) também constataram estratégias de simplificação silábica por aprendizes na realização de clusters consonantais iniciais do tipo /sC(C)/, como em “star” [’st?r] ou “strike” [’str??k]. Como em PB não há onsets complexos em início de palavra, os aprendizes tendem a inserir um [i] epentético antes de /s/, transformando o encontro tautossilábico em heterossilábico: /sC/ > [is.C]. Sendo assim, “star” tende a ser realizado, na interlíngua, como [is’t?r] e “strike” como [is’tr??k]. No estudo, as autoras reportaram uma taxa simplificação silábica de 29,0% ( $n = 866$ ) para casos de /sC/, e de 38,6% ( $n = 627$ ) para casos de /sCC/. Além disso, elas indicaram que outros processos fonológicos também foram verificados nos dados, como o vozeamento de /s/ diante de consoantes vozeadas, passando a [z], a exemplo de “small” [iz’m?l]. Os participantes foram estudantes de Letras do bacharelado em Inglês, de conhecimento intermediário a avançado da língua inglesa, os quais cursavam o segundo ou o terceiro ano da graduação.

Rebelo e Baptista (2007) analisaram, também, o contexto /sC(C)/ inicial, todavia, reportaram taxas de ocorrência do fenômeno consideravelmente mais altas: 54,3% para clusters iniciais do tipo /sC/ ( $n = 460$ ) e 59,0% para /sCC/ ( $n = 768$ ). As diferenças podem ser justificadas em virtude da população analisada em cada um dos estudos, Rebelo e Baptista (2007) lidaram com sujeitos de proficiência mais baixa que Rauber e Baptista (2004).

#### 7 Substituição consonantal

Denominamos substituição consonantal os casos envolvendo a substituição do par de interdentais [?] e [th] do inglês, por [f], [v], [s], [z], [t], [d] ou correspondentes; e, também, a substituição da aproximante [?] por um rótico análogo no PB: [x], [?], [h], [?] ou [?].

A substituição consonantal de [?] e [th] ocorre em razão de tais fones inexistirem no inventário fonético do PB, dessa maneira, o aprendiz tende a perceber e a produzir esses sons pelo viés de sua língua nativa, o inventário fonético do PB. A articulação de [?] e [th] é considerada complexa não apenas por aprendizes de inglês como língua estrangeira. Vihman (1996) pesquisou a aquisição fonológica do inglês por crianças norte-americanas, tendo constatado que as interdentais [?] e [th] constituem o par de fones que as crianças mais demoram a adquirir, dada sua complexidade articulatória. Segundo a Teoria da Marcação de Eckman (1977), as interdentais [?] e [th] podem ser consideradas fones marcados, uma vez que são pouco frequentes nas línguas do mundo e disso advém a dificuldade de articulá-las.

É interessante ressaltar que o aprendiz brasileiro de inglês mapeia [?] e [th] em fones já existentes no PB não de modo aleatório, mas de modo a maximizar a semelhança acústica e articulatória. As consoantes [?] e [th] são fricativas interdentais, e como se mostrará a seguir, elas tendem a ser substituídas por consoantes do PB que mantêm o mesmo modo e/ou ponto de articulação, a exemplo de outras fricativas anteriores, como as labiodentais [f] e [v], ou as alveolares [s] e [z]; ou a exemplo das oclusivas alveolares [t] e [d].

Schadech e Silveira (2013) avaliaram o quanto a produção de [?] e [th] por aprendizes afeta a inteligibilidade da mensagem por nativos. As autoras realizaram um experimento em que tocaram gravações de brasileiros aprendizes de inglês, pronunciando palavras contendo [?] e [th], para dez falantes nativos de inglês. Muitas das gravações continham pronúncias com influência de L1 em L2, de maneira que os aprendizes substituíam [?] e [th] por [f], [v], [s], [z], [t] ou [d]. O grau de inteligibilidade foi mensurado pelos nativos através de questionários, em que deviam marcar, em uma escala variando de “muito fácil” a “muito difícil”, qual o grau de inteligibilidade da gravação. Os resultados indicaram que, de acordo com os nativos, a substituição de [?] tem mais impacto na inteligibilidade que a substituição [th]: [?] foi classificado como de compreensão “não muito fácil” e de [th] como “fácil”.

Reis (2006) investigou a produção das interdentais [?] e [th] por aprendizes de proficiência intermediária-baixa e avançada, em tarefas de leitura de sentenças, textos e retelling, constatando baixas taxas de acerto para ambos os fones. Considerando as três tarefas, os sujeitos de nível intermediário-baixo realizaram [?] em 16,6% dos contextos (n = 489) e [th] em apenas 0,1% dos casos (n = 494). Já os de nível avançado conseguiram produzir corretamente [?] em 41,3% dos casos (n = 499) e [th] em 7,5% (n = 610). Embora os resultados tenham se mostrado estatisticamente significativos, a autora salienta que as baixas taxas de realização de [th] podem ter sido enviesadas em virtude do número reduzido de participantes no estudo: havia 16 informantes de proficiência intermediária-baixa e 8 de avançada. De todo modo, ainda que não se possa ter precisão sobre o percentual de acerto dos fones, os resultados indicam que os aprendizes têm altas taxas de erro na produção das interdentais e que se trata, portanto, de uma dificuldade de pronúncia dos aprendizes brasileiros. Reis (2006) verificou substituições de [?] por [t], [f], [d], [t?], [s] e [t?]; sendo as mais frequentes [t] (45,8%), [t?] (7,5%) e [f] (6,9%); e substituições de [th] por [d], [t?], [d?], [d?], [t?], [?] e [t]; as mais frequentes [d] (85,6%) e [t?] (1,4%)[10].

Trevisol (2010) realizou um estudo sobre a produção da interdental vozeada [th] por professores de inglês. O experimento consistiu da leitura de 20 frases, as quais continham o fone [th] em início e final de palavra. Mesmo nessa população de nível avançado em inglês, as interdentais se mostram como uma dificuldade de pronúncia. Em início de palavra, os sujeitos produziram corretamente [th] em 51,4% das vezes, tendo trocado [th] por [d]

nos demais 48,6% casos ( $n = 220$ ). Em final de palavra, a taxa de acerto verificada foi consideravelmente menor: 26,0% ( $n = 208$ ). O maior número de substituições deu-se pela correspondente desvozeada da interdental [?], que contabilizou 65,5% das realizações ( $n = 208$ ); a seguir, o maior número de substituições foi pela oclusiva [d], com 6,0% dos casos ( $n = 208$ ). Trevisol (2010), também registrou substituições por [v], [f], [t], [t?] e [Ø], no entanto, todas elas são de baixa ocorrência (<1,0%) e podem ser consideradas espúrias. A autora explica a predominância de [?] em final de palavra, em virtude de haver restrições, no PB, para que sons fricativos sejam vozeados em final de palavra. Esse processo será tratado à parte, na Seção 2.1.1.4.4.

No que diz respeito ao [?], tal fone constitui uma aproximante alveolar vozeada e está presente em diversos dialetos do PB. Equivocadamente, a literatura linguística no Brasil convencionou a chamar tal som de “r retroflexo”, embora se trate, em termos articulatórios, de uma aproximante (Rennicke, 2011). A Figura 4 apresenta um mapa da distribuição dos róticos no Brasil, indicando as regiões em que ocorre o [?].

[pic]

Figura 4: Distribuição geográfica dos sons róticos em coda no Brasil - Rennicke (2011) com base em Noll (2008).

Como se nota, a maior parte dos dialetos que contém a aproximante [?] está concentrada na região Centro-Sul do Brasil. Apesar disso, Rennicke (2011) afirma haver estudos que indicam a presença da aproximante [?], em menor concentração, em quase todas as regiões do país. Para os dialetos que possuem a aproximante [?] como parte do inventário fonético, sua percepção e produção na interlíngua não constitui problema, de forma que os aprendizes conseguem, por exemplo, pronunciar car ['k??] e word ['w??d] sem dificuldades.

No entanto, nos dialetos do PB em que [?] não ocorre, os aprendizes têm mais um obstáculo a vencer no aprendizado da língua inglesa. Geralmente, eles acabam por realizar substituições, na interlíngua, mapeando a aproximante [?] em um rótico análogo no PB: [x], [?], [h], [?] ou [?] (Zimmer, 2009).

Osborne (2010) pesquisou a aquisição da aproximante [?] por três brasileiros aprendizes de inglês, com conhecimento de inglês em nível iniciante. O experimento consistiu na leitura de sentenças em voz, as quais continham a aproximante [?] em diversos contextos. Para onset complexos, a exemplo da palavra travel ['træv.?l], [?] foi realizado como [?] em 71,7% dos casos, como [?] em 26,4% e omitido em 1,9% ( $n = 53$ ). Em posição intervocálica, como America [??mer.?k?], [?] foi realizado como [?] em 51,7% das vezes e como [?] nos demais 48,3% ( $n = 29$ ). Em posição de coda, como em park ['p??rk] e war ['w??r], a aproximante [?] apresenta o maior número de variação, sendo realizada ora como [?], [?], [x], [h] ou sendo apagado. Osborne (2010) analisa os dados de coda em duas situações: em meio e final de

palavra. No que diz respeito ao meio de palavra, os resultados obtidos com o [?] foram: [h] 57,6%; [?] 18,2%; apagamento 15,1%; [x] 6,1%; e [?] 3,0% ( $n = 33$ ). Em final de palavra, as realizações são similares, mas não se nota a ocorrência do tepe [?]: apagamento 52,5%; [?] 27,5% [h] 15,0% e [x] 5,0% ( $n = 40$ ). O autor também avaliou a realização do [h] em inglês, em palavras como *huge* ['hju?d?], e do tepe [r], como em *city* [?'s???.i]; no entanto, nenhuma variação foi observada, tendo os aprendizes produzido o padrão esperado em 100% dos casos.

8 Falta de aspiração de oclusivas em posição de início de palavra ou sílaba acentuada

Definimos como falta de aspiração de oclusivas a substituição das consoantes [p?], [t?] e [k?] do inglês, em posição de início de palavra ou sílaba acentuada, por suas correspondentes não-aspiradas [p], [t] e [k].

A aspiração é um fenômeno restrito às consoantes obstruintes. Uma consoante é considerada aspirada quando é articulada de modo que, após a fase de explosão dos articuladores, segue-se a liberação de um sopro de ar. Desde Lisker e Abramson (1964), estudos envolvendo comparações entre segmentos vozeados, desvozeados e aspirados têm sido realizados a partir de medidas de voice onset time (VOT). O VOT consiste no intervalo entre a explosão dos articuladores da consoante e o início do vozeamento da glote. A Figura 5 ilustra as combinações de valores de VOT em uma oclusiva bilabial.

[pic]

Figura 5: Tipos de fonação e valores correspondentes de VOT.

Como se observa, assume-se como ponto de referência, ou ponto zero, a soltura dos articuladores da consoante, a partir disso, calculam-se os valores de VOT. Quando há um intervalo entre a soltura dos articuladores e o início do vozeamento da glote, isto é,  $VOT > 0$ , a consoante é classificada como aspirada, no exemplo: [p?]. Quando o vozeamento se inicia imediatamente após a soltura dos articuladores, no caso de  $VOT = 0$ , a consoante é desvozeada: [p]. Por fim, quando o vozeamento antecede a explosão, ou seja, quando há pré-sonorização,  $VOT < 0$ , a consoante é vozeada: [b].

A relação entre as medidas de VOT e tipo de fonação é dependente de língua, por exemplo, um fone que tenha um determinado valor de VOT pode ser considerado aspirado em uma língua, mas desvozeado em outra. A

., baseada nos dados de Yang (1993), apresenta os valores médios de VOT para as consoantes oclusivas do inglês.

Tabela 1: Valores médios de VOT (em ms) de consoantes oclusivas no inglês (Yang 1993).

[pic]

Yang (1993) opta por reportar os valores de VOT para obstruintes vozeadas na forma A/B, em que A contém a média das amostras com valor negativo de VOT e B as com positivo. O autor argumenta que a diferença de sinal no VOT representa fenômenos diferentes: um VOT < 0 corresponde à pré-sonorização, já um VOT > 0 corresponde à pós- sonorização. Tais fenômenos, de acordo com Yang (1993) devem ser tratados distintamente. A partir Tabela 1, nota-se que, na língua inglesa, o VOT está relacionado ao lugar de articulação da consoante, sendo que o par de alveolares [t?] e [d] apresenta os maiores valores médios, respectivamente 95ms e 20/-91ms. As oclusivas velares [k?] e [g] seguem com 88ms e 32/- 78ms. Por fim, os menores valores médios de VOT são encontrados nas bilabiais [p?] e [b], 77ms e 17/-78ms.

Para o PB, a análise mais extensa de VOT de que temos notícia foi realizada por Klein (1999). Os resultados estão resumidos na Tabela 2.

Tabela 2: Valores médios de VOT (em ms) de consoantes oclusivas no PB (Klein 1999).  
[pic]

Como se nota, no PB, as consoantes desvozeadas apresentam menores valores médios de VOT que no inglês, havendo, portanto, menos aspiração. Além disso, a correlação entre o lugar de articulação da consoante e os valores VOT é mais atenuada, a diferença de VOT entre [p] e [t] é estatisticamente insignificante, sendo que apenas a oclusiva velar [k] apresenta algum nível de aspiração. Ao se comparar as medidas de VOT entre os pares do PB e do inglês [p?] vs. [p], [t?] vs. [t], e [k?] vs. [k], é possível ver que as consoantes desvozeadas da língua inglesa apresentam valores bem mais altos de VOT, o que evidencia, de fato, sua aspiração.

Salienta-se que, diferentemente de Yang (1993), Klein (1999) agrupa os valores de VOT positivos e negativos das consoantes vozeadas, de forma que a média apresentada das vozeadas acaba por se tornar menor. Ainda assim, é possível notar que as oclusivas vozeadas do PB apresentam mais pré- sonorização que as do inglês. A maior diferença é verificada na oclusiva velar, no PB, seu valor médio é de -91ms, enquanto no inglês é de 32/-78ms. A seguir a maior diferença é notada na bilabial [b], -87ms vs. 17/-78ms; por fim, os menores valores são encontrados na alveolar [d], -99 vs. 20/- 91ms. Por conseguinte, conclui-se que as oclusivas vozeadas [b], [d] e [g] do PB possuem mais vozeamento que as do inglês.

Além disso, é possível observar certa sobreposição entre os valores positivos de VOT das oclusivas vozeadas no inglês e das oclusivas desvozeadas no PB. Por exemplo, o valor de VOT positivo de [b] no inglês, 18ms, é bastante similar ao valor de [p] no PB, 17ms. Isso é válido também para os demais lugares de articulação: o valor de VOT da alveolar vozeada [d] do inglês, 20 ms., é muito similar ao da desvozeada [t] do PB, 17 ms.; e o mesmo para as velares [g] e [k], 32ms vs. 38ms. Em outras palavras, as consoantes desvozeadas no PB são articuladas, em certos contextos, de modo muito similar às vozeadas no inglês. Isso

pode trazer problemas na inteligibilidade da pronúncia do aprendiz, uma vez que, caso ele transfira esse padrão para a interlíngua, pode, por exemplo, tentar pronunciar [k] e acabar sendo entendido como [g], em palavras como *caught* ['k??t] e *got* ['g??t]; *coat* ['ko?t] e *goat* ['go?t], etc. Cabe, portanto, ao aprendiz dominar essa diferença, de modo a aumentar a inteligibilidade de sua pronúncia.

Alves (2011) investigou a produção das oclusivas aspiradas [p?], [t?] e [k?] por brasileiros aprendizes de inglês. A autora utilizou valores de VOT na classificação, tendo definido como aspirados os segmentos que apresentavam VOT > 60ms e não-aspirados os que apresentavam VOT < 35ms. A autora observou que [p?] foi realizado sem aspiração pelos aprendizes em 59% das vezes (n = 41), [t?] em 33% das vezes (n = 51) e [k?] em 10% das vezes (n = 96). A diferença no desempenho pode ser explicada em virtude do método de classificação utilizado: apenas duas faixas de valores, VOT < 35ms ou VOT > 60ms; e também em virtude de, no PB, a consoante [k] já possuir maior VOT, dado o lugar de articulação. Entretanto, o estudo contou com um número muito reduzido de participantes (três), de modo que os resultados obtidos devem ser considerados apenas indicativos e não conclusivos.

Prestes (2012) investigou a produção de oclusivas surdas e sonoras por aprendizes brasileiros e falantes nativos de inglês. A autora emprega medidas de VOT na classificação dos segmentos e adota uma postura de tratar o fenômeno em sua gradiência, isto é, considerando-se apenas as medidas de VOT, sem dizer se uma determinada consoante é ou não aspirada. Valores relativos de VOT foram utilizados na análise, mais especificamente, a razão entre a duração do VOT e o tempo total da consoante. Prestes (2012) concluiu que as realizações surdas dos aprendizes apresentaram menor VOT (5,87% VOT/consoante) em relação às dos nativos (2,11% VOT/consoante); e que as vozeadas dos aprendizes apresentam maior VOT (10,15% VOT/consoante) em comparação com as dos nativos (3,89% VOT/consoante) (n = 450). Os resultados indicam, portanto, que brasileiros tendem a realizar [p], [t] e [k], na interlíngua, com menos aspiração que falantes nativos; e [b], [d] e [g] com maior grau de vozeamento. Ressalta-se, todavia, que o estudo também utilizou um número reduzido de participantes, dois brasileiros aprendizes de inglês e dois nativos.

Schwartzhaupt (2012) analisou o impacto que fatores fonético-fonológicos podem ter sobre os valores de VOT, seu experimento foi conduzido com dez brasileiros aprendizes de inglês e cinco nativos. Foram analisados os seguintes fatores fonético-fonológicos: (i) lugar de articulação da consoante, (ii) qualidade da vogal adjacente e (iii) o número de sílabas da palavra-alvo. Os aprendizes possuíam conhecimento intermediário-avançado ou avançado de inglês. Apesar de o foco do estudo não ser a produção dos aprendizes, mas os efeitos dos contextos fonético-fonológicos no VOT, Schwartzhaupt (2012) observou que os aprendizes foram capazes de realizar, na interlíngua, valores de VOT bem próximos aos dos nativos. Ele

salienta o fato de os aprendizes terem conseguido partir de um sistema sem distinção de VOT entre [p] e [t], como é o caso do PB, e terem alcançado, na interlíngua, um sistema em que tal distinção é patente, como é o caso do inglês. Na interlíngua, as realizações de [t?] (77ms) dos aprendizes apresentaram VOT significativamente maior que as de [p?] (61ms), tal como ocorre com falantes nativos.

9 Desvozeamento de obstruintes em posição de final de palavra

O m XXXXX

10 Vocalização de laterais em final de sílaba

Baratieri (2006) investigou a produção da lateral [l] em coda por um grupo de vinte estudantes de inglês como língua adicional, de proficiência intermediária e avançada. Os resultados obtidos indicaram que a vocalização se trata de um fenômeno gradiente, tendo o autor optado por classificar os dados em três categoriais: (a) segmento parcialmente vocalizado; (b) vocalizado [w]; e (c) não vocalizado [l]. A categoria (a) indica os segmentos para os quais o grau de vocalização não é imediatamente perceptível, de maneira que não há um símbolo IPA correspondente. As produções dos aprendizes apresentaram a seguinte distribuição: 2,7% de [l], 35,5% de [w] e 61,8% de segmentos parcialmente vocalizados ( $n = 2134$ ). Como se observa, a taxa de produção do padrão esperado [l] foi extremamente baixa, indicando que a vocalização de laterais perdura mesmo em estudantes de proficiência intermediária ou avançada.

11 Quada da nasal em final de sílaba + nasalização da vogal precedente

Kluge e Baptista (2008) estudaram a produção das nasaís [m] e [n] em posição final de palavra por um grupo de dez aprendizes de nível intermediário de proficiência. A tarefa consistiu na leitura de sentenças em voz alta. O corpus foi formado por 72 sentenças, 36 contendo [m] e 36 contendo [n], em posição final de monossílabo acentuado do tipo (C)CVC. Os estudantes realizaram a nasal alveolar [n] final esperada em 78,6% dos casos ( $n = 359$ ), e a nasal bilabial [m] esperada em 63,9% dos casos ( $n = 357$ ); nos demais, houve apagamento da nasal e a nasalização da vogal precedente. Testes estatísticos indicaram que a diferença de acerto nas realizações de [m] e [n] é significativa, havendo, portanto, mais dificuldade na aquisição do [m] final pelos aprendizes. As autoras justificam que esse resultado pode se dar em virtude de, no PB, as palavras que terminam em nasal tenderem a ser escritas com (ex.: “fim”, “correm”, “amam”), havendo poucas palavras com (ex.: “hífen”, “pólen”, “abdômen”). Sendo assim, o padrão com é reforçado e o aprendiz apresenta mais resistência para superá-lo na interlíngua.

12 Paragoge da consoante oclusiva velar vozeada [g]

O XXXXX

13 Assimilação vocálica

O XXXXX

14 Epêntese interconsonantal (morfema -ed)

O XXXXX

## 2.2.2 Why Consider L1 in L2 Teaching?

Lorem ipsum at nusquam appellantur his, ut eos erant homero concludaturque. Albucius appellantur deterruisset id eam, vivendum partiendo dissentiet ei ius. Vis melius facilisis ea, sea id convenire referrentur, takimata adolescens ex duo. Ei harum argumentum per. Eam vidit exerci appetere ad, ut vel zzril intellegam interpretaris.

Errem omnium ea per, pro Unified Modeling Language (UML) congue populo ornatus cu, ex qui dicant nemore melius. No pri diam iriure euismod. Graecis eleifend appellantur quo id. Id corpora inimicus nam, facer nonummy ne pro, kasd repudiandae ei mei. Mea menandri mediocrem dissentiet cu, ex nominati imperdiet nec, sea odio duis vocent ei. Tempor everti appareat cu ius, ridens audiam an qui, aliquid admodum conceptam ne qui. Vis ea melius nostrum, mel alienum euripidis eu.

## 2.2.3 Common Mispronunciation

**Syllable Simplification** Lorem ipsum

**Consonant Change** Lorem ipsum

**Deaspiration of Voiceless Plosives in Initial or Stressed Positions** Lorem ipsum

**Terminal Devoicing in Word-Final Obstruents** Lorem ipsum

**Delateralization and rounding of lateral liquids in final position** Lorem ipsum

**Vocalization of final nasals** Lorem ipsum

**Velar consonantal paragoge** Lorem ipsum

**Vowel assimilation** Lorem ipsum

**Interconsonantal epenthesis (-ed and -s morphemes)** Lorem ipsum

## 2.3 Automatic Speech Recognition

### 2.3.1 The big picture

Basically, all statistical methods of ASR are dedicated into solving one fundamental equation, which can be described as follows. Let  $O$  be a sequence of observable acoustic feature vectors and  $W$  be a word sequence, the most likely word sequence  $W^*$  is given by:

$$W^* = \arg \max_W P(W|O) \quad (2.1)$$

To solve this equation straightforwardly, one would require a discriminative model, capable of estimating the probability of  $W$  directly from a set of observations  $O$  [? ]. However, HMM are generative models and are not adequate for solving this equation, therefore we apply Bayes' Theorem to Equation 2.1 and end up with:

$$W^* = \arg \max_W \frac{P(O|W)P(W)}{P(O)} \quad (2.2)$$

As one might notice, we can apply a generative model to calculate the conditional probability term of this equation, that is, the probability of the observation sequence  $O$  given a word sequence  $W$ , hence  $P(O|W)$ . At first, it might seem counter-intuitive to conceive a generative model for data analysis, since the data is already available, i.e.  $O$  is known before-hand. As ? ] [? ] points outs, in order to understand how to generative models are used for data analysis, a mental trick is necessary.

[To fix - Fink citation]First one assumes, that the data to be analyzed were generated by a natural process, which obeys similar statistical regularities. Then one tries to reproduce this process with the capabilities of hidden Markov models as closely as possible. If this attempt is successful, on the basis of the artificial model inferences can be drawn on the real process. On the one hand this may concern the probability for generating the available data. On the other hand the inference on the internal processes within the model is at least probabilistically possible. In particular one can determine the state sequence that generated a certain sequence of outputs with highest probability.

For a single audio input, which we want to decode, the audio is already fixed, so the probability of the observable acoustic feature vectors  $P(O)$  is a constant and, therefore, might be discarded. Thus the final fundamental equation is simplified to:

$$W^* = \arg \max_W P(O|W)P(W) \quad (2.3)$$

$P(O|W)$ , the probability of an observable acoustic feature vector given a word sequence, is calculated by an acoustic model. In turn,  $P(W)$ , the *a priori* probability of words is reckoned by a language model.

### 2.3.2 HMM

Markov models consist a set of mathematical models which are suitable for the statistical description of symbol and state sequences [? ]. The simplest form of Markov models are Markov chain models, which represent a system with a set of spaces in which transitions from one state to another occur. Within Markov processes, systems are assumed to be memoryless, that is, the conditional probability of future states is only dependent on the present state. To put it another way, Markov models assume that, given a certain system with states and transitions, the current state does not depend upon the sequence of events that preceded it, the so-called Markov property.

HMMs can be formally described as a 5-tuple  $\lambda = (Q, O, \Pi, A, B)$ , where  $Q = \{q_1, q_2, q_3, \dots, q_N\}$  is a set of  $N$  states.  $O = \{o_1, o_2, o_3, \dots, o_T\}$  is a set of  $T$  observations taken from time  $t = 1$  to  $t = T$ . At each time  $t$  it is assumed that the system will be at a specific state  $q$ , which is hidden, only the observations are directly visible.  $\Pi = \{\pi_i\}$  is a vector with the initial state probabilities, such that

$$\pi_i = Pr(q_i), t = 0 \quad (2.4)$$

$A = [a_{ij}]$  is matrix with the state transition probabilities so that

$$a_{ij} = P(q_t = j | q_{t-1} = i), 1 \leq i, j \leq N \quad (2.5)$$

and  $B = [b_{jt}]$  is a matrix with the emission probability of each state. Assuming a GMM to model the state emission probabilities – the so-called GMM/HMM model; we can define that, for a state  $j$ , the probability  $b_j(o_t)$  of generating  $o_t$  is given by

$$b_j(o_t) = \prod_{s=1}^S \left[ \sum_{m=1}^{M_{js}} c_{jsm} \mathcal{N}(o_t; \mu_{jsm}, \Sigma_{jsm}) \right]^{\gamma_s} \quad (2.6)$$

where  $\gamma_s$  is a stream weight, with default value is one,  $M_{js}$  is the number of mixture components in state  $j$  for stream  $s$ ,  $c_{jsm}$  is the weight of the  $m^{\text{th}}$  component and  $\mathcal{N}(\cdot; \mu_{jsm}, \Sigma_{jsm})$  is a multivariate Gaussian with mean vector  $\mu$  and covariance matrix  $\Sigma$ , that is

$$\mathcal{N}(o; \mu, \Sigma) = (\sqrt{(2\pi)^n |\Sigma|})^{-\frac{1}{2}(o-\mu)^T \Sigma^{-1} (o-\mu)} \quad (2.7)$$

where  $n$  is the dimensionality of  $o$ .

The following constraints apply:

$$a_{ij} \geq 0 \quad (2.8)$$

that is, the probability of moving from state from any state  $i$  to  $j$  is not null, and

$$\sum_{j=1}^N a_{ij} \geq 1, \forall i \quad (2.9)$$

### 2.3.3 PB

#### 15 Reconhecimento Automático de Fala (RAF)

O propósito de um reconhecedor de fala é transformar, de forma eficiente e precisa, o sinal acústico da fala em sua contraparte textual (Rabiner & Schafer, 2007). Se cada palavra da língua fosse pronunciada de forma idêntica por todos os falantes e em todos os contextos, a tarefa de Reconhecimento Automático de Fala (doravante RAF) seria algo banal. Mas isso não acontece: a realidade linguística é demasiado variante, quer inter-, quer intrafalantes. Pode-se, por fim, dizer que uma vogal nunca é pronunciada de uma mesma maneira. Furui (2001) sumariza os problemas do RAF em quatro: (i) dificuldades em lidar com coarticulação e redução; (ii) dificuldades de segmentação da fala; (iii) diferenças inter-individuais; e (iv) insuficiência de conhecimento linguístico. A seguir, cada um desses pontos serão discutidos.

A coarticulação é um fenômeno motor que envolve a realização de gestos articulatórios simultâneos ou sobrepostos (Crystal, 2008). Na fala, a coarticulação pode ocorrer de duas formas: antecipatória (left-to-right) ou preservativa (right-to-left). Na coarticulação antecipatória, um som é realizado tomando as características de outro que lhe sucede. Em “pulo”, por exemplo, a consoante /p/ é realizada com protrusão labial, [p?], em virtude de a vogal que lhe segue, [u], ser arredondada. Por sua vez, na coarticulação preservativa, um som é realizado mantendo-se as características articulatórias de outro que lhe antecede. Em dialetos em que não houve a vocalização da lateral, por exemplo, a consoante /l/ da palavra “sal” pode ser realizada de modo velarizado, [?] em vez de [l], por suceder a vogal [a], que é baixa. De fato, a todo o tempo, os sons da fala estão sujeitos à coarticulação. Por tal razão, frequentemente, utilizam-se no RAF unidades fonéticas dependentes de contexto, a exemplo de trifones.

Um trifone é uma representação fonética contextual, que considera, para um dado fone, o fone anterior e o seguinte. A palavra ‘fala’, por exemplo, pode ser representada, no IPA, pela sequência de fones [’fal?'] ou de trifones [#’fa fal al? l?#], especificando-se as articulações secundárias dos fones. No reconhecimento de fala, por razões de codificação em computadores, em geral, os trifones são especificados no formato “L-X+R”, em que “X”

é um determinado fone, “L” o fone antecedente, e “R” o sucessor. A Figura 6 compara a transcrição da palavra “translate” em fones e trifones.

[pic]

Figura 6: Transcrição da palavra “translate” em fones e trifones de acordo com a convenção do VoxForge (2013).

Quanto às diferenças interindividuais, elas são tamanhas que há até mesmo todo um ramo de investigação da Linguística que se dedica a seu estudo: a Sociolinguística. A Sociolinguística estabelece que a realização linguística é condicionada por diversos fatores sociais: o nível escolar do falante, seu sexo, sua idade, seu estrato social, o lugar onde nasceu e viveu, a situação comunicativa em que está inserido, a formalidade de registro que o contexto demanda, o grau de hierarquia que mantém com seu interlocutor, etc. (Labov, 2008; Weinreich et al., 2006). Para além dos fatores que condicionam a variação, há fenômenos linguísticos que possuem comportamento puramente estocástico: o alçamento vocálico acontece em certas vezes e não em outras, mesmo havendo o mesmo ambiente e considerando- se o mesmo falante.

Não bastasse os fatores sociais, a fala sofre influência também de características anatômicas do indivíduo. A variabilidade linguística começa já quando somos crianças, na fase de desenvolvimento puberal. Das primeiras palavras que balbuciamos até o período de muda vocal, ocorrem grandes mudanças na voz tendo em vista, especialmente, modificações nas configurações das estruturas laríngeas (Guimarães, 2006). Perry et al. (2001) conduziram um estudo sobre vogais produzidas por crianças dos 2 aos 16 anos e verificaram que há diferenças significativas nos valores de frequência fundamental [pic] e nos formantes das vogais [pic]. Ao longo de todo o período de maturação de voz, os valores de frequência dos formantes das vogais [pic] varia significativamente e é distinto para meninos e meninas; após os 12 anos, há grande decréscimo nos valores de frequência fundamental [pic] para os meninos, de modo que a frequência fundamental [pic] passa também a servir perceptualmente para a distinção do sexo das crianças (Perry et al., 2001) (Figura 7).

[pic]

Figura 7: Comparação dos valores de ??0, F1, F2 e F3 para crianças de ambos os sexos dos 2 aos 16 anos (Perry et al., 2001).

Aos sistemas de RAF, cabe a tarefa de reconhecer, ante toda a variabilidade linguística existente, o que há de invariante na fala. Do ponto de vista computacional, a construção de um reconhecedor de fala pode ser vista como uma tarefa que abrange quatro fases principais: preparação de dados, treino, teste e análise (Young, et al., 2006). A preparação de dados consiste na gravação de arquivos de áudio e texto, em sua anotação, transcrição e pré-processamento, de modo a garantir que a entrada seja compatível com o esperado pelo

reconhecedor. O treino constitui a fase em que os dados coligidos são utilizados para criar as componentes do reconhecedor. Em geral, os sistemas de RAF possuem três componentes: (i) um modelo de língua, (ii) um modelo acústico e (iii) um modelo de pronúncia, ou dicionário de pronúncia. Por fim, as fases de teste e análise buscam verificar se os modelos construídos se adéquam à realidade linguística ou ao propósito a que o reconhecedor se dispõe.

#### 16 Representação Digital da Fala

Para desenvolver tecnologias de fala, é preciso antes encontrar formas de codificar, computacionalmente, a informação presente na fala. A fala humana, como constitui um veículo de informação, pode ser vista a partir de uma perspectiva da Teoria Matemática da Comunicação (Shannon, 1948)[11], sendo considerada um sinal acústico. Tecnicamente, sinais são sequências de estados em um sistema de comunicação que codificam uma mensagem.

Sinais constituem o objeto central de estudo da área de Processamento de Sinal, que busca investigar formas de analisar ou modificar os sinais, no intuito de deles extrair informação ou de adequá-los a um determinado fim (Ingle e Proakis, 2011[12]). Embora a maior parte dos sinais que nos rodeia seja analógica, em muitas vezes, seu processamento é feito não de forma analógica, mas digital. De acordo com Ingle e Proakis (2011)[13], os sistemas de Processamento de Sinal Digital (DSP) são vantajosos, pois:

– não necessitam a aquisição de equipamento específico (que são, muitas das vezes, caros), podendo ser desenvolvidos em computadores pessoais;

– baseiam-se apenas em operações numéricas de adição e multiplicação, o que lhes dá estabilidade, não sendo necessário calibragem ou padronização, como é comum nos sistemas analógicos;

– são altamente adaptáveis, de modo que suas operações podem ser modificadas em tempo real, usualmente, através de técnicas de programação simples.

Um esquema de um sistema de processamento de sinal digital é apresentado na Figura 8.

[pic]

Figura 8: Esquema de um sistema de processamento de sinal digital (Ingle e Proakis, 2011).

Embora a maior parte dos sinais do mundo seja analógica - entre os quais a fala, seu processamento, muitas vezes, dá-se gera sinais analógicos,

Embora a maior parte dos sinais com que

O processamento de sinais pode ser feito de forma analógica ou digital.

A área do conhecimento responsável por tal busca é o Processamento de Sinal.

A área responsável por esse

Do ponto de vista da Teoria da Informação, sinais são sequências de estados em um sistema de comunicação que codificam uma mensagem.

(Ingle e Proakis, 2010)

### 17 Tipos de Reconhecedores de Fala

Reconhecedores Automáticos de Fala podem ser agrupados em três categoriais, que se dividem quanto à tarefa de reconhecimento que desempenham: i) reconhecimento de palavras isoladas; ii) de sentenças pré-estabelecidas; iii) reconhecimento de fala contínuo de grande vocabulário (RFCGV)[14] (Rabiner, 1997).

Os sistemas de reconhecimento de palavras isoladas são utilizados, por exemplo, por centrais telefônicas, nas Unidades de Resposta Audível (URA), em menus do tipo: “Fale ‘um’ para ser redirecionado ao setor de cancelamento” ou “Para conversar com um de nossos atendentes, fale ‘atendente’ ”. Reconhecedores de fala do segundo tipo são mais robustos que os do segundo, sendo capazes de reconhecer sentenças pré-definidas ou provenientes de uma gramática pré-estabelecida. Um exemplo de aplicação deste tipo são os módulos de comando por voz de computadores, celulares e os sistemas hands-free em carros, em que se pode dizer sentenças do tipo “ligar o rádio” ou “Siri, google ‘speech recognition’ ”e o comando é reconhecido. Já os sistemas de reconhecimento de fala contínuo de grande vocabulário são os reconhecedores mais abrangentes, sendo capazes de processar a fala espontânea do usuário. Sistemas de diálogo por voz, como os How May I Help You (HMIHY), e sistemas de ditado em editores de textos são desse último tipo.

A seguir, será discutida apenas a arquitetura dos sistemas de RFGV, primeiro, porque consistem nos sistemas de reconhecimento com a arquitetura mais complexa e, segundo, porque é o tipo de reconhecedor que se pretende desenvolver neste projeto.

### 18 Arquitetura Básica de um Reconhecedor de Fala Contínuo de Grande Vocabulário (RFCGV)

O paradigma majoritário em sistemas de RAF é estocástico, destacando-se, especialmente, a utilização de Modelos Ocultos de Markov, ou Hidden Markov Models (HMM) (Huang, et al., 2001). Em tais modelos, a tarefa de reconhecimento é considerada a partir da metáfora do canal ruidoso, ou noisy-channel (Jurafsky & Martin, 2009). O sinal acústico, que constitui a entrada no sistema, é visto como uma deformação da mensagem original, isto é, da sequência de palavras pretendida pelo falante, após passar por um canal com ruído. Assim, o reconhecimento se torna uma tarefa de decodificação, isto é, trata de como recuperar a mensagem original a partir do sinal acústico “ruidoso”. Matematicamente, isso corresponde a estimar, considerando-se uma língua [pic], para uma sequência de palavras [pic], qual é a sequência [pic] mais provável, dado conjunto de estados acústicos observáveis [pic]:

[pic]

Todavia, não é possível calcular [pic] diretamente, sendo necessário aplicar-se o Teorema de Bayes, de modo a obter-se:

[pic]

Como a propósito é buscar a sequência de palavras mais provável para um conjunto já dado de estados acústicos, [pic] se repete a cada cálculo, de maneira que pode ser considerado uma constante de normalização, e a equação pode ser simplificada para:

[pic]

Essa equação fundamenta a base dos sistemas de RAF estocásticos e possui estreita relação com a arquitetura que é por eles compartilhada. Basicamente, os sistemas de RAF contínuo com grande vocabulário possuem três módulos: (i) um modelo de língua, (ii) um modelo acústico e (iii) um modelo, ou dicionário de pronúncia. O modelo de língua é utilizado para estimar [pic], a probabilidade a priori da sequência de palavras. Já o modelo acústico é utilizado para calcular [pic], a verossimilhança da observação. Por fim, o dicionário de pronúncia serve como uma ponte entre o modelo de língua e o modelo acústico, uma vez que possui as palavras que compõem o léxico do reconhecedor, transcritas em forma ortográfica e fonética. A Figura 9 ilustra a arquitetura básica de um sistema de RAF.

[pic]

Figura 9: Arquitetura básica de um Reconhecedor Automático de Fala.

O modelo acústico processa o sinal acústico da fala, de modo a inferir quais são os segmentos sonoros que a compõem, usualmente, empregando fones ou trifones. Em reconhecedores de base em HMM, essa tarefa é feita estimando-se os estados acústicos observados mais prováveis, bem como suas probabilidades de transição. Já o modelo de pronúncia provê a correspondência entre sequências de fones e as palavras da língua. No exemplo, tal modelo mapeia a sequência de fones [fal?] na palavra “falo”. O modelo de língua, por sua vez, estima as ordenações de palavras mais prováveis na língua.

### 2.3.4 A Brief History

Although the task of recognizing words from speech might seem apparently simple beforehand, (after all humans start doing it with as little as four months! (XXXX INSERT CITATION)), the issue is actually very complex one. Over the years, many methods have been proposed to attempt to solve the problem of speech recognition. However until now no solution has been found and machines are still a very long way from performing like humans. Table 2.7 presents a comparison between the performance of humans and machines in some recognition tasks.

As one may observe, humans outperform machines in almost every task, specially the more complex ones. Humans' are indeed the topline for the speech recognition task, the uttermost dream of each speech scientist alive is to build a system capable of performing

Table 2.7 Word error rate comparisons between human and machines on similar tasks [60].

Tasks	Voc. size	Humans	Machines
Connected digits	10	0.009%	0.720%
Alphabet letters	26	1%	5%
Spontaneous telephone speech	2,000	3.8%	36.7%
WSJ with clean speech	5,000	0.9%	4.5%
WSJ with noisy speech (10-db SNR)	5,000	1.1%	8.6%
Clean speech based on trigram sentences	20,000	7.6%	4.4%

similarly to humans. Although this dream is somewhat near for rather simple tasks like connected digits, for other contexts a long path yet lies ahead.

Recognizing spontaneous speech is still a huge barrier for machines, as can be seen from the huge difference in the spontaneous telephone corpus: whereas humans had a WER of 3.8%, for machines this rate is up to 36.7%. Such result is mainly due to linguistic variability. Language varies not only among speakers (the so-called inter-speaker variability), but also within the same speaker (intra-speaker) [8].

Considering inter-speaker differences, factors such as gender, age, social, and regional origin, health and emotional state might have a huge impact on the speech signal [8]. Sociolinguistics has long known that gender affects language usage, in fact, men and women tend to use different language constructions. In her seminal paper in the field, Lakoff [69] [69] found that, in women's speech, strong expressions of feeling are avoided, uncertainty is favored, and means of expression in regard to subject-matter deemed "trivial" to the "real" world are elaborated.

Put aside social aspects, men, women and children's speech are also contrasting because of morphological differences in their vocal tract. Sex and development influence body size, and there is a strong correlation between vocal tract length and body size (either height or weight); in addition to this, the relative proportions of men and women's oral and pharyngeal cavity are unlike [42]. Figure 2.4 presents a comparison between height and vocal tract length for men, women and children. Figure 2.5 presents a model of the vocal tract morphology considering age.

As one can observe, men's vocal tract are longer than women's, followed and obviously children. These morphology differences affect the speech signal thoroughly, specially in what concerns to the Fundamental Frequency (F0). F0 can be defined as the lowest frequency in

the signal counting from zero. Figure 2.6 compares the F0 values between male and females considering aging.

One can notice from Figure 2.6, that no difference is found between male and female voice at a very young age. In fact, boys and girls have roughly the same F0 values. However, when they reach puberty, differences begin to appear. This period is commonly called the voice mutation or voice change, when the F0 for male voice has huge drop, while for female voice the drop is quite small. In terms of perception, this is period when the male voice lowers and gets deeper.

#### XXXXXXXXXXXXXXXXXXXX EXTEND INTER-SPEAKER VARIABILITY XXXXXXXXXXXXXXX DESCRIBE INTRA-SPEAKER VARIABILITY

But let's get back to Table 2.7. It is interesting to notice that humans performed better in all speech recognition tasks, but one: "Clean speech based on trigram sentences". This one task consists of recognizing sentences which were randomly generated using the WSJ trigram language model. Therefore, humans had no advantage over machines in what concerns to syntactic or semantic knowledge. This result highlights one of the most important feature of human hearing, that is, we make a large use of syntactic, semantic and also pragmatic information in order to understand speech. While hearing do not take into account simply the acoustic signal, but the whole context.

Language is a social tool, which aims at successfull interaction. When someone steps into a snack bar and orders an [ais'krim], the vendor has no doubt that this sequence of phones refers to "ice cream" and not "I scream", albeit they are pronounced exactly the same. Although such sequence of phones might be ambiguous in the phonetic level, it is not in higher linguistic levels, such as syntax (the verb "order" is usually followed by a noun), semantics (the object of order has to be something purchasable) and pragmatics (one does not buy his own shout!).

### 2.3.5 HMM-based Speech Recognition

HMM is the most widespread and successfull paradigm in ASR. When HMM were first applied to speech recognition in the late 70's, they were completely revolutionary. Up until recently Deep Neural Network (DNN) seem to be next prominent paradigm in ASR. HMM have been applied to ASR since the late 70's, and they have gathered the best results until recently.

A HMM is a statistical Markov model in which the states are assumed to be hidden, i.e. they are not directly visible, only the state's outputs are observable. Each state has a probability distribution over the possible output tokens, in such a way that output generated

by the HMM states provides some information about the hidden sequence of states which was traversed.

### 2.3.6 Feature Extraction

**Preambulus** Feature extraction is an importart part of speech recognition systems. The feacture extraction phase is responsible for identifying or enhancing the components of the signal that are relevant for recognizing speech sounds, while discarding or diminishing the effect of unuseful information, such as background noise. With respect to speech parameterization, Mel Frequency Cepstral Coefficients (MFCC) are definitely the standard. MFCC have been widely used in ASR systems for almost three decades [29], they are present on the many important speech recognition toolkits, such as Hidden Markov Model Toolkit (HTK), Sphinx, RASR (RASR) and Kaldi. Before we go into further details about these features it is interesting to give a little background about speech recording and coding.

Speech is recorded by using a microphone – nothing new so far! Despite the many types of available microphones (condenser, capacitor, pyezoelectric, laser, etc.) its design remains basically the same as the carbon microphone invented by David Hughes two centuries ago [101]. A microphone is simply an acoustic-to-electric sensor, which converts variations in air pressure (that is, sound) into an electrical signal. Microphones have a very thin membrane, called diaphragm, which vibrates when struck by sound waves. When the diaphragm vibrates, it puts to move a sensitive capsule attached to it, that converts its movement into electrical pulses. Most of the current microphones are dynamic, which means that their capsule consist of .... (XXX VER WIIPEDIA)

After capturing speech through a microphone, one usually wants to store it for later access. In order to store speech digitally on a computer, a coding scheme is mandatory. In the literature, many coding schemes have been proposed, such as linear PCM,  $\mu$ -law, A-law PCM, APCM, DPCM, DM, and ADPCM [60]. The details of each type of speech coder is beyond the scope of this dissertation, the reader can find an description of each scheme in Huang et al. [60] [60] or Furui [46] [46]. Following we will give a brief discussion of linear PCM, which is the standard way of storing audios in digital format.

Pulse Code Modulation (PCM) is a type of analog-to-digital conversion, which constitutes the basis of the WAV digital audio format, together with other lossless formats such as AIF and AU.<sup>4</sup>. PCM coding is based on two properties: (i) a sampling rate of the audio and a (ii) bit depth. The sampling rate determines the number of audio samples that are taken per second from the signal, in turn the bit depth is the number of bits of information in each audio

---

<sup>4</sup>Other types of popular audio files which use lossy data compression, such as MP3, WMA, OGG or AAC (a format common to DivX videos) do not use PCM. Instead

sample. Both values must be constant and should be defined prior to recording (actually coding) an audio. The sampling and the bit depth are closely related to the audio quality, that is, the higher the sampling and the depth the better the fidelity of the digital audio to the analog speech signal. Picture XXX presents an example of a linear PCM representation, at different sampling rates and bit depths, of an audio containing the utterance “Speech recognition”.

Linear PCM assumes that the discrete signal  $x[n]$  is bounded, that is,

$$|x[n]| \leq X_{max} \quad (2.10)$$

and that the quantization step  $\Delta$  is uniform for all consecutive levels of  $x_i$

$$x_i - x_{i-1} = \Delta \quad (2.11)$$

Assuming a binary code, the number of levels which can be represented by PCM is  $N = 2^B$ , where  $B$  is the bit depth, this constitutes the audio resolution. According to [60], speech could be represented in an intelligible way by using 7 bits, however, in practice, applications use values no lower than 11 bits to guarantee communication efficiency. For instance, CDs makes use of 16-bit linear PCM, whereas DVD-Audio and Blu-Ray discs can support up to 24-bit.

Although linear PCM files are able to carry all the necessary auditory information – after all we are able to listen to them and recognize the speech, the music or the noise recorded in them; they are not useful for speech recognition purposes. This occurs because, from the phonological point of view, very little can be said based on the waveform itself [106]. Consider, for instance, the two combinations of 100 Hz, 200 Hz and 300 Hz sine waves, shown in Figure 2.7, which differ only with respect to the relative timing.

As one might notice, disregard of being composed by the same pure tones, the complex waves shown in Figure 2.7 are completely distinct from one another. This happens because the waveform is influenced by phase shifts (also known as phase offsets). Therefore in-phase and out-of-phase waves (Figure 2.8 and Figure 2.9) are represented differently, and this adds too much variability to the waveform, in such way that the signal waveform becomes unsuitable for human analysis and consequently for being used as a raw input in ASR systems.

Another way of representing the audio information, which is more meaningful for human reading or computer analysis is through short-term spectrum. Short-term spectra are obtained by applying a Discrete Time Fourier transform to a windowed signal. At first, the signal is divided into uniformly-spaced periods with a sliding window. For speech recognition,

usually the window size is defined as 25 ms, with a frame shift of 10 ms, audio information is extracted every 10 ms with 15 ms of overlapping among adjacent frames [60]. Figure 2.10 contains an example of a windowing process (in this case, with 50% overlapping).

These windows values are based on two assumptions: (i) that within 25 ms the signal is stationary, i.e. the phonatory system is not moving; (ii) that at least a period of each relevant speech frequency will be captured by this window this windows, that is no relevant are

After windowing the signal a Fourier transform is applied into each window so as to obtain a series of frequency spectra, i.e. a series of representation of the signal in the frequency domain instead of the time domain. As can be noticed in Figure 2.10, since the frame shift is smaller than the window size, the windowing process extracts many redundant information. The intention for doing this will be made afterwards, when we give further details of the Fourier transform. Such transform is based on the Fourier theorem, which states that any periodic waveform can be approximated as closely as desired as the sum of a series of pure sine waves. In other words, the Fourier transform is able to analyse a short-term of the signal, containing a complex wave, and to output which are and what is the amplitude of the pure tones which form this complex wave.

Feature extraction must then be performed in stored audio files in order to extract relevant information from the waveform and discard redundant or unwanted signal characteristics. As already mentioned before, the two most traditional techniques for speech feature extraction, over the past decades, have been the MFCC [29] and the Perceptual Linear Prediction (PLP) [58]. Both parameterization methods are based on the short-term spectrum of speech. For speech recognition purposes, MFCC features usually show better performance when compared to PLP, for this reason in this thesis we are only going to present MFCC features [? ? ].

### 2.3.7 MFCC Features

MFCC is a type of speech parameterization is the result of a cosine transform of the logarithm of the short-term energy spectrum expressed over a mel scale [29]. MFCC features tries to reduce the feature dimensionality of a sound Fourier spectrum, by applying some concepts of Psychoacoustics and Psychophysics in order to extract a vector with relevant values from the spectrum. The aim is to represent speech data in a compressed format, by eliminating information which are not pertinent to the phonetic analysis and to enhance the aspects of the signal which contribute to the detection of phonetic differences [29].

From Psychoacoustics, MFCCs use the notion that humans do not perceive frequency through a linear scale, but through a scale which resembles to be linear-spaced in frequencies

below 1000 Hz and logarithmic in frequencies above 1000 Hz<sup>5</sup>, the so-called mel scale (named after *melody*). The scale is based on experiments with simple tones in which individuals are required to separate frequency values into four equal intervals or to adjust the frequency of a stimulus to be half as high as another reference tone [60]. The reference point between mel scale and a linear frequency scale is 1000 mels, which correspond to a 1000 Hz tone, 40 dB above the absolute threshold of hearing. Since it was first introduced by Stevens et al. [112] [112], the scale has been revisited many times [119], but a common formulation, according to Huang et al. [60] [60] is:

$$M(f) = 1125 * \ln(1 + f/700) \quad (2.12)$$

where  $f$  is the input frequency in Hz. The scale is plotted Figure 2.11.

and also of the physics of speech, such as the fact that human like these systems often have well defined overtones that are harmonic – which is why the MFCCs use the FFT of the FFT)

### 2.3.8 Dealing with Noisy Data

One of the central problems in ASR is how to deal with noisy audio data. It is long known that the performance of speech recognition systems greatly degrade when the environmental or the recording conditions are not controlled, thus allowing unwanted residual sounds to appear in the signal. In acoustics, any type of sound that is not the one you are willing to analyze is considered noise. As a result from this, in speech recognition, the hiss of a fan, the buzz that a computer cooler makes, car horns on the street and so on are all regarded as noise. Even someone's voice can be regarded as noise. Consider, for instance, that you are trying to recognize John's speech in an application, however Mary is close to him talking on the phone, to the extent that traces of her voice are added to the signal. In this scenario, Mary's voice is actually noisy data, since it is undesirable for the given purpose.

### 2.3.9 Types of Speech Recognition Systems

Errem omnium ea per, pro UML congue populo ornatus cu, ex qui dicant nemore melius. No pri diam iriure euismod. Graecis eleifend appellantur quo id. Id corpora inimicus nam, facer nonummy ne pro, kasd repudiandae ei mei. Mea menandri mediocrem dissentiet cu, ex nominati imperdiet nec, sea odio duis vocent ei. Tempor everti appareat cu ius, ridens

---

<sup>5</sup>This is not entirely true. As shown by Umesh et al. [119] [119], in fact, there are no two distinguishable regions in terms of statistical significance. But the idea that we perceive low frequencies better than high ones still hold.

audiam an qui, aliquid admodum conceptam ne qui. Vis ea melius nostrum, mel alienum euripidis eu.

### **2.3.10 The Architecture of a Large Vocabulary Continuous Speech Recognition System**

Non vices medical da. Se qui peano distinguer demonstrate, personas internet in nos. Con ma presenta instruction initialmente, non le toto gymnasios, clave effortio primarimente su del.

### **2.3.11 Computer Assisted Pronunciation Training**

#### **1 Sistemas de Reconhecimento de Pronúncia**

Um reconhecedor de pronúncia nada mais é do que um reconhecedor de fala voltado a uma tarefa específica, qual seja: compreender e analisar a pronúncia de um aprendiz. Como já discutido, um reconhecedor de fala é um sistema computacional que recebe como entrada um sinal acústico de fala e fornece como saída a transcrição textual da informação contida na fala (Rabiner & Schafer, 2007).

Reconhecedores de pronúncia são ferramentas de interesse, especialmente, na área de Computer-Assisted Language Learning (CALL), subárea da Linguística Aplicada que se dedica ao estudo da utilização de computadores para aprendizagem de língua (Beatty, 2002). Os sistemas de CALL que auxiliam no aprendizado ou na prática da pronúncia de outras línguas são os chamados Computer-Assisted Pronunciation Training (CAPT). Tais sistemas são útes, especialmente, por quatro motivos: (i) sua difusão: basta ter acesso a um computador para utilizá-los; (ii) sua capacidade de fornecer feedback individual - nas salas de aula tradicionais, dado o tempo, nem sempre é possível ao professor corrigir verbatim a pronúncia de cada aluno; (iii) seu baixo custo - tais tecnologias possuem baixo custo, se comparadas ao gasto com um curso de pronúncia; (iv) sua possibilidade de propiciar autoestudo assíncrono - o aluno pode treinar sua pronúncia onde e quando quiser, independentemente de um lugar ou horário específico (Witt, 1999).

Sistemas de CAPT, em geral, tentam agregar os seguintes componentes:

listas de pronúncia, material expositivo com informações acústico- articulatórias de cada som, tutoriais e exercícios de transcrição, atividade para prática e avaliação de pronúncia. Dois desses sistemas, notadamente, Accent Master e Macmillan Education Sounds são analisados a seguir.

O Accent Master é um software pago, que busca ensinar a pronúncia do AmE a partir de atividades, exercícios, jogos, vídeos expositivos e animações. A parte de exposição e ensino

de pronúncia do software é bem completa e possui explicações detalhadas de como se produz cada fone do AmE, além de vídeos e animações da posição dos órgãos do aparelho fonador. O software possui versões específicas para a língua nativa do aprendiz (atualmente, há suporte para 21 línguas, incluindo o PB). As versões específicas compreendem uma descrição dos sons da língua nativa, bem como instruções sobre quais aspectos de pronúncia devem ser focados. Porém o Accent Master não possui um sistema de reconhecimento de fala built-in. O treinamento de pronúncia dá-se da seguinte forma: primeiro, o aprendiz ouve uma palavra em inglês, tenta repeti-la, então, o software plota no monitor o oscilograma da palavra ouvida e de sua tentativa, e cabe ao próprio aprendiz comparar se a realização foi similar ou não. No entanto, deixar ao aprendiz a análise do oscilograma não constitui uma boa solução, pois, mesmo considerando a enunciação de uma mesma palavra, a forma de um oscilograma é bastante variável e depende de características do aparelho fonador (Johnson & Mullenix, 1997). Além disso, trabalhos anteriores a respeito de Sistemas de CALL já questionaram o valor pedagógico deste tipo de atividade (Neri et al., 2008). A Figura 10 contém algumas telas da interface do software Accent Master.

[pic] [pic] [pic]

Figura 10: Telas da interface do Accent Master.

O Macmillan Education Sounds: The Pronunciation App é um aplicativo de ensino de pronúncia do inglês para iPhone, iPad e Android. Trata-se de um software com uso gratuito limitado e compras in-app, que foi desenvolvido para treinamento da pronúncia tanto do AmE, quanto do BrE. O app possui exercícios de reading, writing e listening para treinar a transcrição fonética das palavras do inglês. No entanto, não há nenhum tipo de introdução à fonologia ou fonética do inglês, de modo que se pressupõe que o usuário já tenha domínio do Alfabeto Fonético Internacional (AFI) e das convenções de transcrição das palavras do inglês. Sendo assim, a utilização do software acaba por restringir-se a aprendizes intermediários e avançados de inglês que saibam utilizar o Alfabeto Fonético Internacional. O exercício de pronúncia é composto por um dicionário, que contém o áudio das palavras e que possibilita ao aprendiz gravar sua própria fala e comparar com o áudio existente no dicionário. O app não possui reconhecimento de fala e, por conseguinte, não há nenhum tipo de avaliação ou feedback da pronúncia do usuário. A Figura 11 contém exemplos da interface do software gratuito.

[pic] [pic] [pic]

Figura 11: Telas da interface do software Macmillan Education - Sounds.

### 2.3.12 English

Many papers have confirmed the importance of feedback for adequate learning in second language acquisition. Negative feedback has been investigated by (XXXXX) and positive feedback by (XXXXX). For CAPT systems the feedback to the user may be provided by many media: text, voice or video.

Definitely, the more complex and informative form of feedback is through video. Badin et al. [7] [7] developed a 3D talking head, that is able to display the articulation in an augmented mode, by showing all major speech articulators, including those usually hidden such as the tongue or the velum. The talking was called OroFacial Clone, and an example of its display can be found in Figure 2.12.

The model is based on acoustic-to-articulatory inversion and was built upon estimating speech articulators' movement from Magnetic Resonance Imaging (MRI), Computer Tomography (CT) and video data. As one can imagine, building such a model is a very expensive task, which requires not only Computer Graphics (CG) and speech processing expertise, but also access to high-priced medical devices, such as MRI scanners and tomographs.

Part of this system was evaluated by Wang et al. [125] [125] in a CAPT context, by testing the performance of Chinese learners of French. The aim of the test was to analyse the performance of the learners in the production and perception of the French vowels [ø] and [œ], that do not exist in Chinese. The authors found that the students exposed to audiovisual stimuli produced vowels that were closer to the correct ones than the subjects who had access only to the auditory stimuli. After training, the difference between the F1 of [ø] and [œ] was 62 Hz for the audio group and 178 Hz for the audiovisual group. Additionally, the audiovisual group performed better in an [ø] and [œ] perception test, the number of correct responses rose from 61% to 68% for the audio only group, and from 50% to 86% for the audiovisual one.

### 2.3.13 How to Adapt a Speech Recognition to Non-Native Data

#### 2 Adaptação a Dados de Não-nativos

É inegável que as tecnologias de reconhecimento de fala, mesmo as do estado da arte, apresentam problemas. Por tal razão, muitos pesquisadores mostram-se céticos quanto à eficiência do reconhecimento de fala de não-nativos. Entretanto, as críticas que, geralmente, são imputadas a sistemas de reconhecimento de fala de não-nativos, conforme apontam Neri et al. (2003), são fruto da falta de familiaridade com o design de reconhecedores de fala. De fato, caso se tente utilizar um reconhecedor de fala, projetado para nativos, com não-nativos, o desempenho será baixo, tendo em vista que o reconhecedor não está preparado para os

padrões acústico-articulatórios que o não-nativo produzirá. Entretanto, há diversos métodos para se adaptar um sistema de RAF a dados de não-nativos. Conforme apontam Strik e Cucchiarin (1999), variações de pronúncia de falantes não-nativos podem ser adicionadas a qualquer nível do reconhecedor: no modelo acústico, no modelo de língua ou no modelo de pronúncia. Tais modelos são apresentados nas três seções seguintes.

### 3 Adaptação do Modelo Acústico (MA)

No que concerne ao modelo acústico do reconhecedor, três métodos têm sido, comumente, empregados para tratar dados de fala de não-nativos: (i) adaptação ao falante; (ii) construção de modelos bilíngues; e (iii) utilização de modelos combinados, ou de interlíngua (Wang, et al., 2003). Tais métodos distinguem-se quanto à origem dos dados acústicos utilizados para treinar o modelo. Na adaptação ao falante, são utilizados apenas dados de fala de falantes nativos da língua alvo. Na construção de modelos bilíngues, empregam-se no treinamento do modelo acústico dados de fala de falantes nativos tanto da língua alvo, quanto da língua base. Por fim, nos modelos combinados, ou de interlíngua, o modelo acústico é treinado tendo em vista dados de falantes nativos da língua alvo e, também, de aprendizes. A Figura 12 resume as três abordagens disponíveis para adaptar o modelo acústico de um sistema de RAF.

[pic]

Figura 12: Métodos para se adaptar o Modelo Acústico (MA) do reconhecedor a dados de não-nativos.

A fim de melhor esclarecer a distinção que há entre os três métodos, consideremos a criação de um sistema de reconhecimento de pronúncia que tenha por fim reconhecer o inglês americano falado por falantes nativos de PB. Na técnica de adaptação ao falante, para esse reconhecedor, apenas dados de falantes nativos de inglês seriam utilizados no treinamento do modelo. Na abordagem bilíngue, o modelo acústico seria composto de forma dual, possuindo dados de fala de ambas as línguas: tanto do inglês americano, quanto do português brasileiro. Na utilização de modelos combinados ou de interlíngua, o modelo acústico seria alimentado com dados nativos da língua alvo, no caso, americanos falando inglês, e também com dados de aprendizes, isto é, brasileiros falando inglês.

Diversos métodos e algoritmos podem ser empregados para realizar uma das três abordagens de adaptação, a exemplo de: Maximum Likelihood Linear Regression (MLLR) + Maximum A Posteriori (MAP) + Identificação de fones mais informativos (Oh et al., 2006), Phonetic Decision Tree (PDT) (Chen & Cheng, 2012), Polyphone Decision Tree Specialization (PDTS) (Wang et al., 2003), Eigenvoices + MLLR (Tan & Besacier, 2007), Phoneset comum + Modelo multilíngue (Fischer et al., 2002). Nesta dissertação, propomos a utilização de modelos combinados. De tal modo, ne

### 4 Adaptação no Modelo de Pronúncia (MP)

Em sistemas de RAF, os dicionários de pronúncia constituem o módulo que contém as palavras do léxico do reconhecedor, juntamente com sua forma fonética transcrita. Trata-se, em verdade, do componente do reconhecedor que faz a ponte entre as unidades acústicas subpalavras presentes no modelo acústico e as possíveis sequências de palavras especificadas no modelo de língua. O Quadro 5 ilustra a estrutura do CMUdict[15] (Weide, 1998), um dicionário de pronúncia de referência para o inglês americano.

Quadro 5: Exemplo de entradas no CMUdict.

[pic]

Como se observa, um dicionário constitui uma lista de palavras, que contém formas ortográficas e fonéticas pareadas, além de um identificador. A função do identificador é possibilitar a distinção de palavras homógrafas heterófonas, isto é, palavras que possuem mesma grafia mas pronúncia distinta, como gov[e]rno (nome) e gov[?]rno (verbo); bem como a distinção das pronúncias variantes de uma palavra.

No que diz respeito ao RAF de não-nativos, a adaptação que se costuma fazer ao dicionário de pronúncia é a adição das formas variantes de pronúncia do não-nativo, de modo a construir os chamados dicionários multipronúncia. A construção de tais dicionários pode se dar por meio de duas abordagens: (i) baseada em conhecimento ou (ii) baseada em dados (Strik & Cuccharini, 1999). Ambas possuem seus prós e contras.

Na abordagem baseada em conhecimento, variantes de pronúncia são inseridas no dicionário do reconhecedor por meio de regras geradas por um especialista - um linguista. A língua constitui uma heterogeneidade ordenada, isto é, a variação não é um processo que ocorre aleatoriamente, há diversos fatores, sejam eles estruturais, sejam sociais, que condicionam a variação linguística (Weinreich, et al., 2006). Portanto, cabe ao especialista explicitar a estrutura que subjaz à variação linguística, inferindo as regras fonológicas que melhor descrevem as variantes. Tais regras, em geral, assumem o formato:

$$A > B/C_D,$$

em que A representa o elemento a ser modificado, B o elemento já modificado, e C\_D o contexto de aplicação da regra, sendo C o contexto estrutural à esquerda e D o contexto à direita (Crystal, 2008). Regras fonológicas, portanto, são capazes de descrever a variação fonética em uma escala segmental (Wester, 2003). Considere-se, como exemplo, o caso da epêntese de vogais altas anteriores não-arredondadas [i] em sílabas contendo oclusivas em coda, por brasileiros aprendizes de inglês. Por transferência de padrões fonotáticos de L1 para L2, brasileiros tendem a realizar epêntese em palavras que apresentam oclusivas em posição coda, que são proibidas, originalmente, na fonologia do português (Collischonn, 2003; Zimmer & Alves, 2006). Sendo assim, acabam por pronunciar “book” como boo[ki]

em vez de boo[k] e “trip” como tri[pi] em vez de tri[p]. Tal fato pode ser capturado pela regra:

$$\text{COCLUS} > \text{COCLUS} + [i]/\$,$$

em que COCLUS representa uma consoante oclusiva e “\$” a fronteira silábica. Na construção de dicionários multipronúncia baseados em conhecimento, o especialista deve arrolar um conjunto suficiente de regras, de forma a acomodar as diversas variações de pronúncia que um falante não-nativo apresenta.

Trata-se, portanto, de uma tarefa custosa, quer financeira quer temporalmente, uma vez que pressupõe um extenso levantamento e análise de dados linguísticos ou uma consulta dicionários transcritos já compilados. Além disso, apesar de os dados obtidos com um especialista serem fiáveis, frequentemente, o levantamento feito é incompleto, já que diversos fenômenos de variação linguística ainda estão para ser estudados, não havendo literatura disponível (Wester, 2003). Em outras palavras, regras desenvolvidas por um especialista possuem alta precisão (precision), mas não necessariamente alta cobertura (recall). Ademais, a abordagem baseada em conhecimento é dependente de língua ou, em certos casos, de um dialeto; regras desenvolvidas para uma língua ou dialeto não são necessariamente aplicáveis a outras línguas ou dialetos.

A abordagem baseada em dados para a construção de dicionários multipronúncia pode ser classificada como direta ou indireta (Kim, Oh, & Kiem, 2007). Na direta, padrões de variação existentes em um conjunto de teste são analisados e utilizados, imediatamente, para gerar as palavras variantes. Já na indireta, busca-se inferir regras que possam ser aplicadas na geração de uma ou mais variantes para uma palavra. A abordagem direta, portanto, atém-se aos padrões de variação que ocorrem no conjunto de teste, enquanto a indireta é capaz de prever padrões que não vieram a ocorrer.

Um dos problemas da utilização de dicionários multipronúncia é o aumento da incerteza do reconhecedor. Com a adição de variantes ao dicionário de pronúncia, aumenta-se a cardinalidade do espaço amostral de palavras que o reconhecedor deve percorrer na busca e, por conseguinte, aumenta-se sua confusão. Em síntese, ao se adicionar muitas formas variantes ao dicionário de pronúncia, a confusão inserida no modelo pode contrabalancear os ganhos com uma forma fonética mais precisa para as palavras (Compernolle, 2001). Dicionários multipronúncia, de tal forma, devem procurar o ponto ideal entre o número de formas variantes adicionadas e as formas canônicas. Kim et al. (2007) propõem uma métrica para avaliar a confusão de dicionários multipronúncia, baseada no número de variantes de pronúncia de cada palavra, nos fones que a compõem e na distância de Levenshtein entre as palavras. Tal métrica, chamada Confusability Measure (CM), é definida formalmente como: seja  $D$  um dicionário multipronúncia, composto por  $|D|$  palavras, de maneira que cada

palavra [pic] possua [pic] variantes de pronúncia, sendo [pic] a j-ésima variante de pronúncia pertencente à i- ésima palavra, então:

[pic]

onde [pic] é a distância de Leveshtein entre [pic] e [pic], e [pic] é o número de fones da variante de pronúncia [pic], normalizado pelo número total de fones de todas as pronúncias de [pic], tal que:

[pic]

onde [pic] é definido como o número de fones da pronúncia [pic]. O objetivo da CM é sistematizar a construção de dicionários multipronúncia, estabelecendo um critério objetivo para a seleção das palavras que devem compor o modelo. Palavras distintas, mas com contexto fonético similar são favorecidas pela métrica, podendo-se selecioná-las ao estabelecer um threshold.

Em muitas vezes, os ganhos em WER obtidos com a construção de modelos de pronúncia são baixos em razão da adição “cega” de palavras ao dicionário. Jurafsky et al. (2001) demonstraram, por exemplo, que o relaxamento de vogais no inglês e a substituição de segmentos são, automaticamente, capturados por modelos acústicos baseados em trifones, não sendo necessário, portanto, adicionar variantes ao dicionário. Outros tipos de variação, como o apagamento de sílabas, não são bem modelados em trifones, de modo que é necessário tratá-los no dicionário de pronúncia (Jurafsky, et al., 2001).

Diversas técnicas podem ser utilizadas para elaborar um dicionário multipronúncia, a exemplo de: Gaussian Densities Across Phonetic Models (Saraçlar & Khudanpur, 2000), Group Delay Based Segmentation (Brunet & Murthy, 2012), Phones Adaptation and Pronunciation Generalization (Ahmed & Ping, 2011), Automatic Generation of Accented Variants + MLLR (Goronzy & Eisele, 2003), Multi-span Linguistic Parse Tables (Mertens et al., 2011), Decision Trees (Byrne et al., 1998), Pronunciation Mixture Model (McGraw et al., 2008). A seguir, discutimos, em mais detalhe, três desses trabalhos, notadamente: Saraçlar e Khudanpur (2000), Matsunaga et al. (2003) e Kim et al. (2007).

Saraçlar & Khudanpur (2000) propõem uma técnica, a que chamam state level pronunciation model (SLPM), que estima as pronúncias variantes de uma palavra sem a inserção de variantes no dicionário de pronúncia. Os autores elaboraram um método que permite que a representação de um fonema /f/, por exemplo, seja, no modelo acústico, diretamente mapeada nos estados HMM de suas variantes [f1] e [f2], criando assim um modelo misto. Tal técnica é capaz de gerar variantes de pronúncia automaticamente, mas pressupõe a utilização de dados anotados de forma manual para bootstrap do modelo. O state level pronunciation model (SLPM) foi testado com dados do inglês americano, em uma porção do corpus Switchboard (~4 horas, ~100k fones). O incremento em WER reportado foi de 1,7% (valor absoluto),

tal valor corresponde à melhoria no reconhecedor dos valores de WER sem um modelo de pronúncia (39,4%) e com um modelo de pronúncia do tipo state level pronunciation model (SLPM) (37,7%).

Matsunaga et al. (2003) utilizam dicionários multipronúncia e modelos acústicos combinados para tratar o inglês falado por japoneses. Os autores conduziram testes utilizando cinco métodos: (i) léxico do inglês + modelo acústico nativo; (ii) léxico com influência do japonês + modelo acústico japonês; (iii) comparação de ambos os métodos; (iv) léxico japonês e inglês + modelo acústico combinado; (v) léxico combinado + modelo acústico combinado + penalidade para alternância de palavras entrelínguas.

Kim et al. (2007) propuseram um método automático de criação de um dicionário multipronúncia, através de uma abordagem indireta baseada em dados. De modo a criar um sistema de RAF não-nativo para coreanos aprendizes de inglês, os autores expandiram as formas canônicas do reconhecedor e, então, utilizaram uma medida de confusão (CM), baseada na distância de Levenshtein, para atribuir um valor a cada variante de pronúncia e excluir aquelas que apresentavam baixa confusão. As variantes foram geradas através do seguinte procedimento: 1º) as sentenças não-nativas foram reconhecidas através de um reconhecedor de fones; 2º) a saída do reconhecedor foi alinhada com a forma de pronúncia canônica esperada através de um algoritmo de programação dinâmica (algoritmo de Viterbi); 3º) padrões de pronúncia variantes foram obtidos a partir dos dados alinhados; 4º) regras de variação de pronúncia são derivadas dos padrões de variação, por meio de árvores de decisão (algoritmo C4.5); 5º) as regras de variação são aplicadas ao restante do léxico, de modo a gerar possíveis candidatas a variantes de pronúncia. A medida de confusão (CM) é então utilizada, de modo a excluir do modelo de pronúncia as variantes que apresentam baixa confusão. O corpus utilizado no experimento foi uma porção do Wall Street Journal Database (WSJ0) (~7.138 sentenças). A melhoria na taxa de WER reportada foi de 1,34% (valor absoluto), que corresponde a uma variação de 19,92% a 18,58%. O baseline utilizado foi uma porção de 340 palavras do CMU Pronouncing Dictionary.

## 5 Adaptação no Modelo de Língua (ML)

[Esta seção do trabalho será realizada posteriormente, após elaborados os modelos acústico e de pronúncia.]

Fig. 2.4 Height (cm) versus vocal tract length (mm) [42].

Fig. 2.5 Averaged vocal tract morphology [42].

Fig. 2.6 F0 and pitch sigma versus age for males and females [16].

Fig. 2.7 Two complex waveforms generated by the same three pure tone 100 Hz, 200 Hz and 300 Hz sine waves, differing only with respect to their relative timing [68].

Fig. 2.8 Example of in-phase waves.

Fig. 2.9 Example of out-of-phase waves.

Fig. 2.10 Illustration of an original audio recording (the upper waveform) divided into two offset sequences of analysis windows (two lower waveforms) with 50% overlapping frames [80]

Fig. 2.11 Mel scale versus a linear frequency scale.

Fig. 2.12 Example of the OroFacial Clone display, from Badin et al. [7] [7].

# **Chapter 3**

## **Aieouadô’s dictionary and G2P converter**

### **Abstract**

<sup>1</sup>This chapter describes the method employed to build a machine-readable pronunciation dictionary for Brazilian Portuguese. The dictionary makes use of a hybrid approach for converting graphemes into phonemes, based on both manual transcription rules and machine learning algorithms. It makes use of a word list compiled from the Portuguese Wikipedia dump. Wikipedia articles were transformed into plain text, tokenized and word types were extracted. A language identification tool was developed to detect loanwords among data. Words’ syllable boundaries and stress were identified. The transcription task was carried out in a two-step process: i) words are submitted to a set of transcription rules, in which predictable graphemes (mostly consonants) are transcribed; ii) a machine learning classifier is used to predict the transcription of the remaining graphemes (mostly vowels). The method was evaluated through 5-fold cross-validation; results show a F1-score of 0.98. The dictionary and all the resources used to build it were made publicly available.

### **3.1 Introduction**

In many day-to-day situations, people can now interact with machines and computers through the most natural human way of communication: speech. Speech Technologies are present in GPS navigation devices, dictation systems in text editors, voice-guided browsers for

---

<sup>1</sup>This chapter contains an extended version of the following paper, which was originally presented on September 16th 2014, at the 15th INTERSPEECH conference in Singapore: Mendonça, G. and Aluísio, S. (2014). Using a hybrid approach to build a pronunciation dictionary for Brazilian Portuguese. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association INTERSPEECH 2014*, Singapore

the vision-impaired, mobile phones and many other applications [52]. However, for many languages, there is a dire shortage of resources for building speech technology systems. Brazilian Portuguese can be considered one of these languages. Despite being 6<sup>th</sup> most spoken language in the world [73], with about 200 million speakers, speech recognition and speech synthesis for Brazilian Portuguese are far from the current state of the art [87]. In this paper, we describe the method employed in building a publicly available pronunciation dictionary for Brazilian Portuguese which tries to diminish this scarcity.

The dictionary makes use of a hybrid approach for grapheme to phoneme conversion, based on both manual transcription rules and machine learning algorithms, and aims at promoting the development of novel speech technologies for Brazilian Portuguese. Hybrid approaches in grapheme to phoneme conversion have been applied successfully to other languages [27][96][114][124]. They have the benefit of taking advantage from both knowledge-based and data-driven methods. We propose a method in which the phonetic transcription of a given word is obtained through a two-step procedure. Its primary word list derives from the Portuguese Wikipedia dump of 23<sup>rd</sup> January 2014. We decided to use Wikipedia as the primary word list for the dictionary for many reasons: i) given its encyclopedia nature, it covers wide-ranging topics, providing words from both general knowledge and specialized jargon; ii) it contains around 168,8 million word tokens, being robust enough for the task; iii) it makes uses of crowdsourcing, lessening author's bias; iv) its articles are distributed through Creative Commons License. Wikipedia articles were transformed into plain text, tokenized and word types were extracted.

We developed a language identifier in order to detect loanwords among data. It is a known fact that when languages interact, linguistic exchanges inevitably occur. One particular type of linguistic exchange is of great concern while building a pronunciation dictionary, namely, non-assimilated loanwords [17]. Non-assimilated loanwords stand for lexical borrowings in which the borrowed word is incorporated from one language into another straightforwardly, without any translation or orthographic adaptation. These words represent a problem to grapheme-to-phoneme (G2P) conversion since they show orthographic patterns which are not predicted in advance by rules or which are too deviant to be captured by machine learning algorithms. Many algorithms have been proposed to address Language Identification (LID) from text [9][10][117][135]. Since our goal is to detect the language of single words, we employed n-gram character models in the identifier, given its previous success in dealing with short sequences of characters.

Brazilian Portuguese Phonology can be regarded as syllable and stress-driven [107]. In fact, many phonological processes in Brazilian Portuguese are related to or conditioned by syllable structure and stress position [50]. Vowel harmony occurs in pretonic context [12],

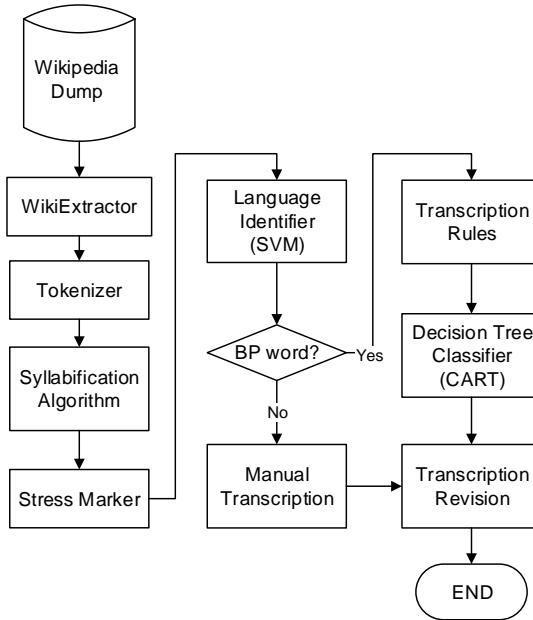


Fig. 3.1 System architecture for building the pronunciation dictionary.

posttonic syllables show a limited vowel inventory [107], nasalization occurs when stress syllables are followed by nasal consonants [97], epenthesis' processes are triggered by the occurrence of non-allowed consonants in coda position [32] and so on and so forth. Therefore, detecting syllable boundaries and stress is of crucial importance for G2P systems, in order to achieve correct transcriptions. Several algorithms have been proposed to deal with the syllabification in Brazilian Portuguese. However most of them were not extensively evaluated nor were made publicly available [91] [123] [87] [102]. For this reason, we implemented our own syllabification algorithm, based directly on the rules of the last Portuguese Language Orthographic Agreement [14].

Word types recognized as belonging to Brazilian Portuguese by the language identifier were transcribed in a two-step process: i) words are submitted to a set of transcription rules, in which predictable graphemes (mostly consonants) are transcribed; ii) a machine learning classifier is used to predict the transcription of the remaining graphemes (mostly vowels). All the data were subsequently revised. Figure 1 summarizes the method.

## 3.2 Method

### 3.2.1 Primary Word List

We used the Portuguese Wikipedia's dump of 23<sup>rd</sup> January 2014 as the primary word list for the pronunciation dictionary. In order to obtain plain text from the articles, we employed WikiExtractor [81]; it strips all the MediaWiki markups and metadata forms. Afterwards, texts were tokenized and unique words types extracted. The Portuguese Wikipedia has about 168,8 million word tokens and 9,7 million types, distributed among 820,000 articles. With the purpose of avoiding misspellings, URLs and other spurious data, only words with frequency higher than 10, which showed neither digits nor punctuation marks were selected.

### 3.2.2 Language Identifier

A Language Identifier module was developed in order to detect loanwords in the pronunciation dictionary. The Identifier consists of a Linear Support Vector Machine Classifier [110] and was implemented in Python, through Scikit-learn [92]. It was trained on a corpus made of the 200,000, containing 100,000 Brazilian Portuguese words and 20,000 words of each of the following languages: English, French, German, Italian and Spanish. All of these words were collected through web crawling News' sites and were not revised. We selected these languages because they are the major donors of loanwords to Brazilian Portuguese [3]. From these words we extracted features such as initial and final bi- and trigraphs; number of accented graphs, vowel-consonant ratio; average mono-, bi- and trigraphs probability; and used them to estimate the classifier. Further details can be found in the website of the Project<sup>2</sup>. After training, we applied the classifier to the Wikipedia word list with the purpose of identifying loanwords among data. The identified loanwords were then separated from the rest of words for later revision, i.e. they were not submitted to automatic transcription.

### 3.2.3 Syllabification algorithm and stress marker

Our syllabification algorithm follows a rule-approach and is based straightforwardly on the syllabification rules described in the Portuguese Language Orthographic Agreement [14]. Given space limitations, rules were omitted from this paper as they can be found in the website of the project, along with all the resources developed for the dictionary. As for the stress marker, once the syllable structure is known in Brazilian Portuguese, one can predict where stress falls. Stress falls:

---

<sup>2</sup><http://nilc.icmc.usp.br/listener/aeiouado>

1. on the antepenultimate syllable if it has an accented vowel <á,â,é,ê,í,ô,ú>;
2. on the ultimate syllable if it contains the accented vowels <á,é,ô> or <i,u>; or if it ends with one of the following consonants <r,x,n,l,z>;
3. on the penultimate syllable otherwise.

### 3.2.4 Transcriber

The transcriber is based on a hybrid approach, making use of manual transcription rules and an automatic classifier, which builds Decision Trees. Initially, transcription rules are applied to the words. The rules covers not all possible graphemes to phoneme relations, but only those which are predictable by context. The output of the rules is what we called the intermediary transcription form. After obtaining it, a machine learning classifier is applied in order to predict the transcription of the remaining graphemes. Figure 2 gives an example of the transcription process.

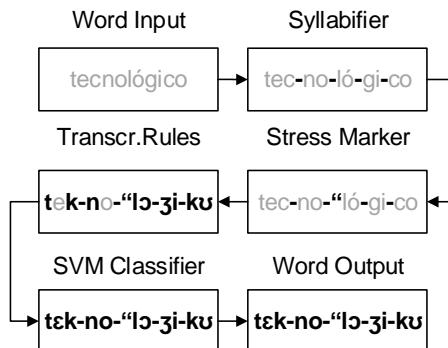


Fig. 3.2 *Example of the transcription procedure – in grey: graphemes yet to be transcribed; in black: graphemes already transcribed.*

The rules' phase has two main goals: guarantee the correct transcription of certain predictable graphemes (mostly consonants) and also ensure the alignment between graphemes and phones for the classifier. They were set in order to avoid overlapping and order conflicts. Long sequences of graphemes, such as triphthongs, contextual diphthongs and general diphthongs are transcribed first (e.g. <x-ce>→[-se]). Then graphemes involving phones that undergo phonological processes are transcribed (e.g. <ti>→[tʃi], <di>→[dʒi]). After that, several contextual and general monophones are transcribed (e.g. <#x>→[ʃ], <#e-x>→[#e-z]).

On what regards to the classifier, it was developed primarily to deal with the transcription of vowels. In Brazilian Portuguese, vowels have a very irregular behavior, specially the mid ones. Therefore the relations between the vowels' graphemes and their corresponding phonemes are hard to predict beforehand through rules. Consider, for instance, the words “teto” (*roof*) and “gueto” (*ghetto*); both are nouns and share basically the same orthographic environment. However the former is pronounced with an open “e” ['tə.tu] and the latter with a closed one ['ge.tu]. The classifier employs Decision Trees, through an optimised version of the CART (Classification and Regression Trees) algorithm and was implemented in Python, by means of the Scikit-learn library [92].

The algorithm was trained over a corpus of 3,500 words phonetically transcribed and manually revised, with a total of 39,934 instances of phones. The feature extraction happened in the following way. After reviewing the data, we obtained the intermediary transcription form for each of these words and aligned them with the manual transcription. Then, we split the intermediary transcription form into its corresponding phones and, for each phone, we extracted the following information: i) the phone itself; ii) 8 previous phones; iii) 8 following phones; iv) the distance between the phone and the tonic syllable; v) word class – parts of speech; v) the manually transcribed phone. We considered a window of 8 phones in order deal with vowel harmony phenomena. By establishing a window with such length, one can assure that pretonic phones will be able to reach the transcription of the vowels in the stressed syllable. The classifier was applied to all 108,389 words categorized as BP words by the Language Identifier module, all of them were cross-checked by two linguists with experience in Phonetics and Phonology.

### 3.3 Results

The Portuguese Wikipedia has about 168,8 million word tokens and 9,7 million types, distributed among 820k articles. After applying the filters to the data, i.e. words with frequency higher than 10, with no digits nor punctuation marks, we ended up with circa 238k word types, representing 151,9 million tokens. Table 1 describes the data.

Table 3.1 *Portuguese Wikipedia Summary – Dumped on 23<sup>rd</sup> January 2014.*

	<b>Word Tokens</b>	<b>Word Types</b>
Wikipedia	168,823,100	9,688,039
Selected	151,911,350	238,012
<b>% Used</b>	90.0	2.4

The selected words covers 90,0% of the Wikipedia content. Although the number of selected word types seems too small at first glance, one of the reasons is that 7,901,277 of the discarded words were numbers (81,5%). The remaining discarded words contained misspellings (*dirijem-se* – it should be *dirigem-se*), used a non-Roman alphabet (λόγω), were proper names (*Stolichno, Zé-pereira*), scientific names (*Aegyptophitecus*), abbreviations or acronyms (LCD, HDMI).

As for the language identifier, we trained and evaluated it with the 200,000 words multilingual corpus. The corpus consists of 100,000 Brazilian Portuguese words and 20,000 words from each of the following languages: English, French, German, Italian and Spanish. All of these words were collected through web crawling News' sites and were not revised. The results obtained for the identifier, through 5-fold cross validation are described in Table 2.

Table 3.2 *Results from the Language Identifier module – Training Phase.*

	Precision	Recall	F1-score	Support
BP words	0.85	0.89	0.87	100,000
Foreign Words	0.88	0.84	0.86	100,000
<b>Avg/Total</b>	<b>0.86</b>	<b>0.86</b>	<b>0.86</b>	<b>200,000</b>

The classifier showed an average F1-score of 0.86. Although such result is not as good as we expected – some authors reported 99% by using similar methods with trigrams probability, the relatively low F1-score can be explained given the nature of the data. In most language identifiers, the input consists of texts or several sentences, in other words, there is much more data available for the classifier. Since we are working with single words, the confusion of the model is higher and the results are, consequently, worse. Additionally, because the word list used to train the identifier was not revised, there is noise among the data. After training and evaluating the classifier, we applied it to the selected word list derived from the Wikipedia, in order to detect loanwords. Table 3 describes the results gathered.

Table 3.3 *Results from the Language Identifier module – Wikipedia word list.*

<b>Wikipedia word list</b>	
BP words	108,370 (46%)
Foreign Words	129,642 (54%)
<b>Total</b>	<b>238,012</b>

As one can observe, although we established a frequency filter to avoid spurious words, many loanwords still remain. More than half of the word list selected from Wikipedia consists

of foreign words. Notwithstanding that, the list of Brazilian Portuguese words is still of considerable size. For instance, the CMUdict [127], a reference pronunciation dictionary for the English language, has about 125,000 word types.

Concerning the syllabification algorithm and the stress marker, we did not evaluate them in isolation, but together with the transcriber since the rules for each of these modules are intertwined. That is to say the transcription rules are strictly dependent on the stress marker module and the syllable identifier. Besides, the Decision Tree Classifier is built upon the output of the transcription rules, so it is entirely dependent on it. The Decision Tree Classifier was trained over a corpus of 3,500 cross-checked transcribed words, containing 39,934 instances of phones. We analyzed its performance through 5-fold cross validation, the results for each individual phone are summarized in Table 4.

As it can be seen, the method achieved very good results, with a F1-score of 0.98. Many segments were transcribed with 100% accuracy, most of them were consonants. As it was expected, the worst results are related to mid vowels [ɛ, e, ɔ, o], specially mid-low vowels, [ɛ] showed a F1-score 0.66 and [ɔ] of 0.71. It can be the case that since the grapheme context is the same for [ɛ, e] and [ɔ, o], the Decision Tree classifier generalizes, in some cases, to the most frequent phone, that is the mid-high vowels [e,o]. The transcriber also had problems with the [k.s] (F1-score: 0.66) and [ʃ] (F1-score: 0.84). This result was also expected, both these phones are related to the grapheme <x> which, in Brazilian Portuguese, shows a very irregular behavior. In fact, <x> can be pronounced as [ʃ, s, z, k.s], depending on the word: “bruxa” (*witch*) [ʃ], “próximo” (*near*) [s]; “exame” (*test*) [z] and “axila” (*armpit*) [k.s].

### 3.4 Final Remarks

We presented the method we employed in building a pronunciation dictionary for Brazilian Portuguese. High F1-score values were achieved while transcribing most of the graphemes in Brazilian Portuguese and the dictionary can be considered robust enough for Large Vocabulary Continuous Speech Recognition (LVCSR) and Speech Synthesis. Although the rules we developed are language-specific, the architecture we used for compiling the dictionary, by using transcription rules and machine learning classifiers, can be successfully replicated in other languages. In addition, the entire dictionary, all scripts, algorithms and corpora were made publicly available.

Table 3.4 *Results from the Transcriber – Training Phase.*

	Precision	Recall	F1-score	Support
<i>syl. boundary</i>	1.00	1.00	1.00	9099
<i>stress</i>	1.00	1.00	1.00	3507
p	1.00	1.00	1.00	760
b	1.00	1.00	1.00	357
t	0.99	0.99	0.99	1135
d	0.99	0.99	0.99	1148
k	0.99	0.99	0.99	978
g	1.00	1.00	1.00	298
tʃ	0.98	0.98	0.97	450
dʒ	0.96	0.96	0.96	243
m	1.00	1.00	1.00	668
n	1.00	1.00	1.00	556
ŋ	1.00	1.00	1.00	69
f	1.00	1.00	1.00	311
v	1.00	1.00	1.00	531
s	0.98	0.98	0.98	2309
z	0.93	0.94	0.93	416
ʃ	0.84	0.84	0.84	138
k.s	0.72	0.64	0.66	41
ʒ	1.00	1.00	1.00	196
l	1.00	1.00	1.00	682
ʎ	1.00	1.00	1.00	58
r	1.00	1.00	1.00	1388
h	0.98	0.99	0.99	737
ɦ	0.97	0.92	0.94	169
w	0.97	0.98	0.97	441
˜w	0.98	0.99	0.99	309
j	0.97	0.95	0.96	223
˜j	0.95	1.00	0.98	110
a	1.00	1.00	0.99	2316
ə	0.99	0.99	0.99	1093
ɛ	0.65	0.68	0.66	275
e	0.93	0.91	0.92	1779
i	0.98	0.99	0.98	2073
ɪ	0.97	0.97	0.97	365
ɔ	0.69	0.75	0.71	220
o	0.93	0.92	0.93	1112
u	0.96	0.96	0.96	488
ʊ	1.00	1.00	1.00	1033
˜a	1.00	1.00	1.00	719
˜e	0.96	0.97	0.97	497
˜i	0.99	0.99	0.99	274
˜o	0.97	0.96	0.97	299
˜u	0.94	0.92	0.93	64
Avg/Total	0.98	0.98	0.98	39934

### 3.5 Acknowledgements

Part of the results presented in this paper were obtained through research activity in the project titled “Semantic Processing of Brazilian Portuguese Texts”, sponsored by *Samsung Eletrônica da Amazônia Ltda.* under the terms of Brazilian federal law number 8.248/91.

# Chapter 4

## Phonetic-based Speller

### Abstract

<sup>1</sup> Recently, spell checking (or spelling correction systems) has regained attention due to the need of normalizing user-generated content (UGC) on the web. UGC presents new challenges to spellers, as its register is much more informal and contains much more variability than traditional spelling correction systems can handle. This chapter proposes two new approaches to deal with spelling correction of UGC in Brazilian Portuguese (BP), both of which take into account phonetic errors. The first approach is based on three phonetic modules running in a pipeline. The second one is based on machine learning, with soft decision making, and considers context-sensitive misspellings. We compared our methods with others on a human annotated UGC corpus of reviews of products. The machine learning approach surpassed all other methods, with 78.0% correction rate, very low false positive (0.7%) and false negative rate (21.9%).

### 4.1 Introduction

Spell checking is a very well-known and studied task of natural language processing (NLP), being present in applications used by the general public, including word processors and search engines. Most of the methods of spell checking are based on large dictionaries to detect non-words, mainly related to typographic errors caused by key adjacency or fast key stroking. Currently, with the recent boom of mobile devices, with small touchscreens and

---

<sup>1</sup>This chapter contains an extended version of the recently submitted paper: Mendonça, G., Avanço, L., Duran, M., Fonseca, E., Volpe-Nunes, M., and Aluísio, S. (2015). Evaluating phonetic spellers for user-generated content in brazilian portuguese. *Proceedings of IJCAI 2015 – International Joint Conference on Artificial Intelligence*

tiny keyboards, one can miss the keystrokes, hitting adjacent keys on the keyboard, thus spell checking has regained attention [35].

Dictionary-based approaches can be ineffective when the task is to detect and correct spelling mistakes which coincidentally correspond to an existing word (real-word errors). Different from non-word errors, real-word errors are context dependent. Several approaches have been proposed to deal with these errors: mixed trigram models [44], confusion sets [Fossati and Di Eugenio], improvements on the trigram-based noisy-channel model [79] and [132], use of GoogleWeb 1T 3-gram data set and a normalized and modified version of the Longest Common Subsequence string matching algorithm [61], a graph-based method using contextual and PoS features and the double metaphone algorithm to represent phonetic similarity [108]. As an example, although MS Word (from 2007 version to on) claims to include a contextual spelling checker, an independent evaluation of it found high precision but low recall in a sample of 1400 errors [59].

Errors due to phonetic similarity also impose difficulties to spell checkers. They occur when a writer knows well the pronunciation of a word but does not know how to spell it. This kind of error requires new approaches to combine phonetic models and models for correcting typographic and/or real-word errors. In [134], for example, the authors use a linear combination of two measures – the Levenshtein distance between two strings and the Levenshtein distance between the Soundex [103] code of two strings. Phonetic errors usually occur due to inconsistent spellings rules, ambiguous word breaking and fast introduction of new words, mainly related to technology jargon, affecting both native and non-native speakers of a language [35].

[?] detail these problems in a typology of spelling errors in the scenery of written language acquisition: errors produced by lack of understanding of letter-to-sound correspondences in written language, by fails in the transcription of the oral language, by breaking rules based on phonology or morphology and by inconsistent spelling rules. Due to these several kinds of error this task is still difficult.

Some applications require interactive spelling correction (e.g. typing a text or a search query), whereas others require fully automatic correction, (e.g. corpus normalization). While in the first kind of applications the spell checker presents several suggestions to the user, the second one requires a spell checker that elects the better suggestion (first hit accuracy), as there is no user to take the decision.

In the last decade, some researchers have revisited spell checking issues motivated by web applications, such as search query engines and sentiment analysis tools based on natural language processing (NLP) of UGC, e.g. Twitter data or product reviews. The search engine Google has turned popular the facility of auto-completing, where suggestions to prefixes

of a query are offered; this is still a process of interactive spelling correction. There is also another kind of facility where a unique suggestion is given after the query is typed, with the link of the documents recovered for it [35], [25]. This facility demands a high precision automatic spelling correction.

Normalization of UGC has received great attention also because the performance of NLP tools (e.g. taggers, parsers and named entity recognizers) is greatly decreased when applied to UGC. Besides misspelled words, this kind of text presents a long list of problems, such as acronyms and proper names with inconsistent capitalization, abbreviations introduced by chat-speak style, slang terms mimicking the spoken language, loanwords from English as technical jargon, as well as problems related to ungrammatical language and lack of punctuation [36, 30, 55, 4]. UGC normalization also requires automatic spelling correction, i.e., there is a need to automatically select the word that will more likely correct the misspelled word, not relying on a list of candidates for human selection.

In [4] the authors propose a spell checker for Brazilian Portuguese (BP) to work on the top of Web text collectors. They have tested their method on news portals and on informal texts collected from Twitter in BP. However, they do not inform the error correction rate of the system. Furthermore, while their focus is on the response time of the application, they do not address real-word errors.

This chapter presents two new spell checking methods for UGC in BP. The first of them deals with phonetically motivated errors, a recurrent problem in UGC not addressed by traditional spell checkers. The second one deals additionally with real-word errors. We present a comparison of these methods with a baseline system and JaSpell over a new and large benchmark corpus for this task. The corpus contains product reviews with 38,128 tokens and 4,083 annotated errors. Such corpus is also a contribution of our study<sup>2</sup>.

This chapter is structured as follows. In Section 4.2 we describe our methods, the setup of the experiments and the corpus we compiled. In Section 4.3 we present the results. In Section 4.4 we discuss related work on spelling correction of phonetic and real-word errors. To conclude, the final remarks are outlined in Section 4.5.

## 4.2 Experimental Settings and Methods

In this Section we present the four methods compared in our evaluation. Two of them are used by existing spellers, one is taken as baseline and the other is taken as benchmark. The remaining two are novel methods developed within the project reported herein. After describing in detail the novel methods, we present the corpus specifically developed to

---

<sup>2</sup>The small benchmark of 120 tokens used in [78] and [2] is not representative of our scenario.

evaluate BP spellers, as well as the evaluation metrics. The first one is the open source spell checker JaSpell for Portuguese (Baseline method). The second is a combination of phonetic rules and Soundex applied to candidates generated by 1 and 2 edit distance (Benchmark method). The third (nome1) and the fourth (nome2) are the two novel methods presented by this paper. The (nome1) is an improvement of the Benchmark method, a pipeline of three phonetic modules applied to candidates generated by 1 and 2 edit distance and by phonetic similarity - a manually built set of phonetic-based rules, a grapheme-to-phoneme converter, and the Soundex method adapted to BP (called here as Grapheme-to-Phoneme method). The (nome2) is a context-sensitive speller, based on machine learning, applied to candidates generated by 1 and 2 edit distance, by phonetic similarity and by word combinations of diacritics (called here as Machine Learning).

If none of these phonetic modules succeed, the second layer chooses the best candidate according to its edit-distance (the lower, the better) and to its frequency in large BP corpus (the bigger, the better); In both methods (Grapheme-to-Phoneme and Rules&Soundex), if none of the phonetic modules succeed, a second layer chooses the best candidate according to its edit-distance (the lower, the better) and to its frequency in large BP corpus (the bigger, the better).

The aim of our experiments is to evaluate the effectiveness in the correction of common misspellings in UGC in BP of different phonetic-based methods applied on a corpus of product reviews. In this research we do not focus on errors related to acronyms, proper names, abbreviations, internet slang, technical jargon or loanwords in English. Instead, our goal is to assess the methods with regard to phonetic-motivated errors and a special group of real-word errors in UGC due to the absence of diacritics.

#### 4.2.1 Method I - Baseline

We use as a baseline the open source Java Spelling Checking Package, JaSpell<sup>3</sup>. JaSpell can be considered a strong baseline and it is employed at the tumba! Portuguese Web search engine to support interactive spelling checking of user queries. JaSpell classifies the candidates for a misspelled word according to the word frequency in a large corpus together with other heuristics, such as keyboard proximity or phonetic keys, provided by the Double Metaphone algorithm [94] for the English language. At the time this speller was developed there was no version of these rules for the Portuguese language<sup>4</sup>.

---

<sup>3</sup><http://jaspell.sourceforge.net/>

<sup>4</sup>Currently, a BP version of the phonetic rules can be found at <http://sourceforge.net/projects/metaphoneptbr/>

Fig. 4.1 Pseudocode: Method II - Benchmark

```

1: function CorrectWord(w)
2:   if w in Unitex then
3:     return w
4:   else
5:     w_trans <- pt_rules(w) # apply PT-Rules to word
6:     w_soundex <- soundex(w) # get word Soundex code
7:     sugs <- lev(Unitex, 2) # get words with dist. 1 or 2
8:
9:     # Look for rule transcription match
10:    for sug in sugs do
11:      sug_trans <- pt_rules(sug)
12:      if w_trans == sug_trans then
13:        return sug
14:
15:     # Look for soundex code match
16:     for sug in sugs do
17:       sug_soundex <- soundex(sug)
18:       if w_soundex == sug_soundex then
19:         return sug
20:   return most frequent suggestion

```

### 4.2.2 Method II - Benchmark

The method presented in [6] is taken as benchmark. It combines phonetic knowledge in the form of a set of rules and the algorithm Soundex. It was inspired by the analysis of errors of the same corpus of products' reviews [56] that inspired our proposals. Furthermore, as such method aims to be used for normalizing web texts, it performs automatic spelling correction.

To increase the accuracy of the first hit, this method relies in some ranking heuristics. The strategies developed by the authors consider the phonetic proximity between the input wrong word and the candidates to substitute it. If the typed word does not belong to the lexicon, a set of candidates is generated by applying one and two edit distances from the original word and the words in the lexicon. Then a set of phonetic rules for Brazilian Portuguese codifies letters and digraphs which have similar sounds in a specific code. If necessary, the next step performs the algorithm Soundex, slightly modified for BP. Finally, if none of these phonetic-based algorithms is able to suggest a correction, the candidate with the highest frequency in a reference corpus among the ones with the least edition-distance is suggested. The lexicon used is the Unitex-PB<sup>5</sup> and the frequency list was taken from Corpus Brasileiro<sup>6</sup>. The pseudocode for the algorithm can be found in Figure 4.1.

<sup>5</sup><http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/>

<sup>6</sup><http://corpusbrasileiro.pucsp.br/cb/>

### 4.2.3 Method III - Grapheme-to-Phoneme based Method (GPM)

By testing the benchmark method, we noticed that many of the wrong corrections were related to a gap between the application of phonetic rules and the Soundex module. The letter-to-sound rules were developed specially for the spelling correction, therefore, they are very accurate for the task but have a low recall, since many words do not possess the misspelling patterns which they try to model. In contrast, the transcriptions generated by the adapted Soundex algorithm are too broad and many phonetically different words are given the same code. For instance, the words "perto" (*near*) and "forte" (*strong*) are both transcribed with the Soundex code "1630", in spite of being very distinct phonetically: "perto" corresponds to ['peh.tu], and "forte" to ['fɔh.tsɪ].

To fill this gap we propose the use of a general-purpose grapheme-to-phoneme converter to be executed prior to the Soundex module. We selected Aeiouado's grapheme-to-phoneme converter [83] for this purpose, since it consists of the state of the art in grapheme-to-phoneme transcription for Brazilian Portuguese. Aeiouado employs a hybrid approach for converting graphemes into phonemes, based on both manual transcription rules and machine learning algorithms. The transcription task is carried out in two stages: i) words are submitted to a set of transcription rules, in which predictable graphemes (mostly consonants) are transcribed; ii) a decision tree classifier is used to predict the transcription of the remaining graphemes (mostly vowels). The method achieved an average F1-score of 0.98 regarding to phone transcription. Since the converter was developed primarily for text-to-speech and automatic speech recognition, its transcriptions are too much detailed for spelling correction purposes. Only few words share exactly the same phone sequence. Therefore, we had to broaden the transcription, by grouping the mid-high and mid-low vowels together, by deleting the difference between nasal and oral vowels, by considering the many rhotic sounds into a single one, etc.

The usage of the grapheme-to-phoneme converter is a bit different from a simple pipeline. According to Toutanova [116], phonetic-based errors usually need larger edit distances to be detected. For instance, the word "durex" (*sellotape*) and one of its misspelled forms "duréquis" have an edit distance of 5 units, despite having very similar or equal phonetic forms: [du'rɛks] ~ [du'rekis]. Therefore, instead of simply increasing the edit distance, which would imply in having a larger number of candidates to filter, we decided to do the reverse process. We transcribed the Unitex-PB dictionary and stored it into a database, with the transcriptions as keys. Thus, in order to obtain words which are phonetic similar words, we transcribe the input word and look it up in the database. Considering the "duréquis" example, we would first transcribe it as [du're.kɪs], and then check if there are any words in the database with this transcription. In this case, it would return "durex", the expected form.

Fig. 4.2 Pseudocode: Method III - GPM

```

1: function CorrectWord(w)
2:   if w in Unitext then
3:     return w
4:   else
5:     w_trans <- pt_rules(w) # apply PT-Rules to word
6:     w_phono <- g2p(w) # apply G2P converter to word
7:     w_soundex <- soundex(w) # get word Soundex code
8:     sugs <- lev(Unitex, 2) # get words with dist. 1 or 2
9:
10:    # Look for rule transcription match
11:    for sug in sugs do
12:      sug_trans <- pt_rules(sug)
13:      if w_trans == sug_trans then
14:        return sug
15:
16:    # Look for G2P transcription in database
17:    if transcription of word in database == w_phono then
18:      return word
19:
20:    # Look for soundex code match
21:    for sug in sugs do
22:      sug_soundex <- soundex(sug)
23:      if w_soundex == sug_soundex then
24:        return sug
25:  return most frequent suggestion

```

The only difference of GPM in comparison with Method II lies in the G2P transcription match, which takes place prior to Soundex. The pseudocode for the algorithm can be found in Figure 4.2.

In spite of being better than the baseline because they tackle phonetic-motivated errors, Method II and GPM have a limitation: they do not correct real word errors. The following method is intended to overcome this shortcoming by using context information.

#### 4.2.4 Method IV – GPM in a Machine Learning framework (GPM-ML)

Method IV has the advantage of bringing together many approaches to spelling correction into a machine learning framework. The architecture of the method is described in Figure 4.3.

The method is based on three main steps: (i) candidate word generation, (ii) feature extraction and (iii) candidate selection. The word generation phase encompasses three modules which produce a large number of suggestions, considering the following aspects: orthographic, phonetic and diacritic similarities. For producing suggestions which are typographically similar, the Levenshtein distance is used. For each input word, we select all words in a dictionary which diverge from the input by at most 2 units. For instance, suppose the user intended to write "mesa" (*table*), but missed a keystroke and typed "meda" instead.

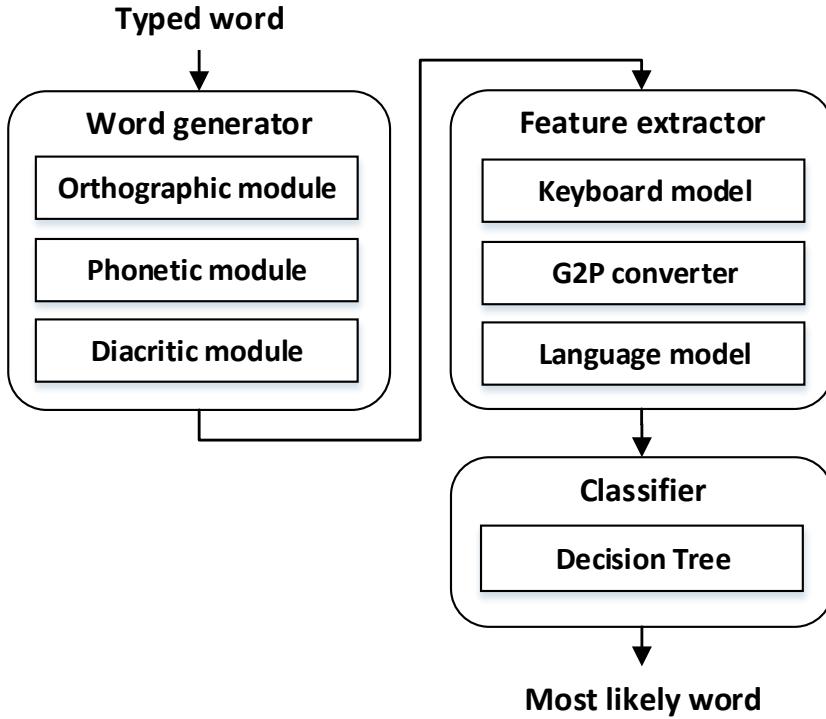


Fig. 4.3 Architecture of the GPM-ML

The Levenshtein module would generate a number of suggestions including an edit distance of 1 or 2, such as "medo" (*fear*), "meta" (*goal*), "moda" (*fashion*), "nada" (*nothing*), "mexe" (*he/she moves*) etc. For computational efficiency, we stored the dictionary in a trie structure, in order to make it quickly searchable. A revised version of the Unitex-PB was employed as our reference dictionary (*circa* 550,000 words)<sup>7</sup>.

As for phonetic similarity, the Aeiouado's grapheme-to-phoneme converter [83] was used to group phonetically related words. We transcribed the Unitex-PB word list phonetically and stored all word transcriptions along with their orthographic form into a database, exactly as we did for GPM. Thus for generating suggestions which are phonetically similar to the word typed by the user, we obtain its phonetic transcription and look it up in the database.

The diacritic module is responsible for generating words which are similar to the word typed by the user with respect to diacritic symbols. This module was proposed since we observed that most of the misspellings in the corpus were caused by a lack or misuse of diacritics. BP has five types of diacritics: accute (́), cedilla (ç), circumflex (^), grave (̀) and

<sup>7</sup>The dictionary is available upon request.

tilde (~). The diacritics often indicate different vowel quality, timbre or stress. However, these symbols are rarely used in UGC, and the reader uses the context to disambiguate the intended word. In order to allow the speller to deal with this problem, the diacritic model generates, given a word input, all possible word combinations of diacritics. Once more, the Unitex-PB is used as reference.

After word generation, the feature extraction phase takes place. This phase is responsible for extracting relevant information from the list of words generated in the previous step. The aim is to allow the classifier to compare these words with the one typed by the user, in such a way that the classifier is able to choose to keep the typed word or to replace it with one of the generated suggestions.

As misspelling errors may be of different nature (such as typographical, phonological or related to diacritics), we try to select features that encompass all these phenomena. For each word suggestion produced in the word generation phase, we extract 14 features:

1. TYPEDORGEN: whether the word was typed by the user or was produced in the word generation phase;
2. ISTYPO: 1 if the word was generated by the typographical module; 0 otherwise;
3. ISPHONE: 1 if the word was generated by the phonetic module; 0 otherwise;
4. ISDIAC: 1 if the word was generated by the diacritic module; 0 otherwise;
5. TYPEDPROB: the unigram probability of the word typed;
6. GENUNIPROB: the unigram probability of the word suggestion;
7. TYPEDTRIPROB: the trigram probability of the word typed;
8. GENTRIPROB: the trigram probability of the word suggestion;
9. TYPOLEVDIST: the levenshtein distance between the typed word and the suggestion;
10. INSKEYDIST: the sum of the key insertion distances;
11. DELKEYDIST: the sum of the key deletion distances;
12. REPLKEYDIST: the sum of the key replacement distances;
13. KEYDISTS: the sum of all previous three types of key distances;
14. PHONELEV DIST: the levenshtein distance between of the phonetic transcription of the typed word and of the suggestion.

The probabilities come from a language model trained over a subset of the Corpus Brasileiro (*circa* 10 million tokens). Good-Turing smoothing is used to estimate the probability of unseen trigrams. After feature extraction, the word selection phase comes into

play. It consists of a Decision Tree Classifier which was trained over the dataset presented in Section 4.2.5, with the features we discussed. The classifier was implemented through scikit-learn [93] and comprises an optimized version of the CART algorithm. Several other classification algorithms were tested, but since our features contain both nominal and numerical data, and since some of them are dependent, the Decision Tree Classifier achieved the best performance.

#### 4.2.5 Dataset

The evaluation corpus was compiled specially for this research and is composed of a set of annotated product reviews, written by users on Buscapé<sup>8</sup>, a Brazilian price comparison search engine. All misspelled words were marked, the correct expected form was suggested and the misspelling category was indicated. Corpus - Erros por classe: 1. Typo (1,027 tokens); 2. Phono (732 tokens: 683 non-contextual and 49 contextual); 3. Diac (2,037 tokens: 1,625 non-contextual and 412 contextual); 4. Other (86 tokens). Number of words in the corpus: 38,128 Number of misspellings: 4,083 (10,7We used snowball sampling to obtain a reasonable amount of data with incorrect orthography. A list of ortographical errors with frequency greater than 3 in the the corpus of product reviews compiled by [56] gathered by Avanço [6] was used to pre-select, from the same corpus, sentences with at least one incorrect word. Among those, 1,699 sentences were randomly selected to compose the corpus (38,128 tokens). All these sentences were annotated by two linguists with prior experience in corpus annotation. The inter-rater agreement for the error detection task is described in Table 4.1.

Table 4.1 *Inter-rater agreement for the error detection task*

		Annot. B		Total
		Correct	Wrong	
Annot. A	Correct	33,988	512	34,500
	Wrong	76	3,559	3,635
	Total	34,064	4,071	38,135

The agreement was evaluated by means of the kappa test [20]. The  $\kappa$  value for the error detection task was 0.915 which stands for good reliability or almost perfect agreement [70]. The final version of the corpus used to evaluate all methods was achieved by submitting both annotations to an adjudication phase, in which all discrepancies were resolved. We noticed that most annotation problems consisted of whether or not to correct abbreviations, loanwords, proper nouns, internet slang, and technical jargon. In order to enrich the annotation and the evaluation procedure, we classified the misspellings into five categories:

<sup>8</sup><http://www.buscape.com.br/>

1. TYPO: misspellings which encompass a typographical problem (character insertion, deletion, replacement or transposition), usually related to key adjacency or fast typing; e.g. "obrsevei" instead of "observei" (*I noticed*) and "memso" instead of "mesmo" (*same*).
2. PHONO: cognitive misspellings produced by lack of understanding of letter-to-sound correspondences, e.g. "esselente" for "excelente" (*excellent*), since both "ss" and "xc", in this context, sound like [s].
3. DIAC: this class identifies misspellings which are related to the inserting, deleting or replacing diacritics in a given word, e.g. "organizacao" instead of "organizaçao" (*organization*).
4. INT\_SLANG: use of internet slang or emoticons, such as "vc" instead of "você" (*you*), "kkkkkk" (to indicate laughter) or ":-)".
5. OTHER: other types of errors that do not belong to any of the above classes, such as abbreviations, loanwords, proper nouns, technical jargon; e.g. "aprox" for "aproximadamente" (*approximately*).

The distribution of each of these categories of errors can be found in Table 4.2. The difference between the total number of counts in Table 4.1 and 4.2 is caused by spurious orthographies which were reconsidered or removed in the adjudication phase. In addition to the five categories previously listed, we also classified the misspellings into either contextual or non-contextual; i.e. if the misspelled word corresponds to another existing word in the dictionary, it is considered a contextual error (or real-word error). For instance, if the intended word was “está” (*he/she/it is*), but the user typed “esta”, without the acute accent, it is classified as a contextual error, since “esta” is also a word in Brazilian Portuguese which means *this FEM*.

The corpus has been made publicly available<sup>9</sup> and intends to be a benchmark for future research in spelling correction for user generated content in BP.

#### 4.2.6 Evaluation Metrics

Four performance measures are used to evaluate the spellers. The *Detection rate* is the ratio between the number of errors detected and the total number of errors. The *Correction rate* stands for the ratio between the number of corrected errors and the total number of errors. *False positive rate* is the ratio between the number of false positives (correct words that are

---

<sup>9</sup>Link omitted for blind review.

Table 4.2 *Error distribution in corpus by category*

Misspelling type		Counts	% Total
TYPO	-	1,027	25.2
PHONO	Contextual	49	1.2
	Non-contextual	683	16.7
DIAC	Contextual	411	10.1
	Non-contextual	1,626	39.8
INT_SLANG	-	201	4.9
OTHER	-	86	2.1
<b>Total/Avg</b>		4,083	100.0

wrongly detected as errors) and the total number of correct words. The *False negative rate* consists of the ratio between the number of false negatives (wrong words that are detected as correct) and the total number of errors. In addition, the correction hit rates are evaluated by misspelling categories. In the analysis, we do not take into account the "int\_slang" and "other" categories, since both show a very irregular behavior and constitute specific types of spelling correction.

### 4.3 Discussion

In Table 4.3, we summarize all methods' results. As one can observe, the GPM-ML achieved the best overall performance, with the best results in at least three rates: detection, correction and false positive. Both methods we proposed in this paper, GPM and GPM-ML, performed better than the baseline in all metrics. However, GPM did not show any improvement in comparison to the benchmark. In fact, the addition of the grapheme-to-phoneme converter decreased the performance in what concerns to the correction rate. By analyzing the output of GPM, we noticed that there seems to be some overlapping information between the phonetic rules and the grapheme-to-phoneme module. Apparently, the phonetic rules were able to cover all cases which could be solved by adding the grapheme-to-phoneme converter. Therefore our hypothesis was not supported.

Table 4.3 *Comparison of the Methods*

Method	Rate			
	Detection	Correction	FP	FN
Baseline JaSpell	74.0%	44.7%	5.9%	26.0%
Benchmark Rules&Soundex	83.4%	68.6%	1.7%	<b>16.6%</b>
GPM	83.4%	68.2%	1.7%	<b>16.6%</b>
GPM-ML	<b>84.9%</b>	<b>78.1%</b>	<b>0.7%</b>	21.9%

All methods showed a low rate of false positives, the best value was found in GPM-ML (0.7%). The false positive rate is very important for spelling correction purposes and is related to the reliability of the speller. If a user knows that a certain word is correct, but the

speller says the opposite, the user is very likely to gradually lose confidence on the speller and, as a consequence, he/she stops using it. In the following we discuss the correction hit rates by misspelling categories. Table 4.4 presents a comparison among all methods.

Table 4.4 *Comparison of Correction Rates*

Misspelling type	Errors	Correction rate by method			
		I	II	III	IV
Typo	1,027	28.3%	<b>56.3%</b>	53.0%	55.4%
Phono Contextual	49	0.0%	0.0%	0.0%	<b>8.1%</b>
Phono Non-contextual	683	48.2%	85.1%	<b>87.1%</b>	81.1%
Diac Contextual	411	9.2%	26.5%	26.5%	<b>64.5%</b>
Diac Non-contextual	1626	64.0%	82.2%	82.4%	<b>96.6%</b>
<b>Total/Weighted Avg</b>	<b>3,796</b>	<b>44.7%</b>	<b>68.6%</b>	<b>68.1%</b>	<b>78.0%</b>

The baseline JaSpell (Method I) presented an average correction rate of 44.7%. Its best results comprise non-contextual diacritic misspellings with a rate of 64.0%. Its worst result is found in contextual phonological errors, not a single case of this type of error was corrected by the speller. The typographical misspellings were also very troublesome for the baseline method, with a correction hit rate of 28.3%. These results indicate that the method is not suitable for real world applications which deal with user generated content. It is important to notice that the JaSpell was not developed specifically for this text domain, so its performance is much influenced by this fact.

The benchmark Rules&Soundex (Method II) achieved a correction rate of 68.6%, a relative gain of 53.4% in comparison to the baseline. The best results are, once more, related to the non-contextual diacritic misspellings (82.2%), which stand for the major class. The best improvements compared to the baseline appear in phonological errors that are influenced by context (85.1%), with a relative increase of 76.6%. These results are coherent with the results reported by [6], since they claim that the method focuses on phonetically motivated misspellings. As already mentioned, GPM (Method III) did not show any gain in comparison with the benchmark. As can be noticed, the grapheme-to-phoneme converter had a small positive impact in what regards to the phonological errors, raising the correction rate of non-contextual phonological misspellings from 85.1% to 87.1% (2.3% gain).

GPM-ML (Method IV) achieved the best performance among all methods in what regards to correction hit rate (78.0%). Some misspelling categories showed a very high correction rate, such as non-contextual diacritic errors (96.6%) and non-contextual phonological errors (81.1%). The trigram Language Model proved to be effective for capturing some contextual misspellings, as can be seen by the contextual diacritic correction rate (64.5%). However, the method was not able to properly infer contextual phonological misspellings (8.1%). We hypothesize that this result might be caused by the few number of contextual phonological

instances in the corpus used for training (there were only 49 cases of contextual phonological misspellings). Such a small number of cases is not adequate for ensuring good performance by machine learning techniques. No significant improvement was found with respect to typographical errors (55.4%) in comparison to the other previous methods.

## 4.4 Related Work

The first approaches to spelling correction date back to Damerau [26] and address the problem by analyzing the edit distance of the words. He proposes a speller based on a reference dictionary and on an algorithm to check for out-of-vocabulary (OOV) words. The method assumes that words which are not found in the dictionary have at most one error, which was caused by a letter insertion, deletion, substitution or transposition. OOV words are then compared to the words from the dictionary. The one error threshold was established to avoid high computational cost. An improved error model for spelling correction, which works for letter sequences of lengths up to 5 and is also able to deal with phonetic errors was proposed by [15]. It embeds a noisy channel model for spell checking based on string to string edits. This model depends on the probabilistic modeling of sub-string transformations. As texts present several kinds of misspellings, no single method will cover all of them, therefore it is very natural to combine methods which supplement each other. This approach was pursued by [116] who included information on pronunciation to the model of typographical errors correction. [116] and also [122] took the pronunciation of the misspelled words into account by using the technology of grapheme-to-phoneme converters. The later proposed the use of triphone analysis as a new correction strategy to combine phonemic transcription with trigram analysis, since they performed better than either grapheme-to-phoneme conversion or trigram analysis alone, in their evaluation. Our GPM method also combines models to correct typographical errors by using information on edition distance, information on pronunciation provided by a set of phonetic rules, on a grapheme-to-phoneme converter and finally on the output of the Soundex method. In this two-layer method these modules are put in sequence, as we take advantage of the high precision of the phonetic rules before trying the converter; typographical errors are corrected in the last pass of the process. We understand that the probabilistic classification framework used by [116] is very interesting and would provide better results to our two-layer method. Therefore, we decided to take advantage of a machine learning approach to decide how to correct a word, by using candidates generated by one and two edit distance, phonetic similarity and word combinations of diacritics. In our GPM-ML proposal, we adapted the output of a grapheme-to-phoneme converter which was developed for automatic speech recognition, and used it together with a keyboard model and a language

model to provide features for a decision tree classifier. We had to broaden the transcriptions in order to deal with real-word errors related to diacritics, since the transcriptions are too much detailed for spelling correction purposes. With this new proposal one can deal with a special group of real-word errors caused by the presence or absence of diacritics, besides phonetic and typographic errors.

## 4.5 Final Remarks

We compared four spelling correction methods for UGC in BP, two of which consist of novel approaches and were proposed in this paper. The Method III (GPM) consisted of an upscale version of the benchmark method. In comparison to benchmark, it contained an additional module with a grapheme-to-phoneme converter. The grapheme-to-phoneme converter was intended to provide the speller with transcriptions that were not so fine-grained or specific as those generated by the phonetic rules and also not so coarse-grained as those created by Soundex. However, it didn't work as well as expected. The Machine Learning version of GPM, the GPM-ML, however, presented a good overall performance, as it is the unique that addresses the problem of real word errors, and surpass all other methods in most situations. It reached 78.0% in correction rate, with very low false positive (0.7%) and false negative (21.9%), thus establishing the new state of the art in spelling correction for UGC in BP. As for future work, we intend to improve GPM-ML by expanding the training database, by testing other language models as well as new phone conventions. In addition, we plan to more fully evaluate it into different testing corpora. We also envisage, in due course, the development of an internet slang module.



# **Chapter 5**

## **A greedy algorithm for the extraction of phonetically rich sentences**

### **Abstract**

<sup>1</sup>A method is proposed for compiling a corpus of phonetically-rich triphone sentences; i.e., sentences with a high variety of triphones, distributed in a uniform fashion. Such a corpus is of interest for a wide range of contexts, from automatic speech recognition to speech therapy. We evaluated this method by building phonetically-rich corpora for Brazilian Portuguese. The data employed comes from Wikipedia’s dumps, which were converted into plain text, segmentized and phonetically transcribed. The method consists of comparing the distance between the triphone distribution of the available sentences to an ideal uniform distribution, with equiprobable triphones. A greedy algorithm was implemented to recognize and evaluate the distance among sentences. A heuristic metric is proposed for pre-selecting sentences for the algorithm, in order to quicken its execution. The results show that, by applying the proposed metric, one can build corpora with more uniform triphone distributions.

### **5.1 Introduction**

In what regards to speech technology, although there are some studies which employ words [115], syllables [47] and monophones [66] to develop Automatic Speech Recog-

---

<sup>1</sup>This chapter contains an extended version of the following paper, which was originally presented on August 20th 2014, at the 14th International Telecommunications Symposium in São Paulo: Mendonça, G., Candeias, S., Perdigão, F., Shulby, C., Tonazzo, R., Klautau, A., and Aluísio, S. (2014). A method for the extraction of phonetically-rich triphone sentences. *Proceedings of ITS 2014 – International Telecommunications Symposium*

nition (ASR) and Text to Speech (TTS) systems, most of the current research widely makes use of contextual phone units, such as triphones and diphones.

The issue of developing a phonetically-rich triphone sentences corpus is of great significance for many areas of knowledge. In many applications of ASR and Speech Synthesis, for instance, rich speech databases are important for properly estimating the acoustic models [98]. In Speech Therapy, phonetically-rich sentences are often employed in reading aloud tasks so as to assess the speech production of patients in various phonetic/phonological contexts [82]. Laboratory Phonologists are also interested in such corpora in order to develop prompts for analyzing speech production and variability [95].

Formally, the task discussed in this work can be described as follows: given a corpus  $K$  with  $s$  sentences, find a subset  $P$  containing  $s_p$  sentences, such that the triphones that compose  $s_p$  holds a uniform distribution as much as possible. Despite its apparent simplicity, in what concerns computational complexity, the task cannot be considered a simple one. Since it has a combinatorial nature, it lacks a polynomial-time solution and should be regarded as an intractable problem [104].

We evaluate the proposed method in building a phonetically-rich triphone sentences corpus for Brazilian Portuguese. The sentences come from the Portuguese Wikipedia dump [131], which was converted into plain text, segmentized and phonetically transcribed. The algorithm employs a greedy approach to select sentences, in a way such that the triphone distribution in the selected sentences is as uniform as possible. In order to expedite its execution, a heuristic metric is proposed to pre-select sentences for the algorithm, favoring the least frequent triphones over the most frequent ones.

The remainder of this paper is organized as follows. In Section 2, we briefly describe the related work available in the literature. In Section 3, we describe the method proposed. In Section 4, we evaluate it by building phonetically-rich corpora for Brazilian Portuguese. The final remarks are outlined in Section 5.

## 5.2 Related Work

Speech can be analyzed in a myriad of forms. The Phonetic or Phonological structure of a language can be described through phones, phonemes, syllables, diphones, triphones, feet, etc. For languages such as Mandarin, in which tones have a phonological value, one must even posit units such as tonemes in order to properly describe speech phenomena [71].

Many methods have been proposed for extracting phonetically-balanced corpora, that is to say corpora made of sentences which reproduce the triphone distribution of a given language [1][105][49][121].

It is known that many linguistic phenomena, including triphone sets, show a Zipfian distribution [77]. A phonetically-balanced corpus, for this reason, is a corpus which follows Zipf's law in representing each triphone inversely proportional to its rank in the frequency table.

These kinds of corpora are important specially for Large Vocabulary Continuous Speech Recognition (LVCSR), where unbalanced triphone representations can achieve better Word Error Rates (WER). However, phonetically-balanced corpora are not adequate for many other tasks, even regarding Speech Recognition. When building a system to assess one's pronunciation quality or to synthesize speech, for instance, more accurate results can be attained by using uniform triphone representations, i.e. phonetically-rich corpora.

Phonetically-rich corpora in our work are those which show sentences with a high variety of triphones, distributed in a uniform fashion regardless its representation in the language. In other words, in order to build such corpora, Zipf's law must be nullified, by favoring less frequent triphones and disfavoring more frequent ones. However, there are studies that consider other definitions and even other basic units to build phonetically-rich corpora.

In Abushariah et al. [1], the concept of "rich" is used in the sense that the set must contain all the phonemes of Arabic language (the chosen language for their study) but without a need for a uniform distribution. The set of sentences was handmade developed by linguists/experts. They used a set of 663 words, also defined by hand, and then Arabic independent sentences have been written using the 663 phonetically rich words. The final database consists of 367 sentences with 2 to 9 words per sentence.

Arora et al.[5] considered syllables as the basic unit to extract automatically phonetically-rich sentences from a large text corpus of Indian language, justifying their choice because a syllable is the smallest segment of utterance. In their process to extract the sentences for a given corpus the chosen set should have the same distribution of syllabic words and also the same distribution of consonant, vowel and other symbols.

Nicodem et al.[89] deals specifically with Brazilian Portuguese (BP) and proposed a method based on genetic algorithms to select a set of sentences for a speech synthesis system. Their goal was to select a recording corpus that would improve the phonetic and prosodic variability of the system. They tried to fulfill the gap of phonetically-balanced corpora available for BP does not consider prosody, that is, the available corpora only deals with phonetic representativeness without considering prosody representativeness.

They have worked with CETENFolha corpus <sup>2</sup> which has circa of 1,5 million sentences in order to gather 4,000 sentences phonetically and prosodically rich, being 1,000 declaratives, 1,000 partial interrogatives, 1,000 total interrogatives, 500 alternatives and 500 exclamatives.

---

<sup>2</sup>[www.linguateca.pt/cetenfolha/](http://www.linguateca.pt/cetenfolha/)

Their approach is composed of 4 stages, including grapheme-to-phoneme conversion, prosodic annotation, feature vector representation, and selection. The authors obtain prosodic features based on the pitch (they use a TTS to obtain the pitch contour), identifying tone events (H+, H-, H, L, and L-, where H and L stands for high and low, respectively) and using N for neutral, for each syllabe. Using these features to represent each sentence, they execute a genetic algorithm to select a subset. Their paper, however, is not clear about how the fitness function meets both constraints (phonetic and prosodic) since their method only includes prosodic features.

### 5.2.1 Heuristic Metric

For the expedition of the sentence extraction through the greedy algorithm, due to its high time complexity order, we set a heuristic metric to pre-select sentences and rank them according to the triphones they contained. The metric uses the probability of the triphones in the Corpus in order to favor the least frequent triphones over the most frequent ones. It consists of a summation of the reciprocal probability for each triphone in the sentence.

Formally, this can be defined in the following way. Consider a corpus  $K$  consisting of a set of sentences  $S = \{s_1, s_2, s_3, \dots, s_n\}$ . Each sentence  $s$  is formed by  $m$  triphones, represented as  $T = \{t_1, t_2, t_3, \dots, t_m\}$ . Given a corpus, the *a priori* probability of the triphones can be calculated straightforwardly: let  $P_K(t_i)$  be the probability of the triphone  $t_i$  in the corpus  $K$ , then  $P_K(t_i)$  is the number of times  $t_i$  occur divided by the total number of triphones in  $K$ . For that matter, a sentence  $s$  can be considered phonetically-rich if it possess many triphones with low probability of occurrence. Therefore, we define the phonetic richness of a sentence  $s$  as the summation of its triphones' reciprocal probabilities:

$$\rho(s) = \sum_{i=1}^m \frac{1}{P_K(t_i)} \quad (5.1)$$

### 5.2.2 Algorithm

Our algorithm for extracting rich sentences was implemented in Python and follows a greedy heuristic. The distance metric is calculated through the SciPy library [63].

Greedy algorithms have been widely used in Computer Science, when the optimum solution of the problem can not be guaranteed [22]. Greedy strategies make locally optimal choices hoping to find the global optimum. Notwithstanding, in many cases, greedy algorithms have been notorious for jams at local maxima, since the best solution for a given problem may not concur with the sum of each partial best choice. However, for the

Fig. 5.1 Pseudocode: Method for extracting phonetically-rich sentences.

```

1: Corpus <- List of available sentences
2: Selected <- [] # List of selected sentences
3: Metrics <- [] # List of tuples with sentences + dist values
4: Ideal <- [] # Ideal corpus, with all equiprobable triphones
5: N <- Number of desired sentences
6:
7: while length(Selected) < N do:
8:   for Sentence in Corpus:
9:     calculate distance between Selected+Sentence and Ideal
10:    append Sentence and its metric to the Metrics list
11:   BestSentence <- select the sentence with the min distance
12:   append BestSentence to Selected
13:   clear the Metrics list

```

extraction of phonetically rich sentences, this approach is suitable, owing to the fact that it is computationally intractable to analyze all possible sets of sentences.

We initialize the algorithm by applying the heuristic metric described in Section 3.2 to all sentences in the corpus. After this, all sentences are ranked in descending order and the first 50,000 sentences with the best values are selected. This metric was proposed because the algorithm has an order of  $O(mn^2)$  time complexity, where  $n$  is the number of sentences and  $m$  the number of selected triphones, and its execution was slow considering all the sentences available in the corpus. Afterwards, the algorithm loops through 50,000 sentences and calculates the euclidean distance between the triphone distribution of the set made up with the selected sentences and the current sentence to an ideal corpus, containing equiprobable triphones. The sentence with the minimum value is appended to a list of selected sentences and removed from the corpus. after, the loop starts over, considering for the calculation of the distance not just each sentence in isolation, but a set comprising each remaining sentence in the corpus together with the sentences already selected in the last step. When the list reaches  $n$  selected sentences, the execution is suspended. The pseudocode for the algorithm is Figure 5.1.

## 5.3 Example Evaluation

### 5.3.1 Corpus

As a proof-of-concept we evaluated our method by building phonetically-rich corpora for Brazilian Portuguese. The original database of sentences consisted of the Wikipedia dump produced on 23<sup>rd</sup> January 2014. Table 1 summarizes the data.

In order to obtain only plain text from Wikipedia articles, we used the software WikiExtractor [81], to strip all of the MediaWiki markups and other metadata.

Then, we segmentized the output into sentences, by applying the Punkt sentence segmentizer [64]. Punkt is a language-independent tool, which can be trained to tokenize sentences. It is distributed together with NLTK [11], where it already comes with a model for Portuguese, trained on the Floresta Sinta(c)tica Treebank [45].

Following, each sentence was transcribed phonetically by using a pronunciation dictionary for each language variety. For Brazilian Portuguese, we employed the UFPAdic 3.0 [87], which contains 38 triphones and 64,847 entries. Given its encyclopedic nature, many sentences in Wikipedia present dates, periods, percentages and other numerical information. For this reason, we decided to supplement the dictionaries, by introducing the pronunciation of numbers from 0 to 2014. The pronunciations were defined manually and embedded into the dictionary.

The transcription task was carried out in the following way: a Python script was developed to loop over each sentence and check if all its belonging words were listed on the dictionary. If all the words were listed, the sentence was accepted, otherwise rejected. Due to the fact that many words which occur in Wikipedia were not registered in the pronunciation dictionary, a large number of sentences had to be discarded. Details are described in Table 2.

Some pilot experiments showed that the metric benefited sentences which were too long, as they had more triphones; or too short, as some of them had very rare triphones. The problem with long sentences is that they can be too complex for a recording prompt, inducing speech disfluencies such as pauses, false starts, lenghtenings, repetitions and self-correction [126]. In addition, the short sentences selected by the algorithm were usually only nominal, containing titles, topics or proper names; therefore, they would not be adequate for sentence prompts. For this reason, we filtered the sentences, selecting only those which had an average size (i.e. between 20 and 60 triphones, and more than four words). Further information is given in Table 3.

After that, we applied the heuristic metric described in Section 3.2, and the top 50,000 sentences were selected.

<b>Articles</b>	<b>Word Tokens</b>	<b>Word Types</b>	<b>Triphone Tokens</b>	<b>Triphone Types</b>
820,000	168,823,100	9,688,039	0	0

Table 5.1 Portuguese Wikipedia Summary – Dumped on 23<sup>rd</sup> January 2014.

<b>Total Sentences</b>	<b>Used</b>	<b>Used/total</b>
7,809,647	1,229,422	15.7%

Table 5.2 Sentences' summary after WikiExtractor and Punkt.

<b>Total Sentences</b>	<b>Short</b>	<b>Average</b>	<b>Long</b>
1,229,422	15,581	873,546	340,295

Table 5.3 Sentences' summary after the length filter.

### 5.3.2 Discussion

For this example evaluation, we discuss the extraction of 250 phonetically-rich sentences. Table 4 describes some triphone statistics for different sets of sentences extracted with the method proposed. The first column presents the number of extracted sentences; the second number of different triphones or triphone types; the third the number of triphone tokens; and the last the triphone type/token ratio which can be used to measure the method’s performance. Owing to the fact that no other methods for the extraction of phonetically-rich triphone sentences were found in the literature, we established a list of random sentences as the baseline for comparison. Table 5 contains the data regarding sentences selected randomly. The list of random sentences derives from the pool of 50,000 sentences described in Section 4.1. Ten different seed states were used in order to ensure randomness, the average of these results are presented in the Table.

As it can be seen through the type/token triphone ratio, the method is capable of extracting sentences in a much more uniform way. For 250 sentences, our method was capable of extracting 4189 distinct triphones, as opposed to 3335 in the random set; a difference of 854 novel distinct triphones. Furthermore, this higher number of distinct triphones was achieved with less triphone tokens (6908 vs. 10482), in a way that the type/token ratio for the method we propose was almost double the baseline: 0.61 in contrast to 0.32. Considering sets with different numbers of sentences, the method outperformed the random selection in all experiments. A Kolmogorov-Smirnov Test (K–S Test) confirms that the sentences selected through our method are closer to a uniform distribution than the ones extracted randomly.

One can observe that, as the number of selected sentences increases, the type/token ratio decreases. It may be the case that, after a huge number of sentences, the method’s output converges to a limit such that no statistical significance can be noticed while comparing to a random selection. However, given time limitations, it was not feasible to analyze such a situation. As the number of selected sentences increases so does the number of triphones for comparison. After a while, the number of triphones for comparison becomes so large that the algorithm’s execution time might not be proper for practical applications.

Additionally, the algorithm’s output needs to be revised. Despite all our caution in the data preparation process, we noticed that some of the sentences selected by the algorithm were, in fact, caused by mistakes from the pronunciation dictionary. Foreign and loan words are known to be a problem for grapheme to phoneme conversion because they do not follow the orthographic patterns of the target language [109]. Several sentences selected by our algorithm contained foreign words which were registered in the dictionary with an abnormal pronunciations, such as *Springsteen* [spr̄igsteē̄], *hill* [iww], *world* [wohwdʒ]. Since no other words are registered with the triphones [e-e+ē̄] or [e-ē̄] except for *Springsteen*, the algorithm

ends up by selecting the sentence in which it occurs. Seeing that our method of comparing triphone distributions is greedy, our algorithm is fooled into believing that these are rare jewels. While this may be the case either way, the algorithm cannot function properly with incorrect transcriptions.

A corpus with 100 revised sentences extracted by this method can be found in the Appendix A.

Sentences	Triphone Types	Triphone Tokens	Type/Token
25	923	928	0.99
50	1485	1541	0.96
75	1965	2151	0.91
100	2389	2774	0.86
125	2736	3384	0.81
150	3091	4075	0.76
175	3390	4736	0.72
200	3715	5477	0.68
225	3991	6200	0.64
250	4189	6908	0.61

Table 5.4 Triphone results from the extraction of sentences through our method.

Sentences	Triphone Types	Triphone Tokens	Type/Token Ratio
25	774	1121	0.69
50	1318	2037	0.65
75	1713	3093	0.55
100	1917	3968	0.48
125	2352	5166	0.46
150	2564	6110	0.42
175	2820	7375	0.38
200	2961	8000	0.37
225	3211	9578	0.34
250	3335	10482	0.32

Table 5.5 Triphone results from the sentences taken randomly.

## 5.4 Final Remarks

We proposed a method for compiling a corpus of phonetically-rich triphone sentences for Brazilian Portuguese. All sentences considered come from the Portuguese Wikipedia dumps, which were converted into plain text and segmentized. Our method consisted of comparing the distance between the triphone distribution of the sentences to a uniform distribution, with equiprobable triphones. The algorithm followed a greedy strategy in evaluating the distance metric. The results showed that our method is capable of extracting sentences in a much more uniform way, while comparing to a random selection. For 250 sentences, we were able to extract 854 new distinct triphones, in a set of sentences with a much higher type/token ratio. However, the method has its limitations. As discussed, it depends entirely on the quality of the pronunciation dictionary. If the pronunciation dictionary has some incorrect words, it might be the case that the algorithm favors such words, if they possess triphone types not registered in other words. As a future work, we intend to define a method that recognizes foreign words and excludes them from the selected sentences. All resources developed in this paper are freely available on the web<sup>3</sup>.

## 5.5 Extracted Sentences Sample

**Triphone Types:** 2307 | **Tokens:** 2959 | **Type/Token Ratio:** 0.78.

1. A ilha fica tão próxima da praia que, quando a maré baixa, pode ser atingida a pé.
2. Diadorim é Reinaldo filho do grande chefe Joca Ramiro traído por Hermógenes.
3. A Sicília tem alguns moinhos ainda em bom estado de conservação que lhe dão beleza e encanto.
4. Em geral, chegaram ao Brasil como escravos vindos de Angola, Congo, e Moçambique.
5. A sardinha é um peixe comum nas águas do mar Mediterrâneo.
6. Possuem esse nome pois costumam viver na plumagem dos pombos urbanos.
7. É brilhante, doce e muito harmônico, sem presença de metal na voz.
8. Para fechar Alessandro Del Piero fez outro aos 121'.
9. Roman Polanski dirige Chinatown com Jack Nicholson.

---

<sup>3</sup>Omitted for blind review.

10. A atriz sabe falar fluentemente espanhol.
11. Eles achavam Getúlio Vargas um problema.
12. Oppenheimer captura cavalo com peão.
13. Um bago tem tamanho médio não uniforme.
14. Segundo relatório da força aérea belga há confrontos com a União Soviética.
15. É irmão do também antropólogo Gilberto Velho.
16. Ganhou sete Oscar e oito Emmy.
17. Qual é minha perspectiva agora?
18. Ela é um fantasma verde, feminino!
19. Justin em seguida volta no tempo.
20. Nós fizemos um álbum do Korn.
21. Desde então Edílson é fã dessas bandas.
22. Há um só senhor uma só fé um só batismo.
23. Ivan Lins faria um show em Mossoró à noite.
24. Cresceram maior que um gato.
25. Há locações disponíveis em Tóquio no Japão.
26. Preso a um tronco nenhum lugar é seguro!
27. Hoje é professor emérito da UFBA.
28. Veio até aqui e não vai mergulhar?
29. Luís Jerônimo é um jovem rico.
30. Na hora pensei: "tenho que fazer isso?"
31. A campanha teve coordenação de Sanches.
32. A mulher que você me deu, fugiu.
33. Eu nunca tive um encontro com Bianca.

34. Homer jura vingança a Burns.
35. Beijo, me liga e amanhã sei lá!
36. Um colégio é como um ser vivo.
37. Sophie é filha de um amigo gay de Alan Greg.
38. Xuxa guarda rancor e é ambiciosa.
39. No mesmo ano conhece Aldir Blanc em Viena.
40. É um imenso painel reunindo um elenco famoso.
41. A Sé integra três belos órgãos.
42. Em ambos, Shannon conquistou medalha.
43. A terra é abundante em recursos como vinagre e óleo vegetal.
44. Faça sua escolha e bom jogo!
45. Quem é que poderia sonhar com algo assim?
46. Ela é ruiva com olhos azuis.
47. Deu a louca na chapeuzinho!
48. De onde venho e para onde vou?
49. Eu choro e sofro tormentas!
50. Um falcão pousa em um pedregulho.
51. Ninguém tenha medo, nem fraqueza!
52. É membro do grupo Monty Python.
53. A sondagem de Senna pela Benetton e a chegada à kart.
54. Isto é um negócio e a única coisa que importa é ganhar
55. Robert é um forte glutão da equipe.
56. Um bárbaro no exército romano?
57. Infância e juventude em Linz.

58. Já ir à argentina era muito bom!
59. Fiquei com inveja dele.
60. Há dragões ao redor do mundo!
61. Edmond é pai do biólogo Jean.
62. A mãe lhe telefonava às vezes.
63. Tonho é tímido, humilde e sincero.
64. André Jung ocupa um lugar central no fórum.
65. Lois pergunta: "você é um homem ou um alienígena?"
66. Sua voz é um assobio fino e longo.
67. Por isso é sempre bom conferir!
68. Celso Lafer recuperou a jóia e devolveu-lhe.
69. É próxima ao Rio Parnaíba.
70. Lendo aquilo fica bem difícil.
71. A faculdade de John Oxford até hoje possui fãs fiéis.
72. Existe uma crença moderna no dragão chinês.
73. Sean Connery já sugeriu que Gibson fosse James Bond.
74. A raiz dos dentes é longa.
75. Essa noite produziu um feito singular.
76. Fim da segunda guerra mundial.
77. –No Zorra, eu fazia humor rasgado.
78. Charles vê um homem ser morto em um tiroteio.
79. Tinham um novo senhor agora.
80. É comum ocorrerem fenômenos ópticos com estas nuvens.
81. Era um cão de pelo escuro e olhos negros.

82. Há títulos na região tcheca da Tchecoslováquia.
83. Raquel Torres vai investigar a área.
84. Clay foge e leva a jovem Jane como refém.
85. Djavan jogou futebol e hóquei no gelo na infância.
86. A origem do fagote é bastante remota.
87. Um jedi nunca usa a força para lucro ou ganho pessoal.
88. Chamavam José Alencar de Zézé.
89. Um código fonte é um sistema complexo.
90. A igreja tem um altar barroco.
91. Luís Eduardo pronunciou a senha: "esgoto".
92. Quanto ao sexo: macho ou fêmea?
93. A rádio Caxias cumpriu esse papel.
94. Roger Lion é um campeão orgulhoso que ama boxe.
95. Um outeiro é menor que um morro.
96. Hitoshi Sakimoto nasceu em Yokohama.
97. Nenhum isótopo do urânio é estável.
98. Chicago é um bairro tranquilo e festivo.
99. Hong Kong continua a utilizar a lei comum inglesa.
100. Só cinco funcionam como museus.



# **Chapter 6**

## **Listener**

O reconhecedor de pronúncia ora proposto será implementado a partir do motor de reconhecimento de fala Julius (Lee & Kawahara, 2009). Nove erros de pronúncia foram selecionados para serem tratados pelo Listener, assumindo-se, como pronúncia padrão, o General American (GA). O modelo acústico será compilado a partir de três corpora. Um de falantes nativos de inglês: TIMIT Acoustic-Phonetic Continuous Speech Corpus[16]; e outros dois de aprendizes: COBAI - Corpus Oral Brasileiro de Aprendizes de Inglês[17] e um corpus coletado especialmente para este trabalho, composto por leitura de sentenças foneticamente balanceadas. O dicionário a ser empregado é o CMU Pronouncing Dictionary, ao qual serão acrescentaremos as hipóteses de pronúncia dos aprendizes, por meio de regras. O modelo de língua será gerado a partir da Simple English Wikipedia em conjunto com um corpus de textos escritos por aprendizes de inglês, o COMAprend, um dos três corpus do projeto COMET da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. A eficiência do reconhecedor será avaliada por meio de medidas de Word Error Rate (WER), Character Error Rate (CER) e matrizes de confusão, aplicadas por meio de ten-fold cross validation sobre os dados dos corpora coligidos. De modo a verificar a viabilidade do método ora proposto, um protótipo do sistema foi elaborado e avaliado, detalhes são apresentados na Seção .

### **4 MATERIAIS E MÉTODOS**

#### **1 Levantamento dos Desvios de Pronúncia**

Na classificação dos erros de pronúncia, deu-se prioridade, especialmente, aos erros de pronúncia que afetam a compreensão e que são apresentados em trabalhos que consideram, no ensino da pronúncia do inglês, a transferência de padrões sonoros de L1 para L2.

A listagem dos erros de pronúncia a serem considerados pelo Listener foi obtida a partir da consulta aos trabalhos de Zimmer (2004), Godoy (2005), Zimmer et al. (2009) e Cristófaro-Silva (2012). Tais trabalhos analisam aspectos de transferência de L1 para L2

e estabelecem um método de ensino de pronúncia que leva em conta essa transferência, a fim de otimizar o aprendizado pelo aluno. De tal maneira, centra-se o estudo no ensino dos padrões de pronúncia que devem ser enfatizados para o falante do PB, a fim de melhor garantir a compreensão de sua pronúncia e sua eficiência comunicativa. No reconhecedor, optou-se por utilizar os nove tipos de erros elencados em Zimmer et al. (2009), por se tratar da investigação mais abrangente sobre o assunto. Os desvios de pronúncia selecionados estão sintetizados no Quadro 5.

Quadro 5: Desvios de pronúncias a ser analisados pelo Listener.

[pic]

## 2 Simplificação silábica

Um conjunto de 30 regras foi definido para gerar as variantes de pronúncia envolvendo casos de simplificação silábica. As regras se baseiam nos exemplos citados por Rauber e Baptista (2004), Rebello e Baptista (2007), Zimmer (2009) e Silveira (2012). A discussão dos contextos consta na Seção 2.1.1.4.1. O pseudocódigo com a implementação das regras está descrito na Figura 13.

Figura 13: Pseudocódigo com regras para geração das variantes de pronúncia envolvendo simplificação silábica.

Como se observa, as regras de simplificação silábica buscam cobrir, majoritariamente, quatro situações: (i) oclusivas em posição final de palavra; (ii) palavras foneticamente terminadas em consoantes, mas com <-e> final na forma escrita; (iii) clusters consonantais em início de palavra do tipo /sC(C)/; (iv) palavras iniciadas por /s ao/, com na forma ortográfica.

## 3 O Motor de Reconhecimento de Fala Julius

Neste trabalho, propomos a utilização do motor de reconhecimento de fala Julius (Lee & Kawahara, 2009), como a base do reconhecedor de pronúncia a ser desenvolvido. Julius é uma engine de alto desempenho e de código aberto para a construção de sistemas de reconhecimento de fala. Ele incorpora grande parte das técnicas do estado da arte em reconhecimento de fala e executa Reconhecimento de Fala Contínuo com Grande Vocabulário (LVCSR). Sua arquitetura está sintetizada na Figura 14.

[pic]

Figura 14: Arquitetura do motor de reconhecimento de fala Julius (Lee & Kawahara, 2009).

Como se observa, ele suporta a entrada de dados de áudio vindo de microfone, de arquivos já gravados, ou de streaming via internet. A saída é composta ou por dados de textos, a exemplo de ditado, ou por uma determinada ação solicitada ao Julius. No caso do

reconhecedor de pronúncia, a saída será constituída pela transcrição da fala e da avaliação de sua pronúncia.

Para se construir um reconhecedor fala através do Julius, são necessários: um modelo acústico, um dicionário de pronúncia e um modelo de língua. Mais detalhes sobre a elaboração desses modelos, de maneira a possibilitar o desenvolvimento do reconhecedor de pronúncia, são fornecidos a seguir.

#### 4 Elaboração do Modelo Acústico

O modelo acústico proposto será elaborado através de HMM e definido para trifones. Julius provê suporte a modelos acústicos de HMM obtidos a partir do HTK Hidden Markov Model Toolkit[18], disponibilizado pelo Speech Vision and Robotics Group, da Universidade de Cambridge. A abordagem utilizada na elaboração do modelo acústico é a de interlíngua. Portanto, ele será estimado a partir de dois tipos de corpora de fala: um de falante nativos do inglês: TIMIT Acoustic-Phonetic Continuous Speech Corpus[19]; outros dois de falantes nativos do PB, aprendizes de inglês como L2: COBAI - Corpus Oral Brasileiro de Aprendizes de Inglês[20] - e um corpus coletado especificamente para o Listener, composto por leitura de sentenças foneticamente balanceadas, isto é, sentenças contendo fones de acordo com sua frequência de ocorrência em uma dada língua. A construção de um modelo acústico interlingual busca contornar a dificuldade de reconhecer a fala de não-nativos, através da inserção de informação da pronúncia não-nativa no processo de treinamento do modelo acústico, por meio de dados com a pronúncia dos aprendizes (Wang, Schultz, & Waibel, 2003).

#### 5 Corpus de Nativos: TIMIT Acoustic-Phonetic Continuous Speech Corpus

Optou-se pela utilização do TIMIT (Garofolo, et al., 1993) como o corpus de falantes nativos inglês por se tratar de: um corpus bem modelado, robusto, foneticamente rico, amplamente utilizado e testado na área de reconhecimento de fala, em cerca de duas décadas de pesquisa, além de cobrir os dialetos majoritários do inglês americano (Lopes & Perdigão, 2011). O corpus TIMIT foi elaborado, conjuntamente, pelo Instituto de Tecnologia de Massachusetts (MIT), SRI Internacional e Texas Instruments Inc. (TI) com o propósito fornecer dados para a realização de estudos de fonética acústica do inglês, bem como para o desenvolvimento de sistemas automáticos de reconhecimento de fala. Ele contém gravações de cerca de 630 falantes, dos oito principais dialetos do inglês americano. As gravações foram elaboradas a partir da leitura de dez sentenças criadas artificialmente, de modo a capturar ambientes fonéticos relevantes. O TIMIT foi verificado manualmente e está transcrito ortográfica e foneticamente, adicionalmente, foi feito o alinhamento temporal entre o arquivo de áudio e as transcrições. Os arquivos estão separados em sentença, amostrados a 16kHz

com 16 bits por amostra. O TIMIT está disponível para venda no Linguistic Data Consortium (LDC) e será adquirido pelo Núcleo Interinstitucional de Linguística Computacional (NILC).

#### 6 Corpus de Não-nativos: Corpus Oral Brasileiro de Aprendizes de Inglês (COBAI)

O COBAI (Mello, Avila, Neder-Neto, & Orfano, 2012) constitui a primeira iniciativa brasileira que busca compilar e distribuir, de forma aberta, um corpus de fala anotado de aprendizes de inglês, falantes nativos do PB. O COBAI integra o Louvain International Database of Spoken English Interlanguage (LINDSEI) e vem sendo organizado pelo Laboratório de Estudos Empíricos e Experimentais da Linguagem (LEEL), da Faculdade de Letras, da Universidade Federal de Minas Gerais (UFMG). O propósito do LINDSEI é a disponibilização de corpora de fala de aprendizes de inglês, com diferentes backgrounds de língua nativa. O COBAI segue as diretrizes de transcrição do LINDSEI, que utiliza padrões XML na anotação. A transcrição é do tipo ortográfica e agrupa informações de: troca de turno, sobreposição de fala, pausas, hesitações, formas reduzidas e algumas indicações fonéticas e prosódicas. Atualmente, cerca de 60% do corpus está anotado. O corpus consiste em 50 gravações de 15 minutos, que incorporam uma narrativa, uma entrevista e uma descrição. Os arquivos estão separados e amostrados a 44kHz com 16 bits por amostra. Todas as gravações foram feitas com falantes nativos do PB, aprendizes de inglês. O grau de conhecimento da língua inglesa dos participantes é variado, havendo desde aprendizes com baixa proficiência até indivíduos proficientes.

Uma contribuição colateral deste projeto será a finalização da transcrição ortográfica do COBAI e a anotação, no corpus, dos nove tipos de erros que serão analisados pelo Listener. De modo a finalizar a transcrição ortográfica do COBAI, pretende-se utilizar um reconhecedor de fala, no caso, o Dragon Naturally Speaking v.12 Premium, de modo a obter uma transcrição ortográfica inicial. A partir disso, realizaremos a revisão da transcrição, no intuito de corrigir os erros e adequar o formato ao que é proposto pelo LINDSEI. Para a transcrição dos desvios de pronúncia, propõe-se o seguinte método: após ter-se obtido a transcrição ortográfica de todo o corpus, será criado um script em Python para percorrer cada palavra transcrita ortograficamente, conferi-la no CMUdict e extrair a transcrição fonética que lá está registrada. De tal forma, obteremos uma versão do COBAI transcrita com a pronúncia canônica do General American (GA), que está registrada no CMUdict. A seguir, será realizada a revisão das transcrições fonéticas, corrigindo-as quando os aprendizes cometem algum dos nove tipos de erros que o Listener avaliará.

#### 7 Corpus de Não-nativos: Corpus de Leitura de Sentenças Foneticamente Balanceadas por Aprendizes

A intenção inicial do projeto era utilizar apenas o COBAI como corpus fala de não-nativos. Porém, após ter-se desenvolvido o protótipo do reconhecedor, conforme será descrito na

Seção , foi possível observar que muitos dados do COBAI terão de ser desconsiderados e, de tal forma, somente o COBAI não será suficiente para fornecer o número de horas necessário para a estimação de um bom modelo acústico. Por isso, decidiu-se criar um corpus específico para o desenvolvimento deste trabalho.

O propósito é compilar um corpus de aprendizes, em situação de leitura de frases pré-definidas, foneticamente balanceadas, o qual seja gravado em um ambiente com isolamento acústico. Os objetivos, com isso, são três: i) assegurar a boa qualidade do áudio, mantendo baixa a relação sinal-ruído; ii) garantir que todas as combinações de trifones estejam presentes na base de dados do modelo acústico, uma vez que as sentenças serão foneticamente balanceadas; e iii) facilitar a tarefa de transcrição, já que a duração da pesquisa de mestrado é curta. Propõe-se seguir as diretrizes de compilação e anotação de corpora, tal como descrito por Hovy e Lavid (2010).

Pretende-se que os detalhes do método de compilação e anotação do corpus sejam definidos em visita técnica a ser realizada na Universidade de Coimbra, sob supervisão da Profa. Sara Candeias, no período de 28 de janeiro a 28 de fevereiro de 2014. O plano de tarefas proposto para a visita inclui:

- a definição do tamanho do corpus a ser compilado;
- a definição do nível de detalhe da transcrição fonética (qual o inventário de fones e xenofones utilizar);
- a definição do tipo de hesitação ou disfluência a ser anotada;
- a discussão de uma métrica de riqueza fonética (?), proposta para a extração de sentenças foneticamente balanceadas;
- o teste e a avaliação da métrica, a partir de um corpus de textos de aprendizes de inglês, o COMAprend.

No que diz respeito às sentenças foneticamente balanceadas, há, para o inglês, diversas listas disponíveis, como as Harvard Sentences (IEEE, 1969), os TIMIT Sentence Prompts (Garofolo, et al., 1993), as MOCHA-TIMIT Sentences (Wrench, 1999), além de diversas listas fornecidas pela Carnegie Mellon University para o motor de reconhecimento Sphinx (Lee, Hon, & Reddy, 1990). No entanto, dado que essas listas foram elaboradas para falantes nativos de inglês, há diversas palavras de baixa frequência, bem como sentenças cuja estrutura sintática é pouco usual. Em um contexto de aprendizes, isso é problemático, pois pode ocasionar disfluências na fala do aprendiz, causando problemas na utilização dos dados, além de padrões de pronúncia altamente irregulares, quando o aprendiz desconhecer a palavra que está lendo.

Para contornar o problema, objetivamos criar um conjunto de sentenças foneticamente balanceadas, a partir de um corpus de texto de brasileiros aprendizes de inglês, como o

COMApred (Tagnin & Fromm, 2009). O COMApred possui textos em formato ortográfico. A fim de se obter a transcrição fonética dos textos, será elaborado um script semelhante ao descrito, anteriormente, na Seção 3.1.3.2.

### 8 Elaboração do Dicionário de Pronúncia

O dicionário de pronúncia será formado com base no CMU Pronouncing Dictionary, o qual será acrescido de transcrições das possíveis pronúncias desviantes dos aprendizes, por meio de regras transformacionais. Dicionários contendo tais características são também chamados na literatura como dicionários multipronúncia (Strik & Cucchiarini, 1999).

### 9 CMU Pronouncing Dictionary

A base do dicionário provirá do CMU Pronouncing Dictionary (também conhecido por CMUDict), disponibilizado no motor de reconhecimento de fala Sphinx, pela Universidade Carnegie Mellon. O CMUDict constitui um dicionário de pronúncia machine readable de referência na área de reconhecimento de fala. Atualmente, nele estão registradas 131.411 entradas, transcritas foneticamente em formato ARPAbet (Zue & Seneff, 1988). O Quadro 6 ilustra a entrada de algumas palavras no dicionário.

Quadro 6: Exemplo de entradas no dicionário de pronúncia do CMU Pronouncing Dictionary.

[pic]

Como se observa, o dicionário possui três campos: (i) um interno, para identificação da palavra, (ii) um com a palavra em sua forma ortográfica, convencionalizada em letras maiúsculas, (iii) e um último campo com a transcrição fonética da palavra, em formato ARPAbet.

### 10 Adição das Formas Variantes de Pronúncia do Aprendiz

Serão utilizadas regras transformacionais para acrescentar ao dicionário as possíveis hipóteses de pronúncia do aprendiz. A utilização de regras transformacionais é de fácil implementação computacional, correspondendo a simples estruturas de seleção, ou construções condicionais. Contexto para aplicação de regras, como os elencados no Quadro 7, serão levantadas, de acordo com a literatura linguística de ASL, e utilizados para as variantes de pronúncias do aprendizes, a partir das palavras do CMUDict. Casos marginais, de contexto muito restrito ou que não se adaptem às regras criadas, serão adicionados, manualmente, em formato de dicionário de exceções.

Quadro 7: Contextos para aplicação de regras de simplificação silábica.

[pic]

### 11 Elaboração do Modelo de Língua

Um modelo de língua será fornecido ao Listener, de modo a possibilitar seu uso em um contexto de ditado. Há diversos modelos de língua para o inglês (como o Gigaword[21],

CSR LM-1[22], HUB4[23]). Porém a grande maioria desses modelos foi gerada a partir de corpora de artigos de jornal e é sabido que textos jornalísticos de jornais tradicionais, para públicos A, B e C tendem a possuir estrutura sintática e vocabulário complexos, dado que as primeiras orações de uma notícia tendem a compactar muita informação e podem trazer jargão não dominado por aprendizes (Canning, 2002). Como a intenção é lidar com a fala de aprendizes, propomos a criação de um modelo de língua que seja mais simplificado e condizente com a sintaxe dos aprendizes. Será elaborado um modelo de língua estatístico, que considera trigramas na análise e se baseia em HMMs.

### 12 Simple English Wikipedia

Como corpus para criação do modelo de língua, utilizaremos a Simple English Wikipedia, cuja proposta é desenvolver uma Wikipedia em inglês de nível básico, com vocabulário e construções sintáticas mais simples, de modo a prover acesso a crianças, estudantes, adultos com baixo nível de letramento e aprendizes de inglês como L2. A versão disponível da Simple English Wikipedia, referente ao dia 16 de janeiro de 2014, possui 108.665[24]. Todos os arquivos estão codificados em XML. A ferramenta SRILM (The SRI Language Modeling Toolkit)[25] será utilizada para auxiliar na criação do modelo de língua.

### 13 Elaboração da Interface Web

A interface web desenvolvida para o Listener emprega conceitos de gamificação, com o propósito de estimular a participação dos usuários e de tornar o processo de aprendizagem mais aprazível.

### 14 Gamificação

Segundo o historiador Huizinga (1938), um jogo constitui:

“uma atividade ou ocupação voluntária, exercida dentro de certos e determinados limites de tempo e de espaço, segundo regras livremente consentidas, mas absolutamente obrigatórias, dotado de um fim em si mesmo, acompanhado de um sentimento de tensão e de alegria e de uma consciência de ser diferente da ‘vida cotidiana’” (Huiziga 1938).

Jogos têm, desde sempre, fascinado as pessoas. Milhões de pessoas vibram e se emocionam ao assistir a uma partida de futebol de seu time preferido. Mestres do xadrez, como Garry Kasparov, chegam a dedicar cerca de seis horas diárias para dominar o jogo (ICC, 1998). Massive Multiplayer Online Role Playing Games mobilizam milhões de pessoas online a um mesmo tempo. Há, também, casos trágicos, como o do taiwanês que teve problemas vasculares e faleceu, logo após jogar Diablo 3 em uma lan-house por 40 horas ininterruptas (Daily Mail, 2012).

De acordo com Lazzaro (2004), as pessoas jogam, basicamente, por um dentre os quatro motivos a seguir: (i) pelo divertimento que o jogo proporciona, (ii) pela competitividade que ele incita, (iii) pelas sensações diversas que ele pode gerar - surpresa, alegria, temor, etc.; e

(iv) pelo pretexto para socializar com os amigos. Jogos bem desenhados tendem sempre a explorar esses pontos, a fim de fidelizar jogadores.

Deburr (2013) define a gamificação a aplicação de estratégias e técnicas usadas no design de jogos em outros contextos, que não jogos. Zicherman e Cunningham (2013) trazem uma definição análoga, para os autores a gamificação é “o processo de utilizar a mecânica dos jogos no intuito de motivar usuários a resolver problemas”, sendo que a mecânica comporta sete aspectos: (i) sistema de pontos, (ii) níveis, (iii) rankings, (iv) atribuição de títulos, (v) desafios/missões, (vi) busca de jogadores e (vii) um loop de estímulo social constante. Explicar cada um desses aspectos

A gamificação tem sido aplicada com sucesso a uma gama de contextos. Há, por exemplo, comunidades de perguntas e respostas, como Stackoverflow e Yahoo! Respostas, que a empregam a fim de motivar seus usuários a participarem da comunidade e a cooperarem entre si. Os usuários que fornecem as melhores respostas recebem pontos no site, sobem níveis, melhoram nos rankings, tornando-se cada vez mais influentes na comunidade. Aplicativos para celulares que incitam a prática de corrida, como o Runstatic e o Nike+, também têm utilizado a gamificação com êxito. Tais aplicativos extraem dados do percurso do usuário via GPS e lhes fornece estatísticas de sua corrida, de maneira que os usuários podem monitorar suas corridas para vencer metas, bater recordes pessoais e, também, competir com os amigos. É possível também desafiar outros usuários, para ver quem corre mais rápido, quem faz um mesmo percurso maior em menor tempo, etc.

No âmbito da educação, a gamificação tem sido explorada no chamado edutainment - palavra amálgama formada a partir de education e entertainment (XXX:XX). Segundo Zicherman e Cunningham (2013), a indústria tem falhado em utilizar a gamificação para aplicações educacionais. Muitas vezes, o aspecto instrucional é ressaltado de maneira exagerada, de modo que os jogos se tornam chatos e as pessoas se sentem desmotivadas para jogá-los. Para os autores, o jogo educacional que melhor explorou os aspectos da gamificação foi “Where in the World Is Carmen Sandiego?”, cujo propósito é ensinar Geografia, mais especificamente, o nome e a localização dos países e de suas capitais. Em “Where in the World Is Carmen Sandiego?”, o jogador encarna um detetive que deve juntar pistas para encontrar a criminosa Carmen Sandiego. Ao longo da jornada, o jogador visita vários países e encontra pessoas que lhe dão pistas sobre onde Carmen Sandiego pode estar escondida. As pistas contêm informações gerais sobre os países, como “Uma pessoa suspeita veio aqui e disse que iria viajar à terra dos Vikings” ou “Eu a vi embarcando em um avião cuja bandeira era vermelha e azul”. A Figura 15 apresenta uma tela do jogo.

Figura 15: Captura de tela do jogo "Where in the World Is Carmen Sandiego?".

“Where in the World Is Carmen Sandiego?” foi lançado em 1985 e, desde então, diversas empresas de edutainment têm tentado repetir o sucesso do jogo, mas sem muito êxito. Merece destaque o lançamento, em 2012, da plataforma para ensino de idiomas Duolingo (XXX:XX), que emprega gamificação e, só no Brasil, já consta com cerca 6,8 milhões de alunos[26]. Como se trata de uma aplicação de ensino de línguas que possui também um módulo de ensino de pronúncia, o Duolingo será tratado em mais detalhes na Seção XXX.

The leaderboard of today has seen some radical redesign since the heyday of pinball machines and quarter arcades. In the era of Facebook and the social graph, leaderboards are mostly tools for creating social incentive

In many instances, such as losing weight or even writing a book, it's difficult for a player to understand where he is at the outset or during early interactions. Moreover, the length and complexity of the overall journey is such that sometimes players can be paralyzed by the seeming lack of progress. Especially in health, education, and other “epic journey” contexts, feedback forms the most important overarching game mechanic, intricately tied to score and progress.

LER Jon Radoff’s Game On

LER Jesse Schell’s The Art of Game Design: A Book of Lenses

Segundo Zicherman e Cunningham (2013), como modelo de negócios, a gamificação transforma a relação empresa-usuário em uma relação simbiótica, cujo benefício para o usuário é o prazer obtido com o jogo e para a empresa é a fidelização do usuário com o produto oferecido.

O corpus de erros induzidos

Para este protótipo, reduziu-se a criação do corpus de leitura de sentenças foneticamente balanceadas a uma tarefa mais simples. Um informante do sexo masculino foi gravado, lendo palavras em isolamento e, propositalmente, enunciando-as com erros de transferência de L1 para L2, a fim de simular variantes de pronúncia que ocorrerão no corpus. As palavras foram selecionadas a partir da lista das 5.000 palavras mais frequentes da língua inglesa, segundo o Corpus of Contemporary American English (COCA)[29]. Cerca de 2h20min de fala foram compiladas. A gravação ocorreu em uma sala fechada, com baixo ruído externo, não isolada acusticamente.

O COCA é o maior corpus de inglês disponível atualmente, tendo sido compilado a partir de cerca de 160.000 textos escritos e transcritos, de vários gêneros, os quais totalizam 450 milhões de tokens. O projeto é coordenado por Mark Davies, professor de Linguística de Corpus da Brigham Young University (BYU). O corpus é gratuito, mas as listas de frequência são pagas, a exceção da menor delas, contendo as 5.000 palavras mais frequentes do inglês, a qual foi utilizada no protótipo.

Um script em Python foi elaborado para aplicar à lista as regras de simplificação silábica elencadas na Figura 18. Quando havia contexto para aplicação de qualquer uma das regras, a palavra era selecionada e adicionada a um banco de palavras. Figura 18 sintetiza o funcionamento do script. Das 5.000 palavras mais frequentes no COCA, 1.855 apresentaram contexto em que é possível haver simplificação silábica, tais palavras, portanto, foram selecionadas para leitura.

[pic]

Figura 18. Fluxograma de funcionamento do script seletor de palavras.

Um microfone condensador de diafragma pequeno (1/2“), com padrão polar cardióide, do tipo Superlux S241/U3 foi utilizado nas gravações. O microfone foi ligado a uma mesa de som analógica Yamaha MG102, através de cabos平衡ados XLR, e alimentado por corrente fantasma (48V). A captação do áudio se deu por um laptop LG A51, via cabos RCA. O software Audacity (Mazzoni & Dannenberg, 2000) foi usado na captação, tendo-se ativado o filtro supressor de ruído. O ambiente de gravação consistiu de uma sala fechada, sem isolamento acústico, em situação de baixo ruído externo.

Os dados foram segmentados através do Adintool, de forma similar ao método descrito na Seção 3.4.1.1. O Appendix II reúne os parâmetros utilizados na segmentação. Os arquivos foram alinhados com sua transcrição ortográfica, manualmente. A seguir, procedeu-se à transcrição fonética das palavras, de forma também manual. Detalhes sobre a transcrição constam, a seguir, na Seção 3.4.2.

O software Praat (Boersma & Weenink, 2014) foi utilizado na transcrição, para visualizar o espectrograma e, também, tocar os arquivos de áudio. O LibreOffice Calc (The Document Foundation, 2014) foi empregado para facilitar a organização das transcrições. Atenção especial foi dedicada à análise da ocorrência ou não do fenômeno de simplificação silábica, erro alvo do protótipo.

[pic]

Figure 1: Workspace utilizado na transcrição fonética do corpus.

#### 4 Elaboração do modelo dicionário de pronúncia

O modelo de pronúncia foi elaborado com base no CMU Pronouncing dictionary, versão 0.7a, de 01 de abril de 2008[30], a qual conta com 133.315 entradas. De modo a manter a compatibilidade com o Julius, 16 palavras com símbolos especiais (“&”, “/”, “;”, etc.) foram retiradas, restando-se 133.304 entradas. O dicionário foi, então, reordenado para manter a ordem esperada pelo motor de reconhecimento.

A seguir, foram selecionadas do dicionário apenas as 1.855 palavras que apresentaram contexto possível de haver simplificação silábica (cf. Quadro 8). A intenção de restringir o modelo de pronúncia às palavras selecionadas para gravação deu-se de maneira a diminuir a

confusão do reconhecedor. Como se trata de um protótipo, treinado a partir de poucas horas de áudio, o modelo acústico não é suficientemente robusto para percorrer um espaço de busca de 133.304 palavras e obter boa acurácia no reconhecimento.

A ferramenta HDMan, do HTK Toolkit, foi empregada na elaboração do dicionário de pronúncia contendo as 1.855 palavras. Tal ferramenta visa a criar novos dicionários a partir de dicionários fontes. Seu funcionamento dá-se da seguintes forma: listas de palavras são fornecidas como entrada, em conjunto com dicionários fontes e um script de edição, os dados dos dicionários-fontes são processados de acordo com as opções especificadas no script e tem-se como saída um novo dicionário, o qual contém as palavras fornecidas na lista, juntamente com as respectivas pronúncias.

Tendo-se obtido o dicionário com as 1.855 palavras utilizadas na gravação, procedeu-se à inserção das variantes de pronúncia. Para esse fim, um script em Python, com 21 regras de transcrição, foi elaborado e aplicado ao dicionário. O script percorreu cada uma das palavras, analisando a sequência de grafemas e de fones, de modo a verificar se havia contexto propenso à simplificação silábica. As regras foram compostas por estruturas condicionais do tipo if... then e seus contextos de aplicação estão descritos no Quadro 7. O conjunto de fones utilizado na elaboração das variantes é constituído pela união do inventário fonético do PB segundo Cristófaro-Silva (2005), e o do AmE segundo Ogden (2012).

Certas regras criam contexto fonético para que outras se apliquem.

Por exemplo, após se aplicar regra 15, [Vm#] > [VN<sub>ASAL</sub>b<sub>i</sub>#] / , é possível se aplicar também a regra 7, [b<sub>#</sub>] > [b<sub>i</sub>#], de forma a gerar uma nova variante a partir de outra já gerada. Aplicando-se a regra 15 à palavra “bomb” [bom], seria possível obter [bomb], com isso, haveria contexto para a aplicação da regra 7, gerando-se também a variante [bombi]. De tal maneira, as regras têm de ser aplicadas iterativamente, até esgotar todas as possibilidades de criar novas variantes.

No caso do dicionário de pronúncia do protótipo, as 21 regras foram aplicadas cinco vezes, até o número de palavras do dicionário se estabilizar, isto é, até a aplicação das regras não adicionar nenhuma palavra nova ao dicionário. O Gráfico 1 descreve o crescimento do número de palavras.

Gráfico 1: Crescimento do número de palavras do dicionário de pronúncia.

[pic]

Como se observa, houve um aumento considerável no número de palavras. Para 1.855 palavras-base, foram geradas 5.742 variantes, de maneira que o dicionário final contabilizava 7.597 possibilidades de pronúncia. O valor médio de pronúncias por palavras foi de 4,1. Certas palavras chegaram a apresentar até 24 possibilidades de pronúncia, como “employment”,

“entertainment”, “independent” e “unemployment”. A Tabela 4 resume as estatísticas do número de variantes por palavra.

Tabela 4: Variantes de pronúncia por palavra.

[pic]

O grande número de possibilidades geradas para algumas palavras não era esperado e pode trazer prejuízos para o reconhecedor. Uma palavra aparentemente simples, como “combined”, apresentou 12 variantes de pronúncia apenas no que diz respeito à simplificação silábica. Considerando- se a influência da escrita na fala, seria possível que o aprendiz pronunciasse [ka] como [ko] e inserisse um [e] em razão de haver na forma escrita, de tal forma, “combined” apresentaria 48 variantes de pronúncia ( $= 4 \times 12$ ). Tendo em vista que objetivo é tratar nove tipos de erros de pronúncia (cf. Seção 3.1.1), é possível que, ao se obter todas as regras, o número de variantes de pronúncia geradas seja tão grande que o reconhecimento seja prejudicado. Caso isso ocorra, uma saída viável seria a criação de dicionários separados para cada tipo de erro ou, mesmo, a utilização de um especialista para cercear as variantes geradas.

### 6.0.1 Evaluation

Since we are proposing a method to build a pronunciation training system, we could evaluate our method in two ways: extrinsic or intrinsically.

The extrinsic evaluation is the one which considers the purpose of the system, that is, it assess if users are appropriately learning with the system, by improving their pronunciation perception and production. Therefore, the extrinsic evaluation of the system would require the development of a longitudinal study, in which a significant portion of individuals would be analyzed for a large amount of time, with regular interviews and tests to check their pronunciation skills. Given time limitations and also the scope of this Master’s thesis, an extrinsic evaluation of the system is not feasible. Instead, we are going to evaluate the method solely in an intrinsic way, which consists of assessing the the pronunciation training system itself, regardless of its practical purpose.

To put another way, our evaluation

### 6.0.2 Evaluation Metrics

The Word Error Rate (WER) measures the performance of an ASR system in terms of how much its output (i.e. the words it recognized) diverges from a given reference. The metric is defined as follows:

$$WER = \frac{S_w + D_w + I_w}{N_w} \quad (6.1)$$

where  $S_w$  corresponds to the number of word substitutions,  $D_w$  is the number of deletions,  $I_w$  is the number of insertions, and  $N_w$  is the number of words in the sentence used as reference, that is to say, the expected output.

The Phone Error Rate (PER) analyses the system's performance in recognizing phones. It is calculated exactly like the WER, but considering phones as units. It is defined as:

$$PER = \frac{S_p + D_p + I_p}{N_p} \quad (6.2)$$

where  $S_p$  corresponds to the number of phone substitutions,  $D_p$  is the phone deletions,  $I_p$  is the phone insertions, and  $N_p$  is the total number of reference phones in the transcription.

Differently from The Real Time Factor (Real Time Factor (RTF)) is a metric used to measure the computational performance of an ASR system. It is defined as:

$$RTF = \frac{P_i}{T_i} \quad (6.3)$$

where  $P_i$  is the time it takes to process an input  $i$ , and  $T_i$  is the duration of  $i$ .

The RTF is a metric used to measure the computational performance of an ASR system. It is defined as:

$$RTF = \frac{P_i}{T_i} \quad (6.4)$$

where  $P_i$  is the time it takes to process an input  $i$ , and  $T_i$  is the duration of  $i$ .

Um reconhecedor de pronúncia pode ser avaliado de dois modos: intrínseca ou extrínseca. A avaliação intrínseca (também chamada *in vitro*) é aquela que se atém à avaliação do reconhecedor em si, isolado de seu fim prático. Em outras palavras, na avaliação intrínseca, o foco de avaliação é a tarefa de reconhecimento, avalia-se a eficiência do reconhecedor em obter, dado um sinal acústico, sua contraparte textual.

Para isso, usam-se, comumente, métricas como Word Error Rate (WER), Character Error Rate e Matrizes de Confusão (Chen, Beeferman, & Rosenfeld, 1998; Goronzy, 2002). Já a avaliação extrínseca (também chama *in vivo*) é aquela se volta à avaliação do propósito para o qual o reconhecedor foi construído, no caso de um reconhecedor de pronúncia, o objetivo final é o aprendizado de pronúncia pelos seus usuários. Métricas utilizadas para nesse tipo de avaliação são a Goodness of Pronunciation (GOP) e a Weighted Goodness of Pronunciation (wGOP), além da verificação do desempenho dos aprendizes em testes de proficiência de língua inglesa (Witt, 1999).

Neste projeto, será realizada apenas a avaliação intrínseca do reconhecedor de pronúncia, através das métricas Word Error Rate (WER), Character Error Rate e Matrizes de Confusão, aplicadas sobre os dados de ambos os corpora coletados, por meio de ten-fold cross validation.

A escolha por este tipo de avaliação se deveu à natureza do projeto: como se trata de um trabalho de mestrado, não haveria tempo hábil para realizar um estudo longitudinal com aprendizes de inglês, de modo a avaliar a eficiência do Listener no ensino de pronúncia.

## 6.1 Tools and Libraries

Lorem ipsum quod dolor sit amet.

### 6.1.1 HTK

Lorem ipsum quod dolor sit amet.

### 6.1.2 Julius

Lorem ipsum quod dolor sit amet.

## 6.2 Speech Corpora

Lorem ipsum quod dolor sit amet.

### 6.2.1 TIMIT

Lorem ipsum quod dolor sit amet.

### 6.2.2 WSJ0

The CSR-I WSJ0 corpus was compiled by Garofolo et al. [48], within the DARPA Spoken Language Program in order to support research on large-vocabulary Continuous Speech Recognition (CSR) systems.

It focuses on American English and contains read speech of texts drawn from a corpus with Wall Street Journal articles. The texts to be read were selected to fall within either a 5,000-word or a 20,000-word subset of the WSJ text corpus. All verbal punctuation is read out aloud and the prompting texts have been pre-filtered to insure unambiguous pronunciations of words. The corpus comprises spontaneous dictation by journalists with varying degrees of experience in dictation, the precise number of speakers is informed in the documentation. As for the recording environment, a Sennheiser close-talking head-mounted microphone was used together with a secondary microphone of varying types.

Table 6.1 describes a summary of the WSJ0 corpus.

For building the acoustic model, we decided to use only a portion of the WSJ0. This decision was made since a considerable number of files in the WSJ recordings had:

1. disfluencies phenomena, such as mispronunciations, verbal deletions, false starts and spoken word fragments;
2. emphatic stress in words which would normally not be stressed due to lexical or syntactic factors;
3. non-speech events (chair squeak, cross talk, door slams, paper rustle, phone ring, etc.)
4. truncated audio files.

All these recordings would degrade the estimation of the acoustic model, giving rise to poor phone or triphone HMMs. On account of this problem, we excluded all these files from the training process.

After that, to check the consistency of the transcription, we used forced alignment with an acoustic monophone model trained over all other English corpora. Forced alignment was performed through a general-purpose Viterbi recognizer, which employed beam search to find the most likely HMM states for an utterance. The beam-width was set to 250. That is, for each audio file, if the transcription provided did not correspond to any alignment found by expanding each node over the best 250 hypotheses, the file was pruned and not considered for training.

The portion of the WSJ0 that we used is detailed in Table 6.2

Table 6.1 Summary of the entire WSJ0 Corpus.

Recorded files	X
Total speech time	X
Average time per file	X
Original format	X
Number of different speakers	X

Table 6.2 Summary of WSJ0 Part We Used.

Recorded files	X
Total speech time	X
Average time per file	X
Original format	X
Number of different speakers	X

### 6.2.3 SpeechDat

Lorem ipsum quod dolor sit amet.

#### 6.2.4 Listener's Corpus

Lorem ipsum quod dolor sit amet.

Table 6.3 Summary of the Listener's Corpus.

Recorded files	6,892
Total speech time	6.8 hours
Average time per file	1.02 seconds
Original format	WAV 16kHz
Number of different speakers	53

### 6.2.5 Oxford Dictionary AmE Corpus

The Oxford Dictionary American English (AmE) corpus was compiled by web crawling specially to this project. Oxford University Press has been making dictionaries for the English language for more than 150 years. Their dictionaries are very traditional and widely known whether in lexicographers' or laymen's circles. Recently, they made the dictionaries publicly available on the web<sup>1</sup>. For the AmE version, one can browse 350,000 words, definitions, and entries, together with over 600,000 *synonyms*<sup>2</sup>. A word example can be found in Figure 6.1.

As one may notice in Figure 6.1, the entry has many information: (i) the word itself, in orthographic form; (ii) word syllabification; (iii) its pronunciation; (iv) the word audio file; (v) Part of Speech (POS) data; (vi) definition; (vii) some example sentences; (viii) and a list of synonyms. The pronunciation follows Oxford's own transcription convention, which can be mapped on the IPA as in Table 6.4.

For compiling the corpus, we built a spider through Scrapy [34], a web crawling framework for Python specially designed to crawl websites and extract structured data from their pages. In total, 49,263 entries were crawled. This number was defined in order to be consonant with the dictionary's legal aspects, which defines that only a fraction of its content might be downloaded either personal or institutional use. For each word, we crawled three fields: the word, its transcription and the audio file. According to Oxford University Press' legal notice, we may not display or distribute any of the crawled content in any media, nor we may use commercially. Therefore, we used the data only for estimating the parameters of the acoustic model.

Some xenophones, i.e. phones from other languages, might be seen in Table 6.4. This happens because Oxford Dictionary also register some loanwords (specially those coming from French) with their original pronunciation, like "Utrecht" and "bon". Since our goal is to deal only with Brazilian-accented English, words such as these were excluded, so no word among the 49,263 entries contain xenophones.

Oxford Dictionary's were recorded by many speakers (both male and female), with high-quality microphones, in sound-isolation rooms. The audios are saved in MP3 format with Variable Bit Rate (VBR), for this reason we had convert each file into WAV and downsample them to 16 kHz. The quality of the audio is excellent with a very high Signal-to-Noise Ratio (SNR), as can be seen in the spectrogram for the word "happiness".

As one may observe in the regions of silence at the beginning and at the end of the utterance, the background noise approaches to zero. A summary of the corpus can be seen in ??.

---

<sup>1</sup><http://www.oxforddictionaries.com/>

<sup>2</sup><http://www.oxforddictionaries.com/words/content-help>

Table 6.4 Oxford Dictionary phone convention.

#	Oxford Phone	IPA Phone	Example	Transcription
1	(h)w	aaaaa	when	trans
2	ä	ɔ	hot	trans
3	ô	ɔ	saw	trans
4	ōo	u	too	trans
5	ā	eɪ	day	trans
6	ē	i	see	trans
7	ī	aɪ	my	trans
8	ō	oʊ	no	trans
9	ə	ə	ago	trans
10	œ (foreign)	œ	Goethe (German)	trans
11	oo	ʊ	put	trans
12	γ (foreign)	ʏ	Utrecht (French)	trans
13	a	æ	cat	trans
14	b	b	bad	trans
15	CH	tʃ	chip	trans
16	d	d	day	trans
17	e	ɛ	bed	trans
18	e(ə)r	ɛr	hair	trans
19	f	f	fight	trans
20	g	g	get	trans
21	h	h	hi	trans
22	i	i	sit	trans
23	i(ə)r	ɪr	near	trans
24	j	dʒ	jar	trans
25	k	k	kick	trans
26	KH	x	loch	trans
27	l	l	lie	trans
28	m	m	man	trans
29	N (foreign)	˜	bon (French)	trans
30	n	n	no	trans
31	NG	n	ring	trans
32	oi	ɔɪ	boy	trans
33	ou	oʊ	how	trans
34	p	p	pie	trans
35	r	r	run	trans
36	s	s	save	trans
37	SH	ʃ	she	trans
38	t	t	time	trans
39	TH	ð	this	trans
40	TH	θ	thin	trans
41	v	v	vow	trans
42	y	y	yes	trans
43	z	z	zoo	trans
44	ZH	ʒ	decision	trans

Fig. 6.1 Entry example in the Oxford Dictionary online.

Table 6.5 Summary of the Oxford Dictionary AmE Corpus.

Recorded files	49,263
Total speech time	4 hours
Average time per file	1.02 seconds
Original format	MP3 (VBR)
Number of different speakers	Unknown

### **6.2.6 Cambridge Dictionary**

  Lorem ipsum quod dolor sit amet.

### **6.2.7 OGI-22**

  Lorem ipsum quod dolor sit amet.

### **6.2.8 Westpoint**

  Lorem ipsum quod dolor sit amet.

### **6.2.9 LapsBM**

  Lorem ipsum quod dolor sit amet.

### **6.2.10 Youtube**

  Lorem ipsum quod dolor sit amet.

Table 6.6 Summary of the Listener's Corpus.

Recorded files	6,892
Total speech time	6.8 hours
Average time per file	1.02 seconds
Original format	WAV 16kHz
Number of different speakers	53

## 6.3 Building the Acoustic Model

### 6.3.1 The Phoneset

Since our goal is to build an ASR system capable of recognizing non-native speech, we propose to use an interlingual phoneset as the basis of the pronunciation model. By doing this, we can define HMM models for estimating phones which are part of both the speaker's native language (BP) and the target language (AmE).

A straightforward approach would be to look up the literature in phonetics and phonology in order to find a BP–AmE interlingual phoneset. However, to the best of our knowledge, no previous works were carried out in this regard. There are papers addressing specific mispronunciations, such as those discussed in subsection 2.2.3, which list some occurring interlingual phones, but there is not a wide-ranging study available about this matter.

Therein we had to develop our own interlingual phoneset. For simplicity, we decided to adapt the union set formed by the phones contained in two machine-readable dictionaries, one for AmE: Carnegie Mellon University Pronouncing Dictionary (CMUdict) [128], and another for BP: Aeiouadô [83]. By doing this, we can cover most of the phone productions that brazilian ESL learners are likely to make. Advanced students will tend to use more properly the English phones, whereas beginners will have a stronger accent, thus producing more phones from the BP phoneset. A brief description of each of these dictionaries is given below, before we go into further details of our interlingual set.

CMUdict [128] is a machine-readable pronunciation dictionary for AmE which has about 125,000 words and their transcriptions. It was designed primarily for speech applications, such as speech recognition and synthesis, and it has been widely tested in both Academia and industry.

Words in CMUdict are transcribed using ARPAbet, a phonetic transcription code developed by the Advanced Research Projects Agency in 1971. It represents each AmE phone with a distinct sequence of American Standard Code for Information Interchange (ASCII) characters. In the dictionary, there are in total 39 phones plus stress marks. The phone convention is described on Table 6.7.

As for Aeiouadô, as described in Section ??, its transcriptions are based on the dialect of São Paulo city and contains 39 different phones. The dictionary makes use of a hybrid approach for converting graphemes into phonemes, which employs both manual transcription rules and machine learning algorithms. Its phone convention is presented in Table 6.8.

In what concerns to our interlingual phoneset, we kept all CMUdict phones, since that is the target language's phoneset the learners are trying to achieve. Then we compared each

Table 6.7 CMUdict phone convention.

#	<b>CMU Phone</b>	<b>IPA Phone</b>	<b>Example</b>	<b>Transcription</b>
1	AA	[ɑ]	odd	AA D
2	AE	[æ]	at	AE T
3	AH	[ə]	hut	HH AH T
4	AO	[ɔ]	ought	AO T
5	AW	[aʊ]	cow	K AW
6	AY	[aɪ]	hide	HH AY D
7	B	[b]	be	B IY
8	CH	[tʃ]	cheese	CH IY Z
9	D	[d]	dee	D IY
10	DH	[ð]	thee	DH IY
11	EH	[ɛ]	Ed	EH D
12	ER	[ər]	hurt	HH ER T
13	EY	[aɪ]	ate	EY T
14	F	[f]	fee	F IY
15	G	[g]	green	G R IY N
16	HH	[h]	he	HH IY
17	IH	[ɪ]	it	IH T
18	IY	[i]	eat	IY T
19	JH	[dʒ]	gee	JH IY
20	K	[k]	key	K IY
21	L	[l]	lee	L IY
22	M	[m]	me	M IY
23	N	[n]	knee	N IY
24	NG	[ŋ]	ping	P IH NG
25	OW	[oʊ]	oat	OW T
26	OY	[ɔɪ]	toy	T OY
27	P	[p]	pee	P IY
28	R	[ɹ]	read	R IY D
29	S	[s]	sea	S IY
30	SH	[ʃ]	she	SH IY
31	T	[t]	tea	T IY
32	TH	[θ]	theta	TH EY T AH
33	UH	[ʊ]	hood	HH UH D
34	UW	[u]	two	T UW
35	V	[v]	vee	V IY
36	W	[w]	we	W IY
37	Y	[y]	yield	Y IY L D
38	Z	[z]	zee	Z IY
39	ZH	[ʒ]	seizure	S IY ZH ER

Table 6.8 Aeiouadô phone convention.

#	Aeiouadô Phone	IPA Phone	Example	Transcription
1	a	[a]	amor	a m o x
2	a~	[ã]	canto	k a~ t U
3	b	[b]	besta	b e s t @
4	d	[d]	da	d a
5	dZ	[dʒ]	dia	dZ i @
6	E	[ɛ]	é	E
7	e	[e]	dedo	d e d U
8	e~	[ẽ]	venda	v e~ d @
9	f	[f]	frio	f 4 i U
10	g	[g]	gula	g u l @
11	G	[ɣ]	carga	carga
12	i	[i]	aí	a i
13	I	[i]	come	k o~ m I
14	i~	[ĩ]	sim	s i~
15	J	[ɲ]	ganho	g a~ J U
16	j	[y]	pai	p a j
17	j~	[ŷ]	parem	p a 4 e~ j~
18	k	[k]	compra	k o~ p 4 @
19	l	[l]	lá	l a
20	L	[ʎ]	palha	p a L @
21	m	[m]	mãe	m a~ j~
22	n	[n]	não	n a~ w~
23	O	[ɔ]	pó	p O
24	o	[o]	gorro	g o x U
25	o~	[õ]	com	k o~
26	p	[p]	pessoa	p e s o @
27	s	[s]	susto	s u s t U
28	s	[ʃ]	chato	S a t U
29	t	[t]	tato	t a t U
30	tS	[tʃ]	noite	n o j t S I
31	u	[u]	durmo	d u G m U
32	U	[ʊ]	cúmulo	k u m u l U
33	u~	[ũ]	um	u~
34	v	[v]	vida	v i d @
35	w	[w]	aula	a w l @
36	w~	[ŵ]	canhão	k a~ J a~ w~
37	x	[x]	rato	x a t U
38	z	[z]	zebra	z e b 4 @
39	4	[r]	arara	a 4 a 4 @
40	@	[ə]	bola	b O 1 @

phone in CMUdict with those contained in Aieouadô, in order to check for missing phones and to analyze whether the overlapping ones really correspond to the same sound.

At a first glance, it seems that BP and AmE share a pool of 24 common phones, comprising fifteen consonantes [b, d, dʒ, f, g, k, l, m, n, p, s, ʃ, t, tʃ, v, z]; seven vowels [ɛ, i, ɪ, ɔ, u, ʊ, ə]; together with two glides [y, w]. However it is worth noticing that this first impression does not hold true.

Despite the fact that both dictionaries present IPA correspondences, such correspondences should not be taken for granted, without previous analysis. In theory, IPA is capable of describing any sound produced by the human vocal tract with exactness. Still IPA transcriptions are biased by the level of detail one wishes to express and by the assumptions of the transcriber. This is the case for some of these overlapping phones. Several of these consonants and vowels, although marked with the same IPA symbol by CMUdict and Aieouadô, in fact, can not be regarded as being the same sound, since they show a very different distribution in English and BP.

For instance, the production of /p, t, k/, in BP and AmE can be quite different. In English, such consonants are generally produced as [p, t, k]. However it is known that when they occur in certain contexts, for example, in word initial position or onset of a stressed syllable, they become aspirated; whence [p<sup>h</sup>, t<sup>h</sup>, k<sup>h</sup>] [74]. On the other hand, this process is not found in BP, where, disregard of the context, [p, t, k] show no relevant levels of aspiration [65].

For that reason, in order to properly estimate the HMM states for the interlingual phones, it is mandatory to create aspirated phone models for these consonants that are different from the non-aspirated ones. In our interlingual phoneset, we decided to keep the distinction between aspirated and non-aspirated phones. Therefore, according to our convention, the /p/ that occurs in a stressed syllable of an English word like “pie” is transcribed as [p<sup>h</sup>], while the /p/ of an unstressed syllable or a BP word is transcribed as [p].

Additionally, some vowels that are described by both dictionaries with the same IPA symbol are not exactly equal. Although English and BP both possess the vowels [ɪ, ə, ʊ], the distribution of these phones between both languages is fairly different. In AmE such vowels hold phonological status, i.e. they are phonemes, so that one could find minimal pairs differing only by [ɪ, ə, ʊ], such as “sheep” ['ʃɪp] vs. “ship” ['ʃɪp]; “cut” ['kʰət] vs. “cat” ['kʰæt]; and “pull” ['pʰʊl] vs. pool ['pʰuł]. As for BP, these vowels exist solely as part of a phonological process. When the tense vowels [i, a, u] occur in unstressed word-final position, they undergo a lenition process and are produced as the lax vowels [ɪ, ə, ʊ], respectively. Hence the BP [ɪ, ə, ʊ] have different formant values [41] and all the typical characteristics of lax vowels, that is, they are short, they have less energy and, consequently, less clear-cut formants [90].

With regard to the missing phones, there were 15 phones which are only present in the BP inventory, these include eight vowels [a, ã, e, ê, ï, o, õ, û]; five consonants [y, þ, ɿ, x, r]; and two nasal glides [ñ, ź].

All eight vowels were added to the interlingual phoneset, for they encompass negative transfer problems. For instance, [ã, ê, ï, õ, û] are related to vocalization of final nasals (*vide* 2.2.3) and [a, e, o] to vowel assimilation (*vide* 2.2.3).

The BP rhotic consonants [r, y, x], were merged onto the same sound, [h] owing to the fact they represent BP dialectal variants not relevant to L1-L2 interphonology.

Furthermore, the BP palatal consonants [þ, ɿ] were not considered in the final interlingual phoneset, since we could not find, in the literature, any negative transfer process in which they occur.

The nasal glides were excluded from the interlingual phoneset, instead we preferred to combine them with their accompanying vowels, in order to create nasal diphthongs, such as [ãɪ, ãʊ, êɪ, õɪ]. This decision was made in accordance with [33], which found that BP nasal diphthongs have a very particular behavior, with articulatory, acoustic and aerodynamic patterns different from the non-nasalized ones. We believe that a single HMM model for each diphthong will be able to better gauge this behavior.

The final phoneset, which we used as the basis for our ASR system, can be found in Table 6.9.

Table 6.9 Interlingual dictionary phone convention. BP word examples are shown in *italics*.

#	Interl. Phone	IPA Phone	Example	Transcription
1	a	[ə]	<i>da</i>	d a
2	aa	[ɑ]	odd	aa d
3	aaa	[a]	<i>dá</i>	d aaa
4	ae	[æ]	cat	k ae t
5	ah	[ə]	but	b ah t
6	ahw	[ɔʊ̯]	xxx	xx xx
7	am	[ã̯]	xxx	xx xx
8	ao	[ɔ̯]	for	f ao r
9	aow	[ɔʊ̯]	xxx	xx xx
10	aw	[aʊ̯]	cow	k aw
11	awm	[āʊ̯]	<i>não</i>	n awm
12	ay	[aɪ̯]	I	ay
13	aym	[aɪ̯]	mæe	m aym
14	b	[b̯]	boot	b uw tt
15	ch	[tʃ̯]	cheek	ch iy kk
16	d	[d̯]	do	d uw
17	dh	[ð̯]	that	dh ae tt
18	e	[e̯]	<i>eu</i>	e w
19	eh	[ɛ̯]	merry	m eh r iy
20	em	[ē̯]	<i>entendi</i>	em tt em jh i
21	ey	[eɪ̯]	April	ey pp r iy ll
22	eym	[eɪ̯]	<i>hein</i>	eym
23	f	[f̯]	fat	f ae tt
24	g	[g̯]	guy	g ay
25	hh	[h̯]	heat	hh iy tt
26	ih	[i̯]	bit	b ih tt
27	i	[i̯]	<i>comi</i>	k o m i
28	im	[i̯]	<i>sim</i>	s im
29	iy	[i̯]	eat	iy tt
30	jh	[dʒ̯]	judge	jh ah jh
31	k	[k̯]	cool	k uw ll
32	kk	[k̯]	cai	kk ay
33	l	[l̯]	lounge	l aa uh n jh
34	ll	[l̯]	fall	f ao ll
35	m	[m̯]	mother	m ah dh ah r
36	n	[n̯]	neat	n iy tt
37	ng	[ŋ̯]	king	k ih ng
38	o	[o̯]	sô	s o
39	om	[ō̯]	<i>conto</i>	kk om tt u
40	ow	[oʊ̯]	no	n ow
41	ohy	[ɔɪ̯]	boys	b oy z
42	oym	[ɔɪ̯]	<i>doações</i>	d o a s oym s
43	p	[p̯]	pity	p ih rd iy
44	pp	[p̯]	<i>pai</i>	pp ay
45	r	[r̯]	run	r ah n
46	rd	[r̯]	city	s ih rd iy
47	s	[s̯]	six	s ih k s
48	sh	[ʃ̯]	shoes	sh uw z
49	t	[t̯]	time	t ay m
50	th	[θ̯]	three	th r iy
51	tt	[t̯]	<i>tudo</i>	tt uw d u
52	u	[ʊ̯]	<i>como</i>	kk om m u
53	uh	[ʊ̯]	could	k uw d
54	um	[ũ̯]	<i>rum</i>	hh um
55	uw	[u̯]	wood	w uw d
56	v	[v̯]	van	v ae n
57	w	[w̯]	what	w aa tt
58	y	[y̯]	union	y uw n y ah n
59	z	[z̯]	<i>zoo</i>	z uw
60	zh	[ʒ̯]	leisure	l eh zh ah r

### 6.3.2 Speech Data

Many of the

### 6.3.3 HMM topology

### 6.3.4 Tree-Based State Tying

Data-driven approaches also show limitations. Since such approaches are generally based on the positive examples which occur in a corpus, rare or non-occurring phenomena are often poorly estimated or even neglected. This is the case for triphone HMM models.

Natural languages have, on average, 30 different phones. The language believed to have the smallest phonetic inventory is Rotokas (East Papuan, New Guinea), with 11 phones, and the one with largest is !Xóõ (Khoisan, Botswana/Namibia), with 160 [57]. English is usually assumed to have 37 to 41 phones, depending on the dialect.

In the CMUDict [128], 39 phones are used to describe the words of AmE. When it comes to triphones, in theory, this number might grow by three orders of magnitude, that is  $39^3$  or 59,319. It is true that due to phonotactic constraints many of the virtually possible triphones never take place in practice. For instance, the triphone sequence [ŋ-s+p] does not exist in English, although it is made of valid and existing monophones [ŋ], [s] and [p].

Kuperman et al. [67] examined the monophone, diphone and triphone frequencies in speech corpora for many languages. In what concerns to English, they analysed the Buck-eye Corpus and reported that it contains 29,804 different occurring triphones (or types), distributed among 431,000 tokens. Although the actual number of triphones for English is almost half the number of possible permutations, it is still a huge number of triphones to model over a corpus. Therefore, in order to build an ASR system, one always has to deal with data scarcity and try to overcome its limitations.

Within HMM ASR, tree-based state tying is a technique to improve the modelling of rare triphones and allow the estimation of non-occurring ones. It was initially proposed by Young et al. [133] and since then has become a standard procedure in HMM ASR systems. The aim of tree-based state tying is to maintain the balance between the model complexity and the available training data by tying the HMM states of acoustically similar triphones.

In contrast to the majority of methods in ASR, tree-based state tying is carried out in a top-down, knowledge-based way. A specialist (generally a speech scientist, phonologist or phoneticist) uses his/her knowledge to build a phonetic decision tree which will organize the phones into sets with similar acoustic parameters.

In most cases, the criteria used to organize phones into a decision tree are the so-called natural classes. Natural classes were proposed within the generative phonology framework. In this framework, phones and phonemes are no longer considered the basic units of analysis, instead it is assumed that they can be broken down into smaller components, which describe aspects of articulation and perception, such as [+nasal], [-continuant], [+strident], etc. For instance, the phone [s] would not be represented in generative phonology as a single phonetic symbol [s], but as a bundle of distinctive features [62]:

$$[a] \rightarrow \left[ \begin{array}{l} +\text{consonantal} \\ -\text{syllabic} \\ -\text{sonorant} \\ +\text{continuant} \\ +\text{anterior} \\ +\text{coronal} \\ +\text{strident} \\ -\text{voiced} \end{array} \right]$$

Distinctive features were probably the most important contribution of generative phonology. They became popular as a model specially because they were able of simplify phonological processes by grouping segments. Consider the case of final-obstruent devoicing. In many pronunciations of Standard German, voiced obstruent consonants become devoiced when they occur in word final position. This is the case for a large number of german nouns which make their plural by the addition of the suffix {-e}. Table 6.10 contains a few examples extracted from Grijzenhout [54].

As one can observe at the provided examples, the plural forms contain only voiced obstruents whilst the singular forms contain their devoiced counterparts. To express this phonological process within structural phonology, one would have to introduce at least four rules:

$$[b] \rightarrow [p]/_{-\#}$$

$$[d] \rightarrow [t]/_{-\#}$$

$$[g] \rightarrow [k]/_{-\#}$$

$$[d] \rightarrow [s]/_{-\#}$$

Whereas using distinctive features, all cases of german final-obstruent devoicing can be explained by a single rule:

$$\begin{bmatrix} +\text{consonantal} \\ -\text{syllabic} \\ -\text{sonorant} \\ +\text{voiced} \end{bmatrix} \rightarrow [-\text{voiced}] / \_ \#$$

The main benefits of distinctive features lie in their capability of making generalizations, that is to say of grouping phones together in a meaningful way. Phonologists have long known that sounds that share the same manner, place of articulation, or voicing level behave similarly. Distinctive features provide a way to express this in an elegant way, by establishing that feature matrices which are not fully specified do form a natural class. Therefore a distinctive features matrix

$$\begin{bmatrix} +\text{consonantal} \\ -\text{syllabic} \end{bmatrix}$$

represents all consonants, whereas

$$\begin{bmatrix} +\text{consonantal} \\ -\text{syllabic} \\ +\text{continuant} \\ +\text{voiced} \end{bmatrix}$$

represents all voiced fricatives, etc. For developing the phonetic decision tree, the specialist employs his/her knowledge of phonetics and phonology in order to define questions about the phonetic environment of a given phone. This phonetic environment is defined by using natural classes, such that an example question would be “is the phone preceded by an obstruent?” or “is there a fricative after the phone?” Technically, the tree is a binary one, i.e a connected acyclic graph such that the degree of each vertex is no more than three. Each internal node of the tree represents a question about the phonetic context a triphone, each branch represents the answer in a yes-no form and leaf nodes define HMM states. Once the tree is built, its structure is used to decide how HMM states will be tied among triphone models. Figure 6.2 presents an example of a phonetic decision tree.

The tree input is each triphone being analysed.

The clustering procedures begins by placing all observations in a single root node. All questions are analised and the one which maximise the likelihood of a single diagonal covariance Gaussian is chosen, then the node is split and child nodes generated. The splitting

goes on until it falls below a threshold. Commonly, a minimum threshold is also set, the usual approach is to consider the frequency of occurrence of the phones in a corpus. That is, if a given phone  $p$  has  $m$  samples and the minimum threshold is  $n$ , such that  $m > n$ , the phone  $p$  is mapped onto a more general and robust phone. Consider that the triphone [i-ŋ+g] appeared 30 times on corpus, and the minimum threshold for splitting the node was 45, then [i-ŋ+g] would be modeled into a more general phone, say, for instance, [i-ŋ+g] or [i-ŋ].

In the example shown in Figure 6.2, the first node (the root of the tree) checks if the left part of the triphone contains a nasal consonant. If positive, the tree examines the right side of the triphone, by questioning whether it is a liquid consonant. If negative, a leaf is reached and the HMM state to be tied is outputted, e.g. “tie the 1<sup>st</sup> emitting HMM state of the analysed triphone A to [Nasal-A+\*]”. Figure 6.3 summarizes the state tying procedure.

For building the phonetic decision tree for Listener, we based ourselves on the distinctive feature set proposed by Jensen [62], which follows the main guidelines of generative phonology by dividing the features into four central classes: i) major class features, ii) manner features; iii) place features; and iv) laryngeal features.

In practice, classes work as following. Major class features distinguish among the most general classes of sound, i.e. vowels, consonants and glides (or semi-vowels). Manner features determine how sounds are articulated, that is do they consist of a stop, a nasal, a fricative, a liquid, a trill or a flap? Place features describe where in the mouth sounds are produced, whether in the labial region, the alveolar region, the velar region, etc. Finally, laryngeal features represent the glottal states of sounds, their basic purpose is to differ voiced sounds from unvoiced ones.

Jensen [62] proposes a set of 17 features to describe most of the world languages.

1. *Major class features*: syllabic, consonantal and sonorant;
2. *Manner features*: continuant, nasal, lateral, strident and delayed release;
3. *Place features*: anterior, coronal, distributed, high, low, back, round and Advanced Tongue Root (ATR).
4. *Laryngeal features*: voice and Heightened Subglottal Pressure (HSP);

The somewhat reduced set for place features is capable of representing a large amount of places of articulation since features which are regularly restricted to vowels (such as high, low, back and round) are also shared with consonants. Besides, once features have a binary nature, therefore, in theory, a number  $n$  of features is able to distinguish up to  $2^n$  phones. In Figure 6.4, a comparison is shown between place features and their corresponding regions of articulation. For a full explanation of each feature, please see Jensen [62].

According to this set of distinctive features, we can arrange the entire phonetic inventory of AmE as in Table 6.11 and that of BP as in Table 6.12.

Table 6.10 Examples of final-obstruent devoicing in German.

<b>Plural form</b>	<b>Singular form</b>	<b>Gloss</b>
Hun[d]e	Hun[t]	dog
Die[b]e	Die[p]	thief
Ber[g]e	Ber[k]	mountain
Mäu[z]	Mau[s]	mouse

Fig. 6.2 Phonetic decision tree for HMM state tying [133].

Fig. 6.3 Tied-state HMM system build procedure [133].

Fig. 6.4 Distinctive features for places of places of articulation [62].

Table 6.11 Distinctive features chart for AmE phones [62].

Table 6.12 Distinctive features chart for BP phones [62].

Table 6.13 Distinctive features chart for Listener phones.

## 6.4 Building the Pronunciation Model

To the best of our knowledge, no previous research has addressed the problem of generating brazilian-accented transcriptions for ASR purposes or has described the mispronunciations phenomena from a computational perspective. Therefore we had to develop our own pronunciation model<sup>3</sup>. There are basically three main approaches we could use to build such model: rule-based methods (XXX CITATION), machine learning methods (XXX CITATION) and hybrid ones (XXX CITATION). For achieving good performance through machine learning or hybrid approaches, one necessary needs a large annotated corpus. That is not the case though. The only Brazilian-accented transcribed corpus we have access is the Listener Corpus, but we carried out some pilot experiments that showed it was not robust enough for the task.

That being so, we decided to make use of a rule-based approach for building the pronunciation model. We reviewed all papers described in ??, that deal with the mispronunciation of English phones by brazilians, in order to find interlingual allophones, such as the English [θ] usually becomes [t], [tʰ] or [f] in beginners' speech [100].

By knowing the allophones, we developed rules in order to generate the mispronunciations in the dictionary. The mispronunciation contexts specified through rewrite rules are thus a contribution of this thesis. It is worth mentioning that, for creating the rules, we took into account the frequency of occurrence of each mispronunciation (when this information was available on the papers). On that account mispronunciations which were reported as being very rare or with no significant probability were excluded.

We implemented the rules through a Python script which made a large usage of the Regular Expression (regex) library. As one familiar with regex knows, regex rules might get too clumsy and hard to understand. For instance, one of the rules in our Python script is:

Listing 6.1 Example of a fully-specified regex rule.

```
re . sub( '( ih | iy ) (m|n|ng) (#|[ ^aeiouyw ])' , r ' i~ \3 ' , pron )
```

To one not familiar with how the dictionary is structured, this rule, in our opinion, could be somewhat meaningless. Therefore, to render the text easier to read, we are going to describe the rules in a pseudocode with a rather flexible notation<sup>4</sup>. Whenever possible, we

---

<sup>3</sup>“Pronunciation models” are also called “pronunciation dictionaries”. In this thesis, we are going to use both terms interchangeably, without any distinction.

<sup>4</sup>For those who are interested in checking the code itself, it can be downloaded on the site of the project: <http://nile.icmc.usp.br/listener>

try to simplify the rules' contexts, so anyone not acquainted with the dictionary format or the regex syntax might understand their content.

## 6.5 Building the Grammars and the Language Model

## 6.6 Results

# **Chapter 7**

## **Conclusions**

### **7.1 Overall Conclusions**

Resultados em reconhecimento de fala, em muitas vezes, são exploratórios. As línguas são sistemas dinâmicos, variando dado o espaço, o tempo, os grupos sociais, as situações comunicativas e o próprio falante. Tendo em vista que cada língua possui aspectos particulares, não se pode assegurar, no PLN, que métodos já testados para outros idiomas, sejam diretamente transplantados para o PB e tenham funcionamento e desempenho similar. Esta dissertação busca propor um método de elaboração de um sistema de reconhecimento de pronúncia para aprendizes de inglês, falantes nativos do PB. Um sistema desse tipo ainda não foi elaborado para o PB e os resultados, portanto, são tentativos. Como se discutiu, tal sistema é de utilidade, tendo em vista a baixa proficiência em inglês dos brasileiros, demonstrada recentemente nos índices da GlobalEnglish (2012) e da Education First (2013).

A literatura pertinente da área foi revisada e métodos que se mostraram promissores foram selecionados para integrar o projeto. Pretende-se elaborador um reconhecedor de pronúncia que seja capaz de tratar nove erros de pronúncia, provendo feedback ao usuário sobre a qualidade de sua pronúncia. Os erros foram selecionados com base nos trabalhos de Zimmer (2004), Godoy (2005), Zimmer et al. (2009) e Cristófaro-Silva (2012), assumindo-se, como pronúncia padrão, o General American (GA). A abordagem de interlíngua foi selecionada para a arquitetura do reconhecedor. O modelo acústico é, assim, alimentado com dados de fala tanto de nativos, quanto de não-nativos, aprendizes de inglês. No caso, para os dados de nativos, será utilizado o TIMIT; e para os de não-nativos, o COBAI e um corpus de leitura de sentenças foneticamente balanceadas, ainda a ser compilado. As sentenças serão extraídas de um corpus de aprendizes, o COMAprend. Um script em Python será utilizado para realizar a conversão grafema-fone das sentenças, tendo por base a pronúncia canônica das palavras registrada no CMU Pronouncing Dictionary. Na abordagem interlingual, também, o modelo

de pronúncia deve ser alimentado com as variantes de pronúncia do aprendiz, de modo a compor os chamados dicionários multipronúncia. Para isso, pretende-se utilizar o CMU Pronouncing Dictionary como base e adicionar as hipóteses de pronúncia dos aprendizes por meio de regras transformacionais. O modelo de língua será constituído por trigramas e gerado a partir da Simple English Wikipedia., de modo a apresentar um sintaxe próxima à produção do aprendiz.

Um protótipo foi elaborado de modo a avaliar a viabilidade do método ora proposto. O cronograma de execução do projeto está dentro do prazo, e as bibliotecas e os softwares necessários para sua execução (HTK, Julius, Adintool, Audacity, Praat, SoX) já foram testados na elaboração do protótipo. Uma porção do COBAI (~3h40min) foi segmentada, alinhada e analisada e utilizada para estimar o modelo acústico. Tendo em vista o grande de número de arquivos do COBAI que teve de ser desconsiderado (apenas ~1h30min do que foi segmentado pode ser utilizado), optou-se por se realizar a coleta de um corpus de leitura de sentenças foneticamente balanceadas especificamente para o desenvolvimento do projeto. De modo a simular, no protótipo, os dados que se obterão com esse corpus, um corpus de erros induzidos (~2h20min) foi gravado, segmentado, transscrito e analisado. Tal corpus também foi utilizado na estimação do modelo acústico do protótipo, juntamente com os dados do COBAI. O método de coleta e anotação do corpus real, de leitura de sentenças foneticamente balanceadas, será definido em visita técnica à Universidade de Coimbra, sob supervisão da Profa. Sara Candeias, no período de 28 de janeiro a 28 de fevereiro de 2014. O dicionário-base do modelo de pronúncia, o CMU Pronouncing Dictionary, já foi testado no protótipo, tendo sido adicionadas variantes de pronúncia para um dos erros selecionados: a simplificação silábica. Foi observado que houve um grande aumento no número de entradas no modelo de pronúncia com a adição das variantes de pronúncia: o dicionário cresceu de 1.855 palavras para 7.597. Sendo assim, é possível, ao se coligir todas as regras para os nove tipos de erros, que o dicionário cresça fortemente, tornando o reconhecimento confuso, bem como consumindo tempo e recursos computacionais. Caso isso ocorra, uma solução possível seria criar dicionários específicos para cada tipo de erro, ou solicitar a um especialista que cerceie o dicionário, eliminando as variantes que ocorrem com menor frequência. Os resultados iniciais obtidos com o protótipo no reconhecimento mostraram-se promissores.

O conteúdo do projeto tem sido publicado na web[31] de modo a dar-lhe visibilidade e angariar possíveis colaboradores. Pretende-se, também, ao final do projeto, disponibilizar um pequeno sistema de treino de pronúncia, como a prova de conceito para uso do reconhecedor de pronúncia. Uma interface vem sendo desenvolvida para disponibilizar o protótipo na web, as Figura 19 e Figura 20 trazem algumas telas de exemplo dessa interface.

[pic]

Figura 19: Interface do protótipo na web - visão geral do site e tela de captura do áudio com espectro de frequência.

| [i] | [ii] | |[pic] |[pic] |

Figura 20: Interface do protótipo na web - [i] palavra reconhecida com transcrição em formato IPA e em alfabeto adaptado; [ii] tela com texto de feedback sobre a pronúncia do aprendiz, após ele reiterar no erro.

Há também a possibilidade de se realizar visita técnica ao Laboratório de Processamento de Sinais (LaPS) da Universidade Federal do Pará (UFPA), na época de avaliação dos resultados finais da dissertação, sob supervisão do Prof. Aldebaro Klautau, co-orientador da pesquisa.

## 7.2 Limitations

## 7.3 Further Work



# References

- [1] Abushariah, M. A. M., Ainon, R. N., Zainuddin, R., Elshafei, M., and Khalifa, O. O. (2012). Phonetically rich and balanced text and speech corpora for Arabic language. *Language Resources and Evaluation*, 46(4):601–634.
- [2] Ahmed, F., Luca, E. W. D., and Nürnberg, A. (2009). Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness. *Polibits*, pages 39–48.
- [3] Alves, I. (2001). *Neologismo: Criação lexical*. Princípios (São Paulo). Editora Atica.
- [4] Andrade, G., Teixeira, F., Xavier, C., Oliveira, R., Rocha, L., and Evsukoff, A. (2012). HASCH: High Performance Automatic Spell Checker for Portuguese Texts from the Web. *Proceedings of the International Conference on Computational Science*, 9(0):403 – 411.
- [5] Arora, K., Arora, S., Verma, K., and Agrawal, S. S. (2004). Automatic extraction of phonetically rich sentences from large text corpus of Indian languages. *INTERSPEECH*.
- [6] Avanço, L., Duran, M., and Nunes, M. G. V. (2014). Towards a Phonetic Brazilian Portuguese Spell Checker. In *Proceedings of ToRPorEsp Workshop PROPOR 2014*, pages 24–31, São Carlos, Brazil.
- [7] Badin, P., Ben Youssef, A., Bailly, G., Elisei, F., and Hueber, T. (2010). Visual articulatory feedback for phonetic correction in second language learning. *Actes de SLATE*, pages 1–10.
- [8] Benzeghiba, M., Mori, R. D., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., and Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10-11):763–786. Intrinsic Speech Variations.
- [9] Bergsma, S., McNamee, P., Bagdouri, M., Fink, C., and Wilson, T. (2012). Language identification for creating language-specific twitter collections. In *Proceedings of the Second Workshop on Language in Social Media*, LSM ’12, pages 65–74, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [10] Bilcu, E. B. and Astola, J. (2006). A hybrid neural network for language identification from text. In *Machine Learning for Signal Processing, 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on*, pages 253–258. IEEE.
- [11] Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media, Inc.

- [12] Bisol, L. (1989). Vowel harmony: a variable rule in brazilian portuguese. *Language Variation and change*, 1:185–198.
- [13] Bisol, L. (2005). *Introdução a estudos de fonologia do português brasileiro*. Edipucrs.
- [14] Brasil (2009). *Acordo ortográfico da língua portuguesa, de 14, 15 e 16 de dezembro de 1990*. Diário do Congresso Nacional da República Federativa do Brasil, Poder Executivo, Brasília, Brazil.
- [15] Brill, E. and Moore, R. C. (2000). An Improved Error Model for Noisy Channel Spelling Correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 286–293.
- [16] Brown Jr, W., Morris, R. J., Hollien, H., and Howell, E. (1991). Speaking fundamental frequency characteristics as a function of age and professional singing. *Journal of Voice*, 5(4):310–315.
- [17] Bussmann, H., Trauth, G., Kazzazi, K., and Bussmann, H. (1996). *Routledge dictionary of language and linguistics / Hadumod Bussmann ; translated and edited by Gregory Trauth and Kerstin Kazzazi*. Routledge, London ; New York :.
- [18] Cagliari, L. C., Laplantine, F., Editora, M. F., Brait, B., Lévy, P., Mattos, R. V., Bosi, A., Hall, E. T., da Graça Nicoletti, M., and Elias, V. M. (2002). Análise fonológica. *São Paulo*.
- [19] Câmara, J. M. (1970). *Estrutura da língua portuguesa*. Editôra Vozes.
- [20] Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.
- [21] Collischonn, G. (2004). Epêntese vocalica e restrições de acento no português do sul do brasil. *Signum: Estudos da Linguagem*, 7(1):61–78.
- [22] Coppin, B. (2010). Inteligência artificial. *Rio de Janeiro: LTC*.
- [23] Cristófaro Silva, T. et al. (2012). Revisitando a palatalização no português brasileiro. *Revista de Estudos da Linguagem*, pages 59–89.
- [24] Crystal, D. (2011). *Dictionary of linguistics and phonetics*, volume 30. John Wiley & Sons.
- [25] Cucerzan, S. and Brill, E. (2004). Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users . In *Proceedings of the Conference on Empirical Methods on Natural Language Processing EMNLP 2004*, pages 293–300, Barcelona, Spain.
- [26] Damerau, F. J. (1964). A Technique for Computer Detection and Correction of Spelling Errors. *Communications of ACM*, 7(3):171–176.
- [27] Damper, R. I., Marchand, Y., Adamson, M., and Gustafson, K. (1998). Comparative evaluation of letter-to-sound conversion techniques for english text-to-speech synthesis. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*.

- [28] Davenport, M., Davenport, M., and Hannahs, S. (2010). *Introducing phonetics and phonology*. Routledge.
- [29] Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366.
- [30] De Clercq, O., Schulz, S., Desmet, B., Lefever, E., and Hoste, V. (2013). Normalization of Dutch User-Generated Content. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 179–188.
- [31] de Medeiros, B. R. (2007). Vogais nasais do português brasileiro: reflexões preliminares de uma revisita. *Revista Letras*, 72.
- [32] Delatorre, F. and Koerich, R. (2005). Production of epenthesis in ed-endings by brazilian efl learners. *Proceedings of the II Academic Forum*, page 8.
- [33] Demasi, R. (2010). *A ditongação nasal no Português Brasileiro: uma análise acústico-aerodinâmica da fala*. University of São Paulo (Unpublished Master's Thesis), São Paulo, SP.
- [34] developers, S. (2014). *Scrapy – A fast high-level screen scraping and web crawling framework*. Scrapinghub. <http://scrapy.org/>.
- [35] Duan, H. and Hsu, B. P. (2011). Online Spelling Correction for Query Completion. In *Proceedings of the 20th International Conference on World Wide Web, WWW '11*, pages 117–126, NY, USA. ACM.
- [36] Duran, M., Avanço, L., Aluísio, S., Pardo, T., and Nunes, M. G. V. (2014). Some Issues on the Normalization of a Corpus of Products Reviews in Portuguese. In *Proceedings of the 9th Web as Corpus Workshop WaC-9*, pages 22–28, Gothenburg, Sweden.
- [37] EducationFirst (2011). *EF English Proficiency Index 2011*. Education First Ltd.
- [38] EducationFirst (2012). *EF English Proficiency Index 2012*. Education First Ltd.
- [39] EducationFirst (2013). *EF English Proficiency Index 2013*. Education First Ltd.
- [40] EducationFirst (2014). *EF English Proficiency Index 2014*. Education First Ltd.
- [41] Fails, W. C. and Clegg, J. H. (1992). A spectrographic analysis of portuguese stressed and unstressed vowels. *Romance Linguistics: The Portuguese Context*. Bergin and Garvey, Westport, pages 31–42.
- [42] Fitch, W. T. and Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 106(3):1511–1522.
- [Fossati and Di Eugenio] Fossati, D. and Di Eugenio, B. I saw TREE trees in the park: How to Correct Real-Word Spelling Mistakes. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC 2008*.

- [44] Fossati, D. and Di Eugenio, B. (2007). A Mixed Trigrams Approach for Context Sensitive Spell Checking. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 4394 of *Lecture Notes in Computer Science*, pages 623–633. Springer.
- [45] Freitas, C., Rocha, P., and Bick, E. (2008). Floresta sintactica: Bigger, thicker and easier. In *Computational Processing of the Portuguese Language*, pages 216–219. Springer.
- [46] Furui, S. (2001). *Digital Speech Processing, Synthesis, and Recognition*. Marcel Dekker, New York, USA, 2<sup>a</sup> edition.
- [47] Ganapathiraju, A., Hamaker, J., Picone, J., Ordowski, M., and Doddington, G. (2001). Syllable-based large vocabulary continuous speech recognition. *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, 9:358–366.
- [48] Garofolo, J., Graff, D., Paul, D., and Pallett, D. (1993a). *CSR-I (WSJ0) Complete LDC93S6A. Web Download*. Linguistic Data Consortium, Philadelphia.
- [49] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1993b). DARPA TIMIT acoustic phonetic continuous speech corpus cdrom.
- [50] Girelli, C. (1990). *Brazilian Portuguese Syllable Structure*. UMI.
- [51] GlobalEnglish (2013). *The 2013 Business English Index & Globalization of English Report*. Pearson Always Learning, Pearson.
- [52] Godwin-Jones, R. (2009). Emerging technologies: Speech tools and technologies. *Language Learning and Technology*, 13-3:4–11.
- [53] Gordon, R. G. and Grimes, B. F. (2005). *Ethnologue: Languages of the world*, volume 15. SIL International Dallas, TX.
- [54] Grijzenhout, J. (2000). Voicing and devoicing in english, german, and dutch; evidence for domain-specific identity constraints. *Theory des Lexikons; Arbeiten des Sonderforschungsbereichs*, 282:1–22.
- [55] Han, B., Cook, P., and Baldwin, T. (2013). Lexical Normalization for Social Media Text. *ACM Trans. Intelligent System Technology*, 4(1):5:1–5:27.
- [56] Hartmann, N., Avanço, L., Balage, P., Duran, M., Nunes, M. G. V., Pardo, T., and Aluísio, S. (2014). A Large Corpus of Product Reviews in Portuguese: Tackling Out-Of-Vocabulary Words. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation LREC’14*, pages 3866–3871.
- [57] Hayes, B. (2011). *Introductory Phonology*. Blackwell Textbooks in Linguistics. Wiley.
- [58] Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752.
- [59] Hirst, G. (2008). An evaluation of the contextual spelling checker of Microsoft Office Word 2007.

- [60] Huang, X., Acero, A., and Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.
- [61] Islam, A. and Inkpen, D. (2009). Real-word spelling correction using Google web 1tn-gram data set. In *In ACM International Conference on Information and Knowledge Management CIKM 2009*, pages 1689–1692.
- [62] Jensen, J. (2004). *Principles of Generative Phonology: An introduction*. Current Issues in Linguistic Theory. John Benjamins Publishing Company.
- [63] Jones, E., Oliphant, T., and Peterson, P. (2014). SciPy: Open source scientific tools for Python. <http://www.scipy.org/>.
- [64] Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.
- [65] Klein, S. (1999). *Estudo do VOT no Português Brasileiro*. Universidade Federal de Santa Catarina (Unpublished Master's Thesis), Florianópolis, SC.
- [66] Kumar, A., Dua, M., and Choudhary, T. (2014). Article: Continuous Hindi speech recognition using monophone based acoustic modeling. *IJCA Proceedings on International Conference on Advances in Computer Engineering and Applications*, ICACEA(1):15–19.
- [67] Kuperman, V., Ernestus, M., and Baayen, H. (2008). Frequency distributions of uniphones, diphones, and triphones in spontaneous speech. *The Journal of the Acoustical Society of America*, 124(6):3897–3908.
- [68] Ladefoged, P. (1995). *Elements of acoustic phonetics*. University of Chicago Press, Chicago.
- [69] Lakoff, R. (1973). Language and Woman's Place. *Language in Society*, 2(1).
- [70] Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):pp. 159–174.
- [71] Lei, X., yuh Hwang, M., and Ostendorf, M. (2005). Incorporating tone-related MLP posteriors in the feature representation for Mandarin ASR. In *In Proc. Eur. Conf. Speech Communication Technology*, pages 2981–2984.
- [72] Lenneberg, E. (1967). *Biological foundations of language*. John Wiley and Sons, Biological foundations of language.
- [73] Lewis, M. Gary, F. and Charles, D. (2013). *Ethnologue: Languages of the World, Seventeenth edition*. Seventeenth edition. SIL International, Dallas, Texas.
- [74] Lisker, L. (1985). How is the aspiration of english /p,t,k/ “predictable”? *Haskins Laboratories, Status report on speech research SR-84*, pages 141–144.
- [75] Long, M. (1983). Does second language instruction make a difference? a review of the research. *TESOL Quarterly*, 17:359–382.

- [76] Lyster, R. and Saito, K. (2010). Oral feedback in classroom sla. *Studies in Second Language Acquisition*, 32:265–302.
- [77] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- [78] Martins, B. and Silva, M. J. (2004). Spelling Correction for Search Engine Queries. In *Proceedings of the 4th International Conference EstAL 2004 – España for Natural Language Processing*, volume 3230 of *Lecture Notes in Computer Science*, pages 372–383.
- [79] Mays, E., Damerau, F. J., and Mercer, R. L. (1991). Context based spelling correction. *Information Processing & Management*, 27(5):517–522.
- [80] McLoughlin, I. (2009). *Applied Speech and Audio Processing – With Matlab Examples*. Cambridge University Press, Cambridge.
- [81] Medialab (2013). Wikipedia extractor. [http://medialab.di.unipi.it/wiki/Wikipedia\\_Extractor](http://medialab.di.unipi.it/wiki/Wikipedia_Extractor).
- [82] Mendes, A. P., Costa, A. N. d., Martins, A. D., Fernandes, A. F. O., Vicente, S. M. D. d. R., and Freitas, T. C. S. (2012). Contributos para a construção de um texto foneticamente equilibrado para o Português-Europeu. *Revista CEFAC*, 14:910–917.
- [83] Mendonça, G. and Aluísio, S. (2014). Using a hybrid approach to build a pronunciation dictionary for Brazilian Portuguese. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association INTERSPEECH 2014*, Singapore.
- [84] Mendonça, G., Avanço, L., Duran, M., Fonseca, E., Volpe-Nunes, M., and Aluísio, S. (2015). Evaluating phonetic spellers for user-generated content in brazilian portuguese. *Proceedings of IJCAI 2015 – International Joint Conference on Artificial Intelligence*.
- [85] Mendonça, G., Candeias, S., Perdigão, F., Shulby, C., Tonazzzo, R., Klautau, A., and Aluísio, S. (2014). A method for the extraction of phonetically-rich triphone sentences. *Proceedings of ITS 2014 – International Telecommunications Symposium*.
- [86] Neri, A., Mich, O., Gerosa, M., and Giuliani, D. (2008). The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, 21:393–408.
- [87] Neto, N., Patrick, C., Klautau, A., and Trancoso, I. (2011). Free tools and resources for Brazilian Portuguese speech recognition. *Journal of the Brazilian Computer Society*, 17(1):53–68.
- [88] Neves, M. H. M. (1999). Gramática do português falado. vol. vii: Novos estudos. são paulo.
- [89] Nicodem, M., Seara, I., Seara, R., Anjos, D., and Seara-Jr, R. (2007). Seleção automática de corpus de texto para sistemas de síntese de fala. *XXV Simpósio Brasileiro de Telecomunicações - SBrT 2007*.
- [90] Nobre, M. A. and Ingemann, F. (1987). Oral vowel reduction in brazilian portuguese. In *Honor of Ilse Lehiste. Foris, Dordrecht*, pages 195–206.

- [91] Oliveira, C., Moutinho, L. C., and Teixeira, A. J. S. (2005). On european portuguese automatic syllabification. In *Proceedings of the Interspeech 2005*, pages 2933–2936. ISCA.
- [92] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011a). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- [93] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011b). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- [94] Philips, L. (2000). The double metaphone search algorithm. *C/C++ Users Journal*, 18(6).
- [95] Pierrehumbert, J. B. P., Beckman, M. E., and Ladd, D. R. (2000). Conceptual foundations of phonology as a laboratory science. In *Phonological knowledge: Conceptual and empirical issues*, pages 273–304. Oxford University Press.
- [96] Polyakova, T. and Bonafonte, A. (2006). Learning from errors in grapheme-to-phoneme conversion. In *INTERSPEECH*.
- [97] Quicoli, A. (1990). Harmony, lowering and nasalization in brazilian portuguese. *Língua*, 80:295–331.
- [98] Rabiner, L. and Schafer, R. (2007). *Introduction to Digital Speech Processing*, volume 1, pages 1–194.
- [99] Reis, J. and Hazan, V. (2011). Speechant: a vowel notation system to teach english pronunciation. *ELTJournal*, 66:156–165.
- [100] Reis, M. (2006). *The perception and production of English interdental fricatives by brazilian EFL learners*. Universidade Federal de Santa Catarina (Unpublished Master's Thesis), Florianópolis, SC.
- [101] Robjohns, H. (2010). A brief history of microphones. <http://microphone-data.com/media/filestore/articles/History-10.pdf>. Last retrieved 11-21-2014.
- [102] Rocha, W. and Neto, N. (2013). Implementação de um separador silábico gratuito baseado em regras linguísticas para o português brasileiro. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pages 108–115.
- [103] Rusell, R. C. (1918). US Patent 1261167 issued 1918-04-02.
- [104] Sedgewick, R. and Flajolet, P. (2013). *An introduction to the analysis of algorithms*. Addison-Wesley-Longman.
- [105] Shen, J.-L., Wang, H.-M., Lyu, R.-Y., and Lee, L.-S. (1994). Incremental speaker adaptation using phonetically balanced training sentences for Mandarin syllable recognition based on segmental probability models. In *ICSLP*. ISCA.

- [106] Shrawankar, U. and Thakare, V. M. (2010). Techniques for feature extraction in speech recognition system : A comparative study. *International Journal Of Computer Applications In Engineering, Technology and Sciences (IJCAETS)*, pages 412–418.
- [107] Silva, T. C. (2005). *Fonética e fonologia do português: roteiro de estudos e guia de exercícios*. Contexto.
- [108] Sonmez, C. and Ozgur, A. (2014). A Graph-based Approach for Contextual Text Normalization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing EMNLP 2014*, pages 313 – 324.
- [109] Steigner, J. and Schröder, M. (2007). Cross-language phonemisation in German text-to-speech synthesis. In *INTERSPEECH 2007*, pages 1913–1916. ISCA.
- [110] Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer.
- [111] Steriade, D. (2000). Paradigm uniformity and the phonetics-phonology boundary. *Papers in laboratory phonology V: Acquisition and the lexicon*, 3:13–334.
- [112] Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.
- [113] Strik, H., Cornillie, F., Colpaert, J., Doremalen, J., and Cucchiari, C. (2009). Developing a call system for practicing oral proficiency: How to design for speech technology, pedagogy and learners. pages 123–125., Warwickshire, England.
- [114] Teixeira, A., Oliveira, C., and Moutinho, L. (2006). On the use of machine learning and syllable information in european portuguese grapheme-phone conversion. In *Computational Processing of the Portuguese Language*, pages 212–215. Springer.
- [115] Thangarajan, R., Natarajan, A. M., and Selvam, M. (2008). Word and triphone based approaches in continuous speech recognition for Tamil language. *WSEAS Trans. Sig. Proc.*, 4(3):76–85.
- [116] Toutanova, K. and Moore, R. C. (2002). Pronunciation Modeling for Improved Spelling Correction. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 144–151.
- [117] Trieschnigg, D., Hiemstra, D., Theune, M., de Jong, F., and Meder, T. (2012). An exploration of language identification techniques for the dutch folktale database. In Osenova, P., Piperidis, S., Slavcheva, M., and Vertan, C., editors, *Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage, LREC 2012*, pages 47–51, Istanbul, Turkey. LREC organization.
- [118] Tsubota, Y., Dantsuji, M., and Kawahara, T. (2004). An english pronunciation learning system for japanese students based on diagnosis of critical pronunciation errors. *ReCALL*, 16:173–180.
- [119] Umesh, S., Cohen, L., and Nelson, D. (1999). Fitting the mel scale. *Proc. ICASSP 1999*, pages 217–220.

- [120] United-Nations (2014). *2014 Human Development Report – Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience*. United Nations Development Programme (UNDP).
- [121] Uraga, E. and Gamboa, C. (2004). VOXMEX speech database: Design of a phonetically balanced corpus. In *LREC*. European Language Resources Association.
- [122] van Berkel, B. and Smedt, K. D. (1988). Triphone Analysis: A Combined Method for the Correction of Orthographical and Typographical Errors. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 77–83, Austin, Texas, USA.
- [123] Vasilévski, V. (2012). Phonologic patterns of brazilian portuguese: a grapheme to phoneme converter based study. In *Proceedings of the Workshop on Computational Models of Language Acquisition and Loss*, pages 51–60, Avignon, France. Association for Computational Linguistics.
- [124] Veiga, A., Candeias, S., and Perdigão, F. (2013). Developing a hybrid grapheme to phoneme converter for european portuguese. *Conf. on Telecommunications - ConfTele*, 1:297–300.
- [125] Wang, X., Hueber, T., and Badin, P. (2014). On the use of an articulatory talking head for second language pronunciation training: the case of Chinese learners of French. In *Proceedings of the 10th International Seminar on Speech Production, ISSP10*, page 4, Koeln, Allemagne.
- [126] Watanabe, M. and Rose, R. (2012). Pausology and hesitation phenomena in second language acquisition. *The Routledge Encyclopedia of Second Language Acquisition*, pages 480–483.
- [127] Weide, H. (1998). The cmu pronouncing dictionary.
- [128] Weide, R. (2008). The CMU Pronouncing Dictionary 0.7a. *Carnegie Mellon University*. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
- [129] Wells, J. (2000). Overcoming phonetic interference. *English Phonetics, Journal of the English Phonetic Society of Japan*, 3:9–21.
- [130] Wiese, R. (2001). The phonology of/r. *Distinctive feature theory*, 2:335.
- [131] Wikimedia (2014). Portuguese Wikipedia database dump backup. <http://dumps.wikimedia.org/ptwiki/20140123/>.
- [132] Wilcox-O’Hearn, A., Hirst, G., and Budanitsky, A. (2008). Real-word Spelling Correction with Trigrams: A Reconsideration of the Mays, Damerau, and Mercer Model. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing’08*, pages 605–616.
- [133] Young, S., Odell, J., and Woodland, P. (1994). Tree-based state tying for high accuracy acoustic modelling. *Cambridge University Engineering Department*.

- [134] Zampieri, M. and Amorim, R. (2014). Between Sound and Spelling: Combining Phonetics and Clustering Algorithms to Improve Target Word Recovery. In *Proceedings of the 9th International Conference on Natural Language Processing PolTAL 2014*, pages 438–449.
- [135] Zampieri, M., Gebre, B. G., and Nijmegen, H. (2012). Automatic identification of language varieties: The case of portuguese. In *Proceedings of KONVENS*, pages 233–237.

# **Appendix A**

## **How to install L<sup>A</sup>T<sub>E</sub>X**

### **Windows OS**

#### **TeXLive package - full version**

1. Download the TeXLive ISO (2.2GB) from  
<https://www.tug.org/texlive/>
2. Download WinCDEmu (if you don't have a virtual drive) from  
<http://wincdemu.sysprogs.org/download/>
3. To install Windows CD Emulator follow the instructions at  
<http://wincdemu.sysprogs.org/tutorials/install/>
4. Right click the iso and mount it using the WinCDEmu as shown in  
<http://wincdemu.sysprogs.org/tutorials/mount/>
5. Open your virtual drive and run setup.pl

or

#### **Basic MikTeX - T<sub>E</sub>X distribution**

1. Download Basic-MiK<sub>T</sub>E<sub>X</sub>(32bit or 64bit) from  
<http://miktex.org/download>
2. Run the installer
3. To add a new package go to Start » All Programs » MikTex » Maintenance (Admin) and choose Package Manager

4. Select or search for packages to install

## **TexStudio - T<sub>E</sub>X editor**

1. Download TexStudio from  
<http://texstudio.sourceforge.net/#downloads>
2. Run the installer

## **Mac OS X**

### **MacTeX - T<sub>E</sub>X distribution**

1. Download the file from  
<https://www.tug.org/mactex/>
2. Extract and double click to run the installer. It does the entire configuration, sit back and relax.

## **TexStudio - T<sub>E</sub>X editor**

1. Download TexStudio from  
<http://texstudio.sourceforge.net/#downloads>
2. Extract and Start

## **Unix/Linux**

### **TeXLive - T<sub>E</sub>X distribution**

#### **Getting the distribution:**

1. TeXLive can be downloaded from  
<http://www.tug.org/texlive/acquire-netinstall.html>.
2. TeXLive is provided by most operating system you can use (rpm, apt-get or yum) to get TeXLive distributions

## Installation

1. Mount the ISO file in the mnt directory

```
mount -t iso9660 -o ro,loop,noauto /your/texlive####.iso /mnt
```

2. Install wget on your OS (use rpm, apt-get or yum install)

3. Run the installer script install-tl.

```
cd /your/download/directory  
.install-tl
```

4. Enter command ‘i’ for installation

5. Post-Installation configuration:

<http://www.tug.org/texlive/doc/texlive-en/texlive-en.html#x1-320003.4.1>

6. Set the path for the directory of TexLive binaries in your .bashrc file

### For 32bit OS

For Bourne-compatible shells such as bash, and using Intel x86 GNU/Linux and a default directory setup as an example, the file to edit might be

```
edit $~/.bashrc file and add following lines  
PATH=/usr/local/texlive/2011/bin/i386-linux:$PATH;  
export PATH  
MANPATH=/usr/local/texlive/2011/texmf/doc/man:$MANPATH;  
export MANPATH  
INFOPATH=/usr/local/texlive/2011/texmf/doc/info:$INFOPATH;  
export INFOPATH
```

### For 64bit OS

```
edit $~/.bashrc file and add following lines  
PATH=/usr/local/texlive/2011/bin/x86_64-linux:$PATH;  
export PATH  
MANPATH=/usr/local/texlive/2011/texmf/doc/man:$MANPATH;  
export MANPATH
```

```
INFOPATH=/usr/local/texlive/2011/texmf/doc/info:$INFOPATH;
export INFOPATH
```

**Fedora/RedHat/CentOS:**

```
sudo yum install texlive
sudo yum install psutils
```

**SUSE:**

```
sudo zypper install texlive
```

**Debian/Ubuntu:**

```
sudo apt-get install texlive texlive-latex-extra
sudo apt-get install psutils
```

# Glossary

## **Application Programming Interface**

Term used in computer science to refer to a list of routines, protocols and tools for building applications. Roughly speaking, an API lists all classes, methods and functions that a given package has.. xvii