Using phonetic knowledge in tools and resources for Natural Language Processing and Pronunciation Evaluation

Gustavo Augusto de Mendonça Almeida

Gustavo Augusto de Mendonça Almeida

Using phonetic knowledge in tools and resources for Natural Language Processing and Pronunciation Evaluation

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Sandra Maria Aluisio

USP – São Carlos
February 2016

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

d539u

de Mendonça Almeida, Gustavo Augusto
Using phonetic knowledge in tools and resources for Natural Language Processing and Pronunciation Evaluation / Gustavo Augusto de Mendonça Almeida; orientadora Sandra Maria Aluisio. -- São Carlos, 2016.
82 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2016.

1. processamento de língua natural. 2. treino de pronúncia. 3. conversão grafema-fonema. 4. corretor ortográfico. 5. avaliação de pronúncia automática. I. Aluisio, Sandra Maria, orient. II. Título.

Gustavo Augusto de Mendonça Almeida

Utilizando conhecimento fonético em ferramentas e recursos de Processamento de Língua Natural e Treino de Pronúncia

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Sandra Maria Aluisio

USP – São Carlos
Fevereiro de 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
February 2016

Acknowledgements

I would like to thank my mother and my brother for their unconditional love, support and understanding through these past three years at USP. Their encouragement has given me the motivation to keep going. They were with me every step of the way – even though physically apart.

I would like to express deepest my gratitude to my adviser Prof. Sandra Aluisio, for the continuous support, patience and never-ending enthusiasm. Not only her guidance helped to develop my research, but also provided the moral support that I needed to move in the roughroad of writing a thesis. Thanks for being so open-minded in your research and willing to take risks. I admire your courage of advising a linguist in a complete new area of study.

My sincere thanks also goes to Prof. Thaïs Cristófaro for being such a passionate, whole-hearted and inspiring scientist. Thanks to her I was introduced to the world of Phonology – and I fell in love with it. Thanks for enlightening me the first glance of research.

I am also very grateful for all my fellow labmates at NILC. My time in São Carlos would not have been the same without them: André Cunha, Erick Fonseca, Fernando Asevedo, Pedro Balage, Lucas Avanço, Nathan Hartmann, Alessandro Bokan, Marco Sobrevilla and Roque Lopez. Juliana Balage and Fernanda Bontempo should also be included in this list, although not being labmates.

Linguists deserve a separate paragraph. Débora Garcia and Amanda Rassi, thanks for the wonderful time and the lovely discussions.

Thanks Christopher Shulby and Vanessa Marquiafável for their friendship, for being an amazing team and for having the boldness to face the unknown. Mari Shulby and Thor Shulby should also be included here.

Also, I would like to thank all my friends and colleagues from the Phonology Lab: Marco Fonseca, Ricardo Napoleão, Maria Cantoni, Jamila Rodrigues, Iiris Renicke, Bianka

Andrade, Leo Almeida, Janaína Rabelo, Ingrid Castro and Fred Baumgratz. A special thanks to Victor Medina, who also helped me with the review.

Several professors have contributed to my journey through Speech Sciences, my gratitude goes to Prof. Sara Candeias, Prof. Aldebaro Klautau, Prof. Fernando Perdigão, Prof. Heliana Mello, Prof. Tommaso Raso, Prof. Thiago Pardo, Prof. Graça Nunes, Prof. Márcia Cançado, Prof. Solange Rezende and Prof. Graça Pimentel.

My gratitude also goes to CNPq and to Samsung Eletrônica da Amazônia Ltda. for funding part of my research.

Finally, I would like to thank Reão Toniazzo for being the most amazing partner, friend and lover I could ever dream of. His everlasting patience and *malemolente* love kept me sane during hundreds of hours of work. Thanks for always being there to listen and to let me get my anxities out, especially when I was too grumpy. You make me a better person.

Gustavo Augusto de Mendonça Almeida

Abstract

This thesis presents tools and resources for the development of applications in Natural Language Processing and Pronunciation Training. There are four main contributions. First, a hybrid grapheme-to-phoneme converter for Brazilian Portuguese, named Aeiouadô, which makes use of both manual transcription rules and Classification and Regression Trees (CART) to infer the phone transcription. Second, a spelling correction system based on machine learning, which uses the trascriptions produced by Aeiouadô and is capable of handling phonologically-motivated errors, as well as contextual errors. Third, a method for the extraction of phonetically-rich sentences, which is based on greedy algorithms. Fourth, a prototype system for automatic pronunciation assessment, especially designed for Brazilian-accented English.

Palavras-chaves natural language processing; pronunciation training; text-to-speech; spelling correction; corpus balancing; automatic pronunciation assessment.

Resumo

Esta dissertação apresenta recursos voltados para o desenvolvimento de aplicações de reconhecimento de fala e avaliação de pronúncia. São quatro as contribuições aqui discutidas. Primeiro, um conversor grafema-fonema híbrido para o Português Brasileiro, chamado Aeiouadô, o qual utiliza regras de transcrição fonética e Classification and Regression Trees (CART) para inferir os fones da fala. Segundo, uma ferramenta de correção automática baseada em aprendizado de máquina, que leva em conta erros de digitação de origem fonética, que é capaz de lidar com erros contextuais e emprega as transcrições geradas pelo Aeiouadô. Terceiro, um método para a extração de sentenças foneticamente-ricas, tendo em vista a criação de corpora de fala, baseado em algoritmos gulosos. Quarto, um protótipo de um sistema de reconhecimento e correção de fala não-nativa, voltado para o Inglês falado por aprendizes brasileiros.

Palavras-chaves processamento de língua natural; treino de pronúncia; conversão grafema-fonema; corretor ortográfico; balanceamento de corpus; avaliação de pronúncia automática.

Table of contents

List of figures

List of tables

List of acronyms

AmE	American English.
API	Application Programming Interface.
ASR	Automatic Speech Recognition.
BP	Brazilian Portuguese.
CALL	Computer Assisted Language Learning.
CAPT	Computer Assisted Pronunciation Training.
CEFR	Common European Framework of Reference.
CFG	Context Free Grammar.
DNN	Deep Neural Network.
EF-EPI	EF English Proficiency Index.
F0	Fundamental Frequency.
G2P	Grapheme-to-Phoneme.
GMM	Gaussian Mixture Model.
HMM	Hidden Markov Model.
HTK	Hidden Markov Model Toolkit.
IPA	International Phonetic Alphabet.
IVR	Interactive Voice Response.
MFCC	Mel Frequency Cepstral Coefficients.

PCM	Pulse Code Modulation.
PLP	Perceptual Linear Prediction.
RASR	RASR.
WER	Word Error Rate.

Chapter 1

Introduction

Setting and Motivation

According to the International Monetary Found (IMF) [14], Brazil was the seventh largest economy in the world in 2015 with a GDP of US$ 2.34 trillions. A survey by The Economist (2013) says that, since 2009, the growth of BRICS accounts for 55% of the entire world economy growth. The current economic scenario is extremely favourable for Brazil to increase its global influence; however, with regard to the ability to communicate globally, Brazil occupies a much more modest position.

In 2015, Brazil ranked 41st out of 70 countries in the English Proficiency Index (EF-EPI) [11], being classified among countries with low English proficiency, with 51.05 points. Scandinavian countries led the very high proficiency ranking, with Sweden (70.94) in the first position, Denmark (70.05) in third place and Norway (67.83) in fourth. Brazil's performance was close to several other Latin American countries, such as Peru (52.46), Chile (51.88), Ecuador (51.67), Uruguay (50.25) and Colombia (46.54). The only exception in Latin America was Argentina, which, despite the country's turbulent economic situation, was ranked 15th, being classified as high proficiency, with a score of 60.26.

The EF English Proficiency Index (EF-EPI) bands are aligned with 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. Brazil's low proficiency rank is analogous to the CEFR B1 level.

As one might notice, the B1 level describes an individual who is usually able to understand familiar matters, deal with travelling situations, describe personal experiences and plans,

Table 1.1 CEFR reference levels.

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

and produce simple texts about subjects of personal interest. Needless to say, this is a very restricted communicative competence, which limits English usage primarily to the personal domain.

With respect to Business English proficiency, Brazil's performance is even more concerning. In the Business English Index (BEI) of 2013 [16], Brazil reached the 71st position out of 77 countries analysed. The country attained a score of 3.27 points, in a scale from 1 to 10, being placed at the "Beginner" range, the lowest range considered by the index. Brazil was close to countries such as El Salvador (3.24), Saudi Arabia (3.14) and Honduras (2.92), which up until recently had experienced civil wars or dictatorship governments. BEI describes individuals at the beginner level as those who "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 task in a company or industry work environment.

Given this scenario, it is clear that we need to improve English language proficiency among Brazilians. Computer Assisted Language Learning (CALL) systems can be of great help in this scenario. These systems have several benefits, such as [47]:

- Providing undivided attention to the user, as opposed to a classrom environment, where teachers need to share attention among all students.
- Being cheap and scaling well as what all students need is access to a computer.
- Enabling asynchronous learning, allowing people with time and place constraints to study at their own pace and schedule.
- High degree of individuality regarding the choice of material which is studied.

Obviously, there are disadvantages too. The limited interaction and feedback is often mentioned as a major problem in CALL. Nevertheless, some studies have shown that students who used CALL for training pronunciation training achieved equivalent results to students who were enrolled in traditional classes, with teacher-led pronunciation training [30, 42].

Gap and objectives

This project seeks to be a primary step towards developing a Computer Assisted Pronunciation Training (CAPT) for Brazilian-accented English. The initial plan was to focus on methods for creating CAPT systems for Brazilian-accented English by embedding as much phonetic

knowledge as possible. As Witt [48] points out in her survey on CAPT from 2012, back then, one of the core challenges in the area was that Automatic Speech Recognition (ASR) systems were not able to provide a reliable signal in terms of phone recognition. Our hypothesis was that by integrating phonetic knowledge in all stages of the pipeline for a CAPT, one would be able to improve the accuracy of the system and provide more trustworthy results.

However, as the project went through, with a broader view of the literature in CAPT and a more profound understading of how ASR works, such plan has shown to be unfeasible in the time given for a Master's Course. Due to the scarcity of resources, it was not possible to replicate for Brazilian-accented English several of the existing methods that were applied to other languages in the literature. For Brazilian-accented English, the only speech corpus currently available is the Corpus Oral de Brasileiros Aprendizes de Inglês (COBAI) [24], which is not suitable for speech recognition, due to the fact that the files are too long, usually consisting of interviews with 5 minutes; for training acoustic models, sentences should have around 10 seconds in order to get correct phone alignments. There were no specialized corpora for speech recognition in Brazilian-accented English, no language models trained from learners' texts and no dictionaries with pronunciation variants.

Instead, we decided to focus on investigating and building resources for Natural Language Processing and Pronunciation Evaluation which, in the future, may somehow help the development of CAPT systems for Brazilian-accented English. We worked particularly with tools and resources for text-to-speech, spelling correction, corpus building and automatic pronunciation assessment. In all cases, we attempted to integrate and embed phonetic knowledge in the pipeline.

Research Hypothesis

Text-to-speech

1. A hybrid approach, which includes both manual rules and machine learning, can be used to achieve state-of-the-art results for Grapheme-to-Phoneme (G2P) conversion in Brazilian Portuguese.

2. Supra-segmental information, such as lexical stress and syllable boundary, are useful for improving the accuracy of a G2P for Brazilian Portuguese (BP), especially with regard to vowel transcription.

3. Part-of-speech tags are able to provide enough information for a hybrid G2P to distinguish between heterophonic homograph pairs, such as "gov[e]rno" (government) "gov[ɛ]rno" (I govern).

Spelling correction

5. A considerable part of the typos that users make are phonologically-motivated, therefore phonetic transcription can be used to improve the coverage of spelling correction systems.

Corpus building

6. Greedy algorithms can help to produce richer speech corpora by making local optimal decisions, through analysing the triphone sequences of the sentences in a corpus and extracting those which keep the triphone distribution as uniform as possible in each iteration.

Automatic pronunciation assessment

7. Multipronunciation dictionaries with hand-written rules for generating variants are a reliable source of pronunciation information for pronunciation assessment.

8. Acoustic models trained on phonetically-rich speech corpora are able to provide more accurate phone models than those trained on balanced corpora.

9. Context-free grammars can be adapted for forced-alignment recognition to list all pronunciation variants of a given word without hindering the performance of the ASR.

10. Combined acoustic models (trained over native corpora + interlanguage data) have phone models which are robust enough to perform speech recognition in all languages used for training.

11. The additional pronunciation variants do not harm the performance.

Contributions of the Thesis

Within this work, we have investigated and developed a set of tools and resources which integrate phonetic knowledge – and benefit from it. Some of these tools were created and tested in tasks related to processing Brazilian-accented English or Brazilian Portuguese, but their architecture and methods are certainly scalable to other languages or scenarios. The full list of contributions is provided below[1]:

[1] All files, resources and scripts developed are available at the project website: *(http://nilc.icmc.usp.br/listener)*. Due to copyright reasons, the corpora used for training the acoustic models cannot be made available.

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 published in Mendonça and Aluísio [25]. *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. [27];

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 presented in Mendonça et al. [28];

4. *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, as presented in Mendonça and Aluísio [26];

Thesis Structure

This Master's thesis is organised into four chapters and follows the structure of a sandwich thesis, consisting of a collection of published or submitted articles. Chapter 2 presents the theoretical foundations, with an introduction to phonetics, as well as automatic speech recognition. Chapter 3 contains all articles that were published throughout the execution of this work or which are in-press. Section 3.1 presents the Aeiouadô's dictionary and G2P converter, which was built uppon a hybrid strategy for converting graphemes into phones, with manual transcription rules, as well as machine learning. Section 3.2 presents a use case of Aeiouadô, namely a phonetic-speller which employs the transcriptions generated by the grapheme-to-phoneme converter. Section 3.3 proposes a method for the extraction of phonetically-rich sentences, which can be used for building more representative speech corpora. Section 3.4 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 Chapter 4, we present the overall conclusions, some limitations we found, together with the next steps for future work.

Chapter 2

Theoretical Foundations

2.1 Phonetics

There is an endless debate about what the boundaries between phonetics and phonology are [43]. 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 corcerned with how these sounds are organised into patterns and systems [8].

For the 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 organised into a system of contrasts which are used distinctively to express meaning [7]. It is interesting to notice that, despite the fact that speech is above everything a continuous phenomenon, both phonetics and phonology will conjecture that speech can be examined through discrete units or segments[1].

Phonetics will analyse the a stream of speech from the viewpoint of a phone, i.e. the smallest perceptible discrete segment in speech [7]. 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 encompasses all sounds that the human vocal tract could possibly produce. For convenience, the IPA chart is plotted in Figure 2.1.

[1] 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.

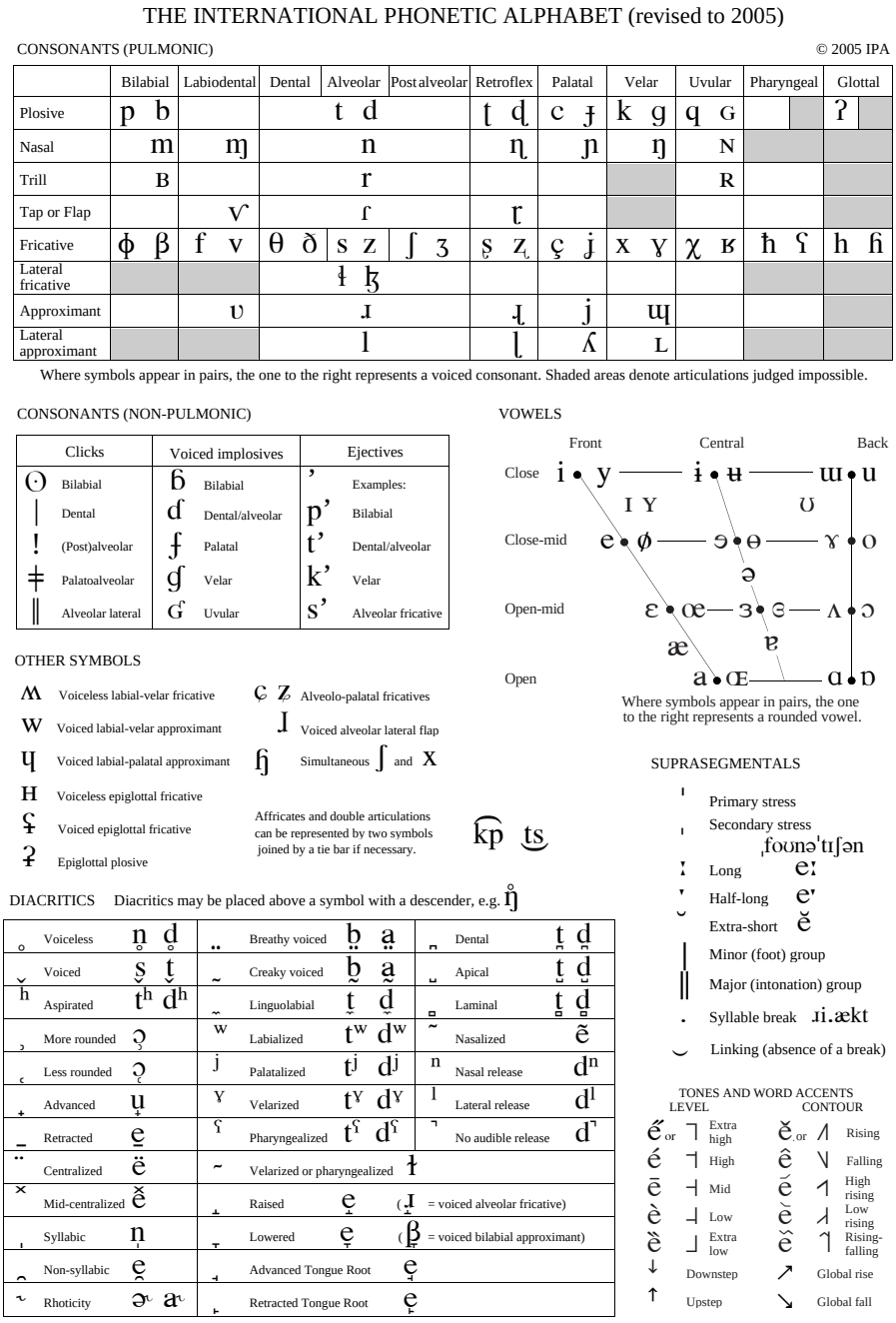


Fig. 2.1 IPA Chart.

The phones in the IPA chart are organised 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 centre-right position.

Pulmonic consonants are organised 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 voiceless (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 a way that the air pressure is so high that it bursts the region where it was blocked and passes through the lips, 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, [ɦ] is a voiced glotal 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 a schematic arrangement of the vowels which summarises 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 the tongue close to the roof of the mouth, such as the segment [u] in ***u****va* (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 ***a****ve* (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 *p****i****pa* (kite), are found in the left part of the vowel diagram; whereas back vowels, like [ɔ] in *ro****ç****a* (small farm), are on the right side.

Vowels and consonants are put together in sequence in order to form words, phrases and sentences. As 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: [ˈtak.sɪ]; in contrast, *aqui* (here) has four letters but only three phones [aˈki]. Despite their close relation, one must not mistaken letters for 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 [2–4, 40, 32], and despite the fact that the analyses usually concur with respect to core questions, there is a lot of disagreement in terms of convention and the usage of different phones. For instance, some authors propose that the postonic "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 [u̯] and [i̯] respectively.

There is not even a consensus as to which standard dialect one refers to when one says "Brazilian Portuguese". As a matter of fact, BP is the native language of nearly 190 million speakers in Brazil [17] 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 [40], since it is widely known and well-established in the area. Silva [40] proposes 46 phones for describing BP (26 consonants and 20 vowels)[2], all segments are grouped into Table 2.1, Figure 2.2 and Figure 2.3.

Table 2.1 Brazilian Portuguese consonants.

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

[2]For simplicity, symbols with optional secondary articulation [lˠ, lʲ] or with alternative notations [ỹ] were omitted.

Fig. 2.2 Brazilian Portuguese oral vowels.

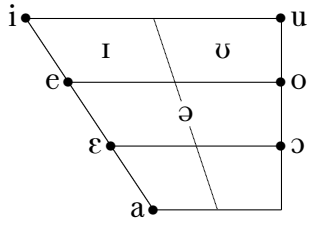
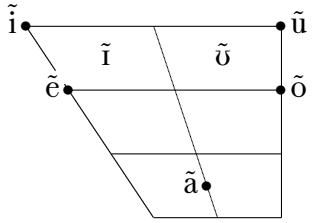


Fig. 2.3 Brazilian Portuguese nasal vowels.



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 so that no air can pass the vocal tract, and then by increasing the pressure in such a way that the bursts through the vocal tract, creating sound. Plosive sounds are also called "stops" or "occlusives". In BP plosive sounds usually occupy 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 velar plosive
[g]	[g]ato	gato	cat	voiced velar 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[.pɪ.]to [5]. A few other examples are shown in Table 2.3:

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. In other words, affricates begin with a stop and then are

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

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

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, ĩ], or when they occupy coda position. For example, in several dialects of BP, the word *tia* (aunt) is realised as [ˈtʃiə] with an initial voiceless postalveolar affricate [tʃ]. Similarly, *dia* (day) is often pronounced as [ˈdʒiə]. This phenomenon is called palatalisation and it results from an overlap among the speech gestures for [t, d] and high vowels [i, ɪ, ĩ]; 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 ***tch****u****tch****uca* (slang:pussycat) or the "dj" in ***Dj****avan* (a personal name). It is worth pointing out that in a few dialects of BP, palatalisation is a much broader phenomenon which affects other contexts as well [6]. A few other examples of words with affricates are provided in Table 2.4.

Table 2.4 Examples of affricate consonants in Brazilian Portuguese.

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

There are three nasal consonants in BP, viz. [m, n, ɲ]. Nasal consonants are produced with the velum lowered, in a way that the air is free to pass through the nose. In current language usage, due to vowel nasalisation, nasal consonants in BP are basically limited to syllable initial position, for example, as the "m" in ***m****ar* (sea), the "n" in ***n****ão* (no) or the "nh" in *rai****nh****a* (queen) respectively. It is important to notice that in words such as *ambos* (both)

or *anta* (tapir), nasalisation 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 a way that the vowel will be entirely nasalized and the nasal consonant will not perceived as a segment [10], i.e. *ambos* will be produced as [ˈãˑbʊs] and *anta* will become [ˈãˑtə] 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 nasalisation.

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

Phone	Transcription	Word	Translation	Description
[m]	ca[m]a	cama	bed	bilabial nasal
[n]	ca[n]a	cana	sugar cane	alveolar nasal
[ɲ]	ba[ɲ]a	banha	fat	palatal nasal
(no nasal cons)	[ã]tônio	Antônio	personal name	nasalized [a]
(no nasal cons)	l[ẽ]brar	lembrar	remember	nasalized [e]
(no nasal cons)	[ĩ]teresse	interesse	interest	nasalized [i]
(no nasal cons)	[õ]bro	ombro	shoulder	nasalized [o]
(no nasal cons)	[ũ]tar	untar	grease	nasalized [u]

The sounds [r, ɾ, ɹ, x, ɣ, h, ɦ] 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 [46].

The first one, called alveolar trill [r] is found in some dialects of BP – especially in 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 rhotic class (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 trill [r] is closely related to the alveolar tap [ɾ], the only difference being 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* [pɛ.ro] vs. *pero* [pɛ.ɾo].

The trill is closely related to the alveolar tap [ɾ], 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* [pɛ.ro] vs. *pero* [pɛ.ɾo]. However, different from the trill, the tap [ɾ] 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 alveolar approximant [ɹ] is the sound which corresponds to the so-called "r-caipira" in BP. It is an approximant consonant, which means that the vocal tract is narrowed, but the level of constriction is not sufficient to generate hiss or turbulence. In the dialects in which

this rhotic sound occur, it is found mostly at the end of a syllable, as the "r" in *amor* (love) or *porta* (door)

The other rhotics variants [x, ɣ, h, ɦ] can be considered free variants or free allophones among themselves, they are also referred to as strong-r, in contrast with the tap. The first two, [x, ɣ] are velar fricative sounds consonants, in other words, they are produced in such a 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 folds, i.e. they are open. On the other hand, [ɣ] is a voiced velar fricative, which means that it puts the vocal folds to vibrate when it is produced. The phones [h, ɦ] are articulated in the region of the glottis and they also show constriction in the air passage, that is why the are called glottal fricatives. Analogously to [x, ɣ], [h, ɦ] also present the voiceless-voiced dichotomy; the vocal folds are open when [h] is produced, but they are closed and vibrate in [ɦ]. Table 2.6 presents some examples of words with rhotic sounds in BP.

Table 2.6 Examples of rhotics in Brazilian Portuguese.

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

Apart from the rhotic ones, BP has six more fricative sounds: [f, v, s, z, ʃ, ʒ]. The first two are named labiodental because they are produced by making the lips touch the upper teeth. As other fricative sounds, the air for [f, v] does not passes freely in the vocal tract, on contrary it finds obstacles thus generating turbulence. With respect to [s, z], both are articulated in the region of the alveolar ridge, that is why they are called alveolar fricatives. Finally, [ʃ, ʒ] are produced more towards the back of the vocal tract, in a place between the alveolar ridge and the hard palate; this is the reason why they are referred to as postalveolar

or palato-alveolar consonants. All these six fricative sounds can be found in all dialects of BP in onset position, as can be seen from the examples in Table 2.7.

Table 2.7 Examples of fricative consonants in Brazilian Portuguese (onset).

Phone	Transcription	Word	Translation	Description
[f]	[f]aca	faca	knife	voiceless bilabial plosive
[v]	[v]aca	vaca	cow	voiced bilabial plosive
[s]	ca[s]a	caça	hunt	voiceless alveolar plosive
[z]	ca[z]a	casa	house	voiced alveolar plosive
[ʃ]	quei[ʃ]o	queixo	chin	voiceless bilabial plosive
[ʒ]	quei[ʒ]o	queijo	cheese	voiceless bilabial plosive

In coda position, [f, v, s, z, ʃ, ʒ] show a different behaviour. The labiodental fricatives [f, v] act similarly to the plosives summarised in Table 2.3, they may occupy the final position of a syllable, e.g. a[f]ta (cold sore), but epenthesis will often take place: a[.fɪ.]ta. Alveolar fricatives are present in coda position in most dialects of BP. Some regions of Brazil have postalveolar fricatives [ʃ, ʒ] instead, notably the dialect spoken in Rio de Janeiro. For [s, z, ʃ, ʒ], anticipatory assimilation more often than not will occur, thus the choice between [s, ʃ] and [z, ʒ] will depend on the following consonant, if it is voiced, then the fricative will also be voiced. For example, the fricative in *rasgar* (to rip) is voiced: ra[z]gar; but the one in *costa* (coast) is not: co[s]ta. The same distinction will be present in dialects with the postalveolar fricatives [ʃ, ʒ]. In Table 2.8, one can find more examples of fricatives in coda in BP.

Table 2.8 Examples of fricative consonants in Brazilian Portuguese (coda).

Phone	Transcription	Word	Translation	Description
[f]	[af.]∼[a.fɪ.]ta	apto	able.MASC	voiceless bilabial plosive
[f]	[of.]∼[o.fi.]talmologia	oftalmologia	ophthalmology	voiceless bilabial plosive
[s]	po[s, ʃ]tar	postar	to post	voiceless bilabial plosive
[s]	ca[s, ʃ]tor	castor	beaver	voiced bilabial plosive
[z]	de[z, ʒ]gaste	desgaste	wear and tear	voiceless alveolar plosive
[z]	tran[z, ʒ]gressivo	transgressivo	transgressive.MASC	voiced alveolar plosive

Glides (also known as semivowels) are phones which are similar to vowels in terms of acoustics or articulation, but which function as consonants in terms of phonotactics, in other words, they are not placed in the nucleus of a syllable. There are two glides in BP, one which has its place of articulation in the velar region [w] and another one which is produced near the hard palate [j]. Acoustically, the velar glide [w] is very similar to the vowel [ʊ], and palatal glide is very close to [ɪ]. The debate whether these sounds should be considered vowels or glides is beyond the scope of this thesis. Table 2.9 presents some examples with glides.

Table 2.9 Examples of glides in Brazilian Portuguese.

Phone	Transcription	Word	Translation	Description
[w]	cé[w]	céu	sky	voiceless bilabial plosive
[w]	pa[w]	pau	stick	voiceless bilabial plosive
[w]	cinq[w]enta	cinquenta	fifty	voiceless bilabial plosive
[j]	fu[j]	fui	(I) was	voiceless bilabial plosive
[j]	pa[j]xão	paixão	passion	voiceless bilabial plosive
[j]	ce[j]a	ceia	supper	voiceless bilabial plosive

BP has two consonants which are articulated by making the air escape the vocal tract around the sides of the tongue: [l, ʎ]; due to this articulatory aspect, these sounds are called lateral consonants. The former [l] is named lateral alveolar since it is produced in the region of the gum ridge. The latter is articulated with the body of the tongue reaching the hard palate, thus [ʎ] is considered a lateral palatal consonant. In terms of context of occurrence, both laterals show a very different distribution.

The alveolar lateral is present in onset position in all dialects of BP. It corresponds to the "l" in words like *lata* (can) or *pular* (to skip). However in coda [l] commonly undergoes vocalisation, and is produced as a vowel [ʊ̯] or glide [w], for example, in *sal* (salt) or *Sol* (sun), both "l" are frequently pronounced as vowels or glides, instead of consonant.

As for the palatal lateral, it is limited to syllable initial position and generally corresponds to the letters "lh" in writing, e.g. *lhama* (llama) or *alho* (garlic). A few other examples of words with laterals can be seen in Table 2.10.

Table 2.10 Examples of lateral consonants in Brazilian Portuguese.

Phone	Transcription	Word	Translation	Description
[l]	sa[l]a	sala	classroom	voiceless bilabial plosive
[l]	[l]ança	lança	spear	voiced bilabial plosive
(l-vocalisation)	sa[w]to	salto	jump	voiceless alveolar plosive
(l-vocalisation)	ca[w]da	calda	syrup	voiced alveolar plosive
[ʎ]	a[ʎ]eio	alheio	someone else's	voiceless bilabial plosive
[ʎ]	o[ʎ]ar	olhar	look	voiceless bilabial plosive

With respect to vowels, as it can be observed in Figure 2.2, BP has ten oral vowels: [i, ɪ, e, ɛ, a, ə, ɔ, o, u, ʊ]. Different from consonants, vowels are described in terms of height, backness and roundedness. Height refers how close the tongue is to the roof of the mouth or how open the jaw is. Backness describes how retracted the tongue is relative to the back of the mouth. Finally roundedness indicates the position of the lips when the vowel is articulated.

BP has four vowels which are produced forward in the mouth [i, ɪ, e, ɛ, a], all of which are not rounded. There are four back vowels [ɔ, o, u, ʊ], which are all rounded, i.e. they are produced with lip protrusion. One central vowel [ə] – also called "schwa" – also exists in BP.

The vowels [i, e, ɛ, a, ɔ, o, u] are considered tense, which means that they occur in pretonic or tonic syllables; whereas [ɪ, ə, ʊ] are relaxed, being found just in postonic contexts. A few examples of vowels in BP can be seen in Table 2.11 and Table 2.12. As one might see from the examples, the distribution of vowels in BP is deeply influenced by the lexical stress, postonic syllables use just a subset of the vowels which are present in pretonic or tonic syllables.

Table 2.11 Examples of vowels in Brazilian Portuguese (pretonic and tonic).

Phone	Transcription	Word	Translation	Description
[i]	S[i]béria	Sibéria	Siberia	high front unrounded vowel
[i]	b[i]co	bico	nib	high front unrounded vowel
[i]	s[i]go	sigo	(I) follow	high front unrounded vowel
[e]	p[e]dalar	pedalar	to pedal	mid-high front unrounded vowel
[e]	p[e]ra	pera	pear	mid-high front unrounded vowel
[e]	p[e]sames	pêsames	condolence	mid-high front unrounded vowel
[ɛ]	p[ɛ]zinho	pezinho	little foot	mid-low front unrounded vowel
[ɛ]	p[ɛ]ste	peste	plague	mid-low front unrounded vowel
[ɛ]	p[ɛ]	pé	foot	mid-low front unrounded vowel
[a]	g[a]linha	galinha	chicken	low front unrounded vowel
[a]	c[a]sa	casa	house	low front unrounded vowel
[a]	ch[a]	chá	tea	low front unrounded vowel
[ɔ]	h[ɔ]rinha	horinha	lit. little hour	mid-low back rounded vowel
[ɔ]	g[ɔ]sto	gosto	(I) like	mid-low back rounded vowel
[ɔ]	s[ɔ]	só	alone	mid-low back rounded vowel
[o]	rod[o]via	rodovia	highway	mid-high back rounded vowel
[o]	g[o]sto	gosto	taste	mid-high back rounded vowel
[o]	b[o]lo	bolo	cake	mid-high back rounded vowel
[u]	[u]tilidade	utilidade	use	high back rounded vowel
[u]	[u]va	uva	grape	high back rounded vowel
[u]	p[u]s	pus	(I) put	high back rounded vowel

2.1.2 The Phonetic Inventory of American English

For the phonetic inventory of American English (AmE), we will assume the analysis proposed by Skandera and Burleigh [41]. Skandera and Burleigh [41] describe the standard dialect of AmE through a set of 44 phones (24 consonants, 12 vowels and 8 diphthongs). Table 2.13, Figure 2.4 list all segments.

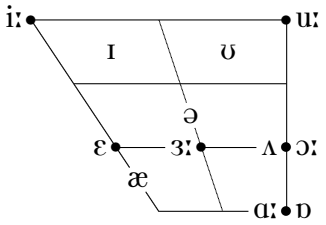
Table 2.12 Examples of vowels in Brazilian Portuguese (postonic).

Phone	Transcription	Word	Translation	Description
[ɪ]	quas[ɪ]	quase	almost	relaxed high front unrounded vowel
[ɪ]	pont[ɪ]	ponte	bridge	relaxed high front unrounded vowel
[ə]	cas[ə]	casa	house	relaxed low front unrounded vowel
[ə]	menin[ə]	menina	girl	relaxed low front unrounded vowel
[ʊ]	menin[ʊ]	menino	boy	relaxed high back rounded vowel
[ʊ]	rat[ʊ]	rato	mouse	relaxed high back rounded vowel

Table 2.13 American English consonants.

	Bilabial	Labiod.	Dental	Alveolar	Postalv.	Palatal	Velar	Glottal
Plosive	p b			t d			k g	
Affricate				tʃ dʒ				
Nasal	m			n			ŋ	
Tap				ɾ				
Fricative		f v	θ ð	s z	ʃ ʒ			h
Approximant				ɹ		j	w	
Lateral Appr.				l				

Fig. 2.4 American English vowels.



English has the same six plosive sounds that are found in BP, namely [p, b, t, d, k, g]. However, in English, the voiceless plosives are not produced the same way as they are in BP; in some contexts these plosive sounds become aspirated, that is to say they are produced with a burst of breadth in the release stage. For instance, the word *time* is not pronounced with a simple alveolar plosive [t], but with an aspirated one [t^h]. It is worth noticing that voiced plosives do not undergo aspiration; [b, d, g] are produced in a similar way to BP, although only with slightly less voicing [37].

In English, all stop consonants may occur in coda position without any epenthesis, as in *stop*, *bob*, *act*, etc. A few more examples of plosives in English can be observed in Table 2.14.

Aspiration does not take place in every context. For instance, when [p, t, k] form a complex onset, they are not aspirated. Instead, the following consonant may become partly voiceless as one might notice from the examples in Table 2.15.

Table 2.14 Examples of plosive consonants in American English (I).

Phone	Transcription	Word	Description
[p^h]	[p^h]ain	pato	aspirated voiceless bilabial plosive
[p]	sto[p]	stop	voiceless bilabial plosive
[b]	[b]ad	bad	voiced bilabial plosive
[b]	ca[b]	cab	voiced bilabial plosive
[t^h]	[t^h]op	top	aspirated voiceless alveolar plosive
[t]	spri[t]	sprite	voiceless alveolar plosive
[d]	[d]ay	day	voiced alveolar plosive
[d]	mo[d]	mode	voiced alveolar plosive
[k^h]	[k^h]at	cat	aspirated voiceless velar plosive
[k]	shrin[k]	shrink	voiceless velar plosive
[g]	[g]ate	gate	voiced velar plosive
[g]	dra[g]	drag	voiced velar plosive

Table 2.15 Examples of plosive consonants in American English (II).

Phone	Transcription	Word	Description
[p]	[p$\underset{\circ}{\text{ɹ}}$]ay	pray	voiceless bilabial plosive
[p]	[p$\underset{\circ}{\text{l}}$]ay	play	voiceless bilabial plosive
[t]	[t$\underset{\circ}{\text{ɹ}}$]rain	train	voiceless alveolar plosive
[t]	[t$\underset{\circ}{\text{j}}$]une	tune	voiceless alveolar plosive
[k]	[k$\underset{\circ}{\text{ɹ}}$]ane	crane	voiceless velar plosive
[k]	[k$\underset{\circ}{\text{l}}$]ock	clock	voiceless velar plosive

Similarly to BP, English has two affricate postalveolar sounds: [tʃ, dʒ]. These are very close to the ones which exist in BP, the only difference is that in English, the fricative phase of the affricate tend to be shorter. For example, the [ʃ] stage of the initial [tʃ] in a word like *cheap* is shorter than the [ʃ] in *sheep*. In contrast with BP, affricates in English are not positional variants of [t] or [d] and may occur both in the onset or in the coda position of a syllable. One can observe examples of affricates in Table 2.16.

Table 2.16 Examples of affricate consonants in American English.

Phone	Transcription	Word	Description
[tʃ]	[tʃ]ange	change	voiceless postalveolar affricate
[tʃ]	ca[tʃ]ing	catching	voiceless postalveolar affricate
[tʃ]	ri[tʃ]	rich	voiceless postalveolar affricate
[dʒ]	[dʒ]oke	joke	voiced postalveolar affricate
[dʒ]	ba[dʒ]es	badges	voiced postalveolar affricate
[dʒ]	a[dʒ]	age	voiced postalveolar affricate

There are three nasal consonants in English, namely [m, n, ŋ]. In comparison to BP, the only difference lies in the last phone [ŋ], which is a velar nasal consonant and does not exist in BP – note that BP has [ɲ] and not [ŋ]. In terms of distribution, unlike BP, nasal consonants may appear in the coda position of a syllable. In addition to this, in English, vowels are usually not nasalized when they are succeded by a nasal consonant, in other words, the [ɛ] in *men* remains the same as in *merry*, with no nasalisation.

As for the velar nasal [ŋ], there is one specific detail about its distribution: it only occurs in coda position. This sound is usually related to the sequence of letters "ng" in English, as in *king* or *studying*. Table 2.17 presents some examples of nasal consonants in English.

Table 2.17 Examples of nasal consonants in American English.

Phone	Transcription	Word	Description
[m]	[m]ay	may	bilabial nasal
[m]	li[m]p	limp	bilabial nasal
[m]	li[m]	limb	bilabial nasal
[n]	[n]ame	name	alveolar nasal
[n]	se[n]d	send	alveolar nasal
[n]	pla[n]	plane	alveolar nasal
[ŋ]	so[ŋ]	song	velar nasal
[ŋ]	wro[ŋ]	wrong	velar nasal

In English, only the alveolar approximant [ɹ] is considered a rhotic, [h, ɾ] are not part of the rhotic class. This mismatch is due to the fact that the rhotic class is defined based on phonological criteria – not phonetic.

All of these sounds have been previously discussed for BP. The alveolar retroflex [ɹ] corresponds to the sound which is usually represented in the English orthography by the letter letter "r", as in *car* or *rat*. The voiceless glottal fricative [h] has one particular characteristic in English: it only occurs in word initial position and corresponds to the letter "h" in written language. Some examples are "huge" and "helmet". Finally, the flap [ɾ] is the same phone that, in BP, occurs in words like *arara* (parrot) or *fruta* (fruit). However, different from BP, in English, such sound is an allophone of [t, d] and is present in intervocalic contexts, in words like *city* or *better*. Table 2.18 presents some other examples of words with rhotics in English.

Apart from the above-mentioned glottal fricative [h], English has seven more fricative sounds: [f, v, θ, ð, s, z, ʃ, ʒ]. The phones [f, v, s, z, ʃ, ʒ] are present in BP and the major difference lies in their context of occurrence. In BP, when [f, v] occur in coda position, epenthesis often takes place. However, in English, these sounds, along with [ʃ, ʒ], may occur in coda position without any constraint, as shown in the examples in Table 2.19.

Table 2.18 Examples of rhotics in American English.

Phone	Transcription	Word	Description
[ɹ]	bee[ɹ]	beer	alveolar approximant
[ɹ]	[ɹ]ound	round	alveolar approximant
[ɹ]	b[ɹ]eath	breath	alveolar approximant
[h]	[h]oney	honey	voiceless glottal fricative
[h]	[h]ave	have	voiceless glottal fricative
[ɾ]	bu[ɾ]er	butter	alveolar flap
[ɾ]	we[ɾ]ing	wedding	alveolar flap
[ɾ]	ci[ɾ]y	city	alveolar flap

Table 2.19 Examples of fricatives in American English.

Phone	Transcription	Word	Description
[f]	[f]ast	fast	voiceless labiodental fricative
[f]	su[f]er	suffer	voiceless labiodental fricative
[f]	lea[f]	leaf	voiceless labiodental fricative
[v]	[v]ery	very	voiced labiodental fricative
[v]	gi[v]	give	voiced labiodental fricative
[v]	he[v]v	have	voiced labiodental fricative
[s]	[s]aid	said	voiceless alveolar fricative
[s]	[s]prite	sprite	voiceless alveolar fricative
[s]	[s]niff	sniff	voiceless alveolar fricative
[z]	[z]ebra	zebra	voiced alveolar fricative
[z]	doe[z]	does	voiced alveolar fricative
[z]	kisse[z]	kisses	voiced alveolar fricative
[ʃ]	[ʃ]ure	sure	voiceless postalveolar fricative
[ʃ]	ca[ʃ]	cash	voiceless postalveolar fricative
[ʒ]	c[ʒ]	car	voiced postalveolar fricative
[ʒ]	ba[ʒ]	butter	voiced postalveolar fricative

Another difference is related to the allowed positions of [s]. In English, this fricative can be followed by a plosive or nasal consonant in word-initial position, giving rise to complex onset clusters, as the [st] in “stop”. In BP, this sequence of sounds would usually undergo epenthesis, by adding an initial [i]. This phenomenon is so stable that it can even be noticed in the orthograhy of Portuguese loanwords which are borrowed from English, e.g. *snorkel* becomes *esnórquel* with an initial letter “e”. A few examples can be seen in Table 2.20 – for all these contexts, Brazilian speakers will tend insert an [i] at the beginning of the word.

English also has two fricative sounds which do not exist in BP, namely [θ, ð]. These phones correspond to interdental fricatives, which means that they are produced with the tip of the tongue below the upper teeth with constriction when air passess through the oral cavity. The former sound, [θ], is a voiceless interdental fricative, whereas the latter, [ð], is its

Table 2.20 Examples of initial complex onsets with alveolar fricatives in American English.

Phone	Transcription	Word	Description
[s]	[s]prite	sprite	voiceless alveolar fricative
[s]	[s]top	said	voiceless alveolar fricative
[s]	[s]chool	said	voiceless alveolar fricative
[s]	[s]small	small	voiceless alveolar fricative
[s]	[s]niff	sniff	voiceless alveolar fricative

voiced counterpart. In terms of letter-to-sound relation, both sounds often correspond to the sequence "th" in writing, as the "th" in "thin" or "that", respectively. Further examples can be found in Table 2.21.

Table 2.21 Examples of interdental fricatives in American English.

Phone	Transcription	Word	Description
[θ]	[θ]ink	think	voiceless interdental fricative
[θ]	au[θ]or	author	voiceless interdental fricative
[θ]	too[θ]	tooth	voiceless interdental fricative
[ð]	[ð]is	this	voiced interdental fricative
[ð]	mo[ð]er	mother	voiced interdental fricative
[ð]	ba[ð]	butter	voiced interdental fricative

There are two glides in English, [w, j][3]. Both are quite similar to those existing in Portuguese. Glides are phones which hold, at the same time, properties of vowels and consonants. In terms of acoustic parameters, the velar glide [w] is very similar to the vowel [ʊ], and the palatal one is very close to [ɪ]. In English, the velar glide usually corresponds to the graphemes "w" or "wh" in writing, as the "w" in *wine* or the "wh" in *where*. The palatal glide, on the other hand, is usually related to the letter "y", as in *yard*, or the letter "u" in *university*. Table 2.22 presents some other examples with glides.

Table 2.22 Examples of glides in American English.

Phone	Transcription	Word	Description
[w]	[w]atch	watch	velar approximant glide
[w]	q[w]iet	quiet	velar approximant glide
[w]	t[w]elve	twelve	velar approximant glide
[j]	[j]ou	you	palatal approximant glide
[j]	T[juː]esday	Tuesday	palatal approximant glide
[j]	[juː]nion	union	palatal approximant glide

English has one lateral sound, the alveolar lateral [l]. Different from BP, this sound is allowed in coda position and does not undergo vocalisation. This sound is represented in

[3]Some analysis also consider [ɾ] a glide [33].

writting with the letter “l” or with double “ll”, in words like *light* or *fell*. A few other cases can be found in Table 2.23.

Table 2.23 Examples of lateral consonants in Brazilian Portuguese.

Phone	Transcription	Word	Description
[l]	[l]ove	love	lateral alveolar approximant
[l]	[l]ift	lift	lateral alveolar approximant
[l]	f[l]ight	flight	lateral alveolar approximant
[l]	Goog[l]	Google	lateral alveolar approximant
[l]	beautifu[l]	alheio	lateral alveolar approximant

English has twelve vowels [iː, ɛ, ɑː, ɒ, ʌ, ɔː, uː, ə, ɜː, ɪ, ʊ, æ] as described in Figure 2.4. As it can be seen, most of these vowels do not occur in BP, which leads to several L1 negative-transference phenomena. In other words, when pronouncing English words and utterances, Brazilian speakers will usually replace the English vowels with those ones which are more phonetically similar in BP. Unlike BP, English vowels also differ in terms of length. The vowels [iː, ɑː, ɔː, uː, ɜː] are long whereas [ɛ, ɒ, ʌ, ə, ɪ, ʊ, æ] are short.

It is worth noticing that what are usually called long and short vowels in English differ not only with respect to duration, but also in vowel quality – i.e. the acoustic parameters of a long high front vowel [iː] are different from those of an [i]. For instance, not only have the vowels in sheep and ship different length, but also they are produced with the tongue in a different position. The former [iː] is produced with the tongue higher and more towards the front of the mouth, in comparison to [ɪ]. Examples of vowels can be found in Table 2.24.

Table 2.24 Examples of vowels in American English.

Phone	Transcription	Word	Description
[iː]	ch[iː]k	cheek	high front unrounded vowel
[iː]	r[iː]ch	reach	high front unrounded vowel
[iː]	b[iː]n	been	high front unrounded vowel
[ɪ]	ch[ɪ]k	chick	high front unrounded vowel
[ɪ]	r[ɪ]ch	rich	high front unrounded vowel
[ɪ]	b[ɪ]n	bin	high front unrounded vowel
[ɛ]	b[ɛ]t	bet	high front unrounded vowel
[ɛ]	m[ɛ]sh	mesh	high front unrounded vowel
[ɛ]	p[ɛ]n	pen	high front unrounded vowel
[æ]	b[æ]t	bat	high front unrounded vowel
[æ]	m[æ]sh	mash	high front unrounded vowel
[æ]	p[æ]n	pan	high front unrounded vowel
[ʌ]	b[ʌ]t	but	high front unrounded vowel
[ʌ]	m[ʌ]sh	mush	high front unrounded vowel
[ʌ]	p[ʌ]n	pun	high front unrounded vowel
[ə]	[ə]mount	amount	high front unrounded vowel
[ə]	ign[ə]rant	ignorant	high front unrounded vowel
[ə]	Chin[ə]	China	high front unrounded vowel
[ɒ]	l[ɒ]st	lost	high front unrounded vowel
[ɒ]	l[ɒ]ck	lock	high front unrounded vowel
[ɒ]	c[ɒ]d	cod	high front unrounded vowel
[ɑː]	l[ɑː]st	last	high front unrounded vowel
[ɑː]	p[ɑː]ss	pass	high front unrounded vowel
[ɑː]	h[ɑː]rd	hard	high front unrounded vowel
[ɜː]	l[ɜː]rks	lurks	high front unrounded vowel
[ɜː]	p[ɜː]rse	purse	high front unrounded vowel
[ɜː]	h[ɜː]rd	heard	high front unrounded vowel
[ɔː]	P[ɔː]l	Paul	high front unrounded vowel
[ɔː]	c[ɔː]rd	cord	high front unrounded vowel
[ɔː]	f[ɔː]rk	fork	high front unrounded vowel
[ʊ]	p[ʊ]ll	pull	high front unrounded vowel
[ʊ]	l[ʊ]k	look	high front unrounded vowel
[ʊ]	sh[ʊ]d	should	high front unrounded vowel
[uː]	p[uː]ll	pool	high front unrounded vowel
[uː]	l[uː]ke	Luke	high front unrounded vowel
[uː]	sh[uː]d	shoed	high front unrounded vowel

2.2 Automatic Speech Recognition

2.2.1 Some context

Although the task of recognising words from speech might seem apparently simple beforehand, after all, humans start doing it when they are as young as four months old, the task is actually very complex one. If each word in a language were pronounced in the same way by all speakers in every situation, the task of automatic speech recognition would be considered solved, since all patterns could be defined in advance. However, the linguistic reality is quite the opposite. In fact, it is no exaggeration to say that a vowel is never pronounced in the exact same way – i.e. with the same acoustic properties – even considering a single speaker [21]. Intra- and inter-speaker variability are inherent to natural languages.

Over the years, many methods have been proposed to attempt to solve the problem of automatic speech recognition, until now, no solution has been found and machines are still a very long way from performing like humans. Back in 2001, Huang et al. [19] did a comparison between the performance of humans and machines in some recognition tasks, the results are summarised in Table 2.25.

Table 2.25 Word error rate comparisons between human and machines on similar tasks [19].

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%

As one may observe, humans outperform machines in almost every task, especially 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 similarly to humans. Although this dream is somewhat near for rather simple tasks such as connected digits, for other contexts a long path still lies ahead. Whereas humans had a WER of 3.8% in recognising spontaneous speech, for machines, this rate was as high as

36.7% in 2001. Some more up-to-date results can be found in Figure 2.5, which shows the performance of ASR until 2014 in many different tasks [20].

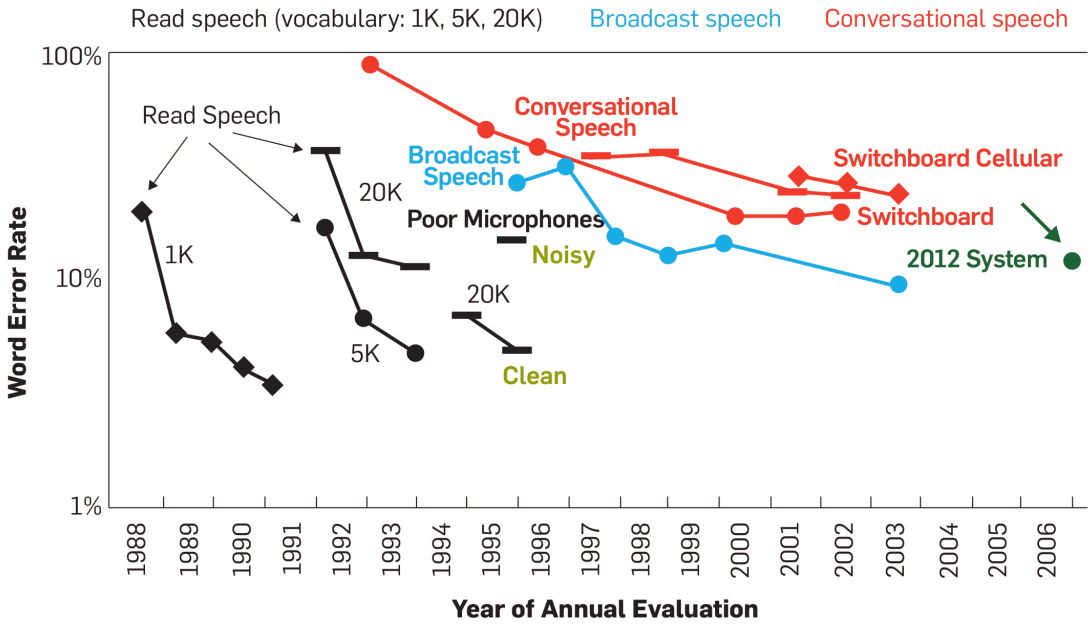


Fig. 2.5 Historical progress of speech recognition word error rate on more and more difficult tasks.

The results reported in Figure 2.5 are often from papers from large IT companies, such as Google, Apple or Microsoft, which have at their disposal not only the best algorithms and computer power in the world, but also the largest databases, some of them with up to 2,000 hours for training [20]. Nonetheless, ASR is such a difficult task, that, despite having all set, the best results reported are still around 15% for conversational speech [20]. This means that a short sentence with 4 words will be fully recognised only 52% of the times.

Such result for conversational speech is mainly due to the large amount of linguistic variability. Language varies not only among speakers (the so-called inter-speaker variability), but also within the same speaker (intra-speaker) [1]. 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 [1]. 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 [22] found that, in women's speech, strong expressions of feeling are avoided, uncertainty is favoured, and means of expression in regard to subject-matter deemed "trivial" to the "real" world are elaborated.

Disregarding social aspects, men, women and children's speech are also contrasting due to 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 [13]. Figure 2.6 presents a comparison between height and vocal tract length for men, women and children. Figure 2.7 presents a model of the vocal tract morphology considering age.

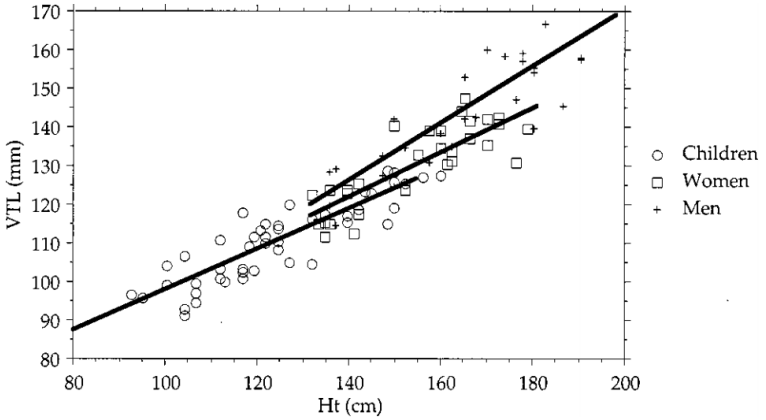


Fig. 2.6 Height (cm) versus vocal tract length (mm).

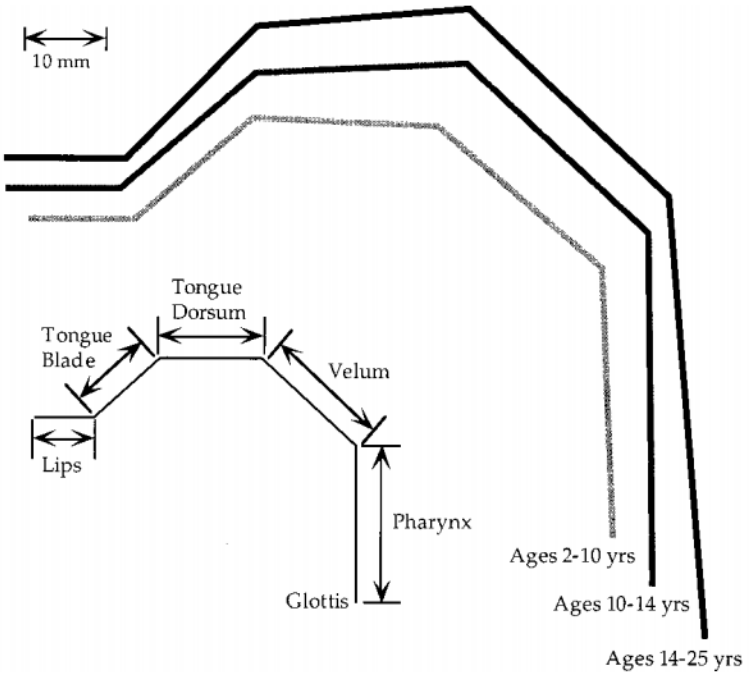


Fig. 2.7 Averaged vocal tract morphology.

As one can observe, men's vocal tract are longer than women's, followed by the children's. These differences affect the speech signal thoroughly, especially in what concerns to the Fundamental Frequency (F0). F0 can be defined as the lowest frequency in the signal counting from zero. Figure 2.8 compares the F0 values between male and females considering aging.

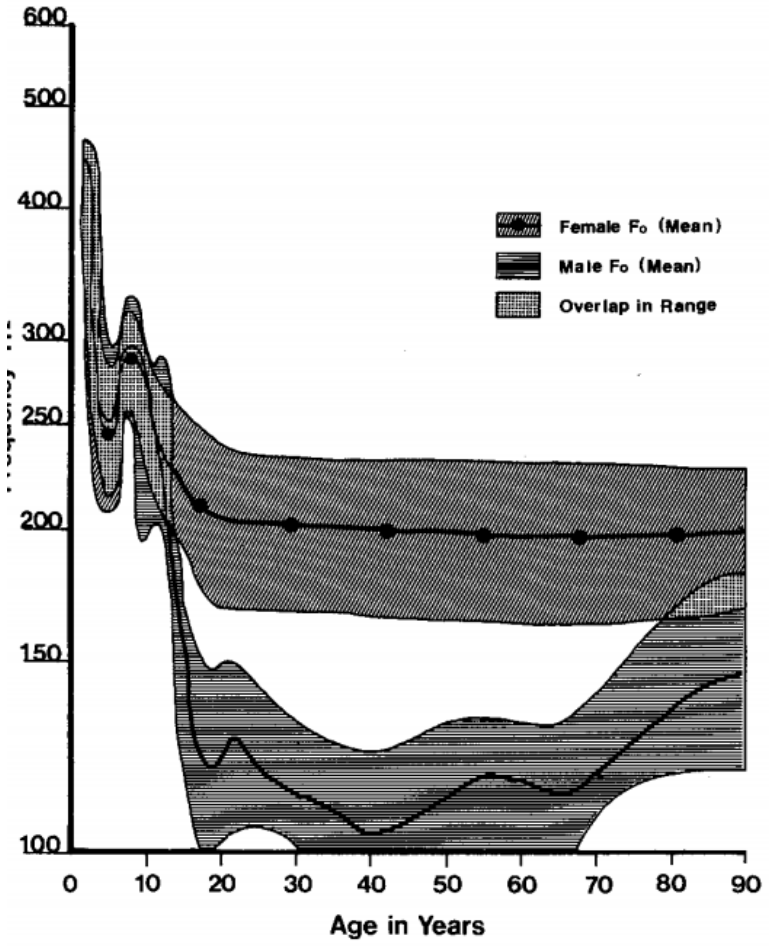


Fig. 2.8 F0 and pitch sigma versus age for males and females.

One can notice from Figure 2.8, 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 undergoes a huge drop, whereas for female voice the drop is quite small. In terms of perception, this is the period when the male voice lowers and becomes deeper.

Back to Table 2.25, it is interesting to notice that humans performed better in all speech recognition tasks, but one: "Clean speech based on trigram sentences". This very task consists of recognising 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 aspects of human hearing, that is, that we humans make a large use of syntactic, semantic and also pragmatic information in order to understand speech. Hearing does not take into account 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 [aɪsˈkrim], the vendor has no doubt that this sequence of phones refers to "ice cream" and not "I scream" [aɪˈskrim], albeit they are very much alike. Such sequences of phones might cause confusion in the phonetic level, but in higher linguistic levels their difference is quite clear, 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!).

2.2.2 The fundamental equation

Roughly speaking the purpose of an ASR system is to transform, in a precise and efficient way, the acoustic parameters of a speech signal into readable text [35]. 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:

$$W* = \arg\max_{W} P(W|O) \tag{2.1}$$

To solve this equation straightforwadly, one would require a discriminative model, capable of estimating the the probability of W directly from a set of observations O [15]. However, HMM are generative models and are not adequate for solving this equation and, therefore, we apply Bayes' Theorem to Equation 2.1, ending up with:

$$W* = \arg\max_{W} \frac{P(O|W)P(W)}{P(O)} \tag{2.2}$$

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. In order to understand how generative models are used for data analysis, a mental trick is necessary [12]. First, one must assume that the data was generated through a process which obeys statistical regularities. Then, a model is trained over the observable data, in order to generate the data itself; in other words, the model is used to calculate the

probability of generating the available data. Assuming that the observable data follows a regular pattern which represents the underlying hidden states, the models which are estimated can be said to encode the information of such hidden states. The estimated models are thus used determine what the state sequence that generates a certain sequence of outputs with the highest probability is.

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:

$$W* = \arg\max_{W} P(O|W)P(W) \tag{2.3}$$

$P(O|W)$, 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.

2.2.3 Architecture of an Automatic Speech Recognition

ASRs can be grouped into three categories that take into account the complexity of the task they perform: (i) isolated-word recognition; (ii) command and control systems, in other words, speech recognition systems which are able to recognise pre-defined sentences or expressions; and (iii) large vocabulary continuous speech recognition [34].

Isolated-word recognition are often used by call centers, in Interactive Voice Response (IVR), with their well-known voice menus: “For recent orders say ’order’; for technical support say ’support’ (...)”.

The second type of ASR is more robust, being able to recognise pre-defined commands or sentences, which are usually represented through grammars, such as a Context Free Grammar (CFG). An example of application of this type involves the use of voice commands in computers, mobile phones and hands-free systems in cars. In those systems, sentences such as “turn on the radio” or “what is the closest Starbucks” are said and the command is recognised.

Large vocabulary continuous speech recognition comprises the most complex type of ASR, being capable of processing users’ spontaneous speech. Large vocabulary continuous speech recognition is present in a wide range of applications, such as voice search, personal assistants, dictation systems, domotic, Computer Assisted Pronunciation Training, etc. These systems usually have three modules: (i) a language model, (ii) an acoustic model and (iii) a pronunciation model or dictionary.

The language model is used to estimate the a priori probability of the sequence of the words. The acoustic model, in its turn, is used to calculate the likelihood of the observation. Finally, the pronunciation model plays the role of a bridge between the language and the acoustic models, as it possesses the words which comprise the lexicon of the recogniser and their corresponding phonetic forms. Figure 2.9 illustrates the basic architecture of an ASR system.

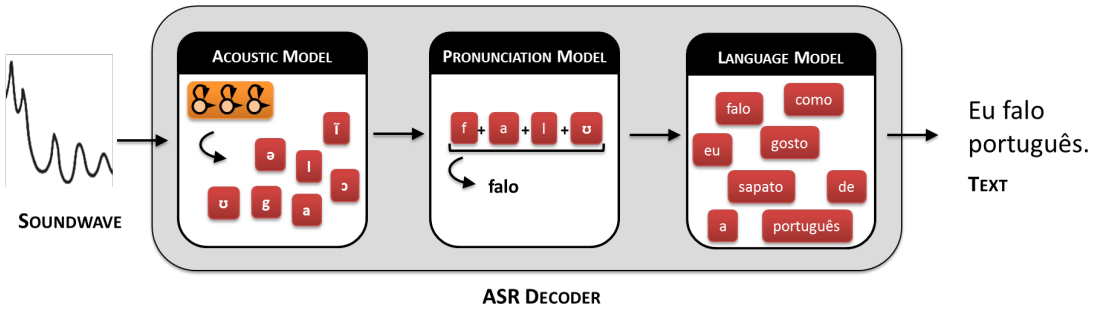


Fig. 2.9 Architecture of a continuous speech recognition system.

The acoustic model processes the acoustic signal of the speech. This model does so in order for it to infer what sound segments comprise the speech, usually by means of using phones or triphones. In HMM-based recognisers, the task of processing the acoustc signal is carried out estimating the most likely observed acoustic states as well as their transition probabilities. On the other hand, the pronunciation model provides us with the correspondence between phones and words sequences in the language. In the example, such model maps on the sequence of phones [falʊ] in the word "falo". The language model, in its turn, estimates the ordering of the most likely words in the language.

2.2.4 Hidden Markov Models in Speech Recognition

HMMs are the most widespread and succesful paradigm in ASR. When HMMs 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 the next prominent paradigm in ASR. This is so due to HMM having been applied to ASR since the late 70's, and having gathered the best results until recently.

An 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 [12]. Each state has a probability distribution over the possible output tokens, in such a way that the output generated by the HMM states provides some information about the hidden sequence of states which was traversed. 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 $Q = \{q_1, q_2, q_3, ..., q_N\}$ is a set of N states. $O = \{o_1, o_2, o_3, ..., o_T\}$ 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, in that

$$\pi_i = Pr(q_i), t = 0 \tag{2.4}$$

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

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

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

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

where γs 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^{th} component and $\mathcal{N}(\cdot; \mu_{jsm}, \Sigma_{jsm})$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ, that is

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

where n is the dimensionality of o.

The following constraints apply:

$$a_{ij} \geq 0 \tag{2.8}$$

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

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

2.2.5 Feature Extraction

Feature extraction is an importart part of speech recognition systems. The feature extraction phase is responsible for identifying or enhancing the components of the signal that are relevant for recognising speech sounds, while discarding or diminishing the effect of unuseful information, such as background noise. With respect to speech parameterisation, Mel Frequency Cepstral Coefficients (MFCC) are definitely the standard. MFCC have been widely used in ASR systems for almost three decades [9], and 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 noteworthy to provide some background information 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, pyezoeletric, laser, etc.), their design remain basically the same as the carbon microphone invented by David Hughes two centuries ago [36]. 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 based on electromagnetic induction (a.k.a dynamic microphones).

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, μ-law, A-law PCM, APCM, DPCM, DM, and ADPCM [19]. The details of each type of speech coder is beyond the scope of this dissertation.

For our purposes, Linear PCM is the only relevant one, since it is the standard way of storing audios in digital format. Pulse Code Modulation (PCM) is a type of analogue-to-digital conversion, which constitutes the basis of the WAV digital audio format, together with other lossless formats, such as AIF and AU. [4]. 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 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 analogue speech signal.

[4] 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

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

$$|x[n]| \leq X_{max} \tag{2.10}$$

and that the quantisation step Δ is uniform for all consecutive levels of x_i

$$x_i - x_{i-1} = \Delta \tag{2.11}$$

Assuming a binary code, the number of levels which can be represented by PCM is $N = 2^B$, where B is the bit depth, which indicates the audio resolution. According to [19], 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 make use of 16-bit linear PCM, whereas DVD-Audio and Blu-Ray discs can support up to 24 bits.

Although linear PCM files are able to carry all the necessary auditory information – after all, we are able to listen to them and recognise 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 [38]. Despite being composed by the same pure tones, complex waves can be completely distinct from one another in terms of their waveform, due to phase shifts, also known as phase offsets. In-phase and out-of-phase waves (Figure 2.10 and Figure 2.11) are represented differently, and this adds much variability to the waveform, in such a way that the signal waveform becomes unsuitable for human analysis and consequently for being used as a raw input in ASR systems.

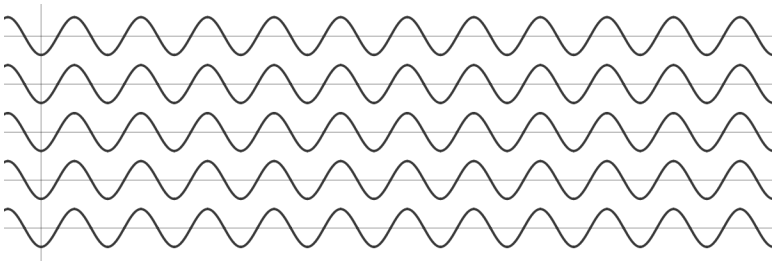


Fig. 2.10 Example of in-phase waves.

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, the window size is usually defined as 25 ms, with a frame shift of 10 ms, the audio information

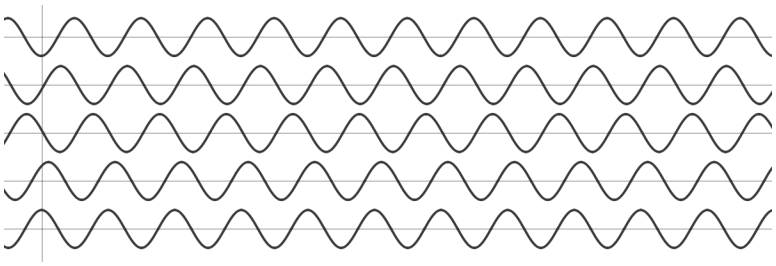


Fig. 2.11 Example of out-of-phase waves.

is extracted every 10 ms with 15 ms of overlapping among adjacent frames [19]. Figure 2.12 contains an example of a windowing process (in this case, with 50% overlapping).

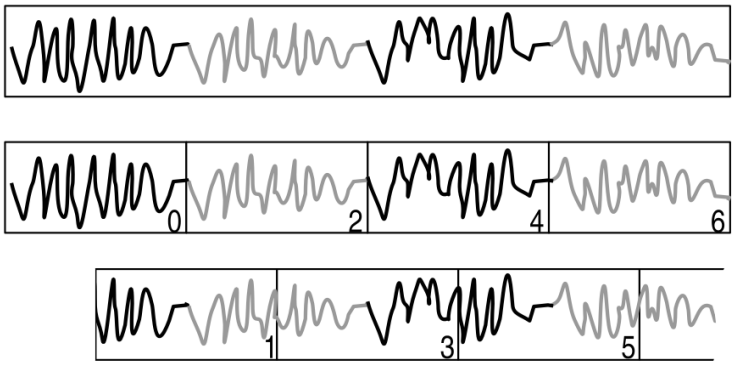


Fig. 2.12 Illustration of an original audio recording (the upper waveform) divided into two offset sequences of analysis windows (two lower waveforms) with 50% overlapping frames [23]

These windows values are based on two assuptions: (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 within this window.

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 it can be noticed in Figure 2.12, as the frame shift is smaller than the window size, the windowing process extracts many redundant information.

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 what the amplitude of the pure tones which form this complex wave is.

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 [9] and the Perceptual Linear Prediction (PLP) [18]. Both parameterisation 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 [29].

2.2.6 MFCC Features

MFCC is a type of speech parameterisation is the result of a cosine transform of the logarithm of the short-term energy spectrum expressed over a mel scale [9]. MFCC features tries to reduce the feature dimensionality of a sound Fourier spectrum, by applying some concepts of Psychoacoustics and Psychophysics in order to the 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 [9].

From Psychoacoustics, MFCCs use the notion that humans do not perceive frequency through a linear scale, but through a scale which seems to be linear-spaced in frequencies below 1000 Hz and logarithmic in frequecies above 1000 Hz[5], the so-called mel scale (named after *mel*ody). 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 [19]. The reference point between a mel scale and a linear frequency scale is 1000 mels, which corresponds to a 1000 Hz tone, 40 dB above the absolute threshold of hearing. Since it was first introduced by Stevens et al. [44], the scale has been revisited many times [45], but a common formulation, according to Huang et al. [19] is:

$$M(f) = 1125 * ln(1 + f/700) \tag{2.12}$$

where f is the input frequency in Hz. For better visualization, the scale is plotted in Figure 2.13.

[5]This is not entirely true. As shown by Umesh et al. [45], 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 prevails.

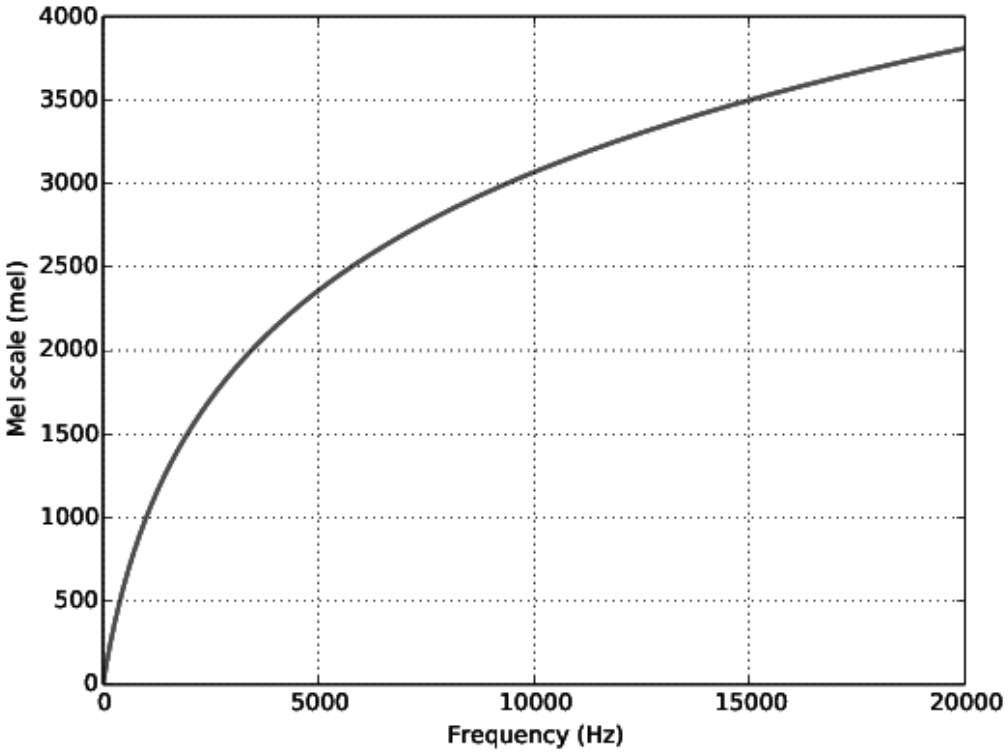


Fig. 2.13 Mel scale versus a linear frequency scale.

2.2.7 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 degrate 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 analyse 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 recognise John's speech in an application, but 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, as it is undesirable for the given purpose.

Chapter 3

Copy of the articles

This chapter presents a copy of all papers which were originally produced within the period of the Master's course. All papers are based on the premise that linguistic knowledge can be used to push the state-of-the-art of NLP tasks one step further. Since our focus was on speech and on pronunciation training, this means phonetic knowledge.

- **Section 3.1:** 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

- **Section 3.2:** Mendonça, G., Candeias, S., Perdigão, F., Shulby, C., Toniazzo, R., Klautau, A., and Aluísio, S. (2014). A method for the extraction of phonetically-rich triphone sentences. *Proceedings of ITS 2014 – International Telecommunications Symposium*

- **Section 3.3:** Mendonça, G., Avanço, L., Duran, M., Fonseca, E., Volpe-Nunes, M., and Aluísio, S. (2016). Evaluating phonetic spellers for user-generated content in Brazilian Portuguese. *PROPOR 2016 – International Conference on the Computational Processing of Portuguese (submitted)*

- **Section 3.4:** Mendonça, G. and Aluísio, S. (2016). Listener: A prototype system for automatic speech recognition and evaluation of brazilian-accented english. *Journal of the Brazilian Computer Society (submitted)*

Using a hybrid approach to build a pronunciation dictionary for Brazilian Portuguese

Gustavo Mendonça, Sandra Aluisio

Instituto de Ciências Matemáticas e de Computação
University of São Paulo, Brazil
gustavom@icmc.usp.br, sandra@icmc.usp.br

Abstract

This paper 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.

Index Terms: pronunciation dictionary, grapheme to phoneme conversion, text to speech

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 the vision-impaired, mobile phones and many other applications [1]. 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 6th most spoken language in the world [2], with about 200 million speakers, speech recognition and speech synthesis for Brazilian Portuguese are far from the current state of the art [3]. 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 [4][5][6][7]. 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 23rd 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 [8]. 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][11][12]. 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 [13]. In fact, many phonological processes in Brazilian Portuguese are related to or conditioned by syllable structure and stress position [14]. Vowel harmony occurs in pretonic context [15], posttonic syllables show a limited vowel inventory [13], nasalization occurs when stress syllables are followed by nasal consonants [16], epenthesis' processes are triggered by the occurrence of non-allowed consonants in coda position [17] 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 [18] [19] [3] [20]. For this reason, we implemented our own syllabification algorithm, based directly on the rules of the last Portuguese Language Orthographic Agreement [21].

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

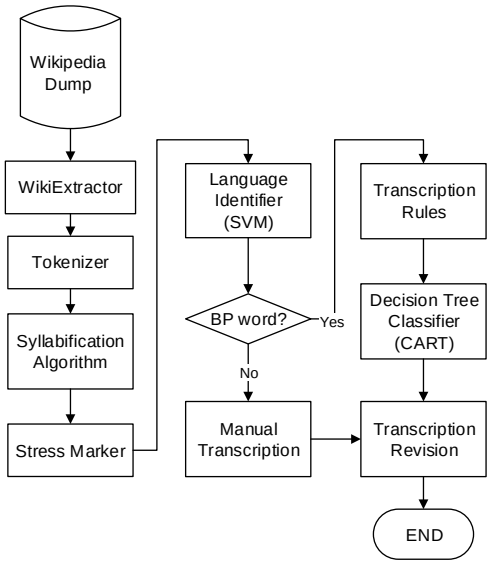


Figure 1: *System architecture for building the pronunciation dictionary.*

the transcription of the remaining graphemes (mostly vowels). All the data were subsequently revised. Figure 1 summarizes the method.

2. Method

2.1. Primary Word List

We used the Portuguese Wikipedia's dump of 23rd January 2014 as the primary word list for the pronunciation dictionary. In order to obtain plain text from the articles, we employed WikiExtractor [22]; 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.

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 [23] and was implemented in Python, through Scikit-learn [24]. 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 [25]. 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[1]. 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.

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 [21]. 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:

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

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.

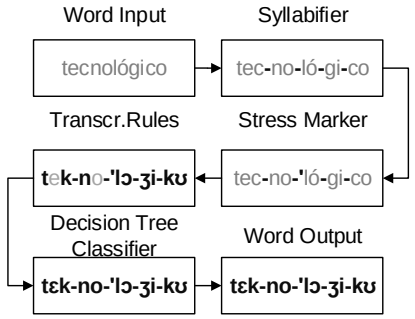


Figure 2: *Example of the transcription procedure – in grey: graphemes yet to be transcribed; in black: graphemes already transcribed.*

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

[1] http://nilc.icmc.usp.br/listener/aeiouado

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" ['tɛ.tʊ] and the latter with a closed one ['ge.tʊ]. 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 [24].

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. 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 1: *Portuguese Wikipedia Summary – Dumped on 23rd January 2014.*

	Word Tokens	**Word Types**
Wikipedia	168,823,100	9,688,039
Selected	151,911,350	238,012
% Used	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 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
Avg/Total	0.86	0.86	0.86	200,000

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

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

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 [26], 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],

Table 4: *Results from the Transcriber – Training Phase.*

	Precision	Recall	F1-score	Support
syl. boundary	1.00	1.00	1.00	9099
stress	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
ɲ	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
ɾ	1.00	1.00	1.00	1388
h	0.98	0.99	0.99	737
fi	0.97	0.92	0.94	169
w	0.97	0.98	0.97	441
w̃	0.98	0.99	0.99	309
j	0.97	0.95	0.96	223
j̃	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
ã	1.00	1.00	1.00	719
ẽ	0.96	0.97	0.97	497
ĩ	0.99	0.99	0.99	274
õ	0.97	0.96	0.97	299
ũ	0.94	0.92	0.93	64
Avg/Total	0.98	0.98	0.98	39934

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].

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.

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.

6. References

[1] R. Godwin-Jones, "Emerging technologies: Speech tools and technologies," *Language Learning and Technology*, vol. 13-3, pp. 4–11, 2009.

[2] F. Lewis, M. Gary and D. Charles, *Ethnologue: Languages of the World, Seventeenth edition*, ser. Seventeenth edition. Dallas, Texas: SIL International, 2013. [Online]. Available: http://www.ethnologue.com

[3] N. Neto, C. Patrick, A. Klautau, and I. Trancoso, "Free tools and resources for brazilian portuguese speech recognition," *Journal of the Brazilian Computer Society*, vol. 17, no. 1, pp. 53–68, 2011.

[4] R. I. Damper, Y. Marchand, M. Adamson, and K. Gustafson, "Comparative evaluation of letter-to-sound conversion techniques for english text-to-speech synthesis," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[5] T. Polyakova and A. Bonafonte, "Learning from errors in grapheme-to-phoneme conversion." in *INTERSPEECH*, 2006.

[6] A. Teixeira, C. Oliveira, and L. Moutinho, "On the use of machine learning and syllable information in european portuguese grapheme-phone conversion," in *Computational Processing of the Portuguese Language*. Springer, 2006, pp. 212–215.

[7] A. Veiga, S. Candeias, and F. Perdigão, "Developing a hybrid grapheme to phoneme converter for european portuguese," vol. 1, pp. 297–300, May 2013.

[8] H. Bussmann, G. Trauth, K. Kazzazi, and H. Bussmann, *Routledge dictionary of language and linguistics / Hadumod Bussmann ; translated and edited by Gregory Trauth and Kerstin Kazzazi*. Routledge, London ; New York :, 1996.

[9] S. Bergsma, P. McNamee, M. Bagdouri, C. Fink, and T. Wilson, "Language identification for creating language-specific twitter collections," in *Proceedings of the Second Workshop on Language in Social Media*, ser. LSM '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 65–74. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390374.2390382

[10] E. B. Bilcu and J. Astola, "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*. IEEE, 2006, pp. 253–258.

[11] D. Trieschnigg, D. Hiemstra, M. Theune, F. de Jong, and T. Meder, "An exploration of language identification techniques for the dutch folktale database," in *Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage, LREC 2012*, P. Osenova, S. Piperidis, M. Slavcheva, and C. Vertan, Eds. Istanbul, Turkey: LREC organization, May 2012, pp. 47–51. [Online]. Available: http://doc.utwente.nl/82013/

[12] M. Zampieri, B. G. Gebre, and H. Nijmegen, "Automatic identification of language varieties: The case of portuguese," in *Proceedings of KONVENS*, 2012, pp. 233–237.

[13] T. C. Silva, *Fonética e fonologia do português: roteiro de estudos e guia de exercícios*. Contexto, 2005.

[14] C. Girelli, *Brazilian Portuguese Syllable Structure*. UMI, 1990. [Online]. Available: http://books.google.com.br/books?id=KRGMnQEACAAJ

[15] L. Bisol, "Vowel harmony: a variable rule in brazilian portuguese," *Language Variation and change*, vol. 1, pp. 185–198, 1989.

[16] A. Quicoli, "Harmony, lowering and nasalization in brazilian portuguese," *Lngua*, vol. 80, pp. 295–331, 1990.

[17] F. Delatorre and R. Koerich, "Production of epenthesis in ed-endings by brazilian efl learners," *Proceedings of the II Academic Forum*, p. 8, 2005.

[18] C. Oliveira, L. C. Moutinho, and A. J. S. Teixeira, "On european portuguese automatic syllabification." in *Proceedings of the Interspeech 2005*. ISCA, 2005, pp. 2933–2936. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2005/i05_2933.html

[19] V. Vasilévski, "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*. Avignon, France: Association for Computational Linguistics, April 2012, pp. 51–60. [Online]. Available: http://www.aclweb.org/anthology/W12-0912

[20] W. Rocha and N. Neto, "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*, 2013, pp. 108–115.

[21] Brasil, *Acordo ortogrfico da lngua portuguesa, de 14, 15 e 16 de dezembro de 1990*. Braslia, Brazil: Dirio do Congresso Nacional da Repblica Federativa do Brasil, Poder Executivo, 2009.

[22] Medialab, "Wikipedia extractor," http://medialab.di.unipi.it/wiki/Wikipedia_Extractor, 2013.

[23] I. Steinwart and A. Christmann, *Support vector machines*. Springer, 2008.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] I. Alves, *Neologismo: Criação lexical*, ser. Princípios (São Paulo). Editora Atica, 2001. [Online]. Available: http://books.google.com.br/books?id=7fIuAAAAYAAJ

[26] H. Weide, "The cmu pronouncing dictionary," 1998. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Evaluating phonetic spellers for user-generated content in Brazilian Portuguese

Gustavo Mendonça, Lucas Avanço, Magali Duran, Erick Fonseca,
Maria das Graças Nunes, and Sandra Aluisio

Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
São Carlos, Brazil
{gustavoauma@gmail.com, avanco89@gmail.com,
magali.duran@uol.com.br, erickfonseca@gmail.com,
gracan@icmc.usp.br, sandra@icmc.usp.br}

Abstract. 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 paper 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%).

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 tiny keyboards, one can miss the keystrokes, hitting adjacent keys on the keyboard, thus spell checking has regained attention [1].

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 [2], confusion sets [3], improvements on the trigram-based noisy-channel model [4] and [5], use of GoogleWeb 1T 3-gram data set and a normalized and modified version of the Longest Common Subsequence string matching algorithm [6], a graph-based method using contextual and PoS features and the double metaphone algorithm to represent phonetic similarity [7]. 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 [8].

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 [9], for example, the authors use a linear combination of two measures – the Levenshtein distance between two strings and the Levenshtein distance between the Soundex [10] code of two strings. 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. 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 [11–14].

In [14] 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 paper presents two new spell checking methods for UGC in BP. The first of them deals with phonetically motivated errors, a recurrent problem in UCG 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[1]. This paper is structured as follows. In Section 2 we describe our methods, the setup of the experiments and the corpus we compiled. In Section 3 we present the results. In Section 4 we discuss related work on spelling correction of phonetic and real-word errors. To conclude, the final remarks are outlined in Section 5.

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 evaluate BP spellers, as well as the evaluation metrics.

[1] The small benchmark of 120 tokens used in [15] and [16] is not representative of our scenario.

2.1 Method I - Baseline

We use as a baseline the open source Java Spelling Checking Package, JaSpell[2]. 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 [17] for the English language. At the time this speller was developed there was no version of these rules for the Portuguese language[3].

2.2 Method II - Benchmark

The method presented in [18] 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 [19] 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[4] and the frequency list was taken from Corpus Brasileiro[5].

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 ['pɛh.tʊ], and "forte" to ['fɔh.tʃɪ].

[2] http://jaspell.sourceforge.net/

[3] Currently, a BP version of the phonetic rules can be found at http://sourceforge.net/projects/metaphoneptbr/

[4] http://www.nilc.icmc.usp.br/nilc/projects/unitex-pb/web/

[5] http://corpusbrasileiro.pucsp.br/cb/

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 [20] for this purpose, since it consists of the state of the art in grapheme-to-phoneme transcription for Brazilian Portuguese.

The usage of the grapheme-to-phoneme converter is a bit different from a simple pipeline. According to Toutanova [21], 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: [duˈrɛks] ∼ [duˈrɛkɪs]. 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 [duˈrɛ.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.

The only difference of GPM in comparison with Method II lies in the G2P transcription match, which takes place prior to Soundex. 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.

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 1.

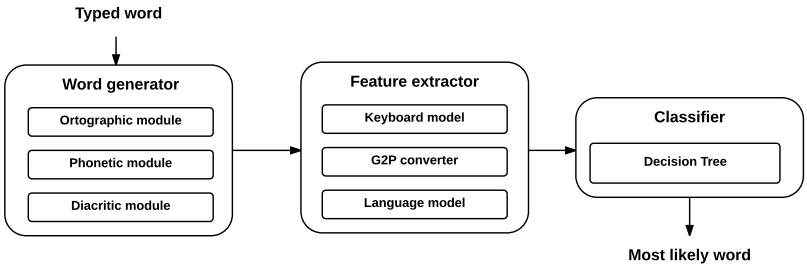


Fig. 1. *Architecture of the GPM-ML*

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. 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)[6].

As for phonetic similarity, the Aeiouado's grapheme-to-phoneme converter [20] 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 (ˆ), grave (`) and tilde (˜). The diacritics often indicate different vowel quality, timbre or stress. However, these symbols are rarely used in UGC, and the reader uses the 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, as described in Table 1.

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 2.5, with the features we discussed. The classifier was implemented through scikit-learn [22] 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.

[6] The dictionary is available upon request.

Table 1. *List of features*

Feature	Description
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. PHONELEVDIST	the levenshtein distance between of the phonetic transcription of the typed word and of the suggestion.

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[7], a Brazilian price comparison search engine. All misspelled words were marked, the correct expected form was suggested and the misspelling category was indicated. We 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 [19] 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 2.

Table 2. *Inter-rater agreement for the error detection task*

		Annot. B		
		Correct	**Wrong**	**Total**
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 [23]. The κ value for the error detection task was 0.915 which stands for good reliability or almost perfect agreement [24]. 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:

[7] 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” (*excelent*), 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 ”organização” (*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 3. The difference between the total number of counts in Table 2 and 3 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[8] and intends to be a benchmark for future research in spelling correction for user generated content in BP.

Table 3. *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
Total/Avg		4,083	100.0

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

[8] Link omitted for blind review.

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 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.

3 Discussion

In Table 4, 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. *Comparison of the Methods*

	Rate			
Method	**Detection**	**Correction**	**FP**	**FN**
Baseline JaSpell	74.0%	44.7%	5.9%	26.0%
Benchmark Rules&Soundex	83.4%	68.6%	1.7%	**16.6%**
GPM	83.4%	68.2%	1.7%	**16.6%**
GPM-ML	**84.9%**	**78.1%**	**0.7%**	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. In the following we discuss the correction hit rates by misspelling categories. Table 5 presents a comparison among all methods.

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 misspelling 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,

Table 5. *Comparison of Correction Rates*

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

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 [18], 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 Related Work

The first approaches to spelling correction date back to Damerau [25] 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 [26]. 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 [21] who included information on pronunciation to the model of typographical errors correction. [21] and also [27] 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 [21] 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.

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.

References

1. Duan, H., Hsu, B.P.: Online Spelling Correction for Query Completion. In: Proceedings of the 20th International Conference on World Wide Web. WWW '11, NY, USA, ACM (2011) 117–126
2. Fossati, D., Di Eugenio, B.: A Mixed Trigrams Approach for Context Sensitive Spell Checking. In Gelbukh, A., ed.: Computational Linguistics and Intelligent Text Processing. Volume 4394 of Lecture Notes in Computer Science., Springer (2007) 623–633
3. Fossati, D., 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
4. Mays, E., Damerau, F.J., Mercer, R.L.: Context based spelling correction. Information Processing & Management **27**(5) (1991) 517–522
5. Wilcox-O'Hearn, A., Hirst, G., Budanitsky, A.: 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 (2008) 605–616
6. Islam, A., Inkpen, D.: Real-word spelling correction using Google web 1tn-gram data set. In: In ACM International Conference on Information and Knowledge Management CIKM 2009. (2009) 1689–1692
7. Sonmez, C., Ozgur, A.: A Graph-based Approach for Contextual Text Normalization. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing EMNLP 2014. (2014) 313 – 324
8. Hirst, G.: An evaluation of the contextual spelling checker of Microsoft Office Word 2007 (2008)
9. Zampieri, M., Amorim, R.: 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. (2014) 438–449
10. Rusell, R.C.: US Patent 1261167 issued 1918-04-02. (1918)
11. Duran, M., Avanço, L., Aluísio, S., Pardo, T., Nunes, M.G.V.: Some Issues on the Normalization of a Corpus of Products Reviews in Portuguese. In: Proceedings of the 9th Web as Corpus Workshop WaC-9, Gothenburg, Sweden (April 2014) 22–28
12. De Clercq, O., Schulz, S., Desmet, B., Lefever, E., Hoste, V.: Normalization of Dutch User-Generated Content. In: Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013. (2013) 179–188
13. Han, B., Cook, P., Baldwin, T.: Lexical Normalization for Social Media Text. ACM Trans. Intelligent System Technology **4**(1) (February 2013) 5:1–5:27
14. Andrade, G., Teixeira, F., Xavier, C., Oliveira, R., Rocha, L., Evsukoff, A.: HASCH: High Performance Automatic Spell Checker for Portuguese Texts from the Web. Proceedings of the International Conference on Computational Science **9**(0) (2012) 403 – 411
15. Martins, B., Silva, M.J.: 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. (2004) 372–383
16. Ahmed, F., Luca, E.W.D., Nürnberger, A.: Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness. Polibits (12 2009) 39–48
17. Philips, L.: The double metaphone search algorithm. C/C++ Users Journal **18**(6) (2000)
18. Avanço, L., Duran, M., Nunes, M.G.V.: Towards a Phonetic Brazilian Portuguese Spell Checker. In: Proceedings of ToRPorEsp Workshop PROPOR 2014, São Carlos, Brazil (2014) 24–31

19. Hartmann, N., Avanço, L., Balage, P., Duran, M., Nunes, M.G.V., Pardo, T., Aluísio, S.: 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. (2014) 3866–3871
20. Mendonça, G., Aluísio, S.: 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 (2014)
21. Toutanova, K., Moore, R.C.: Pronunciation Modeling for Improved Spelling Correction. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02 (2002) 144–151
22. 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., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12** (2011) 2825–2830
23. Carletta, J.: Assessing Agreement on Classification Tasks: The Kappa Statistic. Computational Linguistics **22**(2) (June 1996) 249–254
24. Landis, J.R., Koch, G.G.: The Measurement of Observer Agreement for Categorical Data. Biometrics **33**(1) (1977) pp. 159–174
25. Damerau, F.J.: A Technique for Computer Detection and Correction of Spelling Errors. Communications of ACM **7**(3) (mar 1964) 171–176
26. Brill, E., Moore, R.C.: An Improved Error Model for Noisy Channel Spelling Correction. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. ACL '00 (2000) 286–293
27. van Berkel, B., Smedt, K.D.: Triphone Analysis: A Combined Method for the Correction of Orthographical and Typographical Errors. In: Proceedings of the Second Conference on Applied Natural Language Processing, Austin, Texas, USA (February 1988) 77–83

A Method for the Extraction of Phonetically-Rich Triphone Sentences

Gustavo Mendonça*, Sara Candeias†‡, Fernando Perdigão†, Christopher Shulby*,
Rean Toniazzo§, Aldebaro Klautau¶ and Sandra Aluísio*
*Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo – São Carlos, Brazil.
†Instituto de Telecomunicações,
Universidade de Coimbra – Coimbra, Portugal.
‡Microsoft Language Development Center – Lisbon, Portugal.
§Departamento de Engenharia de Materiais,
Universidade Federal de São Carlos – São Carlos, Brazil.
¶Laboratório de Processamento de Sinais,
Universidade Federal do Pará – Belém, Brazil.
Email: gustavom@icmc.usp.br, saracandeias@co.it.pt, fp@co.it.pt, chrisshulby@gmail.com,
reantoniazzo@gmail.com, a.klautau@ieee.org, sandra@icmc.usp.br

***Abstract*—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, segmented 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.**

I. Introduction

In what regards to speech technology, although there are some studies which employ words [1], syllables [2] and monophones [3] to develop Automatic Speech Recognition (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 [4]. 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 [5]. Laboratory phonologists are also interested in such corpora in order to develop prompts for analyzing speech production and variability [6].

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 to 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 [7].

We evaluate the proposed method in building a phonetically-rich triphone sentences corpus for Brazilian Portuguese. The sentences come from the Portuguese Wikipedia dump [8], which was converted into plain text, segmented 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 II, we briefly describe the related work available in the literature. In Section III, we describe the method proposed. In Section IV, we evaluate it by building a phonetically-rich corpus for Brazilian Portuguese. The final remarks are outlined in Section V.

II. 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 Mandarim, in which tones have a phonological value, one must even posit units such as tonemes in order to properly describe speech phenomena [9].

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 [10][11][12][13].

It is known that many linguistic phenomena, including triphone sets, show a Zipfian distribution [14]. 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 their 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. [10], 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. [15] considered syllables as the basic unit to extract, in an automatic way, phonetically-rich sentences from a large text corpus from Indian languages, justifying their choice because a syllable is the smallest segment of the uterance. 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. [16] deals specifically with Brazilian Portuguese and proposed a method based on genetic algorithms to select a set of sentences for a speech syntesis 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 Brazilian Portuguese, since the available corpora disregards prosodic features. They evaluated it through the CETENFolha corpus (www.linguateca.pt/cetenfolha/) which has circa 1,5 million sentences in order to gather 4,000 sentences phonetically- and prosodically- rich. Their approach is composed of 4 stages, including grapheme-to-phoneme conversion, prosodic annotation, feature vector representation, and selection. The authors obtained prosodic features based on the pitch, therefore identifying tone events for each syllable (N, H+, H-, H, L, and L-, where H and L stands for high and low, respectively, and N for neutral). Using these features to represent each sentence, they developed a genetic algorithm (GA) to select a subset. Their paper, however, does not discuss how the GA fitness function meets both constraints (phonetic and prosodic).

III. Method

A. Unit of analysis

Contextual phone units are extensively applied to speech technology systems given their ability to encompass allophonic variation and coarticulation effects, specially triphones. A triphone is represented as a sequence $(p_{left}$ - p - $p_{right})$, where p_{left} is the phone which precedes p and p_{right} is the one which follows it. Table I presents a comparison of the word *speech* transcribed using monophones and triphones.

Word	Monophone Form	Triphone Form
speech	[s p i tʃ]	[#-s-p s-p-i p-i-tʃ i-tʃ-#]

TABLE I. A comparison between monophone and triphone transcription.

As one might observe, triphones are capable of describing the surronding environment of a given phone and this has a huge impact in the performance of acoustic models for speech recognition or speech synthesis. Given the above reasons, we chose triphones as the unit of analysis for our algorithm.

B. 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 $S = \{s_1, s_2, s_3, ..., s_n\}$. Each sentence s is formed by m triphones, represented as $T = \{t_1, t_2, t_3, ..., t_m\}$. The *a priori* probability of the triphones can be calculated straightforwadly: 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 occurence. Therefore, we define the phonetic richness of a sentence s as the summation of its triphones' reciprocal probabilities:

$$\varrho(s) = \sum_{i=1}^{m} \frac{1}{P_K(t_i)} \tag{1}$$

C. Algorithm

Our algorithm for extracting rich sentences was implemented in Python and follows a greedy strategy. The distance metric is calculated through the SciPy library [17].

Greedy algorithms have been widely used in Computer Science, when the optimum solution of the problem can not be guaranteed [18]. 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 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 III-B 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. Then 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 described below.

```
Corpus <- List of available sentences
Selected <- [] // List of selected sentences
Metrics <- [] //List made of tuples with sentences
              and euclidean distance values
Ideal <- Ideal corpus, with all equiprobable triphones

while length(Selected) < n do:
  for Sentence in Corpus:
    calculate distance between Sentence+Selected and Ideal
    append Sentence and its metric in the list Metrics
  BestSentence <- select the sentence in the loop with the
                  minimum distance
  append BestSentece to Selected
  clear the Metrics list
end.
```

IV. Example Evaluation

A. Corpus

As a proof-of-concept we evaluated our method by building a phoneticaly-rich corpus for Brazilian Portuguese. The original database of sentences consisted of the Wikipedia dump produced on 23rd January 2014. Table II summarizes the data.

Articles	Word Tokens	Word Types
~820,000	168,823,100	9,688,039

TABLE II. Portuguese Wikipedia Summary – Dumped on 23rd January 2014.

In order to obtain only plain text from Wikipedia articles, we used the software WikiExtractor [19], to strip all of the MediaWiki markups and other metadata. Then, we segmented the output into sentences, by applying the Punkt sentence tokenizer [20]. Punkt is a language-independent tool, which can be trained to tokenize sentences. It is distributed together with NLTK [21], where it already comes with a model for Portuguese, trained on the Floresta Sintá(c)tica Treebank [22].

Following, each sentence was transcribed phonetically by using a pronunciation dictionary for each language variety. We employed the UFPAdic 3.0 [23], developed for Brazilian Portuguese, which contains 38 phones and 64,847 entries. Triphones were generated dynamically, based on the transcription registered in the dictionary. Cross-word triphones were considered in the analysis along with cross-word short pause models. Given its encyclopedic nature, many sentences in Wikipedia present dates, periods, percentages and other numerical information. For this reason, we decided to supplement the dictionary, 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 III.

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

TABLE III. Sentences' summary after WikiExtractor and Punkt.

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 [24]. 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 IV. After that, we applied the heuristic metric described in Section IV-A, and the top 50,000 sentences were selected (= 2,340,237 triphone tokens and 10,237 triphone types).

Total Sentences	Short	Average	Long
1,229,422	15,581	873,546	340,295

TABLE IV. Sentences' summary after the length filter.

B. Discussion

For this example evaluation, we discuss the extraction of 250 phonetically-rich sentences. Table V 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 VI contains the data regarding sentences selected randomly. The list of random sentences derives from the pool of 50,000 sentences described in Section IV-A. Ten different seed states were used in order to ensure randomness, the average of these results are presented.

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 V. 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 VI. TRIPHONE RESULTS FROM THE SENTENCES TAKEN RANDOMLY.

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 (40,9% of all types in the corpus), as opposed to 3335 (32,5%) 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 [25]. Several sentences selected by our algorithm contained foreign words which were registered in the dictionary with abnormal pronunciations, such as *Springsteen* [sprĩgsteẽ], *hill* [iww], *world* [wohwdʒ]. Since no other words are registered with the triphones [e-e+ẽ] or [e-ẽ+#] 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.

V. FINAL REMARKS

We proposed a method for compiling a corpus of phonetically-rich triphone sentences. It was evaluated for Brazilian Portuguese. All sentences considered come from the Portuguese Wikipedia dumps, which were converted into plain text, segmented and transcribed. 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 them, 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. We also plan in applying the method to others corpora, e.g. CETENFolha, in order to make the results comparable with other studies for Brazilian Portuguese, such as Nicodem et al. [16]. All resources developed in this paper are freely available on the web[1].

REFERENCES

[1] R. Thangarajan, A. M. Natarajan, and M. Selvam, "Word and triphone based approaches in continuous speech recognition for Tamil language," *WSEAS Trans. Sig. Proc.*, vol. 4, no. 3, pp. 76–85, 2008.

[2] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, vol. 9, pp. 358–366, 2001.

[3] A. Kumar, M. Dua, and T. Choudhary, "Article: Continuous Hindi speech recognition using monophone based acoustic modeling," *IJCA Proceedings on International Conference on Advances in Computer Engineering and Applications*, vol. ICACEA, no. 1, pp. 15–19, March 2014.

[4] L. Rabiner and R. Schafer, "Introduction to digital speech processing," *Foundations and trends in signal processing*, vol. 1, pp. 1–194, 2007.

[5] A. P. Mendes, A. N. d. Costa, A. D. Martins, A. F. O. Fernandes, S. M. D. d. R. Vicente, and T. C. S. Freitas, "Contributos para a construção de um texto foneticamente equilibrado para o Português-Europeu," *Revista CEFAC*, vol. 14, pp. 910–917, 10 2012.

[6] J. B. P. Pierrehumbert, M. E. Beckman, and D. R. Ladd, "Conceptual foundations of phonology as a laboratory science," in *Phonological knowledge: Conceptual and empirical issues*. Oxford University Press., 2000, pp. 273–304.

[7] R. Sedgewick and P. Flajolet, *An introduction to the analysis of algorithms*. Addison-Wesley-Longman, 2013.

[8] Wikimedia, "Portuguese Wikipedia database dump backup," http://dumps.wikimedia.org/ptwiki/20140123/, 2014.

[1] http://nilc.icmc.usp.br/listener

[9] X. Lei, M. yuh Hwang, and M. Ostendorf, "Incorporating tone-related MLP posteriors in the feature representation for Mandarin ASR," in *In Proc. Eur. Conf. Speech Communication Technology*, 2005, pp. 2981–2984.

[10] M. A. M. Abushariah, R. N. Ainon, R. Zainuddin, M. Elshafei, and O. O. Khalifa, "Phonetically rich and balanced text and speech corpora for Arabic language," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 601–634, 2012.

[11] J.-L. Shen, H.-M. Wang, R.-Y. Lyu, and L.-S. Lee, "Incremental speaker adaptation using phonetically balanced training sentences for Mandarin syllable recognition based on segmental probability models." in *ICSLP*. ISCA, 1994. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/icslp1994.html#ShenWLL94

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus cdrom," 1993.

[13] E. Uraga and C. Gamboa, "VOXMEX speech database: Design of a phonetically balanced corpus," in *LREC*. European Language Resources Association, 2004.

[14] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[15] K. Arora, S. Arora, K. Verma, and S. S. Agrawal, "Automatic extraction of phonetically rich sentences from large text corpus of Indian languages." *INTERSPEECH*, 2004.

[16] M. Nicodem, I. Seara, R. Seara, D. Anjos, and R. Seara-Jr, "Seleção automática de corpus de texto para sistemas de síntese de fala," *XXV Simpósio Brasileiro de Telecomunicações - SBrT 2007*, 2007.

[17] E. Jones, T. Oliphant, and P. Peterson, "SciPy: Open source scientific tools for Python," http://www.scipy.org/, 2014.

[18] B. Coppin, "Inteligência artificial," *Rio de Janeiro: LTC*, 2010.

[19] Medialab, "Wikipedia extractor," http://medialab.di.unipi.it/wiki, 2013.

[20] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection." *Computational Linguistics*, vol. 32, no. 4, pp. 485–525, 2006.

[21] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

[22] C. Freitas, P. Rocha, and E. Bick, "Floresta sintá(c)tica: Bigger, thicker and easier," in *Computational Processing of the Portuguese Language*. Springer, 2008, pp. 216–219.

[23] N. Neto, C. Patrick, A. Klautau, and I. Trancoso, "Free tools and resources for Brazilian Portuguese speech recognition," *Journal of the Brazilian Computer Society*, vol. 17, no. 1, pp. 53–68, 2011.

[24] M. Watanabe and R. Rose, "Pausology and hesitation phenomena in second language acquisition," *The Routledge Encyclopedia of Second Language Acquisition*, pp. 480–483, 2012.

[25] J. Steigner and M. Schrder, "Cross-language phonemisation in German text-to-speech synthesis." in *INTERSPEECH 2007*. ISCA, 2007, pp. 1913–1916.

APPENDIX: EXAMPLES OF THE EXTRACTED SENTENCES

Number of Sentences: 100; **Number of Triphone Types:** 2307; **Number of Triphone 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.
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 Zezé.
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.

RESEARCH

Listener: A prototype system for automatic speech recognition and evaluation of Brazilian-accented English

Gustavo A Mendonça* and Sandra M Aluisio

Abstract

Recent surveys have shown that Brazil is among the countries with the lowest knowledge of the English language. Considering that English is the current lingua franca, any effort that seeks to improve such knowledge is worthwhile. This paper presents a project aimed at being a primary step towards developing a pronunciation assessment system for Brazilian-accented English. Lately, with increased computing power and improved speech recognition, research on automated pronunciation error detection and pronunciation assessment has gained attention. The prototype developed herein, called Listener, is based on speech recognition and comprises three modules: (i) a pronunciation model; (ii) an acoustic model; and (iii) a context free grammar. Pronunciation variants were added manually to a lexicon of 1,841 words through transcription rules. The acoustic model was trained on a combined model, which is fed with data from English, Portuguese and Brazilian-accented English. Context free grammars were used to allow forced-alignment. The prototype achieved a true positive rate of 0.90 on detecting mispronunciations in clean speech. However, in noisy data, the performance was severely damaged, being as low as 0.57 for isolated words or 0.31 for words+sentences.

Keywords: pronunciation training; non-native speech recognition; natural language processing

*Correspondence: gustavoauma@gmail.com
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, Brazil
Full list of author information is available at the end of the article

1 Introduction

According to the International Monetary Found (IMF) [1], in 2015, Brazil was the seventh largest economy in the world with a GDP of US$ 2.34 trillions. A survey by The Economist (2013) says that, since 2009, the growth of BRICS accounts for 55% of the entire world economy growth. The current economic scenario is extremely favourable for Brazil to increase its global influence; however with regard to the ability to communicate globally, Brazil occupies a much more modest position.

In 2015, Brazil ranked 41st out of 70 countries in the English Proficiency Index (EF-EPI) [2], classified among countries with low English proficiency, with 51.05 points. Scandinavian countries led the very high proficiency rankings, with Sweden (70.94) in the first position, Denmark (70.05) in third the spot and Norway (67.83) in fourth. Brazil performance was close to several other Latin America countries, such as Peru (52.46), Chile (51.88), Ecuador (51.67), Uruguay (50.25) and Colombia (46.54). The only exception in Latin America was Argentina that, despite the recent great depression, was ranked 15th, being classified as high proficiency, with a score of 60.26.

The EF-EPI bands are aligned to the Common European Framework of Reference for Languages (CEFR) in the following way: 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 level B1, that describes an independent language user with the intermediate communication skills (see Table 1.

As one might notice, the B1 level describes an individual who is usually able to understand familiar matters, deal with traveling situations, describe personal experiences and plans, and produce simple texts about subjects of personal interest. Needless to say, this is a very restricted communicative competence, which limits English usage primarily to the personal domain.

With respect of Business English proficiency, Brazil performance is even more concerning. On the Business

Table 1 CEFR reference level description for B1.

#	Communication skills
1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc.
2	Can deal with most situations likely to arise while traveling in an area where the language is spoken.
3	Can produce simple connected text on topics that are familiar or of personal interest.
4	Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.

English Index (BEI) of 2013 [3], Brazil reached the 71st position out of 77 countries analyzed. We attained a score of 3.27 points, in a scale from 1 to 10, being placed at the "Beginner" range, the lowest range considered by the index. We were close to countries such as El Salvador (3.24), Saudi Arabia (3.14) and Honduras (2.92) which up until recently had experienced civil wars or dictatorship governments. BEI describes individuals at the beginner level as those who "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 task in a company or industry work environment.

Given this scenario, it is clear that we urgently need to improve English language proficiency among Brazilians. This project seeks to be an initial step towards this direction. We developed a prototype system for automatic speech recognition and evaluation of Brazilian-accented English, called *Listener*, which is capable of recognizing utterances in Brazilian-accented English and identifying which are the mispronunciations. The system was designed especially for graduate students who want to improve their pronunciation skills. The interest in this group is due to the fact that graduate students often need to speak English in order to publicize their research, for instance, by attending conferences or giving lectures. The system is based on an Automatic Speech Recognition system which makes use of forced alignment, *HMM/GMM* acoustic models, context free grammars and multipronunciation dictionaries[1].

[1] All files, resources and scripts developed are available at the project website: http://nilc.icmc.usp.br/listener. Due to copyright reasons, the corpora used for training the acoustic models cannot be made available.

This paper is structured as follows. Section 2 presents some fundamental concepts in speech recognition. Section 3 discusses the materials and methods, with an overview of pronunciation models, acoustic models, as well as context free grammars. Results are shown in Section 4. Finally, in Section 5 we present the overal conclusions, discuss the work limitations and present some future work.

2 Automatic Speech Recognition

Automatic Speech Recognition (ASR) can be defined as the task of converting spoken language into readable text by computers in real-time [4].

Speech is certainly the most natural human way of communication. Allowing people to interact with their gadgets through voice may greatly improve the user-experience, especially in a world which is becoming more and more mobile-oriented. Currently, ASR is present in many widely-used applications, such as personal assistants, speech-to-text processing, domotics, call routing, etc. [5]

All state-of-the-art paradigms in ASR are stochastic and they basically try to solve one single equation, which is called the fundamental equation of *ASR*. It 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:

$$W* = \arg\max_{W} P(W|O) \tag{1}$$

To solve this equation straightforwadly, one would require a discriminative model capable of estimating the the probability of W directly from a set of observations O [6]. If we apply the Bayes' Theorem we obtain the following equivalent equation:

$$W* = \arg\max_{W} \frac{P(O|W)P(W)}{P(O)} \tag{2}$$

which is suitable for a generative model. For a single audio input, the probability of the observable acoustic feature vectors $P(O)$ is a constant and, therefore, might be discarded, in such way that we end up with:

$$W* = \arg\max_{W} P(O|W)P(W) \tag{3}$$

$P(O|W)$ is the conditional 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 or through grammars.

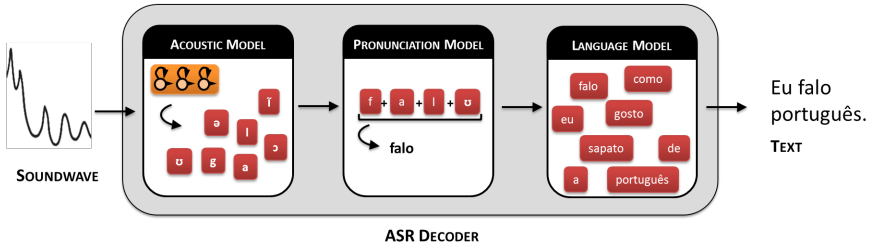


Figure 1 Architecture of a continuous speech recognition system. Note: in the Listener prototype, the language model is replaced by a context free grammar.

Figure 1 illustrates the basic architecture of an ASR system. The acoustic model processes the acoustic signal of the speech. This model does so in order for it to infer what sound segments comprise the speech, usually by means of using phones or triphones. In HMM-based recognisers, the task of processing the acoustc signal is carried out estimating the most likely observed acoustic states as well as their transition probabilities. On the other hand, the pronunciation model provides us with the correspondence between phones and words sequences in the language. In the example, such model maps on the sequence of phones [falU] in the word "falo". The language model, in its turn, estimates the ordering of the most likely words in the language; in the prototype described herein, this step is performed by the grammar, to parse the many pronunciation variants.

3 Materials and Methods

3.1 Architecture of Listener

The prototype of the Listener system has three modules: (i) a pronunciation model; (ii) an acoustic model; and (iii) a context free grammar. In the following subsections each of these modules will be discussed in detail.

3.2 Pronunciation Model

Pronunciation models are lexica with words and their corresponding phonetic transcriptions, according to a given convention. In other words, pronunciation models have the role of linking phones from the acoustic model to words defined in the language model. For speech recognition purposes, phonetic a like ARPAbet or SAMPA are often employed to avoid problems with data formatting or encoding. In ARPAbet, phones are converted into sequences of ASCII characters, in such a way that a word like "speech" ['spiːtʃ] becomes [s p iy1 ch] [7].

For non-native speech recognition, multipronunciation dictionaries are often employed in order to address phenomena of negative-transference from the L1 to L2 [8]. These dictionaries are a type of pronunciation model where pronunciation variants are explicitly added to the lexicon of the ASR [8]. For building the pronunciation model for Listener, the literature on pronunciation training was analyzed and transformation rules were defined based on the most common mispronunciations among Brazilians. The pronunciation model was inspired by several works for pronunciation training, focused on Brazilian-accented English [9–11]. In total, we gathered 13 mispronunciation patterns for Listener, the full list with examples can be found in Table 2.

Table 2 Mispronunciation types selected for the prototype system with examples of the expected pronunciation and the one with negative transfer from L1 to L2.

#	Description	Example	Expect.	Mispron.
1	Initial epenthesis	school	[skuːl]	[iskuːl]
2	Coda epenthesis	dog	[dɑːg]	[dɑːgi]
3	Terminal devoicing	does	[dʌz]	[dʌs]
4	Th-fronting	think	[θɪŋk]	[fɪŋk]
5	Palatalization	teen	[tʰiːn]	[tʃiːn]
6	Deaspiration in plosives	tea	[tʰiː]	[tiː]
7	Vocalization of laterals	well	[wɛl]	[wew]
8	Vocalization of nasals	beam	[biːm]	[bĩ]
9	Velar paragoge	wing	[wɪŋ]	[wɪŋg]
10	Consonantal change	think	[θɪŋk]	[fɪŋk]
11	Vowel change	put	[pʰʊt]	[pʰʌt]
12	General deletion	foot	[fʊt]	[fʊ]
13	General insertion	work	[wɜːrk]	[wɜːrks]

All linguistic contexts described by Zimmer [9] were converted into transcription rules in order to generate the variants for the pronunciation dictionary. The full list of rules can be found in the project's website. A

sample of these rules can be found in Figure 2. These transcription rules are then applied to a base dictionary in order to append it with new pronunciation variants.

Figure 2 Building the pronunciation model. Pseudocode with the rules for generating pronunciation variants (sample).

 # Initial epenthesis
if [s p] in initial position → [iy s p] # sport
if [s t] in initial position → [iy s t] # start
if [s k] in initial position → [iy s k] # skate
if [s m] in initial position → [iy s m] # small
if [s n] in initial position → [iy s n] # snake
...
 # Coda epenthesis
if [p] in final position → [p ih] # stop
if [b] in final position → [b ih] # bob
if [t] in final position → [t ih] # boat
if [d] in final position → [d ih] # and
if [k] in final position → [k ih] # book
if [g] in final position → [g ih] # dog
...
if [m] and ortho ends in <me> → [m ih] # time
if [s] and ortho ends in <ce> → [s ih] # nice
...
 # Th-fronting
if [th] → [f] # think
if [th] → [s] # think
if [th] → [t] # think
...
 # Palatalization
if [t iy] → [ch iy] # teen
if [t ih] → [ch iy] # poetic
...
 # Vocalization of nasal consonants
if [iy m] in final position → [im] # him
if [ae n] in final position → [em] # can
...

It is worth noticing that, in terms of context, there is often overlapping among rules. For instance, in a word like "think" [th ih ng k], there is a rule for converting [th] into [f], another one for converting it into [s], or [t], etc. There are even rules which create context for other ones to apply, for instance, in "boat" [b ow t], if epenthesis takes place, generating [b ow t ih], then [t] could undergo consonantal change/palatalization, thus producing [b ow ch ih].

To make sure that all pronunciation variants are generated, the rules are run inside a while loop, which iterates over each word in the base dictionary generating and adding these new pronunciation to the dictionary; and the loop only stops when there are no new variants.

For the pilot system of Listener, we used the CMUdict [7] as a base dictionary, which contains over 134,000 entries and their pronunciations in American English. However, in the test sets for Listener there are 1,841 unique words, therefore only these were considered in this experiment. These transcription rules were run over these 1,841 unique words and 8,457 new pronunciation variants were generated (totalling 10,298 entries in the final dictionary). In such a way, the average pronunciation per word is 5.6.

3.3 Acoustic Model

Acoustic Models (AM) are used within speech recognition to map the acoustic parameters of into phonemes. AMs are estimated through supervised training over a transcribed speech corpus – often with the Forward-Backward algorithm by modeling phones via Hidden Markov Models (HMM) [12]. Markov models are very suitable for the statistical description of symbol and state sequences [13]. 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, the current state does not depend upon the sequence of events that preceded it. Hidden Markov Models (HMM) are just a special type of Markov processes which contain hidden states.

HMMs are the most widespread models used in ASR [14]. They can be formally described as a 5-tuple $\lambda = (Q, O, \Pi, A, B)$. $Q = \{q_1, q_2, q_3, ..., q_N\}$ represents a set of hidden N states. $O = \{o_1, o_2, o_3, ..., o_T\}$ 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, and only the observations o are directly visible. $\Pi = \{\pi_i\}$ is a vector with the initial state probabilities, such that

$$\pi_i = Pr(q_i), t = 0 \tag{4}$$

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

$$a_{ij} = P(q_t = j | q_{t-1} = i), 1 \leq, i, j \leq N \tag{5}$$

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 in ASR; we can define that, for a state j, the probability $b_j(o_t)$ of generating o_t is given by

$$b_j(o_t) = \prod_{s=1}^{S} \left[\sum_{m=1}^{M_{js}} c_{jsm} \mathcal{N}(o_{st}; \mu_{jsm}, \Sigma_{jsm}) \right]^{\gamma_s} \tag{6}$$

where γs 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^{th} component and $\mathcal{N}(\cdot; \mu_{jsm}, \Sigma_{jsm})$ is a multivariate Gaussian with mean vector μ and covariance matrix Σ, that is

$$\mathcal{N}(o; \mu, \Sigma) = (\sqrt{(2\pi)^n |\Sigma|})^{-e^{-\frac{1}{2}(o-\mu)^T \Sigma^{-1}(o-\mu)}} \quad (7)$$

where n is the dimensionality of o. The following constraints apply to the model:

$$a_{ij} \geq 0 \quad (8)$$

that is, the probability of moving from state from any state i to j is not null, and the sum of all state transitions add up to unity:

$$\sum_{j=1}^{N} a_{ij} = 1, \forall i \quad (9)$$

For building Listener, HMM/GMM were applied to represent triphones. A triphone is a contextual phone, i.e. it is a phonetic unit of analysis which, for a given phone p, takes into account the previous phone $p-1$ and following one $p+1$. For instance, in a word like "speech" [s p iy ch], the phone [iy] would correspond to the triphone $[_{\text{p}}\text{iy}_{\text{ch}}]$, indicating that [iy] occurs after a [p] and before a [ch]. The full of transcription of "speech" in triphones would be $[_{\#}\text{s}_{\text{p}}\ _{\text{s}}\text{p}_{\text{iy}}\ _{\text{p}}\text{iy}_{\text{ch}}\ _{\text{iy}}\text{ch}_{\#}]$, it still has the same number of phone, the only difference is that the phones are now defined context.

For estimating the values and probabilities of the HMM/GMM the CMU Sphinx Toolkit was used [15]. Particularly, the acoustic model was trained over several different corpora, which contained, in total, 40 hours of audio from native speakers of English or Brazilian Portuguese, as well as non-native data in Brazilian-accented English.

Particularly, the following corpora were used for training the models:

1 English: TIMIT [16] and WSJ0 [17];
2 Portuguese: West Point BP [18] and OGI-22 [19];
3 Brazilian-accented English: Listener Corpus (described in Subsection 4.2).

The acoustic model was estimated considering a phonetic inventory of 54 phones, containing 4,000 tied states and 16 gaussian densities per state. The last two values were defined based on a pilot experiment over a sample from the available corpora. The phonetic transcriptions for the English corpora were extracted from the [7]; for the Portuguese corpora, the Aeiouadô grapheme-to-phoneme converter was used [20].

3.4 Context Free Grammars

Context free grammars are formal grammars in which every rule takes the form:

$$A \rightarrow \gamma \quad (10)$$

where A is a nonterminal and γ corresponds to a single or sequence of nonterminal or terminal symbol [21]. In speech recognition, context free grammars were the first attempt to broaden speech recognition to a context larger than digits, letters and menu commands; but they were rapidly replaced by statistical Language Models, as the latter scale better and require much less manual work [6].

For building the prototype of Listener, rules were used in order to build grammars to allow to the pronunciation variants to be recognized. The procedure is similar to the one described by Srikanth and Salsman [22], but instead of using phones as units, we focused on words, in order to be able to detect mispronunciatinos which involve larger context, such as syllables in cases of initial/coda epenthesis or palatalization. Basically, each mispronunciation pattern is defined as a non-terminal, which is then rewritten into the word lemma. Thus, the final prompt is able to recognize all combinations of pronunciation variants. A simplified example of such grammar is shown in Figure 3, which defines the rules for the prompt "I like Apple".

Figure 3 Context free grammars used for recognizing mispronunciations. Each mispronunciation entry is defined as a non-terminal.

```
PROMPT → I LIKE APPLE;
I → I_0;
LIKE → (LIKE_0 |LIKE_1);
APPLE → (APPLE_0 |APPLE_1 |APPLE_2 |APPLE_3);
I_0 → [ay];
LIKE_0 → [l ay k];
LIKE_1 → [l ay k ih];
APPLE_0 → [ae p l];
APPLE_1 → [ae p ow];
APPLE_2 → [eh p l];
APPLE_3 → [eh p ow];
```

4 Results and Discussion

The system was evaluated on three different test sets, all of them were compiled especially for this project and were made publicly available at the project's website. In the following subsections we describe each of them and present the results.

4.1 Test Set I: Corpus of Induced Errors

This test set consists of a speaker-dependent corpus of induced errors in isolated words (6,177 words ∼2 hours). This corpus was recorded by a single male

speaker with good proficiency of English (CEFR: C2), who induced pronunciation errors while reading isolated words in English. The recordings were made with a high-fidelity microphone in a quiet room, in order to reduce background noise. The corpus contains both audios with the expected pronunciation (∼30%) and with pronunciation errors (∼70%).

Results for the Induced test set can be found in Table 3.

Table 3 Recognition results for each phone in the Induced test set. Results are grouped by mispronunciation pattern, and the percentages for True Positives (TP) and Type I (false positive), Type II (false negative) errors are shown.

#	Category	Counts	TP	Type-I	Type-II
0	Expected phone	2265	0.96	0.03	0.02
1	Initial epenthesis	38	1.00	0.00	0.00
2	Coda epenthesis	14	0.79	0.00	0.21
3	Terminal devoicing	0	-	-	-
4	Th-fronting	6	0.67	0.00	0.33
5	Palatalization	197	0.90	0.03	0.07
6	Deaspirtation in plosives	103	0.89	0.05	0.06
7	Vocalization of laterals	11	0.45	0.00	0.55
8	Vocalization of nasals	301	0.86	0.04	0.10
9	Velar paragoge	25	0.96	0.04	0.00
10	Consonantal change	304	0.90	0.06	0.04
11	Vowel change	300	0.92	0.04	0.04
12	General deletion	0	-	-	-
13	General insertion	260	0.97	0.03	0.00
Total/W Avg		3553	0.93	0.03	0.03
Total/Avg (wo Exp.)		1288	0.90	0.04	0.06

As one might observe, the overall phone recognition for the Induced corpus was 0.93. Considering only phones with pronunciation errors, the ratio of true positives was 0.90. The expected phones showed a true positives ratio of 0.96. Initial epenthesis was the mispronunciation which was recognized with the highest accuracy, all 38 cases in the corpus were correctly identified. The system was able to detect initial sequences of [iy s C], as in "stop" [iy s t ao p], with no losses. Following, the best recognition performance was found in cases of general insertion, for instance, when one adds an extraneous phone to the end of a word, e.g. "work" [w ah r k s]. The true positive rate for generation insertion was 0.97. were the ones which were identified with the velar paragoge was the mispronunciation pattern which was recognized in most. Cases of velar paragoge, as in "king" [k ih ng g] or [k ih ng g ih] were accurately inferred in 0.97 cases. The worst results were found for vocalization of laterals (0.45), followed by th-fronting (0.67). The lower performance for these mispronunciations patterns types might be due to the fact that these errors involve phones which are acoustically similar. It could also be due to the fact that there is less data for these cases, but further investigation is needed. The rate of Type-I error, or false positives, was very small. The highest value was found in cases of consonantal change, which had a Type-I error rate of 0.06. Type-II errors occurred more often, vocalization of laterals had a ratio of 0.55, th-froning of 0.33 and coda epenthesis had 0.21.

4.2 Test Set II:Listener Corpus (isolated-words)

The second test set is a subset of the Listener Corpus, which contains only prompts with isolated words. The *Listener Corpus* was compiled especially for this work, in order to be a reference corpus for developing Computer Assisted Pronunciation Training focused on Brazilian-accented English. The corpus was built through crowdsourcing and contains native-speakers of Brazilian Portuguese reading out loud words and phonetically-rich sentences in English, which were extracted according to the method defined in Mendonça et al. [23]. In total, 67 volunteers were recorded (7,208 prompts ∼7 hours). The recording environment was not controled, subjects used their own laptops and personal computers to do the recordings from home. Overall, there is a considerable amount of noise in the channel, due to bad microphones or wrong settings (e.g. too much gain); as well as background noise (music, traffic, fan, animal sounds and so on).

Table 4 contains a summary of the recognition results for this test set.

Table 4 Recognition results for each phone in the Listener Corpus (Isolated-Words). Results are grouped by mispronunciation pattern; the values for True Positives (TP) and Type I (false positive), Type II (false negative) errors are presented.

#	Category	Counts	TP	Type-I	Type-II
0	Expected phone	1144	0.90	0.06	0.04
1	Initial epenthesis	2	0.50	0.50	0.00
2	Coda epenthesis	5	0.60	0.20	0.20
3	Terminal devoicing	0	-	-	-
4	Th-fronting	11	0.36	0.09	0.55
5	Palatalization	10	0.20	0.00	0.80
6	Deaspirtation in plosives	4	0.00	1.00	0.00
7	Vocalization of laterals	20	0.50	0.20	0.30
8	Vocalization of nasals	142	0.61	0.16	0.23
9	Velar paragoge	1	1.00	0.00	0.00
10	Consonantal change	25	0.52	0.24	0.24
11	Vowel change	97	0.60	0.19	0.22
12	General deletion	12	-	-	-
13	General insertion	21	0.81	0.14	0.05
Total/Avg		2518	0.82	0.09	0.09
Total/Avg (wo Exp.)		253	0.57	0.18	0.25

As one may notice, the results for the Listener test set were much lower. Basically, just the recognition results for the expected phones were comparable to the ones reported for the Induced corpus. The expected phones achieved a true positive rate of 0.90. Similarly to the previous test set, the second best performance was found in pronunciation errors with general insertion (e.g. "work" [w ah r k s]), with 0.81. The results for true positives for all other mispronunciation patterns were below 0.62. The average true positive ratio, considering only the data with pronunciation errors,

was 0.57. Two types of errors were never correctly recognized by the system, namely, general deletion (e.g. "foot" [f uh]) and deaspiration in plosives (e.g. "tea" [tt iy]). Other cases of phone replacement, such as palatalization (e.g. "city" [s ih ch i]) or th-fronting (e.g. "think" [f ih ng k] were correctly recognized only in few cases, the true positives rate for these errors were 0.20 and 0.36, respectively.

Our hypothesis is that such low performance is due to two factors: (i) different recording setup; and (ii) wrong transcriptions in the test set. In terms of quality, this test set is quite different from the audios that were used for training the acoustic model. The speech database used for estimating the acoustic model had clean audio, recorded in a controlled recording setup with high quality microphones. Due to the way that the Listener Corpus was compiled, the quality of the audio is quite different, there is both background and channel noise. Because of this mismatch, the acoustic models may not be able to generalize well and recognize the phones approprietaly in this test set.

As the corpus was transcribed one-way [24], there are inconsistencies in the annotations, which inevitably led to losses in the final results. In addition to this, the corpus was transcribed in a phonetic-oriented fashion, in such a way that the transcription have, for instance, some uncommon coarticulation phenomena (e.g. "king" [k iy nh iy]), that are not part of the thirteen mispronunciation patterns which the system is able to recognize.

4.3 Test Set III: Listener Corpus (all)

The last test set includes all types of prompts recorded for the Listener Corpus: isolated words, simple sentences and phonetically-rich sentences. Information about the corpus were already described in the previous subsection. The recognition results for this set are presented in Table 5.

As Table 5 shows, this test set had the worst performance among the three which were considered. The true positive rate was 0.63 on average. Considering just the mispronunciation patterns, i.e. without the results for expected phones, the true positive rate drops to 0.31. Such result is definitely unsuitable for developing Computer Assisted Pronunciation Training applications. What can be inferred from this result is that connected speech deeply affects the recognition results.

Table 5 Recognition results for each phone in the Listener Corpus (all). Results are grouped by mispronunciation pattern; the values for True Positives (TP) and Type I (false positive), Type II (false negative) errors are presented.

#	Category	Counts	TP	Type-I	Type-II
0	Expected phone	9245	0.76	0.17	0.07
1	Initial epenthesis	34	0.12	0.26	0.62
2	Coda epenthesis	49	0.24	0.49	0.27
3	Terminal devoicing	0	-	-	-
4	Th-fronting	179	0.23	0.20	0.57
5	Palatalization	93	0.03	0.13	0.84
6	Deaspirtation in plosives	60	0.03	0.20	0.77
7	Vocalization of laterals	179	0.31	0.35	0.34
8	Vocalization of nasals	1017	0.37	0.29	0.35
9	Velar paragoge	1	1.00	0.00	0.00
10	Consonantal change	260	0.23	0.22	0.55
11	Vowel change	1366	0.37	0.18	0.45
12	General deletion	421	-	-	-
13	General insertion	131	0.30	0.27	0.44
Total/Avg		11669	0.63	0.19	0.19
Total/Avg (wo Exp.)		2424	0.31	0.21	0.48

5 Conclusions

For speech recognition – or, in fact, any supervised machine learning task – the best scenario for training a model is when you have a huge amount of data which is large and diverse enough so that it fully represents population. However, this is usually not the case. The data was not readily available when the project started and all speech corpora were still to be compiled.

The prototype has achieved a very high accuracy in the Induced corpus, with a true positive rate of 0.90, meaning that each mispronunciation pattern that really occurred in the corpus were correctly identified 90% of the time. This result also means that the Multipronunciation dictionary with hand-written rules was a reliable source of phonetic information for pronunciation assessment. Context free grammars based on words were successfully adapted for forced-alignment recognition, in order to list all pronunciation variants of a given word.

However, the Induced corpus is speaker-dependent and contains clean speech data. Considering the Listener Corpus, which has a considerable amount of the noise, the system performed much worse: the true positive rates obtained for both test sets from this corpus (0.57 and 0.31) are unseemly for developing real applications in automatic pronunciation assessment.

As a future work, it might be interesting to evaluate the prototype with data that is not as clean the Induced corpus and that is, also, not as noisy as the test sets from the Listener Corpus. For instance, a corpus recorded from mobile phones would fill both these criteria. It could also be interesting to enlarge the training set, in order to build more robust acoustic models.

6 Acknowledgements

We would like to thank the Instituto de Telecomunicações in Coimbra for for the support with the corpora for training the acoustic models.

References

1. Fund, I.M.: World Economic Outlook. International Monetary Fund, Washington, DC, USA (2015)
2. EducationFirst: EF English Proficiency Index 2015. Education First Ltd., Lucerne (2015)
3. GlobalEnglish: The 2013 Business English Index & Globalization of English Report, p. 15. Pearson Always Learning, Pearson (2013)

4. Huang, X., Acero, A., Hon, H.-W.: Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, 1st edn., p. 980. Prentice Hall PTR, Upper Saddle River, NJ, USA (2001)
5. Candeias, S., Veiga, A.: Chapter twelve the dialogue between man and machine – the role of language theory and technology. In: The Role Of Language Theory And Technology, pp. 215–226 (2014). New Language Technologies and Linguistic Research
6. Gales, M., Young, S.: The application of hidden markov models in speech recognition. Foundations and trends in signal processing **1**(3), 195–304 (2008)
7. Weide, R.: The CMU Pronouncing Dictionary 0.7a. Carnegie Mellon University (2008). http://www.speech.cs.cmu.edu/cgi-bin/cmudict
8. Strik, H.: Pronunciation adaptation at the lexical level. In: ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition (2001)
9. Zimmer, M.: A Transferência do Conhecimento Fonético-Fonológico do Português Brasileiro (L1) Para O Inglês (L2) na Recodificação Leitora: Uma Abordagem Conexionista. Dissertação de Doutorado. Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre (2004)
10. Zimmer, M., Silveira, R., Alves, U.: Pronunciation Instruction for Brazilians: Bringing Theory and Practice Together. Cambridge Scholars, Newcastle (2009)
11. Silva, T.C.: Pronúncia do Inglês Para Falantes do Português brasileiro. Contexto, São Paulo, Brazil (2015)
12. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)
13. Fink, G.A.: Markov Models for Pattern Recognition: from Theory to Applications. Springer, London (2014)
14. Juang, B., Rabiner, L.: In: Brown, K. (ed.) Automatic Speech Recognition - A Brief History of the Technology, p. 24. Elsevier, Amsterdam (2005)
15. Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., Woelfel, J.: Sphinx-4: A flexible open source framework for speech recognition (2004)
16. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S.: Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. NASA STI/Recon Technical Report N **93** (1993)
17. Garofolo, J., Graff, D., Paul, D., Pallett, D.: Csr-i (wsj0) complete. Linguistic Data Consortium, Philadelphia (2007)
18. Morgan, J., Ackerlind, S., Packer, S.: West point brazilian portuguese speech. Linguistic Data Consortium, Philadelphia (2008)
19. Lander, T., Cole, R.A., Oshika, B.T., Noel, M.: The ogi 22 language telephone speech corpus. In: Eurospeech (1995)
20. Mendonça, G., Aluísio, S.: 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 (2014)
21. Jurafsky, D., Martin, J.: Speech and Language Processing : an Introduction to Natural Language Processing Computational Linguistics, and Speech Recognition, 2ª edn. Prentice Hall, New Jersey, USA (2000)
22. Srikanth, R., Salsman, L.B.J.: Automatic pronunciation evaluation and mispronunciation detection using cmusphinx. In: 24th International Conference on Computational Linguistics, p. 61 (2012). COLING
23. Mendonça, G., Candeias, S., Perdigão, F., Shulby, C., Toniazzo, R., Klautau, A., Aluísio, S.: A method for the extraction of phonetically-rich triphone sentences. Proceedings of ITS 2014 – International Telecommunications Symposium (2014)
24. Shulby, C.D.: Iprompter: an Intelligent Pronunciation Tutor for Brazilian Students of Academic English (qualifying Examination Monograph). Universidade de São Paulo, São Carlos, Brazil (September 2015)

Chapter 4

Conclusions

Overall Conclusions

This thesis sought to be an initial step in using deep linguistic knowledge to develop a CAPT for Brazilian-accented English. Back in 2013, when this project started, one of the core challenges of researches in CAPT was that ASR systems were not precise enough in terms of phone recognition [48], what posed serious problems for CAPT since phones are necessarily the basic signal that pronunciation training systems must rely on. Our first hypothesis was that by using as much phonetic knowledge in all stages of the pipeline for a CAPT, we would be able to obtain more robust results in phone recognition. Thus improving the quality of pronunciation assessment and pushing the state of the art forward. The plan was to focus especially in the ASR, not only by training acoustic models that would be representantive of the phones in Brazilian-accented English, but also enriching the pronunciation dictionary with relevant mispronunciations.

However as time went by, this initial project has proven to be undoable for a Master's project not only with respect to the time or resources available, but also in terms of complexity. The project was too ambitious and results in ASR are often exploratory. The truth is that ASR is a huge interdisciplinary area which, at present time, is still not solved. Even large IT companies which have at their disposal the best algorithms, computer power and speech corpora in the world still report Word Error Rate (WER) results around 12% for conversational speech [20]. This WER result means that a short sentence with 4 words is fully recognized only 59% of the times.

Due to the complexity of the task and the scarcity of resources [31], replicating for Brazilian-accented English other methods that were successfully applied to other languages was unfeasiable. There was not enough data for training acoustic models, specialized language models or dictionaries which included entries with common mispronunciations.

Our approach was then more conservative and focused on resources for Natural Language Processing and Pronunciation Evaluation that can help a future development of a CAPT system for Brazilian-accented English. More especifically, our focus was on tasks for text-to-speech, spelling correction, corpus building and automatic pronunciation assessment – our approach in all of them was to include enrich the models with phonetic knowledge in order to increase their performance. As it was shown in the papers presented, improvements were made and we were able to push the state of the art one step further in several occasions[1]:

Text-to-speech

1. The hybrid approach that we proposed to text-to-speech, which makes use of manual rules and machine learning, has proven to be quite efficient. Not only the results are in the state of the art, but also the approach is the quite flexible and scalable. The architecture of Aeiouado allows one to easily work on improving the accuracy/recall of the system by embedding more phonetic knowledge through reviewing the transcription rules, or by enlarging the test set, for instance, providing new examples for training.

2. Supra-segmental information has shown to be useful feature for the machine learning classifier. As can be seen by the results in the paper [25], the model was able to learn the difference between pretonic/tonic vs. postonic vowels. The f1-measure for [ʊ, ə, ɪ] was 1.00, 0.99 and 0.97, respectively.

3. Part-of-speech were able to help the model to learn the difference in heterophonic homograph pairs to a certain extent. The G2P was able to correctly infer heterophonic homograph pairs that were not seen during training, such as "ab[o]rto" (abortion) and "ab[ɔ]rto" (I abort), or "ap[o]sto" (apposition) and "ap[ɔ]sto" (I bet).

Spelling correction

4. The hypothesis that phonetic knowledge could be used to improve the results of the speller due to phonologically-motivated errors was correct. The baseline system which had no phonetic module was able to correct 48.2% of non-contextual phonologically motivated errors in the corpus. In comparison, one of the methods which considered the distance between phonetic transcriptions was able to achieve 87.1% accuracy in the same task.

[1]The resources are available at the project website *(http://nilc.icmc.usp.br/listener)*. Due to copyright reasons, the corpora used for training the acoustic models cannot be made available.

5. Our assumption about the types of mispellings also held. Around 18% of the mispelling errors in user-generated content were phonologically-motivated. We compiled, annotated and released to the public a corpus for spelling correction tasks with 38,128 tokens, 4,083 of which containing mispelled words. It is worth noticing that there were no open corpora to evaluate this task in Portuguese, prior to this work.

Corpus building

6. Our hypothesis that greedy algorithms would be a suitable way for extracting phonetically-rich sentences from a corpora was supported. The results showed that the greedy strategy was capable of extracting sentences in a much more uniform way, while comparing to a random selection. For instance, the method was able to extract 854 new distinct triphones for a sample of 250 sentences, with almost twice the type/token ration of the random sample – 0.61 vs. 0.32, respectively.

Automatic pronunciation assessment

7. The multipronunciation dictionary with manually rules has shown to be a reliable source of phonetic information for pronunciation assessment.

8. Our hypothesis about the acoustic models could be verified as the performance of the prototype was severely affected by noise in the data.

9. Context-free grammars were successfully adapted to perform forced-alignment.

10. The additional pronunciation variants did not harm the performance, as expected.

Limitations and Further Work

Although the research has reached its partial aims in building tools and resources for Natural Language Processing and Pronunciation Evaluation, there were still some unavoidable limitations. We conclude this thesis by discussing the limitations and providing ideas for future research.

Text-to-speech Despite using part-of-speech information to capture the difference in heterophone homograph pairs, this did provide all context that mid vowels need. The worst results were related to the transcription of mid vowels [ɛ, e, ɔ, o]. Particularly, the model was very confused about mid-low ones, [ɛ] showed an F1-score 0.66 and [ɔ] of 0.71. In future

studies, it might be interesting to see if any other features could be used to improve the model performance in distinguishing the difference between [e,o] and [ɛ,ɔ], respectively. Exception dictionaries or post-processing rules can be used to increase the accuracy [39]. It can also be interesting to see if the errors are somehow related to vowel harmony (p[ɛ]r[ɛ]r[ɛ]ca) or to suffixation (p[ɛ] > p[ɛ]zinho). One could also check whether training the models with more data would suffice to solve the issues reported here.

Spelling correction The architecture of the GPM-ML speller was quite complex and used features from many different sources (phonetic transcription, language model, Keyboard model and string edit distance), which, theoretically, would provide enough information for the classifier to learn the mispelling patterns. Although the speller presented results from the state of the art, our initial expectation was that it would have even a better performance, especially with respect to general typos, e.g. “fala” (speech) > “fdala” (speecxh). These cases would generate very low probabilities in the language model, and the keyboard model would be able to indicate that “f” and “d” are adjacent keys, so the pattern for identifying these errors should be available for the classifier. However, in comparison to the other method, which had neither a language model nor a keyboard model, there was just a 2% improvement for general typos. It might be interesting, in a future work, to investigate why the impact was rather small. Further research could also focus in trying to improve the language model in order to tackle contextual errors. We assumed that, by using probabilities from a language model estimated from news texts, the classifier would learn the difference between sequences of words with correct spelling and those with errors. However, this was true just for contextual diacrictic errors, not for those contextual errors which were phonologically-motivated. The speller was able to automatically fix real-errors with diacrictics, such as “ela *e* inteligente” (lit. she and smart) -> “ela *é* inteligente” (she is smart); but it was not able to generalize this to other contexts, such as “lojas do *seguimento*” (lit. stores of the following) -> *“lojas do *segmento*”(related business, lit. stores of the segment). It is worth noticing that the language model was estimated over a subset of the Corpus Brasileiro (circa 10 million tokens) and it was not filtered before estimating the probabilities. The speller cuold benefit from corpora with less noise. More powerful methods of machine learning could also be employed to improve the performance.

Corpus building The method we proposed was able to extract sentences with much more uniform triphones as the type/token ration confirms, however, since the method basically favours rare triphones in each iteration, some of the sentences that were extracted had very awkward reading or uncommon words. Considering that the method is useful especially

for preparing prompts to build speech recognition corpora, this poses a problem: if the prompt is not understood, the voice donor might hesitate or read the sentence in a way that is not natural. In the future, it might be interesting to develop some sort of filter during each iteration of the algorithm, to remove these sentences with problems, while keeping the triphone balance. In addition to this, one of the main reasons for building the algorithm was to test whether phonetically-rich corpora would be more benefitial to phone recognition than phonetically-balanced corpora. Our hypothesis was that since the phones are sampled more equally, acoustic models would be able to better estimate the phone properties, thus leading to improvements in tasks which demand robust phone models, such as phone recognition or forced-alignment. But due to time constraints this hypothesis could not be tested. If supported, this may be quite interesting for application which require very accurate phone models, such as CAPT or speaker verification.

Automatic pronunciation assessment The prototype has achieved a very high accuracy in the Induced corpus, with a true positive rate of 0.90, meaning that each mispronunciation pattern that really occurred in the corpus were correctly identified 90% of the time. However, the Induced corpus is speaker-dependent and contains clean speech data. While evaluating the Listener Corpus, which has a considerable amount of the noise, the system performed much worse: the true positive rates obtained for both test sets from this corpus (0.57 and 0.31) are unseemly for developing real applications in automatic pronunciation assessment. As a future work, it might be interesting to evaluate the prototype with data that is not as clean the Induced corpus and that is, also, not as noisy as the test sets from the Listener Corpus. For instance, a corpus recorded from mobile phones would fill both these criteria. It could also be interesting to enlarge the training set, in order to build more robust acoustic models.

References

[1] 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.

[2] Bisol, L. (2005). *Introdução a estudos de fonologia do português brasileiro*. Edipucrs.

[3] 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*.

[4] Câmara, J. M. (1970). *Estrutura da língua portuguesa*. Editôra Vozes.

[5] Collischonn, G. (2004). Epêntese vocálica e restrições de acento no português do sul do brasil. *Signum: Estudos da Linguagem*, 7(1):61–78.

[6] Cristófaro Silva, T. et al. (2012). Revisitando a palatalização no português brasileiro. *Revista de Estudos da Linguagem*, pages 59–89.

[7] Crystal, D. (2011). *Dictionary of linguistics and phonetics*, volume 30. John Wiley & Sons.

[8] Davenport, M., Davenport, M., and Hannahs, S. (2010). *Introducing phonetics and phonology*. Routledge.

[9] 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.

[10] de Medeiros, B. R. (2007). Vogais nasais do português brasileiro: reflexões preliminares de uma revisita. *Revista Letras*, 72.

[11] EducationFirst (2015). *EF English Proficiency Index 2015*. Education First Ltd.

[12] Fink, G. A. (2014). *Markov models for pattern recognition: from theory to applications*. Springer Science & Business Media.

[13] 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.

[14] Fund, I. M. (2015). *World Economic Outlook.* International Monetary Fund, Washington, DC, USA.

[15] Gales, M. and Young, S. (2008). The application of hidden markov models in speech recognition. *Foundations and trends in signal processing*, 1(3):195–304.

[16] GlobalEnglish (2013). *The 2013 Business English Index & Globalization of English Report*. Pearson Always Learning, Pearson.

[17] Gordon, R. G. and Grimes, B. F. (2005). *Ethnologue: Languages of the world*, volume 15. SIL International Dallas, TX.

[18] Hermansky, H. (1990). Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752.

[19] 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.

[20] Huang, X., Baker, J., and Reddy, R. (2014). A historical perspective of speech recognition. *Communications of the ACM*, 57(1):94–103.

[21] Johnson, K. (2004). Acoustic and auditory phonetics. *Phonetica*, 61(1):56–58.

[22] Lakoff, R. (1973). Language and Woman's Place. *Language in Society*, 2(1).

[23] McLoughlin, I. (2009). *Applied Speech and Audio Processing – With Matlab Examples*. Cambridge University Press, Cambridge.

[24] Mello, H., Avila, L., Neder-Neto, T., and Orfano, B. (2012). Lindsei-br: an oral english interlanguage corpus. volume 1, pages 85–86, Florença, Itália. Firenze University Press.

[25] 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.

[26] Mendonça, G. and Aluísio, S. (2016). Listener: A prototype system for automatic speech recognition and evaluation of brazilian-accented english. *Journal of the Brazilian Computer Society (submitted)*.

[27] Mendonça, G., Avanço, L., Duran, M., Fonseca, E., Volpe-Nunes, M., and Aluísio, S. (2016). Evaluating phonetic spellers for user-generated content in Brazilian Portuguese. *PROPOR 2016 – International Conference on the Computational Processing of Portuguese (submitted)*.

[28] Mendonça, G., Candeias, S., Perdigão, F., Shulby, C., Toniazzo, R., Klautau, A., and Aluísio, S. (2014). A method for the extraction of phonetically-rich triphone sentences. *Proceedings of ITS 2014 – International Telecommunications Symposium*.

[29] Mporas, I., Ganchev, T., Siafarikas, M., and Fakotakis, N. (2007). Comparison of speech features on the speech recognition task. *Journal of Computer Science*, 3(8):608–616.

[30] 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.

[31] 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.

[32] Neves, M. H. M. (1999). Gramática do português falado. vol. vii: Novos estudos. são paulo.

[33] O'Connor, J. D. (1987). *Better English Pronunciation*. Cambridge University Press.

[34] Rabiner, L. (1997). Applications of speech recognition in the area of telecommunications. *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 501–510.

[35] Rabiner, L. and Schafer, R. (2007). *Introduction to Digital Speech Processing*, volume 1, pages 1–194.

[36] Robjohns, H. (2010). A brief history of microphones. http://microphone-data.com//media/filestore/articles/History-10.pdf. Last retrieved 11-21-2014.

[37] Rocca, P. D. A. (2003). Bilingualism and speech: evidences from a study on vot of english and portuguese voiceless plosives. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 19(2):303–328.

[38] 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.

[39] Shulby, C., Mendonça, G., and Marquiafável, V. (2013). Automatic disambiguation of homographic heterophone pairs containing open and closed mid vowels. In *9th Brazilian Symposium in Information and Human Language Technology*, pages 126–137, Fortalez, CE, Brazil.

[40] Silva, T. C. (2005). *Fonética e fonologia do português: roteiro de estudos e guia de exercícios*. Contexto.

[41] Skandera, P. and Burleigh, P. (2005). A manual of english phonetics and phonology. *Tübingen: Gunter*.

[42] Stenson, N., Downing, B., Smith, J., and Smith, K. (2013). The effectiveness of computer-assisted pronunciation training. *Calico Journal*, 9(4):5–19.

[43] Steriade, D. (2000). Paradigm uniformity and the phonetics-phonology boundary. *Papers in laboratory phonology V: Acquisition and the lexicon*, 3:13–334.

[44] 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.

[45] Umesh, S., Cohen, L., and Nelson, D. (1999). Fitting the mel scale. *Proc. ICASSP 1999*, pages 217–220.

[46] Wiese, R. (2001). The phonology of/r. *Distinctive feature theory*, 2:335.

[47] Witt, S. (1999). *Use of Speech Recognition in Computer-assisted Language Learning*. Universidade de Cambridge - Dept. of Engineering, Cambridge, RU.

[48] Witt, S. M. (2012). Automatic error detection in pronunciation training: Where we are and where we need to go. *Proc. IS ADEPT*.