Applied Phonetics: Portuguese Text-to-Speech

Arlo Faria
University of California, Berkeley
Linguistics 110: Prof. Ian Maddieson

Contents

Applied Phonetics: Portuguese Text-to-Speech

Arlo Faria
University of California, Berkeley
Linguistics 110: Prof. Ian Maddieson

May 16, 2003

Abstract

This paper describes a text-to-speech application for a variety of Brazilian Portuguese. After presenting the language's phonetic attributes, the orthographic system is examined and shown to be a function that maps letters to these sounds. Given the orthography's phonological regularity, it is simple to implement the textual analysis portion of a speech synthesis system, as I demonstrate with some simple Perl code.

1 Introduction

For decades, technologists have been predicting that computers will radically revolutionize the way we live. Visionaries have long dreamed of a future in which machines are integrated into nearly every facet of daily life, From artificially intelligent robotic agents to embedded microcomputers in household appliances, it will be virtually impossible to do anything that is not somehow "wired". As the power of computing grows exponentially, no one is even able to imagine the possible applications.

At least one thing is certain: human-computer interaction will be far more intricate than today's mouse, keyboard, and monitor interface. More likely, users will want to relate to machines in a more natural manner – in a more human manner. In this respect, computers of the future will need to be capable of receiving and conveying information through our most preferred channel of communication: speech. Robust speech recognition and natural-sounding speech synthesis will be part of the new interface protocol.

Redefining input and output is just one of the many promising applications of speech and language processing that have only recently become possible. Other examples of speech technologies: learning foreign languages[1], automated telephonic customer service[2], and assistive technologies for the disabled[3].

This paper presents the framework for a text-to-speech system for a common variety of

[1] Computer-Aided Language Learning: `http://www.ocf.berkeley.edu/~arlo/CALL.pdf`
[2] `http://www.tellme.com`
[3] `http://www.perl.com/pub/a/2001/08/27/bjornstad.html`

Brazilian Portuguese. Sparing explicit technical specification, it highlights the linguistic background that is necessary for such a task. Taking advantage of the language's phonological orthography, the process of translating from letter tokens into speech segments is explored in great detail.

I first discuss the phonetic attributes of Portuguese, describing the inventory of sounds along with the language's syllable structure and stress pattern. To illustrate these features, a set of acoustic and auditory measurements are provided. Some of these word lists and figures are included in the Appendix. The recorded speech was preserved as digitized sound files, available online: `www.ocf.berkeley.edu/~arlo/ling110/`.

Instead of a treatment of Portuguese phonology, I then proceed to explain the Portuguese orthographic system. Indeed, the orthography is very phonological: the information needed to derive a phonetic representation is almost entirely contained in the orthography's set of letter-to-phone rules. These rules are listed, thoroughly defining the function mapping a sequence of letters into a sequence of sound segments. Portuguese syllable structure and lexical stress are examined, and it is shown that these are also represented in the orthography.

The final portion of this paper is a demonstration of how to apply the phonetic background and orthography of Portuguese in the implementation of a speech synthesis system. Specifically, the phonological regularity of the Portuguese orthography greatly facilitates the primary stages of the text-to-speech process. An illustrative example is provided, with the letter-to-phone rules being straightforwardly coded into Perl regular expressions.

The text-to-speech implementation that I describe proceeds satisfactorily until the intermediate stage of the process; it is, in effect, a system that translates text into its phonetic realization. For practicality, I show an application that processes a user's text input and displays its reading in IPA transcription. Such a text-to-transcription tool has considerable accuracy – not common for many contemporary speech technologies – and could be very useful to linguists and learners of Portuguese.

Finally, I discuss how one could complete the text-to-speech application by constructing an appropriate audio signal from the phonetic representation. It is evident that this final step presents the greatest technical challenge, and I mention a simple approach to speech synthesis.

This paper thoroughly details the first two of four steps in a text-to-speech system. Building upon this work, the last two – prosodic analysis and signal processing – are areas which I hope to one day explore.

2 The phonetics of Portuguese

Underlying any implementation of speech technology is a thorough understanding of the linguistic systems of the language being targeted. For this Portuguese text-to-speech system, it is necessary to be acquainted with the sounds and patterns of a particular variety of Brazilian Portuguese. This section describes the phonetic inventory utilized in the speech of the author, Arlo Faria.

2.1 Historical Background and Overview

Portuguese is a Romance language related to Castilian and Catalan. A unique type of Latin that was spoken in the isolated northwest part of the Iberian Peninsula during the Middle Ages, the Galician-Portuguese dialects spread south to Portugal during the thirteenth century. After supplanting the Moors, over the next centuries Portuguese sailors would continue to propagate the language. By the seventeenth century, it was spoken in Brazil, African colonies, several island colonies, and ports of India and Southeast Asia.

Today, Portuguese is spoken by 176 million people [2]. The most significant varieties are European Portuguese (10 million speakers) and Brazilian Portuguese (158 million). Because of the size of Brazil (larger than the continental U.S.A.), there are numerous geographical dialects of Portuguese. The sheer vastness of the country, along with the existence of many sociolects, makes it nearly impossible to thoroughly examine and characterize Brazilian Portuguese [1].

For the purposes of this paper, the dialect of Portuguese studied is assumed to be the dialect spoken around Campinas, São Paulo. This is an urban region in the south-central part of the country, which is where the majority of the populace lives. By the estimation of this paper's author[4], the dialect is very similar to that spoken on national television and is close to what might be considered the educated standard[5].

2.2 Vowels

The vowel inventory of Brazilian Portuguese comprises eight sounds of distinct quality. Of these, there are seven principal vowels that can appear in stressed positions ([i e ɛ a ɔ o u]) and one vowel that is always unstressed ([ɐ]). Two approximants allow for all permutations of the principal vowels as diphthongs and some instances of triphthongs. Additionally, nasalization can occur with any type of vowel and is a salient phonemic distinction.

2.2.1 The principal vowels

The principal vowels are symmetrically distributed on a standard vowel chart: three front vowels, two central vowels, and three back vowels. The three back vowels are rounded.

Figure 1 illustrates the relative placements of the Portuguese vowels. Figure 2 is a plot [6] of the first and second formants of these vowels. These measurements were done according to the recording process described in Appendix A. Six data points correspond to each vowel:

[4]Arlo Faria was born in Viçosa, in the rural central state of Minas Gerais. The dialect of Portuguese spoken in this region of the country is markedly different from that spoken in the more urbanized areas. At the age of four, I moved to São Paulo, where I was exposed to a dialect which is probably more associated with a standard Brazilian Portuguese. (Many older texts will consider the dialect of Rio de Janeiro, which is not sociographically too distant.) At eight years of age, I moved to New Jersey; I haved returned to Brazil about once a year since then.

[5]This notion of an "educated standard" elicits contention from many Brazilian linguists. I posed the question to a friend in Rio de Janeiro, Prof. Dr. Mauricio Brito [3] Carvalho. His colleague, Prof. Dr. Luiz Francisco Dias [5], explained that an "educated standard" is an idealistic conceit. Among other factors, there exists a considerable discrepancy between spoken and written language, especially since a significant portion of the population is not capable of the latter.

words were chosen which demonstrated the vowels positioned initially, medially, and finally; each word was spoken twice. There is some personal variation in the observed formants due to the influence of the words' contexts [4].

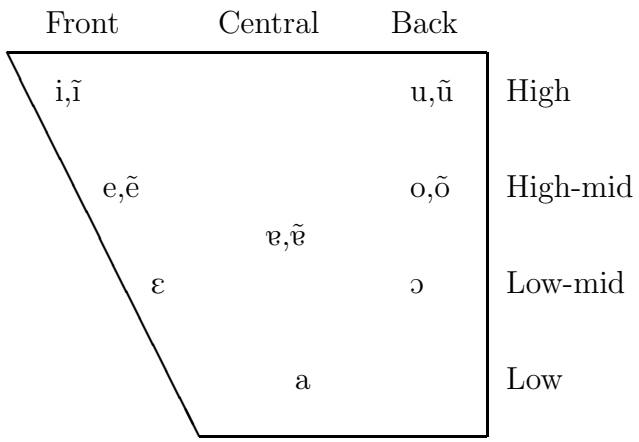


Figure 1: The vowels of Brazilian Portuguese

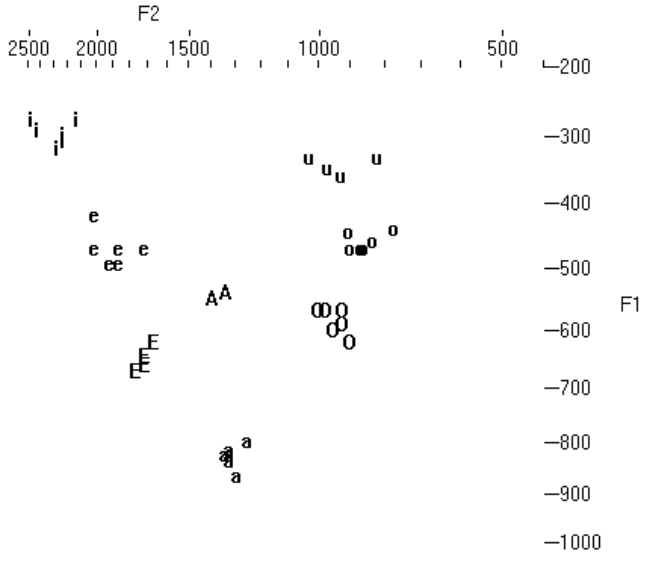


Figure 2: First and second formants of Brazilian Portuguese vowels
Note: In this plot, the symbols E, O, A represent [ɛ ɔ ɐ], respectively.

Table 1: Falling and Rising Diphthongs

Word	Transcription	Translation
riu	ʁiw	laughed
seu	sew	his
céu	sɛw	sky
sal	saw	salt
sol	sɔw	sun
sou	sow	I am
rei	ʁej	king
anéis	aˈnɛjs	rings
vai	vaj	go
herói	eˈɾɔj	hero
dois	dojs	two
fui	fuj	went

Word	Transcription	Translation
série	sɛˈɾje	series
dieta	ˈʤjɛtɐ	diet
comédia	koˈmɛʤja	comedy
maior	maˈjɔ	larger
biologia	bjoloˈʒiɐ	biology
viuvez	vjuˈves	widowhood
qüiproquó	kwiproˈkɔ	*quid pro quo*
lingüeta	lĩgˈwetɐ	little tongue?
cueca	ˈkwɛkɐ	underpants
quarto	ˈkwaɾtu	room
quota	ˈkwɔtɐ	quota
vácuo	ˈvakwo	vacuous

Table 2: Some Portuguese triphthongs

Word	Transcription	Translation
Uruguai	uɾuˈgwaj	Uruguay
enxaguou	ĩʃaˈgwow	rinsed
delinqüiu	delĩˈkwiw	was delinquent (?)
UAU!	waw	WOW!

2.2.2 Diphthongs (and Triphthongs)

Utilizing the palatal approximant [j] and the labio-velar approximant [w], any stressed vowel may form a rising or falling diphthong [7] with the exception of [ij ji] or [uw wu]. It is thus possible to produce twenty-four distinct diphthongs, and a few triphthongs. The diphthongs are illustrated in Table 1. (Many of these examples are from [7].)

It is important to note that the set of rising diphthongs alternates with a sequence of two vowels. For example, *cueca* ("underpants") can be realized as [ˈkwɛ.kɐ] or [ˈku.ɛ.kɐ]. When pronounced with the diphthong, the word is one syllable shorter; this phenomenon seen is typical of spontaneous speech, as opposed to careful speech. While rising diphthongs can alternate with a sequence of two vowels, however, the inverse is not necessarily true. The word *biologia* ("biology") can be realized as [bjo.lo.ˈʒi.a] or [bi.o.lo.ˈʒi.a], but the final syllable is never [ʒja].

In Brazilian Portuguese, the diphthong [ow] is usually pronounced simply as [o], except in very careful speech.

It is possible for a vowel to occur between two approximants, in which case a triphthong results. There are limited examples of these sounds, seen in Table 2.

Table 3: Nasal vowels

Word	Transcription	Translation
sim	sĩ	yes
sempre	ˈsɐ̃pɾi	always
canto	ˈkɐ̃tu	sing
conto	ˈkõtu	tell
um	ũ	one

Table 4: Nasal diphthongs

Word	Transcription	Translation
(no example)	(ĩw̃)	
(no example)	(ɐ̃w̃)	
sem	sɐ̃j̃	without
pão	pɐ̃w̃	bread
pães	pɐ̃j̃s	breads
tom	tõw̃	tone
põe	põj̃	puts
muito	ˈmũj̃tu	many

Falling Diphthongs

Word	Transcription	Translation
qüingentésimo	ˌkw̃ĩʒɐ̃ˈtɛzimu	five-hundredth
cinqüenta	sĩˈkwɐ̃tɐ	fifty
clientela	klj̃ɐ̃ˈtɛlɐ	clientele
quanto	ˈkw̃ɐ̃tu	how much
dianteira	ʤj̃ɐ̃ˈterɐ	front part
(no example)	(w̃õ)	
biombo	ˈbj̃õbu	partition
triunfo	ˈtɾĩũfu	triumph

Rising Diphthongs

2.2.3 Nasal vowels

In literature, there are many competing treatments of nasals in Portuguese. Describing one of these frameworks is beyond the scope of this paper. The most important property to be noted, though, is that nasal vowels are always [-low] [1]. That is, the nasals /ɛ̃ ã ɔ̃/ do not exist; instead, they are likely raised and realized as [ẽ ɐ̃ õ]. Some Portuguese nasal vowels and diphthongs are illustrated in Table 4. (Example words for some nasal diphthongs could not be encountered, but are included in the chart because it is assumed that they are possible.)

As with its non-nasalized counterpart, the diphthong [õw̃] is often pronounced as just [õ]. An example of a nasalized triphthong: [kw̃ɐ̃w̃] *quão* (how). Contrast this vowel quality with the non-nasalized [kwaw] *qual* (which).

2.3 Consonants

The consonants of Brazilian Portuguese are illustrated in Table 5 [11] [7].

The consonants of Portugese can all occur word-medially, and most also occur word-initially (except [ʎ ɲ ɾ])[6]. A very small set of consonants occurs word-finally ([s z ɾ x]).

One of the most variable sounds across dialects of Portuguese is the realization of rhotics[7].

[6]Some notable exceptions: *lhe* [ʎi] 'him/her' (dative); *lhama* ['ʎamɐ] 'llama' (Spanish origin)

[7]By *rhotic* I am referring to the pronunciation of orthographic ⟨r⟩ or ⟨rr⟩

Table 5: The consonants of Brazilian Portuguese

	Bilabial	Labiodental	Alveolar	Palato-alveolar	Palatal	Velar
Plosive	p b		t d			k g
Nasal	m		n		ɲ	
Fricative		f v	s z	ʃ ʒ		x
Affricate				ʧ ʤ		
Tap			ɾ			
Lateral			l		ʎ	
Approximant	w (labio-velar)				j	

These have been transcribed variously as [r ɾ ɽ ɹ ʁ ʁ x χ]. Almost every variety includes the alveolar tap [ɾ] in word-medial contexts. This can be considered the underlying representation [1]. In addition to the tap, my speech includes a voiceless fricative, which is probably velar (though most texts will note uvular [χ]). My speech does not include trills. Word-finally, I have difficulty pronouncing [ɾ], instead saying [ɽ][8] or leaving the syllable open. In spontaneous speech, however, word-final rhotics are often elided[9]. For many speakers, thus, [s z] are the only word-final consonants.

There is some flexibility with the place of articulation for alveolar articulations; sometimes they are realized as dental.

2.4 Syllable Structure

2.4.1 Onsets

Most consonants can occur word-initially, and any single consonant can be a syllable onset.

Unlike the European variety, Brazilian Portuguese has strict limitations on consonant clusters in syllable onsets: only a plosive followed by an alveolar tap or lateral is licensed. Exceptionally, [fɾ] and [fl] can occur. The explanation for this is provided by the Sonority Principle [1] [8], which prohibits onset clusters consisting of consonants with the same degree of sonority.

[8]This is probably influenced by my acquisition of English, or it could be an effect of hypercorrection. Prof. Milton Azevedo [7] notes that the retroflex rhotic – called 'r caipira' ('redneck r') – is becoming more common in urban areas. He attributes this to immigration into cities and the urbanization of the countryside.

[9]Elision of word-final /r/ is a notorious trait of the dialect spoken in Minas Gerais – or so I originally thought. As Prof. Carvalho informed me, "in fact, it is a widespread phenomenon, common in all dialects of Brazilian Portuguese" [3]

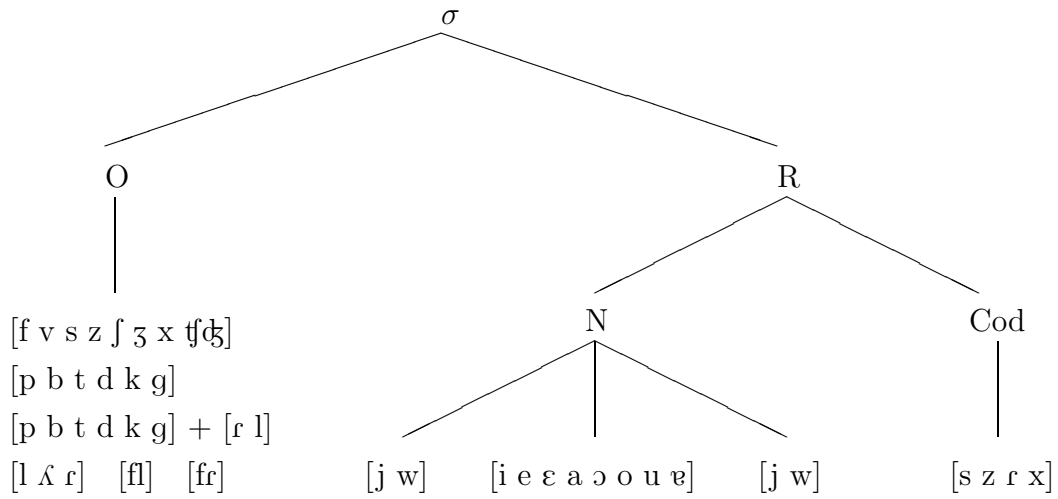


Figure 3: The syllable structure of Brazilian Portuguese.

For other consonant clusters, the existence of an *empty nucleus* is hypothesized [1], such that a vowel, usually [i], is inserted. For example, 'pneu' [piˈnew] 'tire'.

2.4.2 Rhymes

There are no syllabic consonants in Brazilian Portuguese[10]. The rhyme of a Portuguese syllable must include a nuclear vowel; vowels are "the only indispensable elements in syllabic parsing" [1]. A complex nucleus may be formed by adding one or two glides to the nuclear vowel.

The coda of a rhyme is either empty or occupied by one of [s z ɾ x]. As explained previously, the rhotic can be realized as [ɾ x] in a coda environment, or it could be deleted if word-final. Also, in the coda environment, [s z] are allophones, determined by regressive voicing assimilation from segment that follows. For example, *mais tempo* [majs ˈtẽpu] 'more time' vs. *mais dentro* [majz ˈdẽtɾu] 'more inside'. In absolute final position, the coda is unvoiced.

As with unlicensed consonant clusters in syllable onsets, the *empty nucleus* can be present word-finally if a word ends with a consonant other than [s z ɾ x]. This commonly occurs with words borrowed from other languages (especially English): *internet* [ĩterˈnɛʧi] 'internet'.

2.5 Lexical Stress

Portuguese is a language whose stress is determined at the lexical level [9]. That is, every word has a deterministic accent. The unmarked stress is usually on the penult. For

[10]European Portuguese has syllabic consonants, and more consonant clusters. Consequently, the two varieties sound quite different, and Brazilians sometimes joke that speakers in Portugal "eat their vowels".

Table 6: Placement of syllable stress in Portuguese

Orthography	Transcription	Translation
sopa	ˈso.pɐ	soup
suporte	su.ˈpɔɾ.ʧi	support
superlativo	su.ˌpeɾ.la.ˈʧi.vu	superlative
superestrutura	ˌsu.peɾ.ˌes.tɾu.ˈtu.ɾɐ	superstructure
falam	ˈfa.lɐ̃	speak
andaram	ɐ̃.ˈda.ɾɐ̃	walked
sofá	so.ˈfa	sofa
sonâmbulo	so.ˈnɐ̃.bu.lu	sleepwalker
supersônico	ˌsu.peɾ.ˈso.ni.ku	supersonic
supõe	su.ˈpõj̃	suppose.3SG.PRES
amanhã	ˌa.mɐ̃.ˈɲɐ̃	tomorrow
subi	su.ˈbi	climb.1SG.PAST
subirei	ˌsu.bi.ˈɾej	climb.1SG.FUT
subir	su.ˈbi(ɾ)	climb.INF
suor	su.ˈɔ(ɾ)	perspiration

long words, secondary stress normally falls two syllables before the primary stress (and this recurses for further stresses).

When a letter is orthographically marked with a diacritic accent, that vowel is the nucleus of the tonic syllable. This is done to mark stress on the ultima or antepenult[11]. Explicitly marked nasal accents on vowels also indicate the placement of stress; vowels which are nasalized due to proximity to a nasal consonant (in orthography) are not necessarily tonic.

A very notable exception to the general stress rule concerns verbal forms. Infinitives, singular preterites, and futures[12] are tonic on the last syllable; this is not orthographically marked. All infinitives end with with the suffix -r; possibly as a consequence, any word ending with the letter ⟨r⟩ has a tonic last syllable, unless explicitly accented elsewhere[13].

[11] Appending enclitic pronouns onto verbal forms allows a few words to be stressed on the preantepenult: *chamávamos-te* [ʃa.ˈma.va.mos.ʧi] 'we called you' [11]

[12] ...except third-person plural future

[13] *ímpar* [ˈĩpaɾ] 'odd (integer)'

3 The orthography of Portuguese

The previous section described the phonetic inventory of Portuguese, with some reference to the phonological attributes of the language. Unfortunately, a thorough discussion of the phonological processes of Portuguese cannot be provided in this paper. However, a great deal of the conditioning of Portuguese sounds is captured in the logical and legally standardized [10] orthographic system. Because Portuguese orthography is phonological and conservative [1], the mapping from written text to phonetic representation (i.e. an intermediate function of a text-to-speech system) is fairly straight-forward; this task is far simpler than in English.

In fact, reading Portuguese is nearly analogous to a phonological procedure. Orthographic letters are can be thought of as phonemes in an underlying representation. Therefore, reading equates to applying the appropriate phonological rules (in order) to derive the phonetic representation.

Here I present the rules for translation from orthographic tokens, denoted by angle brackets ⟨ ⟩, into sounds, phonetically transcribed within braces [ ]. The transformations are presented in a format suggestive of a grammar for a formal language; I have aimed to prepare the rules of a grammar that is unambiguous and parseable[14].

3.1 Vowels

⟨ a ⟩	→	[a]	
⟨ á ⟩	→	[a]	
⟨ à ⟩	→	[aː]	
⟨ â ⟩	→	[ɐ]	
⟨ a ⟩	→	[ɐ]	word-finally
⟨as⟩	→	[ɐs]	word-finally
⟨ e ⟩	→	[e] or [ɛ]	(see Section 3.1.1)
⟨ e ⟩	→	[i]	word-finally
⟨es⟩	→	[is]	word-finally
⟨es⟩	→	[is]	word-initially, before stressed syllable
⟨ é ⟩	→	[ɛ]	
⟨ ê ⟩	→	[e]	
⟨ i ⟩	→	[i]	
⟨ í ⟩	→	[i]	
⟨ o ⟩	→	[o] or [ɔ]	(see Section 3.1.1)
⟨ o ⟩	→	[u]	word-finally
⟨os⟩	→	[ɐs]	word-finally
⟨ ó ⟩	→	[ɔ]	
⟨ ô ⟩	→	[o]	
⟨ u ⟩	→	[u]	
⟨ ù ⟩	→	[u]	

[14] If this is truly a well-formed grammar, thn these productions and transformations could be ordered and then directly given as input to an automatic parser generator, such as 'Lex', 'YACC', or 'Bison' [12] [13]. This would output a text-to-transcription program, in one easy step.

Five letters ⟨a e i o u⟩ are used to represent the eight principal vowels [a ɐ e ɛ i o ɔ u]. One of these must necessarily be in the nucleus of every syllable. The following rules (at left) apply to the five vowel letters when *there are not two vowels in sequence*.

Note a unique rule: ⟨es⟩ → [is] word-initially, before a stressed syllable. At this point, we have no understanding of the syllable structure or lexical stress. These will be addressed in Sections 3.4 and 3.5. This rule is postponed until later in the analysis; until then, use: ⟨es⟩ → [es] word-initially.

3.1.1 Distribution of high-mid and low-mid vowels

One of the most problematic issues I encountered was the distribution of high-mid and low-mid vowels. After much examination and many attempts at formulating patterns, I have to conclude that the distribution is not perfectly deterministic. When the accent is not marked, it is virtually impossible to explain why some words are pronounced with high-mid or low-mid vowels. For example, some seemingly contradictory pronunciations of ⟨e⟩:

high-mid	low-mid
ele [ˈeli] ‘he’	ele [ˈɛli] ‘ell (letter L)’
	’ela [ˈɛlɐ] ‘she’
pela [ˈpelɐ] ‘by’	pele [ˈpɛli] ‘skin’

Searching through a dictionary [14], I looked at the effect of syllable onsets, both preceding and following. I also considered the influence of stress, and noted the types of vowels in adjacent syllables. I was unable to detect any pattern and have reason to believe that there is no obvious phonological rule. I hypothesize that the distribution of [e o] and [ɛ ɔ] is irregular. This is something that cannot be predicted and must be learned from usage. Two arguments are provided in support of this hypothesis:

Confusion When encountering a word that I have never heard, I am unsure whether to pronounce it with a high-mid or low-mid vowel. If I write a nonsense word with ⟨e⟩ or ⟨o⟩, I find pairs of pronunciations which I judge to be equally plausible. It would be interesting to test other speakers’ judgements of nonsense words.

Homographs Portuguese has a sizeable set of homographs – spellings with multiple pronunciations (heterophones) – that involve pairs of these vowels. There even exists a 76-page dictionary [15] of these words. The vast majority of these homographs involve semantically related words: a singular noun and a verb conjugated in the first-person singular present tense. It is extremely interesting to note that the high-mid vowel always corresponds to the noun and the low-mid vowel is always the verb. In fact, it seems to be true that all present tense verbs (except first-person plural) will employ the low-mid vowel, where by contrast the past tense uses the high-mid vowel.

So while it may be possible to sometimes predict the distribution of these vowels in terms of syntax and semantics, that is certainly not a simple approach to text-to-speech. I will propose another solution to this problem.

It is my speculation that, in Portuguese, higher vowels are “favored”. I believe that Brazilian Portuguese is similar to European Portuguese in that “the usual posture of the

tongue is raised" [11]. A very distinct characteristic of the language is nasalization, which does not occur with low vowels. Also, word-final vowel raising (/e a o/ → [i ɐ u]) seems to signify a tendency towards high vowels.

With regard to the high-mid and low-mid vowel confusion, the higher vowel seems to be underlying. The high-mid vowel seems to occur in a greater number of words, and it is almost exlusive in the word-initial context. In an average speech corpus, however, [e o] does not appear so much more frequently than [ɛ ɔ], because the low-mid vowel is employed in the conjugation of present tense verbs[15].

With respect to this paper's application of text-to-speech (discussed later), I can imagine a few solutions to this uncertainty between high-mid and low-mid vowels:

1. Do a dictionary lookup for word pronunciations. This method will guarantee proper readings for almost all words, except the noun/verb homographs. Such a solution, though, is not computationally efficient as it requires a considerable amount of memory overhead, and I will not pursue this[16].

2. Perform a grammatical analysis of the text and pronounce certain verbal forms with the low-mid vowel; all other words are pronounced with the high-mid vowel. This approach requires an understanding of Portuguese syntax; furthermore, grammatical analysis includes part-of-speech tagging, which involves dictionaries and an understanding of Portuguese morphology.

3. Resolve the ambiguity by choosing one vowel every time, and accepting some degree of error. An experiment could be performed on many listeners to determine the discrimination function for perception of high-mid and low-mid vowels. For speech synthesis, we should use the sound which is most confused; that is, an artificial vowel of middle height. A good prediction would be that this sound is located on a vowel chart at the midpoint between [e] and [ɛ] for front vowels, between [o] and [ɔ] for back vowels. Because it is neither high-mid nor low-mid, the vowel will always sound a slightly incorrect. However, it will hopefully always sound good enough that a listener is able to infer the correct vowel quality from context.

For its ease of implementation, I think the third option is best.

⟨e⟩	→	[e̬]
⟨o⟩	→	[o̬]

or

⟨e⟩	→	[ɛ̝]
⟨o⟩	→	[ɔ̝]

My intuition is that the "halfway" vowel will be closer to high-mid, so [e̬] and [o̬] are more appropriate. To my ears, a mispronounced high-vowel is a more severe error than a mispronounced low-vowel. For details of this experiment, see Appendix A.6.

[15] Present tense is probably the most common tense in everyday speech (I would guess).

[16] Using a pronunciation dictionary would defeat the purpose of this paper: to implement a TTS system that exploits Portuguese orthography's regularity, deriving pronunciations only by using letter-to-phone rules. For more irregular languages, such as English, text-to-speech implementations rely heavily on dictionaries.

3.1.2 A note on vowels and accent marks

Non-nasal accents on vowels primarily function to mark a stressed syllable. The type of accent used can also indicate the quality of a vowel, eliminating ambiguity in some cases.

The acute accent on ⟨í ú⟩ only marks stress and does not affect the vowel height. An acute accent on ⟨é á ó⟩ can be used to denote the lower vowels [ɛ a ɔ]. However, when nasalized by a following ¡m¿ or ¡n¿ (described in later section), these are not lowered; in this instance the acute accent only marks stress.

A circumflex accent on ⟨ê â ô⟩ indicates that the stressed vowel is higher, [e ɐ o]. These vowels are frequently nasal; this is always true for ⟨â⟩.

The grave accent is used on ⟨à⟩ to indicate an orthographic contraction equivalent to ⟨a a⟩. It only occurs in the single syllable words *à* and *às*.

3.1.3 Diphthongs

In the orthography, diphthongs occur when two vowel letters are adjacent. In most cases ⟨i⟩ and ⟨u⟩ represent the glides [j] and [w]. With rising diphthongs, there are quite a few exceptions to this generalization, where two adjacent vowels are in separate syllables [7]. Furthermore, there are a handful of cases where ⟨e⟩ and ⟨o⟩ represent glides.

It is not imprudent to use a broad generalization here. When mispronouncing a two-vowel sequence as a rising diphthong, or vice-versa, a listener may not even notice the error[17]. This is a great advantage to speech synthesis.

However, there is one very common two-vowel sequence which is *not* diphthongized: ⟨ia⟩ in a word-final position. This sound segment is part of the morphemes that mark past imperfect tense and future conditional tense. For these suffixes to be recognized, it is very important that they be pronounced as two distinct vowels, with the primary lexical stress on the penultimate syllable.

It is also important to note that accented ⟨í⟩ and ⟨ú⟩ are *not* glides. Compare: pais [ˈpajz] ‘fathers’ vs. país [paˈiz] ‘country’.

With the ambiguous ⟨iu⟩ and ⟨ui⟩ combinations, I felt that a falling diphthong sounded slightly better than either a rising dipthong or a two-vowel sequence. Triphthongs will not present any complications, so long as the rules are applied to the letters with right-to-left precedence[18].

⟨ i ⟩	→	[j]	when adjacent to another vowel
⟨ u ⟩	→	[w]	when adjacent to another vowel
⟨iu⟩	→	[iw]	
⟨ui⟩	→	[uj]	
⟨ia⟩	→	[ˈiɐ]	word-finally
⟨ei⟩	→	[e]	word-initially or word-medially
⟨ou⟩	→	[o]	([ow] only in careful speech)

These rules apply for the majority of situations. Note that in Brazilian Portuguese, ⟨ei⟩ and ⟨ou⟩ are not always diphthongs, but are pronounce as just [e] and [o].

[17] It should probably not even be considered an “error”, so much as an alternation

[18] This also ensures that word-final vowel raising occurs prior to diphthongization.

3.2 Consonants

The consonants in my variety of Brazilian Portuguese have a fairly wide, but strongly predictable mapping from the orthography. The six plosives can be written with six letters:

⟨p⟩	→	[b]	
⟨b⟩	→	[b]	
⟨t⟩	→	[t]	except before [i]
⟨d⟩	→	[d]	except before [i]
⟨c⟩	→	[k]	except before ⟨i⟩ or ⟨e⟩
⟨g⟩	→	[g]	except before ⟨i⟩ or ⟨e⟩

The last rules are due to the difficulty in articulating a velar plosive before a front vowel. In such an environment ⟨c⟩ and ⟨g⟩ represent alveolar fricatives. Many Brazilian dialects affricate stops before a high front vowel:

⟨t⟩	→	[ʧ]	before [i]
⟨d⟩	→	[ʤ]	before [i]

Note that this is a *phonological* rule, not a strictly orthographic one[19]; hence the phonetic brackets in the conditioning environment shown above. (cf. *bate* ['baʧi] 'hits.3SG.PRES' vs. *bateu* [ba'tew] 'hit.3SG.PAST')

Some velar plosives are written with a vowel ⟨u⟩ appended to a consonant: ⟨qu⟩ or ⟨gu⟩. When these combinations are followed by ⟨i⟩ or ⟨e⟩, the letter ⟨u⟩ is unpronounced; elsewhere, it is a velar approximant. The umlaut ⟨ü⟩ appears in a highly specialized context, representing [w] before ⟨i⟩ or ⟨e⟩.

⟨qu⟩	→	[k]	before ⟨i⟩ or ⟨e⟩
⟨qu⟩	→	[kw]	... elsewhere
⟨qü⟩	→	[kw]	
⟨gu⟩	→	[g]	before ⟨i⟩ or ⟨e⟩
⟨gu⟩	→	[gw]	... elsewhere
⟨gü⟩	→	[gw]	

⟨ç⟩	→	[s]	
⟨c⟩	→	[s]	before ⟨i⟩ or ⟨e⟩
⟨sc⟩	→	[s]	before ⟨i⟩ or ⟨e⟩
⟨ss⟩	→	[s]	
⟨s⟩	→	[s]	word-initially
⟨s⟩	→	[s]	before a voiceless sound
⟨s⟩	→	[s]	absolute final
⟨s⟩	→	[z]	word-medially, before a voiced sound
⟨s⟩	→	[z]	word-finally (following context: voiced)
⟨z⟩	→	[z]	

[19] This has been carelessly misdocumented in some texts [1]

The alveolar fricatives have many conditioning environments, as shown above. The voiceless palatoalveolar fricative is represented by a digraph ⟨ch⟩.

⟨ch⟩	→	[ʃ]	
⟨ j ⟩	→	[ʒ]	
⟨ g ⟩	→	[ʒ]	before ⟨i⟩ or ⟨e⟩

The letter ⟨x⟩ represents the most variable consonant sound in Portuguese. Its pronunciation depends on the letters before and after it. Following ⟨i⟩ and ⟨u⟩, its pronunciation depends on whether it is following a vowel [i u] or an approximant [j w].

⟨x⟩	→	[ʃ]	after ⟨{a,e,o,u}i⟩;	peixe ['pejʃi] 'fish'
⟨x⟩	→	[ks]	after ⟨i⟩;	fixo ['fiksu] 'fixed'
⟨x⟩	→	[s]	after ⟨{a,e,i,o}u⟩;	auxílio [aw'silju] 'auxiliary'
⟨x⟩	→	[ʃ]	after ⟨u⟩;	bruxa ['bruʃɐ] 'witch'

Following ⟨a⟩, the letter ⟨x⟩ will have a somewhat unpredictable distribution. To cover the greatest number of cases, I use one simple rule. There are a few loanwords[20] which need explicit treatment by a text-to-speech system. Nonetheless, ⟨ax⟩ is an environment which occurs with low frequency.

⟨x⟩	→	[ʃ]	after ⟨a⟩;	taxa ['taʃɐ] 'tax'

Following ⟨e⟩ or ⟨o⟩, the letter ⟨x⟩ will have a different pronunciation if it is in a word-initial syllable. Some loanwords[21] stray from the rules below.

⟨x⟩	→	[ks]	after ⟨e⟩	sexo ['sɛksu] 'sex'
⟨x⟩	→	[z]	after word-initial ⟨e⟩	exato [e'ʒatu] 'exact'
⟨x⟩	→	[s]	after ⟨e⟩; before ⟨p⟩ or ⟨t⟩	extra ['estɾɐ] 'extra'
⟨x⟩	→	[ʃ]	after ⟨o⟩	coxa ['koʃɐ] 'thigh'
⟨x⟩	→	[ks]	after word-initial ⟨o⟩	óxido ['ɔksidu] 'oxide'

The word-initial context is straighforward. Also, a rule is needed for the ⟨xc⟩ combination, to prevent other rules from forming a doubled sound [ss].

⟨ x ⟩	→	[ʃ]	word-initially	xampu [ʃɐ̃'pu] 'shampoo'
⟨xc⟩	→	[s]	after ⟨e⟩; before ⟨e⟩ or ⟨i⟩	excelente [ese'lẽtʃi] 'excellent'
⟨xc⟩	→	[sk]	after ⟨e⟩; ... elsewhere	exclusivo [esklu'zivu] 'exclusive'

[20] táxi ['taksi] 'taxi'; sintáxe [sĩ'tasi] 'syntax'; saxofóne [sakso'foni] 'saxophone'

[21] boxe ['bɔksi] 'boxing'; oxalá [oʃa'la] 'God willing'

The letters ⟨l⟩ and the digraph ⟨lh⟩ represent the lateral consonants. In Brazilian Portuguese, word-final ⟨l⟩ is diphthongized[22].

⟨ l ⟩	→	[l]	
⟨ l ⟩	→	[w]	word-finally
⟨lh⟩	→	[ʎ]	

Despite their substantial variation across many dialects, the rhotics fall under well-defined environments.

⟨ r ⟩	→	[x]	word-initially
⟨ r ⟩	→	[ɾ]	word-medially
⟨ r ⟩	→	{silent}	word-finally
⟨rr⟩	→	[x]	

The letters ⟨m⟩ and ⟨n⟩ can represent nasal consonants when preceding a vowel, or nasalization on a vowel that it follows (see Section 3.3). The digraph ⟨nh⟩ represents the palatal nasal and only occurs word-medially.

⟨ m ⟩	→	[m]	before ⟨a⟩, ⟨e⟩, ⟨i⟩, ⟨o⟩, or ⟨u⟩
⟨ n ⟩	→	[n]	before ⟨a⟩, ⟨e⟩, ⟨i⟩, ⟨o⟩, or ⟨u⟩
⟨nh⟩	→	[ɲ]	

Except when indicating palatalization in the digraphs ⟨ch⟩, ⟨lh⟩, or ⟨nh⟩, the letter ⟨h⟩ is silent.

⟨h⟩	→	{silent}	except after ⟨c⟩, ⟨l⟩, or ⟨n⟩

3.3 Nasalization

Orthographic markings for nasalization are given by ⟨m⟩ or ⟨n⟩ following a vowel, and a nasal consonant is not pronounced[23]; this practice is an example of the phonological nature of written Portuguese [1]. Due to the bilabial articulation ⟨m⟩ precedes ⟨p⟩ or ⟨b⟩, while ⟨n⟩ occurs before other consonants.

⟨am⟩	→	[ɐ̃]	before ⟨p⟩ or ⟨b⟩; or word-finally
⟨ an ⟩	→	[ɐ̃]	before other consonants
⟨em⟩	→	[ẽ]	before ⟨p⟩ or ⟨b⟩; or word-finally
⟨ en ⟩	→	[ẽ]	before other consonants
⟨ im ⟩	→	[ĩ]	before ⟨p⟩ or ⟨b⟩; or word-finally
⟨ in ⟩	→	[ĩ]	before other consonants
⟨om⟩	→	[õ]	before ⟨p⟩ or ⟨b⟩; or word-finally
⟨ on ⟩	→	[õ]	before other consonants
⟨um⟩	→	[ũ]	before ⟨p⟩ or ⟨b⟩; or word-finally
⟨ un ⟩	→	[ũ]	before other consonants

[22] European Portuguese has a word-final velarized [ɬ]

[23] A nasal consonant should be pronounced between two vowels. That is, apply the rules from right-to-left, starting at the end of a word. The preceding vowel should still be nasalized (and [-low]): *cama* as [ˈkɐ̃mɐ] instead of [ˈkamɐ].

When word-final, nasal vowels can be diphtongized in some dialects. ⟨n⟩ does not occur word-finally[24].

⟨am⟩	→	[ɐ̃w̃]	word-finally
⟨em⟩	→	[ẽj̃]	word-finally
⟨om⟩	→	[õw̃]	word-finally

The diacritic tilde can also nasalize a vowel, used on ⟨ã⟩ and ⟨õ⟩, often to construct nasal diphthongs[25] such as: ⟨ão⟩ → [ɐ̃w̃]; ⟨ãe⟩ → [ɐ̃j̃]; ⟨õe⟩ → [õj̃].

⟨ã⟩	→	[ɐ̃]
⟨õ⟩	→	[õ]

3.4 Syllable structure

After all the rules from the previous section have been applied to yield phonetic elements, it should be possible to identify the syllable structure as discussed in Section 2.4. This equates to the following productions:

σ	→	(O) R
O	→	{Fric, Plos, Liq}
	→	Plos Liq
	→	[f] Liq
	→	[ɲ]
R	→	N (Cod)
N	→	(Glid) Vow (Glid)
	→	∅
Cod	→	{[s],[z],[x],[ɾ]}
Fric	→	{[f],[v],[s],[z],[ʃ],[ʒ],[x],[ʧ],[ʤ]}
Plos	→	{[p],[b],[t],[d],[k],[g]}
Liq	→	{[ɾ],[l]}
Vow	→	{[i],[e],[ɛ],[a],[ɔ],[o],[u],[ɐ]}
Glid	→	{[j],[w]}

Note that there is a production for the empty nucleus: N → ∅. This phenomenon occurs in Brazilian Portuguese[26] with words of foreign origin that include unlicensed consonant clusters. The empty nucleus is pronounced as [i]. For example: *pneu* [piˈnew] 'tire'; *advogado* [ˌaʤivoˈgadu] 'lawyer'.

To include syllable boundaries in the phonetic transcription, the above rules should be used to parse syllables, and then insert . in between.

$\sigma\sigma$	→	$[\sigma.\sigma]$

[24] *Leblón* [leˈblõ] or [leˈblõw̃] is a neighborhood in Rio de Janeiro, but is a French name.

[25] -ão and -ões are analogs of Spanish morphemes -ion and -iones, which are nasal diphthongs. Portuguese: *nação nações* [naˈsɐ̃w̃ naˈsõjs] 'nation nations'. Spanish: *nación naciones* [naˈsjõn naˈsjõnes] 'nation nations'.

[26] ... as opposed to the European variety, which allows many consonant clusters in the onset.

3.5 Lexical Stress

Lexical stress is well-marked in Portuguese orthography. The primary stress is on the penult, with secondary stress falling every other syllable prior. After parsing the syllable structure described in the previous section, stress can be determined. (# marks a word boundary)

$\sigma\sigma\#$	$\rightarrow$	$[{}'\sigma\sigma]$	Primary stress on penult
$\sigma\sigma{}'\sigma$	$\rightarrow$	$[{}_{,}\sigma\sigma{}'\sigma]$	Secondary stress two syllables before primary
$\sigma\sigma_{,}\sigma$	$\rightarrow$	$[{}_{,}\sigma\sigma_{,}\sigma]$	Secondary stress two syllables before secondary

When marked with an explicit diacritic accent, a vowel is stressed in lieu of the language's general stress pattern. Words that have a tonic ultimate or antepenultimate syllable are usually orthographically accented. A single word cannot have more than one of the acute, circumflex, or tilde accent. (Umlaut ⟨ü⟩ is a glide. Grave ⟨à⟩ only appears in monosyllables.)

$\sigma = (\text{O})\ \acute{\text{N}}\ (\text{Cod})$	$\rightarrow$	$[{}'\sigma]$
$\sigma = (\text{O})\ \hat{\text{N}}\ (\text{Cod})$	$\rightarrow$	$[{}'\sigma]$
$\sigma = (\text{O})\ \tilde{\text{N}}\ (\text{Cod})$	$\rightarrow$	$[{}'\sigma]$

The morphology of Portuguese verbs often imposes irregular stress. As explained in Section 2.5, words that end in ⟨r⟩ are last-syllable tonic because they usually[27] have infinitive suffixes (-ar, -er, -ir). Also, words ending with ⟨i⟩ are stressed finally because they have the morpheme for first-person past (-ei, -ei, -i), first-person future (-arei, -erei, -irei), or third-person present(-i)[28]. Final ⟨u⟩ indicates the third-person past (-ou, -eu, -iu) suffixes.

Aside from the last-syllable tonic conjugations, there are no other words that end with ⟨i⟩ or ⟨u⟩. This is because the penultimately stressed words that end with [i] or [u] are spelled with ⟨e⟩ or ⟨o⟩, and then undergo word-final raising when pronounced. This convention causes a complementary distribution of stress environments, such that word-final ⟨í⟩ or ⟨ú⟩ is somewhat redundant; the orthography allows stress to be inferred when a diacritic mark is omitted from atop word-final ⟨i⟩ or ⟨u⟩ [29].

$\sigma\#$	=	$(\text{O})\ \text{N}\ \langle\text{r}\rangle\#$	$\rightarrow$	$[{}'\sigma]$
$\sigma\#$	=	$(\text{O})\ \langle\text{i}\rangle\ \#$	$\rightarrow$	$[{}'\sigma]$
$\sigma\#$	=	$(\text{O})\ \langle\text{ei}\rangle\ \#$	$\rightarrow$	$[{}'\sigma]$
$\sigma\#$	=	$(\text{O})\langle\text{ou}\rangle\ \#$	$\rightarrow$	$[{}'\sigma]$
$\sigma\#$	=	$(\text{O})\langle\text{eu}\rangle\ \#$	$\rightarrow$	$[{}'\sigma]$
$\sigma\#$	=	$(\text{O})\langle\text{iu}\rangle\ \#$	$\rightarrow$	$[{}'\sigma]$

It should be noted that when the empty nucleus is realized as [i], its syllable $\sigma_{\emptyset}$ is never stressed. The empty nucleus should be ignored when counting syllables to determine primary and secondary stress placement.

[27] A few nouns and adjectives end with ⟨r⟩, but those are also last-syllable tonic.

[28] Irregular: *vai, cai*

[29] The accented ⟨í⟩ is retained following a vowel if it might be confused for a diphthong: *saí* [sa'i] 'depart.1SG.PAST' vs. *sai* ['saj] 'depart.3SG.PRES'

$\sigma_1\ \sigma_\emptyset\ \#$	$\rightarrow$	$[\text{ˈ}\sigma_1\sigma_\emptyset]$	internet [ˌĩ.teɾ.ˈnɛ.ʧi] ‘Internet’
$\#\ \sigma_\emptyset\ \sigma_1\ \#$	$\rightarrow$	$[\sigma_\emptyset\text{ˈ}\sigma_1]$	pneu [pi.ˈnew] ‘tire’
$\sigma_2\ \sigma_\emptyset\ \sigma_1\ \#$	$\rightarrow$	$[\text{ˈ}\sigma_2\sigma_\emptyset\sigma_1]$	pacto [ˈpa.ki.tu] ‘pact’
$\sigma_3\ \sigma_2\ \sigma_\emptyset\ \text{ˈ}\sigma_1$	$\rightarrow$	$[\text{ˌ}\sigma_3\sigma_2\sigma_\emptyset\text{ˈ}\sigma_1]$	readmita [ˌxe.a.ʤi.ˈmi.tɐ] ‘readmits’
$\sigma_3\ \sigma_\emptyset\ \sigma_2\ \text{ˈ}\sigma_1$	$\rightarrow$	$[\text{ˌ}\sigma_3\sigma_\emptyset\sigma_2\text{ˈ}\sigma_1]$	readmissão [xe.ˌa.ʤi.mi.ˈsɐ̃w̃] ‘readmission’

By this point, an analysis of Portuguese orthography has produced information about phonetic elements, syllable structure, and stress placement. Finally, it is possible to evaluate the postponed rule from Section 3.1:

⟨es⟩	$\rightarrow$	[is]	word-initially, before stressed syllable

4 Text-to-speech Implementation

The previous part of this paper discussed how Portuguese orthography can be systematically converted into a phonetic representation. This is essentially the human task better known as 'reading aloud'. For humans, who have a subconscious mastery of language, reading is not an especially difficult process. For a machine, however, natural-sounding speech production is a monumental achievement[30]. This section of the paper describes how the previous portions can be integrated in a computer program.

4.1 Components of text-to-speech synthesis

Text-to-speech synthesis involves a long series of processes that convert digital text input to intelligible audio output.

Preprocessing 1. Before other processing, the input undergoes text normalization. This translates one string of characters into another, expanding literal word tokens such as abbreviations and numerals. This step is "mundane but critical to high-quality synthesis" [16].

Pronunciation 2. Next, a function maps the text to phonemes by using lexicons and orthographic rules. The traditional approach involves involves looking up entries in a pronunciation dictionary. Pronunciations that are not found are then estimated using letter-to-sound rules. For Portuguese, this is sufficient.

Prosody 3.A sequence of adjustments are made to the base pronunciations derived in the previous step. This determines the necessary pitch, duration, and amplitude of each sound unit. Such a process involves a high-level understanding of the language's syntax and semantics. Grammatical analysis can, in some cases, be used to correctly predict proper phrasing. More often, though, prosody is not entirely characterized by text and requires external knowledge. Among the factors that affect the perceived quality of synthesized speech, "prosody is foremost" [17].

Signal 4. The audio output heard by a human must be recognized as speech. There are three approaches to synthesize the sound units [16]:

Articulatory synthesis Physical models of articulators' movements.

Source-filter synthesis (formant synthesis) A tone complex is filtered in ways that mimic the resonance of the human speech organs, to shape the spectral definition of a speech sound.

Concatenative Synthesis Combining pre-recorded segments of speech. The typical unit is a diphone consisting of a consonant and vowel.

Given the orthographic rules of Portuguese, it is possible to implement a text-to-speech system up to the second step. This could then be considered a text-to-transcription system.

[30]The state-of-the-art: IBM and ATT have what I consider the best text-to-speech systems currently available. (`http://research.ibm.com http://research.att.com`)

4.2 Text preprocessing

Given an input character string (ASCII text), the text pre-processing stage of text-to-speech produces a string of only words that are in the proper format for input to the next step, word pronunciation. Such text processing is well suited to a programming (or scripting) language such as Perl [18], which has an extensive facility for processing regular expressions and pattern matching.

To illustrate the process, consider our text input stored in a string:

```
$input = 1 Cetim!
```

First, for simplicity, convert everthing to lowercase. Also, since this system will be insensitive to prosody, we will disregard sentential organization. Strip the input of all non-alphanumeric[31] characters:

```
$input =~ tr/A-ZAÀ-Ý/a-yà-ý/;            #Uppercase to lowercase
$input =~ tr/A-Za-zÀ-Ýà-ý0-9\s\-//cd;    #Delete others
```

Abbreviations are a closed class: a finite set with known mappings. The only way to expand them is to have a database of abbreviations (i.e. a dictionary). Acronyms in Brazilian Portuguese are commonly read out as words instead of individual letters. (*USP* ['uspi] 'University of São Paulo') I think this occurs less frequently in English.

Expand numerals one digit at a time. As a Perl hashtable:

```
%abbreviation_map = qw(dr    doutor
                       sra   senhora
                       eua   estados unidos da america
                       lt    litro);       #and many more...
%digit_map = qw(1 um 2 dois 3 tres 4 quatro 5 cinco
                6 seis 7 sete 8 oito 9 nove 0 zero);
```

There are many other ways, of course, that numbers could be read. A more complete system would treat large numbers, telephone numbers, dates, ordinals, etc...

There are several ways to read a string of numbers: the digit-for-digit system is a 'serial' reading ('one, two, three, four'); a 'combined' reading is a single integer (one thousand, two hundred thirty-four); 'paired' (twelve thirty-four) [19]. For Portuguese, when reading a phone number, the number six, *seis*, is often called *meia* (meaning 'half', referencing the face of a clock).

For each word in the input string, search the database of abbreviations to see if it can be expanded. This is also a good time to check for words which are known to have irregular pronunciations[32]. After iterating for each list, searching in the input string for words that can be expanded, our input string should be prepared for the next step:

```
$input = um cetim
```

[31] ...except the whitespace and the hyphen character

[32] `$input =~ s/sintáxe/sintaze/;`

4.3 Word pronunciation

4.3.1 Letter-to-phone mapping

Implementing the mapping from orthography to phonetic realization is simply a matter of applying letter-to-phone rules. Because Portuguese has a reliable phonological orthography, there is no need to waste memory on a dictionary of pronunciations.

In Perl, regular expression substitutions could be used to translate the input string *in situ*. This has the advantage of being efficient, but requires a system for marking which letters have already been processed. Here I use delimiting braces `[]`to indicate a phoneme:

```
$input =~ s/p/[p]/g;        #  <p> --> [p]
```

The above rule has no conditioning environment. Consider:

```
$input =~ s/c([ie])/[s]\1/g;  # <c> --> [s] before <i> or <e>
```

This rule has an orthographic conditioning environment. Applied to the input, the results is `$input = um [s]etim` Applying some rules to vowels:

```
$input =~ s/e/[e]/g;        #  <e> --> [e]
$input =~ s/i/[i]/g;        #  <i> --> [i]
$input =~ s/u/[u]/g;        #  <u> --> [u]
```

This results in `[u]m [s][e]t[i]m`. The choice of high-mid vowel is convenient, but not always correct (as discussed in Section 3.1.1. The braces allow us to indicate a phonological environment. But in this case it is required that the ordering of the rules be such that the vowel rule for `[i]` was processed prior to the following rules:

```
$input =~ s/t([i])/[{\\textteshlig}\1/g; # t --> tS before [i]
```

This results in `[u]m [s][e][{\textteshlig}][i]m`. The phonetic representation used here is chosen as the keyboard mapping for TIPA[33] fonts for LaTeX typesetting. Perl regular expressions also allow word boundaries in the conditioning environment:

```
$input =~ s/[i]m\b/[\\~{\\i}]/g;    #  <im> --> [~i] word-final
$input =~ s/[u]m\b/[\\~u]/g;        #  <im> --> [~u] word-final
```

After all the above letter-to-phone rules have been applied, we are left with:

```
$input = [\~u] [s][e][{\textteshlig}][\~{\\i}]
```

[33] TIPA= TeX IPA Phonetic Fonts, a powerful font package which I have used in typesetting this paper.

4.3.2 Syllable parsing and stress assignment

Section 3.4 provided a context-free grammar for syllable structure. Such a set of rules can be used to write a LALR(1) parser[34]. The details of parsing are not appropriate to discussion here [**?**] [13], suffice it to say that it is fairly simple to build a parser using an automatic parser generator. Parsing our example according to the transformations in 3.4, and keeping all the information within our character string:

```
$input =  [\~u] [s][e].[{\textteshlig}][\~{\i}]
```

The decision to keep all the information within a character string is arbitrary. It is simpler for the purposes of demonstration, and well-suited to a text-processing language like Perl. Alternatively, the the syllable structure and phonetic elements could be saved in a more object-oriented data structure. For example, to express that the onset of the first syllable in the second word is [s]:

```
Input.Word_array[1].Syllable_array[0].Onset =
     new Phoneme.Consonant.Fricative.Alveolar(Voiced=false);
```

Applying the rules for lexical stress would indicate regular penultimate syllable stress.

```
$input =  [\~u] [s][e]."[{\textteshlig}][\~{\i}]
```

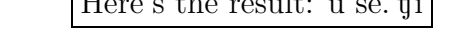
The orthography of Portuguese has provided all the information needed to perform all the letter-to-phone translations, syllable parsing, and placement of lexical stress. After all thi processing, we can filter out all the braces:

```
$input =~  tr/[]//d;       #Delete all braces
```

We have performed a cascade of inline transformations to our input string, resulting in: *input* = *ũse.*"ĩ| This is the phonetic representation of the sound, defining all the sound segments as well as syllabification and stress. Because we conveniently chose to use the phonetic representation of the fonts, this string can be directly written to a file:

```
open(OUT, ">output.tex");                          #Open a .TEX file for writing
print OUT "\\begin{document}\n"                      #Begin document environment
print OUT "\\fbox{Here\'s the result: ";                                #Draw a box
print OUT "\\textipa{\\~u se.\"{\\textteshlig}\\~{\\i}}}\n";          #Sinput
print OUT "\\end{document}"                            #End document environment
```

This is essentially an entire text-to-transcription program! It outputs:



[34] 1-token LookAhead Left-to-right parse, Rightmost derivation. A standard grammar.

4.4 Prosody and Signal Processing

A high quality text-to-speech system will direct the majority of its attention to the prosodic features of speech. It is not enough to determine the pronunciation of words in isolation; especially in Portuguese, intonation is sententially dependent, semantically important, and far from monotonic. See Figure 9 in Appendix A.6 for an example of the rising-falling intonation pattern which occurs in normal speech. Unfortunately, this implementation of text-to-speech is not sophisticated enough to handle Portuguese prosody.

The last component of the text-to-speech chain is a significant technical challenge. Synthesizing an entirely synthetic voice requires a computational model relating sounds to the mechanics of the speech articulators. The concatenative approach to speech synthesis, which can use pre-recorded human speech segments, is less complicated but involves processing a digital signal to properly blend segment boundaries; otherwise the resulting speech sounds choppy.

Given the phonetic representation that is ultimately derived by the system described in this paper, concatenative synthesis could be straightforwardly implemented[35] by assigning a short sound file to each phonetic segment. The files are concatenated in order:

```
%phone-file_map = qw (
    i             high_front.dat   #.dat file is .wav stripped of its header
    \~{\i}        high_front_nasal.dat
    "i            high_front_stressed.dat);                #And many more...
open(OUT, ">>emptyheader.wav");            #Open an output sound file (.wav)
$size = 0;   #The file consist only of a header, and is opened for appending
while ($input) {
  $next = nextphone($input);             #Tokenize the input, get next phone
  $data = $phone-file_map{$next};                #Lookup data for that phone
  print OUT $data;                       #Append the data to the output file
  $size += size($data);                                    #Increment size
}
fixheader(OUT, $size);                      #Adjust the output file's header
close(OUT);                                                   #All done!
```

This procedure will simply glue a bunch of sound files back to back. Without any smoothing of the output, the resulting signal is nearly unintelligible because the jumps between sound segments would be too abrupt. A solution would be to use diphones, sounds which cover two phonetic elements, but only cover the last half of the first sound and the first half of the second sound. This reduces the choppiness of the output since the diphones span segment transitions and are joined at their middles.

5 Conclusion

In this paper, I have described how the phonological attributes of Portuguese orthography may be exploited to develop a text-to-speech system. The phonetic nature of the language

[35] http://www.ocf.berkeley.edu/ arlo/ling110|

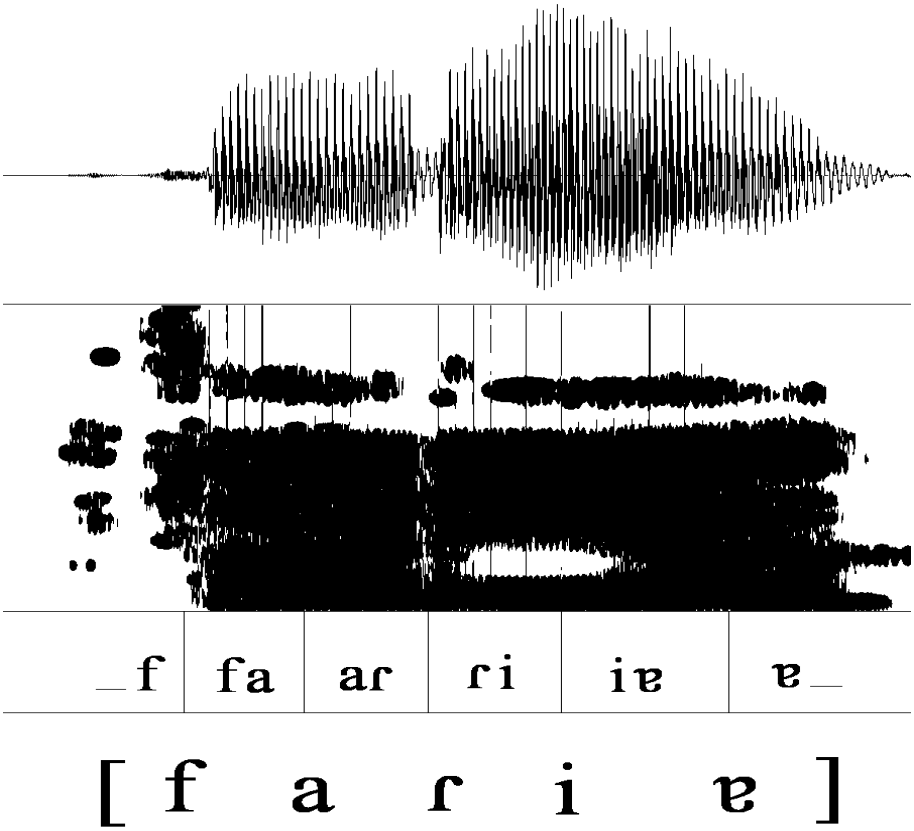


Figure 4: Concatenative analysis of *faria* using diphones

was surveyed, and the set of rules for deriving these sounds using the orthography were made explicit and carefully deconstructed.

I believe that this project has been a very fruitful experience for me. Preparing the first half of a text-to-speech system involved much more research than I anticipated. I did not expect to find so many intricacies in a language that I have been speaking and reading so effortlessly all my life.

Although I was disappointed that I was unable to explore the final signal processing aspect of the task, in the near future I would love to complete the text-to-speech system.

References

[1] M. H. Mateus, E. d'Andrade. *The Phonology of Portuguese*. THE PHONOLOGY OF THE WORLD'S LANGUAGES. Oxford University Press Inc., New York, 2000.

[2] *Ethnologue, Languages of the World, 14th Ed.*. Summer Institute of Linguistics, January 2003. ⟨`http://www.ethnologue.com`⟩

[3] Mauricio Brito Carvalho. Professor of Linguistics. UNIRIO. *Electronic correspondence*. Email messages to Arlo Faria, 10-16 May 2003. ⟨macbrito@centroin.com.br⟩

[4] M. R. D. Martins. *Ouvir Falar: Introdução à Fonética do Português*. Editorial Caminho, Lisbon, 1988.

[5] Luiz Francisco Dias. Professor, Departamento de Linguística, UFMG. *Electronic correspondence*. Email message to M. B. Carvalho, 12 May 2003.

[6] *PlotFormant Version 4.0*. Scicon R D, Inc. Los Angeles, 10 October 2002. ⟨`http://linguistics.ucla.edu/faciliti/facilities/acoustic/acoustic.html`⟩

[7] M. M. Azevedo. *Elementos de Fonética Articulatória*. Unpublished (Distributed as a handout for a UC Berkeley course, Portuguese 135.6: Portuguese Linguistics) Fall 2001.

[8] P. Ladefoged. *A Course in Phonetics*. Harcourt Brace Jovanovich, Inc., Fort Worth, 1993.

[9] G. M. Rio-Torto. *Fonética, Fonologia, e Morfologia do Português*. Edições Colibri, Lisbon, 1998.

[10] S. Farina. *A Nova Ortografia*. Editora Movimento, Porto Alegre, 1971. (Description of Law #5765, passed by the national Congress to standardize orthography)

[11] M. Cruz-Ferreira. *Portuguese (European)*. (From "Illustrations of the IPA", *Handbook of the International Phonetic Association*.) Cambridge University Press, Cambridge, 1999.

[12] A. Appel. *Modern Compiler Implementation in Java*. Cambridge University Press, New York, 1998.

[13] A. V. Aho, R. Sethi, J. D. Ullman. *Compilers: Principles, Techniques, and Tools*. Pearson Higher Education, 1985.

[14] *Langensheidt's Pocket Portuguese Dictionary*. Langensheidt Publishers, Inc., New York, 1989.

[15] P. Pându. *Novo Dicionário de Acentuação da Palavras: Homógrafas e Heterófonas*. Livros do Mundo Inteiro, Rio de Janeiro, 1972.

[16] B. Gold, N. Morgan. *Speech and Audio Signal Processing*. John Wiley Sons, Inc., New York, 2000.

[17] Kim Silverman (Director, Speech Technologies at Apple Computer). Talk on speech synthesis, given at the International Computer Science Center. Berkeley, 21 April 2003.

[18] L. Wall, T. Christiansn, J. Orwant. *Programming Perl, 3rd edition*. O'Reilly Associates, Inc., Sebastopol, CA, 2000.

[19] D. Jurafksy, J. Martin. *Speech and Language Processing* Prentice-Hall, Inc., Upper Saddle River, NJ, 2000.

[20] D. Mazzoni, J. Haberman, M. Brubeck. *Audacity 1.0.0*. Sourceforge, June 2002. ⟨`http://audacity.sourceforge.net`⟩

[21] P. Boersma, D. Weenink. *Praat 4.0.49*. Institute of Phonetic Sciences, University of Amsterdam, 12 March 2003. ⟨`http://www.fon.hum.uva.nl/praat`⟩

R. C. P. da Silveira. *Estudos de Fonética do Idioma Português*. Cortéz Editora, São Paulo, 1982.

R. Gonçalves. *Tratado de Ortografia da Lingua Portuguesa*. Atlântida, Coimbra, 1947.

M. Parreira. *Prontuário Ortográfico Moderno*. Edições Asa, Lisbon, 1985.

I alsoowe many thanks to the people who helped me with this project:

- Prof. Dr. Mauricio Brito Carvalho, a good friend in Rio de Janeiro, who offered his wealth of knowledge, and years of experience studying the Portuguese language.
- Angela, Mauricio's wife, who knew more words than he did.
- Prof. Dr. Luiz Carlos Cagliari, a retired phonetician at UNICAMP, for a great response to my question about Brazilian vowels.
- Prof. Dr. Luiz Francisco Dias, who cleared up some the misconceptions of the "educated standard" dialect of Brazilian Portuguese.
- Julie Larson, my graduate student instructor, for lending me her books.
- Ryan Shosted, Ph.D. student at U.C. Berkeley, for his help with the Portuguese nasals.

A Recorded Data

A.1 Recording Process

The author recorded his own speech and based measurements of Brazilian Portuguese vowels upon this. Additionally, audio recordings of the diphones used in the speech synthesis were extracted from the author's speech. The recordings were made in a non-ideal environment:

- 12'x8'x9' room with concrete walls.
- The speaker was at coordinates (2',2',4').
- An omnidirectional headset microphone was used.
- A personal computer at (1',2',2') contributed noise.

There was a considerable amount of noise in the recordings, which was removed by using noise-cancelling software [20]. This post-capture signal processing satisfactorily removed the background noise, resulting in a signal of quality comparable to one from a soundproof laboratory.

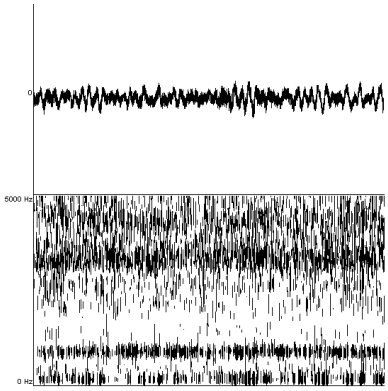


Figure 5: Noise Profile

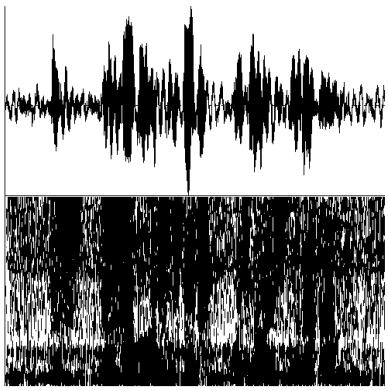


Figure 6: Original Signal

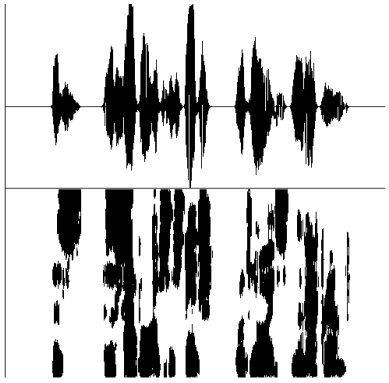


Figure 7: Processed Signal

The recording process is illustrated in the figures above. First, a noise profile (Figure 5) is recorded in the absence of an input speech signal. This noise profile is present in the original signal (Figure 6), but can be removed by the noise-cancelling software. The processed signal (Figure 7) is constructed by inserting complete silence in the portions of the signal which match the noise profile; in the intervals that include the speech signal, it is added to an inverse of the noise signal, removing the noise by destructive interference. An unfortunate side effect of the noise-cancelling algorithm is that it is a smoothing function, causing the loss of high-frequency portions of the signal. For speech, however, this lowpass filtering has a negligible effect.

The sound samples for this section are saved as:

`noise_profile.wav`	A few seconds of noise
`speech_sample.wav`	The original signal: "This is ... speech with noise ... "
`clean_speech.wav`	The processed signal, after noise removal.

Table 7: Words recorded for measurement of Portuguese vowels

Vowel	Portuguese word	Phonetic transcription	English translation
i	isso	ˈisu	that one
	silo	ˈsilu	silo
	quase	ˈkwazi	almost
e	esse	ˈesi	that
	selo	ˈselu	a stamp
	cadê	caˈde	(colloq.) where at?
ɛ	esse	ˈɛsi	ess (the letter 'S')
	selo	ˈsɛlu	I stamp
	Dedé	deˈdɛ	Dedé (TV personality)
a	açо	ˈasu	steel
	chato	ˈʃatu	poophead
	casar	kaˈza	to marry (rural dialect)
ɔ	olho	ˈɔʎu	I look
	jogo	ˈʒɔgu	I play
	avó	aˈvɔ	grandmother
o	olho	ˈoʎu	an eye
	jogo	ˈʒogu	a game
	avô	aˈvo	grandfather
u	uso	ˈuzu	use
	chuto	ˈʃutu	I shoot (a ball)
	caso	ˈkazu	case
ɐ	casa	ˈkazɐ	house

A.2 Measurement of vowels

The acoustic qualities of the Portuguese vowels were extracted from a recording of the words in Table 7, which illustrate the placement of these vowels in different locations. No nasal vowels or diphthongs were used here, for these affect vowel quality. Where possible, minimal pairs were used to isolate vowels in complementary distribution.

`high_front.wav`	[i] high front
`high-mid_front.wav`	[e] high-mid front
`low-mid_front.wav`	[ɛ] low-mid front
`low_central.wav`	[a] low central
`low-mid_back.wav`	[ɔ] low-mid back (rounded)
`high-mid_back.wav`	[o] high-mid back (rounded)
`high_back.wav`	[u] high back
`mid_central.wav`	[ɐ] middle central

A.3 Diphthongs and Nasals

Recordings were made of the diphthongs in Table 1, the triphthongs in Table 2, and the nasals in Table 4. Additionally, a recording was made to contrast the triphthong *qual* from its nasalized counterpart, *quão*.

`falling_diphthongs.wav`	V + [j w]
`rising_diphthongs.wav`	[j w] + V
`triphthongs.wav`	[j w] + V + [j w]
`nasals.wav`	Nasal vowels and diphthongs
`qual_quão.wav`	[kwaw] vs. [kw̃ẽw̃]

A.4 Consonants

The consonants described in Table 5 can all occur as word-initial or word-medial single onsets. A few consonant cluster onsets are possible, and consonants are very restricted in the coda. This is all summarized in the syllable structure shown in Figure 3. Some words illustrating all these consonant positions are listed in Table 8. Where possible, minimal pairs are included for the voiced/voiceless comparison.

`initial_consonants.wav`	Examples of word-initial consonants
`medial_consonants.wav`	Consonants in the onset, word-medially
`coda_consonants.wav`	Consonants in the coda of a syllable

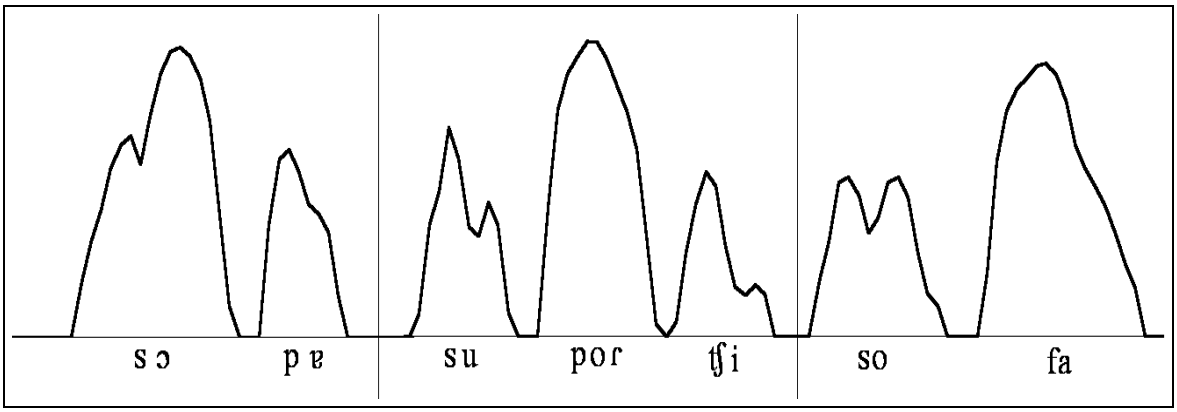
A.5 Stress

A recording of the words in Table 6, demonstrating various stress placements, is saved as `stress.wav`. Figure 8 includes the intensity contours for some selected words.

[5] PET is a loanword from the English acronym for polyethyl terephthalate. This is the kind of plastic used in 2-liter soda bottles. [ˈpɛʧi] is a word that is often heard from people involved in recycling and waste management. (My father works at a plant in São Paulo that recycles plastics.) After exhaustive searching, this was the only minimal pair I found...

Table 8: Words illustrating the consonants in various positions

	Word-initial onset	Word-medial onset	Coda
p	pato [ˈpatu] ‘duck’	campista [kɐ̃ˈpistɐ] ‘camper’	
b	bato [ˈbatu] ‘hit’	cambista [kɐ̃ˈbistɐ] ‘scalper’	
t	tom [tõw̃] ‘tone’	lato [ˈlatu] ‘broad’	
d	dom [dõw̃] ‘donation’	lado [ˈladu] ‘side’	
k	cato [ˈkatu] ‘collect’	vaca [ˈvakɐ] ‘cow’	
g	gato [ˈgatu] ‘cat’	vaga [ˈvagɐ] ‘vacancy’	
ʧ	tia [ˈʧiɐ] ‘aunt’	PET[5] [ˈpɛʧi] ‘PET’	
ʤ	dia [ˈʤiɐ] ‘day’	pede [ˈpɛʤi] ‘ask’	
f	fale [ˈfali] ‘speak’	café [kaˈfɛ] ‘coffee’	
v	vale [ˈvali] ‘worth’	caverna [kaˈvɛrnɐ] ‘cave’	
s	caça [ˈkasɐ] ‘hunt’	preço [ˈpresu] ‘price’	mais tempo [majs ˈtẽpu] ‘more time’
z	casa [ˈkazɐ] ‘house’	prezo [ˈprezu] ‘imprisoned’	mais dentro [majz ˈdẽntru] ‘more inside’
ʃ	chato [ˈʃatu] ‘poophead’	acha [ˈaʃɐ] ‘find’	
ʒ	jato [ˈʒatu] ‘jet’	haja [ˈaʒɐ] ‘has’	
x	rato [ˈxatu] ‘rat’	carro [ˈkaxu] ‘car’	dar [dax] ‘to give’
ɾ		caro [ˈkaɾu] ‘expensive’	dar [daɾ] ‘to give’
			dar [da] ‘to give’
m	mato [ˈmatu] ‘forest’	cama [ˈkɐ̃mɐ] ‘bed’	
n	nato [ˈnatu] ‘innate’	cana [ˈkɐ̃nɐ] ‘cane’	
ɲ		canha [ˈkɐ̃ɲɐ] ‘left-hand’	
l	li [li] ‘read’	galo [ˈgalu] ‘rooster’	
ʎ	lhe [ʎi] ‘him’	galho [ˈgaʎu] ‘branch’	
pɾ	prazo [ˈpɾazu] ‘deadline’	aprazar [apɾaˈzaɾ] ‘to convene’	
bɾ	braço [ˈbɾasu] ‘arm’	abrazar [abɾaˈzaɾ] ‘to scorch’	
tɾ	troca [ˈtɾɔkɐ] ‘exchange’	quatro [ˈkwatɾu] ‘four’	
dɾ	droga [ˈdɾɔgɐ] ‘drug’	quadro [ˈkwadɾu] ‘square’	
kɾ	crave [ˈkɾavi] ‘nails’	lacrimoso [lakɾiˈmozu] ‘tearful’	
gɾ	grave [ˈgɾavi] ‘severe’	lágrima [ˈlagɾimɐ] ‘tear’	
pl	plástico [ˈplasʧiku] ‘plastic’	suplicar [supliˈkaɾ] ‘to supplicate’	
bl	blasfêmia [blasˈfemja] ‘blasphemy’	publicar [publiˈkaɾ] ‘to publish’	
tl		atlas [ˈatlas] ‘atlas’	
dl			
kl	cloro [ˈclɔru] ‘chlorine’	inclemente [ĩkleˈmẽʧi] ‘inclement’	
gl	glória [ˈglɔria] ‘glory’	inglês [ĩˈgles] ‘English’	
fɾ	fralda [ˈfrawdɐ] ‘diaper’	africada [afɾiˈkadɐ] ‘affricated’	
fl	flauta [ˈflawtɐ] ‘flute’	aflito [aˈflitu] ‘afflicted’	



['sɔ.pa] 'soup' suporte [su.'poɾ.ʧi] 'support' sofá [so.'fa] 'sofa'

Figure 8: Relative Intensity of stressed words

Table 9: Alternation of [o] and [ɔ]

Word	Pronounciation	Translation
jogar	['ʒogaɾ]	play.INF
jogo	['ʒogu]	a game
jogo	['ʒɔgu]	play.1SG.PRES
joga	['ʒɔgɐ]	play.3SG.PRES
jogamos	[ʒo'gɐ̃mus]	play.1SG.PRES
joguei	[ʒo'gej]	play.1SG.PAST

A.6 High-mid vs. low-mid vowel perception

The vowels [o] and [ɔ] alternate with noun and verb forms, as in Table 9. Homographs such as *jogo* present a major problem for this text-to-speech implementation. This section describes my attempt to locate the vowel sound that give the most favorable tradeoff between quality and accuracy. Consider my judgements for *Eu jogo um jogo* ('I play a game'):

[ew 'ʒɔgu ũ 'ʒogu]	This is correct.
[ew 'ʒogu ũ 'ʒogu]	This sounds bad, but is still understandable.
[ew 'ʒɔgu ũ 'ʒɔgu]	This is a little worse, but still understandable
[ew 'ʒogu ũ 'ʒɔgu]	This sounds terribly wrong.

These examples demonstrate that choosing a simple rule, for example ⟨o⟩ → [o], will not sound great; nor will it be any better to just use the vowel [ɔ]. But what about some vowel in between those? In this experiment, I insert artificial vowels in the above statement, and then judge the goodness of the speech signal.

Since this text-to-speech implementation will have no prosodic cues, I first record the sentence using a montone voice, with all segments of roughly equal duration and amplitude. This mimics the output of a crude speech synthesizer. See Figure 9.

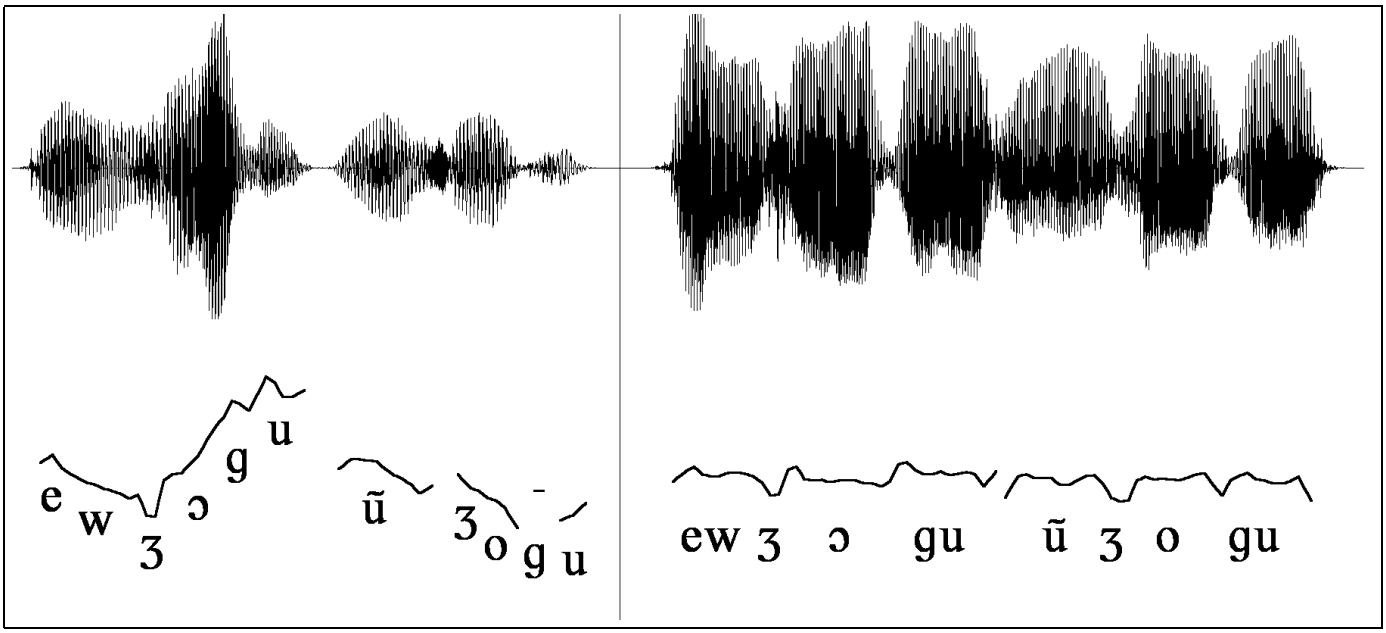


Figure 9: Waveform and pitch of a normal utterance (left) and monotone, machine-like utterance. The sentence is "*Eu jogo um jogo*" ('I play a game').

Spectral analysis of the monotone utterance, and of my vowels in Figure 2, gives these formant frequencies:

	F1 (Hz)	F2 (Hz)
[o]	450	850
[ɔ]	600	950

As a guess, take the linear midpoint of these two vowels:

	F1 (Hz)	F2 (Hz)
[o̞]	525	900

This vowel was synthesized by filtering a multi-frequency tone complex, with filters having a 100 Hz bandwidth, centered at 525 Hz and 900 Hz. The resulting sound was of poor quality and sounded like a [ɔ].