Journal of Portuguese Linguistics

Harmath-de Lemos, S. (2021). Detecting word-level stress in continuous speech: A case study of Brazilian Portuguese. *Journal of Portuguese Linguistics,* 20: 3, pp. 1–43. DOI: https://doi.org/10.5334/jpl.238

RESEARCH PAPER

Detecting word-level stress in continuous speech: A case study of Brazilian Portuguese

Simone Harmath-de Lemos
Department of Linguistics, Cornell University, Ithaca, US
shd57@cornell.edu

This study discusses the detection of primary stress in continuous speech in Brazilian Portuguese (BP) using the West Point corpus (Morgan et al. 2008), and compressed representations of the speech signal (MFCCs, modelled by HMM-GMMs), as implemented in the toolkit Kaldi (Povey et al. 2011). An acoustic model of BP was trained using 5-fold cross validation and tested in three experimental conditions. Fairly high measures of accuracy were achieved in all conditions tested, yielding high *MCCs* and *Kappas*, indicating that the results are neither an effect of imbalanced data sets, nor of chance classification. These results, along with metrics obtained for vowels in pre- and posttonic positions indicate (i) that stress in BP is captured fairly well across speakers and genders by representations of the speech signal that encode spectral features and energy information but which do not directly compute duration or F0; (ii) as captured by the models used herein, there is an asymmetry between pretonic and posttonic vowels; (iii) in a preliminary analysis, *Unstressed* word tokens tend to cluster in prosodically weak positions of the utterance, raising the question of whether stress is consistently realized in these positions; (iv) pending further studies, there is an asymmetry between ultimate, penultimate and antepenultimate words as to how successfully stress is captured by the models used herein.

Keywords: Primary word-level stress; Brazilian Portuguese; Continuous Speech; Automatic Speech Recognition; MFCCs; HMM-GMM

1. Introduction

This paper discusses a method for predicting[1] primary word-level stress placement[2] in continuous speech in Brazilian Portuguese, using Hidden Markov-Gaussian Mixture Models (HMM-GMMs) of the signal as represented by Mel Frequency Cepstral Coefficients (MFCCs), and implemented in the Automatic Speech Recognition (ASR) toolkit Kaldi (Povey et al. 2011). Specifically, I describe the design and the training of an ASR acoustic model of Brazilian Portuguese (BP) which is then tested using three distinct experimental conditions with the purpose of predicting the locus of stressed and unstressed vowels in word tokens drawn from a corpus of continuous speech.

Primary word-level stress—hereafter interchangeably referred to as *primary stress* or simply *stress*—is a structural property of languages, which can be narrowly defined as the relative (acoustic) prominence of parts of a word, typically the syllable or a subpart of it

[1] The term predict is used here in its machine learning sense, where it refers to the assignment of a symbolic label by a machine learning model, given certain features. In the present study, the process of fitting a model has access to the locus of citation stress. When the model is used to predict stress in a word token, it has access only to signal features, and it predicts the locus of stress from those features.

[2] Word-level stress is also called lexical stress in the literature. For the purposes of this study the term lexical stress is avoided, because the study is concerned with the locus of stress in word tokens, rather than with stress in an abstracted lexeme.

(usually, the nucleus or the rhyme). Furthermore, as pointed out in Van der Hulst (2014), even though stress is often used as a cover term for the observable phonetic properties of *accent*—a term which itself refers to a lexical property of words and morphemes that marks the location of such observable correlates—we could expand our understanding of stress to be "… a cover term for *correlates* of accent (rather than just [phonetic] realizations of accent)", in which case, "…we must also include *phonological correlates*…" (Van der Hulst 2014: 5). From that standpoint, primary word-level stress is central to the theoretical understanding of not only phonetic systems, but also of morpho-phonological ones. Within the scope of the present study, the latter considerations are relevant because, beyond building an acoustic model to be used to predict the locus of stress in token words from signal features, a more expansive question pertains to how, and to which extent, computational modeling of the signal side of word stress can inform phonetic and phonological theories of it.

As a linguistic phenomenon, word-level stress is key to a sizable number of language-related processes, such as word segmentation, phonological rules, pitch accent placement, word perception, lexical retrieval, and contrasts between words in the mental lexicon, to name just a few. It follows that, given the overarching significance of stress in language, expanding our grasp of its phonetic and phonological dimensions and mechanisms also in continuous speech is imperative. Yet, undertaking experimental work using continuous speech requires processing ever larger amounts of data, thus rendering the use of computational modeling and machine learning algorithms desirable, and, ultimately, inescapable.

Dealing with relative prominence computationally can however, be a complex endeavor, partly due to the number of distinct acoustic correlates, or combinations thereof, that may express stress phonetically in a given language, and across languages—such as vowel duration, intensity, pitch (F0), and spectral features[3]—and in part because measuring the acoustic correlates of stress is a task that can be accomplished using different criteria (e.g., measuring the duration of the stressed vowel, of the stressed rhyme, or of the entire stressed syllable, measuring peak intensity or relative intensity, spectral balance, tilt or emphasis), which means that distinct ways of taking measurements may potentially provide different insights into the phonetic (and into the phonological) nature of stress. Moreover, it is largely unknown whether different measuring techniques may better represent a given correlate in different languages.[4] Add to these observations intra- and inter-speaker variability in production and the amount of data involved when considering continuous speech, and it follows that, while a vast body of research has looked at word-level stress phonologically, and phonetically in isolated words and in short (mostly carrier-)phrases over the past decades, much less has been done to investigate its nature in continuous speech.[5]

The method described herein aims at furthering the study and understanding of word-level stress in continuous speech while sidestepping some of the complexity introduced by the multi-dimensionality and relational nature of prominence. This is achieved by focusing on attributes which are known to subsume spectral features such as formant locations

[3] See, for example Gordon & Roettger (2017) for a comprehensive list of acoustic correlates that may express stress phonetically across languages.

[4] This is the case, for example, for the various operationalizations of loudness, such as intensity, spectral balance, spectral emphasis, and spectral tilt. There simply aren't, thus far, enough studies that investigate which of these better represents loudness and its role in stress, within and across languages.

[5] Most of the work done with stress in continuous speech comes from other language-related fields, such as Automatic Speech Recognition, Cognitive Science, Neuroscience and Clinical Language Sciences. One notable exception are the works by Barbosa and colleagues for BP (Barbosa 2008; Barbosa, Eriksson, & Åkesson 2013).

and energy information, using mathematical and computational machine-learning models that have relatively high dimensionality.

The work is anchored on the hypothesis that the acoustic information from syllable nuclei alone is sufficient to distinguish between stressed and unstressed syllables in BP, and on the premise that stressed syllables are more carefully—or extremely—articulated than unstressed syllables, which potentially translates in systematic differences related to vowel quality (e.g., more extreme formant frequencies) and/or energy (traditionally captured in phonetics through various operationalizations of loudness, such as intensity, spectral tilt, spectral balance, and spectral emphasis, see Heldner (2001) for a discussion. Thus, the hypothesis is that compressed representations of the speech signal like Mel Frequency Cepstral Coefficients (MFCCs), which encode spectral and energy information, may be used to successfully capture the differences between stressed and unstressed vowels. Stressed vowels are therefore modeled in opposition to unstressed vowels of the same quality, using an HMM/GMM model of MFCCs extracted from the speech signal, as realized in the speech recognition toolkit Kaldi (Povey et al. 2011). Mathematically and computationally, the model parses stress in a token utterance by picking a stress configuration that results in maximal probabilistic weight for the speech token, in the generative probabilistic HMM/GMM model of phonetic realization.

The choice of Brazilian Portuguese[6] is deliberate, because previous literature (e.g., Barbosa, Eriksson & Åkesson 2013; Major 1985; Massini 1991) reports duration to be the most consistent acoustic correlate of primary stress in the language. Since no time-domain representation of the speech signal is used here because the acoustic model relies on compressed representations of spectral and energy features (again, as represented by MFCCs and modeled in HMM/GMMs), robust results would indicate that vowel quality and energy information can be predictors of the acoustic realization of stressed vowels in the language. It is actually not unheard of that languages where duration is shown to be a robust acoustic correlate of stress will also have correlates related to spectral features or energy that also reliably differentiate stressed vowels from unstressed ones: the influential work of Sluijter & van Heuven (1996), for example, showed that spectral balance can reliably distinguish stressed from unstressed vowels in Dutch. For Brazilian Portuguese, Barbosa, Eriksson & Åkesson (2013: 285) showed that posttonic vowels have significant lower values of mean spectral emphasis than pretonic and stressed vowels when the word is in a prominent position of the utterance.[7]

A few of the innovations found in the present work comprise the use of Automatic Speech Recognition (ASR), machine learning classifiers, and a top-down approach to a phonetic study. Less than a handful of phonetic studies made use of these tools to the present date, to the best of the author's knowledge, with only Yuan & Liberman (2009) and Fox (2000) coming to mind, and the present work would be the first one to use classifiers to evaluate a supra-segmental feature of language, at least in the tradition of Linguistics. The study departs from most of the previous literature on ASR-based stress detection in other fields (e.g., Ananthakrishnan & Narayanan 2008; Barros & Weiss 2006; Chen et al.

[6] While the stress systems of Brazilian Portuguese and of European Portuguese (EP) are mostly the same to the best of my knowledge (one exception appears to lie in acronyms, as described in Pereira 2007), it is not to be said that the acoustic model discussed herein can be used with EP speech data, since the present work hinges in part on a multi-pronunciation dictionary that characterizes the possible phonetic realizations of words, and these currently represent data in BP, but not in EP. For the present method to be successfully used with EP speech data, a new phonetic dictionary, representing the phonetic realizations of words in the language, would have to be built.

[7] The authors found that the mean spectral emphasis was higher for the informal interview speaking style than for phrase reading and word list reading speaking styles, and that, moreover, female speakers seem to privilege this parameter over F0 standard deviation.

2004; Deshmukh & Verma 2009; Ferrer et al. 2015; Lai et al. 2006; Li et al. 2013, among others), in a number of ways: where a human usually referees the location of stress for each word token in the data set, I consider the ground truth to be the word's citation stress as understood in phonological theory. Where previous work (e.g., Ananthakrishnan & Narayanan 2008; Barros & Weiss 2006; Chen et al. 2004; Deshmukh & Verma 2009; Ferrer et al. 2015; Lai et al. 2006; Li et al. 2013) resorted to measuring and normalizing one or several of the acoustic correlates of stress and to concatenate the normalized values to the MFCC vectors as additional coefficients, I rely solely on MFCCs of the speech signal, and no additional measurements were taken or used to train the acoustic model. While it is common to use data sets created specifically for each study, a Linguistic Data Consortium (LDC) corpus is used here, both for training and for testing purposes, with all word sizes (in number of syllables) in the corpus being used both during the training and during the testing tasks.

The data set used in the present study is the West Point Brazilian Portuguese Speech corpus (Morgan et al. 2008), a scripted corpus which contains 200 distinct prompt sentences,[8] uttered by 128 speakers of BP, balanced for gender. A vocabulary of phones and a pronunciation dictionary were built specifically to be used in the study.[9] The pronunciation dictionary includes multiple pronunciations for each word entry, as applicable,[10] and mirrors the citation position of stress in each word entry by means of having a vowel labelled with a digit 1.

An acoustic model is trained using the pronunciation dictionary, the vocabulary of phones, and the West Point Corpus. The acoustic model is subsequently experimented with under three distinct conditions, each of which uses a different phonetic dictionary that restricts in a different way the choices the classifier has when predicting whether a vowel is stressed or unstressed for a given word token. To ensure that the model is not over-fitted, a 5-fold cross-validation was performed. Results were computed by looking at the forced alignment files that Kaldi generates based on the models it builds to represent the different phones found in the speech signal and were averaged over the five iterations of each experiment.

For the study, secondary stress (SS), reported to be a part of the phonology of BP (e.g., Major 1985), could be a potential confounding factor, as it could be the case that secondarily stressed vowels are more similar to primarily stressed vowels than they are to their unstressed counterparts (be it represented by MFCCs, duration, pitch, or other acoustic dimensions). Therefore, although the detection of secondary stress itself falls outside of the scope of this paper, SS is taken into account when processing the results.[11]

The remainder of the paper is organized as follows: in section 2, I give a brief review of primary stress assignment in Brazilian Portuguese and describe the structure of the West Point Brazilian Portuguese Speech corpus (Morgan et al. 2008), following that description with an abridged review of the MFCC, HMM-GMM approach to Automatic Speech Recognition and of the Kaldi toolkit (Povey et al. 2011). Section 3 contains a detailed explanation of the methodological aspects of this study, including descriptions

[8] The corpus prompts also include one whole paragraph, not used in the present experiment.

[9] Both the vocabulary of phones and the pronunciation dictionary which were supplied by the LDC along with the West Point corpus are not suitable for the purposes of the work herein, because they did not encompass stress information, and because neither included the diphthongs and the full range of nasal vowels of Brazilian Portuguese.

[10] These multiple pronunciations reflect not only aspects of the phonology of the language, but also regionalisms and phonetic reduction common to continuous speech. The two latter aspects were considered upon audio and visual (spectrogram) inspection of the corpus in its entirety.

[11] Refer to the Background section (2) of this work for a short overview of secondary stress in BP, and to the Methodology section (3) for further discussion on how secondary stress is controlled for in the study.

of the list of phones and the pronunciation dictionary. The three experimental conditions are also detailed in this section, as well as the way in which the data were retrieved from Kaldi's alignment files, then processed and analyzed. Results are shown and discussed in section 4. Section 5 offers concluding remarks and refers to future directions.

2. Background

This section provides background on the three topics pertinent to the present study, namely, stress in Brazilian Portuguese, the West Point corpus (Morgan et al. 2008), and the mechanisms of Automatic Speech Recognition (ASR).

The main aspects of the BP stress system are discussed first, followed by an outline of the structure of the West Point corpus (LDC2008S4–Morgan et al. 2008), and lastly, by an abridged overview of Automatic Speech Recognition systems in general, and of the ASR toolkit Kaldi (Povey et al. 2011) specifically.

2.1. Stress in Brazilian Portuguese

Phonetically, duration has been widely described as the most robust acoustic correlate of primary word-level stress in BP (e.g., Barbosa, Eriksson & Åkesson 2013; Major 1985; Massini 1991), while F0 and intensity are less reliable correlates. In Arantes, Lima & Barbosa (2012: 17) the authors mention that vowels in stressed syllables have higher mean spectral emphasis than those in other positions of the word, which could thus indicate that it is a correlate of primary word stress. In later work, Barbosa, Eriksson & Åkesson (2013) examined duration, F0 standard deviation and spectral emphasis values for three different speaking styles in BP and found out that duration reliably distinguishes stressed vowels from unstressed ones. The authors also found that posttonic vowels have significantly lower mean values of spectral emphasis when compared to the ones in pretonic and stressed syllables (also refer to Endnote 7). In addition, the authors report no difference in the mean duration of pretonic and posttonic vowels in their study. This latest finding could indicate that secondary stress might not be phonetically expressed through differences in vowel duration. With respect to spectral features, when looking at the vowel space of [i e a o u] in polysyllabic words that bear penultimate stress, Arantes (2011) found that such space is maximal for stressed vowels, gradually contracting in pretonic positions.

As mentioned in the introductory section above, predictions about the locus of secondary stress (SS) fall outside of the scope of this inquiry. Nevertheless, within the realm of binary classification, secondary stress can potentially become a confounding factor, for example, if one contemplates the possibility that secondarily-stressed vowels, if or when phonetically realized, may be (acoustically) more similar to primarily-stressed vowels than to unstressed vowels. In this sense, secondary stress becomes relevant herein, warranting a very brief review of the literature.[12] There are competing narratives about both the locus and the phonetic nature of SS in the language: for Major (1985), all pretonic syllables of a word in BP bear secondary stress and all posttonic syllables are unstressed. Collischonn (1994), on the other hand, proposed that secondary stress in BP is binary in nature and that the domain for its assignment is the portion of the word located to the left of the stressed syllable. Barbosa et al. (2004) presented evidence that suggests that secondary stress is equivalent to word-initial prominence only. De Moraes (2003) found out that secondary stress was perceived regularly in words with more than one pretonic syllable and that the locus of perceived SS admitted variation, but the author found no

[12] Note that the review of the literature on secondary stress presented here is far from comprehensive, since it is meant to exemplify that there are competing accounts about both the acoustic nature and the locus of secondary stress in the language.

stable correlate for it (de Moraes 2003: 2066). In Arantes & Barbosa (2008; 2006), the authors propose that secondary stress is best described as phrase-initial prominence and investigate different correlates: duration and pitch accent excursion (2006), and F1 and spectral correlates (2008). Abaurre & Fernandes-Svartman (2008), analyzed text read by 17 speakers and argued for the binary nature of SS in BP, noting that processes of vowel sandhi within the prosodic word, at the lexical boundary, tend to optimize such binary organization. This brief summary illustrates that there is no single answer to the question of where in a word SS falls in Brazilian Portuguese, nor to what its acoustic correlates are in the language. Since, as mentioned above, SS could become a potential confounder for the classifier, the matter needs to be addressed by methodological considerations, which will be discussed in the Methodology section below.

Phonologically,[13] Portuguese is a bounded stress language, where stress is contrastive and predictable in so much as it is (mostly) restricted to the three right-most syllables of the word. Two possible violations to this constraint have been described: Pereira (2007), points out that the three-window constraint is violated when pronouns cliticize to verbs (example (3) below),[14] and Lee (2007), notes that the epenthesis rule of BP, which inserts the front high vowel [i] to break-up disallowed consonant clusters as a means of recovering from phonotactic violations,[15] may also generate words that display pre-antepenultimate stress (example (4) below). The near-minimal triplet of nouns (1) and the minimal triplet of words belonging to distinct lexical categories (2) shown below exemplify contrastiveness in the language.

(1)	*pálido*	[ˈ**pa**. li. dʊ][16]	'pale'$_{N}$
	palito	[pa. ˈ**li**. tʊ]	'tooth pick'$_{N}$
	paletó	[pa. li. ˈ**tɔ**][17]	'suit'$_{N}$

(2)	*sabiá*	[sa. bi. ˈ**a**]	'thrush'$_{N}$
	sabia	[sa. ˈ**bi**. ɐ]	'to know'$_{PST.IPFV.3SG.}$
	sábia	[ˈ**sa**. bjɐ]	'wise'$_{F}$

(3)	*falávamos-te*	[fa. ˈ**la**. vɐ. mos. tɐ]	'to speak'$_{PST.IPFV.1PL.=2SG.Acc.}$
	cantávamo-vo-lo	[kɐ̃. ˈ**ta**. vɐ. mo. vo. lo]	'to sing'$_{PST.IPFV.1PL=2PL.DaT=3SG.Acc.M.}$

[13] I thank an anonymous referee who suggested that the discussion about the phonology of stress in BP, which was a part of the original manuscript, be re-inserted in the final version of this paper. It is a bearing discussion insomuch as previous literature has shown that there is a correlation between phonological weight and duration (i.e., Broselow, Chen & Huffman 1997) and phonological weight and energy (i.e., Gordon 2002; 2006).

[14] Although the author does mention later on in the paper that the violation is apparent, because cliticization is a syntactic operation which happens post-lexical insertion of the word, and stress is marked in the lexicon. It is worth noting that the use of enclitics in Brazilian Portuguese is fairly restricted, mostly limited to a few types of discourse, such as Politics and Law, and to written language, and that proclitics are the more widely used forms in the language.

[15] Lee (2007) mentions the words *técnico*, */ˈtɛkniku/ → [ˈtɛ. ki . ni . kʊ], and *rítmico*, */ˈxitmiko/ → [ˈxi. ʧi. mi. kʊ].

[16] While these transcriptions follow the phonetic dictionary of the *Portal da Língua Portuguesa* (http://www.portaldalinguaportuguesa.org), an anonymous referee pointed out that all posttonic vowels in BP, not only the word-final ones, could be transcribed as [ɪ ɐ ʊ].

[17] The pronunciation [pa. le. ˈtɔ] is as listed in the phonetic dictionary found in the *Portal da Língua Portuguesa*. (http://www.portaldalinguaportuguesa.org) and may be still more common realization of the word *paletó*. The transcription used here is meant to show that the word may be pronounced in this form, which then generates a near-minimal pair with respect to stress placement.

(4)	*logarítmico*	logaRitmiko	→ [lo. ga. ˈ**ɾi**. ʧi. mi. kʊ]	‘logarithmic’$_{\text{ADJ.M.}}$
	étnico	ɛtniko	→ [ˈ**ɛ**. ʧi. ni. kʊ]	‘ethnic’$_{\text{ADJ.M.}}$
	autóctone	aʊtɔktone	→ [aʊ. ˈ**tɔ**. ki. to. nɪ][18]	‘autochthonous’$_{\text{ADJ.}}$

Besides indicating word contrast within and between lexical categories as exemplified above, primary stress also expresses inter-paradigmatic contrast in BP, marking grammatical tense in verbs, as illustrated by the 3rd person plural of the past perfective (penultimate stress) and of the future (ultimate stress), shown in (5).

(5)	*pensaram*	[pẽ. ˈ**sa**. ɾ̃ʊ̃]	‘to think’$_{\text{PST.PFV.3PL.}}$
	pensarão	[pẽ. sa. ˈ**ɾ̃ʊ̃**]	‘to think’$_{\text{FUT.3PL.}}$

Three generalizations about primary stress placement in Portuguese are broadly accepted: (i) antepenultimate is an exceptional pattern, no longer productive in the language; (ii) there is a preference for penultimate stress, unless (iii) the last syllable of the word is heavy—whereby a syllable is heavy if it is closed or if it possesses a bimoraic nucleus, which in the language can be an oral or a nasal diphthong or a nasal vowel (Wetzels 2007).

These generalizations however, do not make the task of describing the mechanism(s) that govern primary stress assignment in the language any more amenable, as made plain by the ample literature debating the topic, which dwells around a number of issues: while stress falls on the penultimate syllable in about 62.5% of 150,000 non-lemmatized words in a (dictionary) corpus study of BP (Araújo et al. 2007: 42), as per generalization (ii), antepenultimate stress words still make up 12.2% of the total number of words in the dictionary and ultimate stress words add to 24.9% (Araújo et al. 2007: 42), so the language is neither straightforward trochaic nor straightforward iambic. With respect to generalization (iii), words such as *sabiá* and *paletó* shown in (1) and in (2) above demonstrate that stress may fall on the ultimate syllable even when it is light. Moreover, there are words in which the last syllable is heavy, but where stress nonetheless falls either on the penultimate or on the antepenultimate syllable, as well as words where the penultimate syllable is heavy where stress falls on the antepenultimate syllable, as illustrated in (6). These examples demonstrate that there are a number of exceptions to generalization (iii).

(6)	*Lúcifer*	[ˈ**lu**. si. fɛɾ]	‘Lucifer’
	pênalti	[ˈ**pe**. naʊ. ʧɪ]	‘penalty’
	nível	[ˈ**ni**. vew]	‘level’

Generalization (i), or the idea that antepenultimate stress is an exceptional pattern, has been more recently brought into question by corpus work such as Araújo et al. (2008), Araújo et al. (2007) and Viaro & Guimarães Filho (2007), all of which provide evidence that shows not only that the number of antepenultimately stressed words is not exactly marginal in the language (about 12%, as previously mentioned), but also that these words have been entering the language as regularly as the other two stress patterns, from the IX to the XX centuries. The same studies also dispute the notion that antepenultimate stress is a by-product of prescriptive grammar, as opposed to an integral part of speakers’ knowledge, pointing out the lack of psycholinguistic evidence to support those claims.

Besides the exceptions found to generalizations (i)–(iii) just discussed, the complexity of the stress system in Portuguese in general, and in BP specifically, displays an additional

[18] The Phonetic Dictionary of the *Portal da Língua Portuguesa* posits [ə] as the quality of the epenthetical vowel for this word, but visual analysis of the word as produced by two speakers of BP showed that the vowel is more similar to [i] than to a schwa.

facet in which, for most accounts, stress assignment in verbs appears to be subject to a set of rules that differs from the one acting upon non-verbs, and which is possibly conditioned by the language's inflectional paradigms, as seen in (5) above and also in (7) below.

(7)			
	pensaram	[pẽ. ˈ**sa**. ɾ̃ɐ̃ũ]	'to think'$_{\text{PST.PFV.3PL.}}$
	pensarão	[pẽ. sa.ˈ**ɾ̃ɐ̃ũ**]	'to think'$_{\text{FUT.3PL.}}$
	pensávamos	[pẽ. ˈ**sa**. vɐ. mʊs]	'to think'$_{\text{PST.IPFV.1PL.}}$

Given all of the facts just listed, it follows then that the task of defining the set of rules that govern primary word-level stress assignment in Portuguese presents a number of challenges. Straight arguments supporting either a syllabic or a moraic trochaic language, even though enticing given the stress system of Latin, are not tenable without added machinery. Sensitivity to quantity (QS), even if positionally restricted, would need to account for the irregular patterns of stress illustrated herein (and for the ones not included in this brief review). Given the nature of the data, while a sizable part of the literature subscribes to the view that stress is mostly predictable in Portuguese, authors like Câmara Júnior (1970) and Morais-Barbosa (1994) proposed that stress is unpredictable in the language, thus specified in the lexicon. With respect to sensitivity to lexical category, authors like Andrade & Laks (1991), Bisol (1992) and Lee (2007), among others, argue that stress is category-blind, while works like Garcia (2017), Hermans & Wetzels (2012), Lee (1997), Pereira (2007), and Wetzels (2007), among others, provide accounts where different rules govern stress in verbs and in non-verbs. Syllable weight is said to play a role for stress placement in non-verbs (Bisol 1992; Garcia 2017; 2019; Hermans & Wetzels 2012; Massini-Cagliari 1995; Wetzels 2007, among others), but it is generally accepted to be non-bearing for the verbal paradigm, where stress is said to be morphologically conditioned.

Some of the additional machinery proposed in previous literature to account for irregular patterns and exceptions herein discussed include: (a) the domain of application of stress rules, which has been proposed to be the word (e.g., Bisol 1992) or the derivational root (e.g., Andrade & Laks 1991), or even a different one for verbs and for non-verbs (e.g., Lee 1997; Pereira 2007); (b) the existence of catalectic consonants (e.g., Bisol 1992) which would explain stress placement in words like *café* [ka.ˈfɛ]; and (c) segmental and syllabic extrametricality, which would explain other irregularities such as antepenultimate stress (the last syllable is extrametrical), irregular patterns of penultimate stress, and stress placement in verbs (e.g., Bisol 1992).

Among recent accounts, Garcia (2017) used the Portuguese Stress Lexicon (PSL, see Garcia 2014)—which was built using the Houaiss dictionary (Houaiss at al. 2001) as its base—to perform a statistical study of stress placement in non-verbs. The author argues that Portuguese has a (QS) stress system, that sensitivity to weight is not restricted positionally in the word and, importantly, that it is gradient, not categorical, as previous analyses proposed. In the study, a negative correlation between weight and antepenultimate stress was found in the lexicon analyzed. Based on these findings, Garcia proposes a grammar where "stress is assigned based on a probabilistic distribution derived from the patterns present in the lexicon" (Garcia, 2017: 75).[19] Most recently, Garcia (2019) used the same PSL to simulate lexica that better approximate the lexicon of an adult native speaker,

[19] Burroni & Harmath-de Lemos (in preparation, presented at LSRL50), propose that stress in Italian and in Portuguese is a morphologically-driven lexical system, subject only to a Basic Accentuation Principle (BAP), in which the right-most accent is the one to surface, thus providing a unified account for primary stress placement in verbs and non-verbs and indicating a lexical statistical grammar learnable from surface forms.

thus creating a lexical baseline to be compared to forced judgements of speakers on a number of conditions related to weight effects in Portuguese. The author finds that both lexical statistics and the grammar have a role in the phonological learning, and that in the case of antepenultimate syllables, lexical statistics and the grammar are at odds, since in the former it was found that there is a negative correlation between heavy syllables in antepenultimate position and antepenultimate stress, but the author's experimental studies showed that speakers did not generalize this negative correlation to the nonce words tested. Naturalness would then be the mediator between lexical statistics and the grammar in the case of antepenultimate stress (Garcia, 2019: 636).

It is worth noting that, in addition to primary and secondary stress, Brazilian Portuguese has been reported to display an asymmetry between pretonic and posttonic syllables (e.g., Major 1985; Câmara Júnior 1970), whereby vowels in posttonic position are always unstressed. This piece of information is important because it plays a key role in determining how the results of the classification experiments are to be computed, a matter discussed in more detail in the Methodology section.

2.2. The West Point Brazilian Portuguese Speech Corpus

The West Point Brazilian Portuguese Speech corpus (LDC2008S04–Morgan et al. 2008), hereafter referred to as WPC, contains digital recordings designed and collected by the Department of Foreign Languages (DFL) at the United States Military Academy at West Point, and at the Center for Technology Enhanced Language Learning (CTELL). Recordings were made at the Brazilian military academy in Brasília in 1999.

The corpus is relatively balanced for gender, with sixty (60) female and sixty-eight (68) male native monolingual and bilingual speakers, who were recorded while reading a script containing two hundred isolated sentences. The WPC contains 131 declarative sentences, 49 interrogatives, and 20 negated sentences. The longest sentence in the prompts has ten words, and the shortest, one. In addition to containing a number of speakers sufficient to perform an ASR study, the corpus was chosen under the working assumption that the number of repetitions of the same word in the exact same prosodic context would be advantageous in training acoustic models of the language's segmental material. Moreover, a corpus of read speech is likely as effective as a corpus of spontaneous speech when looking at the acoustic correlates of stress in BP. For example, the results found in Barbosa, Eriksson & Åkesson (2013: 285) pointed to "…a similar effectiveness of phrase reading and spontaneous styles in uncovering the word stress acoustic correlates in BP, at least for duration and F0 standard deviation."

Out of the 128 speakers in the corpus, data from bilingual speakers—female speakers f40–f45, and male speakers m38–m52—were excluded from the study. Not all 200 sentences in the prompts were uttered by the 99 speakers included in the study, resulting in 7846 utterances, for a total of 39,894 word tokens.

There are 516 distinct word shapes in the WPC's prompts,[20] a figure which is not thought to be a significant limitation to the study given the total number of word tokens in the corpus, adding to roughly 150,000 vowel tokens available to build the acoustic model. A brief summary of the corpus structure is given in **Table 1**.

The distribution of words in the corpus as a function of the number of syllables and of citation stress position is detailed in **Table 2**. Note that the number of syllables reported in this Table refers to a base pronunciation and not necessarily to the surface pronunciation

[20] This number comes after compound words like *quinta-feira* and *guarda-chuva* were included as two separate entries each in the lexicon, *quinta, feira, guarda* and *chuva*.

Table 1: General Information on the West Point Corpus.

Distinct sentences in prompts	200	Utterances Aligned	7846
Shortest sentence (words)	1	Longest Sentence (words)	10
Distinct word shapes in prompts	516	Word tokens in Corpus	39894
Native speakers	99	Declarative Sentences	131
Native female speakers	46	Interrogative Sentences	49
Native male speakers	53	Negations	20

Table 2: Word tokens in the WPC as a function of the number of syllables and stress locus.

Num Syll.	Stress Position				Grand Total
	Monosyll Function	Ultimate	Penultimate	Antepenultimate	
1	11917	3354	–	–	15271
2	NA	4831	10507	–	15338
3	NA	1434	4219	214	5867
4	NA	154	2185	53	2392
5	NA	0	735	98	833
6	NA	0	193	0	193
Grand Total	11917	9773	17839	365	39894

produced by each speaker,[21] which may be subject to epenthesis, syncope, and other phonological processes. Penultimate stress disyllabic words are the most frequent in the corpus, followed by ultimate stress disyllables, and by penultimate stress trisyllables. Note that the dashes (–) mean that the combination is not a logical possibility (as in *antepenultimate monosyllable*, for example), while the zeros (0) mean that there are no occurrences of that particular pattern in the West Point Corpus. In the Table, *NA* means that, although the pattern is a logical possibility, found in the language, and in the West Point corpus, it is not being included in the count for that particular cell in the Table. This is the case, for example, for disyllabic and trisyllabic function words.[22]

2.3. Kaldi and ASR

Kaldi (Povey et al. 2011) is an open source Automatic Speech Recognition (ASR) toolkit, written in C++ and licensed under the Apache License 2.0. To the end user, from an input/output perspective, Kaldi is a collection of commands that can be used to both force-align speech and convert it into text. It is language-independent, and it encompasses normalization and transform algorithms to conduct speaker-independent and speaker-dependent research. The Kaldi toolkit can currently[23] extract standard MFCC (Mel Frequency Cepstral Coefficients) and PLP (Perceptual Linear Predictive) features from the speech signal. It models the acoustic data using either GMM-HMM (Gaussian Mixture

[21] This figure reflects the expected canonical pronunciation in read speech. The number of syllables in certain words may vary according to pronunciation, one example being the word *compreendo*, which may be pronounced [kõ . pre . ẽ . dʊ], with four syllables, or [kõ . prẽ. dʊ], with three syllables.

[22] Since, differently from their monosyllabic counterparts, longer function words are not thought to fall under any special consideration as far as stress is concerned, they are counted together with all other words of the same size (in syllables) in the subsequent columns.

[23] Kaldi is ever changing and evolving, as is its documentation, so the discussion herein reflects not only Povey and colleague's seminal 2011 paper, but also information retrieved from Kaldi's documentation, last accessed in June 2019.

Models-Hidden Markov Models) or SGMM/HMM (Subspace Gaussian Mixture Models-Hidden Markov Models). More recently, three DNN-HMM (Deep Neural Network-Hidden Markov Models) algorithms were added to the toolkit. Kaldi accepts, in principle, any Language Model (LM) that can be compiled into a Weighted Finite State Transducer (WFST).

To convert speech to text, in ASR[24] systems in general and in Kaldi specifically, the speech signal undergoes first feature extraction, a process wherein speech is windowed, typically in 25 millisecond slices, with one such slice being computed every 10 milliseconds. These discrete windows are then converted into high dimension (39 dimensions in Kaldi's MFCCs), fixed size acoustic vectors, a process that can be completed using a few different encoding methods, among them, the two with which Kaldi works, MFCCs and PLPs. Each of the vectors generated in this fashion is known as an observation. Given a sequence *O* of such observations, the job of a speech recognition system is to find the maximally probable sequence of words *W* in language *L* that generated *O*, in other words, the ASR system needs to find the maximum likelihood of *W* given *O*.

$$\hat{W} = argmax_W P(W \mid O) \tag{1}$$

Estimating *P(W|O)* directly—a pattern recognition problem—has not proven successful in the past,[25] so the naïve Bayes' Rule, implemented employing generative models such as HMMs, is used instead.

$$P(W|O) = \frac{P\left(O|W\right)P\left(W\right)}{P\left(O\right)} \tag{2}$$

The probability of an observation, *P*(*O*) is the same for each candidate sentence *W*, so the maximal value can be given by:

$$\hat{W} = argmax_W P(O \mid W)P\left(W\right) \tag{3}$$

The likelihood of observing a given sequence of words *P(W),* where $W \in L$, is a prior calculated based on the Language Model (LM). The probability of a sequence of observations *O* given a sequence of words *W*, or *(P(O|W),* is calculated from the acoustic models (AM) and from the lexicon.

$$P\left(O|W\right) = \sum_Q P\left(O|Q\right)P\left(Q|W\right)P\left(W\right) \tag{4}$$

where $Q = [q_1, q_2, \ldots, q_T]$ is a sequence of acoustic states, one per frame. Resorting to an exceedingly abbreviated description here, *P(O|Q)* is calculated using information from

[24] In addition to the literature cited herein, this sub-section benefited greatly from the many tutorials and lectures on ASR and on Kaldi that I read over time, in special: Daniel Povey's Kaldi Lectures, Andrew Maas' Spoken Language Processing Lectures, and Gilles Boulianne & colleagues' ASR with Kaldi Tutorial.

[25] Newer approaches in pattern recognition, using end-to-end modeling are showing progress in the task of estimating *P (W|O)* directly.

the Mixture of Gaussian probability density functions (PDFs), and $P(Q|W)$ is calculated using information from the Hidden Markov Models (HMMs).

Implementation-wise, in Kaldi (and also in other ASR systems), finding an answer to equation (2) is done by creating a weighted finite-state transducer (WFST) decoding graph $H \circ C \circ L \circ G$, a composition of the HMM (H), the phone Context-dependency Model (C, see following paragraphs), the Language Model or Grammar (G), and the Lexicon (L). The most probable path through the decoding graph will give the most probable sequence of words that generated a given sequence of observations.

Within the architecture just discussed, $P(O|W)$, or the likelihood that a sequence of observations be made given a sequence of words W in language L, is of immediate interest to the present study. In other words, we are interested in the decisions made during the forced alignment process, which is described hereafter in further detail. **Figure 1** summarizes the discussion so far.[26]

One crucial step of ASR is to extract meaningful information from the speech signal (the feature extraction process). During this process, the idea is to extract spectral information that maximizes phone recognition, so characteristics of the source (such as F0 and other details about the glottal pulses), which are not fundamental for phone detection are not directly modelled.

The most popular type of feature used in ASR is the MFCC (Mel Frequency Cepstral Coefficient), among other factors because it approximates human perception of speech better than other feature systems. This is partly because the windowed speech signal is filtered through a Mel-scaled filter bank, which introduces information about the human auditory perception into the model.[27] This process hence steers the model to focus on information that humans would find relevant in the speech signal.

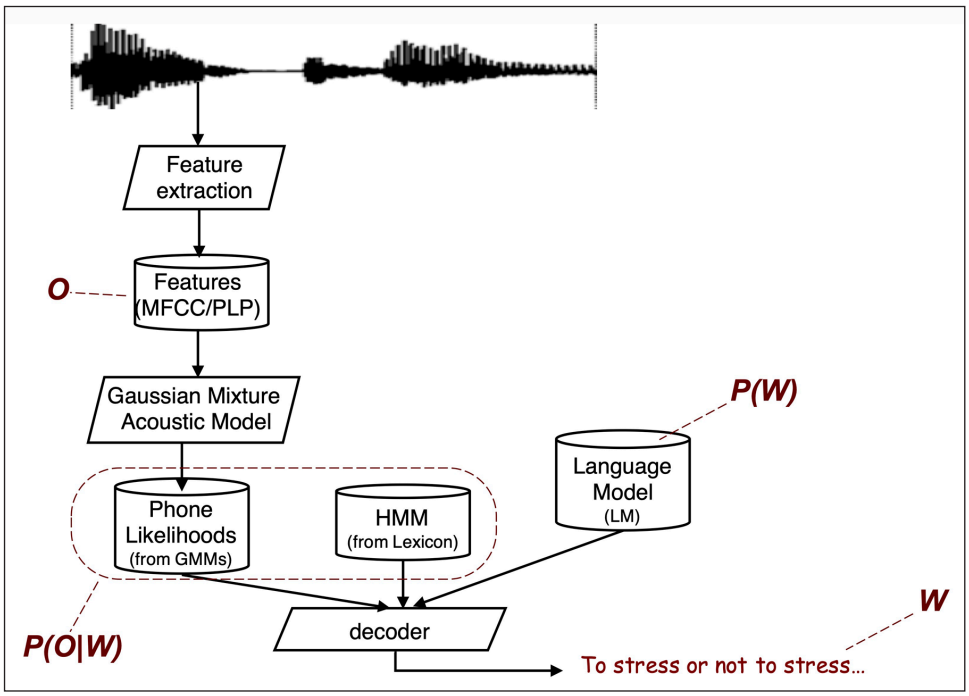

Figure 1: General architecture of an ASR system.

[26] This figure is very similar to the one in Andrew Maas' ASR lecture number 3, Andrew Maas' Spoken Language Processing Lectures.

[27] The idea of using a Mel-scaled filter bank hinges on the fact the humans do not perceive frequency linearly, and the Mel-scale maps the perceptual distance between pitches of different frequencies.

MFCCs are described in some detail here as it is relevant for the present work to understand what information of the speech signal is kept in this compressed representation of it. **Figure 2** shows a summarized block diagram of the extraction process as standardized by ETSI (ETSI 2003). Given this extraction process, MFCCs are subject to significant loss of information present in the original speech signal. The magnitude operation, for example, causes the loss of the phase information, while the Mel-filtering and the vector truncation (post-Discrete Cosine Transform) cause the loss of spectral detail (see, for example, Darch et al. 2005; Darch, Milner & Vaseghi 2006; Darch et al. 2007). In the framework of speech recognition, work such as Zheng & Zhang (2000) showed that speech recognition became much more robust once energy information was added to the MFCC vectors, and thus adding an energy coefficient through the *logE* operation, as seen below, became standard.

Figure 3 below illustrates what the signal would look like if the truncated MFCC vectors were reversed back to a magnitude spectral representation. The blue line is the magnitude spectra derived from the MFCC, while the thinner line represents the original magnitude spectra.

Thus, MFCCs are a smoothed representation of the spectral envelope of the speech signal, and also contain one coefficient that keeps energy information. Still in the framework of speech recognition, it is noteworthy to mention that works by Darch and colleagues (Darch et al. 2005; Darch, Milner & Vaseghi 2006; Darch et al. 2007), have shown that formant frequencies can be accurately estimated (when compared formants calculated using LPC analysis) from MFCCs.

In Kaldi (and in ASR in general), feature vectors have hence 12 MFCC coefficients plus one energy feature. In addition to these 13 dimensions, to add some dynamics to the stationary observations, first order (the *deltas*) and second order (the *deltas of deltas*) derivatives of each MFCC and of the energy coefficient are calculated and added, resulting in a 39-dimensional feature vector.

The acoustic features extracted from the speech signal have to be modeled so that the ASR system can learn which feature vectors correspond to which phones. One of the most widely adopted modeling techniques, used in the present study, is the GMM-HMM acoustic modeling. The basic unit of this acoustic model is a context-independent (CI) phone, or a phone that is modelled independently from its neighboring phones. Alternatively,

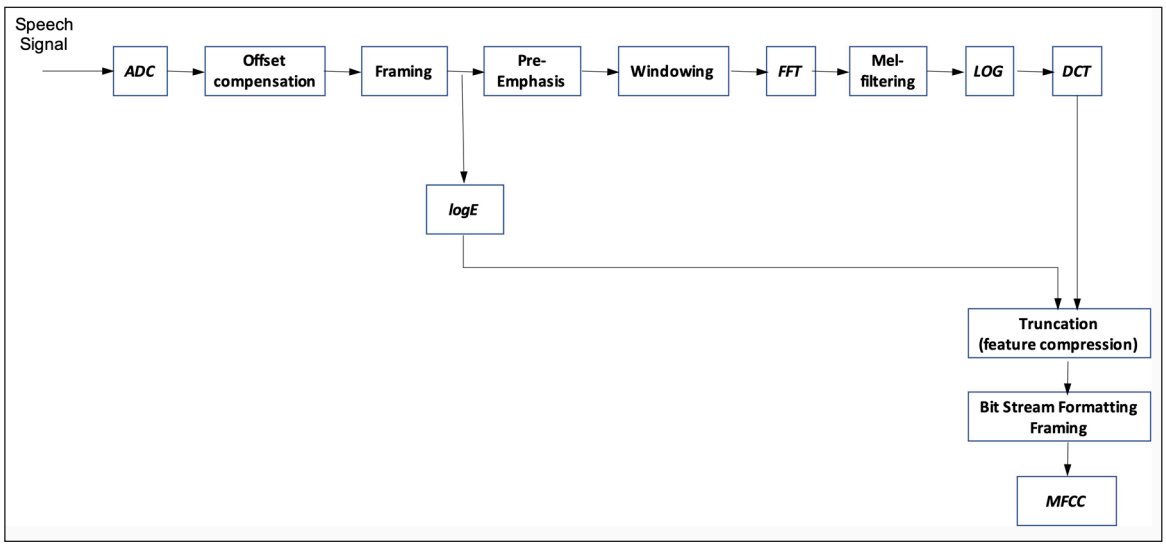

Figure 2: Block diagram of the ETSI Aurora standard for MFCC extraction.

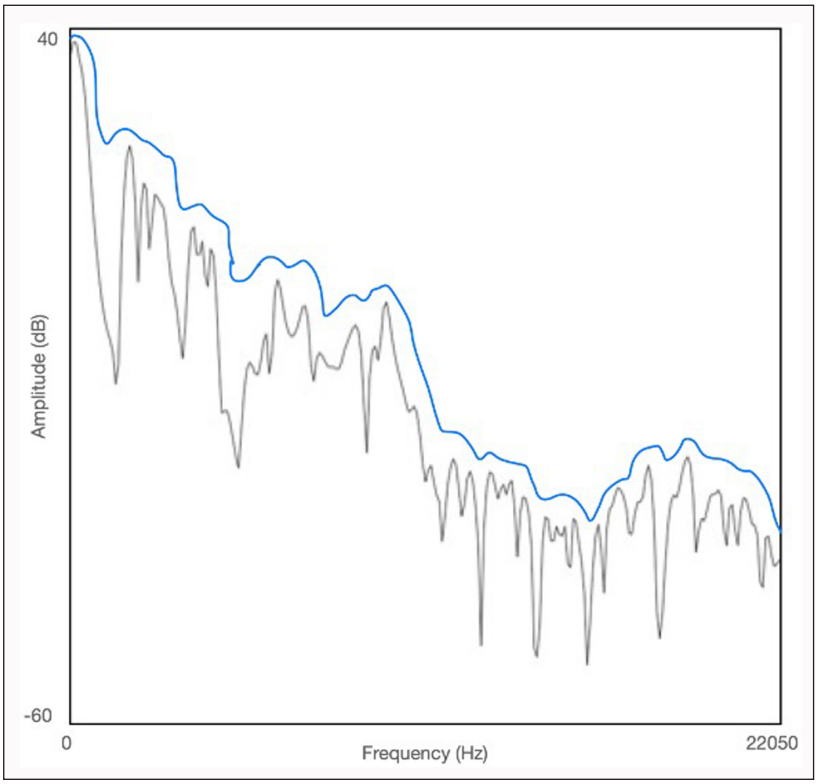

Figure 3: Original (black line) and MFCC-derived (blue line) magnitude spectra.

context-dependent (CD) phones[28] can also be used, but the latter fall outside of the scope of the present study (see comment in the Methodology section below). Each phone is usually modeled as a set of subphones or states, traditionally three of them (plus start and end states), so that in the word 'cat' [k æ t], for example, the phone [æ] would have three states (subphones), [æ]_1, [æ]_2, and [æ]_3. For each sub-phone there are two possible transitions, one to the next state, and one self-loop. Transitions from state *i* to state *j* are assigned probabilities, call them a_{ij}, or the *transition probability*. The word 'cat', would be then represented as in **Figure 4**.[29]

These states are generated using information from the lexicon (phonetic dictionary), and now need to be evaluated, in other words, we need to find out the likelihood of a sequence of observations given a specific HMM. The task is accomplished by associating a likelihood function to each state, modeled using a Mixture of Gaussians that generate probability density functions (PDFs). The continuous density HMM model for 'cat' would look as illustrated in **Figure 5**.

In this subsection, we briefly reviewed how Kaldi (Povey et al. 2011) specifically, and ASR tools in general, find the maximally probable sequence of words *W* in a language *L* to have generated the sequence of (acoustic) observations *O* made from the speech signal. Crucially, we reviewed the information extraction process used to generate MFCCs and work that showed that these vectors encode a smoothed representation of the spectral envelope, from which it is possible to accurately predict formant frequencies (e.g., Darch

[28] In ASR context-dependent (CD) phones were shown to greatly improve speech decoding into text. In these types of phones, a different model is created for a given phone as its neighboring phones change, so the phone [i], for example, would have different models depending on which phones surround it, [b i t] ≠ [m i t] ≠ [m i n]. The most commonly used model is a triphone, which is built taking into consideration the phones immediately to the left and immediately to the right of the phone being modelled, but pentaphone and heptaphone models have also been used.

[29] This is a left-to-right (or Bakis) HMM, which is the structure used by Kaldi. Another possible structure is the Ergodic, a fully connected HMM structure, not shown here.

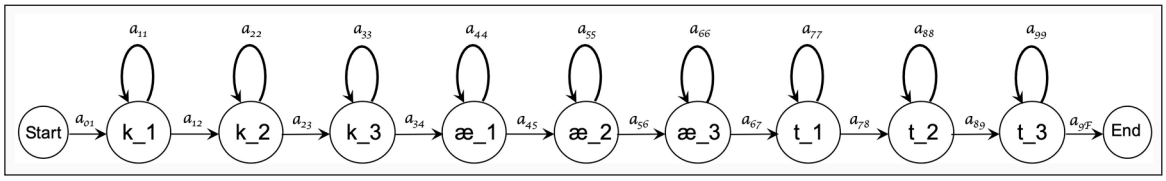

Figure 4: Raw HMM for 'cat'.

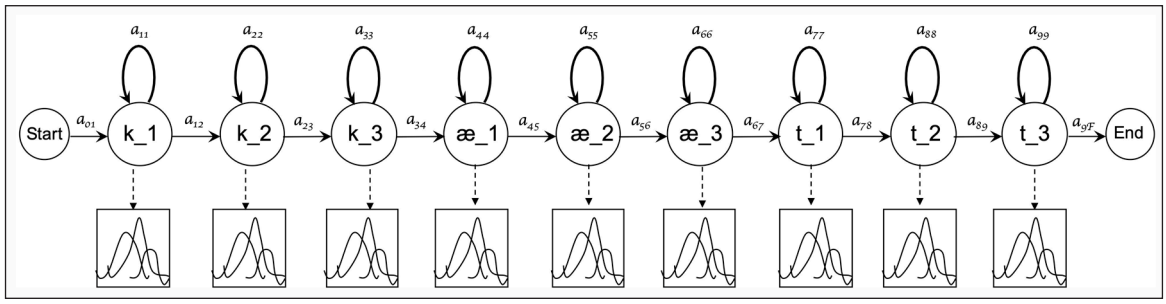

Figure 5: GMM-HMM for 'cat'.

et al. 2005; Darch, Milner & Vaseghi 2006; Darch et al. 2007). Furthermore, MFCCs also encode an energy (*logE*) coefficient (ETSI 2003), which was shown to increase robustness in recognition.

3. Methodology

The first step in the study is to train an acoustic model, which was done here using the ASR toolkit Kaldi (Povey et al. 2011) and speech data from the West Point Corpus (Morgan et al. 2008). To train an acoustic model and to perform the test experiments, Kaldi needs a pronunciation dictionary (herein interchangeably referred to as the lexicon or the phonetic dictionary), a list of phones, a language model (LM) and transcripts of the data. The first two were developed specifically for this study, and were designed having linguistic units in mind as opposed to the units motivated by engineering that are more traditional in the field of ASR. With respect to the LM, because decoding speech into text is not the objective of the study, elaborate language models are not expected to have a bearing on the answers sought here, and therefore a unigram model[30] was used to align the corpus. The transcripts of the data were retrieved straight from LDC distribution for the West Point Corpus (WPC) and were corrected for inaccuracies found out during an auditory inspection of the corpus in its entirety.

To ensure that the acoustic model is not over-fitted, a 5-fold cross-validation process was followed, whereby the corpus was split into five disjoint subsets, balanced for prompt type, and the entire study, including the training pass and the three experiments, was iterated throughout five times. In other words, for each of the five folds, the training pass was done using 4/5 of the corpus, and the three experiments described in this section were performed separately over both the training data set (the 4/5 of the corpus) and the remaining 1/5 of the corpus, the test data set. The corpus-splitting process resulted in five training data sets containing between 6206 to 6376 utterances and in five test data sets, each containing between 1470 to 1640 utterances.

Three experiments were designed to test a classifier's prediction skills about stressed and unstressed vowels when using the acoustic model, the first of which generates baseline performance expectations, as it presents the classifier with the most complex problem to

[30] I thank Mats Rooth for the unigram model used throughout the study.

solve. This experiment is performed using a lexicon where, for each word shape, there are as many entries in the lexicon as there are syllables and combinations thereof in the word shape, the mathematical equivalent to 2^n entries per word shape (where n stands for the number of syllables in a word), each of which is labelled with one of the 2^n possible stress placement configurations. This process is repeated for each of the possible pronunciations given to a specific word shape. This lexicon (phonetic dictionary) configuration is analogous to asking the classifier to predict whether in a given word token, any vowel, no vowel, or any combination thereof, is more similar to the model of a stressed vowel or to the model of an unstressed vowel of its quality, or, which word shape in the 2^n lexicon maximally corresponds to the acoustic observations made from the speech signal for that particular word token.

A second experiment is carried out with a lexicon that contains $n+1$ entries per word shape (again, n being the number of syllables in the word), each of which is labelled for stress in a sole vowel of the word, and an additional entry where no vowels are labelled as stressed. This experimental condition constitutes a way of forcing the classifier to choose exclusively one of the vowels in the word as the stress-bearing vowel, or no vowel at all. The third experiment is conducted using a lexicon that contains n distinct entries for each word shape, each of which is labelled for stress in a sole vowel of the word.

The lexica described above are used in each experiment in such a manner that, to generate the forced alignments, a decision has to be made as to which of the possible renditions of a given word in the lexicon—represented as a sequence of the phones contained in the word—generates the sequence with maximal probability given the observations made.

As mentioned in the Background section, for the purposes herein, the interest lies on the representations generated by context independent (CI) phones. This is because the working premise is that stressed vowels are systematically different from their unstressed counterparts regardless of phonetic context (even though both the former and the latter will vary as a function of the neighboring sound segments). Furthermore, from a modelling standpoint, the lexicon is too small for it to be sensible to work with context-dependent (CD) phones, such as triphones. As a consequence, only monophone alignments (CI phones) are computed throughout the study.

During training, there needs to be a bootstrapping pass to generate the monophone model. This was done using 4000 utterances of the training data sets, and the model was trained using Kaldi's *train_mono.sh* script (4 jobs, one machine). A monophone model for each of the 5 folds is trained using the list of phones and the lexicon labelled for citation stress described in the following subsections (3.1.1, 3.1.2). For each training and test data set in each of the 5 folds, in each of the three experiments, MFCCs are computed using the proper rendition of the lexicon. Following the training pass, also for each of the five folds and three experiments, both the training data set and the test data set are force-aligned using the relevant monophone model, rendition of the lexicon and the corresponding MFCCs. The forced alignments were generated using Kaldi's *align_si.sh* script (8 jobs, one machine).

The results were retrieved from the forced alignment files that Kaldi generates (the *ali** files), using Perl and Python scripts written specifically for the task, and were then averaged over the five folds in each experiment and in each data set. A second set of scripts summarized the data according to the description laid out in the Data Analysis subsection below. A file containing the raw summarized data was then imported into Microsoft Excel 16.43, where the performance metrics were calculated. Monosyllabic function words were tallied separately because no assumptions can be made about them (see section 3.2 and 3.3) with respect to stress.

The remainder of this section describes in further detail the list of phones and the phonetic dictionary constructed for this study. The section ends with a description of how the results were analyzed, which offers important information about the manner in which stressed vowels are compared to unstressed ones, both syntagmatically and paradigmatically.

3.1. The List of Phones

For the purposes of the present study, a list of phones should capture two crucial aspects: the first is related to the nature of the study itself, since, differently from most work done using speech recognition, the objective herein is to represent linguistic units, as opposed to simple acoustic units. So, for example, whereas diphthongs (e.g., [aʊ]) are encoded as two acoustic segments ([a] + [ʊ]) in many ASR studies, they are encoded as one linguistic unit herein (e.g., [aʊ]), under the assumption that neither the unstressed diphthong nor its stressed counterpart are just the sequential combination of the exact monophthongs [a] and [ʊ], in other words, [ˈaʊ] ≠ [ˈa] + [ˈʊ].

The second aspect which the list of phones should capture is related to the number of phones needed in order to represent a reasonable amount of phonetic variation that exists in any language, specifically with respect to vowels. To encode said variation, three sources of information were considered: the phones described in Barbosa and Albano (2004), the phones used in the phonetic dictionary of the 'Dicionário Fonético da Língua Portuguesa' (Correia et al. 2019), and the present author's native intuition.

Implementation-wise, each phone within the list of phones was encoded to mirror the Advanced Research Project Agency's (ARPA) ARPAbet[31] as used in the *Carnigie Mellon University Phoneme Set* whenever possible, and new encodings following the paradigms of the CMU phoneme list were added as needed. Vowels marked with the digit 1 are understood to be vowels that bear stress and vowels with no additional markings are understood to be unstressed.

In the study, we want to keep the models of the stressed and unstressed counterparts of a given vowel quality independent from one another. To accomplish that, the former and the latter do not share the same phonetic symbol, and are each modeled with distinct Gaussian Probability Density Functions. In implementational terms this means that each phone in the phone list is added to a different line of the file *nonsilence_phones.txt*.

In view of the details and constraints explained so far, 25 consonant and 74 vowel phones are represented in the list, totaling 99 phones, a number that includes one stressed and one unstressed counterpart for each vowel quality represented. This is because, even though the distribution of vowels in the language is such that some vowel qualities never bear stress and others, conversely, overwhelmingly bear stress, in the different experimental conditions it is assumed that any vowel quality may be stressed or unstressed, thus these need to be represented in the list of phones.[32] **Table 3** summarizes the distribution of phones in the list per segment type.

[31] Although the Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA) is more representative of the IPA, and could have been used as an encoding method instead, some of the XSAMPA symbols are less tractable for computation purposes (e.g., stress placement is marked by double quotes (") in XSAMPA).

[32] A few examples of this asymmetry are the word-final reduced vowels [ɪ] and [ʊ], which appear only in posttonic position, and the vowels [ɛ] and [ɔ], which occur almost exclusively in stressed position in the language, except for polymorphemic words (e.g., *pezinho* [pɛ .ˈzi . ɲʊ] 'feet'$_{\text{DIM}}$., and *somente* [sɔ .ˈmẽ . ʧɪ] 'only'). Analogously, the oral diphthongs [ɛɪ], [ɔʊ], [ɛʊ], and the nasal diphthongs [ɐ̃ʊ̃], [ɐ̃ɪ̃] are rarely found in unstressed positions of the (monomorphemic) word.

Table 3: Number of phones of each type in the List of Phones.

Consonant	Vowel			
	Monophthong		Diphthong	
	Oral	Nasal	Oral	Nasal
25	22	10	34	8

3.2. The Lexicon

The base phonetic transcription for each word in the lexicon reflects the *Rio de Janeiro* standard pronunciation of the word as given in the *Dicionário Fonético da Língua Portuguesa* of the *Portal da Língua Portuguesa* (Correia et al. 2019). The *São Paulo* standard pronunciation was used when found to be more representative of a given word token in the prompts, after an audio inspection of the corpus.[33] In addition, recall from the introductory section of this study that the phonetic dictionary incorporates explicit pronunciation modeling for each word entry. Because the WPC does not provide data that identifies the geographic provenance of each speaker, the explicit pronunciation modeling used in the dictionary intended to capture three main sources of variation: (i) common, generally widespread phonetic variation processes found in BP, which are not necessarily associated to a specific regional variety; (ii) some regional variation in vowels, found out to be present in the West Point corpus during the audio inspection, and (iii) reduction processes found upon visual inspection of the corpus utterances in *Praat* (Boersma & Weenink 2019). **Table 4** exemplifies some of these variations.[34] Note that while **Table 4** exemplifies some of the processes just mentioned, it is not the comprehensive list of pronunciations encoded in the lexicon for these words.[35]

Following the base (citation) phonetic pronunciation encoded in the lexicon, as described in the beginning of this section, and taking into account the number of word tokens in the corpus, the distribution of phones is approximately as given in **Table 5**. Note that there is a rough balance between the rate of consonant tokens (49.5%) and that of vowel tokens (50.5%). It is important to notice however, that speakers do not necessarily utter the base phonetic pronunciation of any given word token, which can considerably alter the actual number of consonant and vowel tokens in the corpus.

All of the 905 distinct words that formed the lexicon originally embedded in the West Point corpus (of which only 516 are present in the 200 corpus' prompts used in this study) were modeled for explicit pronunciation, generating a lexicon that contains 2,057 word entries labelled for stress in citation position.

There are four renditions of this explicit pronunciation-modeled lexicon: one which is used during the training pass, and three others, used during the experiments described in the initial paragraphs of this section. The rendition of the lexicon used during the training pass contains only words labelled for stress in citation position, which in practice means that the vowels found in stress citation position of a word are marked with a digit 1. With respect to monosyllabic function words—and their corresponding phonetic variants—these were the only words to be encoded two-way in all renditions of the lexicon: one where the (sole) vowel was labelled as stressed (digit 1) and the other where the vowel

[33] This is the case, for example, for the word *água* ('water'), which was more consistently refereed to be a disyllabic word by a linguist native speaker, a pronunciation consistent with the *São Paulo* standard variety [ˈa. gwɐ], as opposed to the *Rio de Janeiro* standard, [ˈa. gu.ɐ].

[34] Note the /S/ represented in words *três*, *paz* and *voz*, as these can surface as [s z], depending on the word that follows and on the syntactic structure of the sentence.

[35] Additional pronunciations for the word *setecentos* (seven hundred), for example, include the affricate allophone [tʃ].

Table 4: Examples of explicit pronunciation modeling included in the Lexicon.

dia	'day'	/dia/ → ['ʤi. ɐ]
dente	'tooth'	/deNte/ → ['dẽ. ʧɪ]
zoológico	'zoo'	[zo. o.'lɔ. ʒi. kʊ] OR [zo.'lɔ. ʒi. kʊ]
gratuita	'free'	[gra.'tuɪ. tɐ] OR [gra. tu.'i. tɐ]
paz	'peace'	/paS/ → [paɪs]
três	'three'	/treS/ → [treɪs]
voz	'voice'	/vɔS/ → [vɔɪs]
janeiro	'January'	/ʒaneɪro/ → [ʒa. 'ne. ɾʊ]
eixo	'axis'	/eɪʃo/ → ['e. ʃʊ]
televisão	'TV'	[te. le. vi. 'zẽʊ̃] OR [tɛ. lɛ. vi. 'zẽʊ̃]
ocupado	'busy'	[o. ku. 'pa. dʊ] OR [ɔ. ku. 'pa. dʊ]
crédito	'credit'	['krɛ. ʤi. tʊ] OR ['krɛʤtʊ]
setecentos	'seven hundred'	[sɛ. te. 'sẽ. tʊs] OR [sɛt'sẽtʊs]

Table 5: Phone distribution in the West Point Corpus.

Consonants	Vowels	Total Phones
76,790	78,466	155,256

had no label (unstressed). This strategy was adopted to reflect the fact while monosyllabic function words are generally destressed, a sizable rate of them in the corpus were uttered bearing focus or are stressed for other reasons.

In the remaining three renditions of the lexicon all word entries were labelled for stress placement according to each experimental condition: for the baseline experiment, each entry in the lexicon has 2^n renditions (n being the number of syllables), and analogously for experiments $n+1$ and n. **Table 6** shows snippets of the lexica just described. In the testing experiments, for each word token, the model chooses among a set of competing pronunciations, such as the pronunciations [a b r ahx1] and [a1 b r ahx] for the dictionary entry *abra1N*. Here the word form is used to determine what pronunciations are in competition in a given experimental condition.

3.3. Data Analysis

The analyses of the predictions made by the classifier in the three experiments (explained at the end of the present section) will offer slightly different insights about stressed and unstressed vowels, from both a syntagmatic and a paradigmatic perspective. Syntagmatically-speaking (are stressed vowels distinct from the surrounding vowels?), experiment n illustrates which position of the word token is occupied by the vowel that best fits the model of a stressed vowel of its quality. Experiment $n+1$ will offer an insight on whether there is one vowel in the word token, or whether there are no vowels at all, thought to best fit the model of a stressed vowel of its quality, in a way akin to (a phonetic version of) Obligatoriness (as in Hyman 2006: 231). An analysis of the results from experiment 2^n, where the choices are unconstrained,[36] will highlight whether is there more than one vowel in the word token that fits more closely to the model of a stressed

[36] Although this experiment could potentially offer an insight into a phonetic version of Culminativity, as understood in Hyman (2006: 231), there isn't sufficient data in the literature to ensure that pretonic vowels, if they bear secondary stress, are phonetically different from their stressed counterparts, nor how they differ, with respect to acoustic correlates.

Table 6: Stress Placement in the different renditions of Lexicon.

	abra1T	a1 b r ahx		abra1N	a b r ahx1
Training	chame1T	sh ah1 m e		abra1N	a1 b r ahx
	chame1T	sh ah1 m ihx		chame1N	sh ah m e1
	vários1T	v a1 r iw s		chame1N	sh ah m ihx1
			Experiment *n*	chame1N	sh ah1 me
	abra2N	a b r ahx		chame1N	sh ah1 m ihx
	abra2N	a b r ahx1		vários1N	v a r iw1 s
	abra2N	a1 b r ahx		vários1N	v a1 r iw s
	abra2N	a1 b r ahx1			
	chame2N	sh ah m e		abra1N1	a b r ahx
	chame2N	sh ah m e1		abra1N1	a b r ahx1
Experiment 2^n	chame2N	sh ah1 me		abra1N1	a1 b r ahx
	chame2N	sh ah1 me1		chame1N1	sh ah m e
	chame2N	sh ah m ihx		chame1N1	sh ah m e1
	chame2N	sh ah m ihx1	Experiment *n+1*	chame1N1	sh ah m ihx
	chame2N	sh ah1 m ihx		chame1N1	sh ah m ihx1
	chame2N	sh ah1 m ihx1		chame1N1	sh ah1 m e
	vários2N	v a r iw s		chame1N1	sh ah1 m ihx
	vários2N	v a r iw1 s		vários1N1	v a r iw s
	vários2N	v a1 r iw s		vários1N1	v a r iw1 s
	vários2N	v a1 r iw1 s		vários1N1	v a1 r iw s

vowel of its quality in that particular word token. Paradigmatically (how do stressed vowels compare to unstressed vowels?), the analysis of the results of each experiment offers insights on how well can unstressed vowels be told apart from stressed vowels when there are stringent restrictions—as in experiment condition *n*, where one vowel in the word token has to be stressed and all others have to be unstressed—and whether these differences hold as the restrictions are relaxed—as in experiment condition 2^n, where any vowel, any combination of vowels, or no vowel in the word token may be stressed (or unstressed).

In addition to the paradigmatic and syntagmatic observations we seek to make, two other relevant pieces of information should be taken into account when outlining the analysis methodology for this study: from the discussion presented in section 2 (Background), recall that secondary stress can potentially be a confounding factor, because it is unclear whether secondarily stressed vowels are acoustically more similar to a primarily stressed vowel or to an unstressed vowel in BP. Moreover, the position occupied by secondary stress in a word is also unclear in the language: recall that there are competing accounts about the locus of SS in a word. It is therefore a possibility that any vowel located to the left of the primarily stressed vowel could potentially bear secondary stress. As a result, encoding information about secondary stress in the phonetic dictionary to eliminate potentially confounding factors is not an option and the alternative is to take secondary stress into consideration when processing the results from all experiments. The second crucial piece of information (also discussed in the Background section) is that posttonic vowels in BP are generally described to be unstressed. These factors, added to the observation made about monosyllabic function words in section 3.2 (The Lexicon) are operationalized through the following considerations:

(i) No ground truth can be assumed about the vowels to the left of the (citation position) stressed vowel.
(ii) Posttonic vowels should be unstressed.
(iii) Primary stress is located in the citation position of the word.
(iv) No assumptions are made about monosyllabic function words.

In practical terms, item (i) above means that results for the vowels located to the left of the stressed vowel are computed separately. Item (ii) means that predictions of stressed vowels in posttonic position should be penalized and (iii) means that predictions of stress in citation position are rewarded. Consideration (iv) means that the results for monosyllabic function words are always counted separately.

Two sets of metrics summarize the complementary perspectives we seek (syntagmatic and paradigmatic). To compute the predictions made by the classifier from the perspective of word tokens as a unit, the data are summarized in the following manner: *Matches* (*M*) are assigned to aligned word tokens where the vowel predicted to be stressed matches citation position for that word in the lexicon (the ground truth) and all other vowels in the token were predicted to be unstressed. *Partial Matches* (*PM*) mean slightly different things for the different experiments: in experiment n and in experiment $n+1$, a *Partial Match* (*PM*) is a prediction of stress on a vowel located to the left of the citation stress vowel,[37] and posttonic vowels are predicted to be unstressed, while for experiment 2^n a *Partial Match (PM)* means that the classifier predicts stress on the vowel located in citation position and on any vowel located to the left of citation position, while all posttonic vowels are predicted to be unstressed. *Mismatches* (*MM*) are the instances of aligned word tokens in which the classifier predicts a posttonic vowel to be stressed, regardless of the predictions made about the other vowels in the word token (even if those are paired with a prediction matching citation position). Lastly, the metric *Unstressed* (*U*) encompasses the counts of word tokens where the classifier predicted that none of the vowels are stressed.

With the counts just described in mind, two accuracy rates and two error rates are calculated: an overall *Composite Accuracy Rate* (*CAR*) is calculated by summing the *Partial Matches* (*PM*) and the Matches (*M*), while an overall *Accuracy Rate* (*AR*) is computed using the count of *Matches* (*M*) only. Notice that these accuracy rates are not the same as the *Accuracy* metric calculated using the confusion matrices described below. The overall *Error Rate* (*ER*) is computed using the count of *Mismatches* (*MM*) only, and the *Composite Error Rate* (*CER*) is calculated through the sum of *Unstressed* (*U*) and *Mismatch* (*MM*) tokens. These metrics are shown below in formulae (5)–(8). Note that Tot_{WT} corresponds to the total number of word tokens being evaluated.

$$AR = \frac{M}{Tot_{WT}}100 \tag{5}$$

$$ER = \frac{MM}{Tot_{WT}}100 \tag{6}$$

$$CAR = \frac{\left(PM + M\right)}{Tot_{WT}}100 \tag{7}$$

[37] This is done because nothing can be assumed about these vowels, and choosing more than one vowel as stressed in these experiments is not a possibility.

$$CER = \frac{\left(MM + U\right)}{Tot_{WT}} 100 \tag{8}$$

The second set of metrics calculated here reflects the discriminatory skills of the predictions made in the experiments (offering a paradigmatic perspective of stressed and unstressed vowels), and it is calculated from the tabulation of confusion matrices, where the data are summarized in the following fashion: for each word token aligned, a prediction of stressed vowel in stress citation position is counted as a *True Positive* (*TP*) and a prediction of unstressed vowel in stress citation position is counted as a *False Negative* (*FN*). Conversely, a prediction of unstressed vowel in pretonic or posttonic position is counted as a *True Negative* (*TN*), and finally, predictions of stressed vowels in pre- or posttonic positions are counted as *False Positives* (*FP*).

To provide a clearer account of the results, taking into consideration differences in class sizes and other imbalances that exist in natural language in general and in the WPC data set specifically, the counts in the confusion matrices are used to calculate the following metrics: *Accuracy, Precision, Sensitivity, Specificity, F1-score*, the *Matthew's Correlation Coefficient* (*MCC*), the *Cohen's Kappa Coefficient* (*Kappa*), the *False Positive Rate* (*FPR*), the *False Discovery Rate* (*FDR*), and the *False Negative Rate* (*FNR*), using the following formulae:

$$Accuracy = \frac{\left(TP + TN\right)}{\left(TP + FP + TN + FN\right)} \tag{9}$$

$$Sensitivity = \frac{TP}{\left(TP + FN\right)} \tag{10}$$

$$Specificity = \frac{TN}{\left(TN + FP\right)} \tag{11}$$

$$Precision = \frac{TP}{\left(TP + FP\right)} \tag{12}$$

$$F1score = \frac{2 * TP}{\left(2 * TP + FP + FN\right)} \tag{13}$$

$$MCC = \frac{\left(TP * TN\right) - \left(FP * FN\right)}{\sqrt{\left(TP + FP\right) * \left(TP + FN\right) * \left(TN + FP\right) * \left(TN + FN\right)}} \tag{14}$$

$$Kappa = \frac{\left(Total\,Accuracy - Random\,Accuracy\right)}{\left(1 - Random\,Accuracy\right)} \tag{15}$$

$$False\ Positive\ Rate\left(FPR\right) = \frac{FP}{\left(FP + TN\right)} \tag{16}$$

$$False\ Discovery\ Rate\left(FDR\right) = \frac{FP}{\left(FP + TP\right)} \tag{17}$$

$$False\ Negative\ Rate\left(FNR\right) = \frac{FN}{\left(FN + TP\right)} \tag{18}$$

Accuracy gives the rate of correct predictions made about both classes of vowels, stressed and unstressed. *Sensitivity* is the rate of vowels correctly predicted to be stressed among all of the vowels expected to be stressed (the ground truth of stress locus for each word), while *Precision* is a measure of the performance of the classifier that shows the rate of vowels that are truthfully stressed among all vowels which were predicted to be stressed. *Specificity* is the rate of vowels correctly identified as unstressed among all vowels expected to be unstressed. *Sensitivity* and *Specificity* are, in a way, a metric of correlation between the predictions of stressed and unstressed vowels and the ground truth of each. The *F1-Score* provides an alternative measure of *Accuracy*, showing the balance between *Precision* and *Sensitivity*. The *False Positive* and the *False Negative* rates present a complementary perspective to *Specificity* and *Sensitivity*, and the *False Discovery Rate* is complementary to *Precision*.

The metrics just described provide complementary insights on the results obtained for stressed and unstressed vowels but are not adjusted for imbalances that may exist in the size of the classes represented in the confusion matrices (*TP, TN, FP, FN*), and do not take in consideration the probability of chance agreement between the predictions made by the classifier and the ground truth. The *Matthews' Correlation Coefficients* (*MCCs*) and the *Cohen's Kappa Coefficients* (*Kappa*) fill in these gaps.

The *MCCs* are a measure of the strength of the correlation between two raters, and they take into consideration the proportion of each class in the confusion matrix (*TP, TN, FP, FN*). *MCCs* range between $-1 \leqslant MCC \leqslant 1$, where –1 means complete disagreement, 0 means chance agreement and 1 means perfect agreement between the predictions and the ground truth.[38]

The *Kappa* shows the chance-corrected standardized measure of agreement between the predictions and the ground truth. A *Kappa* coefficient ranges from $-1 \leqslant \kappa \leqslant 1$, where –1 would mean perfect disagreement between predictions and the ground truth, 0 would mean that the amount of agreement found can be expected from random chance, and 1 represents perfect agreement.[39] For each *Kappa Coefficient* calculated, the confidence interval is also given. A *Kappa* value between 0.61 and 0.8 denotes substantial agreement between predictions and the ground truth, while any value higher than that indicate almost perfect agreement according to the (more conservative) interpretation of given by Landis & Koch (1977). In the interpretation given by Cicchetti & Sparrow (1981), a value of 0.61 or above is interpreted to be excellent agreement.

[38] Note that for imbalanced data sets it is almost impossible for the *MCC* to be really close to 1.

[39] Although values of *Kappa* below zero are possible, they are unlikely in practice. A short article about the *Cohen's Kappa Coefficient* can be found on the NIH page at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/. One criticism of the *Kappa* is that it may lower the rate of agreement excessively.

4. Results and discussion

Out of the 7,846 utterances in the corpus, two utterances, prompt 162, uttered by male speaker 01 (m01), and prompt 041, uttered by male speaker 11 (m11), had empty audio files, yielding 7,844 aligned utterances and totaling 39,888 word tokens (from the initial 39,894 in the corpus) for each experimental condition. Of these word tokens, 11,915 (originally 11,917, minus 2 from prompt 162 above) are monosyllabic function word tokens and are not computed together with the other tokens (see discussion in section 3).

Recall from the discussion in section 3 of this paper that the training data sets contain 4/5 of the data in the WPC and the test data sets contain the remaining 1/5 of the data. The results presented below represent the average of the five folds over the training and the test sets for each experiment. Since the data shown in **Table 7** and in **Figures 6** and **7** show that the results are slightly less optimistic for the test data sets, the subsequent part of the analyses is performed over the five folds of the test data sets (avoiding overly confident inferences).

The results as related to binary classification and to the design of the three experiments are discussed in detail below, then subsection 4.3 examines what these results mean for the syntagmatic and paradigmatic perspectives of stressed and unstressed vowels in Brazilian Portuguese sought herein.

4.1. Results for Word Tokens

Table 7 summarizes *Accuracy* and *Error* rates as a function of how the aligned word tokens compare to the expected word shape given the ground truth described in section 3 (Methodology), across the three experiments. From a classification perspective, it is expected that accuracy rates will decrease as the complexity of the experiment increases (where complexity is defined as the number of options the classifier can choose from in a given experiment), and the error rates will increase in the same direction, expectations which are both borne out from the data shown in the Table.

Analogously, the expectation that there would be only slight differences in performance[40] between the results for the training and the test data sets is confirmed by the data in **Table 7**. These results confirm that the model did not simply memorize the data used during the training pass, but rather learned to generalize over the data found in the training set, and it can thus be successfully used to evaluate new data. For each of the

Table 7: *Accuracy* and *Error* rates for the Training and Test data sets.

Experiment	Data Set	*AR* (%)	*ER* (%)	*CAR* (%)	*CER* (%)
Baseline (2^n)	Training	71.41	4.81	82.63	17.37
	Test	69.70	5.19	81.56	18.44
n+1	Training	83.73	1.30	86.17	13.83
	Test	82.37	1.48	85.28	14.72
n	Training	93.55	2.59	97.41	2.59
	Test	92.73	2.83	97.17	2.83

[40] Had the results from the training data sets been very different from the results from the test data sets, we would be faced with the issue that the model memorized the data instead of generalizing over (learning from) it, and would thus not be effective in evaluating new data.

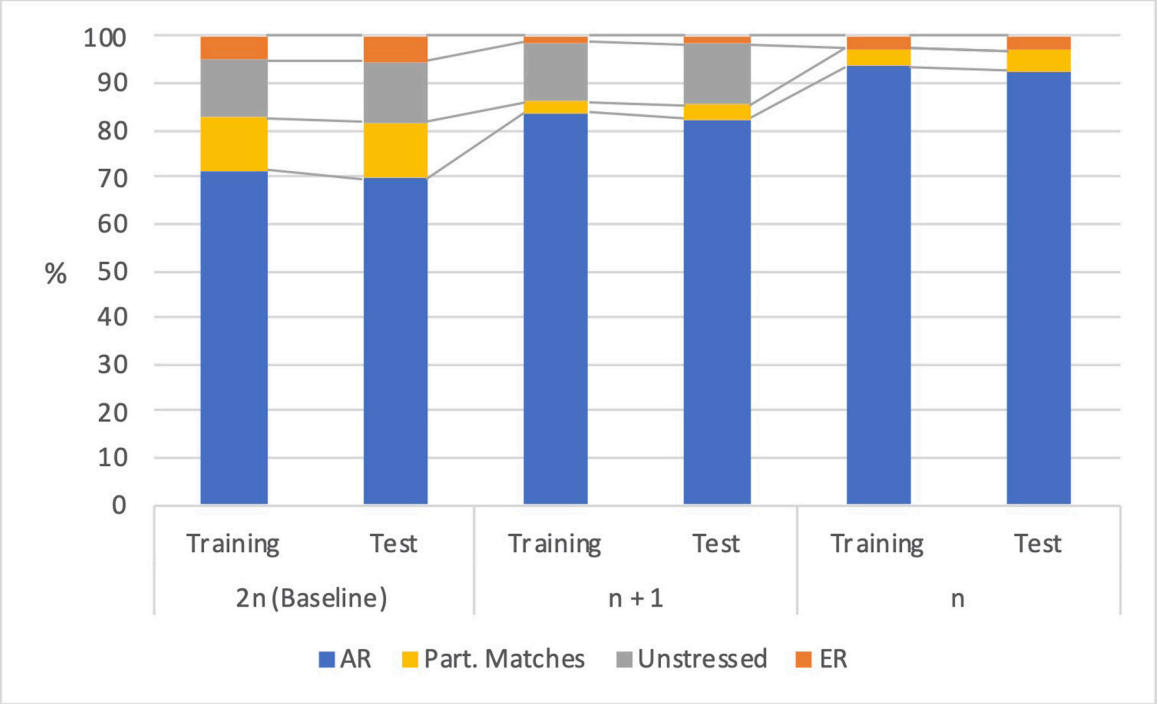

Figure 6: Overall Accuracy, Error, Partial Matches and Unstressed rates (summarized in Table 7).

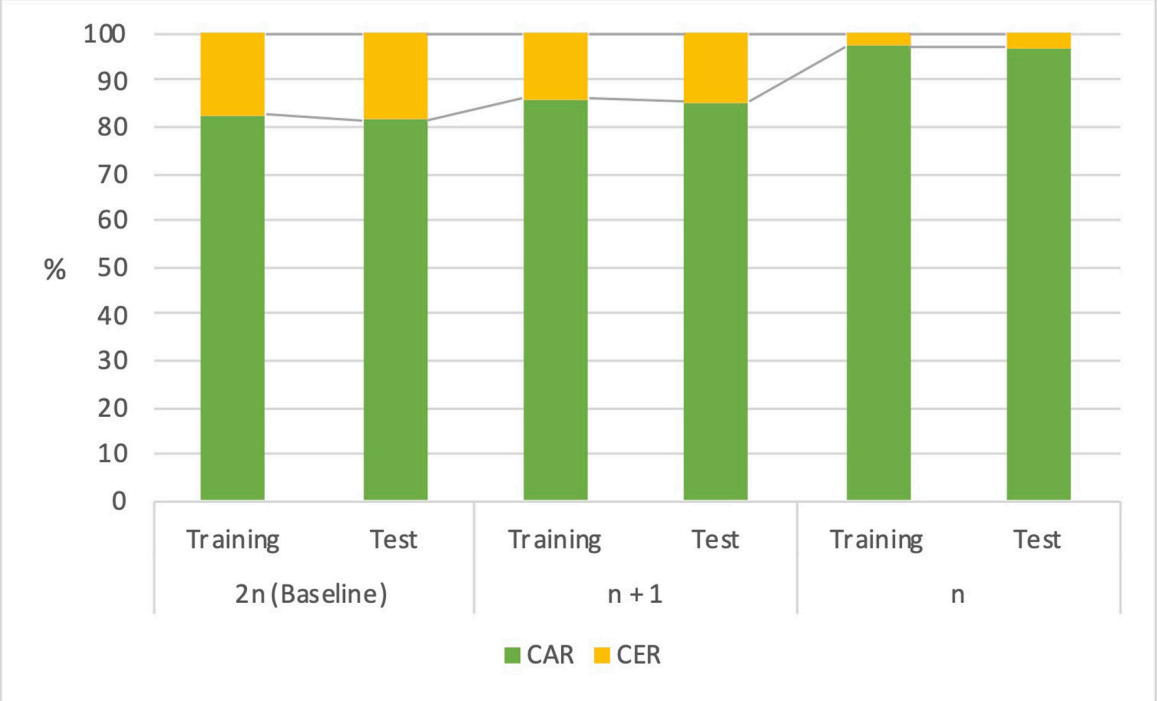

Figure 7: Overall Composite Accuracy and Error rates (summarized in Table 7).

three experiments the accuracy rates (*AR* and *CAR*) are only slightly less optimistic for the test data sets, with differences in *AR* ranging from roughly 0.82 to 1.71 percentage points (for experiments 2^n and n respectively). Conversely, the error rates (*ER* and *CER*), increase only very slightly in the test data sets (a difference that ranges from 0.18 to 0.38 percentage points).

The results presented in **Table 7** are illustrated in **Figures 6** and **7**. Note that for all Figures within this section, the scale of the y-axes may change as best suited to provide a clearer visualization of the data.

The difference between *Accuracy Rate* (*AR*, **Figure 6**) and *Composite Accuracy Rate* (*CAR*, **Figure 7**) across the three experiments illustrates how the classifier expectedly changes predictions as it becomes more restricted with respect to where in a word token a vowel can be predicted to be stressed. In particular, the decrease in the difference between the *AR* and *CAR* from experiment 2^n to experiment $n+1$ is noteworthy, because it indicates that for a sizable number of aligned tokens, the classifier predicted that both the vowel in citation position and some vowel in pretonic position were stressed in experiment 2^n, and that when the choice was restricted to one vowel per word token in experiment $n+1$ (and also in experiment n), the classifier was significantly more likely to predict that the vowel in citation position was stressed (*in lieu* of predicting that a pretonic vowel was stressed).[41]

The visible decrease in the *Error Rate* (*ER*, or the rate of *Mismatches*) from experiment 2^n to experiment $n+1$ reflects the proportion of aligned word tokens where the vowel in citation position and also a posttonic vowel were predicted to be stressed (roughly 2/3 of the total number of the total *Mismatches* in experiment 2^n) and which were then mostly predicted to have a stressed vowel in citation position in experiment $n+1$.

The increase seen in the *Composite Error Rate* (*CER*) from experiment n to the other two experiments is a by-product of how the experiments were designed: in experiment n a word token where none of the vowels bear stress (the *Unstressed* option) is not a possibility. Given the amount of word tokens classified as *Unstressed* in experiments $n+1$ and 2^n, it follows that a substantial number of those fell under the *Matches* and *Partial Matches* categories during experiment n, resulting in a visibly lower *CER* for this experiment.

In experiment 2^n, for approximately 73% of the word tokens counted as *Partial Matches*, as mentioned above (a prediction of a stressed vowel in pretonic position) the vowel in citation position was also predicted to be stressed. This percentage, along with the decrease seen in the rate of *Partial Matches* from experiment 2^n (see **Figure 6**) to experiment $n+1$ (roughly 9%) and to experiment n (approximately 7.4%), illustrates that ¾ of the times in which a vowel in pretonic position was predicted to be stressed the vowel in citation position was also predicted to be stressed, and that for the most part the prediction of stress locus fell back to the vowels in citation position when the choices for stress placement were restricted (experiments n and $n+1$).

A somewhat unexpected and interesting datum found in **Table 7** and in **Figure 6** is the fairly high rate of word tokens predicted to be completely *Unstressed* in experiments $n+1$ and 2^n, totaling roughly 12.5% of the word tokens in the training data set, and about 13.2% of the word tokens in the test data set. Note that the rate remains virtually unchanged in both experiments, indicating that it is not a byproduct of the experiments' design, or of chance classification, or of imbalances in the data sets, warranting a more detailed analysis.

The first thought that comes to mind is the possibility that the rate of *Unstressed* word tokens may be driven by (content) monosyllabic word tokens, which account for a sizeable portion of the corpus data, since for these there are only two choices, the (sole) vowel in the word token is either predicted to be stressed or unstressed. A glance at the distribution of word tokens predicted to be *Unstressed* in experiment 2^n as a function of both the number of syllables and the position of stress as shown in **Table 8** below, may prove informative.

The data in **Table 8** shows that disyllabic and trisyllabic words account for approximately 2/3 of all the word tokens predicted to be *Unstressed*, and that penultimately stressed

[41] Given that the *Error Rate* goes down from experiment 2^n to experiment $n+1$, and that the rate of *Unstressed* word tokens is constant.

Table 8: Word tokens predicted to be *Unstressed* per number of syllables and stress position.

Num Syll	Ultimate (%)	Penultimate (%)	Antepenultimate (%)	Grand Total (%)
1	12.90	N.A.	N.A.	12.90
2	27.75	39.84	N.A.	67.58
3	2.94	12.04	0.57	15.55
4	0.51	2.54	0.05	3.10
5	–	0.81	0.00	0.81
6	–	0.05	–	0.05
Grand Total	**44.10**	**55.28**	**0.62**	**100**

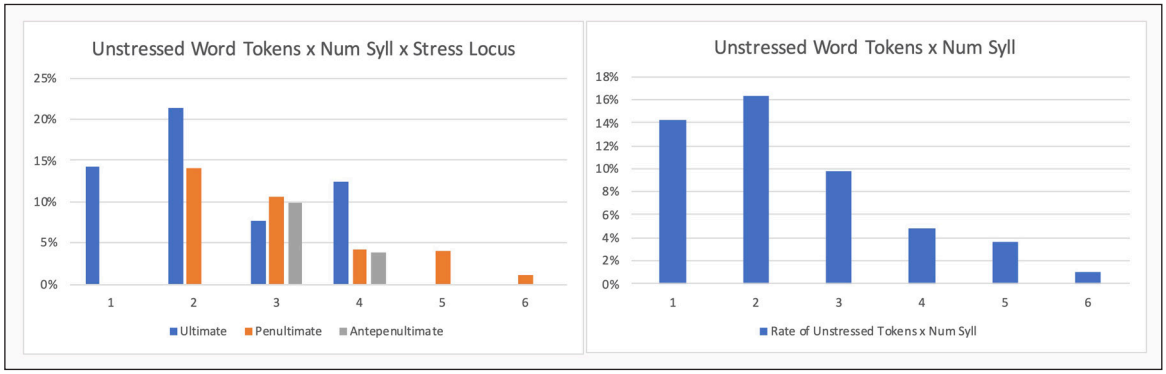

Figure 8: Distribution of Unstressed tokens as a function of the number of syllables and stress position.

words make more than 60% of the tokens predicted to be *Unstressed*. In a complementary perspective, **Figure 8** below shows the rate of word tokens predicted to be *Unstressed* with respect to the total number of word tokens of that size and stress locus.

In the Figure, while roughly 14% of (content) monosyllables were predicted to be *Unstressed*, a little more than 16% of the disyllabic and almost 10% of the trisyllabic word tokens were also predicted to be so. Furthermore, there are also tetrasyllables (4.81% of the total number of tetrasyllabic words) and pentasyllables (3.6% of the total number of 5-syllable words) predicted to be *Unstressed* (and even a very small number of hexasyllabic words). Therefore, it seems that the rate of word tokens predicted to be *Unstressed* is not being driven uniquely by monosyllables, or, put differently, that the rate of word tokens predicted to be *Unstressed* is not a byproduct of the size of a word in syllables, either.

If the rate of word tokens predicted to be *Unstressed* is likely not a byproduct of the experiments' design, nor is it particularly associated to the size of the word in syllables, then this datum must reflect a phenomenon that is linguistic in nature.[42] A few possibilities arise: it may be the case that for these particular word tokens the spectral features and energy information used here alone do not capture stress information, and that perhaps another correlate (such as duration) is used to signal stress more unequivocally in these word tokens. It could also be that, for these particular word tokens, stress does not surface phonetically. These two hypotheses however, as phrased here, seem somewhat unlikely,

[42] Ad hoc bottom-up measurements of various acoustic correlates of the vowels that occupy citation position, as well as analyses of the word type (*content X function*), the grammatical category, and the type of stressed syllable *(open or closed syllable, oral or nasal monophthong, oral or nasal diphthong)* in the word tokens predicted to be *Unstressed* are being conducted as this paper is written, but the results fall beyond the scope of the present study.

since, first, the results so far have shown that spectral features and energy information capture stress fairly robustly, and second, it seems somewhat odd for stress to not surface phonetically in some random fashion. An alternative version for the former hypotheses would be that there is some phonological or prosodic process at work in these word tokens and that, as a consequence, either stress does not surface phonetically, or it is encoded using acoustic correlates that are not captured by the present model (in which case, the change in acoustic correlate in and of itself could potentially signal the phenomenon at hand). With respect to prosody, one contender that comes to mind is the position that the *Unstressed* word token occupies in the utterance. **Figure 9** illustrates this distribution, showing the averages of *Unstressed* word tokens for each position of a given utterance. The labels closest to the X-axis represent the offset of the word within the utterance and the numbers below those represent the size of the utterance in number of words (if a position is missing for a given utterance size it means that all words in that position for all utterances of that size are monosyllabic function words). For each utterance size plotted in the graph (1-word to 10-word long), there are two or more distinct prompt sentences.

In the Figure above there is a trend whereby the rate of word tokens predicted to be *Unstressed* is visibly lower at the end of the utterance and is consistently higher towards mid-utterance position, which seems to indicate that word tokens are more likely to be predicted to be *Unstressed* in the prosodically weak positions of the utterance, and less likely so in positions of nuclear accent.[43] Note also that in the longer utterances (e.g., 8, 9, and 10 words), the pattern seems to repeat itself, in what could be the effects of two Intonational Phrases. Only a fully-fledged analysis of the trends seen in **Figure 9** however,[44] would shed light on the true crux of the matter, which is, is stress in these *Unstressed* word tokens more phonological in nature (i.e., it does not surface phonetically), or does stress surface through different acoustic correlates, or combinations thereof, in the weaker positions of the utterance?

Since the results so far have shown that the outcomes are consistently more conservative (even if just slightly so) for the test data sets, the analyses presented in the remainder of this subsection will subsume these results only. **Figure 10** presents the results for monosyllables in the corpus, summarized into two classes, that of function and of content words (results from experiment 2^n). Although there is a large body of literature describing that monosyllabic function words, but not monosyllabic content words, are destressed cross-linguistically, there is no work in BP, to the best of this author's knowledge, that

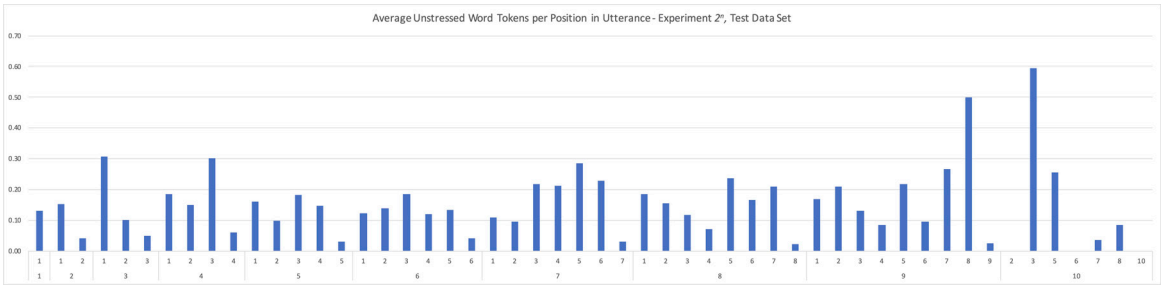

Figure 9: Distribution of *Unstressed* word tokens as a function of the position of the word in the utterance.

[43] See, for example, De Moraes (2007) for a summary of intonational patterns in BP.

[44] A detailed analysis would have to look separately at the prompts in the WPC, controlling the predicted *Unstressed* tokens for stress locus, grammatical category, compound status number of Intonational Phrases in the utterance, type of sentence, to name a few. Such analysis would ideally add instrumental measurements of the acoustic correlates known to subsume stress in the word tokens of interest. This analysis is being conducted in work parallel to the present study.

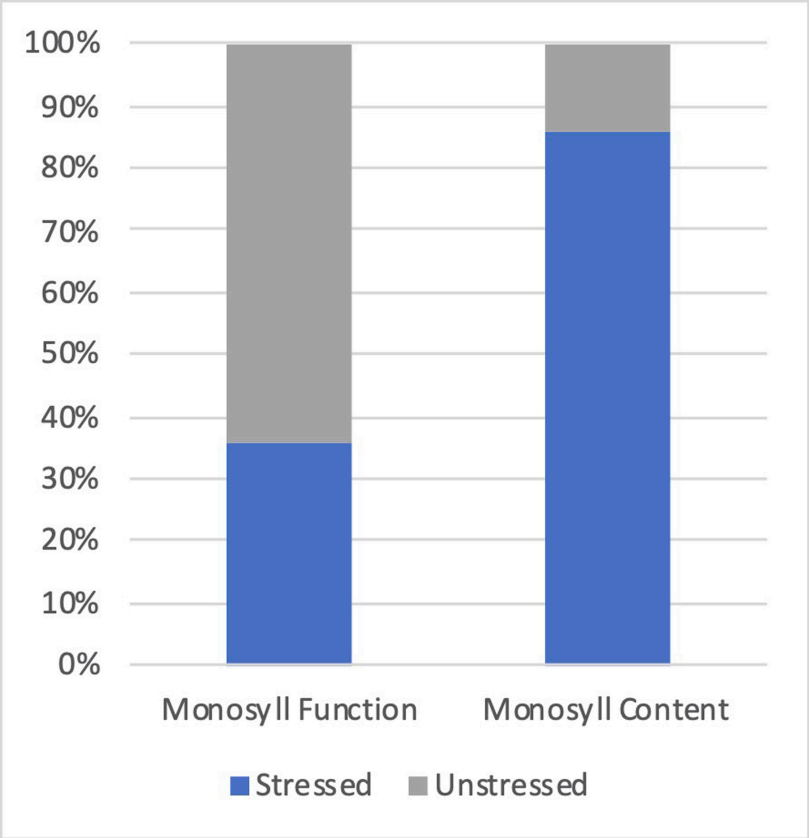

Figure 10: Results (Test data set of experiment 2^n) for *function* and *content monosyllabic* words.

shows this experimentally. The results fit these accounts well, as approximately 86% of the monosyllabic content word tokens were predicted to be stressed, but only around 35% of the function words fell in that prediction pattern. Note that these results are in agreement with phonological accounts whereby function words may be incorporated into adjacent phonological words.

Lastly, **Figure 11** illustrates the breakdown of the results by stress locus, across all three experiments. Monosyllabic content words were not included in the plots, so to show all stress positions without the bias these words may introduce for ultimate stress.

While the results for experiment n and $n+1$ show a trend whereby the *Accuracy Rate* increases the farther the stress locus falls from the right edge of the word, no clear pattern can be inferred form the results for experiment 2^n. From this perspective it would be interesting to perform instrumental measurements of the acoustic correlates of the stressed vowels to see if a pattern indeed emerges with respect to how systematically these can tell apart the stressed vowels in the different positions of the word, in the lines of the word of Delgado Martins (1986, apud Magalhães, 2016) for EP, where it was found that duration and energy can be established instrumentally for ultimate and antepenultimate stress, but less systematically so for penultimate stress. Importantly, these results illustrate that the rate of word tokens predicted to have exactly the same shape as the citation word is very high across experimental conditions, for all three positions of stress.

4.2. The Classifier's Discriminatory Skills

While the data discussed so far outlines how reliably stressed vowels are distinguished from the surrounding vowels in a given word token, they do not provide a detailed picture of discriminatory capacity of the model, or of how well it distinguishes between stressed

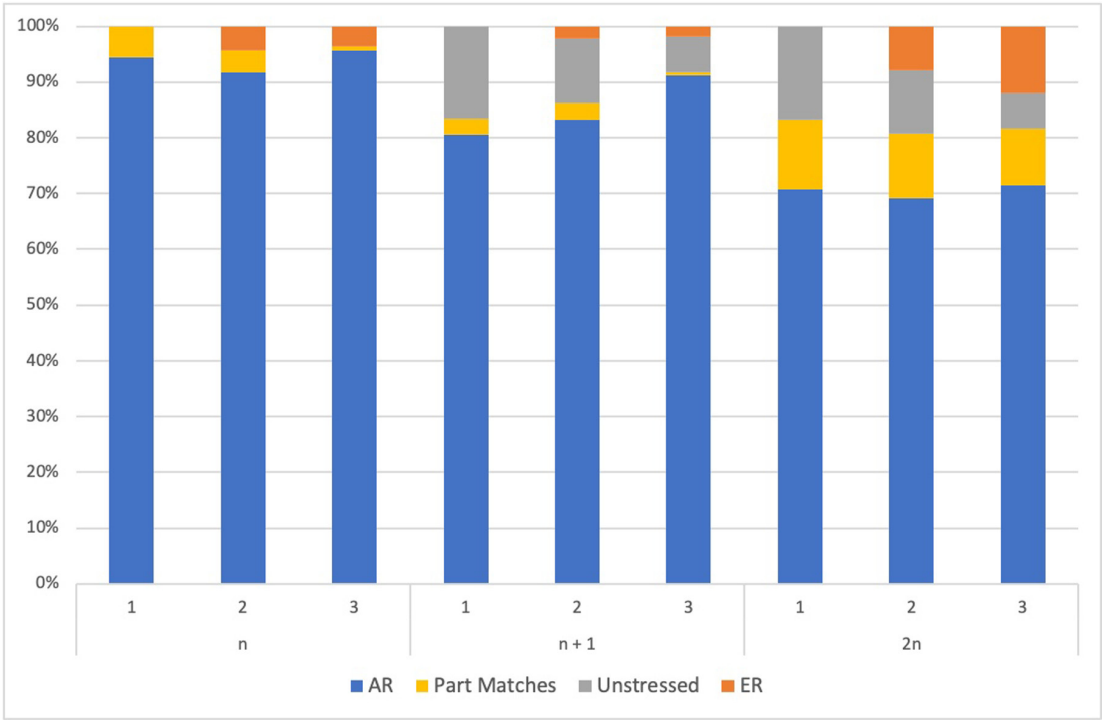

Figure 11: Distribution of aligned word tokens as a function of stress position.

and unstressed vowels given the ground truth. To do so, the data are summarized in confusion matrices and evaluated according to the metrics discussed in the Data Analysis subsection of this paper.

Table 9 shows that, analogously to the metrics discussed in the previous paragraphs, here too, for each of the three experiments, the rate of change between the metrics calculated for the training data sets and those calculated for the test data sets in each experiment is very small, not surpassing 2 percentage points for an individual metric. **Figure 12** illustrates the results graphically.

The accuracy rates show that for all vowels predicted to be stressed and for all vowels predicted to be unstressed, the rate of truly stressed and truly unstressed vowels is very high across experiments and data sets (range 86%–94%). This means that the spectral features and energy information encoded in the MFCCs (modelled in HMM-GMMs) allow for the distinction between a stressed vowel and an unstressed vowel to be made with fairly high accuracy in BP.

The values for *Sensitivity* and *Specificity* indicate that, for all experiments and all data sets, the classifier is slightly better at correctly predicting that a vowel is unstressed when that vowel is expected to be unstressed, than it is at predicting that a vowel is stressed when that vowel is expected to be so. In an oversimplified way, these values indicate that, as represented by spectral and energy information encoded in MFCCs, unstressed vowels are slightly more likely to fit the model of an unstressed vowel than stressed vowels are likely to fit the model of a stressed vowel. *Sensitivity* is highest in experiment n and *Specificity* is highest in experiment $n+1$. Computationally this is a meaningful piece of information because it would help to decide which experiment is best suited if one is more interested in finding stressed or unstressed vowels in a speech corpus.

The results for *Precision* in the test data set show that when the classifier makes a positive prediction (a *stressed vowel* prediction), it is right 82% of the times in experiment 2^n, 94% of the times in experiment n, and 96% of the times in experiment $n+1$. *Precision* is an

Table 9: Evaluation metrics for the classifier's overall discriminatory skills.

Experiment	Data Set	*Accuracy*	*Sensitivity*	*Specificity*	*Precision*	*F1-Score*	*MCC*
2^n (Baseline)	Training	0.86	0.86	0.86	0.83	0.84	0.71
	Test	0.85	0.85	0.85	0.82	0.83	0.69
n+1	Training	0.91	0.84	0.97	0.96	0.89	0.82
	Test	0.90	0.82	0.96	0.95	0.88	0.80
n	Training	0.94	0.94	0.95	0.94	0.94	0.88
	Test	0.93	0.93	0.94	0.93	0.93	0.87

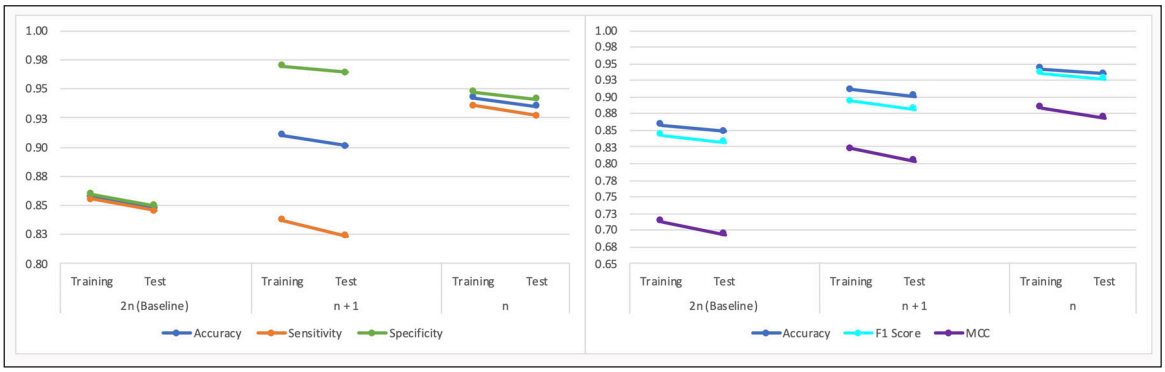

Figure 12: Results for the classifier's overall discriminatory skills.

important metric in as much as it illustrates the performance of the test: it could be, for example, that high values of *Sensitivity* were achieved because the classifier only makes positive predictions, and thus a considerable number of them are bound to be right.

The *F1-scores* provide an alternative measure of the classifier's accuracy, a compromise between Precision and Sensitivity. The *F1-scores*[45] shown in **Table 9** are almost as high as *Accuracy* for all three experiments, indicating that there is good balance between *Precision* and *Sensitivity*.

Recall from the Methodology section that the *MCC* ($-1 \leqslant MCC \leqslant 1$) measures the strength of correlation between the predictions and the ground truth taking into consideration the proportion of each class within the confusion matrix (*TP, TN, FP, FN*). The results achieved for the *MCCs* indicate a very strong correlation between predictions and ground truth across experiments and data sets. Put differently, the strength of correlation we see between predictions and the ground truth is not affected by the fact that language is imbalanced (the number of expected true negatives is much higher than the number of expected true positives).

Table 10 shows the *Cohen's Kappa Coefficients* (*Kappa*), which represent the chance-corrected strength of the agreement between predictions and the ground truth. For each *Kappa* calculated, for the standard errors (SE) listed in the Table, there is a 95% chance that they would fall within the confidence intervals also shown in the Table. The *Kappas* for all experiments and all data sets are considered substantial (from 0.61 to 08) to almost perfect (0.81 to 1) in the interpretation given by Landis & Koch (1977) and are all considered excellent in the view of Cicchetti & Sparrow (1981).

[45] Note that although the *F1-score* is a measure of accuracy, it does not take into account true negatives, but *Accuracy* does.

Table 10: *Cohen's Kappa Coefficients* for the Training and Test data sets.

Experiment	Data Set	*Kappa*
Baseline (2^n)	Training	0.7130, SE = 0.003, 95\% conf int.: 0.706 to 0.719
	Test	0.6930, SE = 0.006, 95\% conf int.: 0.680 to 0.706
n+1	Training	0.8170, SE = 0.003, 95\% conf int.: 0.811 to 0.822
	Test	0.7980, SE = 0.005, 95\% conf int.: 0.787 to 0.809
n	Training	0.8830, SE = 0.002, 95\% conf int.: 0.879 to 0.887
	Test	0.8680, SE = 0.004, 95\% conf int.: 0.86 to 0.877

The results for *MCC*s and *Kappas* for all experiments and all data sets help to ascertain that the strength of the agreement between predictions and the ground truth is neither a product of an imbalanced data set, nor the result of chance agreement.

In the following paragraphs, analogously to what was done in the previous subsection, a breakdown of the metrics calculated from the confusion matrices is detailed for the test data sets only. **Figure 13** shows the metrics as a function of stress locus in a word. Monosyllabic content words were not included in the computations here to avoid bias on the metrics calculated for ultimate stress words (since *False Positives* (*FP*) and *True Negatives* (*TN*) will never occur in a monosyllabic word).

Recall from the results discussed in the previous subsection that there seemed to be a trend in experiments n and $n+1$ whereby *AR* showed that predictions became more accurate as the stress locus got farther from the right edge of the word, but that no pattern in this respect could be inferred from the results in experiment 2^n. The graphs in the two first rows of **Figure 13** show data that appear to corroborate this trend: across all experiments, all of the metrics shown in the graphs improve as the locus of stress moves away from the right edge of the word. In the lines of work like that of Gahl, Yao & Johnson (2012)—who showed that in (English) spontaneous speech words that are infrequent and have low phonological neighborhood density are more carefully articulated[46]—one explanation for this asymmetry could be that penultimate and antepenultimate words in the WPC are more carefully articulated,[47] and as a consequence the vowels in them are better exemplars of both stressed and unstressed vowels, which results in more accurate predictions. An alternative explanation would be that spectral features and energy information capture stress more systematically as it falls farther from the right edge of the word, and instrumental measurements of other correlates of stress may or may not reproduce such asymmetry. Work for European Portuguese (EP) has shown an effect in these lines, in which duration and energy were found to be the most reliable correlates of stress in the language, but were instrumentally less systematic so for penultimate words (Delgado Martins 1986, apud Magalhães 2016) than for ultimate and antepenultimate words. It is worth noticing that, while the asymmetry between ultimate and penultimate words stems from a data set that contains 9,771 word tokens and 145 different word shapes in the former category, and 17,837 word tokens and 345 different word shapes of the latter category, in the case of antepenultimate words, only 365 word tokens and 11 word shapes determined the results here. Therefore, additional work using a data set that

[46] Where the number of phonological neighbors is the number of words in the language that differ in a single phoneme, by means of substitution, addition, or deletion and where articulatory effort is measured through vowel dispersion.

[47] The idea being that words that bear penultimate and antepenultimate stress are inherently (and progressively) longer, and maybe also less likely to have a large number of phonological neighbors.

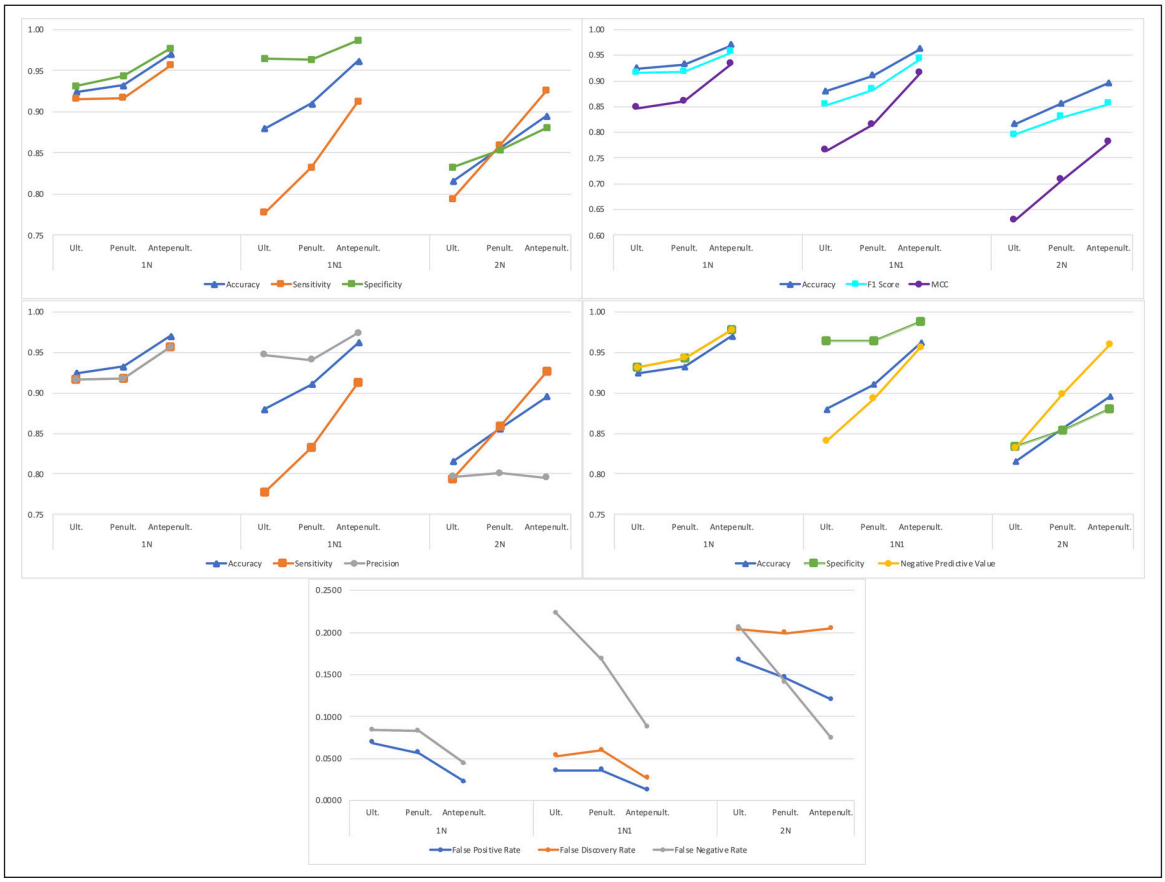

Figure 13: Discriminatory skills as a function of stress locus.

contains a larger number of antepenultimate word tokens and a greater variety of word shapes should shed additional light on this matter.

While *Sensitivity* and *Specificity* in **Figure 13** mostly follow the general patterns seen in the overall results (**Figure 12** earlier in the text), in experiment 2^n, for antepenultimate words only, the rate of stressed vowels correctly identified as such among all vowels expected to be stressed becomes higher than the rate of unstressed vowels correctly identified as unstressed among all the truly unstressed vowels. This pattern likely stems from the design of the experiment, which allows vowels in all positions of a given word to be predicted as stressed. Indeed, in looking at the graphs in the second row for the same experiment, as *Sensitivity* increases for the antepenultimate words, *Precision* decreases, indicating that the ratio of correct positive predictions for all positive predictions made became lower. This trend is also visible on the *F1-score* for that category in the same experiment.

The *F1-scores* show that the balance between *Precision* and *Sensitivity* improves (*F1-score* plot is closer to *Accuracy* plot) for antepenultimate words in relation to penultimate and to ultimate words. The strength of the correlation between predictions and the ground truth, given by the *MCCs*, is also highest for antepenultimate words, as is the chance-corrected agreement between predictions and ground truth shown in yellow here.[48] The latter two metrics confirm, for all three stress loci, that the predictions are not a product of the imbalance in the class sizes, nor a result of chance classification.

[48] Because there are fewer antepenultimate word tokens in the data set, the confidence intervals for the antepenultimate *Kappas* are wider than the confidence intervals for the other stress loci.

On the second row of **Figure 13**, the graph on the right adds a plot for the *Negative Predictive Value*, the rate of correctly made predictions of unstressed vowels for all the unstressed vowel predictions made, which is a tradeoff with *Specificity*, a measure of the tests' performances in finding relevant results for unstressed vowels. The results for *NPV* show that the rate of relevant negative predictions (an *unstressed* prediction) increases as citation stress locus falls farther from the right edge of the word in all experiments. *NPV* is highest for all stress loci in the least complex experiment (n) and is very close in value for the more complex experimental conditions ($n+1$ and 2^n). The values for *Precision* show that there is a less clear trend in what concerns finding relevant positive results for the different stress loci, as the plot slightly dips for penultimate words before rising for antepenultimate words.

Because the various measures of accuracy presented here are very close in values, the differences are often harder to visualize, so the last graph in **Figure 13** shows the *False Positive Rate* (*FPR*), the *False Negative Rate (FNR*) and the *False Discovery Rate* (*FDR*), a complementary view with respect to the accuracy metrics where changes in value are easier to spot. The values for *FPR* show that the rate of vowels incorrectly predicted to be stressed among all vowels expected to be unstressed is lower the farther the stress locus gets from the right edge of the word, for all experiments, meaning it is less likely that we find an incorrect prediction in antepenultimate then in penultimate and in ultimate words when the vowel is expected to be unstressed. The trend for *FNR* follows the same path of *FPR*, but in a much sharper fashion, where it is considerably less likely to find incorrect predictions among the vowels expected to be stressed in antepenultimate words.

Lastly, we move on to look at unstressed vowels separately in pre-tonic and in posttonic positions of the word. Since asymmetries between these positions have been previously reported in the literature (e.g., Câmara Jr. 1970; Major 1985) of Brazilian Portuguese, an analysis of the results obtained herein could potentially shed some additional light on the matter. **Figure 14** summarizes the results for *Specificity* and for *False Positive Rate* for pre- and posttonic vowels. In the graph, recall that there are no posttonic positions in ultimate words and no pretonic positions in monosyllabic ultimate, disyllabic penultimate, and trisyllabic antepenultimate words.

The results shown in **Figure 14** present a fairly clear trend in which, for all experiments, and for all sizes of words within all stress loci, the rate of vowels correctly identified as unstressed among all the vowels expected to be unstressed (given by *Specificity*) is higher for posttonic positions. In other words, it is more likely for unstressed vowels in posttonic positions to be identified as such, than it is for unstressed vowels in pretonic positions of the word. The results from experiment 2^n are the most telling, since in this experimental condition there are no restrictions to predicting that more than one vowel in the same word token is stressed. Indeed, the data indicate a clearer and sharper separation of pre- and posttonic vowels in this experiment. Showing the converse perspective, the graphs on the right side of the figure illustrate that the rate of *False Positives* (*FPR*) is higher for pretonic vowels in all experiments and word sizes within the three stress loci. The difference is again sharper in experiment 2^n. Note that the asymmetry persists when there is imbalance between the number of pretonic and posttonic vowels in a word.

The question of whether the values of *Specificity* and *Sensitivity* in pretonic vowels can help to shed light on the nature and behavior of secondary stress in BP, while certainly a worthy one, falls beside the scope of the present study.[49]

[49] An analysis of the findings related to this matter is underway.

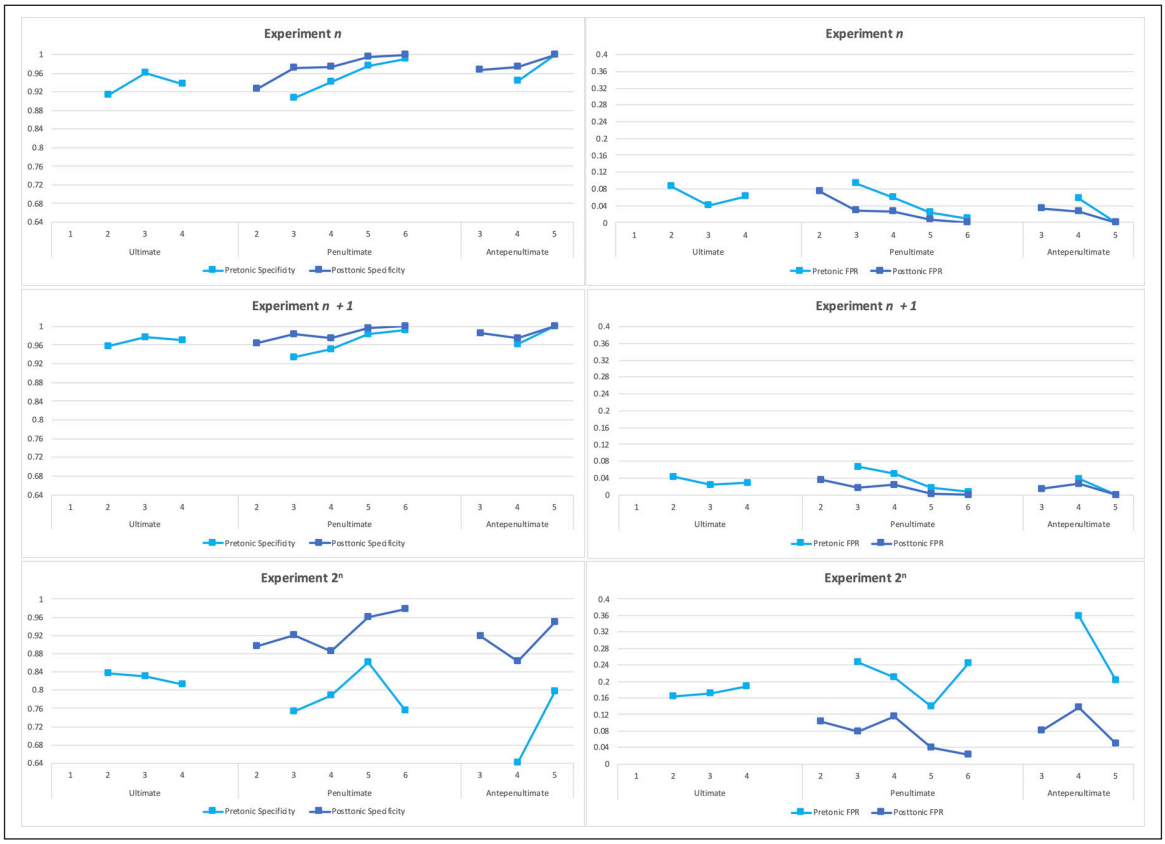

Figure 14: Discriminatory skills for pretonic and posttonic vowels.

4.3. Discussion Synopsis

The results of the three experiments performed for the present study were discussed in this section. Results were averaged over the five folds of the experiments and then summarized and analyzed from two distinct and complementary perspectives, which provide insights on the syntagmatic and the paradigmatic nature of stressed and unstressed vowels as captured by a compressed representation of the speech signal that subsumes spectral features and energy information (MFCCs). The overall results for all metrics in both analyses, were, as anticipated, slightly less optimistic for the test data sets than for the training data sets, prompting the more detailed analyses to be summarized for the test data sets only.

From a syntagmatic perspective, the results have shown robustly that stressed vowels distinguish themselves from the surrounding vowels, even in a complex classification problem (experiment 2^n), as shown by the *Accuracy Rate* (*AR,* or the rate of word tokens predicted to have the exact same shape of the citation form for that word) of 69.7% of the word tokens (19,498 in all). In the less complex experiments, *AR* approximates 93%. For experiments n and $n+1$, and only slightly so for experiment 2^n, this syntagmatic effect is more noticeable the farther the stress locus gets from the right edge of the word—the rate of antepenultimate words that fits perfectly in the citation form of the word is larger than the rate of penultimate, than the rate of ultimate words.

Still from a syntagmatic perspective, the rate of tokens predicted to be completely *Unstressed*, a virtually steady 13% of all the word tokens in the corpus, in both experiment 2^n and $n+1$, came as somewhat of a surprise. Having established that this rate is neither a byproduct of the design of the experiments, nor is it driven by monosyllabic words, a couple of options arise to explain these findings: one possibility is that, for a number of

word tokens, stress cannot be captured using representations of the speech signal that subsume spectral features and energy information. Another possibility is that, in these tokens, stress is more of a phonological property, not necessarily surfacing phonetically. Independently of which of these two possibilities are correct, perhaps the crucial matter lies in trying to understand whether there is some phonological or prosodic process (or processes) at work in these *Unstressed* word tokens. In an attempt to do some inception work to answer this question, a brief analysis of *Unstressed* word tokens as a function of the position they occupy in the utterance was presented. This analysis showed a fairly clear trend in which *Unstressed* tokens are more likely to appear in the prosodically weaker positions of the utterance, for the different utterance sizes (in number of words) in the corpus. A much more detailed analysis is needed though, in trying to determine whether stress is more phonological in nature for these tokens, or whether, in these *Unstressed* tokens, stress is better captured through different acoustic correlates (in which case one might ask what the reason for such change would be: could it be to signal some other prosodic aspect of the utterance?).

A brief analysis of function and content monosyllabic words confirmed the trend widely reported in the literature cross-linguistically, which is that function monosyllabic words are generally destressed (only about 35% of them predicted to be stressed in experiment 2^n), but not content monosyllabic words (approximately 86% of which were predicted to be stressed in experiment 2^n). While these results are expected given the vast literature ascertaining so, there are no experimental studies looking at these two classes in BP, to the best of this author's knowledge.

From a paradigmatic perspective the results proved to be yet more robust, with high values of *Accuracy*, *Sensitivity* and *Specificity* in all three experiments for all stress loci. These results revealed that stressed and unstressed vowels, as captured by MFCC representations of the signal and modelled in the HMM-GMM framework, are systematically different from one another. The values for the *Matthew's Correlation Coefficients* show that the correlation between predictions and the ground truth is still very strong after taking the imbalances in the data set into account, and the *Cohen's Kappa Coefficients* indicated a very strong to almost perfect chance-corrected correlation between predictions and ground truth.

Noteworthy, also from a paradigmatic perspective, is the fact that there is a clearer trend in which stressed vowels and unstressed vowels appear to be more distinctly so as the locus of stress in the citation form of the word moves away from its right edge. Differently put, both stressed and unstressed vowels are more likely to be accurately predicted as such in antepenultimate words then they are in penultimate words, then they are in ultimate words. A parallel asymmetry was shown to exist for penultimate words in European Portuguese in the work of Delgado Martins (1986, apud Magalhães 2016). Whether or not these asymmetries are rooted in cognitive processes, in line with work such as Gahl, Yao & Johnson (2012), and whether they hold after studies on larger data sets establish instrumentally the acoustic correlates of stress in point, are some of the questions that remain open.

Lastly, an analysis of *Specificity* and of *False Positive Rate* (*FPR*) for pre- and posttonic vowels separately revealed an asymmetry between these positions in the word, as previously reported in the literature for BP (e.g., Major 1985; Câmara Jr. 1970). For all experiments and all word sizes within all stress loci, posttonic vowels are more likely to be predicted to be unstressed than pretonic vowels are so. The effect holds despite the imbalance in the number of pre- and posttonic vowels in words of different sizes and stress locus.

Several metrics in both analyses—including the rate of *Unstressed* word tokens, the change in rates of *Partial Matches* and *Error*, *Accuracy* rates for monosyllabic content word tokens, the progression of *Precision*, *Sensitivity* and *Specificity*, among others—indicated inter-experiment agreement where expected. Conversely, other metrics confirmed differences where they would be expected, given the design of each experiment—e.g., a perfect agreement for monosyllabic content words in experiment *n*, lower word token *Accuracy Rate* (perfect matches) in experiment 2^n, and *Accuracy* (from the confusion matrix) of 1 for monosyllabic content words in experiment *n*, among others.

5. Final Thoughts

This study examined primary word-level stress in continuous speech using the LDC West Point Speech corpus (Morgan et al. 2008) as dataset and the ASR toolkit Kaldi (Povey et al. 2011). Specifically, I investigated whether stressed vowels are systematically different from the vowels that surround them (a syntagmatic comparison), and whether stressed vowels are systematically different from unstressed vowels (a paradigmatic comparison) when captured by acoustic features of the speech signal (MFCCs, HMM-GMMs) which are known to subsume spectral features and energy information, but not information about the time-domain (*duration*) or the source characteristics (*F0*) of the signal. This was done by building a linguistically informed list of phones and a phonetic dictionary modelled for explicit pronunciation, which were used to train an acoustic model of Brazilian Portuguese. The model was then tested in three increasingly complex classification experiments, with the purpose of measuring the strength of agreement between the predictions made and ground truth defined by the citation form of each word.

The results obtained showed that stress is robustly realized in most content word tokens in all three experimental conditions and for all three loci of citation stress, across speakers and across genders. These results indicate that spectral features and energy information can systematically capture the differences between stressed vowels and the vowels that surround them in a given word token, and also the differences between stressed and unstressed vowels as two classes. This is a departure from previous literature (e.g., Barbosa, Eriksson & Åkesson 2013; Major 1985; Massini 1991), which showed that only duration robustly distinguishes stressed vowels from unstressed vowels in BP.[50]

In the experimental conditions that allowed for such a choice (condition 2^n and condition $n+1$), a non-negligible rate of word tokens were predicted to be completely *Unstressed* (roughly 13% of all word tokens). While it is possible that stress cannot be captured using representations of the speech signal that subsume spectral features and energy information for these particular word tokens (but maybe another acoustic correlate of stress could), one should also consider the likelihood that in these tokens stress is more of a phonological property, thus not necessarily surfacing phonetically. Importantly, a crucial matter lies in trying to understand whether there are other phonological or prosodic process(es) at work in these *Unstressed* word tokens. A preliminary analysis of *Unstressed* word tokens as a function of their position in the utterance ($1 \leqslant$ *utterance size (words)* $\leqslant 10$) indicates a trend whereby word tokens are more likely predicted to be Unstressed in prosodically weaker positions of the utterance. As discussed in the Synopsis section above however,

[50] I thank an anonymous referee for pointing out that in the work of Arantes, Lima & Barbosa (2012: 17) the authors mention that vowels in stressed syllables have higher mean spectral emphasis than those in other positions of the word, which could thus indicate that it is a correlate of primary word stress. Later work by Barbosa, Eriksson & Åkesson (2013) reported that only duration was the most consistent correlate of stress in BP.

much remains to be done in order to sketch a more comprehensive picture of the possible interaction(s) between phonetic stress and the position of the word in the utterance.

The results and the metrics calculated to evaluate them, revealed two asymmetries with respect to stress in BP. The first asymmetry regards pre- and posttonic vowels (or syllables) and has been reported in previous literature on stress in BP (e.g., Câmara Jr. 1970; Major 1985), but was not corroborated by the results from previous experimental work (e.g., Barbosa, Eriksson & Åkesson 2013). The results presented here showed a trend, for all experimental conditions and for all loci of stress, whereby posttonic vowels are more likely to be predicted to be unstressed than vowels in pretonic positions of the word token. This finding presents evidence that favors the assumption that posttonic vowels are unstressed. These results also indicate that pretonic vowels are more likely to be more similar to stressed vowels (since they are more likely to be predicted to be stressed than posttonic vowels are), raising further questions about the nature secondary stress in BP.

The second asymmetry uncovered in the results of this study regards stress locus. Stressed and unstressed vowels are more likely to be correctly predicted as such in antepenultimate words, followed by penultimate words, followed by ultimate words. While further work is needed to ascertain this finding, both by experimenting with data sets that contain larger number of word shapes (as opposed to the 516 word shapes in this study), and by performing further instrumental measurements of the spectral features and energy information of the vowels that occupy the stress position in the citation form of the word, it is not unheard of that a certain acoustic correlate (or a set of) is more systematic in expressing stress in some positions of the word than it is in others, as discussed in the Results section.

One of the strengths of the present study lies in the number of speakers included (speakers = 99, fairly balanced for gender), in the number of repetitions of each sentence found in the corpus' prompts ($22 \leqslant repetitions \leqslant 98$, for a total of 7,844 utterances and of 39,894 word tokens), and also in the number of word shapes under scrutiny, when compared to traditional phonetic studies. Methodologically, the study is designed to explore boundaries in the use of machine learning and in top-down approaches to conduct phonetic and phonological studies. It innovates in adopting a data analysis method that evaluates stressed and unstressed vowels paradigmatically while also emulating a syntagmatic comparison without performing direct instrumental measurements.

The general approach instantiated here is to put linguistic research on prosody and other aspects of surface phonological form into contact with corpus data using a robust, probabilistic model mapping the linguistic surface form to the signal. There is the potential to apply this methodology to other aspects of the Brazilian Portuguese prosodic system, and to segmental phonology. Pending corroboration of the results reported herein with results of future studies performed in speech corpora containing a greater diversity of word shapes, the method developed for this study illustrates what can be accomplished in speech corpus phonetic/phonology using a relatively small amount of data and computationally cheaper models, such as HMM-GMMs (as opposed to various Neural Network alternatives) and context independent monophones.

Further analyses of the results herein reported are currently being performed to explore how they relate to syllable weight and grammatical category, in the interest of shedding further light on the phonetics/phonology interface of stress in Portuguese. Since it has been shown that phonological weight distinctions are correlated with phonetic duration (see, for example, Broselow, Chen & Huffman 1997) and with the total energy of the syllable rhyme (see Gordon 2002; 2006), it becomes interesting to explore whether the results presented herein echo the (QS) stress system of non-verbs as described in the various accounts for stress placement in Portuguese discussed in section 2.1 herein.

While the study's motivations are rooted in linguistic research, it also has applications in speech recognition technology. For instance, the approach defined here can straightforwardly be used to add stress information to a speech recognition phonetic lexicon that lacks stress information.[51]

The analyses presented throughout the paper were detailed but by no means exhaustive, given the nature and complexity of stress as a variable. As these final thoughts are being written *ad hoc* measurements are being taken to try and better understand the nature of the *Unstressed* word tokens. An analysis of classification for pretonic vowels is also underway, in order to delineate their behavior in the corpus. Analyses of the results as a function of vowel quality, and as a function of the phones neighboring the vowels in stress citation position of a given word can also be potentially informative.

A reservation about the study computationally speaking is that it uses a speech corpus with a few hundred word shapes. Although this is a higher than usual number of word shapes for a phonetic study, in a computational experiment, it is possible that the performance of the model in predicting stress to some extent takes advantage of this. Thus, an important topic for further research is to test the model on corpus data with a much larger set of word shapes. A number of the trends identified throughout section 4 (Results and Discussion), including the more robust ones, will ideally be reproduced in studies with larger corpora and a greater number of word shapes, especially trisyllables and longer. Future studies should also investigate the effects that the incorporation of pitch information and the use of different modeling (such as DNN-HMMs) and adaptation techniques (such as fMLL) would have on the final results.[52]

Acknowledgements

I would like to thank Mats Rooth for invaluable comments and suggestions and Bruce W. McKee for instrumental discussions about binary classification. I also thank two anonymous reviewers for insightful comments that led to improvements in the article. Mistakes are my own. The present work came to life during the 2020 pandemic and would not have seen the light of the day without the invaluable support of the editors of this special issue.

Competing Interests

The author has no competing interests to declare.

References

Abaurre, M. B., & **Fernandes-Svartman, F. R.** (2008). Secondary stress, vowel reduction and rhythmic implementation in Brazilian Portuguese. In L. Bisol & C. R. Brecancini, *Contemporary Phonology in Brazil* (pp. 54–83). Newcastle, UK: Cambridge Scholars Publishing.

Ananthakrishnan, S., & **Narayanan, S. S.** (2008). Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE transactions on audio, speech, and language processing, 16*(1), 216–228. DOI: https://doi.org/10.1109/TASL.2007.907570

[51] In work in progress, we are using this approach to add stress information to the Aeiouadô lexicon described in Mendonça, G., & Aluisio, S. (2014). Using a hybrid approach to build a pronunciation dictionary for Brazilian Portuguese. In *Fifteenth Annual Conference of the International Speech Communication Association*.

[52] These are topics of work in progress, where a Kaldi model of the C-ORAL Brasil corpus (Raso, T., & Mello, H. (2012, April). The C-Oral-Brasil: a reference corpus for informal spoken Brazilian Portuguese. In International Conference on Computational Processing of the Portuguese Language (pp. 362–367). Springer, Berlin, Heidelberg.), is being developed, and studies using pitch information and different modelling techniques are in their planning stages.

Andrade, E. D., & **Laks, B.** (1991). Na crista da onda: O acento de palavra em português. [On the crest of the wave: The word accent in Portuguese]. *Actas do 7º Encontro da Associação Portuguesa de Lingüística*, 15–26.

Arantes, P. (2011). Prosódia e redução do espaço vocálico em português brasileiro. [Prosody and reduction of the vowel space in Brazilian Portuguese]. *Livro de Resumos do III Simpósio sobre Vogais (SIS Vogais)*. Porto Alegre, Novembro 2011. Available at https://sites.google.com/site/sisvogais3/

Arantes, P., & **Barbosa, P. A.** (2006). Secondary stress in Brazilian Portuguese: The interplay between production and perception studies. *Proceedings of the Speech Prosody 2006 Conference.*

Arantes, P., & **Barbosa, P. A.** (2008). F1 and spectral correlates of secondary stress in Brazilian Portuguese. In *Proceedings of the Speech Prosody 2008 Conference*, 559–562. Campinas, Brazil: RG/CNPq.

Arantes, P., Lima, M. L. C., & **Barbosa, P. A.** (2012). Some prosodic correlates of referential status in Brazilian Portuguese. *Revista Diadorim*, *12*, 1–25. DOI: https://doi.org/10.35520/diadorim.2012.v12n0a3969

Araújo, G. A. D., Guimarães Filho, Z. D. O., Oliveira, L., & **Viaro, M. E.** (2007). As proparoxítonas e o sistema acentual do português. [Antepenultimate words and the accent system of Portuguese]. In G. A. D. Araújo, (Org.), *O acento em português: Abordagens fonológicas* (pp. 37–60). São Paulo: Parábola.

Araújo, G. A., Guimarães-Filho, Z. O., Oliveira, L., & **Viaro, M. E.** (2008). Algumas observações sobre as proparoxítonas e o sistema acentual do português. [A few observations about antepenultimate words and the accent system of Portuguese]. *Cadernos de Estudos Lingüísticos*, *50*(1), 69–90. DOI: https://doi.org/10.20396/cel.v50i1.8637239

Barbosa, P. A. (2008, May). Prominence and boundary-related acoustic correlations in Brazilian Portuguese read and spontaneous speech. In *Proceedings of the Speech Prosody 2008 Conference*, 257–260. Campinas, Brazil: RG/CNPq.

Barbosa, P. A., & **Albano, E. C.** (2004). Brazilian Portuguese. *Journal of the International Phonetic Association*, *34*(2), 227–232. DOI: https://doi.org/10.1017/S0025100304001756

Barbosa, P. A., Arantes, P., & **Silveira, L. S.** (2004). Unifying stress shift and secondary stress phenomena with a dynamical systems rhythm rule. In *Proceedings of the Speech Prosody 2004 Conference*, 49–52.

Barbosa, P. A., Eriksson, A., & **Åkesson, J.** (2013). On the robustness of some acoustic parameters for signaling word stress across styles in Brazilian Portuguese. In *INTERSPEECH*, 282–286.

Barros, M. J., & **Weiss, C.** (2006). Maximum entropy motivated grapheme-to-phoneme, stress and syllable boundary prediction for Portuguese text-to-speech. *IV Jornadas en Tecnologías del Habla*, 177–182. Zaragoza, Spain.

Bisol, L. (1992). O acento e o pé métrico binário. [The accent and the binary foot]. *Cadernos de Estudos Lingüísticos*, *22*, 69–80. DOI: https://doi.org/10.20396/cel.v22i0.8636897

Boersma, P., & **Weenink, D.** (2019). Praat: Doing phonetics by computer [Computer Program]. Version 6.0.50. http://www.praat.org/

Broselow, E., Chen, S., & **Huffman, M.** 1997. Syllable weight: Convergence of phonology and phonetics. *Phonology, 14*(1). 47–82. DOI: https://doi.org/10.1017/S095267579700331X

Câmara, J. M. (1970). *Estrutura da língua portuguesa*. [Structure of the Portuguese language]. Petrópolis, Brazil: Editora Vozes.

Chen, K., Hasegawa-Johnson, M., & **Cohen, A.** (2004, May). An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-

prosodic model. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, 509. IEEE. DOI: https://doi.org/10.1109/ICASSP.2004.1326034

Cicchetti, D. V., & **Sparrow, S. A.** (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*(2), 127–137.

Collischonn, G. (1994). Acento secundário em português. [Secondary Stress in Portuguese]. *Letras de Hoje*, *29*(4), 43–53.

Correia, M., Ashby, S., & **Janssen, M.** (2019). Dicionário fonético do portal da língua portuguesa. [Phonetic Dictionary of the Portuguese Language]. Last Accessed: 2020-11-20. http://www.portaldalinguaportuguesa.org/index.php?action=fonetica&act=list

Darch, J., Milner, B., Almajai, I., & **Vaseghi, S.** (2007, April). An investigation into the correlation and prediction of acoustic speech features from MFCC vectors. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, *4*, 465–468. DOI: https://doi.org/10.1109/ICASSP.2007.366950

Darch, J., Milner, B., Shao, X., Vaseghi, S., & **Yang, Q.** (2005, March). Predicting formant frequencies from MFCC vectors [speech recognition applications]. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing-ICASSP'05*, *1*, 941–944. DOI: https://doi.org/10.1109/ICASSP.2005.1415270

Darch, J., Milner, B., & **Vaseghi, S.** (2006). MAP prediction of formant frequencies and voicing class from MFCC vectors in noise. *Speech Communication*, *48*(11), 1556–1572. DOI: https://doi.org/10.1016/j.specom.2006.06.001

de Moraes, J. A. (2003). Secondary Stress in Brazilian Portuguese Perceptual and Acoustical Evidence. In *ICPhS-15*, 2063–2066.

de Moraes, J. A. (2007). *Intonational Phonology of Brazilian Portuguese*. Retrieved August 13, 2018, from https://linguistics.ucla.edu/people/jun/Workshop2007ICPhS/Moraes-BP.pdf

Deshmukh, O. D., & **Verma, A.** (2009). Nucleus-level clustering for word-independent syllable stress classification. *Speech Communication*, *51*(12), 1224–1233. DOI: https://doi.org/10.1016/j.specom.2009.06.006

ETSI, ES. (2003). *201 108 v1.1.3: Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithm; Compression Algorithms. ETSI standard.* Retrieved September 19, 2020, from: https://www.etsi.org/deliver/etsi_es/201100_201199/201108/01.01.03_60/es_201108v010103p.pdf

Ferrer, L., Bratt, H., Richey, C., Franco, H., Abrash, V., & **Precoda, K.** (2015). Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. *Speech Communication*, *69*, 31–45. DOI: https://doi.org/10.1016/j.specom.2015.02.002

Fox, M. A. (2000). Syllable-final/s/lenition in the LDC's CallHome Spanish Corpus. In *Sixth International Conference on Spoken Language Processing–ICSLP-2000, 1*, 556–559.

Gahl, S., Yao, Y., & **Johnson, K.** (2012). Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language*, *66*(4), 789–806. DOI: https://doi.org/10.1016/j.jml.2011.11.006

Garcia, G. D. (2014). Portuguese Stress Lexicon. Available (February 2017) at http://guilhermegarcia.github.io/psl

Garcia, G. D. (2017). Weight gradience and stress in Portuguese. *Phonology*, *34*(1), 41–79. DOI: https://doi.org/10.1017/S0952675717000033

Garcia, G. D. (2019). When lexical statistics and the grammar conflict: Learning and repairing weight effects on stress. *Language*, *95*(4), 612–641. DOI: https://doi.org/10.1353/lan.2019.0068

Gordon, M. (2002). A Phonetically Driven Account of Syllable Weight. *Language*, *78*(1), 51–80. Retrieved February 23, 2021, from http://www.jstor.org/stable/3086645. DOI: https://doi.org/10.1353/lan.2002.0020

Gordon, M. (2006). *Syllable weight: Phonetics, phonology, typology*. Routledge. DOI: https://doi.org/10.4324/9780203944028

Gordon, M., & **Roettger, T.** (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, *3*(1). DOI: https://doi.org/10.1515/lingvan-2017-0007

Heldner, M. (2001). Spectral emphasis as an additional source of information in accent detection. In *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*.

Hermans, B., & **Wetzels, W. L.** (2012). Productive and unproductive stress patterns in Brazilian Portuguese. *Letras & Letras*, *28*(1). 77–114.

Houaiss, A., Villar, M. S., & **Franco, F. M. M.** (2001). *Dicionário eletrônico Houaiss da língua portuguesa*. [Houaiss Electronic Dictionary of the Portuguese Language]. Rio de Janeiro: Objetiva.

Hyman, L. M. (2006). Word-prosodic typology. *Phonology*, *23*(2), 225–257. DOI: https://doi.org/10.1017/S0952675706000893

Lai, M., Chen, Y., Chu, M., Zhao, Y., & **Hu, F.** (2006, May). A hierarchical approach to automatic stress detection in English sentences. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, *1*, 1. DOI: https://doi.org/10.1109/ICASSP.2006.1660130

Landis, J. R., & **Koch, G. G.** (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. DOI: https://doi.org/10.2307/2529310

Lee, S. H. (1997). O acento primário do português. [Primary stress in Portuguese]. *Revista de Estudos da Linguagem*, *6*(2), 5–30. DOI: https://doi.org/10.1109/ICASSP.2006.1660130

Lee, S. H. (2007). O acento primário no português: Uma análise unificada na Teoria da Otimalidade. [Primary stress in Portuguese: A Unified Optimality Theory analysis]. In G. A. D. Araújo, (Org.), *O acento em português: Abordagens fonológicas* (pp. 120–143). São Paulo: Parábola.

Li, K., Qian, X., Kang, S., & **Meng, H.** (2013). Lexical stress detection for L2 English speech using deep belief networks. In *Interspeech*–2013, 1811–1815.

Magalhães, J. (2016). Main Stress and Secondary Stress in Brazilian and European Portuguese. In W. L. Wetzels, J. Costa & S. Menuzzi (Eds.), *The Handbook of Portuguese Linguistics* (pp. 107–124). Oxford: Wiley-Blackwell. DOI: https://doi.org/10.1002/9781118791844.ch7

Major, R. C. (1985). Stress and rhythm in Brazilian Portuguese. *Language*, 61(2), 259–282. DOI: https://doi.org/10.2307/414145

Massini, G. (1991). A duração no estudo do acento e do ritmo do português. [Duration in the study of accent and rhythm in Portuguese]. (Unpublished master's thesis). Campinas, SP: Universidade Estadual de Campinas, Instituto de Estudos da Linguagem. Available at: <http://www.repositorio.unicamp.br/handle/REPOSIP/270351>

Massini-Cagliari, G. (1995). Cantigas de amigo: Do ritmo poético ao linguístico. Um estudo do percurso histórico da acentuação em português. [Songs of a friend: From poetic to linguistic rhythm. A study of the diachronic path of accentuation in Portuguese]. (Unpublished doctoral dissertation). Campinas, SP: Instituto de Estudos da Linguagem da Universidade Estadual de Campinas.

Morais-Barbosa, J. (1994). *Introdução ao estudo da fonologia e morfologia do português*. [Introduction to the study of Portuguese phonology and morphology]. Coimbra: Livraria Almedina.

Morgan, J., Ackerlind, S., & **Packer, S.** (2008, May). West Point Brazilian Portuguese Speech LDC2008s04. *Linguistic Data Consortium*. DOI: https://doi.org/10.35111/yjwc-zx48

Pereira, M. I. (2007). Acento latino e acento em português: Que parentesco. [Accent in Latin and accent in Portuguese: What relationship]. In G. A. D. Araújo, (Org.), *O acento em português: Abordagens fonológicas* (pp. 61–83). São Paulo: Parábola.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & **Vesel, K.** (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Sluijter, A. M., & **Van Heuven, V. J.** (1996). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical Society of America*, *100*(4), 2471–2485. DOI: https://doi.org/10.1121/1.417955

Van der Hulst, H. (Ed.). (2014). *Word stress: Theoretical and typological issues*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139600408

Viaro, M. E., & **Guimarães-Filho, Z. O.** (2007). Análise quantitativa da frequência dos fonemas e estruturas silábicas portuguesas. [Quantitative analysis of the frequency of Portuguese phonemes and syllable structures]. *Estudos Linguísticos*, *36*(1), 27–36.

Wetzels, L. (2007). Primary word stress in Brazilian Portuguese and the weight parameter. *Journal of Portuguese Linguistics*, *6*(1), 9–58. DOI: https://doi.org/10.5334/jpl.144

Yuan, J., & **Liberman, M.** (2009). Investigating /l/variation in English through forced alignment. In *Tenth Annual Conference of the International Speech Communication Association–INTERSPEECH2009*, 2215–2218.

Zheng, F., & **Zhang, G.** (2000). Integrating the energy information into MFCC. In *Sixth International Conference on Spoken Language Processing– ICSLP-2000*, *1*, 389–392.

How to cite this article: Harmath-de Lemos, S. (2021). Detecting word-level stress in continuous speech: A case study of Brazilian Portuguese. *Journal of Portuguese Linguistics,* 20: 3, pp. 1–43. DOI: https://doi.org/10.5334/jpl.238

Submitted: 16 September 2019 **Accepted:** 28 January 2021 **Published:** 19 April 2021