9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

What Automatic Speech Recognition Can Tell Us About Stress and Stress Shift in Continuous Speech

Simone Harmath-de Lemos

Cornell University, USA
shd57@cornell.edu

Abstract

I examine lexical stress and stress shift in contexts of stress clash in Brazilian Portuguese (BP) continuous speech data. I start by investigating whether an automatic speech recognition (ASR) toolkit can detect lexical stress using spectral information, as represented by the Mel Frequency Cepstral Coefficients (MFCCs) of stressed and unstressed vowels. The ASR toolkit was trained using a phonetic dictionary where each entry was labeled for primary stress, a list of phones, transcripts, and a language model (LM). The output acoustic model was then used in two test scenarios, where the task of choosing the stressed vowel in a word token was increasingly complex. Results achieved an overall accuracy rate of 92.57% and 80.97% respectively. To investigate stress shift, I use speech data from a production study recorded in Brazil. In the study, speakers where asked to utter syntactically ambiguous sentences using prosody that would cue for one of two possible meanings (and structures). Stress clash would (potentially) be resolved by means of stress shift in one of the structures. Preliminary results showed apparent stress shift in roughly 20% of the contexts identified by a human referee as having the syntactic structure where stress shift would occur.

Index Terms: lexical stress, stress shift, continuous speech, syntactic ambiguity, Brazilian Portuguese, MFCCs, Kaldi.

1. Introduction

Lexical stress is a structural property of language, generally understood as the relative prominence of a syllable within a word. Crosslinguistically, lexical stress is important insofar as it may cue speakers to a number of phenomena: it can indicate word contrast, as in [ɪm.ˈpɔrt] (VERB) vs. [ˈɪm.pɔrt] (NOUN) and it may trigger segment-level phonological rules, like aspiration of voiceless stops in the first position of stressed syllables in English, e.g., [p^hɑt]. Lexical stress may also trigger phrasal rules, such as stress shift in contexts of stress clash—as in "Cor**nell**" and "**stu**dents", but "**Cor**nell **stu**dents"—as well as attract nuclear intonational tones [5].

Beyond theoretical linguistics, lexical stress is relevant to a number of fields, such as second language acquisition (SLA), computer aided language learning (CALL), and automatic speech recognition (ASR). In SLA for example, studies like [1] have shown that deviance in prosody by non-native speakers of English has a stronger effect in intelligibility rating than deviance in segmental features. Ferrer et al. [3], implemented a lexical stress classifier to be used by CALL systems to help English learners to produce closer-to-target stress, so to increase intelligibility rating of those coming from native languages where stress systems were fairly different from that of the target language.

In terms of phonetic implementation, the acoustic correlates most commonly associated to lexical stress are vowel duration, intensity, pitch (F0), and spectral features, ([4] inter alia), and languages may differ as to which acoustic correlates, or combinations thereof are used. Factors such as intra- and inter-speaker variability in production, the number of correlates involved and the possible interactions between them, as well as methodological questions about measuring—what criteria should be used when measuring different correlates: duration of the stressed vowel, of the syllable coda, or of the entire stressed syllable? Peak intensity or relative intensity?—make lexical stress an elaborate, yet important property to be studied in continuous speech data, warranting the investigation of ancillary methods.

The present study has two parts: experiment 1 investigates whether the automatic speech recognition (ASR) toolkit Kaldi [9] can detect primary lexical stress based on spectral information, as represented by Mel Frequency Cepstral Coefficients (MFCCs) of stressed and unstressed vowels. In a second experiment, using the model generated during the training pass of the first experiment, Kaldi looks for instances of stress shift in contexts of stress clash (e.g., the words "Cor**nell**" and "**stu**dents" would become "**Cor**nell **Stu**dents") in a corpus recorded during a production experiment in Brazil. Both experiments were conducted using speech corpora of Brazilian Portuguese, a language in which duration is reported to be the most consistent acoustic correlate ([7], [2] inter alia), so to test whether a method that uses spectral information to detect lexical stress is suitable to be used with languages where spectral information is not a robust correlate of lexical stress.

In the sections that follow, I briefly review stress assignment in Brazilian Portuguese (BP), describe the system components, the methodology and the two experiments, following with a discussion and final considerations.

2. Background

2.1. Brazilian Portuguese

It is generally accepted that Brazilian Portuguese (BP) has at least two degrees of stress[1]. Primary lexical stress, henceforth primary stress, may fall on one of the three right-most syllables of a word (ultimate, penultimate, and antepenultimate stress), as shown below:

maçã	[ma.ˈsɐ̃]	'apple'
banana	[ba.ˈnɐ.na]	'banana'
pêssego	[ˈpe.se.go]	'peach'

With respect to primary stress assignment, [10] proposes that Brazilian Portuguese is a mixed system, where syllable weight is relevant to primary stress assignment in non-verbal lexical categories. [10] also notes that the unmarked primary stress pattern in non-verbs is the penultimate syllable, but that

[1] Secondary stress is not addressed in the present study.

10.21437/SpeechProsody.2018-199

primary stress falls on the ultimate syllable if it is heavy[2] [3]. Antepenultimate stressed syllables are exceptions in the language, and [10] presents an OT account for primary stress in non-verbs in accordance with these observations. In the verbal paradigm, primary stress is assigned as a function of tense.

2.2. ASR toolkit - Kaldi

Kaldi [9] is an open-source ASR toolkit, licensed under Apache v 2.0. It is written in C++, has code-level integration with finite state transducers (FST), and uses OpenFst as a library. Kaldi is language-independent and speaker-independent, making it suitable to be used in projects that aim to investigate phonological and prosodic aspects of more than one language. It is flexible with respect to extracting acoustic features from the speech signal (Mel Frequency Cepstral Coefficients (MFCCs) or Perceptual Linear Predictive (PLP) features), and with respect to the language model it can use (ARPA or n-gram models). Kaldi can also use a number of methods to model acoustic features, like Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) and Deep Neural Network-Gaussian Mixture Model (DNN-GMM). Kaldi can model phone segments as monophones (independent from phonetic context), or as triphones (context-dependent phones, where context comprises one phone to the left and one phone to the right of the phone under scrutiny). Kaldi builds a decision tree based on the (mono- or tri-)phone models generated during training.

2.3. The West Point Corpus (LDC2008S04)

The West Point Brazilian Portuguese speech corpus [8] contains digital recordings designed and collected by the Department of Foreign Languages (DFL) and the Center for Technology Enhanced Language Learning (CTELL).The data was collected at the Brazilian Military Academy in Brasília, in 1999.

Sixty (60) female and sixty-eight (68) male native and non-native speakers were recorded while reading a script that contained two hundred and ninety six isolated sentences (not in context). In the present study, only data generated by monolingual speakers was used. A breakdown of the corpus is presented in Tables 1-3.

Table 1: *West Point Corpus - Overall*

Summary	
Utterances in corpus	7846
Word tokens in corpus	40007
Phone tokens in corpus	155409
Vowel phone tokens in corpus	78494
Male speakers	53
Female speakers	46

3. Methodology

3.1. Settings - Kaldi

Kaldi used MFCCs to extract acoustic features from speech signal in both experiments. Acoustic data was modeled using the HMM-GMM algorithm. Alignments were generated using a monophone model, because the idea is to test whether the spectral features of stressed and unstressed vowels themselves differ systematically, without additional information from phonetic context in which the vowel is found. Kaldi used a unigram[4] as a language model, because the focus of the study is acoustic data and acoustic modeling, as opposed to speech decoding.

[2] In [10], heavy-syllables are Vl, Vr, Vs, oral diphthongs, nasal diphthongs, and nasal vowels.

[3] There is a fair amount of exceptions for this rule, which are accounted by in [6], as being the result of unproductive stress rules in BP, related to borrowings from Indigenous languages, Bantu, Yoruba languages, French and English.

Table 2: *West Point Corpus - Word Tokens x Syll. in Word*

Syllables	Word Tokens	Syllables	Word Tokens
1	15341	4	2408
2	15385	5	742
3	5935	6	196
Total Tokens			**40007**

Table 3: *West Point Corpus - Word Tokens x Stress Locus*

Stress	Word Tokens
antepenult	366
penult	17889
ult	21752
Total Tokens	**40007**

3.2. Detecting lexical stress in continuous speech

In this experiment I examined whether Kaldi can systematically detect placement of primary lexical stress in Brazilian Portuguese. The experiment consisted of one training pass and two tests. For the training pass, in addition to the speech data and the unigram mentioned above, Kaldi was given list of phones, the transcripts of the corpus, and a dictionary contianing the phonetic transcriptions for each word in the corpus, with multiple pronunciations per word, where appropriate. Each entry in the dictionary was labelled for stress at the target position only, with the exception of monosyllabic function words, which were listed as optionally unstressed.

The acoustic model generated during the training pass was used to run two tests: one where Kaldi could choose any of the vowels in a given word token as the stressed vowel. Call this test $n = syll$, where n is the number of options Kaldi can choose from and $syll$ is the number of syllables in the word. Kaldi was given the same list of phones and unigram LM used during the training pass, but all words in the transcripts were especially marked to map to entries in the phonetic dictionary in which each word had as many options for stress locus as the number of vowels in it, hence giving Kaldi $n = syll$ options to choose stress locus from. This test is illustrated in Figure 1.

In the second test, call it $n = 2^{\text{syll}}$, Kaldi also used the configuration described in section 3.1 above, the alignment model generated during the training pass, the list of phones and the unigram LM used in the training pass, but all words in the transcripts were especially marked to map to entries in the phonetic dictionary where for each word, stress could fall on any of its vowels or on any combination thereof, giving Kaldi 2^{syll}(n is the number of options and $syll$ is the number of syllables in the word) options to choose stress locus from. This is illustrated in Figure 2.

[4] The unigram is a courtesy of Mats Rooth

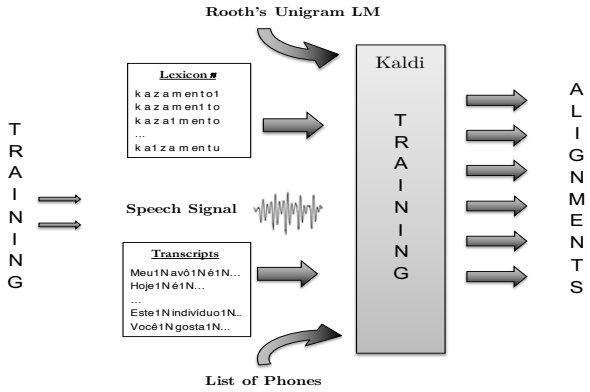

Figure 1: *Test 1 ($n = syll$)*

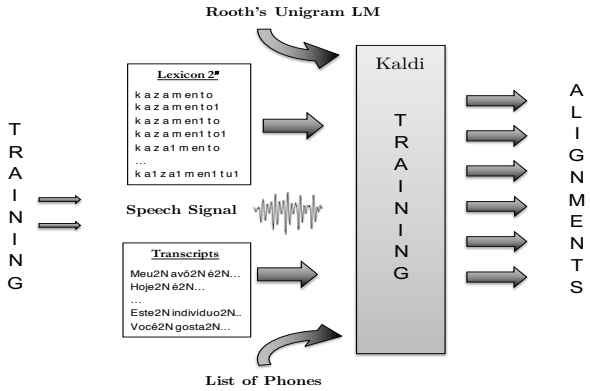

Figure 2: *Test 2 ($n = 2^{syll}$)*

4. Stress shift in continuous speech?

The second experiment in this study examined (potential) stress alternations in Brazilian Portuguese: in [11], it is reported that, in cases of structural ambiguity where the words involved are in a context of stress clash, if such clash is resolved by means of stress shift, the cue is that a single phonological phrase (PhP) is formed, and the interpretation will be that the second word involved in the clash modifies the word that immediately precedes it. If, however, clash is resolved by the insertion of pauses and/or the addition of a tone accent to each of the words involved, two distinct phonological phrases are formed, and the interpretation will be that the second word involved in the clash modifies a word other than the one that immediately precedes it. The expectation in this study would be that stress shift, if it happens to resolve clash in the context of dependency just described, would not be categorical, based on the premise that speakers have to be able to understand what the dependency is in utterances where the words involved were never in a context of stress clash to begin with. Nevertheless, it may be that stress shift happens as a by-product of phonological phrasing.

The speech corpus used in this experiment was recorded during a production study in Brazil, where four (4) participants, native speakers of BP, were presented with fifty (50) instances of structurally ambiguous sentences, along with short contextualizing paragraphs, which would cue speakers to utter a given sentence using prosody corresponding to one of the two possible structures (and meanings). A sample sentence is shown in Figure 3. A linguistically informed human referee judged each of the utterances produced in the experiment, classifying them for meaning, a predictor of what would be the expected course for stress clash resolution—stress shift, in case the second word involved in the clash modified the immediately preceding word, or the insertion of a pause and/or addition of extra pitch accent(s) otherwise.

Kaldi was set up with the configuration described in 3.1 above, and was given the same list of phones, phonetic dictionary, and unigram LM used in experiment 1, as well as the acoustic model generated in the training pass described in 3.2. The transcripts Kaldi was given in this experiment were marked

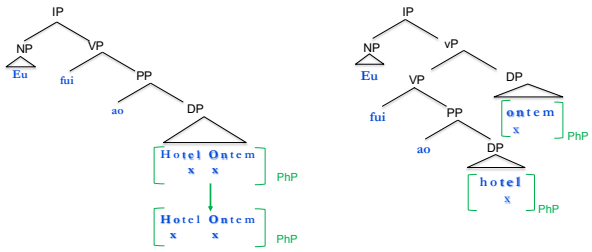

Figure 3: *Experiment 2 – Stress Alternations*

as in test 2 of experiment 1, mapping to entries in the phonetic dictionary in which each word had as many options for stress locus as the number of vowels in it, or any combination thereof, giving Kaldi ($n = 2^{syll}$) options to choose stress locus from.

5. Results

Word tokens, represented by their component phones, were extracted from the forced alignments using Kaldi's *show-alignments* script, and were then processed into text format. For each word token aligned, information was added about number of syllables in the word, expected locus of stress, observed locus of stress, utterance ID, and speaker ID. Raw data files were imported and analyzed in MS Excel, v.15.41.

Results for monosyllabic function words were not included in the counts because Kaldi was trained using a phonetic dictionary in which these words were optionally stressed, so a match/mismatch classification would not make sense here. Results for the remainder word tokens were tabulated under four different categories:

Perfect Matches (M): observed locus of stress matched expected AND nothing else.

Partial Matches (PM): observed locus of stress matched expected, AND any vowel to the left was also deemed stressed.

Unstressed (U): none of the vowels in the word were detected as stressed.

Mismatches (MM): observed locus of stress does not match expected, OR observed and expected loci match, but any vowel positioned to the right of the expected locus of stress was marked as stressed.

Accuracy rate for both tests was calculated as shown in Equation 1 and Partial Match rate (PMR) was calculated as shown in Equation 2.

$$Accuracy = \frac{M}{WordTokens}100 \quad (1)$$

$$PMR = \frac{PM}{WordTokens}100 \quad (2)$$

Table 4: *Accuracy in Stress Detection, Test 1($n = syll$)*

Sylls	**Accuracy** (%)	**Sylls**	**Accuracy** (%)
1	100	4	86.34
2	91.61	5	92.99
3	88.58	6	95.92
Overall			**92.57%**

Table 5: *Accuracy in Stress Detection, Test 2 ($n = 2^{syll}$)*

Sylls	**Accuracy** (%)	**Sylls**	**Accuracy** (%)
1	87.70	4	44.05
2	71.68	5	54.23
3	56.99	6	44.39
Overall			**67.43%**

Table 6: *PMR in Stress Detection, Test 2 ($n = 2^{syll}$)*

Syll.	**PMR** (%)	**Syll.**	**PMR** (%)
1	0.00	4	43.35
2	5.83	5	35.84
3	26.44	6	54.08
Total P. M.			**13.54%**

Table 7: *Stress Shift, Test 2 ($n = 2^{syll}$)*

	No Shift (%)	**Shift** (%)	**Multiple Stress**(%)
Not Expected	39.47	7.89	52.63
Expected	29.76	17.86	52.38
Total (%)	34.62	12.88	52.51

5.1. Experiment 1 - stress detection

Results for test 1, $n = syll$, are summarized in Table 4, as a function of the number of syllables in a word token.

Results for test 2, $n = 2^{\text{syll}}$, are summarized in Table 5, also as a function of the number of syllables in a word. Table 6 shows results for Partial Match rate.

5.2. Experiment 2 - stress shift?

Analysis of the stress alternation experiment included data from three out of the four participants who took part in the original experiment. If an instance of disfluency was found in any of the words involved in the context where stress clash was happening, the utterance was not tabulated in the results. Six (6) utterances in a total of 128 were found to carry instances of disfluency, making for an adjusted total of 122 utterances. Table 7 shows a summary for the experiment.

6. Discussion

In the Stress Detection experiment, results for test 1 ($n = syll$), summarized in Table 4, indicate that what Kaldi is detecting is consistent with primary lexical stress in Brazilian Portuguese. Results for test 2 ($n = 2^{\text{syll}}$), summarized in Table 5 and Table 6, show that the results remain consistent in light of more complex questions. Results shown in Table 6 suggest that especially in words that are 3-syllable or longer, some pretonic vowels are more similar to stressed vowels than they are to unstressed vowels, hinting to the presence of secondary lexical stress. This question may be at least partially resolved if secondary stress is modeled in the phonetic dictionary.

A visual inspection of the results showed that a number of the mismatches occurred in contexts where expected stress fell in the ultimate syllable, and the following word in the utterance started with a phone of the same vowel quality (hinting to instances of Sandhi). A way around this would be to model the reduced vowels of Brazilian Portuguese in the list of phones, and to add pronunciation entries containing the reduced vowels as predicted in the literature. Visual inspection also indicated that disyllabic function words might be another source of mismatches, warranting further investigation of stress in function words, monosyllabic or not. An aspect (number counts not shown here) that requires further investigation is the number of instances of words that Kaldi classified as bearing no stress (tabulated as $Unstressed$), which roughly totaled 11% of the overall number of word tokens in experiment 1, test 2 ($n = 2^{\text{syll}}$). Word tokens classified as $Unstressed$ are especially interesting in words that are 4 syllables or longer, for which the sum of $Accuracy$ and PMR was higher than 87%. Measuring duration of the vowel token where primary stress was expected to fall can shed additional light on the acoustic realization of primary stress in these word tokens.

Two additional aspects (numbers not shown here) are worthy of investigation: a possible effect of word length in syllables—Kaldi is more accurate to detect stress in longer words—and a possible effect of locus of target stress—the farthest expected stress locus is from the right boundary of the word, the more accurate Kaldi gets in detecting stress.

Data shown in Table 4, which summarizes the stress detection experiment, appears to confirm the prediction that the resolution of stress clash by means of stress shift when the second word involved in the clash modifies the first is not a categorical feature. It also appears to be the case that stress shift happens for at least some speakers. However, because the high number of tokens involved in which Kaldi detected multiple vowels as stressed, further acoustic measurements performed on the vowel tokens involved in the instances classified as stress shift would add information to these results.

7. Conclusion and future research

The results obtained through the methodology herein described, using Kaldi and spectral information, as represented by MFCCs, to detect primary lexical stress in continuous speech—in a language in which this feature is not reported as a robust correlate for stress—are encouraging. Future studies could model secondary stress and reduced vowels of Brazilian Portuguese, both as a means of increasing accuracy in detection, and as a means of investigating secondary stress, and its interactions with primary stress in continuous speech. The question of stress alternation needs further examination: results indicate that, even though stress shift in the context herein investigated may not be categorical, some speakers appear to do it in the phonological context described in [11]. Importantly, future studies should use data from typologically different languages, as it would be interesting to learn if comparable results are found.

8. References

[1] J. Anderson-Hsieh, R. Johnson, and K. Koehler, “The relationship between native speaker judgements of non-native pronunciation and deviance in segmentals, prosody and syllable structure," in *Language Learning*, 42, pp. 529–555.

[2] P. A. Barbosa, A. Eriksson, and J. Åkesson, “On the robustness of some acoustic parameters for signaling word stress across styles in Brazilian Portuguese," in *INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association, August 25–29, Lyon, France, Proceedings* 2013, pp. 282–286.

[3] L. Ferrer, et al. “Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," in *Speech Communication*, Volume 69, May 2015, pp. 31–45, ISSN 0167-6393, http://dx.doi.org/10.1016/j.specom.

[4] M. Gordon, and T. Roettger. “Acoustic correlates of word stress: A cross-linguistic survey," in *Linguistics Vanguard*, 3.1 (2017).

[5] B. Hayes, “Metrical stress theory: principles and case studies,” Chicago: University of Chicago Press,1995.

[6] B. Hermans, and W. L. Wetzels, “Productive and unproductive stress patterns in Brazilian Portuguese," in *Letras & Letras*, 28(1).

[7] R. C. Major, “Stress and Rhythm in Brazilian Portuguese," in *Language*, 61(2), pp. 259–282. http://doi.org/10.2307/414145.

[8] J. Morgan, S. Ackerlind, and S. Packer, “West Point Brazilian Portuguese Speech LDC2008S04," Web Download. Philadelphia: Linguistic Data Consortium.

[9] D. Povey, et al. “The subspace Gaussian mixture model?A structured model for speech recognition," in *Computer Speech & Language*, Volume 25, Issue 2, pp. 404–439, ISSN 0885-2308, http://dx.doi.org/10.1016/j.csl.2010.06.003.

[10] W. L. Wetzels, “Primary word stress in Brazilian Portuguese and the weight parameter," in *Journal of Portuguese Linguistics*, 5/6, pp. 9–58.

[11] A. P. Gravina, and F. Fernandes-Svartman, “Interface sintaxe-fonologia: desambiguação pela estrutura prosódica no português brasileiro," in *Alfa: Revista de Linguística*, São José do Rio Preto. 57(2). pp. 639–668. Retrieved April 06, 2015.