The Importance of Word-final Vowel Duration for Non-native Portuguese Speaker Identification by Means of Support Vector Machines

A importância de duração de vogais em posição final da palavra para identificação de falantes não nativos de português usando Máquinas de Vetores Suporte

Suzanne Franks*
University of Georgia
Athens - Georgia / USA

Rommel Barbosa**
Universidade Federal de Goiás (UFG)
Goiânia- Goiás / Brasil

ABSTRACT: This article studies the acoustic characteristics of some oral vowels in tonic syllables of Brazilian Portuguese (BP) and which acoustic features are important for classifying native versus non-native speakers of BP. We recorded native and non-native speakers of BP for the purpose of the acoustic analysis of the vowels [a], [i], and [u] in tonic syllables. We analyzed the acoustic parameters of each segment using the Support Vector Machines algorithm to identify to which group, native or non-native, a new speaker belongs. When all of the variables were considered, a precision of 91% was obtained. The two most important acoustic cues to determine if a speaker is native or non-native were the durations of [i] and [u] in a word-final position. These findings can contribute to BP speaker identification as well as to the teaching of the pronunciation of Portuguese as a foreign language.

KEYWORDS: phonetics, acoustic phonetics, second language acquisition, Brazilian Portuguese, vowels, Support Vector Machines.

* scfweb@uga.edu

** rommel@inf.ufg.br

RESUMO: Este artigo estuda as características acústicas de algumas vogais orais em sílabas tônicas do português brasileiro (PB) e quais aspectos acústicos são importantes para classificar falantes nativos versus não nativos de PB. Gravamos falantes nativos e não nativos de PB com o propósito de realizar uma análise acústica das vogais [a], [i] e [u] em sílabas tônicas. Analisamos parâmetros acústicos de cada segmento usando o algoritmo Máquinas de Vetores Suporte para identificar a qual grupo, nativo ou não nativo, um novo falante pertence. Ao considerar todas as variáveis, obtemos uma precisão de 91%. Os dois índices acústicos mais importantes para determinar se um falante é nativo ou não foram a duração de [i] e [u] na posição final da palavra. Os resultados obtidos podem contribuir à identificação de falantes de PB e ao ensino de pronúncia do português como língua estrangeira.

PALAVRAS-CHAVE: fonética, fonética acústica, aquisição de segunda língua, Português brasileiro, vogais, Máquinas de Vetores Suporte.

1. Introduction

A speaker of a given language can easily recognize a foreign accent from a speech sample. On the other hand, a person who is not familiar with a particular language may not be able to distinguish between a native or non-native speaker of that language. Acoustic analyses of speech samples make identification of the speaker possible, which is particularly valuable when a person who knows a language is not available to perform the task. Furthermore, an acoustic analysis helps identify specific characteristics about individual sounds that distinguish a native speaker from a non-native speaker. Brosseau-Lapré et al. (2013) examine the influence of input variability on the learning process of adults learning French. Due to the highly variable nature of speech, a simple one-on-one comparison of corresponding elements (words, syllables, or segments) by native and non-native speakers does not enable one to isolate the specific sources of variability in perception that may cause deviations from native-like production. Furthermore, the relative importance of different types of variability to language learning, as well as production of non-native speech, is unknown.

In addition to informing foreign language educators, linguistic knowledge regarding particular first language (L1)-second language (L2) pairings can be useful to improve accuracy in voice-controlled information retrieval systems. This study looks at the speech production of a group of native and non-native speakers of Brazilian Portuguese (BP) in order to find which aspects of vowel pronunciation are most important in distinguishing a native from a non-native speaker.

According to the 2010 census, Brazil's population was around 190.8 million (IBGE, 2010). Worldwide, more than two million people speak Portuguese (LEWIS, SIMONS, FENNIG, 2013), of which roughly 80% speak the BP variety. Additionally, courses in Portuguese as a foreign language at American universities have gained popularity in recent years. Phonetic research in BP is still catching up with the language's current level of popularity.

Both English and Portuguese have relatively large vowel inventories; however, the acoustic and perceptual spaces of English vowels differ from the acoustic and perceptual spaces in Portuguese. While the English vowel system has nine monophthongs and six diphthongs (LADEFOGED, 2006), the Portuguese vowel system consists of seven oral monophthongs, five nasal monophthongs, and as many as 23 different diphthongs (SILVA, 1998; WHITLAM, 2011). The current study examines three vowels as produced by native and non-native speakers of BP. Although descriptions of BP's vowel system are available (CALLOU, MORAES, LEITE, 2013; GAMA-ROSSI, 2001; MORAES, CALLOU, LEITE, 1996a; MORAES, CALLOU, LEITE, 1996b; NOBRE, INGEMANN, 1987; RAUBER, 2008; REDENBARGER, 1981), the use of acoustic data to identify speakers as native versus non-native has not been attempted.

In this study, we first present a review of psycholinguistic literature regarding the acquisition of L2 speech. In section two, we move on to explain the materials and methods used for this study's experimental investigation. In section three, we deal with the linguistic theory and machine learning algorithm and calculations relevant to this study. Section four covers the results and discussion regarding the acoustic measures of vowel sounds and support vector machines (SVM) algorithm. Finally, conclusions are drawn about the importance of the duration of the word-final vowels [i] and [u] in Portuguese speaker discrimination.

This study seeks to answer the following research question: What are the acoustic features of the three BP vowels [a], [i], and [u] in syllables bearing primary stress that distinguish a native speaker from a non-native speaker?

1.1. Literature Review

Psycholinguistic studies have addressed the influence of L1 phonemic systems on an L2. Many studies focus on vowel inventories (BOHN, FLEGE, 1992; PISKE et. al, 2002); however, other studies examine consonant phonetic features, such as voice onset time (ECKMAN, 1987; FLEGE, 1987), or suprasegmentals, such as duration (SIMÕES, 1991). For this line of research, SVMs are powerful classification algorithms that have been used successfully in many different areas, including language and speech recognition problems (CAMPBELL et al., 2006).

In this section, we will present a few of the more influential psycholinguistic theories that address non-native accents. For example, the interlanguage structural conformity hypothesis, developed by Eckman et al. (1989), emphasized interlanguages as languages that "can and must be described on their own grounds" (BLEY-VROMAN, 1983 apud ECKMAN, 1991). Referring to the interlanguage structural conformity hypothesis, Eckman (1991) stated that the universal generalizations that are true for native languages are also true for interlanguages. One such characteristic of an interlanguage is that at any point in time a non-native speaker of a language may make use of a phonological system that is neither identical to the target language (TL), nor to the speaker's native language.

The speech learning model (SLM), proposed by Flege (1995), presents another perspective on L2 phonological acquisition and is comprised of seven hypotheses. Flege's SLM hypotheses 2, 3, and 7 are most relevant to our study. Hypothesis 2 reads: "A new phonetic category can be established for an L2 sound that differs phonetically from the closest L1 sound if bilinguals discern at least some of the phonetic differences between the L1 and L2 sounds" (p. 239). Hypothesis 3, by contrast, proposes the following: "The greater the perceived phonetic dissimilarity between an L2 sound and the closest L1 sound, the more likely it is that phonetic differences between the sounds will be discerned" (p. 239).

Phonetic differences are also addressed in SLMs Hypothesis 7, which states: "The production of a sound eventually corresponds to the properties represented in its phonetic category representation" (FLEGE, 1995, p. 239). Because vowel sounds have fairly flexible acoustic targets in any given language, the acquisition of new vowel sounds is more than acquiring new mental representations of appropriate categorical distinctions of TL phonemes. Rather, mental representations of interlanguage phonetic

categories of vowel sounds evolve during the acquisition process to enable the language learner to perceive and produce sounds that are closer to the TL vowel sounds. Often the acoustical measures taken of an interlanguage reveal the production of a non-native speaker to be somewhere between the L1 and the TL.

2. Materials and methods

The current study analyses the production of 11 native speakers (five male and six female) of BP and 22 students (eight male and fourteen female) of BP as a foreign language. All non-native speakers were residents of the state of Georgia (USA) enrolled in university-level Portuguese courses, ranging from second to sixth semester. Seven of the native speakers were from the Southeastern region of Brazil (mostly the state of São Paulo), two were from the South (Rio Grande do Sul), and two were from Northeast (Ceará). The mean age of native speaker participants was 31 years (range: 24-47 years), while the mean age of non-native speaker participants was 22 years (range: 18-33 years). Native speakers were recorded reading 19 sets of randomized Portuguese sentences at a normal speaking pace. Only six of the sentence sets were used for this study; the other sets were used as distracters. Non-native speakers were recorded doing an elicited imitation task of repeating randomized sentences previously recorded by two of the native speaker participants (one male and one female). Three seconds of silence followed each sentence, enough time for participants to repeat what they heard.

A direct repetition technique (SNOW, HOEFNAGEL-HÖHLE, 1977; MARKHAM, 1977) was selected for this study because the alternative elicitation methods of reading or spontaneous speech present more drawbacks than advantages for inexperienced non-native speakers. In a reading task, varying reading abilities may interfere with pronunciation. Furthermore, in spontaneous speech, morphosyntactic or lexical errors can interfere in the non-native speakers' production (PATKOWSKI, 1990; PISKE, MACKAY, FLEGE, 2001). In the present study, all recordings took place in a sound attenuated room in a phonetics laboratory at the University of Georgia. The recordings of native-speaker and non-native-speaker participants were analyzed using PRAAT (BOERSMA, WEENINK, 2012), and measurements were automated using a script for PRAAT (HIRST, 2012).

The token words used in the production experiment were designed to elicit target vowel sounds in BP words in carrier phrases. This study focused on the three oral vowels that correspond to the three corners of a two-dimensional vowel space (obtained from F_1 frequency by F_2 frequency measurements in Hz): the low central vowel [a], the high front vowel [i], and the high back vowel [u]. In first language acquisition research, language-specific corner vowels have been used in experimental research (BOND, PETROSINO, DEAN, 1982; RVACHEW et. al., 2008). An important reason for examining the three corner vowels is to determine the extreme corners of a speaker's acoustic vowel space to discover the limits of a speaker's or a group of speakers' acoustic vowel space. Vowel dispersion is an important part of preserving perceptual contrasts, especially within a language, whether L1 or L2. Baptista (2006) affirms "that listeners perceive each vowel in relation to the speaker's total acoustic vowel space, which they calibrate from the formant frequency patterns in the rest of the ongoing speech" (p. 20). A learner's perception at a systemic level affects how each individual vowel phoneme is produced.

Each token in this study was disyllabic and contained one of the vowels being investigated in a stressed syllable. Tokens were divided evenly between words with penultimate stress and ultimate stress. To facilitate acoustic analysis and to control variation caused by the environment, the vowels under investigation were preceded and/or followed by stops (voiceless) or affricates (voiced or voiceless). Two carrier phrases were used, but only the tokens in the second carrier phrase (second and third token repetitions) were considered in the acoustic analysis. The use of a carrier phrase is common in acoustic-phonetic studies concerned with vowel length, amplitude, and other measures, such as formant frequencies (LIMA-GREGIO et al., 2010; OH, 2008; SHAIMAN, 2001). A set of carrier phrases with a token containing penultimate stress and with the target vowel being surrounded by voiceless stops appears in 1.

(1) Token: ['ka.tɐ]
a. Diga cata, por favor.
'Say cata, please.'
b. Diga cata de novo.
'Say cata again.'
c. Diga cata de novo.
'Say cata again.'

Another sample from the elicited imitation protocol appears in 2. In this case the token carries the ultimate stress, and the target vowel is preceded by a voiceless affricate, or in rare cases the target vowel is preceded by a voiceless stop due to regional dialectal variation. All vowels in tokens with ultimate stress are followed by a voiced affricate [ʤ] or voiced stop [d], depending on each speaker's regional accent.

(2) Token: [ves.'ʧi] ~ [veʃ. 'ʧi] ~ [ves. 'ti]
a. Diga vesti, por favor.
'Say (I) dressed please.'
b. Diga vesti de novo.
'Say (I) dressed again.'
c. Diga vesti de novo.
'Say (I) dressed again.'

The environment surrounding each vowel was controlled in a consonant environment, stress, word length, and placement within a sentence to minimize variation caused by external factors, such as coarticulatory assimilation and prosodic differences that could affect the vowel quality, length, and amplitude.

Target vowels in all recordings were segmented and labeled by hand. Both the waveform and spectrogram were considered when boundaries were placed at the beginning and end of each vowel. That is, boundaries were placed at the nearest zero-crossing on the waveform that agreed with the beginning or end of the vowel formants in the spectrogram.

3. Theory and calculation

Acoustic measures were recorded for the target vowels [a], [i], and [u] in the recordings of both native-speaker and non-native-speaker participants. Six numeric values (variables) were derived from the acoustic signal at the temporal mid-point of each target vowel using a PRAAT script (HIRST, 2012). These variables included duration (in milliseconds), intensity, pitch (f_0 frequency in Hz), F_1 frequency in Hz (inversely related to tongue height), F_2 frequency in Hz (for our purposes, associated with tongue backness), and F_3 frequency in Hz (for our purposes, inversely related to lip rounding). Means of the six numeric values from the second and third repetitions were calculated. For each speaker, there were 36 acoustic numeric values plus a

gender designation, hence 37 variables per speaker. The 36 numeric values consisted of the six values obtained from the temporal mid-point of each target vowel in a stressed position of the disyllabic words. That is, in tokens containing target vowels in which primary stress falls in the penultimate position [cv.cv], the first vowel was analyzed, whereas in tokens with ultimate primary stress [cv.´cv], the second vowel (which is also a vowel in word-final position) was analyzed. The words with penultimate stress were *cata* [´ka.tɐ] 'gather', *tico* [´ti.ku] ~ [´ʧi.ku] 'whit', and *tuco* [´tu.ku] 'worker who removes the earth, in the conservation of the railway bed'. The words with ultimate stress were *está* [es.´ta] ~ [eʃ.´ta] 'to be (3rd sing. pres.)', *vesti* [ves.´ti] ~ [veʃ.´ʧi] 'to dress (1st sing, pret.)', and *tutu* [tu.´tu] 'tutu'.

Since pitch, intensity, and duration, all part of speech prosody, are important sources for accented speech, the f_0 frequency, duration, and intensity of each target vowel were measured (HUANG, JUN, 2011; KOLLY, DELLWO, 2014; MORRISON, 2008; TROFIMOVICH; BAKER, 2006). Vowel quality is another significant factor in accented speech (MORRISON, 2008; PISKE, et al., 2002), which is measured with the frequency bands commonly labeled as F_1, F_2, and F_3.

3.1. Machine learning algorithm used for classification

A classification problem involves identifying to which category – native or non-native speaker – a new person belongs based on a training set of data with variables whose category is already known. The set of variables associated with a person is represented as a vector. Machine learning entails the use of mathematics and computational methods aimed at finding efficient and accurate algorithms for classification.

Learning algorithms for classification have been successfully deployed in a variety of applications, such as: food science (BATISTA et al., 2012), animal science (AGUIAR et al., 2012), speech and language science (CAMPBELL et al., 2006) to name a few. The learning stages of our classification problem are as follows. Starting with an existing collection of labeled examples, we divided the data into a training set and a test set. We used k-fold cross-validation for the model selection and for performance evaluation. In our problem, the original set is partitioned into k equal-size subsets. In k subsets, one subset is set apart for testing, while the others, consisting of k-1 sets, are used for training. This process is repeated k times, with each of the k subsets used exactly once as the test (WITTEN, FRANK,

HALL, 2011). As the number of samples (or participants) in our problem is relatively small, we used the leave-one-out process of cross-validation in which we considered k to be equal to the number of samples. After applying the classification algorithm and analyzing its accuracy using all original given features, we applied the same methods for a reduced number of selected features. This reduced number of features was obtained by a feature selection algorithm. The accuracy and computational complexity of the algorithm can be affected by the total number of variables in a problem. By using a variable selection algorithm, we presuppose that a data set may have variables that do not provide additional information beyond selected features. Here in our work we used the correlation feature selection (CFS) subset algorithm. A search algorithm is used in CFS to evaluate the value of feature subsets. The method of discovery by which CFS measures the suitability of feature subsets considers the usefulness of each variable for finding the class label along with the level of inter-correlation among the variables (HALL, 1998).

The well-known classification algorithm SVM was used in the present study. The software Weka (Waikato Environment for Knowledge Analysis) was used to perform the algorithm. Weka contains a collection of visualization tools and algorithms for data analysis and classification, and is available for free under the GNU General Public License (WITTEN, FRANK, HALL, 2011).

Below, we provide an intuitive explanation of the algorithms used in this study. A more in depth discussion can be found in Schlkopf and Smola (2001).

3.2. Support Vector Machines

SVM is one of the best known and most used classification algorithms. Vapnik introduced SVM in 1963, which was later extended for use in a non-linear case by Doser, Guyon, and Vapnik in 1992. In 1995, Cortes and Vapnik (1995) introduced the use of the soft-margin, which uses slack variables for non-separable linear cases.

SVM is based on a procedure that finds a special type of linear model called the maximum-margin hyperplane. To picture a maximum-margin hyperplane, consider a two-class dataset (in this study, native versus non-native speakers) whose classes are linearly separable, meaning that a hyperplane in the input space classifies all training instances correctly. The maximum-margin hyperplane can also be defined as that which gives the

greatest separation between the classes. The instances that are closest to the maximum-margin hyperplane are called support vectors. Each class has at least one, but often more than one, support vector. What is critical is that the set of support vectors singularly defines the maximum-margin hyperplane for the learning problem (WITTEN, FRANK, HALL, 2011).

The classifier should choose a hyperplane that minimizes errors when classifying new samples. Decision boundaries with greater margins tend to generate fewer errors than those with smaller margins (TAN, STEINBACH, KUMAR, 2006). Therefore, a linear SVM is a classifier that searches for the hyperplane with the widest margin.

Consider the problem of binary classification here (native versus non-native speakers). Each sample can be represented by one n-upla ($\boldsymbol{x}_i$, y_i), $i = 1, \ldots, N$, in which $(x_{i1}, x_{i2}, \ldots x_{id})^t$ corresponds to acoustic features obtained from the i^{th} example and $y_i \in \{-1, 1\}$ (indicating native or non-native speakers).

The decision boundary of the classifier can be expressed in the form:

$$\boldsymbol{w} \,.\, \boldsymbol{x} + b = 0$$

in which $\boldsymbol{w}$ and b are parameters of the model (TAN, STEINBACH, KUMAR, 2006). Thus, for a new sample $\boldsymbol{z}$ data, we can classify this example in the following manner:

$y = 1$, if $\boldsymbol{w} \,.\, \boldsymbol{z} + b > 0$, or
$y = -1$, if $\boldsymbol{w} \,.\, \boldsymbol{z} + b < 0$

The margin of the decision boundary is given by the distance of the hyperplanes that contain the support vectors of each class. This margin is given as:

$$d = \frac{\|w\|^2}{2}$$

In the training phase of SVM, the values of w and b are obtained, as solved by the following optimization problem:

$$\min_{w} \quad \frac{\|w\|^2}{2}$$
$$\text{subject to} \quad y_i(\boldsymbol{w} \,.\, \boldsymbol{x}_i + b) \geq 1, \; i = 1, 2, \ldots, N.$$

In some instances, two classes within a data set may not be divisible on a two dimensional plane. For such cases, a model that draws a decision boundary can be used. This process is known as a *soft margin* (TAN, STEINBACH, KUMAR, 2006). In this case, slack variables (ζ_i) should be added to the constraints of the optimization problem.

$$\boldsymbol{w} \cdot \boldsymbol{x}_i + b \geq 1 - \zeta_i, \text{ if } y_i = 1$$
$$\boldsymbol{w} \cdot \boldsymbol{x}_i + b \leq -1 + \zeta_i, \text{ if } y_i = -1$$
with $\zeta_i > 0, \forall i$.

$$\min_{w} \frac{\|w\|^2}{2} + c \cdot \left(\sum_{i=1}^{N} \zeta_i\right)$$

$$\text{subject to} \quad y_i \cdot (\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \zeta_i$$

When there are no linear decision boundaries, we convert the original data set $\boldsymbol{x}$ into a new space $\varphi(\boldsymbol{x})$ in such a way that, in this new space, a linear decision boundary separates the examples. The linear decision boundary in the newly constructed space can be defined by the equation:

$$\boldsymbol{w} \cdot \boldsymbol{\varphi}(\boldsymbol{x}) + b = 0$$

We now must resolve the optimization problem:

$$\min_{\boldsymbol{w}} \frac{\|w\|^2}{2}$$
$$\text{subject to} \quad y_i(\boldsymbol{w} \cdot \boldsymbol{\varphi}(\boldsymbol{x}_i) + b \geq 1, I = 1, 2, \ldots, N.$$

After solving the optimization problem, a new sample $\boldsymbol{z}$ can be classified using the function:

$$f(\boldsymbol{z}) = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{\varphi}(\boldsymbol{z}) + b)$$

The dot product in the transformed space can be expressed in terms of a similar function in the original space:

$$k(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{\varphi}(\boldsymbol{u}) \cdot \boldsymbol{\varphi}(\boldsymbol{v}) = (\boldsymbol{u} \cdot \boldsymbol{v} + 1)^2,$$

This function k, which is computed in the original space, is referred to as the kernel function.

Some examples of these functions include:

$$k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} \,.\, \boldsymbol{y} + 1)^{p}$$

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^{2}}{2\sigma^{2}}\right)$$

$$k(\boldsymbol{x}, \boldsymbol{y}) = \tanh(k\boldsymbol{x} \,.\, \boldsymbol{y} - \delta)$$

3.3. Experiments and accuracy tests

Tests were performed to measure algorithm performance, as follows: the original 37 attributes were used, and the two most highly ranked elements, selected by the CFS Subset Eval procedure, were used (word-final [i] duration and word-final [u] duration). To compare the models, three performance criteria were employed:

$$\text{Accuracy} = \frac{TruePositives + TrueNegatives}{ruePositives + TrueNegatives + FalsePositives + FalseNegatives}$$

$$\text{Recall} = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$\text{Precision} = \frac{TrueNegatives}{TrueNegatives + FalsePositives}$$

Where, true positive means t forecasted to belong to class A and which can in fact be found in it; false positive means t forecasted to belong to A but which cannot actually be found in it; true negative means t not forecasted to belong to A and which cannot in fact be found in it; false negative means t not forecasted to belong to A but which can in fact be found in it. These values can be represented by a matrix, called the confusion matrix (Table 1).

TABLE 1
Confusion Matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	A (true positive)	B (false negative)
	Negative	C (false positive)	D (true negative)

4. Results and discussion

Information about the acoustic characteristics of individual sounds that can distinguish a native from a non-native speaker through the use of data-mining techniques is useful for the automatic identification of native versus non-native BP speakers. Knowledge about the differences between a speaker's L1 and TL may guide Portuguese teachers in developing pronunciation instruction in the future. Two aspects of this study, not included in the previously cited studies on BP vowels, are: (1) the analyses of the production of native BP speakers and non-native BP speakers and (2) the non-native participants being in a setting of university-level foreign language instruction rather than in a setting where Portuguese is spoken as a native language.

This study sheds light on which acoustic variables of vowels in syllables bearing primary stress are more important in classifying native-speaker versus non-native-speakers. In this research, we sought to answer the question: What are the acoustic features of the three BP vowels [a], [i], and [u] in syllables bearing primary stress that distinguish a native speaker from a non-native speaker? Although vowel quality is perhaps the most studied phonetic aspect of vowels, the formant frequencies of the vowels in this experiment did not stand out as important acoustic features for identifying to which group a speaker belonged. Moreover, the duration of [a], [i], and [u] in the penultimate position and [a] in the word-final position proved insignificant in determining if a speaker was native or non-native. Of the acoustic features measured in this study, only the duration of word-final [i] and [u] proved to be important in classifying native versus non-native speakers.

Acoustic analyses of non-native-speaker production versus native-speaker production agree with theories about the acquisition of similar and new vowel sounds in a second language, as stated in the interlanguage structural

conformity hypothesis (ECKMAN, 1991; ECKMAN, MORAVCSIK, WIRTH, 1989) and Flege's SLM (1995). The vowels in the interlanguage produced by non-native speakers of BP reflect a language system in its own right and a compromise between native BP vowels and the learners' L1. Based on observations of non-native speaker BP production, informants appear to discern phonetic differences between their L1 and TL (SLM Hypothesis 2). Evidence from acoustic analysis shows that non-native speakers have formed new phonetic categories for BP vowels that correspond to phonetic category representations that are acoustically close to native BP vowels (SLM Hypothesis 7). The results from these studies will assist Portuguese foreign language teachers in the development of techniques for teaching more native-like speech rhythms for BP vowel sounds.

To understand how vowel quality attributes vary between native and non-native speaker groups, plots of the mean values of F_1 and F_2 for each BP vowel are presented in Figures 1 and 2. F_1 values correlate (inversely) with tongue height, while F_2 values more or less correlate (inversely) with the tongue backness of vowel articulation. Figure 1 shows the vowels produced by native and non-native male speakers in the ultimate stressed position (or word-final position).

The vowels along the solid line represent mean values for native speakers, whereas those along the dotted line represent mean values for non-native speakers. In these tokens, male non-native speakers pronounced the [i] more fronted and slightly lower than male native speakers. The [a] is almost identical for both groups of male speakers. However, male non-native speakers produced [u] further back and slightly lower than male native speakers. For vowels in the ultimate position produced by female speakers, as shown in Figure 2, trends similar to the male speakers can be observed for [i] and [u]; however, the [a] is higher and slightly further back for female non-native speakers.

The non-native speaker group's vowel space is smaller than the native speaker group's vowel space, and the non-native speaker group's vowel space is shifted toward the front of the mouth. As far as vowel quality is concerned, none of the three vowels stand out as being significantly less accurate when compared to native-speaker pronunciation. These findings reinforce the results produced by the machine learning algorithm, which considered duration, and not vowel quality, as the most important feature in distinguishing native and non-native vowel production.

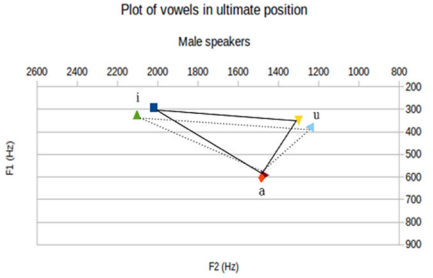

Figure 2: Plot of mean F1 and F2 values for [a], [i], and [u] in the ultimate stressed position by male native speakers (solid lines) and male non-native speakers (dotted lines).

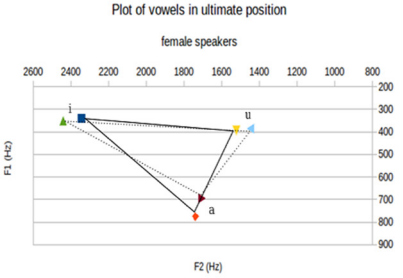

Figure 1: Plot of mean F1 and F2 values for [a], [i], and [u] in the ultimate stressed position by female native speakers (solid lines) and female non-native speakers dotted lines). speakers (dotted lines).

Figures 3 and 4 compare F_1 and F_2 values of vowels produced by BP native speakers (solid lines), BP non-native speakers (dotted lines), and American English (AmEn) native speakers (dashed lines) (AmEn values are from Hillenbrand et al., 1995). In Figure 3 (female speakers) and Figure 4 (male speakers), values for stressed vowels in the penultimate and ultimate positions in this study were averaged for comparison with AmEn speaker data, which grouped together all occurrences of stressed vowels. For each BP vowel, there are two similar AmEn vowels, that is, in the acoustic vowel space for BP [i], there are two AmEn vowels, [i] and [ɪ]. BP [a] is the low central vowel, whereas AmEn has two low vowels, [æ] and [ɑ]. The BP [u] is the high back vowel, while AmEn has two high back vowels, [u] and [ʊ]. In spite of the differences in vowel inventories between the non-native speakers' L1 and L2, for all three BP vowels, the non-native BP speakers produced vowel sounds that were closer to the native BP averages than to any of the corresponding AmEn vowels. Although all speakers, male and female, produced vowels with F_1 and F_2 values of relatively close to the BP native speaker targets, the mean values for male non-native speakers proved to be nearly identical to the native speaker means, as presented in Figure 4.

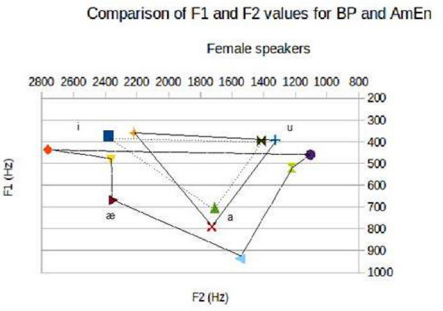

Figure 3: Mean F1 and F2 values for female BP native speakers (solid lines) BP non-native speakers (dotted lines), and American English native speakers (dashed lines). AmEn values from Hillenbrand et al. (1995).

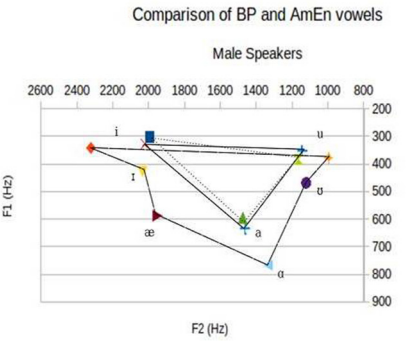

Figure 4: Mean F1 and F2 values for male BP native speakers (solid lines) BP non-native speakers (dotted lines), and American English native speakers (dashed lines). AmEn values from Hillenbrand et al. (1995)

Table 2 shows the Accuracy of the classification algorithms and Table 3 the confusion matrix using all 37 features. Highly accurate results were obtained by using the SVM algorithm, reaching 91% when all original variables were considered. Table 4 shows the confusion matrix using only 2 features – duration of word-final [i] and [u]. In this case, accuracy fell to 84.9% when only these two features were used.

TABLE 2
Classification accuracy

Machine Learning Algorithm	Accuracy (%)	Recall (%)	Precision (%)
Using 37 features	91	100	73
Using 2 features	85	100	55

TABLE 3
Confusion matrix using 37 features

		Predicted	
SVM		Non-native	Native
Actual	Non-native	22	0
	Native	3	8

TABLE 4
Confusion matrix using 2 features

		Predicted	
SVM		Non-native	Native
Actual	Non-native	22	0
	Native	5	6

Temporal features or duration are more important than vowel quality for identifying native versus non-native speakers of BP. In fact, the durations of word-final [i] and [u] were the most important acoustic measures for classifying native-speaker versus non-native-speakers of BP. Native speakers tended to produce a shorter duration for these two segments, whereas non-native speakers produced longer segments. Table 5 presents the median duration of [a], [i], and [u] in the stressed position based on a corpus of 1,000 words spoken by a native speaker from Rio de Janeiro in the first column (SIMÕES, 1991) and the median duration for the same vowels spoken by native and non-native speakers of BP in this study.

TABLE 5
Median duration in milliseconds for each segment in the stressed position

Vowel	Simões (1991)	Native BP Speakers	Non-native BP speakers
[a]	120	152	140
[i]	108	91	121
[u]	112	105	130

Table 6 reports the median duration for the same vowels (this study), but only in the stressed ultimate position, which agree with the trend of non-native speakers taking longer to articulate these sounds. However, Table 7 indicates that median duration for stressed [a] and [i] in the penultimate stress were practically identical for native and non-native speakers.

TABLE 6
Median duration in milliseconds for each segment in the ultimate stressed position

Vowel	Native BP Speakers	Non-native BP speakers
[a]	168	204
[i]	108	191
[u]	112	175

TABLE 7
Median duration in milliseconds for each segment in the penultimate stressed position

Vowel	Native BP Speakers	Non-native BP speakers
[a]	128	128
[i]	73	73
[u]	93	98.5

White and Mattys (2007a, 2007b) used a series of metrics designed to quantify speech rhythm. One of the more successful metrics, referred to as VarcoV (which is the standard deviation of vocalic interval duration divided by mean vocalic interval duration, multiplied by 100), resulted in L2 speakers having "rhythm scores intermediate between scores of their L1 and those for native speakers of the L2" (WHITE. MATTYS, 2007b, p. 245). White and Mattys also found that durational and phonotactic contrasts resulted in relatively strong stressed syllables in English and Dutch. In Spanish and French, however, relative strengths of stress and unstressed syllables were less different. Thus, in general, the stressed vowels in the current study produced by non-native speakers (with English L1) would not only be stronger, but would also have a longer duration. For non-native speakers, the lengthening effect on vowels in word-final stressed syllables was stronger than in penultimate syllables.

The temporal cues used in this study were the duration of [a], [i], and [u] in syllables bearing primary stress. Although other prosodic features contribute to a foreign accent, Kolly and Dellwo (2014) maintain that the most salient of the temporal cues in foreign accent recognition is segment duration. This study agrees that the duration of word-final high vowels, in particular, is an important feature in BP native versus non-native accent recognition. In light of Culter and Butterfield's (1990) study of pre-boundary syllable lengthening in clear speech, the longer word-final [i] and [u] produced by non-native BP speakers is not surprising. All other factors being equal, high vowels tend to be shorter than low vowels because of the extra time needed to reach the articulatory target of low vowels. The magnitude of vowel lengthening at word boundaries, in particular word-final, is language specific, and, in this experiment, the duration difference between native and non-native speech only stands out for high vowels, since low vowels are inherently longer. Although the word-final [a] also exhibited pre-boundary lengthening, the differences between the median lengths of the native versus non-native [a] were not as great as with the word-final [i] and [u].

The use of a direct repetition method for the non-native speakers may have resulted in a production closer to the target language than from other data collection methods. Although one would expect an L2 speaker to have a slower speech rate, the direct repetition technique diminished these differences in such a way that other phonetic details, such as slight variations in segment duration, could be detected. In fact, the duration results were robust enough to affirm that the phonetic differences in speaker production are sufficient for distinguishing native versus non-native speech.

5. Conclusions

This paper describes, for the first time, the values of 36 acoustic features, plus one nominal feature, in native and non-native speakers of BP in the USA. In general, native speakers talk faster than non-native speakers. However, the duration of two of the three vowel sounds, analyzed only in word-final positions, were deemed to be important differences in classifying native versus non-native speakers, suggesting that speech rate does not fully account for the longer duration in word-final [i] and [u]. Rather, non-native speakers are applying, to some extent, the L1 (English) phonological and rhythmic rules to their BP speech. Although English does not have

contractive vowel length, allophonic lengthening is a well-studied feature of English vowels. Regarding syllable weight in English, when all other things are equal, vowels in stressed syllables are longer, while vowels in open syllables tend to be longer than vowels in closed syllables (LADEFOGED, 2006). Another phenomenon that affects vowel length in the word-final position is preboundary lengthening, that is, longer syllable rhymes are cues for word boundaries (CULTER, BUTTERFIELD, 1990). In fact, regardless of stress placement in a word or phrase, vowels in word-final positions are lengthened in English utterances. Furthermore, stress timing characterizes English, whereas syllable and mora timing characterize BP (FROTA, 2001). Although some similar allophonic and rhythmic phenomena that occur in English also occur in BP, the variation in syllable weight among stress and unstressed syllables, as well as among syllables in different positions in words, does not vary as much in BP as in English because of the different rhythm classes that characterize the two languages.

The SVM, a well-known algorithm for classification, was employed in this study, the results of which show very accurate results (91%) when all variables analyzed were considered. However, a slightly worse result (84.9%) was acquired when only two features selected by a feature selection algorithm were used. Acoustic features of language production with data mining techniques may be used as a powerful and versatile alternative tool when a native speaker of a language is not available.

For future research, the methodology used in this paper can be expanded upon and replicated in other acoustic phonetic studies of language acquisition, language variation, and so forth. An acoustic analysis of L1 and L2 production, with special attention paid to temporal measures of all segments, including complete vowel inventories as well as consonants, may well be useful for speaker identification. Similar studies can be useful in identifying language varieties, such as regional and sociocultural varieties. Studies with more participants and/or longer recordings may be useful for identifying levels of language proficiency.

References

AGUIAR, G.F.M. et al. Determination of trace elements in bovine semen samples by inductively coupled plasma mass spectrometry and data mining techniques for identification of bovine class. In: *Journal of Dairy Science*, v. 95, n. 12 Amsterdam, December, 2012. p. 7066–7073.

BAPTISTA, B.O. Adult phonetic learning of a second language vowel system. In: BAPTISTA, B.O.; WATKINS, M.A. (Orgs.) *English with a Latin beat: Studies in Portuguese/Spanish-English Interphonology***.** Amsterdam: John Benjamins Publishing Company, 2006. p. 19-40.

BATISTA, B.L. et al. Multi-element determination in Brazilian honey samples by inductively coupled plasma mass spectrometry and estimation of geographic origin with data mining techniques. In: *Food Research International*, v. 49,Amsterdam, November, 2012. p. 209-215.

BLEY-VROMAN, R. The Comparative fallacy in interlanguage studies: the case of systematicity. In: *Language Learning*, v. 33, n. 1, Oxford, March, 1983. p. 1-17.

BOERSMA, P.; WEENINK, D. *Praat: doing phonetics by computer [Computer program]. Version 5.3.35.* Available at: <http://www.praat.org/> Retrieved November 18, 2013.

BOHN, O.; FLEGE, J.E. The production of new and similar vowels by adult German learners of English. In: *Studies in Second Language Acquisition*, v. 14, n. 2, Cambridge, June, 1992. p. 131-158.

BOND, Z. S.; PETROSINO, L.; DEAN, C.R. The emergence of vowels: 17 to 26 months. In: *Journal of Phonetics***,** v. 10, n. 4, Amsterdam, October,1982. p. 417-422.

BROSSEAU-LAPRÉ, F. et. al. Stimulus variability and perceptual learning of nonnative vowel categories. In: *Applied Psycholinguistics***,** v. 34, n. 3, Cambridge, May, 2013. p. 419-441.

CALLOU, D.; MORAES, J.A.; LEITA, Y. As vogais orais: Um estudo acústico-variacionsita. In: ABAURRE, M.B.M. (Org.). *Gramática do português culto falado no Brasil, Vol. VII: A construção fonológica da palavra.* São Paulo: Contexto, 2013. p. 75-93.

CAMPBELL, W.M. et. al. Support vector machines for speaker and language recognition. In: *Computer Speech and Language*, v. 20, n. 4, Amsterdam, October, 2006. p. 210-229.

CORTES, C.; VAPNIC, V. Support-vector networks. In: *Machine Learning,* v. 20. n. 3, Berlin, September, 1995. p. 273-297.

CULTER, A.; BUTTERFIELD, S. Durational cues to word boundaries in clear speech. In: *Speech Communication*, v. 9, Amsterdam, December, 1990. p. 485-495.

ECKMAN, F.R. The reduction of word-final consonant clusters in interlanguage. In: JAMES, A.; LEATHER, J. (Org.), *Sound patterns in second language acquisition***.** Dordrecht: Foris Publications, 1987. p. 143-162.

ECKMAN, F.R. The structural conformity hypothesis and the acquisition of consonant clusters in the interlanguage of ESL learners. In: *Studies in Second Language Acquisition*, v. 13, n. 1, Cambridge, March, 1991. p. 23-41.

ECKMAN, F.R.; MORAVCSIK, E.A.; WIRTH, J.R. Implicational universals and interrogative structures in the interlanguage of ESL learners. In: *Language Learning*, v. 39, n. 2, Oxford, June, 1989. p. 173-205.

FLEGE, J.E. Production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalent classification. In: *Journal of Phonetics*, v. 15, Amsterdam, January, 1987. p. 47-65.

FLEGE, J.E. Second language speech theory: theory, findings, and problems, In: STRANGE, W. (Org.). *Speech perception and linguistic experience: issues in cross-language research.* Baltimore: Timonium, York Press, 1995. p. 233-277.

FROTA, S.; VIGÁRIO, M. On the correlates of rhythm distinction: The European/Brazilian Portuguese case. In: *Probus*, v. 13, n. 2, Madrid, July, 2001. p. 247-275.

GAMA-ROSSI, A. The perception-production relationship in the acquisition of vowel duration in Brazilian Portuguese. In: *Letras de Hoje*, v. 36, n. 3, Porto Alegre, September, 2001. p. 177-186.

HALL, M. A. **Correlation-based feature subset selection for machine learning**. (Ph.D. Dissertation). Hamilton, New Zealand: University of Waikato, 1998.

HILLENBRAND, J. et. al. Acoustic characteristics of American English vowels. In: *Journal of the Acoustical Society of America*, v. 97, n. 3, Melville, May, 1995. p. 3099-3111.

HIRST, D.J. **Analyse_tier.praat.** Aix-en-Provence: PRAAT script, 2012.

HUANG, B.H.; JUN, S. The effect of age on the acquisition of second language prosody. In: *Language & Speech***,** v. 54, n. 3, London, September, 2011. p. 387-414.

Instituto Brasileiro de Geografia e Estatística (IBGE). 2010 population census. This government document provides the data regarding census of the Brazilian population in 2010. November 2010. Available at: <http://www.ibge.gov.br/english/estatistica/populacao/censo2010/default.shtm> Retrieved January 25, 2014.

KOLLY, M.; DELLWO, V. Cues to linguistic origin: The contribution of speech temporal information to foreign accent recognition. In: *Journal of Phonetics*, v. 42, Amsterdam, January, 2014. p. 12-23.

LADEFOGED, P. *A course in phonetics***.** 5 ed. Boston: Thomason Wadsworth, 2006.

LEWIS, M. P.; SIMONS, G.F.; FENNIG, C.D. (Org.). *Portuguese. Ethnologue: Languages of the World, 7th edition.* Available at: <http://www.ethnologue.com>. Retrieved November 18, 2013.

LIMA-GREGIO, A.M. et. al. Spectral findings for vowels [a] and [ɐ̃] at different velopharyngeal openings. In: *PRO-FONO: Revista de Actualização Cientifica*, v. 22, n. 4, São Paulo, December, 2010. p. 515-520.

MARKHAM, D. *Phonetic imitation, accent, and the learner***.** Lund: Lund University Press, 1997.

MORAES, J.A.; CALLOU, D.; LEITE, Y. O Sistema vocálico no português do Brasil: Caracterização acústica. In: KATO, M.A. *Gramática do português falado, vol. V***.** Campinas: UNICAMP/FAPESP, 1996a. p. 33-53.

MORAES, J.A.; CALLOU, D.; LEITE, Y. O vocalismo do português do Brasil. In: *Letras de Hoje***.** v. 31, n. 2, Porto Alegre, June, 1996b. p. 27-40.

MORRISON, G.S. L1-Spanish Speakers' Acquisition of the English /i/–/ɪ/ contrast: Duration-based perception is not the initial developmental stage. In: *Language & Speech*, v. 51, n. 4, London, December, 2008. p. 285-315.

NOBRE, M.A.; INGEMANN, F. Oral vowel reduction in Brazilian Portuguese. In: CHANNON, R.; SHOCKEY, L. (Org.), *In honor of Ilse Lehiste***.** Dordrecht: Foris Publications, 1987.

OH, E. Coarticulation in non-native speakers of English and French: An acoustic study. In: *Journal of Phonetics*, v. 36, n. 2, Amsterdam, April, 2008. p. 361-384.

PATKOWSKI, M.S. Age and accent in a second language: a reply to James Emil Flege. In: *Applied Linguistics*, v. 11, Oxford, March, 1990. p. 73-89.

PISKE, T. et. al. The production of English vowels by fluent early and late Italian–English bilinguals. In: *Phonetica***,** v. 59, Basel, January-March, 2002. p. 49–71.

PISKE, T.; MACKAY, I.R.A.; FLEGE, J.E. Factors affecting degree of foreign accent in an L2: a review. In: *Journal of Phonetics*, v. 29, n. 2, Amsterdam, April, 2001. p. 191-215.

RAUBER, A.S. An acoustic description of Brazilian Portuguese oral vowels. In: *Diacrítica, Ciências da Linguagem***,** v. 22, n. 1, Braga, January, 2008. p. 229-238.

REDENBARGER, W.J. *Articulator features and Portuguese vowel height***.** Cambridge: Department of Romance Languages and Literatures, Harvard University Press, 1981.

RVACHEW, S. et. al. Emergence of the corner vowels in the babble produced by infants exposed to Canadian English or Canadian French. In: *Journal of Phonetics*, v. 36, n. 4, Amsterdam, October, 2008. p. 564-577.

SCHÖLKOPF, B.; SMOLA, A.J. *Learning with kernels: support vector machines, regularization, optimization, and beyond.* Cambridge: The MIT Press, 2001.

SHAIMAN, S. Kinematics of compensatory vowel shortening: the effect of speaking rate and coda composition on intra- and inter-articulatory timing. In: *Journal of Phonetics*, v. 29, n. 1, Amsterdam, January, 2001. p. 89-107.

SILVA, T.C. *Fonética e fonologia de Português: Roteiro de estudos e guia de exercícios*. São Paulo: Editora Contexto, 1998.

SIMÕES, A.R.M. Towards a phonetics of discourse. In: *Cadernos de estudos linguísticos—Institudo do Estudo da Linguagem.* v. 21, Campinas. July/ December, 1991. p. 59-78.

SNOW, C.E.; HOEFNAGEL-HÖHLE, M. Age differences in pronunciation of foreign sounds. In: *Language & Speech***,** v. 20, London, October-December, 1977. p. 357-365.

TAN, P.; STEINBACH, M.; KUMAR, V. *Introduction to data mining*. Boston: Addison-Wesley: Pearson PLC, 2006.

TROVIMOVICH, P.; BAKER, W. Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. In: *Studies in Second Language Acquisition*, v. 28, Cambridge, March, 2006. p. 1-30.

WHITE, L.; MATTYS, S.L. Calibrating rhythm: First language and second language studies. In: *Journal of Phonetics*, v. 35, n. 4, Amsterdam, October, 2007a. p. 501-522.

WHITE, L.; MATTYS, S.L. Rhythm typology and variation in first and second languages. In: PRIETO, P.; MASCARÓ, J.; SOLÉ, M. (Org.). *Current Issues in Linguistic Theory: Segmental and prosodic issues in Romance phonology*. Amsterdam: John Benjamins Company, 2007b. p. 237-257.

WHITLAM, J. *Modern Brazilian Portuguese grammar: a practical guide*. New York: Routledge: Taylor & Francis Group, 2011.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data mining: practical machine learning tools and techniques.* 3rd edition. Burlington: Elsevier-Morgan Kaufmann Publishers, 2011.

Recebido em 25/01/2014. Aprovado em 05/05/2014.