Int J Speech Technol (2015) 18:583–592
DOI 10.1007/s10772-015-9299-z

Automatic prominent syllable detection with machine learning classifiers

David O. Johnson[1] · **Okim Kang**[1,2]

Received: 29 April 2015 / Accepted: 5 August 2015 / Published online: 10 September 2015

Abstract In this paper, we examine the performance of automatically detecting Brazil's prominent syllables using five machine learning classifiers and seven sets of features consisting of three features: pitch, intensity, and duration, taken one at time, two at a time, and all three. Prominent syllables are the foundation of Brazil's prosodic intonation model. We found that using pitch, intensity, and duration as features produces the best optimal results. Our findings also revealed that in terms of accuracy, F-measure, and Cohen's kappa coefficient that bagging an ensemble of decision tree learners performed the best (accuracy = 95.9 ± 0.2 %; F-measure = 93.7 ± 0.4; $\kappa = 0.907 \pm 0.005$). The performance of our current model proves to be significantly better than any other automatic detection software that exists or that of human transcription experts of prosody.

Keywords Prominent syllable detection · Machine learning · Brazil's prosodic intonation model · ToBI

1 Introduction

Prosody conveys crucial information in speech. It reflects various features of the speaker as well as the utterance: emotional state of speaker, form of the utterance, the presence of irony or sarcasm, emphasis, contrast, and focus, or other elements of language that may not be encoded by grammar or by choice of vocabulary. Prosody extends over one single sound segment in an utterance and covers other paralinguistic aspects of speech such as pitch, tone, duration, intensity, and voice quality (Chun 2002). Several intonational models for representing prosodic features have been practiced. However, they often face various challenges and limitations in terms of adequately interpreting actual, natural human discourse (Xu 2012).

Two principle frameworks are often used for representing and interpreting prosodic features: Brazil (1997) and Pierrehumbert (1980). Both the Pierrehumbert and Brazil models provide important information about the prosodic features of human discourse. Pierrehumbert's framework is widely used to model text-to-speech synthesis and has been realized quantitatively. Brazil's framework has been used in discourse analysis and language teaching and learning (Cauldwell 2012). The former model does not account for the meaning of intonation in naturally occurring discourse sufficiently (Dilley 2005).

Dilley (2005) proposed a tone interval theory, which captured the intonational and rhythmic aspects of speech. Her theory provided experimental evidence that showed how Pierrehumbert's framework (Pierrehumbert 1980; Pierrehumbert and Beckman 1988) did not account for intonation meaning. However, such tone interval theory still does not explain how tone units are used in discourse. This is the motivation of the current study choosing Brazil's model, which offers efficient and meaningful interpretation of natural discourse (e.g., tone choices). It emphasizes the idea of interactional significance of prosody and the achievement of the communicative functions in discourse (Brazil 1997). For example, Brazil's model stresses the first and last prominent syllables and comprises pitch concord, which are essential distinctions across varieties of languages (Pickering 1999). It uses tone

✉ Okim Kang
okim.kang@nau.edu

[1] Northern Arizona University, Flagstaff, AZ, USA

[2] Northern Arizona University, Liberal Arts Building #18, Room 140, PO Box 6032, Flagstaff, AZ 86011, USA

choices for the interpretation of speakers' intention, meaning, emotion, or other communicative purposes in the discourse (Pickering 2009).

In a monologic speech, rising tones are used for showing solidarity or expressing known or shared knowledge or indicating uncertainty or lack of power. Falling tones are for presenting a topic closure or expressing new information, or showing speakers' authority. Level tones are more for focusing on action rather than discourse or indicating the continuation of discourse. In addition, rising contours are associated with anger, fear, and joy whereas falling contours are connected to sadness and tenderness (Juslin and Laukka 2004). The patterns of these tones can be understood in the relationship between the final tone unit of one move and the initial key choice of the next move, called pitch concord (Brazil 1997).

The fundamental unit of Brazil's model is the prominent syllable. Brazil is very clear in his work that the importance of prominence is on the syllable and not the word. He provided examples of words with more than one prominent syllable and words whose prominent syllable varied depending on the intonational meaning the speaker was imparting. Although the rest of Brazil's model is easy to quantify, what makes a syllable prominent, is difficult to compute. Brazil states that prominent syllables are recognized by the hearer as having more emphasis than other syllables. A trained analyst can easily identify prominent syllables by listening to an utterance. However, quantifying the difference between a prominent syllable and a non-prominent syllable is not so straightforward. Brazil further notes in his description of prominent syllables that prominence should be contrasted with word or lexical stress. Lexical stress focuses on the syllable within content words that is stressed. However, prominence focuses on the use of stress to distinguish those words that carry more meaning, more emphasis, more contrast, in utterances. Thus, a syllable within a word that normally receives lexical stress may receive additional pitch, length, or loudness to distinguish meaning (Brazil 1997). Alternatively, a syllable that does not usually receive stress (such as a function word) may receive stress for contrastive purposes.

Brazil's definition of prominence is similar, but more specific, than more commonly known definitions of prominence. Terken (1991) stated prominence is the attribute of a linguistic unit which makes it stand out from its environment perceptually. However, he did not precisely specify the linguistic unit as the syllable as did Brazil. But, like Brazil, he said prosodic prominence is connected to the suprasegmental pitch, duration, and intensity attributes of speech.

The purpose of this paper is to determine the best machine learning classifier and set of features, chosen from pitch, length (i.e., duration), or loudness (i.e., intensity) to automatically detect Brazil's prominent syllables. Specifically, we will assess the performance of five machine learning classifiers and seven sets of features consisting of three features: pitch, intensity, and duration, taken one at time, two at a time, and all three, in automatically detecting Brazil's prominent syllables.

Section 2 reviews existing research in the area of prominent syllable and prominent word detection. Section 3 describes the speech corpus, classification features, machine learning classifiers, and experimental methods used in the current research. In Sect. 4, we present the results of our experiments, followed by a comparison with other research findings in the field of speech science along with conclusions in Sect. 5.

2 Prominent syllable and word detection research

In the field of speech production and engineering, Brazil's (1997) framework has been hardly utilized. In contrast, there is a large body of research on detecting Pitch Accents and Boundary Tones as defined by the ToBI standard. The tones and break indices (ToBI) is a system for labeling prosodic events in spoken utterances (Wightman et al. 1992; Beckman and Elam 1997). This standard specifies three types of prosodic events: Pitch Accents, Boundary Tones and Break Indices. Pitch Accents refer to the prosodic function of prominence. Boundary Tones and Break Indices refer to the prosodic function of phrasing. Although pitch accents are defined as a function of prominence, there are usually more pitch accents in an utterance than Brazil's prominent syllables. This is due to the fact, that Pierrehumbert did not make a distinction between lexical stress and what Brazil calls, "prominence". Thus, there is no one-to-one correspondence between Brazil's concept of prominent syllables and Pitch Accents, Boundary Tones, or Break Indices.

The ToBi-related research uses a variety of intensity, duration, and pitch measurements along with lexical or syntactic cues (i.e., features) to detect prosodic events. Ludusan and Dupoux (2014) investigated using several duration and pitch features, by themselves and in a combination, without any lexical or syntactic cues to detect prosodic boundaries. They found that a combination of all the cues compared well with previous work. Ni et al. (2011) detected ToBI Pitch Accents with an accuracy of 91.4 % and Boundary Tones with an accuracy of 95.2 % utilizing pitch, duration, intensity, and lexical and syntactic cues. Later, they applied the same techniques to detect Mandarin stress (Ni et al. 2012) with an accuracy of 89.9 %. Jeon and Liu (2009) also used pitch, duration, intensity, and lexical and syntactic features to detect ToBI Pitch Accents and achieved an accuracy of 89.8 %. Likewise, Ananthakrishnan and Narayanan (2008) detected ToBI accent (86.75 % accuracy) and prosodic phrase boundaries (91.6 % accuracy) with pitch, duration, intensity, and lexical and syntactic cues.

To classify both Pitch Accents and Boundary Tones, González-Ferreras et al. (2012) used a number of acoustic features (pitch, energy, and vowel nucleus duration), lexical and syntactic features (part-of-speech tags), and pitch contour features with fusion of pairwise coupled neural network and decision trees classifiers and applied the Viterbi algorithm to find the best tone sequence to achieve classification accuracies of 70.8 % (pitch accents) and 84.2 % (boundary tones). With pitch, duration, intensity, and lexical and syntactic features, Sridhar et al. (2008) detected ToBI Pitch Accents (86 % accuracy) and Boundary Tones (93.1 % accuracy).

Rosenberg and Hirschberg (2009) compared pitch accent identification at the syllable, vowel, and word level, and found that a word level approach is superior to syllable or vowel level identification achieving an accuracy of 84.2 %.

Silipo and Greenberg (1999) concluded that intensity and duration are the most important acoustic parameters underlying prosodic stress in casually spoken American English, and that pitch plays only a minor role in the assignment of stress. In a later study (Silipo and Greenberg 2000), they reexamined this conclusion using both the range and average level of pitch to determine whether there were circumstances in which pitch figures importantly in prosodic stress. They found in the later study that pitch range is slightly more effective than average pitch. They explained that this finding was most likely a consequence of duration-related information intrinsic to pitch range, and was thus consistent with their early finding that pitch played a relatively minor role in stress assignment in naturally spoken American English. Kochanski et al. (2005) studied seven dialects of British and Irish English and three different styles of speech to find acoustic correlates of prominence. They found pitch played a minor role in distinguishing prominent syllables from the rest of the utterance. Instead, speakers primarily marked prominence with patterns of intensity and duration. Rosenberg and Hirschberg (2006) studied the correlation between intensity and pitch accent of four native speakers of Standard American English. They were able to predict pitch accent in read speech with an accuracy of 81.9 % using only intensity.

There are also other non-ToBI research initiatives examining "prominence", where "prominence" in this case also includes lexical stress. These research initiatives combine intensity, duration, pitch, and other acoustic features (i.e., no lexical or syntactic cues) to automatically identify syllabic prominence. In relatively recent years, various studies have attempted to detect such prominence types. Avanzi et al. (2010) detected syllabic prominence in French with pitch, duration, and pause. Similarly, Streefkerk et al. (1997) identified prominence in Dutch with intensity and duration and no pitch. Prominence in English was detected with pitch, intensity, and duration by Mahrt et al. (2011, 2012a, b). Tamburini (2006) used pitch movements, overall syllable energy, syllable nuclei duration, and mid-to-high-frequency emphasis to detect syllabic prominence in English. Some researchers only used pitch and intensity to detect syllabic prominence in French and Italian (Ludusan et al. 2011). Finally, Cutugno et al. (2012) included pitch, intensity, and duration of a syllable and its neighbors to detect syllabic prominence in English and Italian.

The research above shows that a good number of machine learning classifiers and features have been employed to detect various types of prominence. In this paper, we will determine the best machine learning classifier and set of features to automatically detect Brazil's prominent syllables. We will only test pitch, duration, and intensity because those are the features 'prominence' is comprised of in Brazil (1997) terms. Specifically, we will examine the performance of five machine learning classifiers (neural network, decision tree, support vector machine, bagging, and boosting) and seven sets of features consisting of three features: pitch, intensity, and duration, taken one at time, two at a time, and all three.

3 Methods

3.1 TIMIT corpus

The DARPA TIMIT Acoustic–Phonetic Continuous Speech Corpus (TIMIT) of read speech was designed to provide speech data for the acquisition of acoustic–phonetic knowledge and for the development and evaluation of automatic speech recognition systems (Garofolo et al. 1993). TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. The text material in the TIMIT prompts consists of two dialect sentences, 450 phonetically-compact sentences, and 1890 phonetically-diverse sentences. The dialect sentences were intended to reveal the dialect of the speakers and were read by all 630 speakers. The phonetically-compact sentences were designed to provide a good coverage of pairs of phones, with extra occurrences of phonetic contexts thought to be either of particular interest or difficult. Each speaker read five of these sentences and each text was spoken by seven different speakers. The phonetically-diverse sentences were selected to maximize the variety of allophonic contexts found in the texts. Each speaker read three of these sentences, with each sentence being read only by a single speaker. The corpus includes hand corrected start and end times for the phones, phonemes, pauses, syllables, and words. The TIMIT corpus includes definitions for 60 phones. The TIMIT phones are used by other corpora. For this research, we used a subset of the corpus consisting of 84 speakers speaking four dialects. There were 836 utterances in our subset containing 10,657 syllables. Table 1 shows the distribution of speakers by gender and dialect.

Table 1 Distribution of TIMIT speakers by gender and dialect used in this research

Dialect	Male	Female	Total
New England	7	4	11
Northern	18	8	26
North Midland	23	3	26
South Midland	5	16	21
Total	53	31	84

We augmented the corpus by identifying the prominent syllables in the experimental subset. The prominent syllables were identified by a trained analyst who coded them both by listening to the audio files and by using the MultiSpeech and CSL Software (KayPENTAX 2008) to view the pitch, intensity, and duration of the syllables. Roughly, 10 percept of the samples were analyzed by a second trained analyst to verify the reliability of prominent syllable coding. The inter-coder reliability between the two human coders was around 85–87 %, which were relatively acceptable rates as seen in other similar coding protocols (e.g., Kang 2010) particularly using Brazil's (1997) framework. The two analysts reviewed any inconsistencies and resumed coding the samples until they agreed on the coding. The first analyst then completed the analysis independently for the remaining speech samples. The analyst identified 3536 prominent syllables in the speech samples. This coding method has been widely practiced as a reliable labeling scheme in other studies (Kang 2010; Kang et al. 2010; Pickering 1999) in applied linguistics.

Although the TIMIT corpus consists of isolated sentences, we chose it for prominent syllable detection because it contained a wide variety of speakers and dialects, many more than the six speakers and one dialect in the Boston University Radio News Corpus (Ostendorf et al. 1995), which is another commonly used corpus for intonation studies. The current study only used a subset of the TIMIT corpus. However, 84 speakers speaking four dialects with over 10,000 syllables and over 3500 prominent syllables proved to be sufficient for the identification of an appropriate prominent syllable classifier and feature set.

3.2 Classification features

As input to the classifiers, we used seven sets of features for each syllable consisting of combinations of three features: pitch, intensity, and duration, taken one at time, two at a time, and all three. The pitch feature was calculated by taking the median of the pitch contour of the syllable extracted by Praat (Boersma and Weenink 2014). The intensity feature was calculated by taking the maximum of the intensity contour of the syllable extracted using the Matlab *audioread* function (MathWorks 2013). The duration feature was calculated by using the syllable start and stop times from the TIMT corpus (Garofolo et al. 1993). In other words, the prominent syllable classifiers used the syllable boundaries given in the corpus. Pitch, intensity, and duration vary across speakers. They can be different even for the same speaker due to various idiosyncratic factors. To ameliorate the effect that this variation might have on prominent syllable detection, the features within a run (i.e., a run is the speech between two pauses, where a pause is defined as a silence longer than 100 ms; the lengths of the pauses were provided by the corpus) were normalized with Z-scores and scaled to the interval [−1 1] as follows:

$$f_i = \text{feature value for syllable i} \tag{1}$$

$$f_{mean} = \text{mean of } f_i \text{ for all syllables in the run} \tag{2}$$

$$f_{std} = \text{standard deviation of } f_i \text{ for all syllables in the run} \tag{3}$$

$$fnorm_i = (f_i - f_{mean})/f_{std} \tag{4}$$

$$fnorm_{max} = \text{maximum } fnorm_i \text{ for all syllables in the run} \tag{5}$$

$$fnorm_{min} = \text{minimum } fnorm_i \text{ for all syllables in the run} \tag{6}$$

$$fnorm_{scale} = \max(fnorm_{max}, |fnorm_{min}|) \tag{7}$$

$$fscaled_i = fnorm_i/fnorm_{scale} \tag{8}$$

Z-score normalization provides a zero-mean, unit-standard deviation normalization of the input data. For this to be valid there is an assumption that the underlying data be normally distributed. The [−1, 1] interval normalization is extremely sensitive to outliers. However, we tried three other methods of normalization: (1) no normalization, (2) dividing the feature values by the mean feature value of the run, and (3) Z-score normalization without interval normalization; and found all of them provided worse performance in terms of accuracy, F-measure, and κ.

3.3 Classifiers

We used five standard machine-learning classifiers to detect prominent syllables: neural network, support vector machine, decision tree, bagging, and boosting.

In machine learning, neural networks are a group of statistical learning models motivated by the biological neural networks in animal brains (Happel and Murre 1994). They are utilized to estimate or approximate functions (e.g., prominent syllable detection) that can depend on a number of unknown inputs (e.g., pitch, duration, and intensity). Neural networks are commonly portrayed as arrangements of interconnected nodes that send messages to each other, representing the interconnection of neurons in the brain. The connections have numeric weights that are tuned with a set of training data, allowing neural nets to adjust to inputs and capable of learning. We employed the Matlab *fitnet* function with ten hidden nodes (i.e., neurons) to implement the neural network classifier (MathWorks 2013).

Support vector machines are machine learning models with associated learning algorithms that recognize patterns (i.e., pitch, intensity, and duration of prominent syllables) (Cortes and Vapnik 1995). Provided with a set of training examples (e.g., pitch, intensity and duration of syllable), each denoted as belonging to one of two categories (e.g., prominent syllable or non-prominent syllable), a support vector training algorithm constructs a model that designates new examples as belonging to one category or the other. It is a non-probabilistic binary linear classifier. A support vector model is a depiction of the examples as points in space (e.g. pitch, intensity, and duration as points in 3-dimensional space), plotted so that the examples of the separate classes (e.g., prominent syllable and non-prominent syllable) are partitioned by a well-defined gap that is maximally broad. A new example is then plotted into that same space and determined to belong to a class depending on which side of the gap it is on. We utilized the Matlab *svmtrain* function to implement the support vector machine classifier (MathWorks 2013).

Decision tree learning is a machine learning technique that makes use of a decision tree as a predictive model to map observations (e.g., pitch, intensity, and duration) about an item (e.g., syllables) to conclusions about the item's target value (e.g., prominent or non-prominent) (Quinlan 1999). It is a predictive modeling approach frequently found in statistics and data mining. Classification trees are models where the target variable can take a finite set of values. Leaves of the tree represent class labels (e.g., prominent and non-prominent) and branches are combinations of features that lead to those class labels (e.g., pitch, intensity, and duration). Decision tree learning is one of the more successful machine learning techniques. The decision tree classifier was implemented with the Matlab *ClassificationTree* function (MathWorks 2013).

Bagging and boosting are ensemble classifiers that combine the results of weak classifiers (typically decision trees) to improve their performance. Ensemble prediction usually entails more calculations than predicting with a single model, thus ensembles may be considered as a means to make up for poor learning algorithms with extra computation (Opitz and Maclin 1999). An ensemble is itself a machine learning technique, because it is trained and then applied to make predictions. Ensembles are more flexible in the functions they can model. This flexibility can lead to them over-fitting the training data more than a single model would. To compensate for this, ensemble classifiers employ techniques that reduce problems related to over-fitting of the training data.

Bagging stands for *bootstrap aggregation* (Breiman 1994). Bagging is accomplished by replicating portions of the training data and constructing multiple decision trees from the replicated data. The output of the ensemble is the average of the predictions from the individual trees. Bagging was implemented with the Matlab *fitensemble* function using 100 decision tree learners (MathWorks 2013).

Most boosting algorithms entail repetitive training of weak classifiers and adding them to a final strong classifier (Breiman 1996). When they are added, they are usually weighted in a manner that is typically connected to the weak learners' accuracy. When a weak learner is added, the outputs of the other weak learners are reweighted. Instances that are misclassified lose weight and instances that are classified appropriately gain weight. Thus, future weak learners concentrate more on the instances that prior weak learners classified incorrectly. Several boosting algorithms also decrease the weight of examples that are continually classified incorrectly. Boosting was realized with the Matlab *fitensemble* function using the AdaBoostM1 booster and 100 decision tree learners (MathWorks 2013).

Resubstitution error is the variation between the actual responses (i.e., prominent and non-prominent) in the training data and the responses the tree predicts based on the input training data (i.e., pitch, intensity, and duration). If the resubstitution error is high, you cannot expect the predictions of the tree to be good. A common method of determining the number of decision trees to use in an ensemble is to plot resubstitution error versus number of trees and use a number of decision trees well past the knee of the curve (MathWorks 2013). Figure 1 illustrates that 100 decision tree learners is sufficient.

None of the classifiers were optimized beyond the standard settings for the Matlab functions.

3.4 Experimental design

In all the experiments we applied five-fold cross-validation. The folds were created by randomly assigning the 84

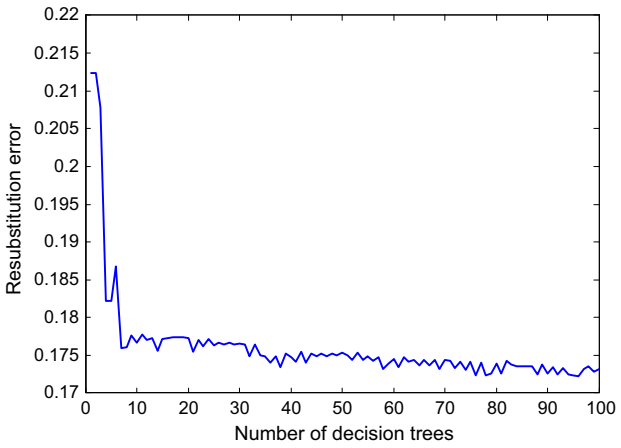

Fig. 1 Number of decision trees versus resubstitution error

speakers to folds. Speakers were randomly assigned to folds rather than the utterances to ensure that training and testing on the same speaker did not bias the experiments. Thirty-five experiments were conducted: one for each combination of the five classifiers (i.e., neural network, decision tree, support vector machine, bagging, and boosting) and seven combinations of features (i.e., pitch, intensity, and duration taken one at a time, two at a time, and all three at a time).

4 Results

The purpose of this research is to determine which machine classifier and set of features, chosen from pitch, duration, and intensity, is the best to automatically detect Brazil's prominent syllables. In 35 experiments, we examined the performance of five machine learning classifiers and seven sets of features consisting of three features: pitch, intensity, and duration, taken one at time, two at a time, and all three at time, in automatically detecting Brazil's prominent syllables. To evaluate the performance of the five classifiers and the seven sets of features, accuracy, F-measure, and Cohen's kappa coefficient (κ) (Cohen, 1960) were used. Accuracy and F-measure are calculated as follows:

$$\text{TP} = \text{number of syllables where both the computer and the human identified it as prominent} \quad (9)$$

$$\text{TN} = \text{number of syllables where both the computer and the human identified it as not prominent} \quad (10)$$

$$\text{FP} = \text{number of syllables where the computer identified it as prominent and the human identified it as not prominent} \quad (11)$$

$$\text{FN} = \text{number of syllables where the computer identified it as not prominent and the human identified it as prominent} \quad (12)$$

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (13)$$

$$\text{F-Measure} = 2\text{TP}/(2\text{TP} + \text{FP} + \text{FN}) \quad (14)$$

The confidence interval was assumed to be symmetrical and was calculated as follows, where n is the number of folds (5) and σ is the standard deviation of the folds:

$$\text{Confidence interval} = \pm\frac{\sigma}{2\sqrt{n}} \quad (15)$$

Table 2 shows the performance of five classifiers: neural network, decision tree, support vector machine, bagging, and boosting, using seven different sets of the features: duration, intensity, and pitch. The accuracy, F-measure, and Cohen's kappa coefficient (κ) are the mean of the fivefolds.

Bagging is clearly the best classifier for identifying prominent syllables and the best feature set is duration, intensity, and pitch by all three measures (accuracy = 95.9 ± 0.2 %; F-measure = 93.7 ± 0.4; κ = 0.907 ± 0.005). The second best classifier is Decision Tree, which is what would be expected, since Bagging is a method for improving classification results by using an ensemble of 100 Decision Trees. Comparisons between human coder agreements and the machine are further provided in Sect. 5.

5 Discussion

The results showed that our computer program could detect prominent syllables with an accuracy of 95.9 % (±0.2 %), an F-Measure of 93.7 (±0.4), and a κ of 0.907 (±0.005) when compared with humans. These results can be interpreted through those from other related computer programs. They can also be discussed with those between other human experts.

First, even though, cross-corpus comparisons are not always reliable, there are other related computer programs where prominence was identified automatically. Avanzi et al. (2010) reported F-measures of three French syllabic prominence detectors: ANALOR (69.7), PROSOPROM (71.7), and IRCAMPROM (75.4). We achieved an F-measure of 93.7 (±0.4), which is significantly better than ANALOR, PROSOPROM, and IRCAMPROM. Obin et al. (2009) proposed an approach for detecting prominence in a corpus of French read speech which obtained an F-measure of 87.5 and an accuracy of 90.4 %. Christodoulides and Avanzi (2014) trained and evaluated four classifiers on a

Table 2 Accuracy, F-measure, and Cohen's kappa coefficient (κ) and confidence interval for different classifiers and different sets of features sorted by accuracy

Classifier	Features	Accuracy (%)	F-measure	κ
Bagging	Duration, intensity, pitch	95.9 ± 0.2	93.7 ± 0.4	0.907 ± 0.005
Bagging	Duration, intensity	95.2 ± 0.3	92.7 ± 0.4	0.891 ± 0.006
Bagging	Intensity, pitch	95.1 ± 0.3	92.5 ± 0.5	0.888 ± 0.007
Bagging	Duration, pitch	94.2 ± 0.2	91.2 ± 0.4	0.869 ± 0.006
Bagging	Intensity	93.3 ± 0.4	90.0 ± 0.5	0.850 ± 0.008
Bagging	Duration	90.4 ± 0.4	86.1 ± 0.5	0.787 ± 0.008
Bagging	Pitch	89.9 ± 0.4	84.5 ± 0.6	0.769 ± 0.009
Decision tree	Duration, intensity, pitch	89.4 ± 0.2	83.8 ± 0.4	0.760 ± 0.005
Decision tree	Duration, intensity	88.7 ± 0.3	82.6 ± 0.4	0.742 ± 0.006
Decision tree	Intensity, pitch	87.7 ± 0.4	81.1 ± 0.5	0.720 ± 0.008
Decision tree	Intensity	86.5 ± 0.3	79.4 ± 0.4	0.694 ± 0.006
Decision tree	Duration, pitch	85.7 ± 0.3	77.9 ± 0.4	0.674 ± 0.006
Decision tree	Duration	83.2 ± 0.2	75.0 ± 0.2	0.624 ± 0.003
Boosting	Duration, intensity, pitch	82.5 ± 0.1	72.3 ± 0.2	0.596 ± 0.003
Boosting	Duration, intensity	82.4 ± 0.1	71.9 ± 0.2	0.592 ± 0.003
Neural network	Duration, intensity, pitch	82.3 ± 0.1	71.4 ± 0.1	0.587 ± 0.002
Neural network	Duration, intensity	82.2 ± 0.1	71.2 ± 0.2	0.584 ± 0.003
Support vector machine	Duration, intensity	81.1 ± 0.1	72.5 ± 0.2	0.581 ± 0.002
Decision tree	Pitch	80.1 ± 0.4	69.4 ± 0.6	0.546 ± 0.009
Boosting	Intensity, pitch	79.2 ± 0.1	66.7 ± 0.1	0.517 ± 0.001
Neural network	Intensity, pitch	79.1 ± 0.1	65.6 ± 0.2	0.508 ± 0.002
Boosting	Intensity	78.7 ± 0.1	67.9 ± 0.2	0.519 ± 0.003
Neural network	Intensity	78.7 ± 0.1	65.4 ± 0.1	0.501 ± 0.001
Neural network	Duration, pitch	76.3 ± 0.2	62.4 ± 0.4	0.452 ± 0.005
Boosting	Duration, pitch	76.3 ± 0.2	62.5 ± 0.3	0.452 ± 0.005
Support vector machine	Intensity, pitch	75.9 ± 1.0	54.8 ± 6.1	0.414 ± 0.046
Support vector machine	Duration, intensity, pitch	75.2 ± 1.5	43.8 ± 8.0	0.349 ± 0.064
Support vector machine	Duration, pitch	74.5 ± 0.2	66.4 ± 0.1	0.464 ± 0.003
Support vector machine	Intensity	73.6 ± 0.7	56.0 ± 2.2	0.377 ± 0.024
Boosting	Duration	73.2 ± 0.1	55.3 ± 0.4	0.365 ± 0.004
Neural network	Duration	73.0 ± 0.1	57.8 ± 0.4	0.380 ± 0.005
Boosting	Pitch	67.6 ± 0.3	35.6 ± 0.7	0.167 ± 0.007
Support vector machine	Duration	67.4 ± 0.3	10.1 ± 4.5	0.057 ± 0.026
Neural network	Pitch	67.3 ± 0.2	27.3 ± 1.2	0.119 ± 0.008
Support vector machine	Pitch	65.7 ± 0.2	40.1 ± 0.3	0.169 ± 0.004

corpus of spontaneous French speech and found the neural network classifier was the best with and accuracy of 84.2 % and an F-measure of 79.1. Rosenberg and Hirschberg (2010) found classifiers trained on Mandarin L1 English could automatically detect prominence in Mandarin L1 English with an accuracy of 87.2 % and an F-measure of 86.6 while those trained on native English speech detected prominence with an accuracy of 74.8 % and F-measure of 82.4. The current results are higher than all of these studies with an F-measure of 93.7 (±0.4) and an accuracy of 95.9 % (±0.2 %).

The machine classifier performances shown in Table 2 can also be compared to the inter-rater agreement between two human experts. Price et al. (1988) conducted an inter-rater agreement study on a set of three stories from the Boston University Radio News Corpus (Ostendorf et al. 1995) containing 1002 words. They found agreement on presence versus absence for 91 % of the words. Boundary tone agreement was 93 % for the 207 words marked by both labelers with an intonational phrase boundary, and similarly there was 91 % agreement for 280 phrase accents. Ludusan et al. (2011) reported an inter-rater agreement of 91.5 % on syllabic prominence. In this case, the human–human inter-rater reliability was less than the human–computer inter-rater reliability of 95.9 % for the best classifier shown in Table 2.

In another example, where the human–computer inter-rater reliability of 95.9 % was greater than the human–human inter-rater reliability, Kang (2010) found the inter-rater agreement between two phonetic analysts was 86 % or lower in identifying Brazil's prominent syllables. The main problem raised in her study was human coder's subjectivity and tiredness involved in the labor-intensive procedure of prominence analysis. Indeed, prominence analyses are subject to perceptual limitations (Kang and Pickering 2013). Discrepancies between the two human analysts tend to take place in determining the location of prominent syllables in a series of discourse. Therefore, a calibrating procedure having two human analysts reach consensus is required to ensure the reliability of the analysis; however, this process is often known to be difficult. Accordingly, the current method achieving 95.9 % (±0.2 %) agreement between the human rater and the computer is very promising. Such obtainment of high agreements was possible due to the consistency of a computer program, once it was trained on the basis of protocols used for human coders. While certain ambiguous parts of speech could involve inconsistent judgments between two human coders, a computer program can make it constant and coherent throughout the speech. This computer-based prominence detection suggests a useful resource to supplement human coding in the field of speech science.

The inter-rater agreement between two human experts can likewise be contrasted to the machine classifier performances shown in Table 2 with Cohen's kappa coefficient. Escudero-Mancebo et al. (2014) noted that in the current state of art for ToBI research, κ ranges from 0.51 (Yoon et al. 2004) to 0.69 (Syrdal and McGory 2000). Breen et al. (2012) reported κ values of 0.52 and 0.77 for RaP research. The RaP (Rhythm and Pitch) system is a method of labeling the rhythm and relative pitch of spoken English. It is an extension of ToBI that permits the capture of both intonational and rhythmic aspects of speech (Dilley and Brown 2005). It is based on tone interval theory proposed by Dilley (2005). Nevertheless, the current method achieved a much greater κ value (0.907 ± 0.005) for inter-rater agreement between the computer and a human than either the ToBI research or RaP research.

The performances of the support vector machine and the neural network are very low. For the support vector machine, this is probably because the support vector machine is an older machine learning method, which has been surpassed in performance by more modern methods, such as ensembles of decision trees. Escudero-Mancebo et al. (2014) also found that support vector machines underperformed neural networks and decision trees. A more likely reason for the poor performance of the support vector machine and the neural network is that machine learning techniques (e.g., support vector machines, neural networks, decision trees, bagging, and boosting) perform differently in different applications. There is no machine learning technique that works best in all applications. That is the reason we compared the performance of more than one machine learning technique.

6 Conclusions

Overall, the current research has shown that it is possible to detect the fundamental element of Brazil's model, prominent syllables, with an accuracy exceeding that of two human analysts and other programs that measure prominence. This is an important achievement because detecting prominent syllables is the foundation of Brazil's theory. As we discussed earlier, Dilley (2005) showed that Pierrehumbert's framework (Pierrehumbert 1980; Pierrehumbert and Beckman 1988) does not account for the meaning of intonation in natural discourse. On the other hand, Brazil's model offers a meaningful interpretation of natural discourse by emphasizing the interactional communicative functions in discourse (Brazil 1997). Thus, automatically detecting prominent syllables is the important first step in automatically interpreting the interactional aspects of natural discourse.

The next steps are finding classifiers and algorithms with appropriate feature sets for automatically detecting the other elements of Brazil's model (i.e., tone unit, tone choice, relative pitch, and pitch concord). Automatic interpretation of natural discourse has many applications in automatic speech recognition (Bocklet and Shriberg 2009; Hämäläinen et al. 2007; Litman et al. 2000; Ostendorf 1999), text-to-speech synthesis, speaker verification and identification (Shriberg et al. 2005; Escudero-Mancebo et al. 2014), human-robot interaction (Nadel et al. 2006), automatic speech scoring systems (Kang et al. 2010; Kang and Wang 2014), computer-aided language learning, forensics, and early childhood diagnosis of autism (Frith and Happé 1994; Fine et al. 1991; Paul et al. 2005; Shriberg et al. 2001; McCann and Peppé 2003). The current study demonstrated the potential of exploring a new discourse-based intonation model, i.e., Brazil's (1997) intonation discourse framework, to better understand natural discourse in various contexts.

References

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.

Avanzi, M., Lacheret-Dujour, A., & Victorri, B. (2010). A corpus-based learning method for prominence detection in spontaneous speech. In *Proceedings of prosodic prominence, speech prosody 2010 satellite workshop*, Chicago, 10 May.

Beckman, M., & Elam, G. (1997). Guidelines for ToBI labelling. http://www.ling.ohio-state.edu/research/phonetics/E_ToBI.

Bocklet, T., & Shriberg, E. (2009, April). Speaker recognition using syllable-based constraints for cepstral frame selection. In *IEEE international conference on acoustics, speech and signal processing, 2009 (ICASSP 2009)* (pp. 4525–4528). IEEE.

Boersma, P., & Weenink, D. (2014). Praat: Doing phonetics by computer (version 5.3.83). [Computer program]. Retrieved August 19, 2014.

Brazil, D. (1997). *The communicative value of intonation in English*. Cambridge: Cambridge University Press.

Breen, M., Dilley, L. C., Kraemer, J., & Gibson, E. (2012). Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch).

Breiman, L. (1994). *Bagging predictors*. Technical Report 421. Department of Statistics, University of California at Berkeley.

Breiman, L. (1996). *Bias, variance, and arcing classifiers*. Technical Report 460. Department of Statistics, University of California at Berkeley.

Cauldwell, R. (2012). RIAS VAN DEN DOEL, How friendly are the natives? An evaluation of native-speaker judgements of foreign-accented British and American English. Utrecht: Netherlands Graduate School of Linguistics (LOT), 2006. pp. xii + 341. ISBN-10: 90-78328-09-6, ISBN-13: 978-90-78328-09-4. *Journal of the International Phonetic Association, 42*(02), 213–215.

Christodoulides, G., & Avanzi, M. (2014). An evaluation of machine learning methods for prominence detection in French. In *Fifteenth annual conference of the International Speech Communication Association*.

Chun, D. M. (2002). *Discourse intonation in L2: From theory and research to practice*. Amsterdam: John Benjamins.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273–297.

Cutugno, F., Leone, E., Ludusan, B., & Origlia, A. (2012). Investigating syllabic prominence with conditional random fields and latent-dynamic conditional random fields. In *INTERSPEECH*.

Dilley, L. C. (2005). *The phonetics and phonology of tonal systems*. Doctoral dissertation, Massachusetts Institute of Technology.

Dilley, L. C., & Brown, M. (2005). The RaP (Rhythm and Pitch) labeling system. Unpublished manuscript.

Escudero-Mancebo, D., González-Ferreras, C., Vivaracho-Pascual, C., & Cardeñoso-Payo, V. (2014). A fuzzy classifier to deal with similarity between labels on automatic prosodic labeling. *Computer Speech & Language, 28*(1), 326–341.

Fine, J., Bartolucci, G., Ginsberg, G., & Szatmari, P. (1991). The use of intonation to communicate in pervasive developmental disorders. *Journal of Child Psychology and Psychiatry, 32*(5), 771–782.

Frith, U., & Happé, F. (1994). Language and communication in autistic disorders. *Philosophical Transactions of the Royal Society B: Biological Sciences, 346*(1315), 97–104.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon Technical Report N, 93, 27403.

González-Ferreras, C., Escudero-Mancebo, D., Vivaracho-Pascual, C., & Cardeñoso-Payo, V. (2012). Improving automatic classification of prosodic events by pairwise coupling. *IEEE Transactions on Audio, Speech, and Language Processing, 20*(7), 2045–2058.

Hämäläinen, A., Boves, L., de Veth, J., & Bosch, L. T. (2007). On the utility of syllable-based acoustic models for pronunciation variation modelling. *EURASIP Journal on Audio, Speech, and Music Processing, 2007*(2), 3.

Happel, B. L., & Murre, J. M. (1994). Design and evolution of modular neural network architectures. *Neural Networks, 7*(6), 985–1004.

Jeon, J. H., & Liu, Y. (2009). Automatic prosodic events detection using syllable-based acoustic and syntactic features. In *IEEE international conference on acoustics, speech and signal processing, 2009 (ICASSP 2009)* (pp. 4565–4568). IEEE.

Juslin, P. N., & Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research, 33*(3), 217–238.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System, 38*(2), 301–315.

Kang, O., & Pickering, L. (2013). Using acoustic and temporal analysis for assessing speaking. In A. Kunnan (Ed.), *Companion to language assessment* (pp. 1047–1062). Hoboken: Wiley-Blackwell.

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal, 94*(4), 554–566.

Kang, O., & Wang, L. (2014). Impact of different task types on candidates' speaking performances and interactive features that distinguish between CEFR levels. *ISSN 1756-509X*, 40.

KayPENTAX. (2008). *Multi-speech and CSL software*. Lincoln Park, NJ: KayPENTAX.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America, 118*(2), 1038–1054.

Litman, D. J., Hirschberg, J. B., & Swerts, M. (2000). Predicting automatic speech recognition performance using prosodic cues. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference* (pp. 218–225). Association for Computational Linguistics.

Ludusan, B., & Dupoux, E. (2014). Towards low-resource prosodic boundary detection.

Ludusan, B., Origlia, A., & Cutugno, F. (2011). On the use of the rhythmogram for automatic syllabic prominence detection (pp. 2424–2427). In *INTERSPEECH*.

Mahrt, T., Cole, J., Fleck, M. M., & Hasegawa-Johnson, M. (2012a). F0 and the perception of prominence. In *INTERSPEECH*.

Mahrt, T., Cole, J., Fleck, M., & Hasegawa-Johnson, M. (2012b). Modeling speaker variation in cues to prominence using the Bayesian information criterion. In *Speech prosody 2012*.

Mahrt, T., Huang, J. T., Mo, Y., Fleck, M. M., Hasegawa-Johnson, M., & Cole, J. (2011). Optimal models of prosodic prominence using the Bayesian information criterion (pp. 2037–2040). In *INTERSPEECH*.

MathWorks, Inc. (2013). MATLAB release 2013a. [Computer program]. Retrieved February 15, 2013.

McCann, J., & Peppé, S. (2003). Prosody in autism spectrum disorders: A critical review. *International Journal of Language & Communication Disorders, 38*(4), 325–350.

Nadel, J., Simon, M., Canet, P., Soussignan, R., Blancard, P., Canamero, L., & Gaussier, P. (2006). Human responses to an expressive robot. In *Proceedings of the sixth international workshop on epigenetic robotics*. Lund University.

Ni, C. J., Liu, W., & Xu, B. (2011). Automatic prosodic events detection by using syllable-based acoustic, lexical and syntactic features. In *INTERSPEECH* (pp. 2017–2020).

Ni, C., Liu, W., & Xu, B. (2012). From English pitch accent detection to Mandarin stress detection, where is the difference? *Computer Speech & Language, 26*(3), 127–148.

Obin, N., Rodet, X., & Lacheret-Dujour, A. (2009). A syllable-based prominence detection model based on discriminant analysis and context-dependency. In *SPECOM* (pp. 97–100).

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research, 11*, 169–198.

Ostendorf, M. (1999, December). Moving beyond the 'beads-on-a-string' model of speech. In *Proceedings of IEEE ASRU workshop* (pp. 79–84). Piscataway, NJ: IEEE.

Ostendorf, M., Price, P. J., & Shattuck-Hufnagel, S. (1995). The Boston University radio news corpus. *Linguistic Data Consortium*, 1–19.

Paul, R., Augustyn, A., Klin, A., & Volkmar, F. R. (2005). Perception and production of prosody by speakers with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 35*(2), 205–220.

Pickering, L. (1999). *An analysis of prosodic systems in the classroom discourse of native speaker and nonnative speaker teaching assistants*. Unpublished doctoral dissertation, University of Florida, Gainesville.

Pickering, L. (2009). Intonation as a pragmatic resource in ELF interaction. *Intercultural Pragmatics, 6*(2), 235–255.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation*. Doctoral dissertation, Massachusetts Institute of Technology.

Pierrehumbert, J., & Beckman, M. (1988). Japanese tone structure. *Linguistic Inquiry Monographs, 15*, 1–282.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S., & Veilleux, N. (1988). A methodology for analyzing prosody. *The Journal of the Acoustical Society of America, 84*(S1), S99.

Quinlan, J. R. (1999). Simplifying decision trees. *International Journal of Human-Computer Studies, 51*(2), 497–510.

Rosenberg, A., & Hirschberg, J. (2006). On the correlation between energy and pitch accent in read English speech. In *INTERSPEECH*.

Rosenberg, A., & Hirschberg, J. (2009). Detecting pitch accents at the word, syllable and vowel level. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the Association for Computational Linguistics, Companion Volume: Short Papers* (pp. 81–84). Association for Computational Linguistics.

Rosenberg, A., & Hirschberg, J. B. (2010). Production of English prominence by native mandarin Chinese speakers.

Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., & Stolcke, A. (2005). Modeling prosodic feature sequences for speaker recognition. *Speech Communication, 46*(3), 455–472.

Shriberg, L. D., Paul, R., McSweeny, J. L., Klin, A., Cohen, D. J., & Volkmar, F. R. (2001). Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome. *Journal of Speech, Language, and Hearing Research, 44*(5), 1097–1115.

Silipo, R., & Greenberg, S. (1999). Automatic transcription of prosodic stress for spontaneous English discourse. In *Proceedings of the XIVth international congress of phonetic sciences (ICPhS)* (Vol. 3, p. 2351).

Silipo, R., & Greenberg, S. (2000). Prosodic stress revisited: Reassessing the role of fundamental frequency. In *Proceedings of NIST speech transcription workshop*.

Sridhar, V. R., Bangalore, S., & Narayanan, S. S. (2008). Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(4), 797–811.

Streefkerk, B. M., Pols, L. C., & Ten Bosch, L. F. (1997). Prominence in read aloud sentences, as marked by listeners and classified automatically. In *Proceedings of the Institute of Phonetic Sciences*, University of Amsterdam (Vol. 21, pp. 101–116).

Syrdal, A. K., & McGory, J. T. (2000). Inter-transcriber reliability of ToBI prosodic labeling. In *INTERSPEECH* (pp. 235–238).

Tamburini, F. (2006). Reliable prominence identification in English spontaneous speech. *Proceedings of speech prosody 2006.*

Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *The Journal of the Acoustical Society of America, 89*(4), 1768–1776.

Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. In *Proceedings of the 1992 international conference on spoken language processing, ICSLP* (pp. 12–16).

Xu, Y. (2012). Speech prosody: A methodological review. *Journal of Speech Sciences, 1*(1), 85–115.

Yoon, T., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In *INTERSPEECH*.