Int J Speech Technol (2016) 19:95–109
DOI 10.1007/s10772-015-9327-z

Automatic prosodic tone choice classification with Brazil's intonation model

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

Received: 25 September 2015 / Accepted: 21 November 2015 / Published online: 2 December 2015

Abstract This paper examines the performance of automatically classifying five tone choices (i.e., falling, rising, rising-falling, falling-rising, and neutral) of Brazil's intonation model. We tested two machine learning classifiers (neural network and boosting ensemble) in two configurations (multi-class and pairwise coupling) and a rule-based classifier. Three sets of acoustic features built from the TILT and Bézier pitch contour models and a new four-point pitch contour model we introduced here were investigated. Tone choices are one of the key elements of Brazil's prosodic intonation model. We found the rule-based classifier, which was built on our four-point model, achieved better results than the others with an accuracy of 75.1 % and a Cohen's kappa coefficient of 0.73. This research proves that it is possible to classify tone choices with an accuracy reaching close to the percentage of agreement between two human analysts. The findings further concluded that our four-point model was better for classifying Brazil's tone choices than both of the TILT or Bézier models.

Keywords Tone choice classification · Machine learning · Brazil's prosodic intonation model · ToBI · TILT model · Bézier model

1 Introduction

The pattern of stress and intonation in a language is called prosody. There are many application domains that might benefit from automatic detection of prosody. It can be utilized in text-to-speech synthesis to model intonation for computerized and robot speech. Shriberg et al. (2005) and Escudero-Mancebo et al. (2014) demonstrated that prosodic models improve speaker identification and verification. Irregular prosody is one of the symptoms of autism and other related developmental disorders (Frith and Happé 1994; Fine et al. 1991; Paul et al. 2005; Shriberg et al. 2001; McCann and Peppé 2003). Computer programs that detect irregular prosody automatically have been employed to diagnose autism (Xu et al. 2009; Ringeval et al. 2011; Oller et al. 2010; Diehl & Paul 2012; Van Santen et al. 2010). Suprasegmental measures derived from the elements of Brazil's model have been shown to explain half of the variance in oral proficiency and comprehensibility ratings of non-native speakers (Kang et al. 2010; Kang and Wang 2014). A number of studies have concluded that the inclusion of prosodic elements enhances automatic speech recognition (Bocklet and Shriberg 2009; Hämäläinen et al. 2007; Litman et al. 2000; Ostendorf 1999).

This study examines automatic detection of tone choice which is one of the fundamental elements of Brazil's (1997) model of prosody (see Sect. 2 for further details on Brazil's model). The purpose of this paper is to determine the best machine learning algorithm and the associated acoustic feature set, for classifying tone choice. We analyzed the accuracy and κ of two machine learning classifiers (neural network and boosting ensemble) in two configurations (multi-class and pairwise coupling) and a rule-based classifier. We tested three sets of acoustic features created from the TILT and Bézier models and a new four-point model we have introduced in this paper. Then, we explained how we decided on the classifiers and acoustic feature sets to test. We also described the methods employed to determine the best machine learning algorithm and the best acoustic feature set for classifying the tone

✉ Okim Kang
okim.kang@nau.edu

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

choice of a termination prominent syllable. Finally, after presenting the results, we compared the current findings with those of other research in the field of speech science.

2 Brazil's intonation model

Prosody is described by a variety of speech models. Brazil's (1997) model and Pierrehumbert's (1980) model are two that are used often in the fields of linguistics and applied linguistics. Pierrehumbert's model is often utilized to model prosody for synthesized speech in text-to-speech applications (Wennerstrom 2001). Brazil's model is frequently applied to language teaching (Cauldwell 2012). Using Brazil's model is an innovative aspect of the current study because as far as we know it has not been applied to computational linguistics before. Brazil's model defines pitch concord in an interactive dialog between two persons. Pitch concord matches the relative pitch of the key (first) and termination (last) prominent syllables between two speakers. For instance, high pitch on the termination of one speaker's statement is matched with a high pitch on the key of the next speaker's statement. Likewise, a mid termination is paired with a mid key. Pitch concord is a powerful predictor of speaking proficiency in non-native speakers (Pickering 1999). If we assume the goal of computational linguistics is more human-like speech production and interaction, then it is necessary to explore and adopt a model with a more thorough interpretation of intonation at a discourse (i.e., dialog) level.

The basis of Brazil's theory is the tone unit. Brazil explains a tone unit as a portion of a discourse that a listener can distinguish as having a rising and falling pitch pattern that is distinctive from those of otherwise alike tone units having other patterns of pitch. Every tone unit has one or more prominent syllables, which can be identified from three properties of the syllable: pitch (fundamental frequency in Hz), duration (length in seconds), and intensity (amplitude in dB) (Chun 2002). Brazil asserts (as others have) that the importance of prominence is on the syllable, and not the word. Brazil differentiates prominence from lexical stress. He explains that lexical stress denotes the syllable inside content words that is stressed; however, prominence is the use of emphasis to add more meaning, importance, or contrast to words in a discourse. Accordingly, a syllable that is typically not stressed (e.g., a function word) may be accented to make it prominent. Conversely, a syllable that is customarily stressed lexically may be delivered with additional pitch, duration, or intensity to highlight its meaning, importance, or contrast. Every tone unit contains a key (first) and a termination (last) prominent syllable. If a tone unit has a single prominent syllable, then it is considered to be equally the key and termination prominent syllable. The termination syllable is also referred to as the tonic syllable. The relative pitch of the key and termination prominent syllables and the tone choice of the termination prominent syllable define the tone unit's intonation pattern. Brazil postulated three evenly balanced scales of relative pitch: low, mid, and high, and five tone choices: falling, rising, rising-falling, falling-rising, and neutral as illustrated in Fig. 1.

The Brazil model covers both constrained and unconstrained speech in monologues and dialogs. Thus, the elements of the model (e.g., tone choice) apply equally to all types of speech.

3 Related research

In this section, we will review related research to identify techniques that can be applied to solving the problem of classifying tone choice. Brazil's (1997) model has not been exploited in the field of computational linguistics. However, there is a large body of research on classifying ToBI Pitch Accents and Boundary Tones from which we identified candidate machine learning classifiers and acoustic feature sets for our experiments. The tones and break indices (ToBI) is a system for labeling prosodic events in speech (Wightman et al. 1992; Beckman and Elam 1997). ToBI defines three prosodic events: pitch accents, boundary tones, and break indices. Of these, pitch accents and boundary tones are the most closely related to Brazil's tone choices. Pitch accents serve as cues for prominence, while boundary tones serve as cues for intonational phrasing. Although pitch accents are cues for prominence, there are usually more pitch accents in a dialog than there are Brazil's prominent syllables. Boundary tones match closely with Brazil's concept of key prominent syllables (i.e., initial boundary tones and phrasal tones) and termination prominent syllables (i.e., final boundary tones). ToBI defines eight types of pitch accents and nine types of boundary tones. There is not a one-to-one correspondence between Brazil's tone choices and either pitch accents or boundary tones. Nonetheless, the methods of classifying them and Brazil's tone choices are similar.

We compared several ToBI experiments involving pitch accents and boundary tones based on the accuracy to determine the candidate classifiers and feature sets we utilized in our experiments. We applied three constraints to the experiments we considered: (1) The experiment had to involve multiple speakers because single speaker classification is somewhat trivial and our goal is speaker independent recognition of Brazil's tone choices; (2) the experiment had to classify with only acoustic features; and (3) the experiment had to include five or more classes since there are five tone choices.

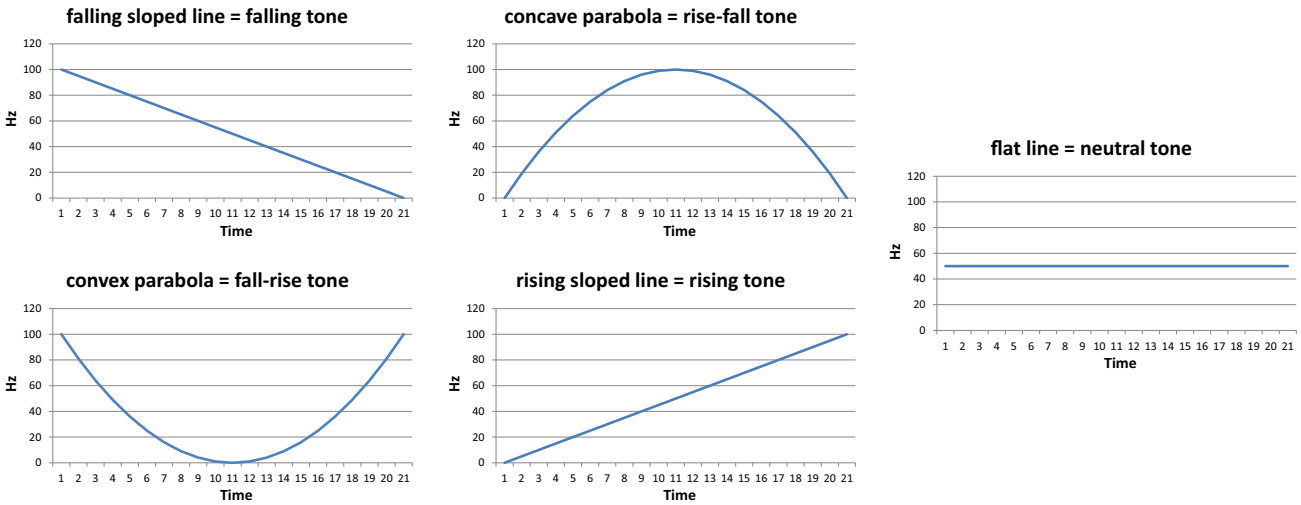

Fig. 1 Brazil's five tone choices

There are three pitch contour models, which have been employed in the ToBI investigations. In the TILT model, intonation is characterized by parameters representing amplitude, duration, and tilt, where tilt is a measure of the shape of the pitch contour (Taylor 2000). The Bézier model is an approximation of pitch contours with Bézier functions (Escudero-Mancebo and Cardeñoso-Payo 2007). The Quantized Contour Model (QCM) (Rosenberg 2010a, b) quantizes the pitch contour of a word in the time and pitch domains, generating a low-dimensional representation of the contour. Each of these models produces a set of acoustic features, which can be classified with machine learning.

Table 1 presents the accuracy of several recent ToBI experiments along with what was classified (pitch accent or boundary tones), the number of classes classified out of the total number of classes, the number of speakers out of the total number of speakers, the pitch contour model, and machine learning classifier. Also indicated is whether the experiment met two of our constraints, i.e., multiple speakers and five or more classes. None of the experiments met our constraint of acoustic features only. All of the experiments made use of the Boston University Radio News Corpus (Ostendorf et al. 1995), except Li et al. (2010). Their corpus data was a set of 20 male and 20 female speakers from an L2 English speech corpus read by native Mandarin speakers. The speakers were asked to read 29 prompted sentences and instructed to read with a rising or falling intonation, according to an indicator next to each sentence.

AuToBI is a tool for automatic ToBI annotation (Rosenberg 2010a, b). Rosenberg reported on the performance of AuToBI in classifying pitch accents and boundary tones utilizing various classifiers and features in 2010 and then again in 2012. In 2010, he described the operation of AuToBI on the Boston Directions Corpus and the Columbia Games Corpus. Utilizing SVMs, AuToBI classified pitch accents of the spontaneous portion of the Boston Directions Corpus with a combined error rate of 0.284, intonational phrase final tones with 55.0 % accuracy, and intermediate phrase ending tones with 68.6 %. He did not give the pitch accent classification results on the Columbia Games Corpus, but stated the intonational phrase final tones were classified with 35.34 % accuracy, whereas intermediate phrase ending phrase accents were classified with 62.21 % accuracy. In 2012, Rosenberg examined a number of features and classifiers to improve the capability of AuToBI to classify pitch accents and boundary tones. He found the AdaBoost classifier implemented with weka did the best at classifying pitch accents (60.91 % accuracy) and that the Random Forest classifier implemented with weka was the best at classifying pitch accent (47.44 %) and pitch accent/boundary tones (74.47 %).

From the experiments that met our constraints, we chose the neural network and decision tree classifiers as candidates for our experiments. We augmented the decision tree classifier with boosting. Boosting is a machine learning ensemble method designed to improve the performance of decision tree classifiers. We did not choose to use a Naïve Bayesian classifier because of all machine learning techniques Naïve Bayesian classifiers are typically the weakest (Caruana and Niculescu-Mizil 2006). We also selected two classification configurations: multi-class and pair-wise coupling. In the multi-class configuration, the classifier makes a 1-of-n choice. Multi-class classifiers generally function worse than binary classifiers. Pairwise coupling is a method of breaking a multiple classification problem into a number of more accurate binary classification problems

Table 1 Summary of several recent ToBI experiments sorted by constraints met and accuracy (Acc)

ToBI experiment	Classified	Classes/total	Speakers/total	Model	Classifier	Acc (%)	Constraint Met	
							Multiple speakers	Five or more classes
González-Ferreras et al. (2012)	Boundary tones	9/9	6/6	Bézier	Pair-wise DT	73.4	Yes	Yes
(Rosenberg 2010a, b)	Boundary tones	5/9	6/6	QCM	Multi-class Bayesian	72.9	Yes	Yes
González-Ferreras et al. (2012)	Boundary tones	9/9	6/6	TILT	Pair-wise DT	72.3	Yes	Yes
Rosenberg (2010a, b)	Pitch accents	5/8	6/6	QCM	Multi-class Bayesian	64.0	Yes	Yes
González-Ferreras et al. (2012)	Pitch accents	8/8	6/6	TILT	Pair-wise DT	57.7	Yes	Yes
González-Ferreras et al. (2012)	Pitch accents	8/8	6/6	Bézier	pair-Wise DT	55.7	Yes	Yes
González-Ferreras et al. (2012)	Boundary tones	9/9	6/6	TILT	Pair-wise NN	55.1	Yes	Yes
González-Ferreras et al. (2012)	Boundary tones	9/9	6/6	Bézier	Pair-wise NN	55.0	Yes	Yes
González-Ferreras et al. (2012)	Pitch accents	8/8	6/6	TILT	pair-Wise NN	48.7	Yes	Yes
González-Ferreras et al. (2012)	Pitch accents	8/8	6/6	Bézier	Pair-wise NN	46.9	Yes	Yes
Ananthakrishnan and Narayanan (2008)	Pitch accents	4/8	6/6	TILT	Multi-class NN	56.4	Yes	No
Ananthakrishnan and Narayanan (2008)	Boundary tones	2/9	6/6	TILT	Multi-class NN	67.7	Yes	No
Li et al. (2010)	Pitch accents	2/8	40/40		Multi-class Bayesian	91.2	Yes	No
Sun (2002)	pitch Accents	4/8	1/6		Boosting ensemble of DT learners	87.2	No	No
Levow (2005)	Pitch accents	4/8	1/6		SVM	81.3	No	No
Ross and Ostendorf (1996)	Pitch accents	4/8	1/6		Multi-class DT	72.4	No	No
Ross and Ostendorf (1996)	Boundary tones	3/9	1/6		Multi-class DT	66.9	No	No

DT decision tree, *NN* neural network, *SVM* support vector machine

(Hastie and Tibshirani 1998). For feature set models, we picked the TILT and Bézier model. We did not select the Quantized Contour Model because the low number of classes in Rosenberg (2010a, b) experiment may have over-inflated the accuracy of it compared with the TILT and Bézier model experiments.

In addition to the candidate classifiers and feature sets that we identified from the ToBI experiments, we also considered another classifier and another pitch contour model. The rule-based classifier is further detailed in Sect. 4.3. The other pitch contour model, which we call the four-point model in this paper, was derived for the rule-based classifier. This pitch contour model is the generalization of any pitch contour, i.e., every pitch contour has a first, last, minimum, and maximum pitch point. Section 4.2.1 contains a more in-depth description of the four-point model.

4 Experimental procedure

In summary, in this paper we have compared the accuracy and κ of two candidate machine learning classifiers (neural network and boosting ensemble) in two configurations (multi-class and pairwise coupling) in automatically classifying the five tone choices of Brazil's intonation model. For each of the four combinations of classifier and configuration, we have considered three sets of features derived from three pitch contour models: TILT, Bézier, and our four-point model. We have also made comparisons of these twelve combinations with our rule-based classifier, which is founded on the four-point model.

4.1 TIMIT corpus

The DARPA TIMIT Acoustic–Phonetic Continuous Speech Corpus (TIMIT) of read speech provides 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.

Table 2 Distribution of TIMIT speakers in this research by gender and dialect

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

The TIMIT corpus is composed of constrained (i.e., short read sentences) monologic speech. We chose the TIMIT corpus over others (e.g., Boston University Radio News Corpus) because of the large number of speakers and dialects spoken.

The TIMIT corpus includes definitions for 60 phones. The TIMIT phones are used by other corpora. For our experiments, we utilized a subset of the corpus consisting of 84 speakers speaking four dialects. There were 825 utterances in our subset containing 10,512 syllables of which 994 of those were terminating prominent syllables. Table 2 presents the distribution of speakers by gender and dialect.

We augmented the corpus by identifying the prominent syllables and tone choices on the termination (last) prominent syllables in the experimental subset using the syllable demarcations provided with the corpus. The prominent syllables and tone choices were identified by a trained linguist who coded them both by listening to the audio files and by using Praat (Boersma and Weenink 2014), a computerized speech analysis program, to confirm the movement of the pitch contour. Approximately, ten percent of the samples were analyzed by a second trained linguist to confirm the consistency of the coding. The inter-rater reliability between the two linguists was 85 to 87 %, which is a satisfactory rate found in other similar studies (e.g., Kang 2010) utilizing Brazil's (1997) prosody model. The two linguists revised any discrepancies and continued coding the data until there were no more discrepancies. The first linguist then finished coding the rest of the speech files alone. This method of annotation has been employed as a reliable labeling technique extensively in other applied linguistics studies (Kang et al. 2010; Kang and Wang 2014; Pickering 1999). The linguist identified the tone choice of 994 terminating prominent syllables in the speech samples. The distribution of tone choices is depicted in Table 3.

Table 3 Distribution of tone choices

Tone Choice	Count
Fall	432
Rise	141
Rise-Fall	50
Fall-Rise	37
Neutral	334

Initially the analysts examined the pitch contours with the Multi-Speech and Computerized Speech Laboratory (CSL) Software (KayPENTAX 2008), while the computer analyzed them using Praat (Boersma and Weenink 2014). We discovered significant differences between the pitch contours displayed by the two. This discrepancy resulted in substantial disagreement between the tone choices classified by the computer and those by the human expert. In addition, further differences in pitch contour were found even between various versions of Praat. More differences were identified between the same versions of Praat running on different computers. Maryn et al. (2009) also reported this difference among Multi-Speech and CSL Software and Praat. They declared that pitch and intensity values were not comparable. Amir et al. (2009) also noted this discrepancy and added that the findings from Multi-Speech and CSL Software and Praat should not be combined. To ensure these variations did not affect our results, the analyst re-conducted the tone choice annotations that were utilized to train the classifiers using the same version of Praat of which the computer made use.

4.2 Parameterization of the F0 contours

We investigated three sets of classification features, each derived from a different model of the pitch contour: four-point model, TILT model (Taylor 2000), and a model proposed by Escudero-Mancebo and Cardeñoso-Payo (2007), which consists of Bézier parameters. The pitch contour was extracted with Praat (Boersma and Weenink 2014).

4.2.1 Four-point model features

The four-point model is of our own design, which we are proposing here. The four-point model has two sub-models as depicted in Fig. 2: rise-fall-rise and fall-rise-fall.

The rise-fall-rise sub-model is applied if the maximum pitch point is earlier in time than the minimum pitch point; the fall-rise-fall sub-model is applied if the minimum pitch point is earlier in time than the maximum pitch time. The features for the sub-models are built on the following four points (from which the model derives its name): *first* is the pitch of the first point in the pitch contour (Hz); *last* is the pitch of the last point in the pitch contour (Hz); *max* is the maximum pitch in the pitch contour (Hz); and *min* is the minimum pitch in the pitch contour (Hz). The features for the rise-fall-rise sub-model are first-rise (*r1*), first-fall (*f1*), and second-rise (*r2*) and they are calculated as follows:

$$r1 = max - first \tag{1}$$

$$f1 = max - min \tag{2}$$

$$r2 = last - min \tag{3}$$

The features for the fall-rise-fall sub-model are first-fall ($f1$), first-rise ($r1$), and second-fall ($f2$) and they are calculated as follows:

$$f1 = first - min \tag{4}$$

$$r1 = max - min \tag{5}$$

$$f2 = max - last \tag{6}$$

We apply this model because it is the generalization of any pitch contour; i.e., every pitch contour has a first, last, minimum, and maximum pitch point. In some cases, some or all of the four points might coincide. For example, the maximum may also be the first point. Theoretically, the classifiers should determine the tone choice by the significance or insignificance of the rises and falls. The significance of the rises and falls is determined during the classifier training. For instance, as depicted in Fig. 3, in the

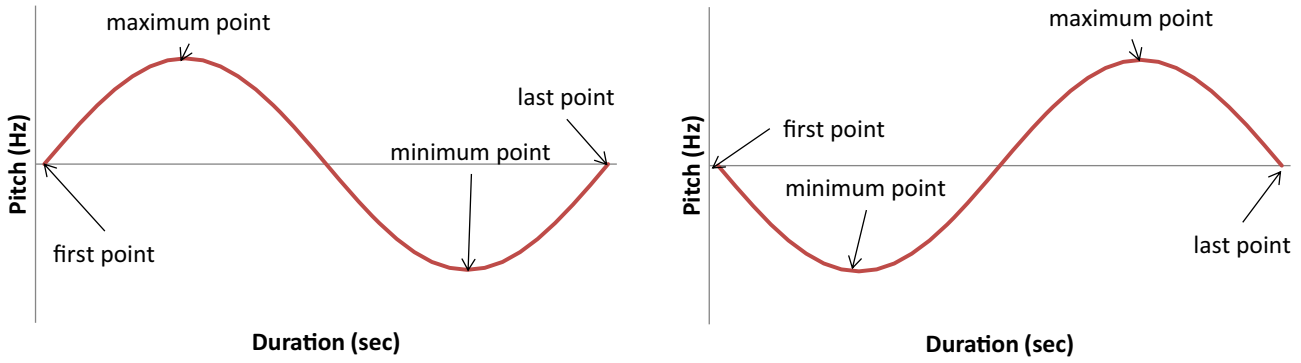

Fig. 2 Four-point model sub-models: rise-fall-rise (*left*) and fall-rise-fall (*right*)

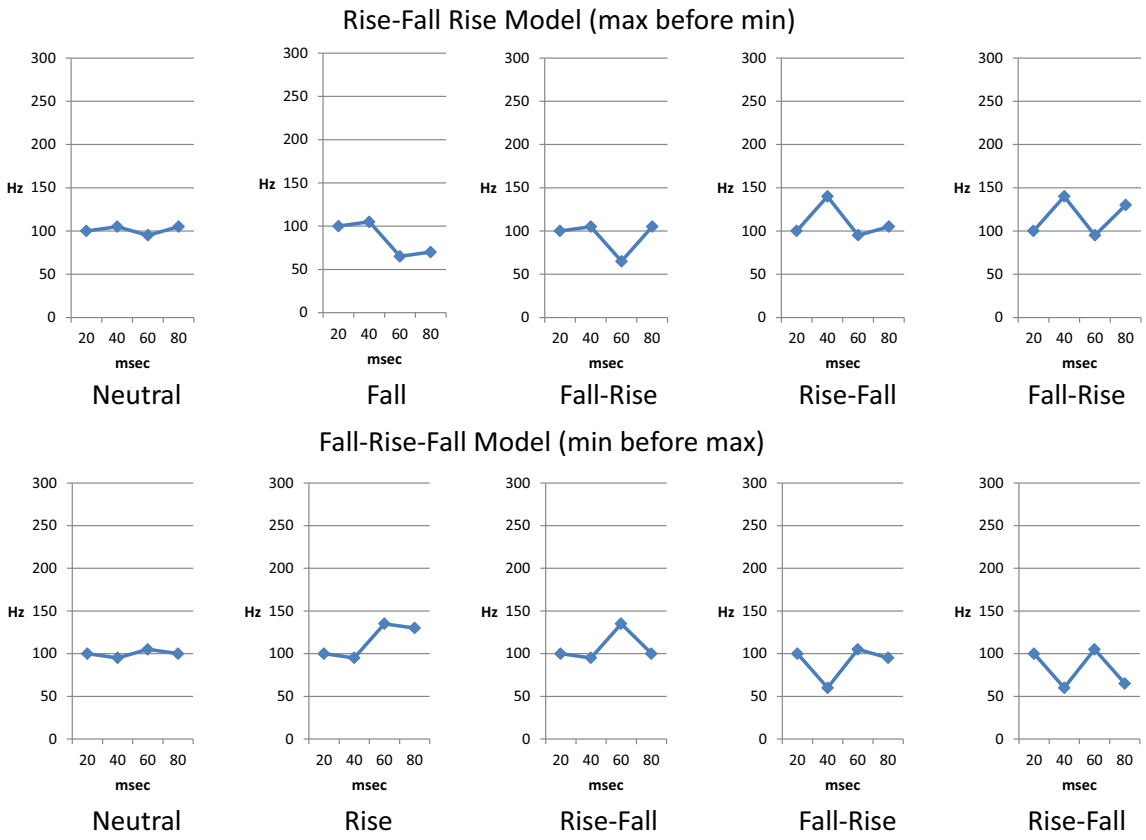

Fig. 3 Examples of how the significance of the rises and falls determines the tone choice

Table 4 Truth table for all possible combinations of significant and insignificant rise and falls; 0 = rise/fall is insignificant (i.e., it is less than a threshold); 1 = rise/fall is significant (i.e., it is more than a threshold)

Rise-fall-rise sub-model				Fall-rise-fall sub-model			
r1	*f1*	*r2*	Tone Choice	*f1*	*r1*	*f2*	Tone Choice
0	0	0	Neutral	0	0	0	Neutral
0	0	1	Rise	0	0	1	Fall
0	1	0	Fall	0	1	0	Rise
0	1	1	Fall-rise	0	1	1	Rise-fall
1	0	0	Rise	1	0	0	Fall
1	0	1	Rise	1	0	1	Fall
1	1	0	Rise-fall	1	1	0	Fall-rise
1	1	1	Fall-rise	1	1	1	Rise-fall

rise-fall-rise sub-model, if all the rises and falls are insignificant, then the tone choice is neutral. If *r*1 is insignificant, *f*1 is significant, and *r*2 is insignificant, the tone choice is fall.

Table 4 specifies the truth table for all possible combinations of significant and insignificant rise and falls, which are illustrated in Fig. 3. In the last row of Table 4, all of the rises and falls are significant so the tone choice could be either fall-rise or rise-fall. In these two cases, the tone choice of the last two significant rise and fall was applied, i.e., *f*1 and *r*2 for the rise-fall-rise sub-model and *r*1 and *f*2 for the fall-rise-fall sub-model.

4.2.2 *TILT model features*

TILT is one of the more popular models for parameterizing pitch contours (Taylor 2000). The model was developed to automatically analyze and synthesize speech intonation. In the model, intonation is represented as a sequence of events, which are characterized by parameters representing amplitude, duration, and tilt. Tilt is a measure of the shape of the event, or pitch contour. A popular public domain text-to-speech system, Festival (The Centre for Speech Technology Research 2014) applies this model to synthesize speech intonation. The model is illustrated in Fig. 4. Three points are defined: start of the event, the peak (the highest point), and the end of the event.

Each event is characterized by five RFC (rise/fall/connection) parameters: rise amplitude (difference in pitch between the pitch value at the peak and at the start, which

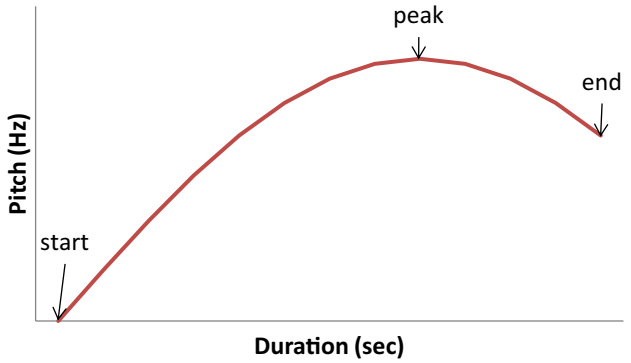

Fig. 4 Parameters of the RFC model in the TILT model of a pitch contour

is always greater than or equal to zero), rise duration (distance in time from start of the event to the peak), fall amplitude (pitch distance from the end to the peak, which is always less than or equal to zero), fall duration (distance in time from the peak to the end), and vowel position (distance in time from start of pitch contour to start of vowel). The TILT representation transforms four of the RFC parameters into three TILT parameters: duration (sum of the rise and fall durations), amplitude (sum of absolute values of the rise and fall amplitudes), and tilt (a dimensionless number which expresses the overall shape of the event). The TILT parameters are calculated as follows:

$$s = \text{start of event} \tag{7}$$

$$p = \text{peak (the highest point)} \tag{8}$$

$$e = \text{end of event} \tag{9}$$

$$a_{rise} = \text{difference in pitch between the pitch value at the peak } (p) \text{ and at the start } (s),\ \geq 0 \tag{10}$$

$$d_{rise} = \text{distance in time from start}(s)\text{of the event to the peak}(p) \tag{11}$$

$$a_{fall} = \text{pitch distance from the end } (e) \text{ to the peak } (p),\ \leq 0 \tag{12}$$

$$d_{fall} = \text{distance in time from the peak } (p) \text{ to the end } (e) \tag{13}$$

$$d = \text{duration} = d_{rise} + d_{fall} \tag{14}$$

$$a = \text{amplitude} = |a_{rise}| + |a_{fall}| \tag{15}$$

$$t = \text{tilt} = \frac{|a_{rise}| - |a_{fall}|}{2(|a_{rise}| + |a_{fall}|)} + \frac{d_{rise} - d_{fall}}{2(d_{rise} + d_{fall})} \tag{16}$$

$$c = \text{pitch contour} = \{f_i, f_{i+1}, \ldots, f_{i+N}\} \tag{17}$$

$$f_i = \text{frequency (Hz) of } i\text{th point in pitch contour} \tag{18}$$

$$f_v \in c \tag{19}$$

$$v = \text{index of the beginning of the vowel} \tag{20}$$

In our experiments, duration (d), amplitude (a), tilt (t), and vowel position (v) were the input features to the classifiers.

4.2.3 *Bézier model features*

Escudero-Mancebo and Cardeñoso-Payo (2007) proposed an alternative to the TILT model that is constructed from the approximation of the pitch contours with Bézier functions as illustrated in Fig. 5.

Similarly we used Bézier functions to approximate the pitch contour of the terminating prominent syllable, where:

$$\mathbf{P} = \text{pitch contour} \tag{21}$$

$$\mathbf{P}_i = (f_i, t_i) = \text{F0 (Hz) at time (s)} t_i \tag{22}$$

$$n = |\mathbf{P}| - 1 \tag{23}$$

$$b = \text{number of Bézier points} = 4 \tag{24}$$

$$x = \left(0, \frac{1}{b-1}, \frac{2}{b-1}, 1\right) \tag{25}$$

$$j = (1, 2, 3, 4) \tag{26}$$

$$\mathbf{B}(x_j) = (p_j, x_j) \tag{27}$$

$$p_j = \text{Bézier approximation of F0 (Hz) at time (s)} x_j \tag{28}$$

$$\mathbf{B}(x) = \sum_{i=0}^{n} b_{i,n}(x)\mathbf{P}_i, \quad 0 \leq x \leq 1 \tag{29}$$

$$b_{i,n}(x) = \binom{n}{i} x_i (1-x)^{n-i} \quad i = 0, \ldots, n \tag{30}$$

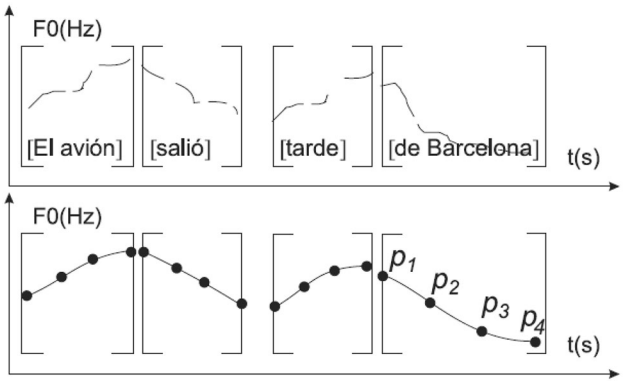

Fig. 5 Example of the Bézier function fitting stylization from Escudero-Mancebo and Cardeñoso-Payo (2007)

The resulting four Bézier parameters (p_1, p_2, p_3, and p_4) are the features on which the tone choice classifiers are trained.

4.3 Classifiers

We tested two standard machine learning classifiers to classify tone choices: neural network and boosting. We employed the Matlab *patternnet* function with ten hidden nodes and the Levenberg–Marquardt optimization network training function to implement the neural network classifier (MathWorks 2013). Boosting is an ensemble classifier that combines the outcomes of weak classifiers (typically decision trees) to improve their accuracy. Boosting was implemented with the Matlab *fitensemble* function using the AdaBoostM1 (binary classifier) or AdaBoostM2 (multi-class classifier) booster and 100 decision tree learners (i.e., weak classifiers).

We also utilized a rule-based classifier that implemented the four-point model truth table specified in Table 4 above. The thresholds for significance versus insignificance of each rise and fall (i.e., rise-fall-rise sub-model: $r1$, $f1$, and $r2$; fall-rise-rise sub-model: $f1$, $r1$, and $f2$) were determined during training. A simple brute-force method of trying every combination of unique rises and falls in the training data as thresholds determined the set of thresholds (TH^*_{rfr}, TH^*_{frf}) that maximized the accuracy as follows:

$$TC = \{1,2,3,4,5\} \text{ corresponding to tone choices } \{\text{rise, neutral, fall, fall-rise, rise-fall}\} \quad (31)$$

$$T = \{\text{human classified tone choices for training data}\} \quad (32)$$

$$T_{rfr} = \{\text{human classified tone choices for training data for rise-fall-rise sub-model}\} \quad (33)$$

$$T_{frf} = \{\text{human classified tone choices for training data for fall-rise-fall sub-model}\} \quad (34)$$

$$T = T_{rfr} \cup T_{frf} \quad (35)$$

$$\emptyset = T_{rfr} \cap T_{frf} \quad (36)$$

$$N = \left|T_{rfr}\right| \quad (37)$$

$$M = \left|T_{frf}\right| \quad (38)$$

$$T_{rfr} = \{t_1, \ldots, t_N\} \quad (39)$$

$$T_{frf} = \{t_1, \ldots, t_M\} \quad (40)$$

$$t_i \in TC \quad (41)$$

$$tr1_i = r1 \text{ for } i\text{-th training sample} \quad (42)$$

$$tf1_i = f1 \text{ for } i\text{-th training sample} \quad (43)$$

$$tr2_i = r2 \text{ for } i\text{-th training sample} \quad (44)$$

$$tf2_i = f2 \text{ for } i\text{-th training sample} \quad (45)$$

$$F_{rfr} = \{(tr1_1, tf1_1, tr2_1), \ldots, (tr1_N, tf1_N, tr2_N)\} \quad (46)$$

$$F_{frf} = \{(tf1_1, tr1_1, tf2_1), \ldots, (tf1_M, tr1_M, tf2_M)\} \quad (47)$$

$$\emptyset = F_{rfr} \cap F_{frf} \quad (48)$$

$$r1_{rfr} = \{tr1_1, \ldots, tr1_N\} \quad (49)$$

$$f1_{rfr} = \{tf1_1, \ldots, tf1_N\} \quad (50)$$

$$r2_{rfr} = \{tr2_1, \ldots, tr2_N\} \quad (51)$$

$$R1_{rfr} = !\exists r1_{rfr} \quad (52)$$

$$F1_{rfr} = !\exists f1_{rfr} \quad (53)$$

$$R2_{rfr} = !\exists r2_{rfr} \quad (54)$$

$$I = \left|R1_{rfr}\right| \quad (55)$$

$$J = \left|F1_{rfr}\right| \quad (56)$$

$$K = \left|R2_{rfr}\right| \quad (57)$$

$$r1_i \in R1_{rfr} \quad (58)$$

$$f1_j \in F1_{rfr} \quad (59)$$

$$r2_k \in R2_{rfr} \quad (60)$$

$$\lambda_{rfr}\left(F_{rfr}, \left(r1_i, f1_j, r2_k\right)\right) = \text{rule-based classifier applying rise-fall-rise sub-model of Table 4} \quad (61)$$

$$\lambda_{rfr}\left(F_{rfr}, \left(r1_i, f1_j, r2_k\right)\right) \in TC \quad (62)$$

$$f1_{frf} = \{tf1_1, \ldots, tf1_M\} \quad (63)$$

$$r1_{frf} = \{tr1_1, \ldots, tr1_M\} \quad (64)$$

$$f2_{frf} = \{tr2_1, \ldots, tr2_M\} \quad (65)$$

$$F1_{frf} = !\exists f1_{frf} \quad (66)$$

$$R1_{frf} = !\exists r1_{frf} \quad (67)$$

$$F2_{frf} = !\exists f2_{frf} \quad (68)$$

$$F = \left|F1_{frf}\right| \quad (69)$$

$$G = \left|R1_{frf}\right| \quad (70)$$

$$H = \left|F2_{frf}\right| \quad (71)$$

$$f1_f \in F1_{frf} \quad (72)$$

$$r1_g \in R1_{frf} \quad (73)$$

$$f2_h \in F2_{frf} \quad (74)$$

$$\lambda_{frf}\left(F_{frf}, \left(f1_f, r1_g, f2_h\right)\right) = \text{rule-based classifier applying fall-rise-fall sub-model of Table 4} \quad (75)$$

$$\lambda_{frf}\left(F_{frf}, \left(f1_f, r1_g, f2_h\right)\right) \in TC \quad (76)$$

$$X = \{\text{rule-based classifier tone choices for training samples for a sub-model}\} \quad (77)$$

$$Y = \{\text{human tone choices for training samples for a sub-model}\} \tag{78}$$

$$Z = |X| = |Y| \tag{79}$$

$$x_i \in X \tag{80}$$

$$y_i \in Ys \tag{81}$$

$$a_i = \begin{cases} 1, & x_i = y_i \\ 0, & x_i \neq y_i \end{cases} \tag{82}$$

$$A(X,Y) = Accuracy = \frac{\sum_{i=1}^{Z} a_i}{Z} \tag{83}$$

$$TH^*_{rfr} = \arg\max_{1 \le i \le I, 1 \le j \le J, 1 \le k \le K} \left[A\left(\lambda_{rfr}\left(F_{rfr}, \left(r1_i, f1_j, r2_k\right)\right), T_{rfr}\right)\right] \tag{84}$$

$$TH^*_{frf} = \arg\max_{1 \le f \le F, 1 \le g \le G, 1 \le h \le H} \left[A\left(\lambda_{frf}\left(F_{frf}, \left(f1_f, r1_g, f2_h\right)\right), T_{frf}\right)\right] \tag{85}$$

4.4 Classifier configurations

We analyzed two different configurations of the neural network and boosting ensemble classifiers: multi-class and pairwise coupling. We employed fivefold cross-validation in each of the experiments to tune the parameters of the machine learning classifiers (i.e., training) and then test them. The method for determining the classifier outputs is described below for each combination of classifier and configuration.

4.4.1 Neural network multi-class

The multi-class neural network provides five outputs, one for each of the possible tone choices. The outputs are real numbers in the range from zero to one. The output with the highest value is selected as the tone choice. There is one multi-class neural network for the TILT model; one for the Bézier; and one for each of the four-point model sub-models.

4.4.2 Boosting ensemble multi-class

The multi-class ensemble provides one output, which is from the set {1, 2, 3, 4, 5} corresponding to the set of tone choices {rise, neutral, fall, fall-rise, rise-fall}. There are four multi-class ensembles; one for the TILT model; one for the Bézier; and one for each of the two four-point model sub-models.

4.4.3 Neural network pairwise coupling

The neural network pairwise coupling configuration consists of ten neural networks trained to classify each combination of tone choices: rise versus neutral, rise versus fall, rise versus fall-rise, rise versus rise-fall, neutral versus fall, neutral versus fall-rise, neutral versus rise-fall, fall versus fall-rise, fall versus rise-fall, and fall-rise versus rise-fall. There are ten neural networks for the TILT model; ten for the Bézier; and ten for each of the four-point model sub-models. The output of each classifier is a real number between zero and one. The outputs are treated as probabilities. The probabilities are combined as follows and the one with the highest probability is the tone choice selected.

$$T = \{1,2,3,4,5\} \text{ corresponding to tone choices } \{\text{rise, neutral, fall, fall-rise, rise-fall}\} \tag{86}$$

$$t \in T \tag{87}$$

$$o_{i,j} = \text{output of classifier trained to classify tone choice } i \text{ vs } j \tag{88}$$

$$o_{i,j} \in \mathbb{R} \tag{89}$$

$$0 \le o_{i,j} \le 1 \tag{90}$$

$$p_1 = Pr(\text{tone choice} = \text{rise}) = o_{r,n} \cdot o_{r,f} \cdot o_{r,fr} \cdot o_{r,rf} \tag{91}$$

$$p_2 = Pr(\text{tone choice} = \text{neutral}) = o_{n,f} \cdot o_{n,fr} \cdot o_{n,rf} \cdot \left(1 - o_{r,n}\right) \tag{92}$$

$$p_3 = Pr(\text{tone choice} = \text{fall}) = o_{f,fr} \cdot o_{f,rf} \cdot \left(1 - o_{r,f}\right) \cdot \left(1 - o_{n,f}\right) \tag{93}$$

$$p_4 = Pr(\text{tone choice} = \text{fall - rise}) = o_{fr,rf} \cdot \left(1 - o_{r,fr}\right) \cdot \left(1 - o_{n,fr}\right) \cdot \left(1 - o_{f,fr}\right) \tag{94}$$

$$p_5 = Pr(\text{tone choice} = \text{rise - fall}) = \left(1 - o_{fr,rf}\right) \cdot \left(1 - o_{r,rf}\right) \cdot \left(1 - o_{n,rf}\right) \cdot \left(1 - o_{f,rf}\right) \tag{95}$$

$$t* = \arg\max_{1 \le t \le 5} p_t. \tag{96}$$

4.4.4 Boosting ensemble spairwise coupling

The boosting ensemble pairwise coupling configuration consists of ten boosting ensembles trained to classify each combination of tone choices: rise versus neutral, rise versus fall, rise versus fall-rise, rise versus rise-fall, neutral versus fall, neutral versus fall-rise, neutral versus rise-fall, fall versus fall-rise, fall versus rise-fall, and fall-rise versus rise-fall. There are ten ensembles for the TILT model; ten for the Bézier; and ten for each of the four-point model sub-models. The output of each classifier is from the set {1, 2, 3, 4, 5} corresponding to the set of tone choices {rise, neutral, fall, fall-rise, rise-fall}. For example, the output of the rise versus neutral classifier would be either 1 or 2. The accuracy of the classifier classifying the training data correctly is treated as the probability that the classifier output

is correct. The probabilities are combined as follows and the one with the highest probability is the tone choice selected.

$$T = \{1,2,3,4,5\} \text{ corresponding to tone choices } \{\text{rise, neutral, fall, fall-rise, rise-fall}\} \quad (97)$$

$$o_{i,j} = \text{ output of classifier trained to classify tone choice } i \text{ vs } j \quad (98)$$

$$o_{i,j} \in T \quad (99)$$

$$a_{i,j} = \text{accuracy of classifier trained to classify tone choice } i \text{ vs. } j \quad (100)$$

$$X_{i,j} = \{\text{tone choices for training samples from classifier classifying tone choice } i \text{ vs } j\} \quad (101)$$

$$Y_{i,j} = \{\text{tone choices for training samples from human classifying tone choice } i \text{ vs } j\} \quad (102)$$

$$Z_{i,j} = \left|X_{i,j}\right| = \left|Y_{i,j}\right| \quad (103)$$

$$x_i \in X_{i,j} \quad (104)$$

$$y_i \in Y_{i,j} \quad (105)$$

$$b_i = \begin{cases} 1, & x_i = y_i \\ 0, & x_i \neq y_i \end{cases} \quad (106)$$

$$a_{i,j} = \frac{\sum_{i=1}^{Z} b_i}{Z} \quad (107)$$

$$t \in T \quad (108)$$

$$p(o_{i,j}, a_{i,j}, t) = \begin{cases} a_{i,j}, & o_{i,j} = t \\ 1 - a_{i,j}, & o_{i,j} \neq t \end{cases} \quad (109)$$

$$\begin{aligned} p_1 &= Pr(\text{tone choice} = \text{rise}) \\ &= p\left(o_{r,n}, a_{r,n}, 1\right) \cdot p\left(o_{r,f}, a_{r,f}, 1\right) \cdot p\left(o_{r,fr}, a_{r,fr}, 1\right) \\ &\quad \cdot p\left(o_{r,rf}, a_{r,rf}, 1\right) \end{aligned} \quad (110)$$

$$\begin{aligned} p_2 &= Pr(\text{tone choice} = \text{neutral}) \\ &= p\left(o_{r,n}, a_{r,n}, 2\right) \cdot p\left(o_{n,f}, a_{n,f}, 2\right) \cdot p\left(o_{n,fr}, a_{n,fr}, 2\right) \\ &\quad \cdot p\left(o_{n,rf}, a_{n,rf}, 2\right) \end{aligned} \quad (111)$$

$$\begin{aligned} p_3 &= Pr(\text{tone choice} = \text{fall}) \\ &= p\left(o_{r,f}, a_{r,f}, 3\right) \cdot p\left(o_{n,f}, a_{n,f}, 3\right) \cdot p\left(o_{f,fr}, a_{f,fr}, 3\right) \\ &\quad \cdot p\left(o_{f,rf}, a_{f,rf}, 3\right) \end{aligned} \quad (112)$$

$$\begin{aligned} p_4 &= Pr(\text{tone choice} = \text{fall-rise}) \\ &= p\left(o_{r,fr}, a_{r,fr}, 4\right) \cdot p\left(o_{n,fr}, a_{n,fr}, 4\right) \cdot p\left(o_{f,fr}, a_{f,fr}, 4\right) \\ &\quad \cdot p\left(o_{fr,rf}, a_{fr,rf}, 4\right) \end{aligned} \quad (113)$$

$$\begin{aligned} p_5 &= Pr(\text{tone choice} = \text{rise-fall}) \\ &= p\left(o_{r,rf}, a_{r,rf}, 5\right) \cdot p\left(o_{n,rf}, a_{n,rf}, 5\right) \\ &\quad \cdot p\left(o_{f,rf}, a_{f,rf}, 5\right) \cdot p\left(o_{fr,rf}, a_{fr,rf}, 5\right) \end{aligned} \quad (114)$$

$$t^* = \arg\max_{1 \le t \le 5} p_t \quad (115)$$

4.5 Experimental design

We employed fivefold cross-validation in each of the experiments. The 84 speakers were randomly allocated to folds. Speakers were randomly allotted to folds instead of the utterances to guarantee that training and testing on the identical speaker did not prejudice the trials. Thirteen experiments were conducted: one for each combination of the two classifiers (neural network and boosting ensemble), two configurations (multi-class and pairwise coupling), and three sets of features (from the TILT, Bézier, four-point models), plus one experiment for the rule-based classifier.

5 Results

In 13 experimental setups, we examined the performance of combinations of three classifiers in two configurations and three sets of features in automatically classifying the tone choice of a termination prominent syllable. We calculated accuracy and Cohen's kappa coefficient (κ) (Cohen 1960) to evaluate the thirteen approaches of classifying tone choice. Accuracy is calculated as follows:

$$H_{test} = \{\text{human tone choices for test samples}\} \quad (116)$$

$$M_{test} = \{\text{machine tone choices for test samples}\} \quad (117)$$

$$N = \left|M_{test}\right| = \left|H_{test}\right| \quad (118)$$

$$h_i \in H \quad (119)$$

$$m_i \in M \quad (120)$$

$$a_i = \begin{cases} 1, & m_i = h_i \\ 0, & m_i \neq h_i \end{cases} \quad (121)$$

$$Accuracy = \frac{\sum_{i=1}^{N} a_i}{N} \quad (122)$$

Cohen's kappa coefficient (κ) is calculated as follows:

$$Pr(a) = \text{ relative observed agreement between human and machine } = Accuracy \quad (123)$$

$$T = \{1,2,3,4,5\} \text{ corresponding to tone choices } \{\text{rise, neutral, fall, fall-rise, rise-fall}\} \quad (124)$$

$$t \in T \tag{125}$$

$$b_{t,i} = \begin{cases} 1, & h_i = t \\ 0, & h_i \neq t \end{cases} \tag{126}$$

$$c_{t,i} = \begin{cases} 1, & m_i = t \\ 0, & m_i \neq t \end{cases} \tag{127}$$

$$Pr(h_i = t) = \frac{\sum_{i=1}^{N} b_{t,i}}{N} \tag{128}$$

$$Pr(m_i = t) = \frac{\sum_{i=1}^{N} c_{t,i}}{N} \tag{129}$$

$$Pr(e) = \text{probability of chance agreement between human and machine} \tag{130}$$

$$Pr(e) = \prod_{t=1,i=1}^{5,N} b_{t,i} \prod_{t=1,i=1}^{5,N} c_{t,i} + \prod_{t=1,i=1}^{5,N} (1 - b_{t,i}) \prod_{t=1,i=1}^{5,N} (1 - c_{t,i}) \tag{131}$$

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} \tag{132}$$

Table 5 displays the accuracy and Cohen's kappa coefficient (κ) of the three feature models: four-point, TILT, and Bézier; using two classifiers: neural network and boosting; in two configurations: multi-class and pairwise coupling. It also presents these metrics for the rule-based classifier. The accuracy and Cohen's kappa coefficient (κ) are the mean of the five folds.

The rule-based classifier, which is built on our four-point model, classified better than the others with an accuracy of 75.1 % and a Cohen's kappa coefficient of 0.73 (bolded in Table 5). We believe this happened because the four-point model, on which the rule-based classifier is founded, is a more general model of pitch contour than either the TILT or Bézier models. Our initial hypothesis was that a more general model was needed to model the more complex pitch contours of Brazil's tone choices.

From a model perspective, our four-point model was the best with a mean classifier accuracy of 74.1 % and a mean classifier κ of 0.71, followed by the Bézier model (71.0 %, 0.68) and the TILT model (67.4 %, 0.65). The TILT model may have functioned poorly because it did not account for Brazil's fall-rise tone choice. From a machine learning classifier point of view, the boosting ensemble was better than the neural network with a mean classifier accuracy of 71.6 versus 69.9 % and a mean κ of 0.69 versus 0.67.

The findings of the multi-class configuration versus the pairwise coupling configuration were mixed. The multi-class configuration worked better for the neural network in all models. It also achieved better results with the multi-class boosting ensemble when our four-point model was employed. However, the pairwise coupling configuration improved more in terms of accuracy and κ than the multi-class configuration for the other two models with the boosting ensemble.

6 Discussion

The study evaluated two machine learning classifiers (i.e., neural network and boosting ensemble) in two configurations (i.e., multi-class and pairwise coupling) in automatically classifying the five tone choices of Brazil's intonation model. For each of the four combinations of classifier and configuration, we considered three sets of features drawn from three pitch contour models: TILT, Bézier, and our four-point model. We have also compared these twelve combinations with our rule-based classifier which is established on the four-point model. We assessed the performance in terms of accuracy and Cohen's kappa coefficient.

The outcomes of our study provide evidence that a computer can classify tone choices of terminating

Table 5 Accuracy and Cohen's kappa coefficient (κ) for different feature models, classifiers, and configurations

Feature Model	Classifier	Configuration	Accuracy (%)	κ
Four-point	Rule-based		**75.1**	**0.73**
	Neural network	Multi-class	74.0	0.72
		Pairwise coupling	72.4	0.70
	Boosting	Multi-class	74.8	0.70
		Pairwise coupling	74.0	0.72
TILT	Neural Network	Multi-class	66.1	0.63
		Pairwise coupling	64.1	0.61
	Boosting	Multi-class	68.6	0.66
		Pairwise coupling	70.7	0.68
Bézier	Neural Network	Multi-class	72.5	0.70
		Pairwise coupling	70.0	0.67
	Boosting	Multi-class	69.9	0.67
		Pairwise coupling	71.7	0.69

prominent syllables with an accuracy of 75.1 % and a κ of 0.73 when compared with a human expert. There is no other research on classifying Brazil's tone choices automatically to make a comparison at the current stage. Thus, our work sets the standard for future efforts.

At the same time, the agreement between a computer and a human found in our study can be compared with the inter-rater agreement between two humans. A common inter-rater agreement measure is 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 investigations. The Rhythm and Pitch (RaP) 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), based on a tone interval theory proposed by Dilley (2005). In our experiments, as can be seen in Table 5, κ was generally higher than this, ranging from 0.61 to 0.73. Cross-corpora comparisons are dubious, but in this case we are comparing the human annotation of corpora using two different models of prosody, ToBI and RaP, with our computer annotation using the Brazil model. Although not conclusive, it does show that our computer annotation is in the range of inter-rater agreement between two humans.

Our study can also be contrasted with other research from the perspective of models, classifiers, and configuration. From a model view point, our four-point model functioned the most successfully, followed by the Bézier model, and the TILT model. The Bézier model performed better than the TILT model in other studies, too (Escudero-Mancebo and Cardeñoso-Payo 2007; González-Ferreras et al. 2012). From the perspective of a machine learning classifier, the boosting ensemble classifies tone choices better than the neural network. González-Ferreras et al. (2012) also support this view that the boosting ensemble is better than a neural network for classifying ToBI boundary tones and pitch accents. Unlike our mixed findings of the multi-class configuration versus the pairwise coupling configuration, after testing the TILT and Bézier models, González-Ferreras et al. (2012) reported that pairwise coupling is better at classifying ToBI boundary tones and pitch accents than multi-class in every case.

7 Conclusions

These experiments assessed the performance, in terms of accuracy and Cohen's kappa, of two machine learning classifiers (i.e., neural network and boosting ensemble) in two configurations (i.e., multi-class and pairwise coupling) of classifying the five tone choices of Brazil's intonation model with three sets of features extracted from three pitch contour models: TILT, Bézier, and our four-point model. These twelve combinations of classifiers, configurations, and feature sets were also contrasted with our rule-based classifier which is founded on the four-point model.

The findings reported in this paper offer empirical evidence that a computer can classify terminating prominent syllable tone choices specified in Brazil's (1997) model of intonation with an accuracy approaching that of two human analysts. They also demonstrate that our four-point model is a better one for Brazil's tone choices than either the TILT or Bézier model. Automatic classification of tone choices is an important achievement because tone choices are one of the key elements of Brazil's model. Brazil's model deals with the intonational and rhythmic aspects of speech and explains how they convey meaning that goes beyond what the sentences communicate (Brazil 1997). Accordingly, automatically classifying tone choices is another vital step in automatically deducing the intonational and rhythmic facets of speech.

Examining other classifiers (e.g., linear classifiers, support vector machines, lazy learning algorithms, random forests, meta-algorithms) as a means of improving tone choice classification is an area for further study. Since TIMIT is only read speech we cannot generalize the results to unconstrained, conversational, or any other type of speech. Thus, another area to explore is the use of other training corpora containing spontaneous, dialogic, and other types of speech.

The results reported in this paper reaffirm the potential of investigating Brazil's (1997) intonation discourse theory as a means of better comprehending natural discourse in different environments that we found in earlier work.

References

Amir, O., Wolf, M., & Amir, N. (2009). A clinical comparison between two acoustic analysis softwares: MDVP and Praat. *Biomedical Signal Processing and Control, 4*(3), 202–205.

Ananthakrishnan, S., & Narayanan, S. (2008). Fine-grained pitch accent and boundary tone labeling with parametric f0 features. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008.* (pp. 4545–4548). IEEE.

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

Bocklet, T., & Shriberg, E. (2009). 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).

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161–168). ACM, New York.

Cauldwell, R. (2012). Brazil, David. *The encyclopedia of applied linguistics.*

Chun, D. M. (2002). *Discourse intonation in L2: From theory and research to practice (Vol. 1)*. Philadelphia, PA: John Benjamins Publishing.

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

Diehl, J. J., & Paul, R. (2012). Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorders. *Research in Autism Spectrum Disorders, 6*(1), 123–134.

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., & Cardeñoso-Payo, V. (2007). Applying data mining techniques to corpus based prosodic modeling. *Speech Communication, 49*(3), 213–229.

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 and 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, 346*(1315), 97–104.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continuos 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.

Hastie, T., & Tibshirani, R. (1998). Classification by pairwise coupling. *The Annals of Statistics, 26*(2), 451–471.

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

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.

Levow, G. A. (2005). Context in multi-lingual tone and pitch accent recognition. In *INTERSPEECH* (pp. 1809–1812).

Li, K., Zhang, S., Li, M., Lo, W. K., & Meng, H. (2010). Detection of intonation in L2 English speech of native Mandarin learners. In *2010 7th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (pp. 69–74). IEEE.

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.

Maryn, Y., Corthals, P., De Bodt, M., Van Cauwenberge, P., & Deliyski, D. (2009). Perturbation measures of voice: a comparative study between Multi-Dimensional Voice Program and Praat. *Folia Phoniatrica et Logopaedica, 61*(4), 217–226.

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 and Communication Disorders, 38*(4), 325–350.

Oller, D. K., Niyogi, P., Gray, S., Richards, J. A., Gilkerson, J., Xu, D., & Warren, S. F. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences, 107*(30), 13354–13359.

Ostendorf, M. (1999). 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: Gainesville: University of Florida.

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

Ringeval, F., Demouy, J., Szaszak, G., Chetouani, M., Robel, L., Xavier, J., & Plaza, M. (2011). Automatic intonation recognition for the prosodic assessment of language-impaired children. *IEEE Transactions on Audio, Speech, and Language Processing, 19*(5), 1328–1342.

Rosenberg, A. (2010). Classification of prosodic events using quantized contour modeling. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 721–724). Association for Computational Linguistics.

Rosenberg, A. (2010). AutoBI-a tool for automatic toBI annotation. In *INTERSPEECH* (pp. 146–149).

Rosenberg, A. (2012). Modeling intensity contours and the interaction of pitch and intensity to improve automatic prosodic event detection and classification. In *2012 IEEE Spoken Language Technology Workshop (SLT)* (pp. 376–381). IEEE.

Ross, K., & Ostendorf, M. (1996). Prediction of abstract prosodic labels for speech synthesis. *Computer Speech and Language, 10*(3), 155–185.

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.

Sun, X. (2002). Pitch accent prediction using ensemble machine learning. In *INTERSPEECH*.

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

Taylor, P. (2000). Analysis and synthesis of intonation using the tilt model. *The Journal of the Acoustical Society of America, 107*(3), 1697–1714.

The Centre for Speech Technology Research, University of Edinburgh (2014), The Festival Speech Synthesis System, [Computer Program]. Retrieved September 15, 2014, from http://www.cstr.ed.ac.uk/projects/festival.

Van Santen, J. P., Prud'hommeaux, E. T., Black, L. M., & Mitchell, M. (2010). Computational prosodic markers for autism. *Autism, 14*(3), 215–236.

Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. New York: Oxford University Press.

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, D., Gilkerson, J., Richards, J., Yapanel, U., & Gray, S. (2009, September). Child vocalization composition as discriminant information for automatic autism detection. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. EMBC 2009.* (pp. 2518–2522). Minneapolis: IEEE.

Yoon, T., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. *Proceedings of the International Conference on Spoken Language Processing* (pp. 2729–2732). Japan: Nara.