Comparison of Inter-rater Reliability of Human and Computer Prosodic Annotation Using Brazil's Prosody Model

Okim Kang[1] & David O. Johnson[1]

[1] Department of English, Northern Arizona University, Flagstaff, Arizona, USA

Correspondence: Okim Kang, Department of English, Liberal Arts Building 18, Room 140, Northern Arizona University, Flagstaff, AZ, 86011-6032, USA. Tel: 928-523-2059.

Received: November 29, 2015 Accepted: December 12, 2015 Online Published: December 15, 2015

doi:10.5430/elr.v4n4p58 URL: http://dx.doi.org/10.5430/elr.v4n4p58

Abstract

The current study examined whether the computer annotations of prodody based on Brazil's (1997) framework were comparable with human annotations. A series of statistical tests were performed for each prosodic feature: tone unit (two accuracy scores and Pearson's correlation), prominent syllable (accuracy, F-measure, and Cohen's kappa), tone choice (accuracy and Fleiss' kappa), and relative pitch (accuracy, Fleiss' kappa, and Pearson's correlation). We considered one population to be the inter-rater reliability scores between the three human coders and the other population to be the inter-rater reliability scores between the computer and the three humans. If the differences between these two populations were significant, then the computer and human annotations were considered not comparable, but if the differences were not significant, then the computer and human annotations were considered comparable. The results indicated that the computer and human annotations were comparable for tone choice and not comparable for prominent syllable. For tone unit, two of the t-tests provided evidence that they were comparable and one did not. The relative pitch t-tests showed a significant disparity between the estimates of relative pitch by the humans and the computer's actual relative pitch calculation.

Keywords: Inter-rater reliability, Brazil's prosody model, Automatic computer prosodic annotation, ToBI, RaP

1. Introduction

1.1 Overview

Prosody, which is the relationship between accentuation and inflection in a language, shapes language research in a variety of fields, including linguistics, computer science, and psychology. Researchers in these fields frequently depend on human prosodic annotation to create exemplars and prove hypotheses. Computer prosodic annotation is important because it is faster, cheaper, and more consistent than human prosodic annotation. Although computer annotation is arguably more consistent, it does not necessarily mean it is more correct than human annotation. The goal of computer annotation is to make it consistent and correct.

1.2 Problem Statement

In this paper, we compare computer and human prosodic annotations based on Brazil's (1997) intonation model. The *tone unit* is the first building block of Brazil's model. Brazil specifies a tone unit as a fragment of an utterance that a hearer can perceive as exhibiting a tone pattern that is not the same as those of other tone units with different tone patterns. The second building block of Brazil's model is the *prominent syllable*, which is differentiated from other syllables by three attributes: pitch (fundamental frequency of a syllable in Hz), intensity (amplitude of the syllable in dB), and duration (timespan of the syllable in seconds) (Chun, 2002). Each tone unit has one or more prominent syllables. The first one is called the *key* prominent syllable and the last one is called the *termination* prominent syllable. The *relative pitch* of the key and termination syllables and the *tone choice* of the termination syllable typify the intonation pattern of a tone unit. Brazil enumerated three evenly divided gradations of relative pitch: low, mid, and high, and five tone choices: falling, rising, rising-falling, falling-rising, and neutral.

1.3 Study Objective

The current study is guided by the following research question: Is the computer annotation of Brazil's model elements, specifically tone unit, prominent syllable, tone choice, and relative pitch, comparable with human annotation? We begin the paper with a description of our methods including the corpora, human annotation, computer annotation,

inter-rater reliability metrics we employed, and how we compared the two with t-tests. Next, we present the inter-rater reliability results between the humans, between the humans and the computer, and the results of comparing them with t-tests. Finally, we discuss the comparisons and offer some conclusions.

2. Literature Review

Research on inter-rater reliability of annotations using Brazil's prosody model has been rare. Johnson and Kang (2015a, 2015b) reported the inter-rater reliability amongst two human annotations of Brazil's (1997) prominent syllables and tone choices in a subset of the TIMIT corpus (Garofolo, Lamel, Fisher, Fiscus, & Pallett, 1993) was 85%-87%. However, this agreement was measured on only about 80 samples.

A few studies examined inter-rater reliability annotations utilizing ToBI. The tones and break indices (ToBI) is a method for marking prosodic occurrences in discourse (Silverman et al., 1992; Beckman & Ayers, 1997). ToBI defines three prosodic elements: Pitch Accents, Boundary Tones and Break Indices. The prosodic concept of prominence is denoted by Pitch Accent. The prosodic notion of intonational phrasing is symbolized by Boundary Tones and Break Indices. Although seemingly similar, there is no one-to-one correspondence between the elements of Brazil's intonational model (i.e., tone unit, prominent syllables, tone choice, and relative pitch) and Pitch Accents, Boundary Tones, and Break Indices.

In the earliest ToBI study, 26 labelers applied the ToBI system to 489 words taken from both read and spontaneous speech corpora (Pitrelli, Beckman, & Hirschberg, 1994). Criticisms of this study are that it utilized an inter-rater reliability metric which did not allow for the probability of random agreement and that the corpus was tiny and uncharacteristic of normal speech (Breen, Dilley, Kraemer, & Gibson, 2012). Syrdal and McGory (2000) employed six labelers who annotated 645 words. Like Pitrelli et al. (1994), the corpus was small and consisted of only two speakers. Unlike Pitrelli et al. (1994), however, it allowed for random agreement by applying Cohen's kappa. A later study by Yoon, Chavarria Cole, and Hasegawa-Johnson (2004) examined a larger corpus of 1,600 words of unconstrained speech articulated by 79 speakers using Cohen's kappa as an inter-rater reliability metric, which was annotated by two transcribers.

In a more recent comprehensive study, Breen et al. (2012) compared the annotations by four trained, but inexperienced, transcribers and four experienced transcribers using Cohen's kappa. The inexperienced transcribers annotated six speakers articulating 5,939 syllables of the Boston Radio News corpus (BURNC) of read professional broadcast news speech (Ostendorf, Price, & Shattuck-Hufnagel, 1995) and another six speakers uttering 3,680 syllables of the CALLHOME corpus of spontaneous nonprofessional speech from telephone conversations (Canavan, 1997). In addition to transcribing the speech with ToBI, the inexperienced transcribers annotated six speakers uttering 2,638 syllables of CALLHOME and six speakers uttering 2,889 syllables of BURNC using RaP. RaP (Rhythm and Pitch) is a technique for tagging the relative pitch and rhythm of English discourse. It is an augmentation of ToBI that facilitates the portrayal of both intonational and rhythmic facets of language (Dilley & Brown, 2005), founded on a tone interval theory offered by Dilley (2005). The experienced ones annotated a new smaller corpus which had not been annotated as part of the first study. The new corpus contained utterances consisting of unconstrained speech from the CALLHOME corpus and of read speech from the BURNC. The utterances were spoken by seven speakers and contained a total of 1,533 syllables. The experienced coders annotated the new corpus utilizing both ToBI and RaP.

Although not quite the same, findings from an investigation, utilizing a rubric for evaluating oral prosody proficiency, had an inter-rater agreement of 0.857 (Rasinski, Rikli, & Johnston, 2009). The rubric encompassed four sections with associated rating explanations: phrasing, volume, pace, and smoothness. The overall prosody rating for each utterance varied between four and sixteen with each section being rated from a low of one to a high of four. Agreement was specified as within two rating points.

3. Methods

3.1 Corpora

A machine learning classifier has been developed to automatically score the English proficiency of speakers using unconstrained speech (Johnson, Kang, & Ghanem, 2015). The classifier calculated the proficiency score from a set of segmental and suprasegmental measures based on Brazil's prosody model (1997). The segmental and suprasegmental measures were computed from the output of an ASR that identifies phones instead of words and other software which ascertains the elements of Brazil's model. The Pearson's correlation between the computer's calculated proficiency scores and the official Cambridge English Language Assessment (CELA) scores was 0.68. CELA is a globally accepted set of tests and qualifications for learners and teachers of English. Two of the elements of Brazil's model, prominent syllable and tone choice, were identified by machine learning classifiers which were

trained on the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al., 1993). In an effort to improve the correlation of 0.68, we examined training the classifiers with different combinations of TIMIT and another corpus, the Boston University Radio News Corpus (BURNC) (Ostendorf et al., 1995). This effort succeeded in improving the correlation to 0.72 (Johnson et al., 2015). The BURNC was annotated by three trained analysts, which gave us the opportunity to study the inter-rater reliability between three human coders and between them and the computer.

This research utilized two corpora. The one called TIMIT in this paper is a subset of 839 utterances from the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al., 1993) containing 11,537 syllables and 7,277 words. Table 1 presents the distribution of the speakers of the 839 utterances by gender and dialect.

Table 1. 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 complete TIMIT corpus includes 6,300 utterances of read speech comprising ten utterances from 630 speakers representing eight main dialect areas of the United States. The read speech is made up of three groups of sentences. The first group consists of two sentences designed to identify the dialect of the speakers. The two dialect sentences were read by all 630 speakers. The second group is 1,890 phonetically-diverse sentences composed of a large diversity of allophonic combinations. Each speaker read three sentences from the second group and each of the sentences was read by only one speaker. The third group is 450 phonetically-compact sentences intended to include most common phone pairs plus additional phonetic combinations that are challenging. Five of the third group were spoken by each speaker and they were read by seven different speakers. The corpus also contains hand rectified start and end times for the phones, phonemes, syllables, words, and pauses.

The other corpus is called BURNC which includes 144 utterances from the Boston University Radio News Corpus (Ostendorf et al., 1995) consisting of 16,253 syllables and 10,566 words. BURNC is a corpus of professionally read radio news stories (Ostendorf et al., 1995). The corpus includes more than seven hours of speech from three female and four male radio announcers. Each story is split into paragraph size portions of several sentences. The paragraphs are annotated with orthographic transcriptions, phonetic alignments, part-of-speech tags, and prosodic labels. The phonetic alignments were generated using constrained speech recognition for the stories considered not noisy and then manually corrected (Ostendorf et al., 1995). The 144 utterances are composed of 24 from each of the six speakers in the Boston University Radio News Corpus. Each utterance represents a paragraph of a story. Table 2 shows the stories selected for each speaker in the study. The stories were selected based on the quality of the recordings and the number of paragraphs (i.e., some stories had more paragraphs than others).

Table 2. BURNC stories for each speaker in the Boston University Radio News Corpus

Speaker	Stories
f1a	s01, s02, s03, s04, s05
f2b	s03, s05, s06, s09, s10
f3a	s01, s02, s03, s04, s05, s07, s08, s09, s10, rrl, trl
m1b	s01, s09, s10, s03, s02
m2b	s01, s02, s04, s03
m3b	jrl, prl, rrl, trl

The syllable boundaries were established automatically by utilizing the dictionary included with the corpus. The dictionary was hand-corrected to correspond to the phonetic alignments for the cases where they did not match.

3.2 Human Annotation

The tone units, prominent syllables, tone choices, and relative pitch of the TIMIT corpus were annotated by a trained analyst (designated as Analyst X in this study). She coded them both by listening to the audio files and by using Praat

(Boersma & Weenink, 2014) and Multi-Speech and Computerized Speech Laboratory (CSL) Software (KayPENTAX, 2008), computerized speech analysis programs, to examine the pitch contours, intensity, and duration of the syllables. Roughly ten samples were annotated by a second trained analyst to verify the reliability of acoustic analyses. The two analysts reviewed any discrepancies and resumed coding more samples until they were in agreement. Analyst X then annotated the rest of the corpus independently. This annotation method has been extensively applied as a dependable labeling technique in other studies (Kang, 2010; Kang, Rubin, & Pickering 2010; Pickering, 1999) in applied linguistics.

Three trained analysts (designated as Analysts X, Y, and Z) annotated the tone units, prominent syllables, tone choices, and relative pitch of BURNC using the Praat (Boersma & Weenink, 2001) computerized speech analysis program to observe the pitch contours, intensity, and duration of the syllables. Analyst X is the same analyst that annotated TIMIT. To establish the reliability of the acoustic analyses, the three analysts compared ten samples, discussed any discrepancies, and reached a consensus. Then, the analysts performed the annotation independently for the rest of the speech samples.

3.3 Computer Annotation

The computer annotations were generated specifically for this study or were collected from other unpublished studies as follows. The computer's tone unit annotations were produced for this paper by a computer algorithm that scrutinized the silent pauses. Tone units were delimited by silent pauses that lasted longer than 200 ms or by ones that lasted between 150 ms and 200 ms and occurred in conjunction with a pitch reset or slow pace. Pitch reset signifies that the relative pitch of the three-phone-window in front of the silent pause is high and the relative pitch of the three-phone-window following it is low, or just the opposite (i.e., low in front of and high following). Slow pace means the duration of the three-phone-window after the silent pause is greater than the normal duration of a three-phone-window. The normal duration of a three-phone-window is determined by adding together the mean duration of the phones in the three-phone-window. The mean duration of each phone is calculated over the entire utterance.

The computer identified the prominent syllables utilizing various supervised machine learning classifiers which analyzed various combinations of syllable pitch, duration, and intensity to determine if the syllable was prominent or not (Johnson & Kang, 2015a). The classifiers were trained and tested on various combinations of the TIMIT and BURNC corpora annotated by Analyst X as follows: five-fold cross-validation of TIMIT, TIMIT trained and BURNC tested, BURNC trained and TIMIT tested, three-fold cross-validation of BURNC, and six-fold cross-validation of a combine corpus of TIMIT and BURNC.

Similarly, the computer employed a number of supervised machine learning classifiers to evaluate the pitch contours of the prominent syllables to ascertain their tone choice (Johnson & Kang, 2015b). The same combinations of the Analyst X annotated TIMIT and BURNC corpora, which were drawn on to test and train the prominent syllable classifiers, were also made use of in testing and training the tone choice classifiers. The computer annotations for relative pitch were calculated using Brazil's (1997) algorithm which divides the pitch range of an utterance into three equal scales: low, mid, and high.

3.4 Inter-Rater Reliability between the Human Raters and the Computer

We used eleven inter-rater reliability metrics to compare the human raters amongst themselves and with the computer. Table 3 gives the metrics and describes how they were measured.

Table 3. Inter-rater reliability metrics

Brazil Element	Inter-rater Reliability Metric	Description
Tone Unit	Accuracy by syllables	Each syllable of an utterance is marked with the tone unit it belongs to. Accuracy is then the percent of syllables in the corpus for which the two raters agreed on the tone unit it belong to.
	Accuracy by number of tone units	Percent of utterances where the two raters agreed on the number of tone units.
	Correlation between number of tone units	Pearson's correlation of the number of tone units per utterance between the two raters.
Prominent Syllable	Accuracy	Percent of prominent syllables on which the two raters agreed.
	F-measure	F-measure of prominent syllables identified by the two raters.
	Cohen's kappa	Cohen's kappa of prominent syllables identified by the two raters.
Tone Choice	Accuracy	Percent of tone choices where the two raters agreed out of the prominent syllables on which the two raters agreed.
	Fleiss' kappa	Fleiss' kappa of tone choice annotations of the prominent syllables on which the two raters agreed.
Relative Pitch	Accuracy	Percent of relative pitches where the two raters agreed out of the prominent syllables on which the two raters agreed.
	Fleiss' kappa	Fleiss' kappa of relative pitch annotations of the prominent syllables on which the two raters agreed.
	Correlation	Pearson's correlation of the ordinal value of the relative pitch annotations of the prominent syllables on which the two raters agreed.

In addition to the two inter-rater reliability measures, accuracy (a.k.a., $Pr(a)$ or joint-probability of agreement) and kappa (Cohen's and Fleiss'), used in the ToBI and RaP studies discussed above, we used F-measure and correlation. F-measure, like kappa, punishes agreement that results from random chance. It is calculated as follows: TP (true-positives) is the number of syllables where both the computer and the human identified it as prominent; TN (true-negative) is the number of syllables where both the computer and the human identified it as not prominent; FP (false-positive) is the number of syllables where the computer identified it as prominent and the human identified it as not prominent; FN (false-negative is the number of syllables where the computer identified it as not prominent and the human identified it as prominent; and F-Measure = 2TP / (2TP + FP + FN). Pearson's correlation can be used to compare inter-rater agreement where the values being rated are ordinal (e. g., the number of tone units, relative pitch).

3.5 Comparison of Human and Computer Annotations

We employed the two-tailed two-sample t-test assuming unequal variances to answer the research question of whether the computer annotations of Brazil model elements, specifically tone unit, prominent syllable, tone choice, and relative pitch, was comparable with human annotations. For the t-test, we considered one population to be the human-human inter-rater reliability scores between Analysts X and Y, Analysts X and Z, and Analysts Y and Z; the other population is the several human-computer inter-rater reliability scores between Analyst X and the computer collected from other studies or is the human-computer inter-rater reliability scores calculated for this study between the computer and the three analysts, X, Y, and Z. If the differences between these two populations were significant ($p < \alpha$), then the computer and human annotations were considered **not comparable**, but if the differences were not significant ($p > \alpha$), then the computer and human annotations were considered **comparable**.

Taking tone choice as an example, we calculated inter-rater reliability scores between the three human analysts using Fleiss' kappa, which gives us a population ($n = 3$) of human-human inter-rater reliability scores. From other studies described above, we have a population ($n = 5$) of human-computer inter-rater reliability scores (also Fleiss' kappa) between the computer and Analyst X. A two-tailed two-sample t-test assuming unequal variances will indicate if the human-human inter-rater reliability scores are comparable with the human-computer inter-rater reliability scores, providing us with an answer to the research question for tone choice.

4. Results

Table 4 gives the tone unit inter-rater reliability scores between each pair of human annotators (i.e., X:Y, X:Z, and Y:Z) and between each human annotator and the computer (i.e., X:C, Y:X, and Z:C). All of the tone unit inter-rater reliability scores compared annotations of BURNC

Table 4. Tone unit inter-rater reliability scores

Between	Corpus	Accuracy by syllables	Accuracy by number of tone units	Correlation between numbers of tone units
X:Y	BURNC	74.6%	63.2%	0.911
X:Z	BURNC	61.4%	47.2%	0.841
Y:Z	BURNC	61.9%	47.9%	0.881
X:C	BURNC	53.0%	31.9%	0.873
Y:C	BURNC	46.7%	29.9%	0.855
Z:C	BURNC	47.5%	34.0%	0.781

The prominent syllable inter-rater reliability scores are presented in Table 5. The prominent syllable annotations of a variety of combinations of TIMIT and BURNC are contrasted. For the human-computer ones, different computer algorithms were used to generate the computer annotations, thus the different scores for the same corpus.

Table 5. Prominent syllable inter-rater reliability scores

Between	Corpus	Accuracy	F-Measure	Cohen's kappa
X:Y	BURNC	91.0%	95.3	0.727
X:Z	BURNC	90.3%	94.9	0.702
Y:Z	BURNC	90.9%	95.2	0.724
X:C	TIMIT	95.9%	93.7	0.907
X:C	TIMIT	79.3%	70.8	0.550
X:C	BURNC	86.3%	61.7	0.540
X:C	BURNC	82.3%	63.7	0.530
X:C	BURNC+TIMIT	83.4%	63.1	0.530
X:C	BURNC	86.4%	61.0	0.530
X:C	BURNC	82.3%	62.4	0.510
X:C	BURNC	82.3%	62.2	0.510
X:C	BURNC+TIMIT	77.9%	63.2	0.480

Table 6 lists the inter-rater reliability scores for tone choice. Like the prominent syllable annotations, the scores compare tone choice annotations of a variety of combinations of TIMIT and BURNC.

Table 6. Tone choice inter-rater reliability scores

Between	Corpus	Accuracy	Fleiss' kappa
X:Y	BURNC	78.1%	0.761
X:Z	BURNC	66.7%	0.636
Y:Z	BURNC	74.5%	0.721
X:C	TIMIT	75.1%	0.730
X:C	TIMIT	70.4%	0.680
X:C	BURNC	66.6%	0.640
X:C	BURNC	66.1%	0.630
X:C	BURNC+TIMIT	71.9%	0.691

The relative pitch inter-rater reliability scores are provided in Table 7. Annotations of BURNC were evaluated to calculate the relative pitch inter-rater reliability scores.

Table 7. Relative pitch inter-rater reliability scores

Between	Corpus	Accuracy	Fleiss' kappa	Correlation
X:Y	BURNC	81.2%	0.799	0.640
X:Z	BURNC	76.0%	0.744	0.634
Y:Z	BURNC	78.4%	0.769	0.653
X:C	BURNC	61.6%	0.593	0.455
Y:C	BURNC	60.6%	0.582	0.466
Z:C	BURNC	53.4%	0.504	0.395

Table 8 summarizes the t-test results for each of Brazil's intonation model elements. A figure is given for each result showing the mean and standard deviation of both populations, human-human (H-H) and human-computer (H-C).

Table 8. T-test results

Brazil Element	Inter-rater Reliability Score	Figure	Comparable? ($p > 0.05$)
Tone Unit	Accuracy by syllables	1	No
	Accuracy by number of tone units	1	Yes
	Correlation between number of tone units	1	Yes
Prominent Syllable	Accuracy	2	No
	F-measure	2	No
	Cohen's kappa	2	No
Tone Choice	Accuracy	3	Yes
	Fleiss' kappa	3	Yes
Relative Pitch	Accuracy	4	No
	Fleiss' kappa	4	No
	Correlation	4	No

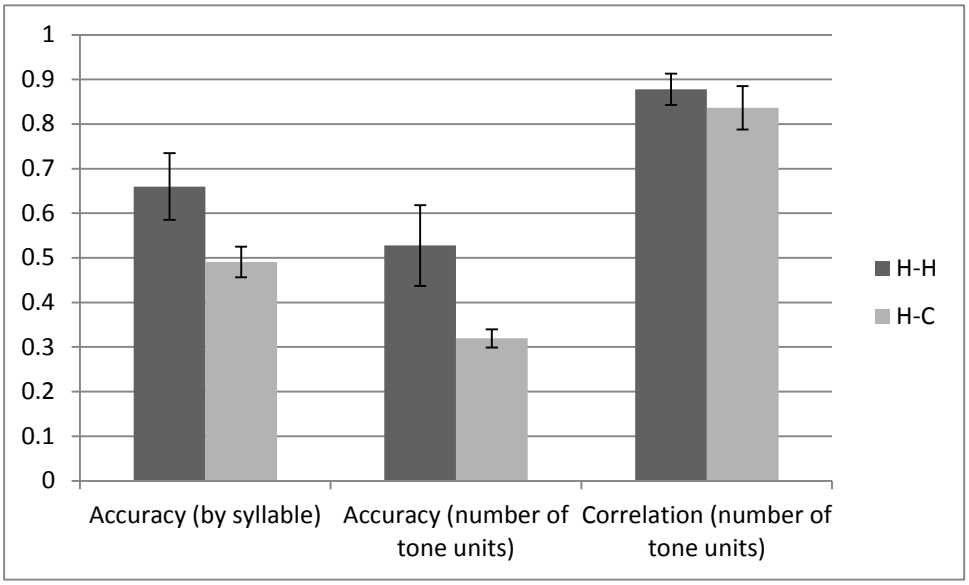

Figure 1. Tone Units (Accuracy by syllable), H-H (M=0.66, SD=0.07) and H-C (M=0.49, SD=0.03); t(3)=3.18, p = 0.038; (Accuracy by number of tone units), H-H (M=0.53, SD=0.09) and H-C (M=0.32, SD=0.02); t(2)=4.30, p = 0.060; (Correlation between number of tone units), H-H (M=0.88, SD=0.03) and H-C (M=0.84, SD=0.05); t(4)=2.78, p = 0.301

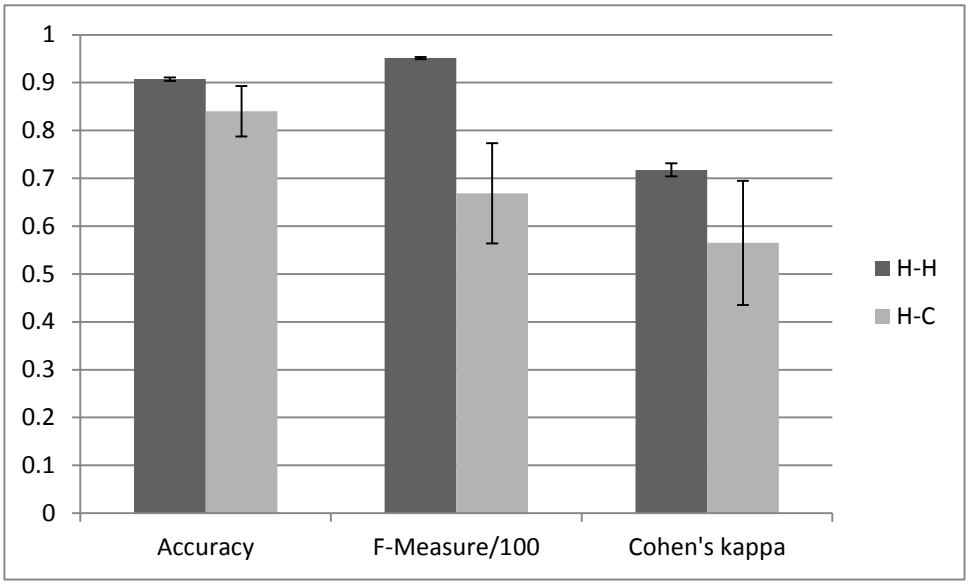

Figure 2. Prominent Syllables (Accuracy), H-H (M=0.91, SD=0.00) and H-C (M=0.84, SD=0.05); t(8)=2.31, p = 0.005; (F-measure), H-H (M=95.13, SD=0.21) and H-C (M=66.87, SD=10.46); t(8)=2.31, p = 0.000; (Cohen's kappa), H-H (M=0.72, SD=0.01) and H-C (M=0.57, SD=0.13); t(8)=2.26, p = 0.007

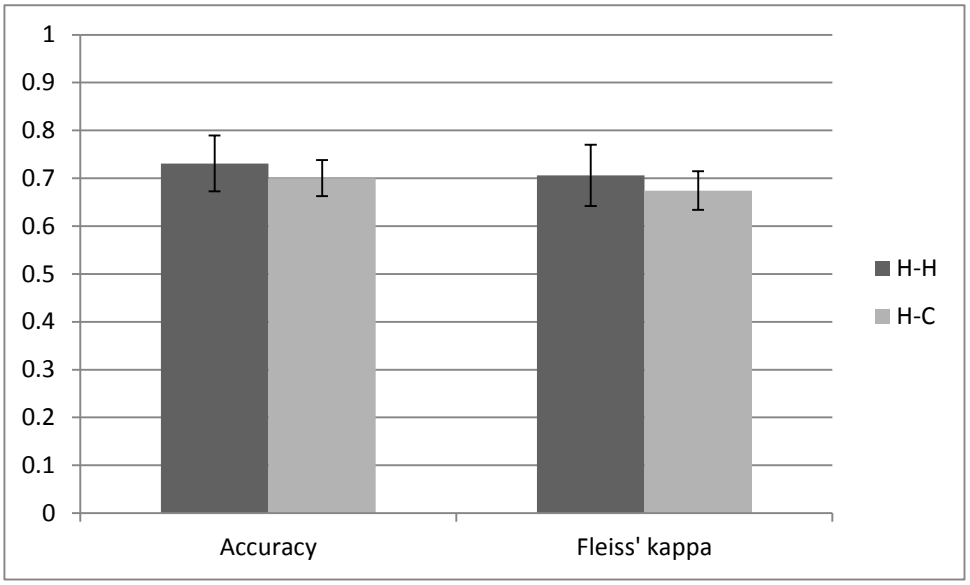

Figure 3. Tone Choice (Accuracy), H-H (M=0.73, SD=0.06) and H-C (M=0.70, SD=0.04); t(3)=3.18, p = 0.473; (Fleiss' kappa), H-H (M=0.71, SD=0.06) and H-C (M=0.67, SD=0.04); t(3)=3.18, p = 0.495

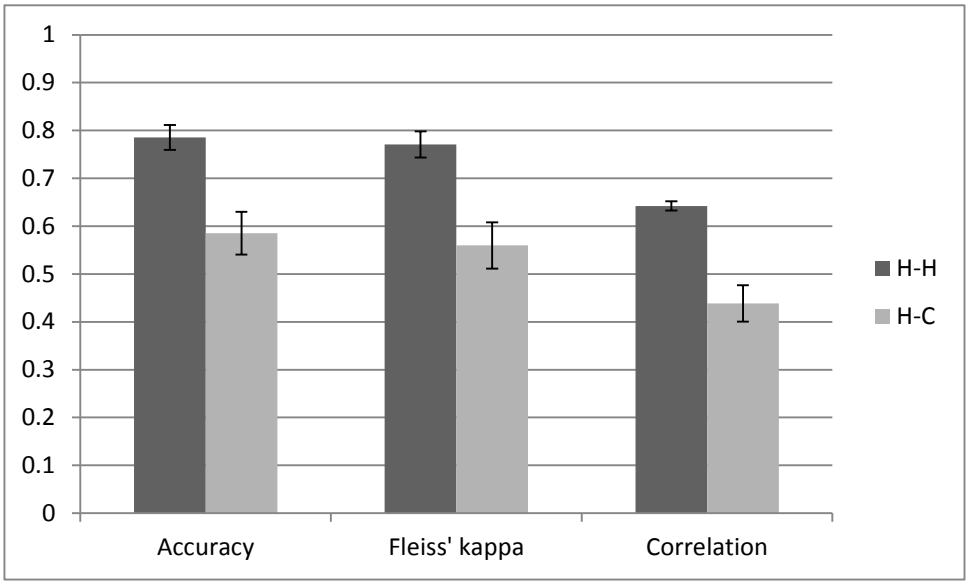

Figure 4. Relative Pitch (Accuracy), H-H (M=0.79, SD=0.03) and H-C (M=0.59, SD=0.04); t(3)=3.18, p = 0.007; (Fleiss' kappa), H-H (M=0.77, SD=0.03) and H-C (M=0.56, SD=0.05); t(3)=3.18, p = 0.007; (Correlation), H-H (M=0.64, SD=0.01) and H-C (M=0.44, SD=0.04); t(2)=4.30, p = 0.012

5. Discussion

In order to generate models and validate theories, researchers in the fields of linguistics, computer science, and psychology regularly utilize human prosodic transcriptions of speech corpora. Computer prosodic transcriptions can be quicker, less expensive, and more reliable in terms of consistency than human prosodic transcription. Even though computer annotation might be more reliable, it may not be more accurate than human annotation. There is a dearth of research on inter-rater reliability of corpus annotations using Brazil's (1997) prosody model. In this paper, we performed an inter-rater reliability study of three analysts who annotated a portion of the BURNC utilizing Brazil's prosody model. Next, we conducted an inter-rater reliability study between the analysts' annotations and those of a computer program. Then, we used the t-test to answer the research question: Is the computer annotation of Brazil's model elements, specifically tone unit, prominent syllable, tone choice, and relative pitch, comparable with human annotation? Another way to consider the research question is: At what point is the computer annotation as good as human annotation and can be trusted as much as the human annotation? We posit that it is when the human-computer inter-rater reliability metrics are comparable to the human-human ones. By comparable we mean there is no significant difference between the human-computer and human-human inter-rater reliability measures. The generally accepted method of measuring significance between any two phenomena is the t-test. Thus, if the t-test shows no significant difference between the inter-rater reliability scores, then we can say they are comparable. Now, the inter-rater reliability scores and t-test results for each element of Brazil's (1997) model will be discussed individually.

The mean tone unit, accuracy by syllable, between the human raters is 66% and 49% between the humans and the computer (Figure 1). Among the human annotators, the average accuracy by number of tone units is 53% and 32% among the computer and the humans (Figure 1). Comparing the number of tone units detected by the computer and the humans, the mean correlation is 0.84; comparing the number between the humans, the mean correlation is 0.88 (Figure 1). In the first study of Breen et al. (2012), they reported the accuracy of human-human annotations of various ToBI elements ranging from 77% to 87% and various RaP elements ranging between 72% and 92%. In the second study, they reported 80% to 91% accuracy for ToBI elements and 75% to 90% accuracy for RaP elements. These accuracies are better than the accuracy by syllable and accuracy by tone unit which were found in our study. Although our experiments, the ToBI experiments, and the RaP experiments were all carried out with the BURNC, the differences in the Brazil, ToBI, and RaP intonation models suggests this comparison may not be meaningful. The t-test shows the difference between the human-human accuracy by syllables scores and the human-computer scores is significant implying that the human and computer tone unit annotations are not comparable. On the other hand, the t-tests for the accuracy by number of tone units and correlation between the number of tone units indicate that the differences in scores are insignificant, suggesting that the human and computer tone unit annotations are comparable.

With regard to prominent syllable annotation, the mean accuracy, F-measure, and kappa between the human annotators is 0.91, 95.13, and 0.72, respectively; among the humans and the computer they are 0.84, 66.87, and 0.57, respectively (Figure 2). Escudero-Mancebo, González-Ferreras, Vivaracho-Pascual, and Cardeñoso-Payo (2014) observed that in ToBI studies, human inter-rater kappa ranges from 0.51 (Yoon et al., 2004) to 0.69 (Syrdal & McGory, 2000). Breen et al. (2012) stated human inter-rater kappa values of 0.52 and 0.77 for RaP research. In our experiments, the mean human inter-rater kappa of 0.72 was above the ToBI range and at the high end of the RaP range. As with the tone unit scores our experiments, the ToBI experiments, and the RaP experiments were all carried out with the BURNC, but the differences in the Brazil, ToBI, and RaP intonation models insinuate that this contrast may be meaningless. Specific to prominence vs. non-prominence, Breen et al. reported ToBI kappas of 0.71 and RaP kappas of 0.77 for their first study and kappas of 0.77 and 0.78 for their second study of ToBI and RaP, respectively. Our mean human inter-rater kappa of 0.72 is at the low end of their two studies. But, the ToBI and RaP definition of prominence varies from that of Brazil and so this may not be a meaningful comparison. The t-tests for accuracy, F-measure, and kappa reveal that the disparities between the human-human and human-computer values are significant; implying that the human and computer prominent syllable annotations may not be comparable.

Concerning tone choice annotation, the average accuracy and kappa among the human annotators is 0.73 and 0.71 (Figure 3), respectively; they are 0.70 and 0.67 among the humans and the computer, respectively. The mean human inter-rater kappa of 0.71 was higher than the range for the ToBI experiments and at the top end of the range for the RaP ones. And as before, the dissimilarities in the Brazil, ToBI, and RaP prosodic models means this contrast may not be consequential. The t-tests for accuracy and kappa bring to light that the differences in the human-human and human-computer values are not significant, pointing toward the human and computer tone choice annotations being comparable.

The results for the relative pitch need to be interpreted differently than those for the tone unit, prominent syllable, and tone choice because relative pitch is an objective measure (i.e., three equal ranges: low, mid, and high) whereas the others are subjective. Comparing the human relative pitch annotations with those of the computer is analogous to comparing human visual estimates of distance with distance measurements made with a ruler. Thus, the human-human relative pitch inter-rater reliability measurements of accuracy, kappa, and correlation really only measure how well the humans estimated the relative pitch in comparison to each other. Whereas, the human-computer relative pitch measurements actually measure how well the humans estimated the relative pitch in comparison to the actual relative pitch that was measured by the computer. Thus, the mean human-human relative pitch inter-rater reliability measures of accuracy, kappa, and correlation of 0.79, 0.77, and 0.64, respectively, show that although there is moderate agreement between the human annotaters, they did a fairly poor job of estimating the relative pitch as evidenced by the mean human-computer relative pitch measures of accuracy, kappa, and correlation which are 0.59, 0.56, and 0.44, respectively (Figure 4). There is no objective measure in ToBI or RaP with which to compare relative pitch. The t-tests also show there is a significant difference between the human's estimates of relative pitch and the actual relative pitch as measured by the computer. The conclusion here is the fairly obvious one that the computer is better at measuring relative pitch than the humans are at estimating it.

Other major differences between our studies and the ToBI and RaP studies are the size of the corpora used and the metrics utilized to measure inter-rater reliability. We employed much larger corpora than any of the ToBI or RaP studies discussed above: 11,537 syllables (TIMIT) and 16,253 syllables (BURNC) vs. 3,680 syllables (CALLHOME) and 5,939 syllables (BURNC). In addition to accuracy and kappa applied in the ToBI and RaP studies, we measured correlation and F-Measure.

6. Conclusion

In summary, the t-tests showed that the computer and human annotations were comparable for tone choice and not comparable for prominent syllable. For tone unit, two of the t-tests indicated they were comparable and one did not. The t-tests for relative pitch pointed out a significant difference between the human's estimates of relative pitch and the actual relative pitch measured by the computer; they also indicate that the humans did a fairly poor job of estimating relative pitch.

As far as future work, these results insinuate more effort should be spent in improving automatic prominent syllable detection and possibly tone unit detection. For tone choice, these outcomes suggest that computer annotation of tone choice is comparable to human annotation. The interpretation of the relative pitch results is different. In this case, the computer is correct because it is following a simple algorithm. Thus, the findings imply that it is difficult for a human to estimate relative pitch accurately using Praat.

References

Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling. *The OSU Research Foundation, 3*.

Boersma, P. & Weenink, D. (2001). *Praat, a system for doing phonetics by computer* (version 5.3.83). [Computer program]. Retrieved August 19, 2014.

Brazil, D. (1997). *The Communicative Value of Intonation in English Book*. 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). 277-312. http://dx.doi.org/10.1515/cllt-2012-0011

Canavan, A., Graff, D., & Zipperlen, G. (1997). Callhome american english speech. *Linguistic Data Consortium*.

Chun, D. M. (2002). *Discourse intonation in L2: From theory and research to practice*. (Vol. 1). John Benjamins Publishing. ANSI z39.48-1984. http://dx.doi.org/10.1075/lllt.1

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. http://dx.doi.org/10.1016/j.csl.2013.08.001

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

Johnson, D. O., Kang, O., & Ghanem, R. (2015). Language proficiency ratings: Human versus Machine. Presentation at *Pronunciation in Second Language Learning and Teaching 2015, Dallas, TX, USA, October 16-17, 2015*.

Johnson, D. O., & Kang, O. (2015a). Automatic prominent syllable detection with machine learning classifiers. *International Journal of Speech Technology*, *18*, (4), 583-592. http://dx.doi.org/10.1007/s10772-015-9299-z

Johnson, D. O., & Kang, O. (2015b). Automatic prosodic tone choice classification of Brazil's intonation model. *International Journal of Speech Technology*. http://dx.doi.org/10.1007/s10772-015-9327-z

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, no. 4. 554-566. http://dx.doi.org/10.1111/j.1540-4781.2010.01091.x

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System 38*, no. 2. 301-315. http://dx.doi.org/10.1016/j.system.2010.01.005

KayPENTAX (2008). *Multi-Speech and CSL Software*. Lincoln Park, NJ: KayPENTAX.

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

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

Pitrelli, J. F., Beckman, M. E., & Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the tobi framework. In *ICSLP*.

Rasinski, T., Rikli, A., & Johnston, S. (2009). Reading fluency: More than automaticity? More than a concern for the primary grades?. *Literacy Research and Instruction*, 48(4), 350-361. http://dx.doi.org/10.1080/19388070802468715

Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., ... & Hirschberg, J. (1992). TOBI: a standard for labeling English prosody. In *The Second International Conference on Spoken Language Processing, ICSLP 1992, Banff, Alberta, Canada, October 13-16, 1992*.

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

Yoon, T., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using toBI. In *INTERSPEECH*, 2729- 2732.