Predicting confidence and doubt in accented speakers: Human perception and machine learning experiments

Xiaoming Jiang¹,², Marc D. Pell¹

¹School of Communication Sciences and Disorders, McGill University, Canada
²Department of Psychology, School of Humanity, Tongji University, China
xiaoming.jiang@mail.mcgill.ca; marc.pell@mcgill.ca

Abstract
Speech prosody provides salient and reliable cues to facilitate social communication. What computational mechanism underlies social judgment towards “out-group” speakers is unclear. This paper focused on Speaker Confidence, a factor affecting one’s trustworthiness, persuasiveness and feeling of (un)knowing, and Speaker Accent, a factor marking one’s identity. We demonstrate that native Canadian-English listeners can recognize confident and doubtful expressions in foreign- and regional-accented speakers. A stronger impression of confidence was shown towards the native speakers. The acoustic analysis demonstrated that speakers systematically varied the mean fundamental frequency to indicate confident and doubt regardless of accent. The out-group speakers varied more on intensity height and variation to achieve certain level of confidence. Machine learning experiments showed above-chance accuracies in all accents to classify vocal expression based on global acoustic cues, highlighting the role of acoustic regularities at utterance level in confidence encoding. Moreover, the classification rate was higher when the model trained in native accent was tested on the native than the regional accent, highlighting an in-group bias of predicting novel vocal expression of confidence from acoustic cues. These findings lend support to the dialect theory of vocal expression recognition while demonstrating a computational mechanism underlying inter-cultural/inter-group confidence perception via speech prosody.

Index Terms: machine learning, ensemble tree model, speaker accent, vocal confidence, cross-cultural perception

1. Introduction
Most studies on vocal communication have focused on the role of emotional state on the use of verbal and nonverbal cues in social communication. Not until recently, growing behavioral and neuroimaging evidence has revealed an exciting new line of research - how the speaker’s mental or meta-cognitive state (e.g. confidence and doubt) shapes the listener’s impression towards the speaker. Confidence (or feeling of knowing) reveals to what extent the speaker commits to what they know and what they think (e.g. very sure or less certain). These studies highlighted the role of acoustic cues (such as pitch, intensity, duration and voice quality), contextual factors (the congruency verbal cues) and listener’s characteristics in predicting the perceptual outcome and underlying neurocognitive mechanisms [1-5]. One intriguing question is how speaker identity has an impact on the encoding their feeling of (un)knowing, and how the recognition of who the speaker is would bear any perceptual consequences on the listener, who shares or does not share the identity of the speaker [6-7]. Here, we focused on the speaker accent, a linguistic fingerprint that marks a person to be different from others in respect to the geographic location, social class, cultural orientation, etc. While independent studies showed that a foreign accented speaker increased one’s sensitivity towards the global and contextual cues in understanding speaker meaning, and may shift one’s strategy to deal with the counterpart’s interactive point, the evidence on regional accent suggests that a familiar, intelligible and native dialect (e.g. American English) can be immediately used by the listener to recognize who is speaking and what is said. These findings highlight the linguistic or cultural bias of the listener in recognizing and interpreting how the listener conveys their knowledge and opinion in their native tone, as compared with that from an “out-group” speaker. However, whether and how the in- and out-group perception would shape how listeners decode speaker confidence from the vocal cue is an unaddressed question [8-9].

This study focuses on what acoustic cues are essential in predicting one’s feeling of (un)knowing in a simple statement spoken in the same or a different accent from the listener. We are also interested in how computational mechanisms based on these cues are influenced by the speaker out-group accent. To classify confidence and doubt in speech, we trained novel machine learning (ML) models (e.g. the ensemble tree algorithm) to make predictive inference on the samples that the models never encountered, a mechanism that simulates the rapid social perception based on novel vocal cues during human vocal communication [7]. The classification rate on the perceptually-valid speaker confidence is calculated to compare how accurate the decision is made based on vocal cues of an accent that is familiar or new to the model.

2. Perceptual-acoustic study

2.1 Perceptual study
Material and paradigm: Auditory recordings of three accent groups were judged by Canadian-English speakers. The procedure of eliciting vocal confidence in native Canadian-English speakers was followed to obtain the recording in the accented speakers [4][10]. Statements of native accents were produced by two Canadian-English speakers. Two native Quebecois-French speakers (‘foreign accent’) and two Australian-English speakers (‘regional accent’) who stayed in Canada for less than 1 year were invited to produce the statements in English. 151 short sentences of communicative functions (Stating fact: He has a good sense of humor; Expressing intentions of an action: I can fix it for you; Judging a person or an object: He likes all types of seafood) were produced in scenarios highlighting the speaker is very
confident, almost confident but not 100% sure, very unconfident, or feel no emotion about what is stated.

Sixty sets of statements, including very confident, very unconfident and neutral-intending expressions were pre-selected by a native Canadian-English speaker, based on the quality of the recording and the match of the intended level of confidence. On each utterance, eighteen native Canadian-English listeners (10 Female/ 8 Male; Mean ± SD Age: 21.39 ± 2.59; Years of Education: 14.67 ± 2.03) were asked to 1) judge whether the speaker conveyed some level of confidence; 2) and to rate how confident the speaker is when portraying that utterance on a 5pt scale (1- the least confident). The linear mixed effects modeling (LMEM) were built with the intended level of confidence and speaker accent as two fixed factors and vocal stimulus as the random effect.

Result: The LMEM revealed a significant effect of accent, F(2, 471)=68.48, p<.0001, and an effect of confidence, F(2, 471)=1063.84, p<.0001. The perceived confidence increased from unconfident, neutral to confident-intending expressions (p<.0001). The native accent was more confident as compared with the regional and the foreign accent (b=-.27, t=-6.28, p<.0001; b=-.51, t=-11.69, p<.0001, native as baseline). There was also an interaction between confidence and accent, F(4, 471)=44.17, p<.0001, with the native accent perceived as more confident than other accents in the confident-intending (Regional: b=-.57, t=7.32, p<.0001; Foreign: b=-1.08, t=13.93, p<.0001) and neutral-intending expressions (b=-.47, t=6.47, p<.0001; b=-.71, t=-98, p<.0001), and less confident in the unconfident-intending expression (b=2.2, t=2.54, p=.01; b=2.7, t=3.14, p=.002). These findings suggest that the neutral-intending expression was perceived as close-to-confident in both native and accented speakers, and that an ingroup bias led to a stronger impression of confidence for the native than the accented speakers when they were intended to be confident.

2.2 Acoustic analysis

The perceptual ratings informed that speakers with out-group accents were naturally perceived as less expressive even when they followed the exact same procedure of eliciting different levels of confidence in their native language (Regional, Australian). However, when the recordings of the regional accent were perceived by a group of Australian-English speakers (n = 6), the perceived confidence showed a pattern similar to the judgment from the Canadian-English speaker. To identify acoustic cues that are associated with a perceived meaning, perceptually valid stimuli were selected with the constraints that 1) all were perceived as conveying some level of confidence by at least 13 out of 18 listeners in the perceptual studies (>72%); 2) the mean confidence rating was above 3.1 for the higher confidence condition (“High”); 3) the mean confidence rating was below 2.9 for the lower confidence condition (“Low”). These constraints were applied to all accents. We did not select those with mean confidence rating higher than 4.1 given the unbalanced number of trials between accents. The resulting stimuli were 63 (High) and 53 (Low) for the native-accented speakers, 57 (High) and 50 (Low) for the foreign-accented speakers, and 65 (High) and 53 (Low) for the regional-accented speakers. Speaker gender was balanced across levels of confidence and accents. LMEM models were built on each acoustic parameter with accent and perceived confidence as two fixed factors, the item within a combination of the fixed factor as the random factor.

The global acoustic parameters were extracted from each utterance per speaker: including the mean fundamental frequency (mean f0), the variation of fundamental frequency (range of f0), the mean amplitude (mean amplitude), the variation of amplitude (range of amplitude), duration of utterance (duration), the mean harmonic-noise-ratio (HNR) and the standard deviation of HNR. The pitch floor and pitch ceiling were customized for female and male speakers to extract global pitch values. The f0 and amplitude values were normalized per speaker with the average of the minimum f0 across utterances as baseline [4]. Differences between high and low perceived confidence are demonstrated in Figure 1 for Foreign accent and Figure 2 for Native accent.

![Figure 1: Global acoustic patterns on the Foreign accent](image1.png)

**Figure 1: Global acoustic patterns on the Foreign accent**

**Mean f0:** There were significant main effects of accent, F(2, 263)=5.87, p=.003, and perceived level of confidence, F(1, 263)=141.99, p<.0001. The mean f0 was higher in the native speakers than in the foreign accented speakers (b=-.06, t=-3.82, p=.0002, native as baseline). The mean f0 was also higher in the vocal expression perceived to be less confident (b=-.15, t=-11.89, p<.0001, low-confidence as baseline). No interaction was found between accent and perceived confidence.

**Range of f0:** Both effects of accent, F(2, 263)=3.41, p=.03, and perceived confidence, F(1, 263)=9.80, p=.002, were significant. Foreign accented speaker showed a lower range of f0 as compared with the native speaker (b=-.20, t=-2.02, p=.05). The expression perceived as more confident showed lower range of f0 than those of higher confidence (b=-.25, t=-3.13, p=.002). No interaction was found between accent and perceived confidence.

**Mean amplitude:** There was a main effect of accent, F(2, 263)=286.99, p<.0001 and perceived confidence, F(1, 263)=11.96, p<.0001. The mean amplitude was lower in the native accent than in Regional and Foreign accents (b=17,
terms of the model, the difference is how we train them. XGBoost provides a more regularized model formalization to control over-fitting, which gives it better performance. All predictive modeling was applied with the caret package within R 3.3.3.

The data sample per accent was split into 70% and 30% of the total sample utterances for training and testing purposes. To reduce the variation of testing accuracy for the training models, a 10-fold cross-validation method was implemented such that the training sample was further divided into 10 subsets and each dataset served in validation set once and in training set 9 times to allow an optimization of the training models. Such cross-validation method was repeated three times. The tuning parameters included the combination of the number of trees and the maximum depth per tree. The highest accuracy of validation was used to determine the best model. Nine machine learning experiments were performed in which training and testing were conducted on each of the accent (Table 1).

We included all global factors given that these measures were determined to be associated with vocal expression of different emotions or attitudes in acted speeches [4][12-15], and had been demonstrated as critical cues to predict confidence and doubt in the native speaker [4]. The same feature sets were used for all models to enable comparisons of results across models.

### Table 1: Classification Rates (% Accuracy, sensitivity and specificity of XGBoost Tree models as a function of the training and testing samples: (a) training on native speaker and testing on native speaker; (b) training on Foreign-accented speaker and testing on native speaker; (c) training on Regional-accented speaker and testing on native speaker; (d) training on Foreign-accented speaker and testing on Foreign-accented speaker; (e) training on native speaker and testing on Foreign-accented speaker; (f) training on Regional-accented speaker and testing on Foreign-accented speaker; (g) training on Regional-accented speaker and testing on Foreign-accented speaker; (h) training on native speaker and testing on Regional-accented speaker; (i) training on Foreign-accented speaker and testing on Regional-accented speaker.

<table>
<thead>
<tr>
<th>Training/Testing sample</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Native/ Native</td>
<td>81</td>
<td>.87</td>
<td>.88</td>
<td>.36</td>
</tr>
<tr>
<td>(b) Foreign/ Native</td>
<td>81</td>
<td>.63</td>
<td>.93</td>
<td>.44</td>
</tr>
<tr>
<td>(c) Regional/ Native</td>
<td>.74</td>
<td>.50</td>
<td>.51</td>
<td></td>
</tr>
<tr>
<td>(d) Foreign/ Foreign</td>
<td>.72</td>
<td>.60</td>
<td>.67</td>
<td>.53</td>
</tr>
<tr>
<td>(e) Native/ Foreign</td>
<td>.78</td>
<td>.88</td>
<td>.67</td>
<td>.47</td>
</tr>
<tr>
<td>(f) Regional/ Foreign</td>
<td>.68</td>
<td>.76</td>
<td>.60</td>
<td>.56</td>
</tr>
<tr>
<td>(g) Regional/ Regional</td>
<td>.79</td>
<td>.73</td>
<td>.84</td>
<td>.45</td>
</tr>
</tbody>
</table>

### 3. Machine learning experiments

Supervised machine learning models were set up to examine what acoustic cues are essential in predicting the speaker confidence in accented speakers. Boosted trees algorithm (i.e. the extreme gradient boosting tree, XGBoost Tree) was applied to train the models to classify vocal expressions based on a set of acoustic cues. Outperforming other models (e.g. random forests and support vector machine, SVM) in supervised learning, the boosted decision trees algorithm is claimed to be less sensitive to outliers and is more suitable for models with fewer more powerful predictive features [11]. XGBoost is a tree ensemble model, making a stack of predictions from a set of classification and regression trees (CART). The final score of each observation is the sum of the prediction score of each individual tree. An important fact is that the trees try to complement each other. Boosted trees and related models such as random forests are not different in

\[ t=10.41, \ p<.0001; \ b=.39, \ t=23.64, \ p<.0001 \]. The mean amplitude was also lower when the vocal expression was perceived as more confident, \( b=-.05, \ t=3.47, \ p=.001 \). A significant interaction between accent and perceived confidence existed, suggesting that the reduced amplitude for the confident expression was only observed in Foreign and Regional accents (\( b=-.11, \ t=-4.14, \ p<.0001; \ b=-.05, \ t=-2.39, \ p=.02 \)).

**Range of amplitude**: There was a significant main effect of accent, \( F(2, 263)=327.86, \ p<.0001 \), suggesting that the native accent was lower in the range of amplitude than Regional and Foreign accents (\( b=.47, \ t=19.95, \ p<.0001; \ b=.57, \ t=23.64, \ p<.0001 \)). There was no main effect of perceived confidence, but a significant effect of interaction between accent and perceived confidence, \( F(2, 263)=5.92, \ p=.003 \). The analysis on each accent revealed a wider range of amplitude only in Foreign accent for the confident than the unconfident expression (\( b=-.09, \ t=-3.32, \ p=.001 \)). The amplitude was also lower when the vocal expression was perceived as higher in confidence and doubt in the native speaker [4] . The same factors were determined to be associated with vocal expression of different emotions or attitudes in acted speeches [4][12-15], and had been demonstrated as critical cues to predict confidence and doubt in the native speaker [4]. The same feature sets were used for all models to enable comparisons of results across models.
Overall classification rates for each machine learning model were shown in Table 1. The classification rate was significantly above chance level for the model both trained and tested on the native accent (a), the model trained on foreign and tested on native accent (b), the model both trained and tested on the foreign accent (d), the model trained on the native and tested on the foreign accent (e), the model both trained and tested on the regional accent (g).

The accuracy rate was significantly higher in the model both trained and tested on the native accent (model (a)) than the rate in the model trained on the native accent and tested on the regional accent (model (c), z=2.09, p<.05), which was also the lowest among all the models (Figure 3). The model (a) also revealed a significantly higher classification rate than the model both trained and tested on the foreign accent (model (e), z=1.75, p<.05).

<table>
<thead>
<tr>
<th>Acoustic cue</th>
<th>Native</th>
<th>Foreign</th>
<th>Regional</th>
</tr>
</thead>
<tbody>
<tr>
<td>normalized mean f0</td>
<td>100.00</td>
<td>100.00</td>
<td>98.67</td>
</tr>
<tr>
<td>duration</td>
<td>83.22</td>
<td>0.00</td>
<td>22.24</td>
</tr>
<tr>
<td>normalized range f0</td>
<td>65.09</td>
<td>7.07</td>
<td>27.81</td>
</tr>
<tr>
<td>normalized range amplitude</td>
<td>38.12</td>
<td>19.85</td>
<td>100.00</td>
</tr>
<tr>
<td>SD HNR</td>
<td>13.72</td>
<td>7.44</td>
<td>0.00</td>
</tr>
<tr>
<td>Mean HNR</td>
<td>3.41</td>
<td>4.18</td>
<td>2.53</td>
</tr>
<tr>
<td>normalized mean amplitude</td>
<td>0.00</td>
<td>38.02</td>
<td>13.16</td>
</tr>
</tbody>
</table>

Figure 3: Receiver Operating Characteristics (ROC) curves showing the accuracy of all models (areas below the curve). The relations between true positive rate (sensitivity) and the false positive rate (specificity) were illustrated. The training and the testing on the native accents revealed a curve following closest to the left-hand border and the top border of the ROC space.

The improvement of predictive accuracy by each acoustic predictor was shown for training models per accent in Table 2. The importance score was determined by the accuracy change before and after permuting a certain acoustic predictor. As demonstrated the normalized mean f0 was ranked top for all three accents. The combination of normalized mean f0 and duration was the highest contributor to the accuracy in the native accent. The combined mean f0 and mean amplitude contributed the most to the classification rate in the foreign accent. The combination of mean f0 and range amplitude contributed the most to the classification rate in the regional accent.

Table 2: The importance score of each acoustic parameter in each training model per accent. Variable importance evaluation functions were calculated for each acoustic predictor and were scaled to have a maximum value of 100. The acoustic parameter is ranked in a descending order according to the native accent.

<table>
<thead>
<tr>
<th>Importance Score (100%)</th>
<th>Native</th>
<th>Foreign</th>
<th>Regional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acoustic cue</td>
<td>Native</td>
<td>Foreign</td>
<td>Regional</td>
</tr>
<tr>
<td>normalized mean f0</td>
<td>100.00</td>
<td>100.00</td>
<td>98.67</td>
</tr>
<tr>
<td>duration</td>
<td>83.22</td>
<td>0.00</td>
<td>22.24</td>
</tr>
<tr>
<td>normalized range f0</td>
<td>65.09</td>
<td>7.07</td>
<td>27.81</td>
</tr>
<tr>
<td>normalized range amplitude</td>
<td>38.12</td>
<td>19.85</td>
<td>100.00</td>
</tr>
<tr>
<td>SD HNR</td>
<td>13.72</td>
<td>7.44</td>
<td>0.00</td>
</tr>
<tr>
<td>Mean HNR</td>
<td>3.41</td>
<td>4.18</td>
<td>2.53</td>
</tr>
<tr>
<td>normalized mean amplitude</td>
<td>0.00</td>
<td>38.02</td>
<td>13.16</td>
</tr>
</tbody>
</table>

4. Discussion

The acoustic analysis informed common and differential use of acoustic cues in encoding confidence and doubt by the native and accented speakers. Both native and accented speakers relied on the mean pitch, variation of pitch and vocal quality measures (HNR) to indicate confidence and doubt. The accented speakers varied more on the intensity values of the global utterance to signify the different levels of speaker meanings. Interestingly, the change in the pattern of harmonic-noise ratio in expressing one’s feeling of (un)knowing was affected by the accent: while the increased vocal fold control led to a clearer phonation for the unconfident voice in native English accents, the increased clarity in the phonation was observed in the confident voice in the foreign accent [16].

The machine learning models showed an above-chance classification rate of vocal expression of confidence and doubt when the training and testing were conducted on the global acoustic cues from the same accent, highlighting the role of acoustic regularities at utterance level in confidence encoding. Moreover, the “in-group” classification was more accurate: the classification rate was higher when training was conducted on the native accent than the regional accent, suggesting an in-group bias of predicting novel vocal expression of confidence from acoustic cues. The speaker accent seemed to modulate the “in-group” classification rate, showing a reduced accuracy in classifying confidence from foreign accented speaker than from the native speaker. This finding was consistent with very recent evidence showing that the speaker categorization affects rapid extraction of social information from non-native speech, with a higher weight placed on the contextual cues or in-depth analysis of acoustic cue against social meaning [6].

Finally, we showed the importance score per acoustic parameter in the training model: regardless of speaker accent on which the model was trained, the mean fundamental frequency was reliably used by the Canadian-English listener to predict the speaker confidence. However, while durational feature contributed to a larger extent in the native accent, the mean and range of speech intensity contributed more in the foreign and regional accent in differentiating the confidence and doubt. These findings highlight computational mechanisms underlying inter-cultural/group confidence perception in speech communication and lend support to the dialect theory of vocal expression recognition [17-21].

5. Acknowledgements

We are grateful to the Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (to MDP, RGPIN/2016-04373) and to the McLaughlin Scholarship from McGill University.
6. References