Quantifying and Evaluating the Impact of Prosodic Differences of Foreign-Accented English

Hansjörg Mixdorff 1 and Murray J. Munro 2

1 Department of Computer Science and Media, Beuth University Berlin, Germany
2 Department of Linguistics, Simon Fraser University, Vancouver, Canada
mixdorff@beuth-hochschule.de, mjmunro@sfu.ca

Abstract

The identification and correction of prosodic deviations in second-language speech still poses a significant challenge for computer-aided language learning. With this ultimate goal in mind, the current study compares utterances by Cantonese speakers of Canadian English with those of native English subjects through both acoustic analysis and perceptual evaluation. We aim to find measurable prosodic differences accounting for the perceptual results. Our outcomes indicate, inter alia, that unstressed syllables are relatively longer compared to stressed ones in the Cantonese corpus than in the Canadian English corpus. Furthermore, the correlations of syllabic durations in utterances of one and the same sentence are much higher for Canadian English subjects than for Cantonese speakers. The latter use a similar range of F0, but produce more and longer pitch-accents than Canadian English speakers. In a perception study we found that applying native durations together with F0 contours to the foreign-accented speech led to significantly improved listener judgments of prosodic goodness. Adjustments to duration alone also tended to yield better ratings, though the effect was not statistically significant. When durations of native English utterances were adjusted to those of Cantonese speakers, significant decrements in ratings were observed.

Index Terms: foreign accent, prosodic analysis, perception tests

1. Introduction

Although foreign accent is commonly attributed to segmental deviations from the native norm, prosodic differences certainly account for many difficulties in understanding accented speech (see, for instance, [1][2]). In the current paper we examine speech from Cantonese users of Canadian English collected for two earlier studies [3][4]. We perform an acoustic prosodic analysis of the material and compare their speech with corresponding utterances by native Canadian English subjects in order to establish objective parameters that best reflect foreign accent, and are correlated with listeners’ judgments of prosodic accuracy. Whereas English is often classified as a stress-timed language, Cantonese is a syllable-timed tone language, a contrast which poses a number of prosodic problems for learners of the other language.

Although pedagogical specialists often identify accurate prosody as a critical aspect of pronunciation teaching [5], so far only a handful of empirical studies have evaluated its role in non-native speech. Hahn [6] reported that accurate stress placement enhanced listeners’ recall of main ideas from L2 speakers’ utterances; Derwing, Munro & Wiebe [7] found that instruction on prosodic features improved comprehensibility of L2 narrative productions, while segmental instruction did not; and Tajima, Port, and Dalby [8] observed that temporal adjustments to L2 speech increased intelligibility.

In the first part of the current paper we present a comparative acoustic analysis of prosodic features of Cantonese and Canadian English speakers. The second part explores the perceptual impact of native prosody in terms of duration and F0 transplanted to the utterances of Cantonese speakers, as well Cantonese duration characteristics applied to Canadian English speakers.

2. Speech Material and Method of Analysis

The original corpus consists of readings of a short English passage adapted from [9] by 77 native adult speakers of Cantonese with English pronunciation skills ranging from poor to good. All were born in Hong Kong and had moved to Canada after the age of 16, where they had been residing for 1 to 4 years. We selected this pre-existing corpus for the large number of available subjects as well as the fact that the reading-style material facilitated sentence-wise comparisons.

Recordings were made in a sound-treated booth and sampled at 22.05 kHz/16 bit. To identify a suitable subset of items for the current study, the second author rated the prosodic performance (rhythm and intonation) of individual participants on a scale from 1–9 (1 = excellent; 9 = very poor). This preliminary assessment was designed only to ensure that a final stimulus set with considerable variability in prosodic goodness would be obtained. The same study was read by 32 native speakers of Canadian English. For the purposes of this study we selected 41 speakers of Cantonese representing the full range of performance ratings, 20 male and 21 female, as well as 30 Canadian English speakers, 15 male and 15 female. Results presented in this paper are based on the first five sentences from the short story with a total duration of between 11 and 22 seconds. The part of the data uttered by Cantonese speakers shall henceforth be referred to as CANT, that of the Canadian English speakers as CNDE.

In the first step, all recordings were force-aligned at the syllable-level using an HTK [10] based system provided by Yuanfu Liao, NTUT, Taipei, trained on the TIMIT corpus [11]. The target of alignment was a canonical Sampa transcription of the underlying text, produced with the grapheme phoneme converter of the first author’s multilingual TTS system [12]. Strictly speaking, due to reductions, especially in the native speakers’ utterances, not all of the identified segments corresponded to phonetic syllables. However, for the sake of the following acoustic comparison it was required that the label sequence be identical for all utterances of the same sentence. The automatic segmentations were converted to PRAAT TextGrid format [13] and syllabic boundaries hand-corrected. We were not interested in the identity and exact boundaries of phones actually realized, but rather in the rhythmic structure of the utterances.
In order to analyse the intonational properties of the two corpora, F0 values were extracted at a step of 10 ms using the PRAAT default pitch extraction settings and subjected to manual inspection and correction. Syllables exhibiting creaky voice were marked on the syllabic tier.

All utterances were subjected to Fujisaki model [14] parameter extraction [15] as shown in Figure 1 (sentence 1, produced by Cantonese speaker CP004 and Canadian English speaker EN02, both male). The figure displays the following, from the top to the bottom: the speech waveform, the F0 contour (+ signs: extracted, solid line: model-based), the text, the underlying phrase and accent commands. This methodology has the great advantage of providing an accurate fit of the original F0 contour, while breaking it down into a limited number of parameters which can then be related to phonologically relevant prosodic landmarks such as accents and boundary tones. It is also very useful when resynthesizing stimuli with modified prosody, as it provides F0 values for each point in time. The amplitudes Ap of the box-shaped accent commands (see Figure 1) are correlated with the amount of emphasis given to an accented syllable, whereas the magnitudes Ap of the impulse-wise phrase commands reflect the amount of F0 reset at the onset of new prosodic phrase.

3. Prosodic Parameters and Results of Analysis

The objective of the analysis was to identify systematic differences between the CANT and CNDE data at the rhythmic as well as the intonational levels and relate them to the proficiency ratings.

Timing: When we examined how the number of speech pauses in an utterance influenced the expert rating, we observed a Pearson $r = .54$ ($p < .01$). The correlation is slightly less with the total duration of speech pauses in the passage ($r = .42$).

For further quantitative analysis, the syllabic labels from the CANT and CNDE corpora were compared with respect to means and standard deviations of durations, as well as rhythmic properties of the utterances. Analysis showed a considerably higher rate of 5.7 syllables/second for the CNDE speakers against 4.4 syllables/second for the CANT group. If we correlate the syllable rate of the learners with the expert rating on their prosodic performance, we obtain a Pearson $r$ of -.67, suggesting fluency as a key factor in the judgment.

We next investigated whether the syllable-timed nature of Cantonese as opposed to the stress-timing of English also affected the realizations of the Cantonese speakers. Based on the text underlying the utterances we categorized all syllables as belonging to one of three classes: Unstressed, stressed (verbs) and stressed (other content words). Due to the fact that verbs often become deaccented in a sentence context this classification was mainly based on the superordinate part-of-speech. Although individual realizations varied, results showed that stressed syllables in the CNDE corpus were relatively longer than unstressed syllables than in the CANT corpus. The following table shows the results of comparison.

<table>
<thead>
<tr>
<th>Group</th>
<th>Syllable type</th>
<th>Mean [ms]</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>CANT</td>
<td>unstressed</td>
<td>181</td>
<td>1230</td>
</tr>
<tr>
<td></td>
<td>stressed (verbs)</td>
<td>280</td>
<td>451</td>
</tr>
<tr>
<td></td>
<td>stressed (others)</td>
<td>310</td>
<td>451</td>
</tr>
<tr>
<td>CNDE</td>
<td>unstressed</td>
<td>121</td>
<td>840</td>
</tr>
<tr>
<td></td>
<td>stressed (verbs)</td>
<td>219</td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>stressed (others)</td>
<td>271</td>
<td>315</td>
</tr>
</tbody>
</table>

The ratio of mean durations unstressed/stressed is 0.45 for the Canadian English speakers, whereas it is 0.58 for the Cantonese speakers. The values are 0.55 versus 0.65 for the stressed syllables in verbs. This suggests that Cantonese speakers are aware of the English stress system, yet they tend to produce syllables of more uniform lengths than the Canadian English speakers.

Looking more closely at the rhythmic patterns of individual sentences we correlated the syllabic durations in one realization of a sentence with the syllabic durations in all the other realizations of the same sentence. The advantage of this approach is that the effect of the speech rate on this measure is rather small. In fact, this measure was previously used for evaluating the quality of a duration model for text-to-speech synthesis [16] and later applied to the analysis of foreign-accented English [17].

By averaging the inter-utterance correlations over all utterances of a given sentence for each group (CANT, CNDE) we obtain a measure of similarity within that group, as well as the mean inter-group correlation. Results indicate that the CNDE realizations (mean $r = 0.91$) are much more similar in their rhythmic structure (more highly correlated) than the CANT ones (mean $r = 0.75$). Also the cross-correlation between the two groups is only moderate (mean $r = 0.77$).

In order to test whether the observed sentence-based correlations were valid indicators of prosodic goodness, we calculated the centroid of all CNDE utterances for each sentence. That is, for each syllable in a given sentence we averaged over all observed instances in the CNDE data set, yielding prototypical syllabic durations for each sentence.

Figure 1: Example of analysis of sentence 1, uttered by a Cantonese (left) and Canadian English speaker (right), both male: “On Wednesday morning at 11 o’clock I was walking down Main Street.” From the top to the bottom: speech waveform, F0 contour (+extracted, -modeled), underlying phrase and accent commands. Vertical lines indicate boundaries of text syllables. Fb is denoted by the horizontal line underneath the F0 pattern.
(“duration norm”). Subsequently we calculated the correlations between each of the Cantonese utterances and their corresponding Canadian English duration norm. Statistical analysis showed that this rhythmic correlation was significantly \((r = -0.479, p < 0.01)\) correlated with the original prosodic goodness rating on the 9-point scale. We have to take into account that the Cantonese data spans a range of proficiency levels so that some speakers might already have attained very high rhythmic proficiency, with others exhibiting almost “Cantonese” rhythm.

**Intonation:** As mentioned earlier, the extracted \(F0\) contours were parameterized using the Fujisaki model in order to establish the differences between the \(CNDE\) and \(CANT\) data sets. To this effect, automatic parameter extraction was performed [15]. Then the analysis results were inspected and, if necessary, corrected using the interactive FujiParaEditor [18]. As can be seen from the examples in Figure 1, the resulting model \(F0\) contours are very close copies of the original natural ones, the RMSE being less than 2 semi tones.

If we look at mean \(F0\) we find that it is generally higher for Cantonese than for Canadian English speakers, for both male (127 vs 115 Hz) and female subjects (220 vs 197 Hz). 3.5\% of syllables in \(CNDE\) exhibit creaky voice, and 3.1\% in \(CANT\). This difference, however, is mostly due to the Canadian male speakers.

Some numerical results of analysis are displayed in Table 2, which shows means and standard deviations of accent command amplitude \(Aa\) and duration for the \(CANT\) and \(CNDE\) data. As can be seen - though mean amplitudes of \(Aa\) are quite similar - the \(CANT\) group produce longer accent commands than the Canadian English speakers. These appear as plateau-like gestures spanning several syllables at high pitch (compare Figure 1, left)

<table>
<thead>
<tr>
<th>group</th>
<th>(Aa)</th>
<th>duration [ms]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNDE</td>
<td>mean</td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td>s.d.</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>443</td>
</tr>
<tr>
<td>CANT</td>
<td>mean</td>
<td>.26</td>
</tr>
<tr>
<td></td>
<td>s.d.</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>785</td>
</tr>
</tbody>
</table>

However, the Cantonese speakers rephrase more frequently, on the average once every 5.9 syllables compared to 7.1 syllables for the Canadian English speakers. This result is partly due to the higher speech rate of the \(CANT\) group (compare Figure 1), as well as the lower fluency of the \(CANT\) group, who insert additional phrase boundaries.

**4. Stimuli and Design of the Perception Study**

Stimuli for perceptual evaluation were created by applying a number of processing steps to the original speech recordings. We selected seven \(CANT\) speakers, as well as two \(CNDE\) subjects, EN11 and EN14, to serve as “donors” of the prosodic features to be transplanted to the Cantonese-accented recordings, while utterances from four other \(CNDE\) subjects were added to serve as anchoring points for the listeners. Table 5 shows a list of the speakers whose utterances were chosen for prosodic manipulations, as well as some of their prosodic characteristics. It also lists the expert prosodic goodness ratings for the Cantonese speakers.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Sex</th>
<th>Rating</th>
<th>Mean Aa [syll. dur. (ms)]</th>
<th>Mean Ap [corr. w. dur. norm]</th>
<th>Mean Fb [Hz]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CP015</td>
<td>M</td>
<td>7</td>
<td>2.84</td>
<td>0.57</td>
<td>0.17</td>
</tr>
<tr>
<td>CP023</td>
<td>F</td>
<td>2</td>
<td>2.07</td>
<td>0.39</td>
<td>0.20</td>
</tr>
<tr>
<td>CP025</td>
<td>F</td>
<td>6</td>
<td>2.12</td>
<td>0.70</td>
<td>0.14</td>
</tr>
<tr>
<td>CP027</td>
<td>F</td>
<td>5</td>
<td>2.05</td>
<td>0.91</td>
<td>0.21</td>
</tr>
<tr>
<td>CP028</td>
<td>M</td>
<td>1</td>
<td>1.83</td>
<td>0.95</td>
<td>0.16</td>
</tr>
<tr>
<td>CP034</td>
<td>M</td>
<td>8</td>
<td>2.90</td>
<td>0.78</td>
<td>0.35</td>
</tr>
<tr>
<td>CP053</td>
<td>M</td>
<td>8</td>
<td>2.66</td>
<td>0.92</td>
<td>0.40</td>
</tr>
<tr>
<td>EN11</td>
<td>F</td>
<td>native</td>
<td>1.86</td>
<td>0.96</td>
<td>0.36</td>
</tr>
<tr>
<td>EN14</td>
<td>M</td>
<td>native</td>
<td>1.78</td>
<td>0.95</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Table 2: Mean and standard deviation of accent command duration.

Table 3: Means and standard deviations of accent command amplitude \(Aa\) associated with stressed syllables in verbs and other content words.

<table>
<thead>
<tr>
<th>stress</th>
<th>group</th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>stressed</td>
<td>(CANT)</td>
<td>485</td>
<td>.21</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td>(CNDE)</td>
<td>307</td>
<td>.20</td>
<td>.15</td>
</tr>
<tr>
<td>stressed</td>
<td>(CANT)</td>
<td>424</td>
<td>.23</td>
<td>.16</td>
</tr>
<tr>
<td></td>
<td>(CNDE)</td>
<td>291</td>
<td>.25</td>
<td>.18</td>
</tr>
</tbody>
</table>

If we look at the frequency of accent commands there are 1.75 commands per second in the \(CNDE\) group, but 1.60 for the

**Table 4:** Mean and standard deviation of phrase command magnitude \(Ap\).

<table>
<thead>
<tr>
<th>group</th>
<th>mean (Ap)</th>
<th>S. D.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>(CNDE)</td>
<td>.29</td>
<td>.14</td>
<td>206</td>
</tr>
<tr>
<td>(CANT)</td>
<td>.29</td>
<td>.13</td>
<td>363</td>
</tr>
</tbody>
</table>

The syllable-based frequency is one command every 2.7 syllables in the \(CANT\) group, but one command every 3.3 syllables in the \(CNDE\) data.

If we calculate mean \(Aa\) for stressed syllables in verbs as opposed to other content words we find that these values are quite similar within the Cantonese group, as the means differ by only 8\%, but for the Canadian English speakers, the difference is almost 29\%, that is, verbs generally receive lower prominence than other content words (see Table 3).
The following types of stimuli (see Table 6) were created:

(1) **Plain resynthesis:** In the PR condition, all utterances were resynthesized using the Fujisaki model-based \( F_0 \) contours, employing the FujiParaEditor and the PRAAT PSOLA resynthesis capability.

(2) **Duration modification A:** Durations of Cantonese stimuli were adjusted to match those of the Canadian English target speakers EN11 and EN14 (conditions DA11 and DA14, respectively), and the Canadian English duration norm DNE. In order to compensate for the change in \( F_0 \) slope due to the higher target speech rate, the amplitudes of the Fujisaki model parameters for the CANT utterances were rescaled accordingly.

(3) **Duration and \( F_0 \) modification:** Utterances from (2) whose duration characteristics now matched those of EN11 and EN14, were further manipulated with respect to \( F_0 \), by applying the original Fujisaki model parameters of EN11 and EN14, respectively, giving conditions DF011 and DF014. The base frequency \( F_0 \) was kept at the original value of the Cantonese speakers. Hence the resulting \( F_0 \) contour had the shape of the donor speaker’s production, while the \( F_0 \) values remained within the range of the Cantonese speaker. This also facilitated the “cross-gender” modifications.

(4) **Duration modification B:** The durations of EN11 and EN14 were adjusted to those of Cantonese speakers CP015, CP034 and CP053 as in (2), giving the DB condition. They were also adjusted to half-way between source and target; that is EN11 and EN14 were slowed down and syllable duration ratios and pause structures of the Cantonese targets applied, but the resulting utterance duration was the average between the original Canadian English and the targeted Cantonese utterance. This approach, the DBH condition, was taken because further deceleration tended to render the CNDE utterances extremely slurred and unnatural.

Finally, in order to equalize the quality of the stimuli and tone down some artifacts of time stretching and pitch shifting, all stimuli were filtered to telephone bandwidth (340-4000 Hz) and resampled at 8k Hz.

The recordings were prepared for perceptual evaluation by separating the five sentences in all stimuli into individual tokens. These were informally prescreened by two research associates, who deemed that any distortions and artifacts in the CANT tokens were minimal. Several of the CNDE tokens in the DB condition, however, were judged to sound too unnatural for use. On the basis of the prescreening, only the fourth sentence from speakers CP015 and CP053 was selected for inclusion in both the DB and DBH conditions. The final stimulus set thus consisted of the 248 stimuli summarized in Table 6.

Twelve native speakers of Canadian English, all with training in phonetics, were recruited as judges. During prescreening it was found that the 9-point evaluation scale used by the expert rater for the original stimulus set was too large to allow satisfactory ratings of the items in the subset. We therefore refined our approach such that the judges rated each production for prosodic goodness on a 5-point scale (1 = poor, 5 = excellent) by focusing on intonation and temporal characteristics. Stimuli were blocked on sentence and in self-paced sessions via headphones under quiet conditions. On hearing each item, the listeners responded via a keyboard press. After a 12-item warm-up session, each token was judged twice in a randomized presentation; the order of blocks was counterbalanced across listeners. Prior to assigning each rating, listeners were allowed to replay the stimulus up to three times. Total time for the task was about 35 minutes, though rest breaks between blocks were permitted.

### 5. Results of the Perception Study

Mean ratings for the two presentations of each item were computed for each listener. The intraclass correlation for the resulting scores was .924 (\( p < .001 \)), indicating a high level of inter-judge reliability. For the purposes of the following analyses, ratings were pooled over the five sentences to yield a mean score by speaker for each condition. A paired samples \( t \)-test revealed that the plain resynthesized (PR) CNDE tokens were rated significantly higher (\( M = 4.7 \)) than the CANT \( PR \) (\( M = 2.8 \)) productions, \( t(11) = 13.49, p < .001 \).

Figure 2 shows the mean goodness ratings on the CANT productions for each stimulus condition. In all manipulated conditions, the values were higher than in the PR condition. For the simultaneous manipulations of duration and \( F_0 \), DF014 and DF011, increases of .78 and .71, respectively, were observed, while for the duration-only manipulations, DNE, DA14, and DA11, the increases were smaller, with values of .17, .2, and .17, respectively. A one-way repeated measures ANOVA revealed a significant effect of condition, \( F(5, 55) = 15.43, p < .001 \). However, according to post hoc Bonferroni-adjusted \( t \)-tests (\( p < .05 \)), only the DF014 and DF011 conditions yielded significantly higher scores. Also, scores in DF014 and DF011 were both significantly higher than those for the three other manipulated conditions. No other pairwise differences proved significant.

![Figure 2: Mean goodness ratings (with standard error) in the 6 conditions for the CANT productions](image-url)
manipulations. Figure 3 presents mean ratings across all conditions for the speakers with the highest and lowest scores in the PR condition. Data for speaker CP028, who was rated almost as high as some of the native English speakers in the PR condition, indicate only small differences in ratings, while those for the lowest-rated speaker in PR, CP015, show greater variation.

![Figure 3. Mean scores for (in order) the PR, DNE, DA14, DA11 and DF011 conditions for the speakers with the highest (CP028) and lowest (CP015) ratings in PR.](image)

Mean goodness ratings for the CNDE productions are given in Figure 4. In all cases, ratings of the manipulated stimuli were lower, with differences from the PR condition ranging from .93 to 1.64. A repeated measures ANOVA once again yielded a significant effect of stimulus condition, F(4,44) = 16.86, p < .001, with post hoc tests indicating that all manipulated conditions had significantly lower ratings than PR. However, none of the other pairwise differences were significant.

![Figure 4. Mean goodness ratings (with standard error) in the 5 conditions for the CNDE productions](image)

6. Discussion and Conclusions

The current study concerned the prosodic analysis of accented English speech data produced by Cantonese users of English. We found that the number of pauses in an utterance, a typical indicator of disfluency, as well as the speech rate, were correlated with expert perceptual judgments of prosodic goodness. At the rhythmic level, Cantonese learners of English produced relatively longer unaccented syllables than Canadian English speakers, which suggests that their rhythm was influenced by the syllable-timed structure of Cantonese. The syllabic durations in the Canadian English group were more uniform than within the Cantonese group expressed by the durational correlations between individual productions of the same sentence. At the intonational level, Cantonese speakers produced comparable excursions of F0 and used a similar range of F0 as the Canadian English controls. They placed pitch accents more frequently and exhibited less vocal fry than their Canadian English counterparts. We found that the degree of prominence in terms of accent command amplitude, Aa, assigned to stressed syllables was more uniform in the Cantonese subjects, whereas it was considerably lower for the Canadian English speakers in the case of verbs. This might suggest that though the Cantonese learners were aware of the stress system, they did not replicate the information structure the same way as the native speakers.

In general, the results of the perceptual study indicated that transplanting native-like prosodic characteristics on Cantonese-accented English speech led to better judgments from phonetically sophisticated listeners. For the sample of speakers under study here, statistically significant improvement was observed for simultaneous manipulation of duration and F0, but not for duration alone, although there was a tendency toward better ratings in the duration-only manipulations. It is important to note that the degree of improvement appears to have depended on the nature of the speech prior to manipulation, with speech that was already prosodically good showing less improvement than speech that was especially poor. Expansion of the stimulus set to include more Cantonese speakers with poor prosody may uncover more subtle, but statistically significant, effects of duration manipulations alone. Also worthy of note is the fact that the effects of the manipulations did not vary according to donor speakers. Again, expansion of the study will be useful in establishing whether this is generally the case. The magnitude of improvement yielded by transposing durations and F0 is comparable to that found in an earlier study [19] (.8 on a 5 point scale).

In the case of the duration manipulations of the native English productions, a significant decrement was seen in the ratings in all conditions. Despite a tendency for the items that were completely slowed to Cantonese-accented rates to be judged somewhat worse than those slowed to a rate intermediate between native English and Cantonese-accented, the differences between manipulated conditions were not significant. However, only a limited number of stimuli were used because of the difficulty in creating natural-sounding stimuli in the completely slowed condition. Once again, there was no indication of an effect of donor speaker.

Future work will entail perceptual experiments with segmentally and prosodically manipulated stimuli in order to identify the factors that contribute most to the perceptions of strong foreign accent and reduced intelligibility. In addition, the effects of the manipulations on phonetically untrained listeners should be evaluated to determine whether comparable patterns of perception are observed. In future work, we plan to evaluate the applicability of our findings for computer-aided pronunciation training. Work by Pfitzinger et al. has shown that re-synthesizing speech in the learner’s voice, but with corrected prosody can assist in training compound accents in German, for instance [20]. A variety of common language laboratory exercises could potentially make use of resynthesized utterances, including discrimination and
identification tasks, and shadowing. The use of the learners’ own speech in such tasks may prove especially beneficial in addressing individual speakers’ pronunciation needs.

7. Acknowledgements

This work was supported by DFG international collaboration grant no. Mi 625-17/1 funding Mixdorff’s stay at Simon Fraser University. We thank Gloria Mellesmoen for assistance with stimulus preparation and perceptual data collection. Thanks also go to Yuanfu Liao for providing the force-alignment segmentations.

8. References
