INTRODUCTION

Our language experience often involves face-to-face interaction through simultaneous perception of the speaker’s voice as well as facial and body movements. While it has been well established that watching a speaker’s lips in conversation can enhance speech perception (Sumby & Pollack, 1954; Massaro, 1987), recent research also shows facilitative effects of co-speech gestures and head movements (Krahmer & Swerts, 2007; Wagner, et al., 2004). Particularly relevant to the present research is the perception of prosodic (suprasegmental) features. While prosody has traditionally been believed to be acoustic/auditory, recent studies show that prosody may have visual articulatory or other facial correlates (Chen & Massaro, 2008; Lansing & McConkie, 1999). For example, movements of the head, neck and mouth are found to be informative for the perception of lexical tone (Burnham et al., 2001; Chen & Massaro, 2008), particularly when a listening condition was poor (Mixdorff et al., 2005).

In the perception of second language (L2) speech sounds, information from multimodalities can be particularly beneficial, as nonnative perceivers may need additional channels of input for processing challenging nonnative sounds (Hazan et al., 2006; Kelly et al., 2008; Wang et al., 2009). Indeed, co-speech hand gestures have also been shown to assist L2 learners’ perception (Kelly et al., 2009), especially when auditory speech is inaccessible to learners (Church, et al., 2004). However, the extent to which gestures may facilitate learning of L2 prosody has been less conclusive. For example, while beat gestures can aid L2 learners in parsing words into syllables (McCafferty, 2006), they are not as effective in discriminating durational differences (Hirata & Kelly, 2010). Moreover, the addition of gestural input may also be inhibitory as learners may experience excessive cognitive load (Hirata & Kelly, 2010).

In the present study, we trained adult Canadian English speakers to perceive Mandarin tones with one of the following three types of input modalities (1) audio-visual-gestural (AVG) training, using hand gestures in conjunction with the head and facial features that can accompany tonal speech, (2) audio-visual (AV) training with no gesture, or (3) audio-gestural (AG) training with no facial information. The gesture used here is the metaphorical gesture, which traces an imaginary tone as it changes in pitch along the dimension of time (duration) and height (pitch), as is commonly used in Chinese tone teaching environments. Our goal was to examine the extent to which multimodal speech input with gestural information facilitates L2 tone perception. We would expect gesture to be beneficial if AG training resulted in significant improvement in tone perception. On the other hand, if the addition of gestural input caused excessive cognitive load, we would expect AVG training to be less effective than AG or AV training. In this paper, we report the preliminary results of our ongoing research.

METHODS

As described above, three types of perceptual training were administered: The AV training group was presented with both speaker voice and facial movements for the tone word productions. The AG group was presented with the same auditory tonal stimuli along with the speaker’s arm/finger gestures exemplifying the contour level of the tones as they were spoken, while their face and body were blocked off. The AVG group perceived the stimuli from all three modalities. In order to assess the trainees’ improvements and learning trajectories, the program included a pretest before training and a posttest after training, along with three intersession tests each after two training sessions.

Participants

Speakers: Six speakers of Mandarin Chinese participated in our recordings. Four of these speakers (2 male, 2 female) provided the stimuli for training. They were instructors of standard Mandarin with five to fifteen years of Mandarin teaching experience, chosen because of their familiarity with training students and knowledge of the Mandarin tones. Additionally, there were also two standard Mandarin speakers (one male, one female) for the pre- and posttest stimuli.

Trainees: The participants for the training portion so far were 12 college students. All were native speakers of Canadian English and lived in the metro Vancouver area in the last ten years. These participants had no prior experience with any tonal languages or substantial musical training experience (with fewer than five years of musical experience, Cooper & Wang, 2012). All participants indicated by questionnaire that their vision was normal or correct-to-normal at the time of participation. Hearing and speech were also self-reported as normal. They were randomly assigned to AV, AG, or AVG training groups.
**Stimuli and Recording**

The training word list contained eighty Mandarin monosyllabic real words (20 tone quadruplets), which was a combination of the one used in Liu & Samuel (2004), as well as Wang et al. (1999). This list of eighty words was recorded twice for all speakers: first without hand gestures, and then a second time with the accompanying movements. The stimuli used for the pre- and posttests were 60 additional Mandarin monosyllabic real words (15 tone quadruplets) used in Wang et al. (1999).

For the recording, each word was presented in pinyin above the corresponding simplified Chinese character using Microsoft PowerPoint. To get the speakers familiarized to the task, each speaker recorded six real word quadruplets found in Wang et al (1999), with and without gestures, which were later excluded. All vocabulary was presented in a randomized order to avoid exaggeration of contrast in the production, following Liu & Samuel (2004). The speakers’ bodies from their collarbones to the tops of their heads were clearly visible, including the entire neck and face. Training stimuli speakers were instructed to trace an acetate graph representation (Chao, 1948) of the tones on the feedback screen of the digital camera with their right index finger so that the gestures would be standard across speakers. Speakers were asked to simultaneously speak the word and trace the contour lines for the same duration. Each speaker was allowed to repeat words as often as they wanted before advancing to the next word. The video clips were then edited with Final Cut Pro and segmented by syllable. Extrinsic noise was cut out and the video was mirrored so that the tone contour was presented in the correct direction for the trainees. One second of silence preceded and followed the syllable in each clip. Stimuli were not normalized for duration, as duration is one of the critical intrinsic properties of tone (Blicher, et al., 1990).

Five of the speakers and one native Mandarin speaker who did not participate in the recordings were brought in for an additional session to judge the goodness of each example using an identification task (but did not judge their own productions). All the stimuli used in training were correctly identified.

**Test and Training Procedures**

Training and testing were sequenced as follows:

Familiarization: Participants were familiarized with the Mandarin tones by listening to an audio-only, syllable quadruplet /ma/. This was immediately followed by the presentation of one syllable quadruplet and four-alternative forced choice task by pressing labeled keyboard keys to familiarize with the task.

Pre/posttest: Prior to and after training, participants were tested with auditorily presented tone words described above. Their task was four-alternative forced choice identification.

Training took place over the course of two weeks with six sessions of 40 minutes each, taking care to schedule participants at least one hour between each session. In total, 960 trials were used for training, including 6 training sessions x 4 speakers x 40 words, with the 80 words split into two sessions. Each session was balanced across tones, syllables and speakers. Training was presented in AV, AG or AVG, depending on the condition.

An intersession test was presented after each two training sessions, similar to the pretest. It used eighty words from the training blocks and was audio-only.

**RESULTS**

The trainees’ percent correct identification scores at pre- and post-training tests were compared using a two-way repeated measures analysis of variance (ANOVA), with Training Group (AV, AG, AVG) as the between-subject factor, and Test (pretest, posttest) as the repeated measure.

The ANOVA yielded significant effects of Test [F(1, 10)=29.029, p<.0001], showing that, across training groups, tone identification at posttest (85%) was more accurate than at pretest (51%). A significant Group difference was also observed [F(2,9)=6.34, p<.019], with Bonferroni post hoc analyses showing that, across tests, the AV group (81%) scored significantly higher than the AG group (54%, p<.018), whereas there was no significant difference between the AV and AVG (69%), or between AG and AVG groups. However, the results revealed no significant interaction between Test and Group [F(2,9)=6.34, p>.696], suggesting that all groups improved to a similar degree. The trainees’ mean correct identification scores at pre- and posttests are displayed in Figure 1.
FIGURE 1. Percent correct identification scores of the pre- and posttest by the three training groups (AV, AG, and AVG).

Given the small sample size of this preliminary study, individual data were examined. As shown in Table 1, two participants in the AG group had particularly lower than average pretest scores (25%, 13%) and one of them had a lower than average posttest score as well (35%). These individual differences may account for the group difference, particularly at pretest.

TABLE 1. Individual results (in % accuracy) of the pre- and posttests by the three training groups (AV, AG, and AVG).

<table>
<thead>
<tr>
<th>Training group</th>
<th>Participant</th>
<th>Pretest</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td>AV</td>
<td>501</td>
<td>68</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>502</td>
<td>60</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>503</td>
<td>55</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>511</td>
<td>71</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>512</td>
<td>82</td>
<td>97</td>
</tr>
<tr>
<td>AG</td>
<td>601</td>
<td>57</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>602</td>
<td>39</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>603</td>
<td>25</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>611</td>
<td>13</td>
<td>88</td>
</tr>
<tr>
<td>AVG</td>
<td>701</td>
<td>38</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>704</td>
<td>52</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>711</td>
<td>70</td>
<td>91</td>
</tr>
</tbody>
</table>

DISCUSSION AND CONCLUDING REMARKS

The preliminary results showed substantial improvement between pre- and posttests for all groups, consistent with the previous findings that speech input from multimodalities benefits speech learning (Hazan et al., 2006). Particularly, gestural information appears to have a comparable effect on perception as visual information given the comparable improvements for the AG and AV groups. Moreover, improvement for the AVG group was similar to the AV and AG groups as well. This provides evidence against perceptual overload from multimodality presentation, as hypothesized. While beat gestures have shown to hinder the perception of L2 durational contrasts (Hirata & Kelly, 2010), gestures tracing pitch trajectories appear to enhance the nonnative perception of tonal contrasts. The discrepancy of these results reveals the complex roles of gestures in speech perception. Unlike visual speech information which involves anticipated and fixed articulatory configurations for the resultant speech sounds, co-speech gestures may come in different forms which are not necessarily bound to speech, be it iconic, deictic metaphoric, or beat (McNeill, 1992). The effectiveness of gestures may thus be constrained by their specific relationship with speech features.
It is important to note that the positive role of hand gestures in the current report should be interpreted with caution. First and foremost, the three training groups did not start at a similar level; as the AG group scored much lower in the pretest than the two other groups. Testing the true level of improvement across training modalities requires starting levels of all groups to be similar, and the current results may be skewed due to the outliers in the AG group and the small sample sizes of all groups. It is suspected that this difference would dissipate in the full-scale study with larger sample sizes.

ACKNOWLEDGMENTS

We would like to thank Anthony Chor, Sonya Gill, and Courtney Lawrence from the Language and Brain Lab (Simon Fraser University, SFU) for their assistance in participant recruitments and data collection. We also thank the valuable input from Drs. Yukari Hirata and Spencer Kelly as well as the Language and Brain Lab members in developing this project. This research has been funded by research grants from the SFU Vice President Academic and the Social Sciences and Humanities Research Council of Canada to YW. KE, BH, and YW have made equal contributions to this paper.

REFERENCES


