Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

Brain Research

Research Report

Effects of production training and perception training on lexical tone perception – A behavioral and ERP study

Shuang Lu[a,*], Ratree Wayland[b], Edith Kaan[b]

[a]Laboratório de Fonética & BabyLab (CLUL), Departamento de Linguística, Faculdade de Letras, Universidade de Lisboa, Alameda da Universidade, 1600-214 Lisboa,Portugal

[b]Department of Linguistics, University of Florida, 4131 Turlington Hall, Gainesville, FL 32611-5454, USA

ARTICLE INFO

Article history:
Accepted 12 July 2015
Available online 20 July 2015

Keywords:
Speech perception
Speech production
Lexical tone
Training
Discrimination
ERP

ABSTRACT

The present study recorded both behavioral data and event-related brain potentials to examine the effectiveness of a perception-only training and a perception-plus-production training procedure on the intentional and unintentional perception of lexical tone by native English listeners. In the behavioral task, both the perception-only and the perception-plus-production groups improved on the tone discrimination abilities after the training session. Moreover, the participants in both groups generalized the improvements gained through the trained stimuli to the untrained stimuli. In the ERP task, the Mismatch Negativity was smaller in the post-training task than in the pre-training task. However, the two training groups did not differ in tone processing at the intentional or unintentional level after training. These results suggest that the employment of the motor system does not specifically benefit the tone perceptual skills. Furthermore, the present study investigated whether some tone pairs are more easily confused than others by native English listeners, and whether the order of tone presentation influences non-native tone discrimination. In the behavioral task, Tone2-Tone1 (rising-level) and Tone2-Tone4 (rising-falling) were the most difficult tone pairs, while Tone1-Tone2 and Tone4-Tone2 were the easiest tone pairs, even though they involved the same tone contrasts respectively. In the ERP task, the native English listeners had good discrimination when Tone2 and Tone4 were embedded in strings of Tone1, while poor discrimination when Tone1 was inserted in the context of Tone2 or Tone4. These asymmetries in tone perception might be attributed to the interference of native intonation system and can be altered by training.

*Corresponding author.
E-mail addresses: shuanglu2013@outlook.com (S. Lu), ratree@ufl.edu (R. Wayland), kaan@ufl.edu (E. Kaan).

http://dx.doi.org/10.1016/j.brainres.2015.07.014
0006-8993/

1. Introduction

In language, pitch differences may signal different levels of prosodic contrasts. In intonation languages, such as English, pitch variation is used to distinguish different sentence types (question or statements) and to flag information as new or unpredictable. In tonal languages, such as Thai, Chinese, and Vietnamese, pitch differences are used to distinguish lexical meaning and to carry grammatical distinction (Chao, 1948). For example, in Mandarin Chinese the syllable [t^hu] produced with four different tones (Tone1: high-level; Tone2: high-rising; Tone3: low-dipping; and Tone4: high-falling; henceforth referred to as T1, T2, T3 and T4 respectively) will result in four different words: [t^hu 1] 'bald', [t^hu 2] 'chart', [t^hu 3] 'earth' and [t^hu 4] 'vomit'.

Previous studies have shown that non-native speakers of tonal languages have difficulties in both comprehending and producing lexical tones (e.g., Gandour, 1983; Wang et al., 1999; White, 1981). It has been found that native speakers of non-tonal languages tend to pay more attention to the pitch onset, offset and the average pitch when exposed to lexical tones, while native speakers of tone languages focus more on the pitch contour (Gandour, 1983; Krishnan et al., 2005; Chandrasekaran et al., 2007). The perceptual and production difficulties have also been attributed to the native language interference. For example, Shen (1989) reported that English speakers are likely to consider Mandarin T4 (high-falling tone) as the falling intonation in the final position of statement sentences. Therefore, they usually produced T4 at a lower pitch onset and with a less steep falling slope than the native Chinese speakers. However, these perceptual and production difficulties that confront native speakers of non-tonal languages are not insuperable. Several behavioral and electrophysiological studies have shown that short-term perceptual and production training are effective in improving the comprehension and production of lexical tones by native speakers of non-tonal languages (Kaan et al., 2007, 2008; Leather, 1990; Song et al., 2008; Wang et al., 1999; Wayland and Guion, 2004; Wayland and Li, 2008). Moreover, the effects of perception training and production training may be inter-transferable (Wang et al., 2003; Leather, 1990). For instance, Leather (1990) reported that after a production training native Dutch speakers were able to perceive the differences in Mandarin tones. But in this study only one syllable was used in both the training and the post-training tasks, so it is unclear whether the effect can generalize to novel stimuli.

To our knowledge, very few studies have directly compared the effectiveness of production training and perception training on lexical tone perception. To bridge this gap, the current study examines the effects of the two types of training on the perception of lexical tones by native English listeners. If speech perception is driven by speech production, as is the basis of, e.g., the Motor Theory of Speech Perception (Liberman and Mattingly, 1985), training naïve learners to produce sounds should be more effective for improving their perceptual ability of these sounds than training people to only perceive sounds. This hypothesis has been supported by some neuroimaging studies, which demonstrated that the speech motor system is activated during speech perception tasks (Binder et al., 2004; Callan et al., 2010; Chevillet et al., 2013; Pulvermüller et al., 2006; Wilson et al., 2004) and may facilitate speech perception through sensorimotor integration (Callan et al., 2004; Du et al., 2014; Hickok et al., 2011; Wilson and Iacoboni, 2006). On the contrary, if speech production is not mandatory for speech perception, as suggested by Dual Stream Model of Speech Processing (Hickok and Poeppel, 2000, 2004, 2007); or if speech production relies on speech perception, as claimed by, e.g., the Directions into Velocities of Articulators Model (DIVA) (Guenther et al., 2006), and the Speech Learning Model (Flege, 1995), production training may be equally/less effective for improving people's perceptual ability as/than perception training. This hypothesis has been supported by a magnetoencephalography (MEG) study by Levelt et al. (1998), who showed that the initiation of articulatory processes was slightly slower than the auditory cortical activations during speech production.

Previous behavioral studies on perception training and production training have also provided evidence for both hypotheses. Some research showed that participants' ability to discriminate/identify novel phonemic contrasts improved through short-term laboratory production training (e.g., Hirata, 2004), suggesting that speech perception might be channeled through the production pathway. Adank et al. (2010) also found that vocal imitation significantly improved the comprehension of unfamiliar accent, regardless of whether people could hear their own voice. Other studies demonstrated that learning in the production domain cannot transfer to the perception domain, even though the participants' production reached native-like accuracy after the production training (e.g. Hattori, 2010).

One concern regarding this line of research is that speech production could not be completely isolated from speech perception in the experimental design. Therefore, some recent studies have compared perception-only training to perception-plus-production training to see whether the additional production task would provide more benefits to participants' perceptual learning. Unexpectedly, most of these studies showed that participants' perceptual learning was hindered by the additional production task and the perception-only training was more effective to improve the participants' perception of L2 phonemic contrasts than the perception-plus-production training (e.g., Herd, 2011; Baese-Berk, 2010). However, most of these training studies focused on the perception of novel contrasts at the segmental level. Very little research has been conducted to examine the effects of perception training and production training on the perception of suprasegmental phonological contrasts (e.g. stress, intonation and lexical tones). Furthermore, no study has investigated which stage of lexical tone processing can be affected by production training and perception training. Will the production and perception training have a different effect on lexical tone perception at the behavioral (intentional) level? Or will the two types of training also differentially affect unintentional processing (i.e. when the auditory stimuli are not relevant to the participants' task)? The current study aims to answer these questions by examining the effects of perception-plus-production training and perception-only training on the intentional and unintentional processing of lexical tones using both behavioral and neurophysiological methods (EEG).

Event-related potentials (ERPs) are a good method to study the unintentional processing of lexical tones while the auditory stimuli are presented to participants. ERPs do not require

participants' attention and can provide high temporal resolution. The participants can, for instance, watch a movie while their brains' responses to the auditory stimuli are recorded. In contrast, behavioral tasks provide data regarding participants' intentional responses after they fully process the stimuli. Therefore, a passive ERPs task, combined with a separate behavioral task can give a thorough view of how people intentionally and unintentionally process lexical tones, and indicate which level of processing is affected by the type of training (i.e. perception vs. production training). Moreover, previous studies on speech-sound training have shown that changes in ERPs may precede the behavioral improvement (e.g., Tremblay et al., 1998). Recording neurophysiological measures may therefore be more informative than only recording behavioral measures, since perceptual changes after training may occur at the unintentional level but not (yet) at the intentional/behavioral level. Some ERP components have been found to be relevant to automatic auditory perception. The Mismatch Negativity (MMN) is a negative deflection in the difference wave obtained by subtracting the event-related potential to frequently presented standard stimuli from that to infrequently presented deviant stimuli. The MMN usually occurs between 100 and 300 ms after the onset of deviant stimuli (e.g., Näätänen and Alho, 1995), and has been claimed to reflect pre-attentive processing (although this component may be sensitive to intentional manipulations, e.g., Woldorff et al., 1991). Many studies have shown that the behavioral improvement in speech perception is usually accompanied by an increased MMN (e.g. Tremblay et al., 1997; Kaan et al., 2007). Sometimes the increased MMN is observed even before the behavioral improvement occurs (e.g. Tremblay et al., 1998). The Late Negativity is another negative wave that occurs around 350–600 ms after the onset of deviant stimuli. The late negativity has been associated with the reorientation of attention (Shestakova et al., 2003). Some studies have shown that the late negativity became smaller after training (e.g. Kaan et al., 2008), suggesting that after training the deviant stimuli became easier to distinguish and required less attention.

Another aim of the current study is to investigate which tone pairs would be more difficult for native speakers of non-tonal languages to discriminate and which tone pairs are more resistant/easier for improvement after the perception-only and the perception-plus-production training respectively. Previous studies have suggested that non-native listeners do not perceive the four Mandarin tones equally well, and some tones are more confusing and more resistant to improvement than others (e.g. Gottfried and Suiter, 1997; Shen, 1989). So and Best (2010) showed that English listeners had more difficulties in discriminating two tones if they share similar phonetic features (i.e. T1–T2, T1–T4 and T2–T3), than if they have no similar feature (i.e. T1–T3, T2–T4 and T3-T4). Wang et al. (1999) trained English listeners to identify the four Mandarin tones and reported that the participants showed great improvements on tone pairs T2–T3, T2–T4 and T1–T2 after training. However, the tone pair T1–T4 was most resistant to improvement. They also noticed that the direction of the tone-pair confusion was asymmetric. For instance, they found that T2 was more likely to be misperceived as T3 than the reverse. Francis and Ciocca (2003) also demonstrated that the order of the stimulus presentation could influence the discrimination of lexical tones. But they argued that this tone presentation order effect might be language specific. They found that Cantonese listeners were more accurate at discriminating the level tone pairs if the first syllable had a lower F0 than the second syllable than when the first syllable had a higher F0. Nevertheless, the English listeners who had no experience with Cantonese did not show such an effect. Also, when both the Cantonese and English listeners were tested with non-speech tokens that had the same F0 patterns, the effect disappeared. Taking previous research into account, we predict that T1-T4 may be the most difficult tone pair for native speakers of non-tonal languages to discriminate. This pair type may also show less improvement after training compared with the other tone pairs. However, if the employment of the motor system benefits perceptual skills, the perception-plus-production group may show more improvement and an increased sensitivity to this tone contrast after training than the perception-only group. In addition, if the order of stimulus presentation affects the discrimination of lexical tone, the tone pairs T1–T4 and T4–T1 may also show different discrimination accuracies as a result of training.

In the present study, native English listeners were trained and tested over the course of three consecutive days. On the first day, participants' brain waves (ERPs) were recorded as baseline, using a passive oddball paradigm. After the ERP recording, participants did a behavioral same/different discrimination task. On the second day, participants received either a perception-only or a perception-plus-production training, which lasted for about one hour.[1] The trainings for both groups were exactly the same except that the perception-plus-production training required participants to imitate the stimuli, while the perception-only training had participants utter a word unrelated to the stimuli. On the third day, both groups did the same ERP and behavioral tasks as on the first day. In the behavioral task, stimuli were eight monosyllables ([p^ha], [p^hi], [k^hɛ], [k^ho], [t^ha], [t^hi], [t^hɛ] and [t^ho]) associated with three linear tones that resemble Mandarin T1 (high-level), T2 (high-rising), and T4 (high-falling). The syllables [t^ha], [t^hi], [t^hɛ], [t^ho], [k^ho] and [p^ha] were used in the pre- and post-training tasks, and the syllables [k^hɛ] and [k^ho], [p^ha], [p^hi] were used in the training session. The syllables [t^ha], [t^hi], [t^hɛ] and [t^ho] were excluded from the training session because we wanted to test whether participants could generalize the improvements gained through training to untrained stimuli. Through the behavioral task, we investigate whether the perception-plus-production training is more effective than the perception-only training in facilitating lexical tone perception at the intentional level. In the ERP task, syllable [t^hu] associated with the three tones were presented in three types

[1]Originally, we planned to conduct two sessions of training during the course of two consecutive days, with each session took about one hour. However, our pilot behavioral data on two participants showed that the participants' discrimination abilities improved significantly even after the first day of training. Moreover, both participants complained that the whole experiment was too long and they got bored with the training content very easily during the second training session. Therefore, we decided to only include one hour of training. This decision may result in some unexpected results, which will be discussed in Section 3.1.

Table 1 – Mean *d*' scores for the trained and untrained stimuli in the pre- and post-training tasks for each group.

	Perception-only		Perception-plus-production	
	Trained stimuli	Untrained stimuli	Trained stimuli	Untrained stimuli
Pre-training	1.74 (0.91)	1.84 (0.98)	1.98 (1.05)	2.16 (1.08)
Post-training	2.53 (1.35)	2.54 (1.22)	2.88 (1.34)	2.95 (1.55)

Standard deviations are in parentheses.

Table 2 – Mean response times of the trained and untrained stimuli in the pre- and post-training tasks for each group.

	Perception-only		Perception-plus-production	
	Trained stimuli	Untrained stimuli	Trained stimuli	Untrained stimuli
Pre-training	761 (358)	756 (305)	955 (392)	993 (347)
Post-training	791 (520)	822 (550)	955 (362)	925 (330)

Standard deviations are in parentheses.

of blocks: (1) T1 was the standard, and both T2 and T4 were the deviants; (2) T2 was the standard, with T1 and T4 being the deviants; (3) T4 was the standard, and T1 and T2 were the deviants (Näätänen et al., 2004). We focus on the MMN and Late Negativity components to examine whether perception-plus-production training affects the unintentional perception of lexical tones differently from perception-only training. We expect that both the behavioral and ERP measures would show that the perception-plus-production training is more beneficial than the perception-only training to native English listeners' perception of lexical tones. Moreover, we anticipate that native English listeners have more difficulties in discriminating tone pair T1–T4 than the other tone pairs, and they may show asymmetrical tone discrimination after training.

2. Results

2.1. Behavioral task

2.1.1. Accuracy

The trained stimuli included the syllables [k^ho] and [p^ha], which were used in the pre-training task, the training session and the post-training task. The untrained stimuli were the syllables [t^ha], [t^hi], [$t^h\varepsilon$] and [t^ho], which were only tested in the pre- and post-training tasks. Mean d' scores with the standard deviations of the trained and untrained stimuli in the pre- and post-training tasks for the perception-only and the perception-plus-production groups are presented in Table 1.

The mean *d*' scores for each participant were submitted to a 2 (test time: pre- and post-training) × 2 (stimulus type: trained and untrained) × 2 (group) repeated measures ANOVA, with test time and stimulus type as the within subject factors and group as the between-subject factor. The result yielded a main effect of test time [F (1, 20) = 53.51, $p < .001$]. Both groups discriminated the tone pairs more accurately in the post-training task than in the pre-training task as a result of training. The test time × group interaction was not significant [F (1, 20) = .22, $p = .65$], suggesting that the two groups did not differ from each other in the extent of improvement. There was no difference between the mean d' scores for the trained and untrained stimuli [F (1, 20) = 2.94, $p = .10$], indicating that the participants generalized the improvements gained through the trained stimuli to the untrained stimuli.[2]

2.1.2. Reaction times

Mean reaction times were computed for all the correct responses in the pre-training task, the training session and the post-training task for the perception-only and production-plus-perception groups. The reaction times were measured from the offset of the second stimulus in each trial. The mean response times of the trained and untrained stimuli in the pre- and post-training tasks for the perception-only and the perception-plus-production are illustrated in Table 2.

The reaction times for each participant were submitted to a 2 (test time: pre- and post-training) × 2 (stimulus type: trained and untrained) × 2 (group) repeated measures ANOVA, with test time and stimulus type as the within-subject factors and group as the between-subject factor. The results showed that there was a significant interaction of test time × stimulus type × group [F (1, 20) = 5.39, $p < .05$]. This interaction was due to the fact that before training the perception-only group discriminated both the trained and untrained stimuli numerically faster than the perception-plus-production group. No other significant effect or interaction was found. Further

[2] In order to exclude the interpretation of general test-retest effect, we performed a separate analysis on the d' scores of the trained stimuli (syllables [k^ho] and [p^ha]) during the pre-training, training and post-training tasks, with test time as within subject factor and group as between-subject factor. The results only yielded a significant main effect of test time [F (2, 40) = 24.72, $p < .001$]. Pairwise comparisons revealed that the participants discriminated the trained stimuli more accurately in the training task than in the pre-training task ($p < .001$). But they discriminated the two syllables equally well in the training and the post-training task ($p > .22$). These results indicated that the training procedures were effective and the improvement gained through the training retained in the post-training task on the third day of experiment. If it was a general test-retest effect, the discrimination accuracy would be higher in the post-training task than in the training task, due to repetition.

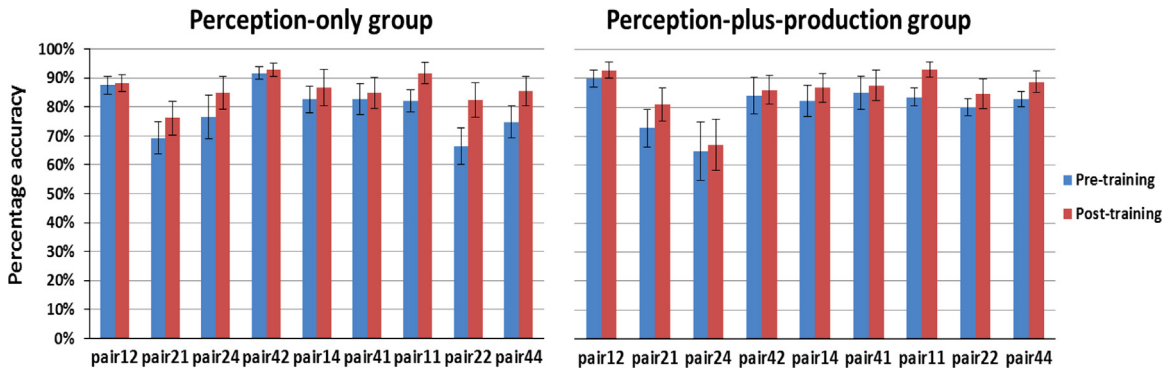

Fig. 1 – Mean percentage accuracy with standard errors of each tone-pair in the pre- and post-training tasks for the perception-only and perception-plus-production groups.

paired samples *t* tests on the trained and untrained stimuli did not show any significant difference between the reaction times in the pre- and post-training tasks for either group. In sum, neither the perception-only group nor the perception-plus-production group discriminated tones more quickly in the post-training task than in the pre-training task.

2.1.3. *Percentage accuracy for individual tone-pairs*

In order to examine which tone pair showed the greatest improvement and which pair type was most resistant to improvement, we calculated the percentages of accurate discriminations for each tone pair (pair12: T1–T2; pair21: T2–T1; pair14: T1–T4; pair41: T4–T1; pair24: T2–T4; pair42: T4–T2; pair11: T1–T1; pair22: T2–T2; pair44: T4–T4).

Fig. 1 depicts the mean percentage accuracy of each tone-pair in the pre- and post-training tasks for the perception-only and perception-plus-production groups. We performed further analyses on the 'different' and 'same' tone pairs separately. The accuracy percentages for the 'different' tone pairs were submitted to a 2 (test time: pre and post-training task) × 6 (pair) × 2 (group) repeated measures ANOVA with test time and pair as the within-subject factors and group as the between-subject factor. The result yielded main effects of test time [F (1, 20) = 12.38, p = .002], and pair [F (5, 100) = 9.72, p < .001]. However, the interaction of test time × pair was not significant [F (5, 100) = .94, p = .42]. Pairwise comparisons showed that the participants' discrimination of pair21 was less accurate than the discrimination of pair12 (p = .001), pair14 (p < .05), pair41 (p < .01), and pair42 (p < .001). The discrimination of pair24 was less accurate than the discrimination of pair42 (p < .01) and pair14 (p < .05). The pair14 and pair41 were discriminated equally well (p = 1.00). Moreover, the interaction of pair × group almost reached significance [F (5, 100) = 2.82, p = .053]. This interaction was due to the fact that the two groups showed different discrimination accuracies for different tone pairs. The perception-only group discriminated pair42 the most accurately but pair21 the least accurately; the perception-plus-production group discriminated pair12 the most accurately but pair21 and pair24 the least accurately.

The analysis on the 'same' tone pairs also yielded main effects of test time [F (1, 20) = 4.55, p < .05], and pair [F (2, 40) = 32.08, p < .001], while the interaction of test time × pair was not significant [F (2, 40) = .28, p = .68]. Pairwise comparisons revealed that the participants discriminated pair11 more accurately than pair22 (p < .001) and pair44 (p < .01), with pair22 being discriminated the least accurately (pair22 vs. pair44: p < .001). In addition, there was a significant interaction of pair × group [F (2, 40) = 4.08, p < .05]. In the perception-only group the discrimination of pair11 was more accurate than pair44 (p < .01), which was more accurate than pair 22 (p < .01). In the perception-plus-production group, pair11 and pair44 were discriminated equally well (p > .05) and both were discriminated more accurately than pair22 (*p*s < .01).

2.2. *ERP*

Fig. 2 presents the ERPs of the frontal electrodes (F3, Fz and F4), the central electrodes (C3, Cz and C4) and the parietal electrodes (P3, Pz and P4) collapsed over all six conditions (T12: deviant T1 in standard T2; T14: deviant T1 in standard T4; T21: deviant T2 in standard T1; T24: deviant T2 in standard T4; T41: deviant T4 in standard T1; and T42: deviant T4 in standard T2) in the pre- and post-training tasks for the perception-only and perception-plus-production groups respectively. Descriptively, a MMN component was elicited for the deviant versus standard stimuli, with a prominent frontal distribution between 200 and 400 milliseconds. A late negativity component was also observed at the frontal and central electrodes between 500 and 800 milliseconds. Fig. 3 shows the deviant minus standard difference waves of the Fz electrode for the six conditions in the pre- and post-training tasks for the two groups.

2.2.1. *MMN*

Fig. 4 presents the ERPs of the Fz electrode for the six conditions in the pre- and post-training tasks for the perception-only and perception-plus-production groups respectively. Fig. 5 displays the isovoltage maps for the MMN. The mean amplitudes of the difference waves between 200 and 400 ms were computed for two regions: right-frontal (RF) included the electrodes F4, F6, F8, FC4, FC6 and FT8; and left-frontal (LF) included the electrodes F3, F5, F7, FC3, FC5 and FT7. The mean amplitudes were submitted to a 2 (test time) × 2 (hemisphere: right and left) × 6 (condition: T12, T14, T21, T24, T41, T42) × 2 (group) repeated measures ANOVA with test time, hemisphere and condition as the within-subject factors and group as the between-subject factor. The results showed that the intercept was significant [F (1, 20) = 13.66, p = .001], which means that the negativity for the deviants

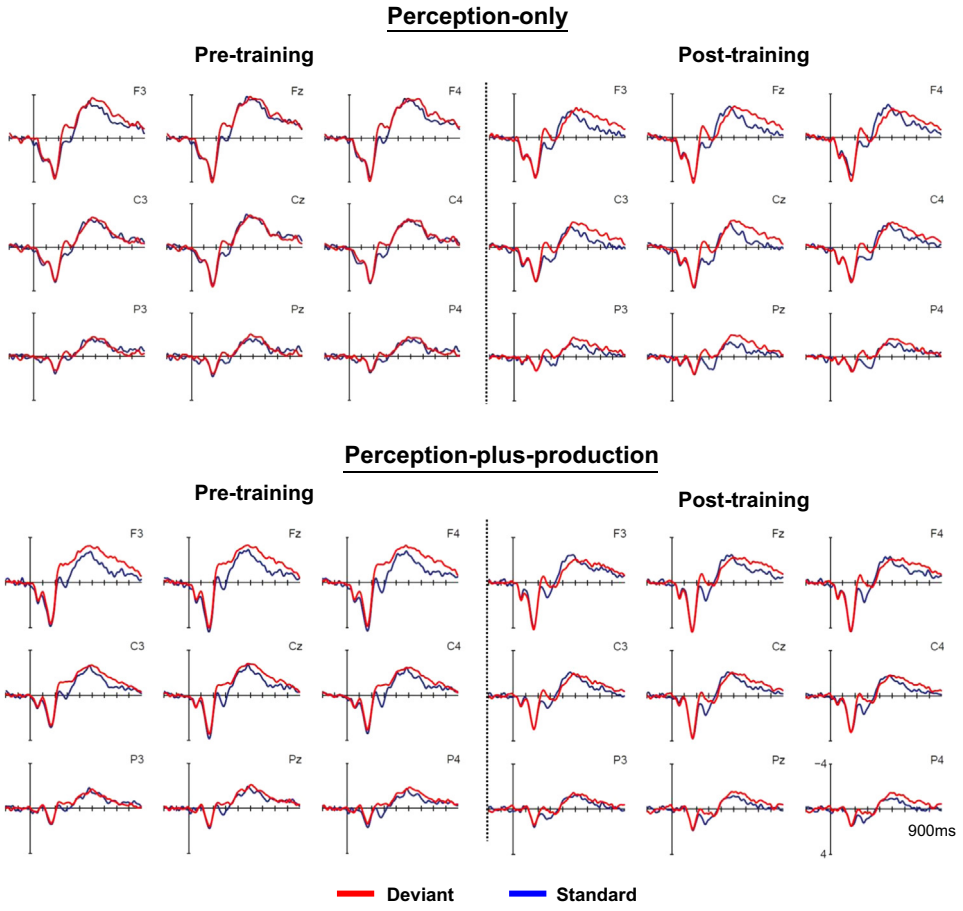

Fig. 2 – The ERPs to standard (blue line) and deviant tones (red line) for three frontal electrodes (F3, Fz and F4), central electrodes (C3, Cz and C4) and parietal electrodes (P3, Pz and P4) collapsed over all six conditions, in the pre- and post-training tasks for the perception-only and perception-plus-production groups respectively. Negative is plotted up in this and the following figures.

was significantly larger than that for the standards. In addition, there was a main effect of test time [F (1, 20) = 6.68, p = .018]. The MMN was *smaller* in the post-training task than in the pre-training task. The effect of condition was also significant [F (5, 100) = 5.13, p = .005]. The MMN was more prominent for the T41 condition than for the T14 condition (p = .05), T24 (p < .01), and T42 (p < .01). No other main effects or interactions involving condition or group were significant. Furthermore, the mean MMN amplitude at the frontal electrodes (F4, F6, F8, FC4, FC6, FT8, F3, F5, F7, FC3, FC5, FT7, Fz and FCz) for the six conditions in the pre- and post-training tasks were compared with a hypothetical zero through one-sample t tests. In the perception-only group, the MMN was only present for the T41 condition before training (p < .03), but was absent after training (p > .05). In the perception-plus-production group, the MMN was significant in the T21 (p < .03) and T41 (p < .01) conditions before training, but was not significant in either condition ($ps \geq .05$) after training.

2.2.2. *Late negativity*

Fig. 3 illustrates that in the perception-only group the late negativity increased in the post-training task compared to the pre-training task, especially for the T12, T14, T24 and T41 conditions. However, in the perception-plus-production group, the late negativity was less prominent in the post-training task than in the pre-training task, especially for the T12, T14 and T21 conditions. Fig. 6 presents the isovoltage maps for the late negativity in the 500–800 ms window for each condition and each group.

The mean amplitudes of the difference waves between 500 and 800 ms were computed for the four regions: right-frontal (RF), left-frontal (LF), right-posterior (RP) and left-posterior (LP). These mean amplitudes were submitted to a 2 (test time) × 2 (hemisphere: right and left) × 2 (anteriority: anterior and posterior) × 6 (condition: T12, T14, T21, T24, T41, T42) × 2 (group) repeated measures ANOVA with test time, hemisphere, anteriority and condition as the within-subject factors and group as the between-subject factor. The result yielded a significant intercept [F (1, 20) = 17.83, p < .001], meaning that the negativity for the deviants was significantly larger than that for the standards. The interaction of hemisphere × group was significant [F (1, 20) = 6.41, p = .02]. The negativity was larger over the right hemisphere in the perception-plus-production group [t (10) = 2.65, p = .024], but was bilaterally distributed in the perception-only group [t (10) = -0.75, p = .47]. Besides, there was a main effect of anteriority [F (1, 20) = 8.19, p = .01] and a significant interaction of test time × anteriority [F (1, 20) = 9.94, p = .005]. The negativity was larger over the frontal electrodes than over the posterior electrodes before training [t (21) = −3.95,

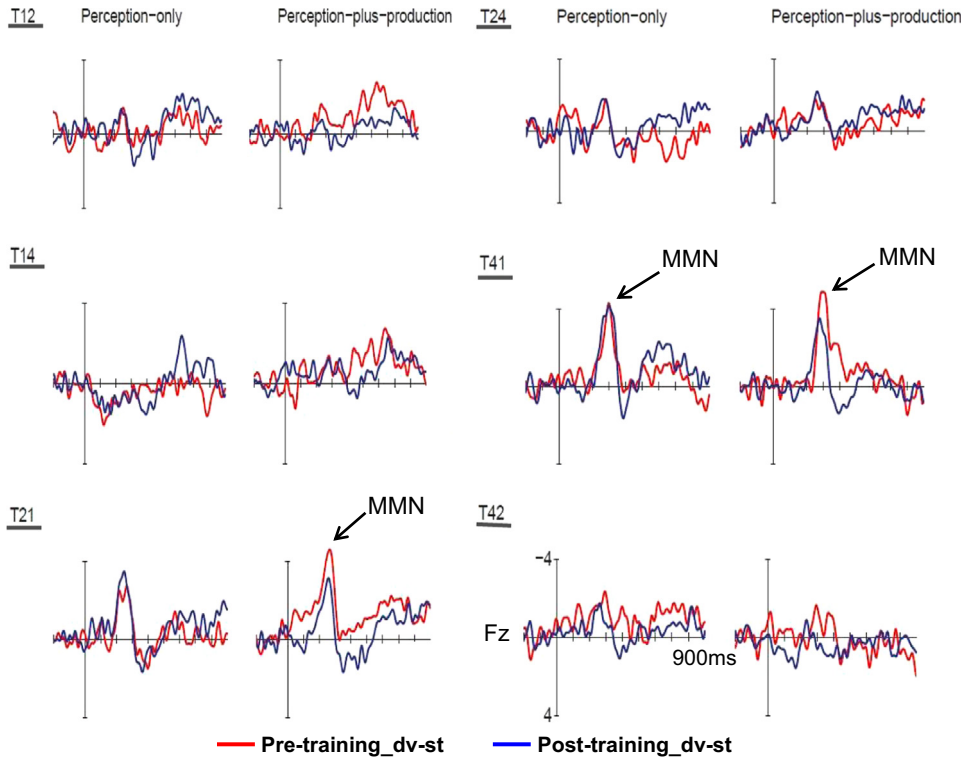

Fig. 3 – Deviant minus standard difference waves of the Fz electrode for the six conditions in the pre-training (red line) and post-training tasks (blue line) for the two groups. T12 condition: deviant T1 in standard T2; T14 condition: deviant T1 in standard T4; T21 condition: deviant T2 in standard T1; T24 condition: deviant T2 in standard T4; T41 condition: deviant T4 in standard T1; T42 condition: deviant T4 in standard T2.

$p=.001$], but was equally distributed over the frontal and posterior regions after training [t (21) = −.85, $p=.41$]. The interactions of hemisphere × anteriority × condition [F (5, 100) = 2.70, $p=.046$] and test time × hemisphere × anteriority × condition [F (5, 100) = 2.76, $p=.048$] were also significant.

Separate analyses were performed for each condition. For the T12 condition, the effect of anteriority [F (1, 20) = 6.48, $p=.019$] and the interaction of test time × hemisphere × anteriority × group [F (1, 20) = 5.96, $p=.024$] were significant. For the T14 condition, there was only a main effect of hemisphere [F (1, 20) = 5.07, $p=.036$]. For the T21 condition, the main effect of anteriority [F (1, 20) = 4.69, $p=.043$], the interaction of test time × anteriority [F (1, 20) = 9.64, $p=.006$], and the interaction of test time × hemisphere × anteriority [F (1, 20) = 5.10, $p=.035$] were significant. For the T24 condition, the main effect of group was significant [F (1, 20) = 5.30, $p=.032$]. The negativity was larger in the perception-plus-production group than in the perception-only group. For the T41 condition, the effect of anteriority [F (1, 20) = 6.61, $p=.018$], the interaction of hemisphere × group [F (1, 20) = 4.79, $p=.041$], and the interaction of hemisphere × anteriority × group [F (1, 20) = 17.77, $p<.001$] were significant. No significant main effect or interaction was found for the T42 conditions.

An overview of the effects for the ERP data as well as the behavioral data is given in Table 3. To sum up, in the 200–400 ms window the MMN was smaller in the post-training task than in the pre-training task. Both groups showed the MMN for the T41 condition before training, but not after training. In addition, the perception-plus-production group displayed the MMN in the T21 condition before training, but not after training. In the 500–800 ms window, the negativity had a frontal distribution before training, but was equally distributed over the frontal and posterior regions after training. Moreover, the two groups differed in the hemispheric distribution of the negativity component. The negativity was larger over the right hemisphere in the perception-plus-production group, but was bilaterally distributed in the perception-only group. For the T24 condition, the negativity was larger in the perception-plus-production group than in the perception-only group.

3. Discussion

The current study collected both behavioral and electrophysiological (ERPs) data to compare the effectiveness of a perception-only training and a perception-plus-production training procedure on the intentional and unintentional processing of tones by native English speakers. If production indeed facilitated perception, as implied by the Motor Theory (e.g., Liberman and Mattingly, 1985), we anticipated that the perception-plus-production training should be more effective than the perception-only training to improve participants' lexical tone perception at both intentional and unintentional levels. On the contrary, if production relies on perception, as claimed by, e.g. the Speech Learning Model (Flege, 1995), we would expect that the perception-only training might be more effective than the perception-plus-production training to improve participants' intentional and unintentional

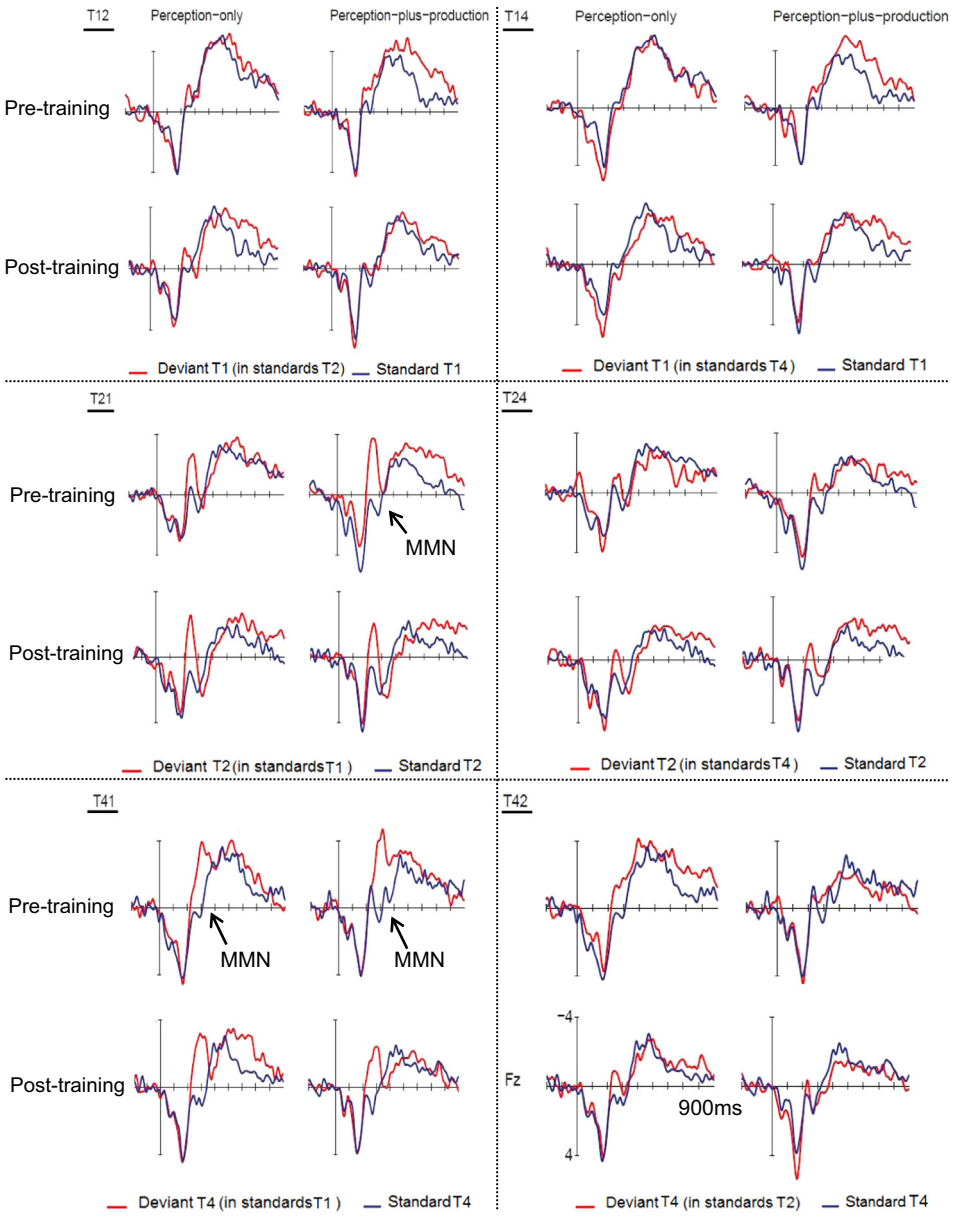

Fig. 4 – The ERPs to standard (blue line) and deviant tones (red line) of the Fz electrode for the six conditions in the pre- and post-training tasks for the two groups.

discrimination of tones. Our results showed that both types of training improved native English listeners' tone discrimination abilities. However, neither the behavioral nor ERP measures yielded a training type difference. Another aim of the current study is to investigate whether some tone pairs are more easily confused than others to native English listeners, and whether they have asymmetrical tone discrimination. We found that pair21 and pair24 were the most difficult tone pairs for the English listeners to discriminate behaviorally, while pair42 and pair12 were the easiest tone pairs. The ERP measures also showed that the English listeners discriminated lexical tones asymmetrically at the unintentional level: they had good discrimination when T2 and T4 were embedded in strings of T1, while poor discrimination when T1 was inserted in the context of T2 or T4. We will discuss these findings in more detail below.

3.1. *Perception-only training vs. perception-plus-production training*

In the behavioral tasks, both the perception-only and the perception-plus-production groups showed improved tone discrimination abilities after the training session. In addition, the participants in both groups generalized the improvements gained through the trained stimuli to the untrained stimuli. However, the two groups did not differ from each other in either the extent of improvement or the response times.

In the ERP task, both groups showed a smaller MMN in the post-training task than in the pre-training task. This result was comparable with Kaan et al. (2008), which demonstrated that after training the MMN decreased for the low-falling tone in English speakers. Kaan et al. (2008) claimed that the larger MMN in the pre-training task was due to the differences in F0

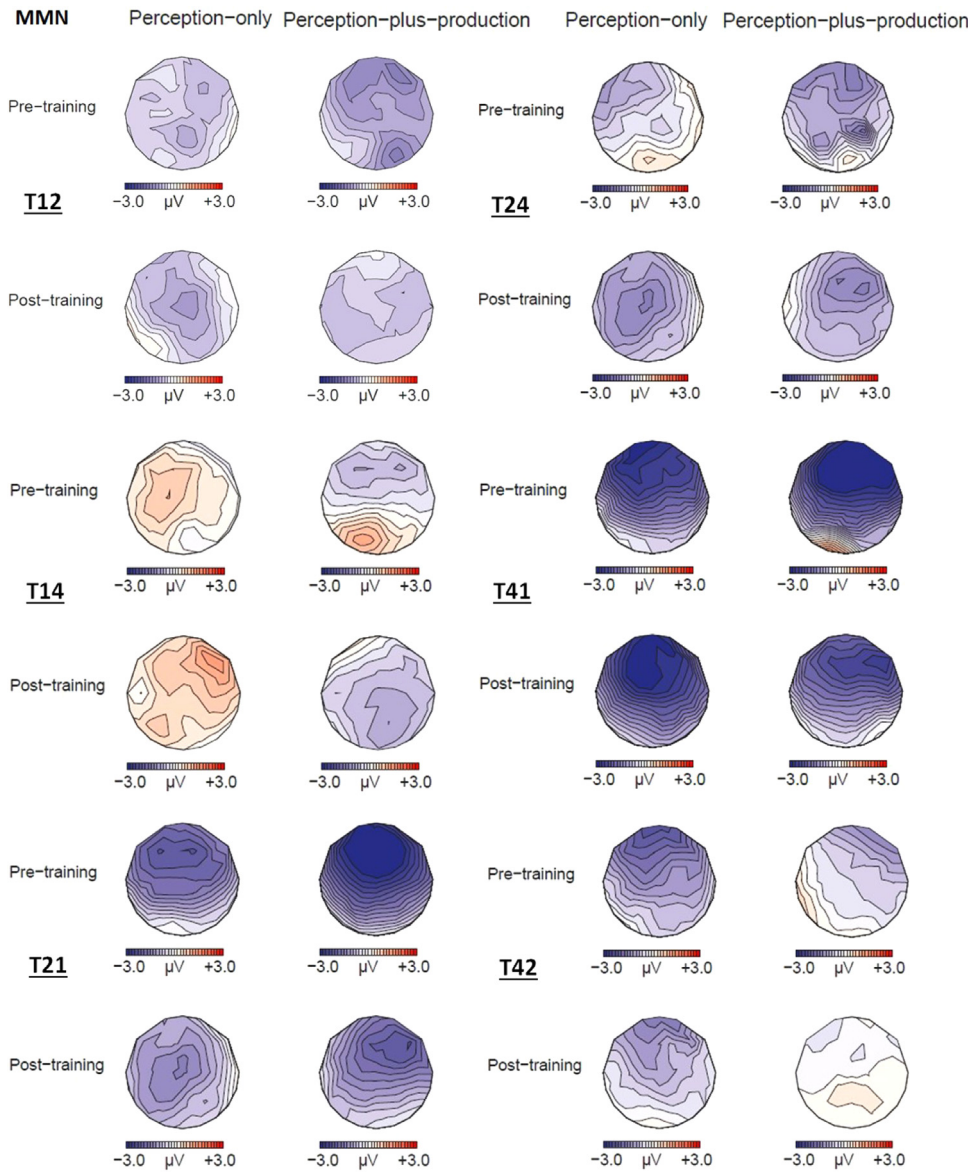

Fig. 5 – Isovoltage maps of the MMN effect, defined as the mean amplitude of the deviant minus standard difference waves in the 200–400 ms for each condition and each group.

onset rather than the F0 direction. The English speakers were more sensitive to the F0 onset differences (e.g. Gandour, 1983) before training and started paying more attention to the F0 direction after training. However, the detection of F0 direction was harder than the detection of F0 onset for the English speakers, therefore, the MMN amplitude decreased in the post-training task compared to the pre-training task.

Furthermore, the two groups in the present study differed in the hemispheric distribution of the late negativity component. The late negativity was right lateralized in the perception-plus-production group, but was equally distributed in both hemispheres in the perception-only group. Even though the scalp distributions do not reflect exact location of neural sources, a difference in the distribution of the late negativity component may indicate that the underlying processes of lexical tones in the two groups are different in some respects. According to Poeppel (2001), the left hemisphere is sensitive to rapid acoustic transients, whereas the right hemisphere is suited for longer/slower changes. Our data may suggest that the perception-plus-production group paid more attention to the gradual change in F0 contour, rather than the abrupt change in F0 onset, hence, involved more right hemisphere regardless of training. In contrast, the perception-only group may use both cues (i.e. F0 level and F0 direction) to distinguish tones, thus involved both hemispheres.

For the T24 condition the perception-plus-production group showed a larger late negativity than the perception-only group. Some previous studies (e.g. Zachau et al., 2005) have associated the late negativity with neural processes of auditory rule extraction. Under this interpretation of the late negativity component, the larger late negativity in the perception-plus-production group may indicate that to some extent the perception-plus-production group is better at tone extraction at the unintentional level compared with the perception-only

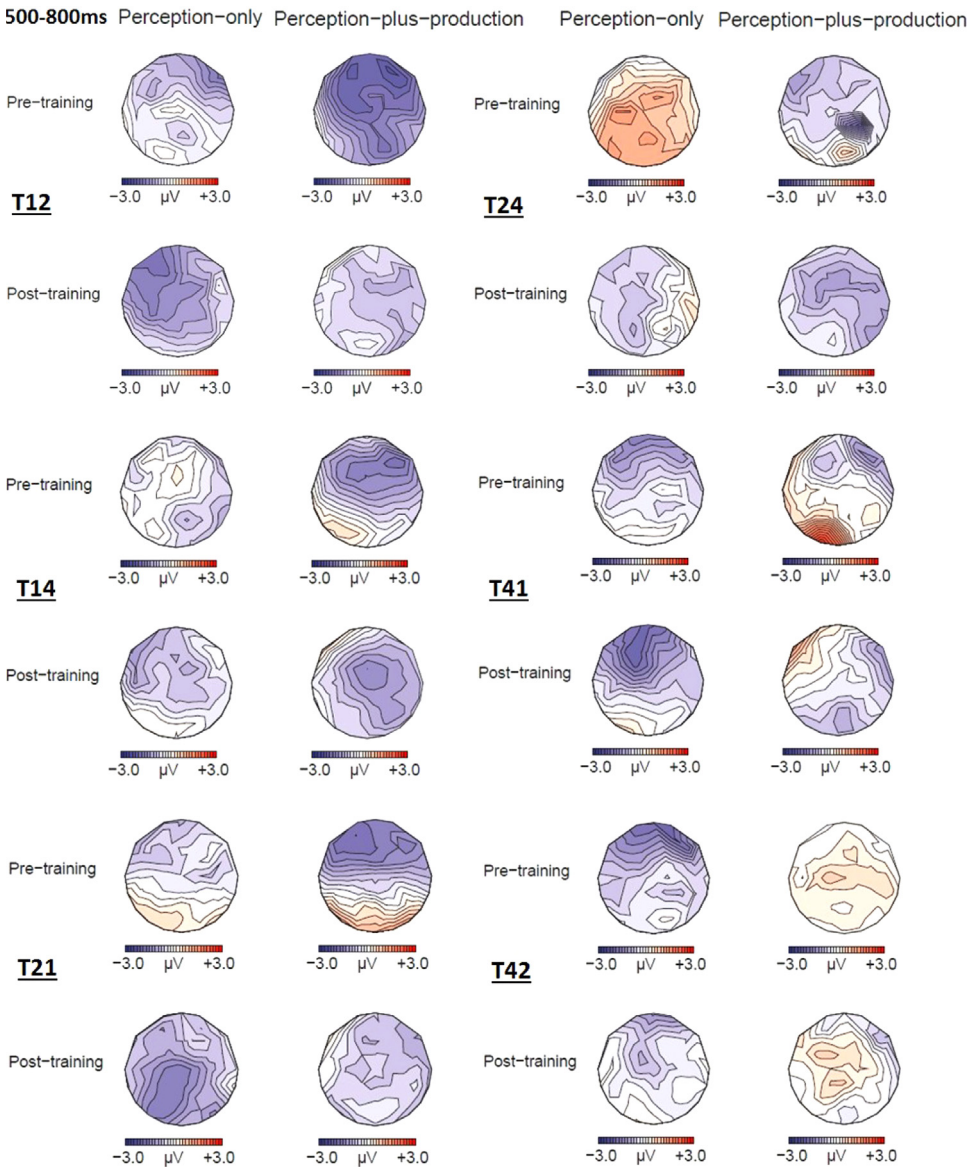

Fig. 6 – Isovoltage maps for the 500–800 ms window for the deviants minus standards for each condition and each group.

group. However, neither of the two groups showed a training effect in terms of the late negativity component, which was incompatible with previous studies. For example, Kaan et al. (2008) reported a decreased late negativity after training and claimed that the smaller late negativity in the post-training task may suggest less intentional reorientation from deviant stimuli to standard stimuli.

In spite of all the subtle differences mentioned above, the perception-plus-production training and the perception-only training did not show different effects on English listeners' intentional or unintentional processing of lexical tones. It should be also noted that our training procedure may have been somewhat biased toward the perception-plus-production group, since the imitation task in the perception-plus-production training may require the participants to pay more attention to the stimuli than the perception-only training. But, even so, the participants who received the perception-plus-production training did not demonstrate more improvement than the participants who received the perception-only training. This result seems to suggest that the employment of the motor system does not specifically benefit the tone perceptual skills. The lack of additional improvement in the perception-plus-production group can be accounted for in two ways. First, the participants did not receive any feedback on their productions during the perception-plus-production training. Accordingly, the perceptual learning through the production process may have been rather limited. Further research should provide feedback to participants' productions (e.g. visual feedback) in order to really facilitate perceptual learning. Second, our training session might have been too short to show the effect of the production component. In other training studies (e.g. Wang et al., 1999; Wayland and Li, 2008; Baese-Berk, 2010), participants usually received at least two sessions of training with each session lasting at least one hour. In our study, the perception-plus-production training only took one hour, during which the participants were only allowed to imitate each

Table3 – Main effects and interactions for the behavioral and ERP data. See text for degrees of freedom, *F*, and *p* values.

Condition	Behavioral		MMN	Late Negativity
	'Different'	'Same'		
Overall	Test time***	Test time*	Test time*	Hemisphere × group*
	Pair***	Pair*** Pair × group*	Condition**	Anteriority* Test time × anteriority** Hemisphere × anteriority × condition* Test time × hemisphere × anteriority × condition*
Pair12/ T21		–	Perception-plus-production: before training*	Anteriority* Test time × anteriority** Test time × hemisphere × anteriority*
Pair21/ T12	<Pair12** <Pair14* <Pair41** <Pair42***	–	Ns	Anteriority* Test time × hemisphere × anteriority × group*
Pair14/ T41		–	Perception-only & perception-plus-production: before training*	Anteriority* Hemisphere × group* Hemisphere × anteriority × group***
Pair41/ T14		–	Ns	Hemisphere*
Pair24/ T42	<Pair42** <Pair14*	–	Ns	Ns
Pair42/ T24		–	Ns	Group*
Pair11	–		–	–
Pair22	–	<Pair11*** <Pair44***	–	–
Pair44	–	<Pair11**	–	–

Ns: not significant.
* $p<.05$.
** $p<.01$.
*** $p<.001$.

stimulus once. This probably explains why they did not gain extra benefit from the production component.

3.2. *Asymmetrical perception of tone pairs*

Consistent with previous studies (e.g., So and Best, 2010; Wang et al., 1999), our behavioral and ERP results demonstrated that some tone pairs are more easily confused than others. However, inconsistent with Francis and Ciocca (2003), our data showed that native English listeners also discriminated tones asymmetrically. In the behavioral task, the English listeners discriminated pair21 and pair24 the least accurately, while pair42 and pair12 the most accurately. Francis and Ciocca (2003) reported that Cantonese listeners were more sensitive to low-high pairs of level tones than the acoustically identical high-low pairs, but native English listeners did not show such preference. In the present study, the native English listeners indeed demonstrated asymmetric tonal perception, yet in the opposite direction. Their discrimination was significantly more accurate if the F0 onset of the first tone was higher than the F0 onset of the second tone (i.e. high-low pairs: pair12 and pair42), compared with the reverse order (i.e. low-high pairs: pair21 and pair24). In addition, they discriminated pair14 and pair41 equally well in the behavioral task, because the two tones in these two pairs have the same F0 onset. In other words, the English listeners were most sensitive to the high-low pairs and least sensitive to the low-high pairs with the high-high pairs in between. These results echo previous studies (Francis et al., 2008; Gandour, 1983), which found that native English listeners tend to pay more attention to the F0 onset than the F0 direction in tone perception. Moreover, these results seem to suggest that native English listeners tend to hear the first tone as having a high F0 onset, which result in that pair21 was treated more like a 'same' pair (i.e. pair11), thus was poorly discriminated. This assumption also explains why pair22 was discriminated least accurately compared with the other two 'same' pairs. Since the first tone was heard as having a high F0 onset, pair22 was more likely to be treated as pair12 and consequently be discriminated as 'different' pair. As for pair24, it was very probably to be heard as pair14, therefore, was discriminated less accurately than pair42. However, this presumption could not explain why the discrimination of pair24 was also less accurate than pair14. There may be other factors contributing to the pattern of asymmetries in tone perception than simple F0 onset.

Our ERP data partially echo the behavioral data, demonstrating that the MMN component only appeared in the T21 (T2 as deviant and T1 as standard) and T41 conditions, but not in the T12, T14 conditions, even though they involved the same tone contrasts. Previous studies suggested that the asymmetries in

tone perception might be language-specific (e.g. Francis and Ciocca, 2003) and might result from increasing exposure to a tone language (Yeung et al., 2013). However, in the present ERP task the native English listeners displayed the asymmetric perception before training, but not after training. One way to account for this asymmetric pattern might be the interference of native intonation system. In the T21 condition, the deviant T2 (rising tone) was preceded by the standards T1 (level tone), resulting in a level-level-rising tone combination; in the T41 condition, the deviant T4 (falling tone) followed the standards T1 (level tone), resulting in a level-level-falling tone combination. These two types of tone combinations are very similar to the intonation patterns in the declarative and interrogative sentences in English. Therefore, the native English speakers were sensitive to the deviants T2 and T4 in the T21 and T41 conditions before training. After training the native language interference was weakened, yet the native English listeners had not formed tone categories properly. This might be why the MMN decreased and the asymmetric perception disappeared in the post-training task. For the T24 and T42 conditions, neither falling-falling-rising nor rising-rising-falling is a possible intonation contour in English. Thus, the native English listeners did not show any preference for either of the conditions.

An alternative account of these asymmetries appeals to *perceptual magnet effect* (Kuhl, 1991). The magnet effect argues that tokens that are close to phonetic category prototypes appear to be harder to discriminate than equally spaced tokens that are further away from the category prototypes, since the prototypes tend to pull neighboring tokens toward them. The effect has been commonly reported in vowel perception, where adults and infants often have difficulty discriminating vowel A in the context of vowel B, but not vice versa (e.g., Iverson and Kuhl, 1995; Kuhl, 1991; Kuhl et al., 1992; Polka and Werker, 1994). In our tone data, before training the native English listeners showed good unintentional discrimination when T2 and T4 were embedded in strings of T1, while displayed poor unintentional discrimination when T1 was inserted in the context of T2 and T4 respectively. This was probably due to the fact that before training the native English listeners perceived tones with respect to their native intonation categories and considered T2 and T4 as prototypes of rising and falling intonation categories respectively. These prototypes "attract" less prototypical T1 and make T1 less discriminable. After training, although the English listeners started making a difference between the native intonation categories and the non-native tone categories, they have not established clear tone categories yet. Thus they were more likely to perceive T1, T2 and T4 as peripheral exemplars of the tone categories and demonstrated poor unintentional discrimination. However, there is one significant problem with such an account. If T2 was interpreted as prototype of the native rising intonation category and T4 as prototype of the native falling intonation category before training, the discrimination between these tones should be excellent (Perceptual Assimilation Model, Best, 1995; Best and Tyler, 2007). Nevertheless, in the present experiment neither T24 nor T42 condition displayed MMN before training. Further research would be necessary to examine why native English listeners exhibited different asymmetric patterns in tone perception at the intentional (behavioral) and unintentional levels and whether the asymmetries can be affected by an extended period of training.

In conclusion, the present study showed that after a perception-only training or a perception-plus-production training the native English listeners improved in tone discrimination ability at the intentional but not unintentional level. Moreover, the participants who received the perception-plus-production training did not show more improvements compared to the participants who received the perception-only training. These results imply that the employment of motor system does not specifically benefit the tone perceptual skills. Lastly, the native English listeners perceived tones asymmetrically, which might be attributed to native language interference. The data of the present study suggest that there is a transfer threshold between perceptual learning and production learning (Roye et al., 2013; Seitz and Dinse, 2007). Training in one modality needs to be sufficient to drive cognitive and behavioral improvements in the other modality. Furthermore, we surmise that reinforcement and feedback of training can boost learning in one modality to surpass this transfer threshold, whereas native language knowledge can interfere or facilitate learning in both modalities. Future research on the effectiveness of perceptual training on production will be required to test this hypothesis.

4. Methods and materials

4.1. *Participants*

Twenty-two native speakers of American English were recruited in the present study. Participants were randomly assigned to two groups: 11 female participants received a perception-only training and the other 11 participants (2 males and 9 females) received a perception-plus-production training. All participants were between the ages of 18 and 25 years old ($M=19.14$, $SD=1.28$), and were undergraduate students at the University of Florida. None of the participants had previous exposure to any tone languages. All participants were right-handed according to Edinburgh Handedness inventory (Oldfield, 1971), had normal vision and had no history of speech or neurological impairment according to self-report. All participants had normal hearing (between 250 Hz and 8 kHz at 25 dB) as tested on-site. In addition, participants were tested on auditory working memory (forward and backward digit spans). There was no significant difference between the overall auditory memory scores for the two groups [$t\ (20)=1.38$, $p=.18$]. Moreover, none of the participants had received more than two years of formal musical training and had not been performing music within the past five years at the time of their participation. Participants' musical aptitude was also tested on site using the Advanced Measures of Music Audiation (AMMA, Gordon, 1989). The participants in the two groups did not differ in the musical tone perception [$t\ (20)=-0.62$, $p=.54$], or the music rhythm perception [$t\ (20)=-0.35$, $p=.73$]. An additional four participants were recruited, but were excluded from data analysis due to excessive eye or body movement artifacts, or technical problems. All participants received financial compensation or course credit for their participation. The experimental procedures were approved by the University of Florida Institutional Review Board.

4.2. Stimuli

4.2.1. Behavioral task

Stimuli consisted of eight monosyllabic syllables ([p^{h}a], [p^{h}i], [k^{h}ɛ], [k^{h}o], [t^{h}a], [t^{h}i], [t^{h}ɛ] and [t^{h}o]) associated with three linear tones that resemble Mandarin T1 (high-level), T2 (high-rising), and T4 (high-falling). T3 (low-dipping) was not included because it has been shown to be the most confusable tone for both native Mandarin speakers and non-native speakers (e.g. Kirkham et al., 2011). The syllables [t^{h}a], [t^{h}i], [t^{h}ɛ], [t^{h}o], [k^{h}o] and [p^{h}a] were used in the pre- and post-training tasks, and the syllables [p^{h}a], [p^{h}i], [k^{h}ɛ] and [k^{h}o] were used in the training session. The syllables [t^{h}a], [t^{h}i], [t^{h}ɛ] and [t^{h}o] were not included in the training session because they were used to test whether participants could generalize the improvements gained through training to untrained stimuli. All of these syllables are legal syllable structures in American English, thus American speakers could focus their attention on the lexical tones and not be distracted by unfamiliar syllables. Two female native speakers of American English were asked to produce each syllable twice with a high pitch in a sound-attenuated booth, yielding four tokens for each syllable. Only female speakers were recorded for stimuli because adult female speakers have been found to be more intelligible than adult male speakers (Bradlow et al., 1996). Three linear pitch contours were superimposed starting from the voiced part of each syllable, using the pitch-synchronous overlap and add (PSOLA) function in the Praat software (Boersma and Weenink, 2010). The procedures of stimulus manipulation were similar to Wong and Perrachione (2007). The onset value of T1 was the mean fundamental frequency (F0) of all syllables produced by the two speakers. The offset of T1 was identical to its onset. T2 had the same ending point as T1, and its starting point was 26% lower than its ending point. Originally the onset of T4 was set to 10% higher than that of T1 and dropped by 82%, according to the values of Mandarin T4 reported in Shih (1988). Pilot data of seven participants, however, displayed a ceiling effect, i.e. there was little room for participants to improve through the training session. Then we decided to set the onset of T4 to be the same as T1 and its offset at 26% lower than its onset. Therefore, the onset of T2 was identical to the offset of T4. All stimuli were then normalized to the same peak intensity. Except for F0 and intensity, all other acoustic features (e.g. duration and voice quality characteristics) were kept identical to the speakers' original productions. As mentioned previously, each syllable was produced twice by two speakers, resulting in four tokens for each syllable. The three tones were generated for each token, which means that (1) within each token the three tones had the same acoustic features except for F0; (2) the same tones for different tokens had the same F0, but different duration and voice quality. All stimuli were judged as perceptually natural by three native Mandarin speakers and three native American English speakers. A second pilot experiment with 8 participants was carried out to ensure that the discrimination of the tones in untrained native English speakers did not show a ceiling effect. Fig. 7 shows the waveforms and spectrograms with the pitch contours of the T1, T2 and T4 of one token of the syllable [t^{h}u].

4.2.2. ERP

In the ERP experiment, the stimuli were the [t^{h}u] syllable associated with the three tones (T1, T2 and T4). These stimuli were not included in the behavioral tasks or the training sessions. The stimulus recording and manipulation were exactly the same as described in Section4.2.1. The mean duration of all the tokens was 553 ms (SD=54.66). On average, pitch contour started at 108 ms (SD=14.18) after the stimulus onset.

4.3. Procedure

The experiment took place over the course of three consecutive days. On the first day, baseline EEG data were recorded on the first day from all participants in both groups. The ERP experiment used the passive oddball paradigm, in which participants were watching a silent movie in a sound-attenuating booth while the stimuli were presented binaurally through inserted ear buds at a constant and comfortable hearing level. Participants were asked to ignore the sounds and focus their attention on the movie.

The EEG experiment consisted of three types of blocks: in the first block, T1 was the standard, and both T2 and T4 were the deviants; in the second block, T2 was the standard, with T1 and T4 being the deviants; in the third block, T4 was the standard, and T1 and T2 were the deviants (Näätänen et al., 2004). Within each block each token of the deviants was presented 25 times, and each token of the standards was presented 250 times, resulting in 200 deviants (25 × 2 tokens/speaker × 2 speakers × 2

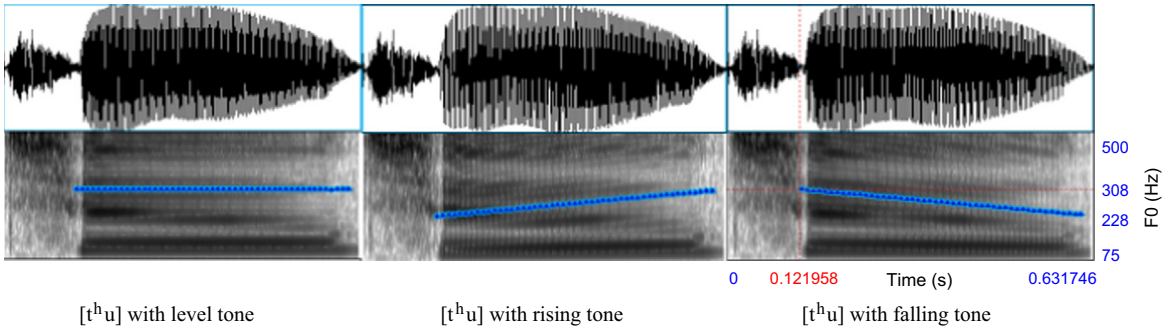

Fig. 7 – Spectrograms of Tone1, Tone2 and Tone4 for the syllable [t^{h}u]. For Tone1, the F0 onset and offset values are 308 Hz. For Tone2, the F0 onset and offset values are 228 Hz and 308 Hz respectively. For Tone4, the F0 onset and offset values are 308 Hz and 228 Hz respectively. (Note: the mean duration of the stimuli was 553 ms; the blue lines indicate the pitch contours of the tones; on average, the pitch contours started at 108 ms (SD=14.18) after the stimulus onset).

tones) and 1000 standards (250 × 2 tokens/speaker × 2 speakers × 1 tones). The stimuli were presented in a pseudo random order, with at least two standards preceding each deviant. Within each block we selected 100 standards (25 × 2 tokens/speaker × 2 speakers) which were not immediately preceded or followed by any deviants. These standards were compared with the deviant stimuli in data analysis. The offset-to-onset inter-stimulus interval randomly varied between 500 and 650 ms to prevent participants' automatic anticipation of stimulus onset and synchronization of EEG rhythms with the presentation. Each block was split equally into two sub-blocks, with each sub-block lasting for about 12 min. The order of the six sub-blocks was randomized across participants, with a different order being presented before and after the training session. For each participant, the order of stimulus presentation in each sub-block was the same before and after training. In order to familiarize the participants with the experimental setting, before the experimental blocks all participants received a practice block, in which each token of the three tones was equally presented for 40 times. The practice block was excluded from data analysis.

After each block, participants received comprehension questions regarding the silent movie. In order to control for potential attentional differences across groups, participants were also asked to count the number of times a particular object occurred in the movie. Only participants who answered more than 60% of the questions correctly, and were able to produce the correct count (plus or minus 3) were included for further data analysis. The movies in the pre- and post-training sessions were different. The order of the two movies was counterbalanced across participants, with half of the participants watching the first movie in the pre-training session while the other half of the participants watching the first movie in the post-training session. Participants were encouraged to take breaks after each block to maintain concentration during blocks and to prevent fatigue.

Continuous EEG was recorded from 39 Ag/AgCl scalp electrodes (Fp1/2, F7/8, F5/6, F3/4, Fz, FT7/8, FC5/6, FC3/4, FCz, T7/8, C5/6, C3/4, Cz, TP7/8, CP5/6, CP3/4, FPz, P7/8, P5/6, P3/4, Pz, O1/2) at 512 Hz. The electrodes were mounted in an elastic cap (Easy-Cap, Falk Minow, Herrsching-Breitbrunn, Germany) and an ANT amplifier (ANT Software b.v., Enschede, The Netherlands). The EEG was referenced to the right mastoid. Horizontal eye movements were recorded by electrodes on the left and right outer canthi, while the vertical eye movements were recorded by electrodes placed above and below the right eye. Impedances were kept below 5 kΩ.

After the ERP recording, participants did a behavioral discrimination task. The discrimination task used a forced-choice AX paradigm, in which the stimuli (the syllables [t^ha], [t^hi], [$t^h\varepsilon$], [t^ho], [k^ho] and [p^ha]) were presented in pairs and were separated by 500 ms inter-stimulus-interval (ISI). The ISI was set to 500 ms because previous studies have shown that participants were sensitive to nonnative sound contrasts with a 500 ms ISI (e.g. Cowan and Morse, 1986; Pisoni, 1973; Van Hessen and Schouten, 1992). In the different pairs the two stimuli in each pair had two different tones which were generated from the same token of the same syllable produced by one speaker. In the same pairs the two stimuli in each pair had the same tones, but were generated from different tokens of the same syllable produced by one speaker. The participants were asked to focus only on the tones and ignore other acoustic features of the stimuli and indicate whether the two stimuli in each pair had the same or different tones by pressing a button on a button box. For half of the participants the leftmost button on the button box indicated 'same' and the rightmost button indicated 'different', and the vice versa for the other half of the participants. This design ensured that the results were not biased by participants' button pressing preference. A total of 288 trials (6 syllables) were presented, with 144 'same' and 144 'different' trials. The order of the trials was randomized for each participant. The experimental trials were preceded by a short tutorial on the three tones and 6 practice trials which were excluded from data analysis. The participants were encouraged to respond as quickly and accurately as possible. Their accuracies and reaction times were logged by the E-prime 2.0 software (Psychology Software Tools, Pittsburgh, PA). No feedback was given through this task. A break of at least 30 s was given halfway through the task, in order to prevent fatigue.

On the second day, the participants received either a perception-only or a perception-plus-production training, which took about one hour. Note that testing production without perception is hardly possible; the two training types were therefore similar in terms of auditory exposure to the stimuli, but differed in the production component: the production training required participants to imitate the stimuli while the perception training had participants utter a word unrelated to the stimuli. In this way, the production training and perception training could be directly compared, with participants in both groups receiving the same exposure and attention to the same stimuli. In perception-only training, participants (1) first heard one stimulus; (2) heard another stimulus after 500 ms; (3) determined whether these two stimuli had the same or different tones; (4) received correct answer feedback with the tone type and graphic indication of each stimulus; (5) heard the first stimulus again and saw the tone and the graphic indication of the tone simultaneously; (6) Said 'next'; (7) heard the second stimulus again and saw the tone and graphic indication at the same time; and (8) Said 'Next'. The next trial started three seconds after the participant's production in (8). Participants in the production group received exactly the same training as participants in the perception-only group, except that they were asked to imitate the stimulus instead of saying 'Next' in steps (6) and (8). They were encouraged to imitate the tones as accurately as they could, and were allowed to try only once. In order to match the production training to the perception training, the participants in the production group did not receive any feedback about their imitations. All participants' productions were recorded by a Marantz PMD660 digital recorder for assessment of the participants' performance, but these data were not used in data analysis.

For each participant, the layout of the 'same' and 'different' buttons on the button box was identical in the pre- and post-training tasks and the training session. Regardless of the training type, a total of 192 trials (the syllables [p^ha], [p^hi], [$k^h\varepsilon$] and [k^ho]) were presented in the training session, with 96 'same' and 96 'different' trials. The order of the trials was randomized for each participant. The experimental trials were preceded by 6 practice trials which familiarized the participants with the training process and were excluded from data analysis. The participants were encouraged to respond as quickly and

accurately as possible. Their accuracy and reaction times in step (3) were logged by the E-prime 2.0 software (Psychology Software Tools, Pittsburgh, PA). A break of at least 30 s was given halfway through the training session. After the 30 s break, the participants could take another break as long as they wanted and started the second half of the training session by pressing any button on the button box.

On the third day, both groups did the same ERP and behavioral tasks as described under Day1. After the participants finished all tasks, they were asked to self-evaluate their performance and give suggestions on ways to improve the quality of participation experience. The entire experiment took about 10 h for each participant.

4.4. *Data analysis*

4.4.1. *Behavioral task*

Data obtained from the pre- and post-training discrimination tasks and the training session were converted to d' scores, which were calculated as the difference between the z-transforms of hit rates (correct responses in the 'different' trials) and false alarm rates (incorrect responses in the 'same' trials). Hit rates of 1 were corrected with 1-1/2N, where N equaled the total number of the 'different' trials. False alarm rates of 0 were replaced by 1/2N, where N equaled the total number of the 'same' trials (Macmillan and Creelman, 2005). Separate *d*' scores for the trained and untrained stimuli in the pre- and post-training tasks were also calculated, in order to examine participants' generalization abilities. Furthermore, the percentage of accurate responses for each tone pair-type was calculated in order to explore which tone pair-type showed the greatest improvement and which type was the most resistant for improvement for each group.

Participants' reaction times (RTs) in the pre- and post-training tasks and the training session were also calculated to see whether the participants responded faster in the post-training tasks than in the pre-training tasks. The RTs were computed only for the correct trials. RTs that greater than the mean RT of all the correct responses plus 2.5 standard deviations for an individual participant were replaced by this value. After the outlier replacement, the mean RTs of the total stimuli, trained stimuli (the syllables [k^ho] and [p^ha]) and untrained stimuli (the syllables [t^ha], [t^hi], [$t^h\varepsilon$] and [t^ho]) for each participant in the pre- and post-training tasks were computed.

4.4.2. *ERP*

The EEG was off-line band-pass filtered from 0.16 to 30 Hz, and arithmetically re-referenced to the mean of both mastoids after recording. Artifact-free epochs were analyzed from −200 to 900 ms relative to the onset of the stimulus, using the 200 ms window preceding the onset as a baseline. On average, 66 trials (58%) for each condition were included in data analysis after artifact rejection in the pre- and post-training tasks for the perception-only and perception-plus-production groups respectively.

The grand-averaged difference waves were generated for each of six conditions by subtracting the average responses to the clean standard stimuli from average responses to the corresponding deviant stimuli. The six conditions were (1) deviant T1 in standards T2 (T12); (2) deviant T1 in standards T4 (T14); (3) deviant T2 in standards T1 (T21); (4) deviant T2 in standards T4 (T24); (5) deviant T4 in standards T1 (T41); and deviant T4 in standards T2 (T42). The mean amplitude of the time windows spanning the MMN and the late negativity of the difference waves were computed for each channel, participant and condition in the pre- and post-training tasks. Based on visual inspection of the raw ERPs and previous studies, the mean amplitude of the MMN was computed between 200 and 400 ms. The MMN is typically largest over the frontal electrodes (e.g. Näätänen and Michie, 1979; Alain et al., 1998). Therefore, our analyses on the MMN only focused on the frontal electrodes (F4, F6, F8, FC4, FC6, FT8, F3, F5, F7, FC3, FC5, FT7, Fz and FCz). The mean amplitude of the late negativity was computed between 500 and 800 ms after stimulus onset for four regions: right-frontal (RF) included the electrodes F4, F6, F8, FC4, FC6 and FT8; left-frontal (LF) included the electrodes F3, F5, F7, FC3, FC5 and FT7; right-posterior (RP) included the electrodes CP4, CP6, CP8, P4, P6 and P8; and left-posterior (LP) included the electrodes CP3, CP5, TP7, P3, P5, and P7.

All the *p*-values and the *F*-values were adjusted using the Greenhouse–Geisser correction (Greenhouse and Geisser, 1959) and the post-hoc paired *t*-tests were adjusted using the Bonferroni correction for multiple comparisons.

Acknowledgements

We would like to thank Eric Holgate for his assistance in this research. This work was supported by *Language Learning* Dissertation grant to the first author.

REFERENCES

Adank, P., Haggort, P., Bekkering, H., 2010. Imitation improves language comprehension. Psychol. Sci. 21 (12), 1903–1909.

Alain, C., Woods, D.L., Knight, R.T., 1998. A distributed cortical network for auditory sensory memory in humans. Brain Res. 812, 23–37.

Baese-Berk, M.M., 2010. An Examination of the Relationship Between Speech Perception and Production (Doctoral dissertation). Northwestern University.

Best, C.T., 1995. A direct realist view of cross-language speech perception. In: Strange, W. (Ed.), Speech Perception and Linguistic Experience: Issues in Cross-Language Research. Baltimore, York, pp. 171–204.

Best, C.T., Tyler, M.D., 2007. Nonntaive and second-language speech perception. In: Bohn, O. -S., Munro, M.J. (Eds.), Language experience in second language speech learning: In honor of James Emil Flege. J. Benjamins, Amsterdam, pp. 13–34.

Binder, J.R., Liebenthal, E., Possing, E.T., Medler, D.A., Ward, B.D., 2004. Neural correlates of sensory and decision processes in auditory object identification. Nat. Neurosci. 7 (3), 295–301.

Boersma, P., & Weenink, D., 2010. Praat: Doing Phonetics by Computer [Computer program]. Version 5.1.25. Retrieved from ⟨http://www.praat.org/⟩ (31.03.10).

Bradlow, A.R., Torretta, G.M., Pisoni, D.B., 1996. Intelligibility of normal speech I: Global and fine-grained acoustic phonetic talker characteristics. Speech Commun. 20, 255–272.

Callan, D., Callan, A., Gamez, M., Sato, M.A., Kawato, M., 2010. Premotor cortex mediates perceptual performance. NeuroImage 51 (2), 844–858.

Callan, D.E., Jones, J.A., Callan, A.M., Akahane-Yamada, R., 2004. Phonetic perceptual identification by native- and second-

language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory auditory/orosensory internal models. NeuroImage 22 (3), 1182–1194.
Chandrasekaran, B., Krishnan, A., Gandour, J., 2007. Neuroplasticity in the processing of pitch dimensions: a multidimensional scaling analysis of the mismatch negativity. Restor. Neurol. Neurosci. 25, 195–210.
Chao, Y.R., 1948. Mandarin Primer. Harvard University Press, Cambridge.
Chevillet, M.A., Jiang, X., Rauschecker, J.P., Riesenhuber, M., 2013. Automatic phoneme category selectivity in the dorsal auditory stream. J. Neurosci. 33 (12), 5208–5215.
Cowan, N., Morse, P.A., 1986. The use of auditory and phonetic memory in vowel discrimination. J. Acoust. Soc. Am. 79, 500–507.
Du, Y., Buchsbaum, B.R., Grady, C.L., Alain, C., 2014. Noise differentially impacts phoneme representations in the auditory and speech motor systems. Proc. Natl. Acad. Sci. 111 (19), 7126–7131.
Flege, J.E., 1995. Second language speech learning: Theory, findings and problems. In: Strange, W. (Ed.), Speech Perception and Linguistic Experience: Issues in Cross-language Research. Baltimore, York, pp. 233–277.
Francis, A.L., Ciocca, V., 2003. Stimulus presentation order and the perception of lexical tones in Cantonese. J. Acoust. Soc. Am. 114 (3), 1611–1621.
Francis, A.L., Ciocca, V., Ma, L., Fenn, K., 2008. Perceptual learning of Cantonese lexical tones by tone and non-tone language speakers. J. Phon. 36 (2), 268–294.
Gandour, J., 1983. Tone perception in Far Eastern languages. J. Phon. 11, 49–175.
Gottfried, T.L., Suiter, T.L., 1997. Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones. J. Phon. 25 (2), 207–231.
Greenhouse, S.W., Geisser, S., 1959. On methods in the analysis of profile data. Psychometrika 24, 95–112.
Gordon, E.E., 1989. Manual for the Advanced Measures of Music Audiation. GIA, Chicago.
Guenther, F.H., Ghosh, S.S., Tourville, J.A., 2006. Neural modeling and imaging of the cortical interactions underlying syllable production. Brain Lang. 96, 280–301.
Hattori, Kota, 2010. Perception and Production of English /r/ -/l/ by Adult Japanese Speakers (Doctoral dissertation). University of College London.
Herd, W., 2011. The Perceptual and Production Training of /d, ɾ, r/ in L2 Spanish: Behavioral, Psycholinguistic, and Neurolinguistic Evidence (Doctoral dissertation). University of Kansas.
Hickok, G., Houde, J., Rong, F., 2011. Sensorimotor integration in speech processing: computational basis and neural organization. Neuron 69 (3), 407–422.
Hickok, G., Poeppel, D., 2000. Towards a functional neuroanatomy of speech perception. Trends Cogn. Sci. 4 (4), 131–138.
Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99.
Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402.
Hirata, Y., 2004. Computer assisted pronunciation training for native English speakers learning Japanese pitch and durational contrasts. Comput. Assist. Lang. Learn. 17, 357–376.
Iverson, P., Kuhl, P.K., 1995. Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. J. Acoust. Soc. Am. 97 (1), 553–562.
Kaan, E., Barkley, C., Bao, M., Wayland, R., 2008. Thai lexical tone perception in native speakers of Thai, English and Mandarin Chinese: an event-related potentials training study. BMC Neurosci. 9, 53.
Kaan, E., Wayland, R., Bao, M., Barkley, C., 2007. Effects of native language and training on lexical tone perception: an ERP study. Brain Res. 1148, 113–122.
Kirkham, J., Lu, S., Wayland, R., Kaan, E., 2011. Comparison of vocalists and instrumentalists on lexical tone perception and production tasks. In: Proceedings of the 17th International Congress of Phonetic Sciences, pp. 1098–1101.
Krishnan, A., Xu, Y, Gandour, J.T., Cariani, P., 2005. Encoding of pitch in the human brainstem is sensitive to language experience. Cogn. Brain Res. 25, 161–168.
Kuhl, P.K., 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, Monkeys do not. Attent. Percept. Psychophys. 50, 93–107.
Kuhl, P.K., Williams, K.A., Lacerda, F., Stevens, K.N., Lindblom, B., 1992. Linguistic experience alters phonetic perception in infants by 6 months of age. Science 255 (5044), 606–608.
Leather, J, 1990. Perceptual and productive learning of Chinese lexical tone by Dutch and English speakers. In: Leather, L., James, A. (Eds.), New Sounds 90: Proceedings of the Amsterdam Symposium on the Acquisition of Second Language Speech. University of Amsterdam, Amsterdam, pp. 305–341.
Levelt, W.J.M., Praamstra, P., Meyer, A.S., Helenius, P., Salmelin, R., 1998. An MEG study of picture naming. J. Cogn. Neurosci. 10, 553–567.
Liberman, A.M., Mattingly, I.G., 1985. The motor theory of speech perception revised. Cognition 21 (1), 1–36.
Macmillan, N.A., Creelman, C.D., 2005. Detection Theory: A User's Guide, 2nd ed. Lawrence Erlbaum Associates, Mahwah, NJ.
Näätänen, R., Alho, K., 1995. Mismatch negativity—a unique measure of sensory processing in audition. Int. J. Neurosci. 80, 317–337.
Näätänen, R., Michie, P.T., 1979. Early selective-attention effects on the evoked potential: a critical review and reinterpretation. Biol. Psychol. 8, 81–136.
Näätänen, R., Pakarinen, S., Rinne, T., 2004. The mismatch negativity (MMN): towards the optimal paradigm. Clin. Neurophysiol. 115 (1), 140–144.
Oldfield, R.C., 1971. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia 9, 97–113.
Pisoni, D.B., 1973. Auditory and phonetic memory codes in the discrimination of consonants and vowels. Percept. Psychophys. 13, 253–260.
Poeppel, D., 2001. Pure word deafness and the bilateral processing of the speech code. Cogn. Sci. 25, 679–693.
Pulvermüller, F., Huss, M., Kherif, F., del Prado Martin, F.M., Hauk, O., Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. 103 (20), 7865–7870.
Polka, L., Werker, J.F., 1994. Developmental changes in perception of nonnative vowel contrasts. J. Exp. Psychol.: Hum. Percept. Perform. 20, 421–435.
Roye, A., Jacobsen, T., Schröger, E., 2013. Discrimination of personally significant from nonsignificant sounds: a training study. Cogn. Affect. Behav. Neurosci. 13, 930–943.
Seitz, A.R., Dinse, H.R., 2007. A common framework for perceptual learning. Curr. Opin. Neurobiol. 17, 1–6.
Shen, X.S., 1989. Toward a register approach in teaching Mandarin tones. J. Chin. Lang. Teach. Assoc. 24, 27–47.
Shestakova, A., Huotilainen, M., Ceponienè, R., Cheour, M., 2003. Event-related potentials associated with second language learning in children. Clin. Neurophysiol. 114, 1507–1512.
Shih, C., 1988. Tone and intonation in Mandarin. Work. Pap. Cornel. Phon. Lab. 3, 83–109.
So, C.K., Best, C.T., 2010. Cross-language perception of non-native tonal contrasts: effects of native phonological and phonetic influences. Lang. Speech 53 (2), 273–293.

Song, H.J., Skoe, E., Wong, P.C.M., Kraus, N., 2008. Plasticity in the adult human auditory brainstem following short-term linguistic training. J. Cogn. Neurosci. 20 (10), 1892–1902.
Tremblay, K., Kraus, N., Carrell, T.D., McGee, T., 1997. Central auditory system plasticity: generalization to novel stimuli following listening training. J. Acoust. Soc. Am. 102 (2), 3762–3773.
Tremblay, K., Kraus, N., McGee, T., 1998. The time course of auditory perceptual learning: neurophysiological changes during speech-sound training. Neuroreport 9 (16), 3557–3560.
Van Hessen, A.J., Schouten, M.E.H, 1992. Modeling phoneme perception: II. A model of stop consonant discrimination. J. Acoust. Soc. Am. 92, 1856–1868.
Wang, Y., Jongman, A., Sereno, J.A., 2003. Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. J. Acoust. Soc. Am. 113 (2), 1033–1043.
Wang, Y., Spence, M.M., Jongman, A., Sereno, J.A., 1999. Training American listeners to perceive Mandarin tone. J. Acoust. Soc. Am. 106, 3649–3658.
Wayland, R., Guion, S., 2004. Training native English and native Chinese speakers to perceive Thai tones. Lang. Learn. 54 (4), 681–712.
Wayland, R., Li, B., 2008. Effects of two training procedures in cross-language perception of tones. J. Phon. 36, 250–267.
White, C.M., 1981. Tonal perception errors and interference from English intonation. J. Chin. Lang. Teach. Assoc. 16, 27–56.
Wilson, S.M., Iacoboni, M., 2006. Neural responses to non-native phonemes varying in producibility: evidence for the sensorimotor nature of speech perception. NeuroImage 33 (1), 316–325.
Wilson, S.M., Saygin, A.P., Sereno, M.I., Iacoboni, M., 2004. Listening to speech activates motor areas involved in speech production. Nat. Neurosci. 7 (7), 701–702.
Woldorff, M.G., Hackley, S.A., Hillyard, S.A., 1991. The effects of channel-selective attention on the mismatch negativity wave elicited by deviant tones. Psychophysiology 28 (1), 30–42.
Wong, P.C.M., Perrachione, T.K., 2007. Learning pitch patterns in lexical identification by native English-speaking adults. Appl. Psycholinguist. 28, 565–585.
Yeung, H.H., Chen, K.H., Werker, J.F., 2013. When does native language input affect phonetic perception? The precocious case of lexical tone. J. Mem. Lang. 68, 123–139.
Zachau, S., Rinker, T., Körner, B., Kohls, G., Maas, V., Henninghausen, K., Schecker, M., 2005. Extracting rules: early and late mismatch negativity to tone patterns. Neuroreport 16, 2015–2019.