Published in final edited form as:

Lang Cogn Neurosci. 2018 ; 33(6): 734–749. doi:10.1080/23273798.2017.1421317.

Adaptation in Mandarin tone production with pitch-shifted auditory feedback: Influence of tonal contrast requirements

Yongqiang Feng^a, Yan Xiao^b, Yonghong Yan^a, and Ludo Max^{c,d}

^aInstitute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China

^bInstitute of Information Science, Beijing Jiaotong University, Beijing 100044, China

^cDepartment of Speech and Hearing Sciences, University of Washington, Seattle, WA 98105-6246, USA

^dHaskins Laboratories, New Haven, CT 06511, USA

Abstract

We investigated Mandarin speakers' control of lexical tone production with F0-perturbed auditory feedback. Subjects produced high level (T1), mid rising (T2), low dipping (T3), and high falling (T4) tones in conditions with (a) no perturbation, (b) T1 shifted down, (c) T1 shifted down and T3 shifted up, or (d) T1 shifted down and T3 shifted up but without producing other tones. Speakers and new subjects also completed a tone identification task with unaltered and F0-perturbed productions. With only T1 perturbed down, speakers adapted by raising F0 relative to no-perturbation. With simultaneous T1 down and T3 up perturbations, no T1 adaptation occurred, and T3 adaptation occurred only if T2 was also produced. Identification accuracy with stimuli representing adapted productions was comparable to baseline, but with simulated non-adapted productions it was reduced for T2 and T3. Thus, Mandarin speakers' adaptation to F0 perturbations is linguistically constrained and serves to maintain tone contrast.

Keywords

auditory feedback; tone production; sensorimotor adaptation; tone identification

Auditory feedback plays an important role in speech production. Feedback about the acoustic end product is critical not only for online control of phonation and articulation but also for new sound learning and long-term performance maintenance. Perturbations to the auditory feedback have been widely used in studies on the sensorimotor mechanisms underlying speech production. In general, the experimental paradigms used by these studies can be divided into two types. In one paradigm, subjects are not able to predict the perturbation because its occurrence is sudden and random. Acoustic parameters that have been manipulated in such studies include formant frequencies (e.g., Purcell & Munhall, 2006) and the fundamental frequency (F0) (e.g., Burnett, Freedland, Larson, & Hain, 1998; Chen, Liu, Xu, & Larson, 2007; Liu, Chen, Larson, Huang, & Liu, 2010; Xu, Larson, Bauer,

& Hain, 2004). Analyses in these studies focus on immediate, within-trial compensatory responses. In the second type of paradigm, which will be adopted here for the present study, auditory feedback is manipulated in a predictable manner across multiple trials (Feng, Gracco, & Max, 2011; Houde & Jordan, 1998; Jones & Munhall, 2000; Max, Wallace, & Vincent, 2003; Villacorta, Perkell, & Guenther, 2007). Regardless of whether or not subjects become aware of the perturbation, and even when instructed to ignore the perturbation, they gradually show adaptive changes—opposite to the direction of the perturbation—to compensate for the altered feedback. Such adaptation occurs for formant frequency perturbations (e.g., Cai, Ghosh, Guenther, & Perkell, 2010; Feng et al., 2011; Houde & Jordan, 1998; Jones & Munhall, 2000; MacDonald, Purcell, & Munhall, 2011; Max et al., 2003; Villacorta et al., 2007) as well as F0 perturbations (e.g., Jones & Munhall, 2000, 2002, 2005; Keough, Hawco, & Jones, 2013). This type of adaptive learning is typically interpreted in the context of acquiring or updating internal models—stored mappings of the relationship between motor commands and their sensory consequences for a given environment (Houde & Jordan, 2002; Jones & Munhall, 2005; Max, 2004).

Adaptation to a single auditory perturbation (i.e., one perturbation that is applied to all productions) indicates that subjects are able to update a single internal model of the motor-to-auditory mapping of the vocal tract (for formants) or the laryngeal mechanism (for F0). In several studies of limb motor control, researchers have introduced *two different* perturbations and demonstrated that both human subjects and nonhuman primates are able to switch between two learned motor-to-sensory mappings when contextual or postural cues are available (Bock, Worringham, & Thomas, 2005; Galea & Miall, 2006; Gandolfo, Mussa-Ivaldi, & Bizzi, 1996; Krouchev & Kalaska, 2003; Mistry & Contreras-Vidal, 2004; Woolley et al., 2007). In speech research, very few experimenters have taken this approach, with exceptions found in a formant-shift study by Rochet-Capellan and Ostry (2011) and a F0-shift study by Keough and Jones (2011). In the latter study, trained singers reproduced a sequence of target notes (A4-G4-F4 during the first half of trials and F4-G4-A4 during the second half of trials for one group of participants and *vice versa* for another group) while a F0 up-shift and F0 down-shift were applied to the auditory feedback for the first and third notes, respectively. Thus, the three specific notes within the sequence served as cues for the type of perturbation (in the order F0 up, no perturbation, and F0 down). Subjects showed adaptation for both the up-shift and down-shift perturbations, and Keough and Jones (2011) concluded that trained singers are able to acquire multiple frequency-specific internal models for vocal pitch control if contextual cues (in this case, order in a sequence) are available. Additional insights into the ability to learn multiple internal models of the phonation system may be gained by studying *typical speakers* in studies that apply a set of perturbations based on *implicit cues* such as the specific sound or word to be produced.

Another limitation within the extant literature is that studies of F0 adaptation have not explored the influence of a speaker's entire language *system*, including phonological and lexical constraints that require a speaker's acoustic output for both perturbed and non-perturbed sounds to reach distinct, non-confusable targets (for formant adaptation, on the other hand, see Mitsuya, Macdonald, Purcell, & Munhall, 2011; Mitsuya, Samson, Ménard, & Munhall, 2013). That is, although experiments have included Mandarin speakers to investigate within-trial reflexive responses to unanticipated pitch perturbations (Chen et al.,

2012; Liu, Wang, Chen, Liu, Larson, & Huang, 2010; Liu, Chen, Larson, Huang, & Liu, 2010; Ning, Shih, & Loucks, 2014; Xu et al., 2004), the unique opportunities offered by tone languages to study the *learning* of F0 control within the context of an overall tone system have not been fully exploited¹. In an experiment by Jones and Munhall (2002), F0 in the auditory feedback was raised or lowered in a predictable manner while subjects sustained high level tones. After trials with unaltered feedback, F0 in the feedback was gradually raised or lowered, then held at a maximum shift of +1 or –1 semitone, and finally returned to normal. Subjects increased F0 relative to baseline in both conditions but more so in the F0-down condition than in the F0-up condition. In a follow-up study (Jones & Munhall, 2005), adaptation to F0-up perturbations for high level tones also generalized to mid rising tones. Those authors concluded that speakers use an internal representation for the long-term calibration of vocal pitch. In short, F0 adaptation has been previously documented for sustained Mandarin tones (Jones & Munhall, 2002; 2005), but not for typical tone productions in existing words. Furthermore, when studying typical tone production, it will be of critical importance to also take into account the interdependence among tone categories (i.e., considering the role of tonal contrasts in distinguishing lexical words) rather than ignoring the linguistic purpose of tones. Here, we take such a linguistically-meaningful approach with a novel paradigm that applies simultaneous, selective perturbations to different tone categories during the production of real words to allow for an investigation of the covariation of tones at a *system* level.

Specifically, in the present study, we aim to answer four questions: (a) do Mandarin speakers adapt to a F0 perturbation applied to only one tone category when the other three tone categories are also produced but remain unperturbed?, (b) do Mandarin speakers adapt to two simultaneous perturbations applied to two different tone categories?, (c) in the presence of perturbations on one or two tones, do Mandarin speakers show accompanying changes in the production of other, non-perturbed tones such that the overall system of distinct tones remains intact?, (d) how accurate is Mandarin speakers' tone identification when hearing the auditory feedback signal from non-adapted and adapted pitch-shifted tones, and can perceptual identification constraints explain the pattern of sensorimotor adaptation addressed in the preceding questions. We investigated the first three questions in Experiment 1, a tone adaptation experiment, and the fourth question in Experiment 2, a tone identification experiment.

¹There are four lexical tones in Mandarin: high level, mid rising, low dipping, and high falling tones, conventionally referred to as Tones 1, 2, 3, & 4, respectively. Tone 3 has the most intrinsic variation in that the F0 falls and then rises in citation form (i.e., produced in isolation) or in phrase-final positions, but it only falls (without the subsequent rise) in non-final positions in connected speech (Liu & Samuel, 2004; Xu, 1994). Although the low dipping tone in citation form has a dynamic contour, a large body of studies has shown that the initial fall and final rise are not important for identifying the tone (Liu & Samuel, 2004; Whalen & Xu, 1992). In addition, a low flat pitch is used as one of the variants in running speech, so that the tone is essentially a low level tone according to the definition of level tones (Maddieson, 1978). Some speakers use a creaky voice quality during the middle portion of low dipping tones and at the offset of high falling tones, & low dipping tones are sometimes even characterized by a loss of voicing at their lowest point which results in irregularity or discontinuity in the F0 contour (Belotel-Grenié & Grenié, 2004; Chao, 1956; Davison, 1991; Gårding, Kratochvil, Svantesson, & Zhang, 1986; Liu & Samuel, 2004). The primary acoustic cue for these tones is F0, with the two dimensions height and contour. Secondary cues include amplitude, duration, & voice quality (Lee, 2009; Liu & Samuel, 2004; Whalen & Xu, 1992). Unlike other tonal languages such as Cantonese (Peng et al., 2012), Taiwanese (Lin & Repp, 1989) or Thai (Abramson, 1978), the dominant feature for tonal contrasts in Mandarin is the distinctive pitch contours and there are no tone pairs that differ only in pitch height. Nevertheless, a previous study revealed partial effect of register on the perception of Mandarin tone segments (Whalen & Xu, 1992), and a tone perception study suggested that both pitch height and pitch contour have a significant effect on distinguishing the high level and mid rising tones (Massaro, Cohen, & Tseng, 1985). Thus, F0 manipulation in the current study is shifting the pitch height of Mandarin tones.

Experiment 1

Participants

Nine adult native speakers of Mandarin Chinese (4 men, 5 women) ranging from 23 to 32 years of age ($M = 26$, $SD = 3.8$) participated in all experimental conditions of Experiment 1. All subjects were born and raised in Northern China and reported no history of hearing, speech, language, or neurological disorders.

Procedure

Each subject completed a passage reading task without any auditory perturbations and one condition of monosyllabic word reading on the first day and then one additional condition of monosyllabic word reading per day for each of the following three days (i.e., a within-subjects design with sessions spread out over four days). Experimental conditions differed in the type of auditory feedback perturbation and/or the tone categories that were produced. The order of completing the four conditions was counter-balanced across subjects.

For the passage reading task, subjects read aloud a text of approximately 300 words at their habitual speaking rate and intensity. Each subject's acoustic speech output was recorded, and the data were used to calculate the standard deviation of F0 values. This variability index was obtained to explore a possible correlation between a speaker's F0 variability and the extent of auditory-motor adaptation during tone production with altered auditory feedback.

In the feedback perturbation conditions, subjects read aloud the monosyllabic word /ma/ with different lexical tones while the auditory feedback signal for some tones was manipulated such that F0 for those productions sounded unaltered, shifted up, or shifted down during the perturbation phase of the conditions. Specifically, three 40-min conditions and one 20-min condition were included: no F0 shift for any of the four tones (*Control Condition*); F0 down shift for the high level tone T1 and no shift for the mid rising tone T2, the low dipping tone T3, or the high falling tone T4 (*Condition 1*); F0 down shift for the high level T1, F0 up shift for the low dipping T3, and no shift for the mid rising T2 and the high falling T4 (*Condition 2*); the same F0 down shift for the high level T1 and F0 up shift for the low dipping T3 but in the absence of any productions of T2 or T4 (*Condition 3*).

The words with T1, T2, T3, and T4 were presented in Chinese scripts that mean “mother,” “hemp,” “horse,” and “scold,” respectively. Each trial started with a word displayed on a monitor for a period of 2 s, and subjects were to produce the word within this time window. The monitor turned blank for 1 s until the next trial started. The selection of only high level T1 and low dipping T3 tones as the targets for F0 feedback manipulation was based on considerations that the experiment would be too long for the subjects if all possible combinations of tone category and F0 shift direction were included. In addition, T1 and T3 essentially correspond to the upper and lower pitch boundaries, respectively, of the overall tone space. Thus, shifting auditory feedback for T1 and/or T3 toward the center of the tone space (i.e., a downward shift for T1 and/or an upward shift for T3) results in a more compact tone space with less distinction between these particular tones.

One cycle of producing each word of the training word set (presented in random order) is referred to here as an epoch (i.e., a block of 4 words in the Control Condition and in Conditions 1 and 2, or a block of 2 words in Condition 3). There were 192 epochs of training words for a given condition: 8 epochs in a baseline phase during which auditory feedback was unaltered, 24 epochs in a ramp phase during which the magnitude of a feedback perturbation increased gradually across trials, 144 epochs in a max-perturbation phase during which the perturbation was held at its maximum level, and 16 epochs in a post-perturbation phase during which normal auditory feedback was restored.

In addition to training words, test words were included to examine the generalization of aftereffects. To examine generalization across different words, syllables /fæp/ with all four tones were produced once in the baseline and once in the post-perturbation phase of the Control Condition and of Conditions 1 and 2. To examine generalization across tones, the syllable /ma/ produced with T2 and T4 served as the test words in Condition 3 (where T2 and T4 had not been produced as training words). Thus, two epochs of test words were added per condition. Each subject produced 776 trials in the Control Condition and in Conditions 1 and 2, and 388 trials in Condition 3 (see Figure 1). There was a 1-min break after every 192, 194, or 196 trials.

Auditory feedback manipulation

The set-up for the experiment is shown in Figure 2. Subjects' speech was transduced by a microphone (Beta-53, Shure) and amplified before being routed through a signal processor (VoiceOne, TC Helicon). The signal processor altered F0 in real-time (approximately 10 ms delay) under control of a notebook computer. F0 was raised or lowered in steps of 4 cents during the ramp-perturbation phase, and remained at a constant level of +100 or -100 cents (i.e., +1 or -1 semitone) during the max-perturbation phase. The conversion formula from Hz to cent is:

$$c = 1200 \times \log_2 (f/f_r) \quad (1)$$

where c and f are frequency variables in cents and Hz, respectively, and f_r is the reference frequency in Hz. Output of the signal processor was routed to a headphone amplifier (S-phone, Samson), and fed back to the subject via insert earphones (ER-3A, Etymotic Research). The subject's speech intensity was monitored, and color-coded feedback on the computer screen prompted the subject to maintain an intensity in the range from 72 to 78 dB SPL at the microphone that was positioned 15 cm from the mouth (based on ear-level recordings of the feedback intensity during speech production, the entire system was calibrated for a 75 dB SPL input at the microphone to result in 72 dB SPL in the earphones; see Cornelisse, Gagné, & Seewald, 1991). Subjects' produced speech and the auditory feedback signal were recorded on notebook computers at a sampling rate of 44.1 kHz.

Data processing and analysis

All acoustic recordings were down-sampled to 10 kHz, and the fundamental frequency was estimated in the PRAAT software (Boersma & Weenink, 2012) using an analysis window

step of 10 ms and a frequency range from 75 to 500 Hz. For each subject's passage reading data, the mean F0 was calculated as a reference frequency and all F0 values from the reading task were normalized (in cents) relative to this mean F0. The standard deviation (SD) of the normalized F0 was computed to represent the typical F0 variability for each subject.

For the monosyllabic word task, each production's vowel onset and offset (defined as the initial and final oscillations, respectively, in the waveform) were manually labeled, and F0 samples within this time interval were treated as the pitch contour of the tone associated with the trial (Howie, 1974). Following conventional procedures for measuring Mandarin tones (Barry & Blamey, 2004; Wong, 2012; Xu, 2001; Xu & Wang, 2001; Zhou & Xu, 2008), one pitch target was extracted for level tones whereas two targets were extracted for dynamic tones. Specifically, the F0 mean was determined for T1 (T1m), the F0 minimum was determined for T3 (T3n), and both the F0 maximum and minimum were determined for T2 (T2x, T2n) and T4 (T4x, T4n) (see Figure 3). F0 contour smoothing algorithms in the PRAAT software prevented the occurrence of extreme F0 estimations (e.g., half or double F0 jumps). When creaks occurred, especially for T3, F0 estimates by PRAAT were missing or unreliable due to the irregularly spaced pitch period in creaky portions of the acoustic signal. In such cases, only the segments with reliable F0 values were used in the pitch target extraction for each trial (examples shown in the Results section). For a given condition, all pitch targets were normalized in cents relative to the mean of the corresponding pitch targets from the baseline phase.

Significant changes in F0 during a given condition's max-perturbation phase as compared with baseline would indicate adaptation, at least in the absence of similar changes in the Control Condition. Thus, given that the purpose of the study was to determine for which tones in which conditions adaptation did or did not occur (rather than comparing F0 values across multiple conditions), paired *t*-tests of max-perturbation vs. baseline were chosen for the statistical analyses. Thus, for each subject, the mean pitch target values during the max-perturbation phase were calculated to quantify adaptation extent for each tone. Using the data from all subjects, paired *t*-tests were then conducted to compare those values with baseline (effectively one-sample *t*-tests because the baseline mean was always zero). However, when statistically significant changes in a certain pitch target also occurred in the Control Condition (i.e., thus when subjects adjusted their F0 even in the absence of an auditory perturbation), it was necessary to carry out statistical tests directly comparing the relevant experimental conditions with that Control Condition. Hence, paired *t*-tests comparing mean F0 during the max-perturbation phase in Conditions 1, 2, and 3 with the Control Condition were conducted for targets that showed significant changes in the Control Condition. Lastly, paired *t*-tests (experimental vs. control) were also applied to the training words and test words in the post-perturbation phase to examine aftereffects (i.e., continuing compensation after unaltered auditory feedback is restored) and generalization (i.e., transfer of adaptation from training words to test words), respectively. The statistical significance level was set at 0.05. Besides *t* and *p* values, Cohen's *d* was calculated as an index of effect size.

Results

1. Baseline and control condition (unaltered auditory feedback)—Representative F0 contours of the tones produced with normal auditory feedback in the monosyllabic word reading task are shown in Figure 3. The contour of each tone in Figure 3 was obtained, separately for female and male subjects, by averaging across all non-creaky productions during the baseline phase in all four conditions and across speakers. Averaging was accomplished after first temporally stretching or compressing (via upsampling or downsampling, respectively) each contour to match the mean duration for its tone category. The vertical dashed lines mark the boundary between /m/ and /a/. The /a/ portion of the F0 track primarily carries the tone contour. The contour patterns for T1, T2, T3, and T4 are consistent with their conventional characterization—high and steady, rising from mid level, low dipping, high falling, respectively. In particular, both T1 and T4 start with high frequencies, but T1 remains flat in the high frequencies whereas T4 falls into the low frequencies. T2 and T3 are mid-low-onset tones and resemble one another in that both show a succession of falling and rising frequencies. The apparent distinctions between these two tones include a smaller degree of F0 drop, an earlier “reversal” point, and a higher F0 offset for T2 than for T3. On average, T3 is characterized by the longest duration whereas T4 is the shortest in duration. There are no major sex differences in pitch pattern except that the mean F0 of the tone space is ~210 Hz for females and ~110 Hz for males and that all four tones are longer in duration for females than for males. Additionally, the overall pitch range, delimited by T4x and T3n, was approximately 7 semitones for the female speakers and 8 semitones for the male speakers.

Only a minority of T3 productions from the baseline phase showed this tone’s canonical form (a continuous contour that is first falling and then rising, typically referred to as dipping) whereas the remaining baseline T3 trials were produced with a creaky voice quality. Creaky voice quality in T3 production has been reported in previous studies (Belotel-Grenié & Grenié, 2004; Chao, 1956; Liu & Samuel, 2004). The waveforms and spectrograms of two creaky T3 production trials of syllable /ma/ are shown in Figure 4. The nonmodal phonation is obvious in the trial on the left produced by a male subject, and a complete loss of voicing is noticeable in the trial on the right produced by a female subject. As represented by the yellow lines imposed on the spectrograms, the pitch contours were interrupted near the middle of the productions due to an irregular pitch period or absence of voicing. A local minimum within the middle portion, usually close to either the creak onset or offset, was marked as the pitch target for a creaky T3 trial (shown with the red arrows in Figure 4).

Statistical results for all six pitch targets in the Control Condition are listed in the upper left of Table 1. The two pitch targets of T2 (T2x, T2n), the pitch target of T3 (T3n), and the two pitch targets of T4 (T4x, T4n) did not deviate significantly in the max-perturbation vs. baseline phase for this Control Condition without auditory perturbation. However, the mean pitch target for T1 (T1m) did show a statistically significant deviation from its baseline value ($t=-3.319$, $p=0.011$, $d=1.1$). In particular, T1m decreased gradually across trials throughout the Control Condition (see grey solid line CCT1m in Figure 5). A similar change in T4x, albeit statistically nonsignificant, was apparent with a medium effect size ($d=0.7$) (see grey

dashed line CCT4x in Figure 5). Relative to the baseline phase, the mean F0 for T1m and T4x in the max-perturbation phase decreased by 73 cents and 54 cents, respectively. Pearson's linear correlation analysis for mean T1m and mean T4x during the max-perturbation phase indicated a strong correlation between these two pitch targets ($r=0.855$, $p=0.003$). These results indicate that, in this condition without any auditory perturbation, the “upper limit” of the F0 range—represented by the mean of T1 and the onset of T4—gradually drifted down across trials whereas there was no drift for the “lower limit” of the F0 range.

2. Single perturbation—For Condition 1, in which F0 in the auditory feedback was shifted down only for the high flat tone T1, the statistical results listed in the upper right part of Table 1 indicate that none of the pitch targets showed significant changes during the max-perturbation phase as compared with the baseline. As described above, however, high level T1m decreased significantly during the Control Condition in which no feedback alteration was applied. Hence, for T1, adaptation to a down-shift F0 perturbation could consist of resisting this typical decrease in F0 that occurs over time even with unaltered feedback. Therefore, paired t -tests comparing the max-perturbation phases of each of the Conditions 1, 2, 3 versus the Control Condition were needed to determine possible adaptation in T1m. Results from these tests are shown in Table 2.

The difference between T1m in Condition 1 with down-shifted pitch feedback vs. the Control Condition with unaltered feedback was statistically significant ($t=2.364$, $p=0.046$, $d=0.8$). This difference is illustrated in Figure 5 (black solid line C1T1m vs. gray solid line CCT1m), and suggests that subjects did adapt to the perturbation by changing F0 for T1 productions in the opposite direction of the perturbation when they received normal auditory feedback for the other three tones.

The onset F0 of the high-falling T4 tone (T4x) showed a nonsignificant but similar pattern, even though the auditory feedback for these T4 productions had not been altered at all. The mean changes in F0 for T1m and T4x averaged across subjects was 75 cents and 52 cents, respectively. In fact, there was a strong and statistically significant positive correlation ($r=0.887$, $p=0.001$) between changes in T1m and T4x when both were calculated as an F0 difference between Condition 1 and the Control Condition. This statistically significant correlation suggests that adaptive learning for the high T1 in response to altered feedback for this tone *generalized* to productions of T4 which also starts with a high F0.

For T1m and T4x, aftereffects (i.e., continuing compensation after normal feedback had been restored in the post-perturbation phase) was examined with paired t -tests again comparing Condition 1 with the Control Condition (given the above discussed F0 changes in the Control Condition, even without any auditory perturbation). Results indicate that there were statistically significant aftereffects for both T1m ($t=4.063$, $p=0.004$, $d=1.4$) and T4x ($t=4.141$, $p=0.003$, $d=1.4$). Moreover, paired t -tests completed on the data from the test words (rather than training words) indicate that adaptation to the F0-down perturbation that had been applied only to the T1 version of the syllable /ma/ also generalized to T1 (T1m: $t=2.373$, $p=0.045$, $d=0.8$) and T4 versions (T4x: $t=2.925$, $p=0.019$, $d=1.0$) of the syllable /

fa₀/). Lastly, there were no statistically significant correlations between F0 variability (SD) in the passage reading task and adaptation extent of the T1m pitch target.

3. Two different perturbations—Similar to the situation for Condition 1, the results from max-perturbation vs. baseline *t*-tests for T2x, T2n, T3n, T4x, T4n (Table 1) and, when necessary, experimental conditions vs. Control Condition *t*-tests for T1m (Table 2) were used to examine possible F0 adjustments during Conditions 2 and 3 in which auditory feedback was manipulated for both high level T1 (F0 shifted down) and low dipping T3 (F0 shifted up). In this case, however, T1m – which changed in a statistically significant manner even during the Control Condition – changed in a similar manner during Conditions 2 and 3 (i.e., there were no statistically significant differences for Conditions 2 and 3 vs. the Control Condition). Thus, only statistically significant changes within Conditions 2 and 3 (Table 1) indicate adaptation.

The changes in T3n ($t=-3.207$, $p=0.012$, $d=1.1$) and T4n ($t=-3.504$, $p=0.008$, $d=1.2$) observed during max-perturbation vs. baseline in Condition 2 (T1 shifted down, T3 shifted up) reached statistical significance. As shown in Figure 6, subjects decreased T3n during the max-perturbation phase relative to baseline (mean adaptation extent 82 cents) as well as relative to the corresponding phase in the Control Condition. A similar decrease in F0 was also observed in T4n (mean adaptation extent 55 cents). However, the lack of a statistically significant correlation between the adaptation extents of T3 and T4 implies that the generalization from T3 to T4 was not tightly coupled.

Neither T3n nor T4n of the training words or testing words in Condition 2 showed a significant difference between the post-perturbation phase and the baseline phase—hence, no aftereffects or generalization were observed. In addition, neither of these two measures showed a strong correlation between adaptation extent and F0 variability (SD) during the passage reading task. That is, the pitch range of continuous speech was not related to how much the speaker adapted in T3 or T4 production.

Interestingly, in the absence of interspersed T2 and T4 productions (Condition 3), subjects did not compensate for the perturbations in either T1 or T3.

Discussion

Pitch contours produced with normal auditory feedback—The overall results for pitch patterns of tone production with unaltered auditory feedback are in line with those from previous investigations in terms of pitch contour, height, and sex differences (Chao, 1956; Xu et al., 2004). However, to our knowledge, there have been no prior reports of downward drift in high flat tones or in the onset of falling tones throughout the course of repeated production of all four lexical tones in citation form. Given that our subjects produced a much greater number of trials (768 per condition) than in previous studies, one might speculate that vocal fatigue or decreased attention may have played a contributing role in the observation of such gradual F0 changes over time in the control condition with unaltered auditory feedback. Unfortunately, our present data do not allow any strong conclusions to be drawn about these potential explanations, and further research is warranted

to identify the actual reasons for speakers' changes in F0 over the time course of our Control condition with unaltered auditory feedback.

Adaptation to a single perturbation—Previous studies have reported adaptation to F0 shifts in the auditory feedback when vocal pitch serves no phonological purpose in productions by English speakers (Jones & Munhall, 2000) or when it is used to signal morphemes in productions by Mandarin speakers (Jones & Munhall, 2002, 2005). The present results are consistent with those previous findings, and further indicate that a perturbation applied to a single tone category is compensated for in a linguistically contrastive context where the other three tone categories are produced with normal auditory feedback.

Nevertheless, some discrepancies in aftereffects and generalization exist between our own results and those from prior studies. Jones and Munhall (2002) found that Mandarin speakers lowered F0 of sustained high level tones (relative to the F0 of the final few trials in the training phase) after the auditory feedback manipulation was suddenly removed in a shift-down condition, similar to the aftereffects observed in sustained vowel production by English speakers (Jones & Munhall, 2000). However, our finding did not show such an F0 drop in the same tone during the post-perturbation phase. After feedback was returned to normal in our single-perturbation condition (i.e., shift-down on high level tones), speakers held F0 at approximately the same level as during the baseline phase and throughout the perturbation phases—essentially raising F0 relative to the corresponding phases in the control condition. In other words, aftereffects in the current study show a persistent compensation, indicating that sensorimotor learning occurred after a short-period of exposure to the perturbations. The discrepancy in aftereffects between these two studies might result from the methodological difference related to the number of produced tone categories (a single tone vs. all four tone categories) as well as the manner in which these tones were produced (sustained tones vs. word production).

Jones and Munhall (2005) observed that adaptation for the high level tone (T1) in response to an upward shift of F0 generalized to the mid rising tone (T2). Similar to the aftereffect for high level tones, their Mandarin speakers increased the F0 of rising tones after feedback was returned to normal. In the present study, adaptation for the high level tone (the training syllable produced with altered feedback) transferred to the same tone in the test syllable (not produced with altered feedback), indicating that auditory-motor learning occurred for a tone *category* rather than for a specific word. Moreover, this learning also generalized to the onset of high falling tones (T4) in both training and test syllables (substantiated by the positive correlation in F0 change extent between T1 and T4 and by the aftereffects), but adaptive changes were not seen in rising (T2) or dipping (T3) tones. Again, this discrepancy regarding generalization to T2 between Jones and Munhall's (2005) studies and our own data might be related to the different training sets (one tone vs. four tones) or production mode (prolonged vs. typical production in words). At any rate, our findings suggest that sensorimotor learning is not restricted to the perturbed tone only, and also affects other tones that are closely related in terms of laryngeal configuration at phonation onset.

Adaptation to two different perturbations—While hearing their T2 and T4 productions with normal auditory feedback, subjects compensated for an F0 upward shift in the auditory feedback for T3 by lowering F0 in the T3 productions, and this was accompanied by a similar F0 decrease in the offsets of T4 productions. On the other hand, the simultaneous F0 downward shift perturbation in the auditory feedback for T1 productions remained uncompensated. Together with the absence of any aftereffects, these findings suggest that it may be difficult for Mandarin speakers to simultaneously adapt to two different F0 shifts that are applied to high level and low dipping tones.

Our finding that Mandarin speakers failed to simultaneously compensate for two different perturbations is at odds with Keough and Jones's (2011) results based on sustained musical notes produced by trained singers, despite the fact that the stimuli in both studies served as contextual cues implicitly signaling the direction of the feedback perturbation. When two different perturbations were applied in our study, only the upward F0 shift applied to low dipping tones elicited compensation in Mandarin speakers whereas simultaneous compensation for two perturbations was observed in trained singers by Keough and Jones (2011). On the other hand, there was generalization to an unaltered tone category (here falling tones) in our study, but no generalization occurred to the unaltered music note in the study with trained singers. One explanation for these discrepant findings may lie in the fundamental difference between the two vocalization tasks: reproducing target musical notes when the targets are presented acoustically (Keough & Jones, 2011) versus producing lexical tones when the targets are presented visually as words with lexical contrasts. It is also possible, however, that the singers' professional training resulted in finer control over the laryngeal motor system and/or greater acuity for tone perception.

Interestingly, one possible reason for the observation that, in Condition 2, subjects compensated for the F0 up-shift on T3 but not the F0 down-shift on T1 may be that the T3 up-shift (if left uncompensated) could reduce the contrast between T2 and T3 and lead to confusion of these two tones. Thus, the observed motor behavior may have a foundation in the perceptual domain and the maintaining of linguistically-relevant tonal contrasts. Previous work has indeed shown that T2 and T3 are the most confusable tone pair (Li & Thompson, 1977; Liu & Samuel, 2004). This hypothesis is supported by the observation that adaptation to the T3 up-shift was eliminated when no T2 (or T4) tones were produced (Condition 3). We therefore conducted a separate tone perception experiment (Experiment 2) to test the hypothesis that adaptation to tone perturbations is partially driven by the need to maintain tonal contrasts.

Experiment 2

Participants

Of the group of nine speakers who participated in the tone production experiment (i.e., Experiment 1), two female and three male subjects (hereafter the *speaker-listener group*) were available to return to the laboratory four months later for a tone identification experiment (i.e., Experiment 2). The purpose of this second experiment was to determine whether the F0 adaptation observed in Experiment 1 served primarily to maintain appropriate contrasts within the entire tone system. In addition, we aimed to investigate

whether the perception of such tone contrasts differs between the speakers who had produced the tones (and, thus, had been exposed to pitch-shifted auditory feedback and their own adaptive responses) and random listeners (who had not previously been exposed to the speakers, to pitch-shifted feedback, or to the adaptive responses). Therefore, another group of five women and four men without any prior involvement in Experiment 1 (hereafter the *listener-only group*) also participated in the tone identification experiment. These subjects ranged from 22 to 30 years of age ($M = 27$, $SD = 3.1$), had a Northern China dialect background, and reported no history of any speech or hearing problems.

Stimuli

All acoustic stimuli for the identification experiment were derived from a subset of the monosyllabic word production trials collected in Experiment 1. Specifically, the pool of productions for stimulus generation included all 9 speakers' baseline trials of all four conditions (i.e., Control Condition, Condition 1, Condition 2, and Condition 3) and the max-perturbation trials of Condition 2 in which subjects had shown adaptation in the form of lowered F0 values for T3 when F0 in the auditory feedback for this tone was shifted up. Each individual stimulus was created prior to the experiment by routing the acoustic signal of a recorded production through a set-up similar to the one used for Experiment 1. In particular, a notebook computer played back the production, and this acoustic signal was amplified and routed to the signal processor (VoiceOne) that either manipulated F0 or kept F0 intact depending on the tone category of the production and the type of target stimulus to generate. The VoiceOne output was sent to a headphones amplifier and recorded on another notebook computer. Using the productions of each speaker, three types of stimuli were generated as follows.

Type 1 stimuli were prepared to represent unaltered auditory feedback signals as perceived during normal tone production. Thus, productions from the baseline phase of all four conditions in Experiment 1 were used as the source signal for playback and the signal processor did not implement any alterations. For a given speaker and a given tone, 32 productions (8 baseline trials $\times$ 4 conditions) were played back through the equipment and recorded as stimuli for the perception task.

Type 2 stimuli simulated what a speaker in the production experiment would have heard if he or she had shown *no adaptation* in tone production when two different F0 perturbations were simultaneously applied to tones 1 and 3 while either no perturbations were applied to tones 2 and 4 (as in Condition 2) or tones 2 and 4 were not produced (as in Condition 3). In other words, this stimulus type represented—when assuming an absence of adaptation—the auditory feedback signals when F0 was altered for T1 and T3 but unaltered for T2 and T4. The productions from the baseline phase of all four conditions were used again as the playback source signals, but the signal processor shifted F0 down by 100 cents for T1 and up by 100 cents for T3 while leaving F0 unaltered for T2 and T4. There were again 32 stimuli for each speaker and each tone.

Type 3 stimuli corresponded to the auditory feedback signal perceived when a speaker *did show adaptive changes* in tone production in response to the two different pitch-shift perturbations that had been applied to T1 and T3. All actual productions from the max-

perturbation phase of Condition 2 were played back and the signal processor manipulated the F0 in these adapted productions for each tone category in the same manner as described above for Type 2 stimuli. For a given speaker and a given tone, 144 stimuli of Type 3 were created².

Procedure

In this tone identification experiment, all subjects completed a four-alternative-forced-choice task. Each subject in the speaker-listener group was presented with the stimuli derived from his or her own productions, whereas each subject in the listener-only group was presented with the stimuli generated from the recordings of a sex-matched speaker. An experimental session involved three blocks of trials with two half-minute inter-block breaks. Each block consisted of 200 stimuli (50 stimuli for each tone category) and was designated Type 1, 2, or 3 depending on whether the stimuli within the block were chosen from the aforementioned Type 1, 2, or 3 stimuli, respectively. A Type 1 or 2 block was created by selecting items randomly (at least once but no more than twice) from Type 1 or 2 stimuli (for a total of 50 stimuli from a pool of 32 different items for each tone category), whereas a Type 3 block was created by randomly selecting items once from Type 3 stimuli (50 stimuli out of a pool of 144 items for each tone category). Each subject listened to three blocks (once for each block type, 600 trials in total), and the order of block type and trials within blocks was randomized.

Each trial started with the presentation of an acoustic stimulus via insert earphones at 72 dB SPL at both ears and the display of four Chinese scripts (the same scripts as in the production experiment) on a computer monitor. The subject was asked to indicate the perceived tone category by clicking with a computer mouse on the corresponding Chinese script. The subject was allowed to take as much time as needed before clicking on a confirm button to proceed to the next trial. No information was provided about who produced the words that were used to create the stimuli.

Data processing and analysis

To avoid response bias in our measure of tone identification, *d-prime* (d') was calculated as a sensitivity index for each tone category within a block and for each subject separately. Specifically, d' was defined as follows:

$$d' = z(h) - z(f) \quad (2)$$

where z represents z-score function, h hit rate, and f false alarm rate. A greater d' indicates that a subject is more accurate in identifying the tone category. For each tone category within a block of trials completed by a given subject, the hit rate was computed as the

²Type 1 and 3 stimuli both simulated auditory feedback signals actually perceived in Experiment 1. For two reasons, we chose here in Experiment 2 to play back and alter the speakers' productions rather than playing back recordings of the feedback signals themselves. First, our set of recordings of the actual auditory feedback signals from Experiment 1 was incomplete due to some technical issues with this aspect of the experiment. Second, Type 2 stimuli were not perceived as feedback signals in any conditions of Experiment 1. Thus, to ensure consistency across the three types of stimuli, all stimuli were generated by playing recordings of the original production through the same equipment circuitry.

number of correctly identified trials divided by the number of all trials for that tone (i.e., 50), and the false alarm rate was calculated as the number of trials incorrectly identified as the tone divided by the total number of trials for the other three tone categories (i.e., 150). Note that all extreme values of hit rate or false alarm rate were adjusted to prevent infinite z -scores—in particular, zeros were replaced by 0.001 and ones were replaced by 0.999.

Separate repeated measures analyses of variance (ANOVA) were conducted for the speaker-listener group and the listener-only group. In each analysis, d' was the dependent variable, and tone category and stimulus type were within-subjects variables. Degrees of freedom (df) were corrected based on Huynh-Feldt epsilon for any violations of the sphericity assumption (Max & Onghena, 1999). Additionally, *post hoc* paired t -tests were conducted to determine whether there were any significant differences in d' between Type 1 and Type 2 stimuli or between Type 1 and Type 3 stimuli. The significance level was set to 0.05 for all tests.

Results

1. Variation in identification performance across tones—The ANOVA results revealed a statistically significant main effect of tone category both for the speaker-listener group ($F(2.833, 11.331) = 3.926, p = 0.040$) and for the listener-only group ($F(2.877, 23.015) = 3.585, p = 0.031$). Thus, tone identification performance varied among the four lexical tones regardless of whether the listener perceived his or her own productions or someone else's productions. Indicated by the d' grand average across stimulus types and groups, T4 was the most sensitive tone ($d' = 6.017$) followed by T1 ($d' = 5.778$), with tones 2 and 3 having the lowest sensitivity ($d' = 5.308$ and $d' = 5.390$, respectively).

2. Identification of unaltered baseline productions (Type 1 stimuli)

Identification results for the unaltered tones from baseline productions of the listener-only group and the speaker-listener group are listed as confusion matrices on the left side of Tables 3 and 4, respectively. Each confusion matrix shows the correct identification percentages (elements on the diagonal) and confusion percentages (elements off the diagonal) across all subjects in the respective group.

The listener-only group identified all four tone categories of unaltered baseline productions with high accuracy (>97%). The most noticeable confusion in the matrix is that 2.9% of T2 stimuli were perceived as T1. The majority of these misidentifications occurred on the stimuli derived from one speaker whose F0 contour was characterized as high and flat in shape with a brief rising close to the offset. Other misidentifications (1.3%) affected tones 2 and 3. The confusion is asymmetric, with T3 more frequently misidentified as T2 than *vice versa*.

The speaker-listener group identified their own four tone categories with even greater accuracy. Specifically, 99.6% of T1, T2, and T3 were perceived correctly, and 100% of T4 were perceived correctly. Many confusion errors were even smaller than in the listener-only group.

3. Identification of baseline productions with T1 and T3 pitch shift (Type 2 stimuli)

—The confusion matrices and d' means for the identification of Type 2 stimuli

(representing feedback that would have been heard in Experiment 1 if no adaptation had occurred) are included in the middle of Table 3 for the listener-only group and in the middle of Table 4 for the speaker-listener group. Identification accuracy remained generally high, but both groups of listeners showed a descriptive increase in the number of instances in which T3 was perceived as T2 as compared with the above discussed results for Type 1 stimuli. Although this main effect of stimulus type was found to be statistically nonsignificant ($p = .083$), we further examined the observed discrepancies between the confusions for baseline productions with a downward pitch shift applied to tone 1 and an upward pitch shift applied to tone 3 (Type 2 stimuli) vs. those for unaltered baseline productions (Type 1 stimuli).

For the listener-only group, *post hoc* paired *t*-tests confirmed that there were no significant differences in d' between Type 1 and Type 2 stimuli for any tone category. For the speaker-listener group, however, the results of some *post hoc* paired *t*-tests indicated statistically significant effects of stimulus type (upper half of Table 5). In particular, the difference in d' between Type 2 and Type 1 stimuli was statistically significant for tone 2 ($t = 3.505$, $p = 0.025$) and tone 3 ($t = 3.735$, $p = 0.020$). As can be seen in the bottom row of Table 4, the speaker-listener group's d' group means for Type 1 stimuli vs. Type 2 stimuli were 5.727 vs. 5.145 for T2, and 5.973 vs. 4.984 for T3. In other words, identification performance for T2 and T3 was degraded for the type of stimuli (Type 2) that included unaltered T2 but F0-up-shifted T3 as compared with stimuli (Type 1) in which both T2 and T3 productions were unaltered. Considering stimuli identified as T2, the correct identification rate within the Type 1 and Type 2 stimulus types was 99.6% in both cases. However, errors perceiving T3 as T2 and perceiving T1 as T2 occurred more frequently within the Type 2 stimuli than within the Type 1 stimuli (3.2% vs. 0.4% and 1.2% vs. 0.4%, respectively). In addition, the correct identification of T3 decreased from 99.6% for Type 1 stimuli to 96.8% for Type 2 stimuli even though no other tones were misidentified as T3.

4. Identification of pitch-shifted adapted productions (Type 3 stimuli)—When subjects from either the speaker-listener group or the listener-only group judged adapted productions with F0-shifts applied (i.e., the feedback that subjects in Experiment 1 actually heard), identification performance did not differ significantly from that observed when judging unaltered baseline productions. The statistical results of paired *t*-tests (Type 1 vs. Type 3 stimuli) for the speaker-listener group are listed in the bottom half of Table 5. No statistically significant difference was found for any tone category. The confusion matrix on the right side of Table 4 shows the same group's performance for identifying the manipulated adapted tones. Tones 1 and 4 were identified perfectly, and the asymmetric confusion between tones 2 and 3 remained as the main errors. As shown in the confusion matrix on the right side of Table 3, the performance of the listener-only group showed a similar pattern, except for the error of T2 being mislabeled as T1 due to the idiosyncrasy of one speaker's productions.

Discussion

Perception of unaltered and altered adapted tones—The results of the tone perception experiment show that subjects' performance was comparable in perceiving

unaltered tones and altered adapted tones, and that identification accuracy varied across tone categories. The more accurate perception for tones 1 and 4 vs. tones 2 and 3 is consistent with observations from previous tone perception studies. For example, it has been shown that the performance of identifying synthetic tones was nearly perfect when the F0 range of the tone space (i.e., span from the peak of the falling tone to the valley of the dipping tone) was set to 80 Hz, and that the overall identification rate remained high—but with increasing confusion between T2 and T3—when the F0 range was reduced to 2 Hz (unpublished work by Zue reported in Klatt, 1973).

It has been well documented that T2 and T3 are the most confusable tone pair in Mandarin (Chuang, Hiki, Sone, & Nimura, 1971; Liu & Samuel, 2004; Shen & Lin, 1991). Our finding that the confusion is asymmetric, with T3 more frequently misidentified as T2 than *vice versa*, is also in keeping with those previous studies (Chuang et al., 1971; Liu & Samuel, 2004). Although Shen and Lin (1991) found a reverse directional confusion (i.e., more T2 mislabeled as T3 than T3 mislabeled as T2) across 10 different syllables, the pair of tones 2 and 3 associated with the syllable /ma/ (i.e., the same syllable as used in the present study) actually showed more errors of T3 being misperceived as T2. Furthermore, it has been shown that native Mandarin-speaking children acquire tones 2 and 3 later and with more difficulty (in particular substituting one of these tones for the other tone; Li and Thompson, 1977) and that T2 and T3 are the tone pair with the greatest confusability in both perception and production for second language learners (Kiriloff, 1969; Wang, Jongman, & Sereno, 2003).

Perception of altered non-adapted tones—Results of the tone identification experiment also indicate that the F0 manipulations that were applied to T1 and T3 selectively influenced subjects' perception of their tone system, in particular among the subjects who had produced the original stimuli. When subjects judged F0 down-shifted T1 stimuli that were modified versions of their own prior productions, identification performance did not differ as compared with judging the original, unaltered T1 productions. Thus, it appears that high level tones are allowed to vary in F0 within a certain range (with a tolerance of at least 100 cents below the typical value) without losing their own perceptual identity or interfering with the perception of the other three tone categories. On the other hand, with regard to judging F0 up-shifted T3 stimuli that were modified versions of their own prior productions, performance deteriorated significantly: there was an increase in the number of errors whereby the manipulated T3 was misidentified as T2.

It is noteworthy that the manipulations of tones 1 and 3—which pushed these two tones toward the center of the tone space and, thus, closer together—did not ever result in confusion between these two tone categories. Possible explanations are that tones 1 and 3 have distinctive contours or that the distance in F0 height between these tones is much greater than the F0 manipulation magnitude. Lastly, T4 retained the greatest sensitivity index in the presence of F0 manipulations imposed on two of the other three tone categories. This might be due to T4's dynamic pitch contour—a sweep from the high frequencies to the low frequencies—which may remain sufficiently distinct from the contours of all other tone categories regardless of whether or not the F0 of those other tones is shifted up or down. In

fact, previous work has suggested that the prevalence of falling tones in tonal languages might be due to physiological factors as well as their perceptual saliency (Hombert, 1975).

Difference in perceiving tones produced by oneself and others—The observation that the speaker-listener group performed better than the listener-only group, particularly when perceiving unaltered and altered adapted tones, suggests that identifying one's self-produced tones is easier than identifying someone else's tones. Perceiving tones requires an estimation of the pitch range within which each tone may be located, a process referred to as speaker normalization (Moore & Jongman, 1997; Peng et al., 2012). This estimation appears less accurate for tones produced by another speaker vs. self-produced tones even without explicit information regarding the speaker's identity. However, experimentally manipulating pitch height for high level tones and low dipping tones might degrade the speaker normalization progress to a greater extent for the subjects who produced the original stimuli. If so, this may in turn reduce their perception performance to a level comparable to that of listeners who did not produce the stimuli.

General discussion

Variation and interconnection in the Mandarin tone system

Variation patterns within a tone system can be described based on the start, end, and inflection points of F0 landmarks (Barry & Blamey, 2004; Wong, 2012; Xu, 2001; Xu & Wang, 2001; Zhou & Xu, 2008). One form of Mandarin tone F0 variation observed in the present work is that both the overall height of high level tones T1 and the onset of high falling tones T4 drifted downward over many repeated trials when produced with unaltered auditory feedback. By the end of the 40-minute experimental sessions, both these pitch targets had dropped by nearly 1 semitone, and the onset of falling tones approximated the offset of mid rising tones. This lowering of F0 decreased the upper limit of the pitch range and, as a result, also compressed the overall pitch range as there was no downward drift in the lower limit (corresponding to the turning point of dipping tones and the offset of falling tones).

In addition to such F0 variation in tone categories, the present study also revealed interdependence among some of the Mandarin tones. The finding that T1 adaptation generalized to T4 adjustments with the same direction and similar magnitude suggests that these two tones with high onset are linked to a certain degree. Covariation in the onset F0 of these two tones is consistent with previous investigations that found, at tone onset, similarities in voice quality measures (e.g., spectral tilt, F1 bandwidth, amplitude difference between the first two harmonics or H1-H2) that are also closely related to glottal configuration (Keating & Esposito, 2007; Lee, 2009).

Another observation of interdependence in the present data relates to the inflection point of dipping tones and the offset of high falling tones. The initial portion (roughly 50%) of a dipping tone is characterized by a falling pitch contour with similar slope and time course as seen in a high falling tone even though the actual F0 level is different. The fact that creaks occur frequently at dipping tones' inflection point and occasionally at falling tones' offset (Belotel-Grenié & Grenié, 2004; Davison, 1991; Liu & Samuel, 2004) suggests that

speakers produce the pitch targets for these two tone categories in a similar manner (e.g., in nonmodal mode) although some studies have reported that creaks are not critical perceptual cues (Gårding et al., 1986). It is possible that these interconnections among tone categories reflect shared laryngeal mechanisms and configurations such that adjustments in one tone category also result in changes in other linked tones.

Adaptation in tone production serves to maintain tonal contrasts

Findings from the tone adaptation experiment revealed not only that Mandarin speakers adjust pitch production in response to auditory feedback perturbations, but the pattern of results across the different conditions also suggested that such adaptation is influenced by the high-level linguistic role of tone contrasts. Perceptual data from the subsequent tone identification experiment provided further support for this interpretation, at least if one accepts that perceiving played-back productions shares underlying processes with monitoring one's own ongoing speech through auditory feedback. Combined, the production and perception results suggest that F0 adaptation in Mandarin serves to avoid confusability and to maintain spectral contrast, an interpretation that has also been offered for articulatory adaptive changes in formant shift studies (Houde & Jordan, 2002; Mitsuya et al., 2011, 2013).

When a single auditory feedback perturbation (F0 down-shift) was applied during the production of high level tones in the adaptation experiment, the vocalization control system was able to form or update the required internal model to compensate for the perturbation in order to achieve typical pitch heights. On the other hand, when two different perturbations (F0 down-shift on T1 and F0 up-shift on T3) were applied to the auditory feedback signal, the control system apparently did not form the necessary internal models to adapt to both perturbations simultaneously. Again based on the perception data, it seems that when speakers produced only T1 and T3 and no other tones, the resulting variation in F0 was tolerated because it did not cause confusion between the two produced tones. The perceptual data also suggest, however, that when the two different perturbations (F0 down-shift on T1 and F0 up-shift on T3) were applied when speakers produced these two affected tones interspersed with rising (T2) and falling (T4) tones, not adapting would have resulted in (a) the F0 shift-up on dipping tones causing confusion with rising tones but (b) the F0 down-shift on high level tones not causing a significant increase in confusion with any other tones. With the control system unable to adapt to both perturbations simultaneously, it adapted only to the perturbation that would have yielded significantly more confusions. Thus, F0 in productions of dipping tones was lowered to compensate for the feedback alteration.

Taken together, the results from Experiment 1 and Experiment 2 strongly suggest that adjustments in vocalization in response to pitch shift perturbations are both goal-driven and corpus-dependent. In tasks such as producing a single vowel or tone, particularly in sustained phonation where pitch does not serve a linguistic purpose, speakers may adapt to pitch-shifted auditory feedback merely to achieve the sensory consequences (here perceived pitch) that are typically associated with the specific motor actions (Jones & Munhall, 2000, 2002, 2005). However, in tasks where speakers produce two or more tone categories and pitch serves a critical role in signaling lexical contrasts, adaptation to pitch-shifted auditory

feedback affects the entire tone system in such a manner that pitch contrasts are maintained and perceptual confusion among tones is avoided.

Acknowledgments

Funding

This work was supported by the Institute of Acoustics, Chinese Academy of Sciences [grant Y154221431] and the National Institute on Deafness and Other Communication Disorders [grants R01DC007603 and R01DC014510].

References

- Abramson AS. Static and dynamic acoustics in distinctive tones. *Language and Speech*. 1978; 21:319–325. DOI: 10.1177/002383097802100406 [PubMed: 750791]
- Barry JG, Blamey PJ. The acoustic analysis of tone differentiation as a means for assessing tone production in speakers of Cantonese. *Journal of the Acoustical Society of America*. 2004; 116:1739–1748. DOI: 10.1121/1.1779272 [PubMed: 15478441]
- Belotel-Grenié A, Grenié M. The creaky voice phonation and the organization of Chinese discourse. *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*; 2004. 5–8.
- Bock O, Worringham C, Thomas M. Concurrent adaptations of left and right arms to opposite visual distortions. *Experimental Brain Research*. 2005; 162:513–519. [PubMed: 15754178]
- Boersma P, Weenink D. Praat: doing phonetics by computer, [Computer program]. Version 5.3.32, retrieved 17. 2012. Oct, from <http://www.praat.org/>
- Burnett TA, Freedland MB, Larson CR, Hain TC. Voice F0 responses to manipulations in pitch feedback. *Journal of the Acoustical Society of America*. 1998; 103:3153–3161. DOI: 10.1121/1.423073 [PubMed: 9637026]
- Cai S, Ghosh SS, Guenther FH, Perkell JS. Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong /iau/ and its pattern of generalization. *Journal of the Acoustical Society of America*. 2010; 128:2033–2048. DOI: 10.1121/1.3479539 [PubMed: 20968374]
- Chao YR. For Roman Jakobson. The Hague: Mouton; 1956. Tone, intonation, singsong, chanting, recitative, tonal composition, and atonal composition in Chinese; 52–59.
- Chen SH, Liu H, Xu Y, Larson CR. Voice F0 responses to pitch-shifted voice feedback during English speech. *Journal of the Acoustical Society of America*. 2007; 121:1157–1163. DOI: 10.1121/1.2404624 [PubMed: 17348536]
- Chen Z, Liu P, Wang EQ, Larson CR, Huang D, Liu H. ERP correlates of language-specific processing of auditory pitch feedback during self-vocalization. *Brain and Language*. 2012; 121:25–34. DOI: 10.1016/j.bandl.2012.02.004 [PubMed: 22377260]
- Chuang CK, Hiki S, Sone T, Nimura T. The acoustical features and perceptual cues of the four tones of standard colloquial Chinese. *Proceedings of the 7th International Congress on Acoustics*; 1971. 297–300.
- Cornelisse LE, Gagné JP, Seewald RC. Ear level recordings of the long-term average spectrum of speech. *Ear and Hearing*. 1991; 12:47–54. [PubMed: 2026288]
- Davison DS. An acoustic study of so-called creaky voice in Tianjin Mandarin. *UCLA Working Papers in Phonetics*. 1991; 78:50–57.
- Feng Y, Max L, Gracco V. Integration of auditory and somatosensory error signals in the neural control of speech movements. *Journal of Neurophysiology*. 2011; 106:667–679. DOI: 10.1152/jn.00638.2010 [PubMed: 21562187]
- Galea JM, Miall RC. Concurrent adaptation to opposing visual displacements during an alternating movement. *Experimental Brain Research*. 2006; 175:676–688. [PubMed: 16835793]
- Gandolfo F, Mussa-Ivaldi FA. Motor learning by field approximation. *Proceedings of the National Academy of Sciences of the United States of America*. 1996; 93:3843–3846. [PubMed: 8632977]

- Gårding E, Kratochvil P, Svantesson JO, Zhang J. Tone 4 and tone 3 discrimination in modern standard Chinese. *Language and Speech*. 1986; 29:281–293. DOI: 10.1177/002383098602900307
- Hombert J-M. The perception of contour tones. *Proceeding of the First Annual Meeting of the Berkeley Linguistics Society*; 1975. 221–232.
- Houde JF, Jordan MI. Sensorimotor adaptation in speech production. *Science*. 1998; 279:1213–1216. DOI: 10.1126/science.279.5354.1213 [PubMed: 9469813]
- Houde JF, Jordan MI. Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research*. 2002; 45:295–310. DOI: 10.1044/1092-4388(2002/023)
- Howie JM. On the domain of tone in Mandarin. *Phonetica*. 1974; 30:129–148. DOI: 10.1159/000259484
- Jones JA, Munhall KG. Perceptual calibration of F0 production: evidence from feedback perturbation. *Journal of the Acoustical Society of America*. 2000; 108:1246–1251. DOI: 10.1121/1.1288414 [PubMed: 11008824]
- Jones JA, Munhall KG. The role of auditory feedback during phonation: studies of Mandarin tone production. *Journal of Phonetics*. 2002; 30:303–320. DOI: 10.1006/jpho.2001.0160
- Jones JA, Munhall KG. Remapping auditory-motor representations in voice production. *Current Biology*. 2005; 15:1768–1772. DOI: 10.1016/j.cub.2005.08.063 [PubMed: 16213825]
- Keating PA, Esposito C. Linguistic voice quality. *UCLA Working Papers in Phonetics*. 2007; 105:85–91.
- Keough D, Hawco C, Jones JA. Auditory-motor adaptation to frequency-altered auditory feedback occurs when participants ignore feedback. *BMC Neuroscience*. 2013; 14:25.doi: 10.1186/1471-2202-14-25 [PubMed: 23497238]
- Keough D, Jones JA. Contextual cuing contributes to the independent modification of multiple internal models for vocal control. *Journal of Neurophysiology*. 2011; 105:2448–2456. DOI: 10.1152/jn.00291.2010 [PubMed: 21346208]
- Kirilloff C. On the auditory perception of tones in Mandarin. *Phonetica*. 1969; 20:63–67. DOI: 10.1159/000259274
- Klatt DH. Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *Journal of the Acoustical Society of America*. 1973; 53:8–16. DOI: 10.1121/1.1913333 [PubMed: 4687672]
- Krouchev NI, Kalaska JF. Context-dependent anticipation of different task dynamics: rapid recall of appropriate motor skills using visual cues. *Journal of Neurophysiology*. 2003; 89:1165–1175. DOI: 10.1152/jn.00779.2002 [PubMed: 12574490]
- Lee CY. Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study. *Journal of the Acoustical Society of America*. 2009; 125:1125–1137. DOI: 10.1121/1.3050322 [PubMed: 19206887]
- Li CN, Thompson S. The acquisition of tone in Mandarin speaking children. *Journal of Child Language*. 1977; 4:185–199. DOI: 10.1017/S0305000900001598
- Lin HB, Repp BH. Cues to the perception of Taiwanese tones. *Language and Speech*. 1989; 32:25–44. DOI: 10.1177/002383098903200102 [PubMed: 2533648]
- Liu H, Wang EQ, Chen Z, Liu P, Larson CR, Huang D. Effects of tonal native language on voice fundamental frequency responses to pitch feedback perturbations during sustained vocalizations. *Journal of the Acoustical Society of America*. 2010; 128:3739–3746. DOI: 10.1121/1.3666047 [PubMed: 21218905]
- Liu P, Chen Z, Larson CR, Huang D, Liu H. Auditory feedback control of voice fundamental frequency in school children. *Journal of the Acoustical Society of America*. 2010; 128:1306–1312. DOI: 10.1121/1.3467773 [PubMed: 20815465]
- Liu S, Samuel AG. Perception of Mandarin lexical tones when F0 information is neutralized. *Language and Speech*. 2004; 47:109–138. DOI: 10.1177/00238309040470020101 [PubMed: 15581188]
- MacDonald EN, Purcell DW, Munhall KG. Probing the independence of formant control using altered auditory feedback. *Journal of the Acoustical Society of America*. 2011; 129:955–965. DOI: 10.1121/1.3531932 [PubMed: 21361452]

- Maddieson I. Universals of tone. In: Greenberg J, editor *Universals of Human Language*. Stanford: Stanford University Press; 1978. 335–366.
- Massaro DW, Cohen MM, Tseng CY. The evaluation and integration of pitch height and pitch contour in lexical tone perception in Mandarin Chinese. *Journal of Chinese Linguistics*. 1985; 13:267–289.
- Max L. Stuttering and internal models for Sensorimotor control: a theoretical perspective to generate testable hypotheses. In: Maassen B, Kent R, Peters HF, van Lieshout P, Hulstijn W, editors *Speech Motor Control in Normal and Disordered Speech*. Oxford: Oxford University Press; 2004. 357–388.
- Max L, Onghena P. Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research*. 1999; 42:261–270. DOI: 10.1044/jslhr.4202.261
- Max L, Wallace ME, Vincent I. Sensorimotor adaptation to auditory perturbations during speech: acoustic and kinematic experiments. *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*; Barcelona, Spain. 2003. 1053–1056.
- Mistry S, Contreras-Vidal JL. Learning multiple visuomotor transformations: adaptation and context-dependent recall. *Motor Control*. 2004; 8:534–546. [PubMed: 15585906]
- Mitsuya T, Macdonald EN, Purcell DW, Munhall KG. A cross-language study of compensation in response to real-time formant perturbation. *Journal of the Acoustical Society of America*. 2011; 130:2978–86. DOI: 10.1121/1.3643826 [PubMed: 22087926]
- Mitsuya T, Samson F, Ménard L, Munhall KG. Language dependent vowel representation in speech production. *Journal of the Acoustical Society of America*. 2013; 133:2993–3003. DOI: 10.1121/1.4795786 [PubMed: 23654403]
- Moore CB, Jongman A. Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America*. 1997; 102:1864–1877. DOI: 10.1121/1.420092 [PubMed: 9301064]
- Ning L-H, Shih C, Loucks TM. Mandarin tone learning in L2 adults: A test of perceptual and sensorimotor contributions. *Speech Communication*. 2014; 63–64:55–69. DOI: 10.1016/j.specom.2014.05.001
- Peng G, Zhang C, Zheng HY, Minett JW, Wang WSY. The effect of intertalker variations on acoustic-perceptual mapping in Cantonese and Mandarin tone systems. *Journal of Speech, Language, and Hearing Research*. 2012; 55:579–595. DOI: 10.1044/1092-4388(2011/11-0025)
- Purcell DW, Munhall KG. Compensation following real-time manipulation of formants in isolated vowels. *Journal of the Acoustical Society of America*. 2006; 119:2288–2297. DOI: 10.1121/1.2173514 [PubMed: 16642842]
- Rochet-Capellan A, Ostry DJ. Simultaneous acquisition of multiple auditory-motor transformation in speech. *Journal of Neuroscience*. 2011; 31:2657–2662. DOI: 10.1523/JNEUROSCI.6020-20.2011 [PubMed: 21325534]
- Shen XS, Lin M. A perceptual study of Mandarin tones 2 and 3. *Language and Speech*. 1991; 34:145–156. DOI: 10.1177/002383099103400202
- Villacorta VM, Perkell JS, Guenther FH. Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *Journal of the Acoustical Society of America*. 2007; 122:2306–2319. DOI: 10.1121/1.2773966 [PubMed: 17902866]
- Wang Y, Jongman A, Sereno JA. Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustical Society of America*. 2003; 113:1033–1043. DOI: 10.1121/1.1531176 [PubMed: 12597196]
- Whalen DH, Xu Y. Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica*. 1992; 49:25–47. DOI: 10.1159/000261901 [PubMed: 1603839]
- Wolpert DM, Kawato M. Multiple paired forward and inverse models for motor control. *Neural Network*. 1998; 11:1317–1329. DOI: 10.1016/S0893-6080(98)00066-5
- Wong P. Acoustic characteristics of three-year-olds' correct and incorrect monosyllabic Mandarin lexical tone productions. *Journal of Phonetics*. 2012; 40:141–151. DOI: 10.1016/j.wocn.2011.10.005

- Woolley DG, Tresilian JR, Carson RG, Riek S. Dual adaptation to two opposing visuomotor rotations when each is associated with different regions of workspace. *Experimental Brain Research*. 2007; 179:155–165. [PubMed: 17119942]
- Xu Y. Production and perception of coarticulated tones. *Journal of the Acoustical Society of America*. 1994; 95:2240–2253. DOI: 10.1121/1.4777217 [PubMed: 8201120]
- Xu Y. Sources of tonal variations in connected speech. *Journal of Chinese Linguistics*, monography series #17. 2001:1–31.
- Xu Y, Larson CR, Bauer JJ, Hain TC. Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *Journal of the Acoustical Society of America*. 2004; 116:1168–1178. DOI: 10.1121/1.1763952 [PubMed: 15376682]
- Xu Y, Wang QE. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*. 2001; 33:319–337. DOI: 10.1016/S0167-6393(00)00063-7
- Zhou N, Xu L. Development and evaluation of methods for assessing tone production skills in Mandarin-speaking children with cochlear implants. *Journal of the Acoustical Society of America*. 2008; 123:1653–1664. DOI: 10.1121/1.2832623 [PubMed: 18345853]

Figure 1.
Timeline of the four phases in an experimental condition.

Figure 2. Block diagram of instrumentation set-up. The auditory feedback loop is shown with thick lines and arrows.

Figure 3.

F0 contours for each of the four tones averaged across baseline productions from all conditions (left: female subjects, right: male subjects). Pitch targets used for statistical analysis are marked with open circles, except for T1 which was quantified by its average frequency. The dashed vertical line marks the “boundary” between the word-initial nasal consonant and the target vowel, and, thus, defines the beginning of the tone.

Figure 4.

Waveforms and spectrograms of two T3 productions with creaky voice (left: male subject, right: female subject). Estimated F0 tracks in yellow lines are imposed on the spectrograms and the selected pitch targets are marked with red arrows.

Figure 5.

Group means (across subjects and blocks of 4 epochs) for targets T1m (mean F0 for high level T1) and T4x (maximum F0 for high falling T4) in the Control Condition (CC; no auditory perturbation) and Condition 1 (C1; F0 in the auditory feedback shifted down for T1, no shift for T2, T3, T4). Baseline and max-perturbation phases (which were compared with paired *t*-tests) are marked on the x-axis. T1m decreased gradually across trials throughout the Control Condition, and a similar—but statistically nonsignificant—decrease was observed for T4x. Relative to this Control Condition, high level T1m was raised in Condition 1 where auditory feedback was manipulated by implementing a downward pitch shift on T1 productions. F0 changes in the high falling T4x (although always produced without feedback alteration) were strongly correlated with those in T1m.

Figure 6.

Group means (across subjects and blocks of 4 epochs) for targets T3n (minimum F0 for low dipping T3) and T4n (minimum F0 for high falling T4) in the Control Condition (CC; no auditory perturbation) and Condition 2 (C2; F0 in the auditory feedback shifted down for high level T1, shifted up for low dipping T3, and no shift for T2 and T4). T3n and T4n decreased significantly from baseline to max-perturbation (phases marked on the x-axis).

Table 1

Results for normalized F0 targets during max-perturbation vs. baseline in the Control Condition (CC) and experimental conditions (C1, C2, C3).

	t	p	d	t	p	d
CC	T1m	-3.319	0.011*	T1m	0.056	0.957
	T2x	-1.507	0.170	T2x	0.051	0.960
	T2n	-1.211	0.261	T2n	0.082	0.936
	T3n	-0.999	0.347	T3n	-0.068	0.947
	T4x	-2.013	0.079	T4x	-0.066	0.949
	T4n	-0.092	0.929	T4n	-0.188	0.855
						0.1
C2	T1m	-1.708	0.126	T1m	-1.064	0.318
	T2x	-0.856	0.417			0.4
	T2n	-0.834	0.429			
	T3n	-3.207	0.012*	C3	T3n	-0.896
						0.3
	T4x	-2.045	0.075			
	T4n	-3.504	0.008*			

Note:

* $p < 0.05$.

Table 2

Results for normalized F0 targets T1m during the max-perturbation phase in Conditions 1, 2, & 3 (C1, C2, C3) vs. Control.

		<i>t</i> (<i>df</i> =8)	<i>p</i>	<i>d</i>
C1	T1m	2.364	0.046 [*]	0.8
C2	T1m	1.650	0.138	0.6
C3	T1m	1.250	0.247	0.4

Note:

^{*}
 $p < 0.05$.

Table 3

Confusion matrices and d' for all four tones averaged across 9 subjects in the listener-only group.

	Type 1 stimuli (unaltered baseline productions)				Type 2 stimuli (altered baseline productions)				Type 3 stimuli (altered adapted productions)			
	perceived (%)				perceived (%)				perceived (%)			
	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4
T1	99.6	0.2	0	0.2	100	0	0	0	100	0	0	0
T2	2.9	97.1	0	0	4.4	94.7	0.9	0	2.4	96.2	1.3	0
T3	0	1.3	98.7	0	0.4	5.3	94.2	0	0.2	1.8	98.0	0
T4	0	0.4	0	99.6	0.4	0	0	99.6	0.2	0	0	99.8
d'	5.671	5.328	5.652	5.881	5.637	5.004	5.064	6.031	5.754	5.108	5.233	6.065

Table 4

Confusion matrices and d' for all four tones averaged across 5 subjects in the speaker-listener group.

Type 1 stimuli (unaltered baseline productions)					Type 2 stimuli (altered baseline productions)				Type 3 stimuli (altered adapted productions)				
perceived (%)					perceived (%)				perceived (%)				
	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4	
stimulus	T1	99.6	0.4	0	0	98.8	1.2	0	0	100	0	0	0
	T2	0.4	99.6	0	0	0.4	99.6	0	0	0	99.2	0.8	0
	T3	0	0.4	99.6	0	0	3.2	96.8	0	0	2.4	97.6	0
	T4	0	0	0	100	0.8	0	0	99.2	0	0	0	100
d'	5.850	5.727	5.973	6.180	5.575	5.145	4.984	5.766	6.180	5.538	5.436	6.180	

Table 5

Statistical results for identification of stimuli of Type 1 vs. 2 and Type 1 vs. 3 based on paired *t*-tests applied to the speaker-listener group's *d'* for four tones.

Type of stimuli	Tone	<i>t</i> (<i>df</i> = 4)	<i>p</i>
Type 1 vs. 2 (unaltered baseline vs. altered baseline)	T1	1.272	0.272
	T2	3.505	0.025 *
	T3	3.735	0.020 *
	T4	1.633	0.178
Type 1 vs. 3 (unaltered baseline vs. altered adapted)	T1	-1.551	0.196
	T2	0.488	0.651
	T3	0.937	0.402
	T4	N/A **	N/A **

Note:

* $p < 0.05$;

** *t* cannot be computed because the standard error of the difference is 0.