Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences

Yi Xu[a)]
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511

Charles R. Larson and Jay J. Bauer
Department of Communication Sciences and Disorders, Northwestern University, 2240 Campus Drive, Evanston, Illinois 60208

Timothy C. Hain
Departments of Neurology, Otolaryngology, and Physical Therapy/Human Movement Sciences, Northwestern University, 645 N. Michigan, Suite 1100, Chicago, Illinois 60611

(Received 7 November 2003; revised 28 April 2004; accepted 30 April 2004)

Recent research has found that while speaking, subjects react to perturbations in pitch of voice auditory feedback by changing their voice fundamental frequency (F_0) to compensate for the perceived pitch-shift. The long response latencies (150–200 ms) suggest they may be too slow to assist in on-line control of the local pitch contour patterns associated with lexical tones on a syllable-to-syllable basis. In the present study, we introduced pitch-shifted auditory feedback to native speakers of Mandarin Chinese while they produced disyllabic sequences /ma ma/ with different tonal combinations at a natural speaking rate. Voice F_0 response latencies (100–150 ms) to the pitch perturbations were shorter than syllable durations reported elsewhere. Response magnitudes increased from 50 cents during static tone to 85 cents during dynamic tone productions. Response latencies and peak times decreased in phrases involving a dynamic change in F_0. The larger response magnitudes and shorter latency and peak times in tasks requiring accurate, dynamic control of F_0, indicate this automatic system for regulation of voice F_0 may be task-dependent. These findings suggest that auditory feedback may be used to help regulate voice F_0 during production of bi-tonal Mandarin phrases. © *2004 Acoustical Society of America.*
[DOI: 10.1121/1.1763952]

PACS numbers: 43.70.Aj, 43.75.Bc [AL] Pages: 1168–1178

I. INTRODUCTION

The demands of normal speech production require syllable rates of 5–7/sec, each usually consisting of a consonant and a vowel. Additionally, for speakers of a tone language like Mandarin Chinese, characteristic pitch contours over the voiced portion of the syllables (i.e., lexical tones) have to be produced to distinguish words that are otherwise phonologically identical. For example, the syllable /ma/ in Mandarin can mean "mother," "hemp," "horse," or "to scold" if it is said with either a High (H), Rising (R), Low (L), or Falling (F) tone, as shown in Fig. 1(a). Understanding speech motor control requires knowledge of the mechanisms controlling production of strings of consonants, vowels and tones. Figure 1(b) displays mean fundamental frequency (F_0) tracings across four, 5-syllable Mandarin sentences produced by a male native speaker in Xu (1999). Each 5-syllable sentence carries a tone sequence of H*x*RHH, where *x* varies across H, R, L, and F. The local variations in pitch (indicated by standard deviation bars) are small in comparison to the F_0 changes for the different tones and reveal consistent productions within each tone sequence.

These consistent patterns suggest that in addition to a need for a control process to direct the implementation of the sequential lexical pitch targets (Xu and Wang, 2001), there may also be one or more processes that ensure that the production is executed reliably and precisely. In the present paper we address a candidate second process in the form of the reliance on auditory feedback to stabilize voice F_0 around a pitch trajectory or target. The role of kinesthetic feedback in the rapid control of speech movements has been demonstrated by previous research (Abbs and Gracco, 1984; Kelso *et al.*, 1984). The role of auditory feedback in the online control of natural speech, however, has not been clearly demonstrated before. Experiments using pitch-shifted auditory feedback presented during the nonsense words "ta:tatas" (Donath *et al.*, 2002; Natke *et al.*, 2003; Natke and Kalveram, 2001) demonstrated that if the first syllable was stressed, there was a response to the pitch-shift stimulus that persisted into the next syllable. Jones and Munhall (2002) presented pitch-shifted feedback during unnaturally prolonged vowels during Mandarin speech and also showed an effect on F_0. In both of these experiments the speech was not normal, and so the question remains whether auditory feedback is used on-line for control of F_0 during normal speech. Moreover, previous pitch-shifting studies suggest that auditory feedback contributes to long-term parametric adaptation rather than to online control of voice F_0 (Jones and Munhall, 2000; Jones and Munhall, 2002). That is, com-

a)Corresponding author: Yi Xu, Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511. Telephone: (203) 865-6163, ext. 210; fax: (203) 865-8963; electronic mail: xu@haskins.yale.edu

 0001-4966/2004/116(2)/1168/11/$20.00

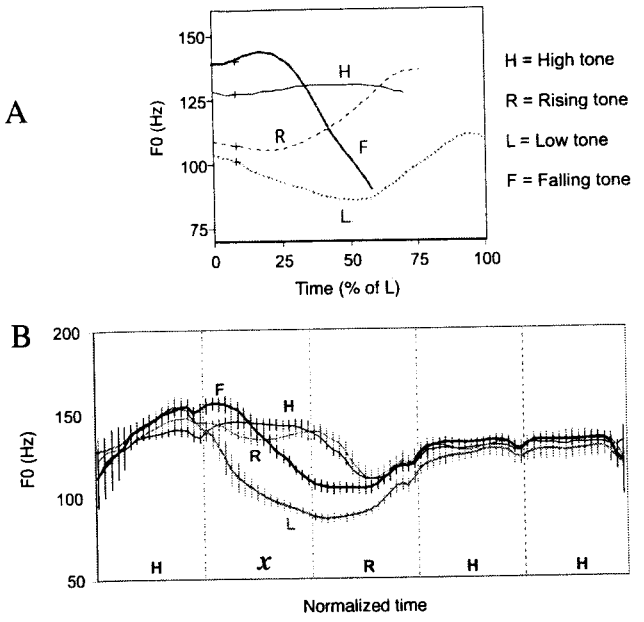

FIG. 1. (a) Mean F_0 contours of Mandarin syllable /ma/ spoken with four lexical tones: High (H), Rising (R), Low (L), and Falling (F). The syllables mean "mother," "hemp," "horse," or "to scold," respectively. Data averaged over 48 repetitions by eight male speakers (Xu, 1997). (b) Mean F_0 curves of the Mandarin tone sequences HxRHH where x varies across H, R, L, and F (which changes the meaning of the first word from "catty" to "cat-fan," "cat-rice," or "cat-honey." The vertical grids mark the syllable boundaries. The short vertical bars depict +/− one standard deviation about the mean. Data averaged across five repetitions produced by one speaker from Xu (1999).

pensatory reactions to shifted auditory feedback tend to occur *after*, rather than during articulation of a linguistic unit such as a consonant or a vowel. A possible reason for not using auditory feedback for online control of articulation is that it is slow and may exceed the temporal domain of the targeted linguistic unit. In the case of lexical tones in Mandarin, the time interval available for each tone is the duration of the syllable (Xu and Wang, 2001). As found in Xu (1997), the mean duration of a simple CV syllable in Mandarin (where C is a nasal consonant like /m/, and V is a simple vowel or diphthong like /a/, /i/, /ao/ or /ai/) is about 180 ms. This means that for auditory feedback to be effective for the online control of voice F_0 in Mandarin speech, the system must respond significantly sooner than 180 ms.

Recent research suggests that this time constraint can be met in some situations. It has been demonstrated that German speakers compensate for the mismatch between intended F_0 and feedback pitch during production of nonsense syllables (Donath *et al.*, 2002; Natke *et al.*, 2003; Natke and Kalveram, 2001). The observed response latencies were about 150 ms, which is 30 ms shorter than the syllable durations reported by Xu (1997, 1999), and therefore may be quick enough for the correction to take effect within a single syllable. However, the authors of the study were convinced that these latencies were still too slow for the system to be effective in controlling F_0 within single syllables. Thus, they concluded that "the purpose of the auditory-vocal system is not to control voice F_0 precisely within single syllables, but rather on a supra-segmental level in the context of prosody (Donath *et al.*, 2002)." Furthermore, in a study that investigated the role of auditory feedback on Mandarin, Jones and Munhall (2002) found that Mandarin speakers responded to pitch-shifted auditory feedback with compensatory changes in voice F_0 with a response latency around 200 ms. This is longer than the mean syllable duration in Mandarin reported before (Xu, 1997, 1999), although slightly shorter than the 215 ms reported by Duanmu (1994). Previous research thus suggests that neither Mandarin nor German speakers respond to pitch-shifted feedback quick enough for auditory feedback to modulate F_0 within syllables.

In studies of voice F_0 compensation using nonspeech tasks, however, response latencies shorter than those reported by the aforementioned German and Mandarin studies have been found. In a study in which trained singers received pitch-shifted feedback during glissandos, response latencies were as short as 76 ms (Burnett and Larson, 2002). In studies where non-trained singers produced sustained vowels at constant F_0, latencies of 114 ms (Hain *et al.*, 2000) and 130 ms (Larson *et al.*, 2001) have been reported. If it can be shown that the system can also respond to perturbed auditory feedback with latencies shorter than syllable duration during normal speech, it would suggest that auditory feedback may be used to control voice F_0 online during a syllable, not just for suprasegmental features of longer duration. It therefore becomes necessary to explore possible explanations for the discrepancies in reported latencies between speech and nonspeech tasks.

One possibility is that the discrepancies in response latency across these studies are due to methodological differences. In studies on sustained vowels (Burnett *et al.*, 1998; Hain *et al.*, 2000; Larson *et al.*, 2001) or prolonged vowels during Mandarin speech (Jones and Munhall, 2002), latency was defined as an F_0 trajectory that exceeded +/−2 S.D.'s of the prestimulus mean F_0. As has been pointed out, this technique likely overestimates the actual time when the voice F_0 begins to respond to the pitch-shifted feedback (Donath *et al.*, 2002). Moreover, the technique requires a steady F_0 level prior to the stimulus, which is not appropriate for speech because the prestimulus F_0 trajectory varies. In addition, the latency reported by Jones and Munhall may also have been skewed since the speakers produced long sustained vowels during the experiment, which is not typical of normal Mandarin speech. Because of the discrepancies in latencies from previous studies (Burnett *et al.*, 1998; Burnett and Larson, 2002; Donath *et al.*, 2002; Hain *et al.*, 2000; Jones and Munhall, 2002; Larson *et al.*, 2001; Natke *et al.*, 2003; Natke and Kalveram, 2001) and the critical role this plays in defining functions of auditory feedback during speech, we adopted a technique similar to that developed by Donath and Natke (Donath *et al.*, 2002; Natke *et al.*, 2003) for measuring latency. We used a statistical test to determine whether the F_0 trajectories following the stimulus differ significantly from control (nonstimulated) trajectories during the production of disyllabic phrases in normal Mandarin speech.

Another issue, especially relevant to studies of speech in nontone languages such as German, is that the latency and magnitude of the compensatory responses of the audio-vocal system to pitch-shifted feedback vary depending on the

strength of the demand of the vocal task. In a study in which subjects were instructed to modulate their voice F_0 when they perceived pitch-shifted feedback, the response latencies were reduced (Hain *et al.*, 2000). Also, Natke *et al.* (2003) showed that the pitch-shift response is larger in magnitude in singing compared to previous studies of sustained vowels in nonsingers. These findings suggest that the response latencies and magnitudes reported in previous studies observed with sustained vowels may be slower and smaller than those involving dynamic pitch movement in natural speech.

The present study was therefore designed to test the following hypotheses using Mandarin speech. First, compensatory responses to pitch-shifted feedback occur during the natural production of tonal sequences in the language. Second, the compensatory responses are fast enough to serve as part of the online sensory-motor control mechanisms for tone production in speech. Third, the latency and magnitude of the compensatory responses vary with the demand of the tonal production, and are generally faster and larger for speech in which accurate F_0 control is necessary. To test these hypotheses, we devised a technique that introduces pitch-shifted stimuli at specific times during a disyllabic speech production.

The primary goal of the experiment was to see whether and how speakers react to the pitch-shifted auditory feedback during three bi-tone sequences representative of Mandarin speech: High-High (H-H), High-Rising (H-R), and High-Falling (H-F). The pitch-shift stimuli were timed so that they would occur either during the first syllable or during the transition between the first and second syllables. The object of testing these two timing variables was to test the effects of pitch-shifted feedback during relatively steady state F_0 contours and during dynamic F_0 contours. Both upward and downward pitch-shift stimuli were presented to assess whether stimulus direction interacted with the direction in the change in the tone associated with each bi-tonal sequence. For all three independent variables, we predicted that response magnitudes would be larger and response latencies and peak times would be shorter than those previously observed in a static vowel condition. These effects would suggest the need for a rapid response of sufficient magnitude to correct for production errors within the syllable.

II. METHODS

A. Subjects

Subjects consisted of six speakers whose first language was the Beijing dialect of Mandarin Chinese (four females and two males; ages 20–40). Subjects reported no history of hearing loss, neurological deficits and/or speech-language disorders.

B. Apparatus and stimuli

Subjects were seated comfortably in a small acoustically shielded chamber and asked to read aloud the experimental stimuli at approximately 70 dB SPL (self-monitored visually with a Dorrough Loudness Monitor model 40-A). Voice output was transduced through an AKG boom-set microphone (model HSC 200) at a microphone-mouth-distance of 3 cm. The microphone signal was amplified by a Mackie mixer (model 1202) and then processed for pitch-shifting through an Eventide Ultraharmonizer (H3000 SE). The microphone signal was then mixed with 70 dB SPL pink masking noise (Goldline Audio Noise Source, model PN2, spectral frequencies 1 to 5000 Hz) using a Mackie mixer (model 1202-VZL), routed through HP 350 dB attenuators, and presented to the subject over AKG headphones (model HSC 200) at 80 dB SPL after amplification by a Crown D75-A amplifier. Acoustical equipment was calibrated with a Brüel & Kjær 2203 sound level meter (A weighting). During each utterance a MIDI software program (Max v. 4.1 by Cycling '74) directed the Ultraharmonizer to randomly pitch-shift the voice feedback upwards or downwards by 200 cents for a duration of 200 ms, or to leave the feedback unchanged (control). The Ultraharmonizer automatically shifts pitch in units of cents because this scale is logarithmically related to F_0 and is constant relative to the absolute F_0 of a given subject. The subject's voice onset automatically activated the MIDI program using a locally fabricated Schmitt trigger circuit that detected a positive voltage ($\sim$100 mv) on the leading edge of the amplified vocal waveform and produced a TTL pulse with less than 1 ms delay. The output of this circuit then was fed to a modified Macintosh mouse to simulate a "mouse click." The pointer of the mouse was kept stationary during an experiment over a "button" on the MIDI software program, and when a vocalization began, the trigger circuit activated the mouse to initiate the MIDI program. The MIDI program then presented a randomized signal to the Harmonizer. The variability in the timing of the MIDI output from the onset of the pulse from the vocal detection circuit was about 25 ms. The MIDI signal to the harmonizer was either a command to produce no shift in pitch feedback, an increase or a decrease in pitch feedback to the subject. The variability in the delay time for the harmonizer to present a pitch-shift stimulus was about 15 ms. Thus, the total variability in the delay time between onset of vocalization and the presentation of the pitch-shift stimulus to the subject was about 40 ms.

Subjects read a randomized list of disyllabic nonsense Mandarin phrases (/ma ma/) printed in Chinese characters, produced without a carrier frame, consisting of 3 bi-tonal patterns at a comfortable rate of about 0.5 s per phrase. Similar syllable sequences were used before in an acoustic study (Xu, 1997). The tonal patterns, High-High (H-H), High-Rising (H-R), and High–Falling (H-F), were chosen to place different linguistic demands on speech production seen in Mandarin (see Fig. 1). In the H-H phrase, the variation in F_0 is rather minimal as compared to both the H-R and H-F phrases. For the H-R phrase, the speaker must maintain a relatively high F_0 followed by a rapid F_0 drop with a subsequent rise. Thus to be perceived as H-R, there must be a distinct drop in F_0 to enable the subsequent rise, and both of these changes must differ from the F_0 of the "High" tone. The H-F phrase should require greater precision than the H-H phrase because the "Falling" tone must be clearly differentiated from the "High" tone. However, accuracy at the end of the "Falling" tone is inconsequential because glottal-

ization frequently occurs with the drop in F_0 when it is at the end of an utterance.

The second procedural variable we manipulated was the timing of the stimulus. Stimuli were either presented relatively early in the first syllable (~100 ms after vocalization onset) or near the beginning of the second syllable (~250 ms after vocal onset). The actual timing of the stimuli varied from 75–115 ms for the "100 ms" condition and 250–290 ms for the "250 ms" condition. The early timing condition was designed to disrupt the first (High) tone and possibly the transition into the second tone (High, Rising or Falling), while the later timing would primarily disrupt dynamic transitions into the second tone. The timing of the stimulus with respect to the syllable boundaries varied as a function of the speed of speaking but was generally within +/−50 ms of the syllable transition.

The final stimulus parameter we manipulated was the direction of the pitch-shift stimulus. Pitch-shift stimuli were either upward (increase in pitch of voice feedback) or downward (decrease in pitch of voice feedback). In keeping with previous studies, we anticipated that most responses would be opposite in direction to the stimulus (compensatory) and would occur in both stimulus directions. In addition, we anticipated that larger responses would be observed in the H-R and H-F phrases when the direction of the stimulus was opposite to the ongoing F_0 modulation at the time of the stimulus presentation. We predicted that for the H-R phrase, a downward stimulus in the 250 ms timing condition would elicit a smaller response than an upward stimulus because the latter would be perceived as a failure to reach the desired low F_0 trajectory, and subjects would respond with a greater response magnitude. We also predicted that downward stimuli occurring during the elevation prior to the tone drop in the H-F phrase would elicit a larger response than an upward stimulus.

For each phrase and timing condition, the subject produced about 20 repetitions as control trials, 20 with an increase in voice pitch feedback, and 20 with a decrease in pitch feedback. Prior to actual data collection, subjects produced several practice phrases, and their temporal patterns were measured for consistency. Consistency was evaluated in two ways. Our Mandarin speaking experimenter (YX) listened to the speech to be certain that the words were spoken with correct pronunciation. Second, we measured the temporal patterns of the waveforms on a computer screen to make sure the durations of each syllable were consistent across trials. We allowed for variations in timing of individual syllables of approximately 50 ms. If the productions lacked a consistent temporal pattern, subjects were instructed again on how to produce the phrases consistently. Consistent patterns were necessary to adjust the MIDI software to deliver the stimuli at the same relative time in the phrase for each subject. Between sets of trials, subjects rested and hydrated themselves.

C. Analysis

The subject's voice output and auditory feedback was low pass filtered at 5 kHz and digitized on-line onto a laboratory computer at 10 kHz (12-bit sampling). A control signal representing the onset and direction of the pitch-shift stimulus was digitized as well. In off-line analysis, voice and auditory feedback signals were low pass filtered (digital filter, 5th order) at the mean F_0 level for each subject, differentiated so as to equalize the waveform amplitude, and then smoothed with a five-point binomial, sliding window. A voice F_0 analog wave was then extracted using a software algorithm (Igor Pro v. 4.06 by Wavemetrics) that detected positive-going threshold-voltage crossings, interpolated the time fraction between each pair of sample points that constituted a crossing, and calculated the reciprocal of the period defined by the center points. The resulting F_0 analog (Hz) was further transformed into cents, using the following equation: $\text{cents} = 100\ (39.86 \log_{10}(f2/f1))$, where f1 is an arbitrary value of 196 Hz and f2 is the voice signal in Hertz. The conversion of all F_0 analog signals to cents allowed for a comparison of F_0 across different pitch levels and subjects. An interactive program was then used to generate event-related averages for each experimental condition. Each syllable phrase was time-aligned to the start of each vocalization to reduce the dispersion of temporal variations in the F_0 trajectory, marked as to the type of bi-tonal pattern (H-H, H-R, or H-F), and sorted based on the pitch-shift stimulus condition (up, down, or control). An average waveform of the F_0 analog was then generated for each bi-tonal pattern and stimulus condition per subject. Measures of the average response to the pitch-shift stimuli were made by a comparison with the average of the control wave for the specific condition for that subject.

An additional analysis was performed to estimate the response latency by determining if the averaged test waves differed significantly from the averaged control waves. For this procedure, F_0 analog waves were first decimated to 100 Hz. Then a t-test (equal variance; two tailed) was performed comparing all test trials of the F_0 analog wave for each condition with the corresponding set of control trials on a point-by-point basis. That is, t-tests were performed between test and control trials in 10 ms intervals. The result of this operation produced a wave comprised of a probability value (p) between a set of test and control waves for each 10 ms epoch of the ensemble average. Figure 2 illustrates the relation between the average waves (A), "p" values (B), and the difference wave (C) for one subject. Responses in which p values failed to reach a significance level of at least 0.02 and remain significant for at least 50 ms were defined as nonresponses. Latency was defined as the first p value that occurred at a significance level of 0.05 [the circled value in Fig. 2(b)]. Although a Bonferroni correction would normally be warranted for multiple t-tests, we believe that rejecting responses that do not last at least 50 ms is a more conservative approach to control for type I errors. The logic for this argument is that the twitch contraction times for most laryngeal muscles are less than 30 ms (Alipour-Haghighi *et al.*, 1987; Kempster *et al.*, 1988; Larson *et al.*, 1987), and a response that lasts at least 50 ms is more likely to represent a neuromuscular event than non-muscular signal transients. The response magnitude and the time of the peak response magnitude were measured from the maximal point on the difference wave, if these values occurred within a time frame

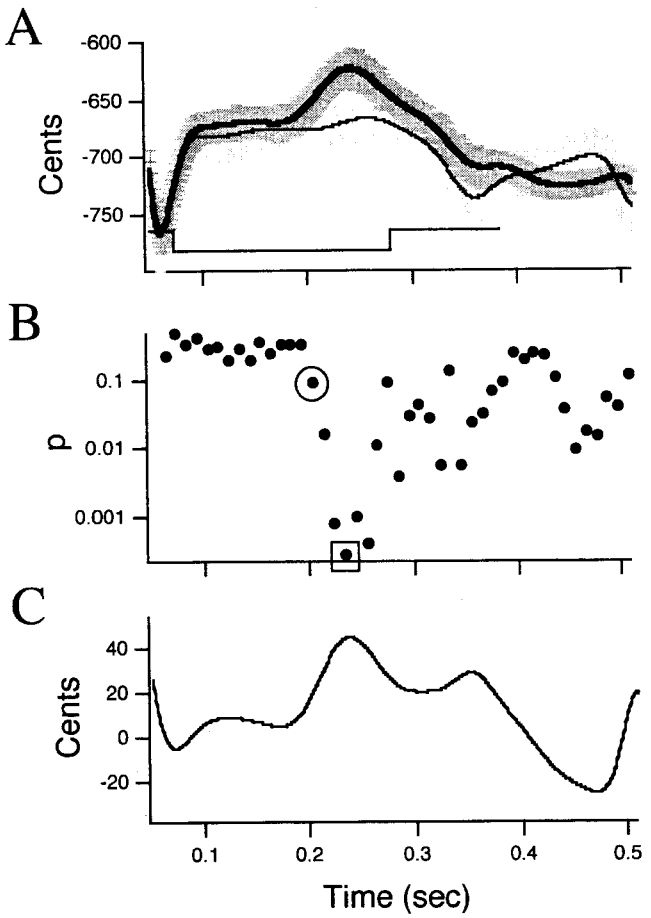

FIG. 2. (a) Averaged test wave (heavy black line) superimposed on standard error of the mean (SE) (dark gray wide line) in response to a downward pitch-shift stimulus. Control average wave (thin black line) superimposed on SE (light gray wide line). The square wave at the bottom indicates time and direction of stimulus (vertical dimension not to scale). (b) Probability (p) values resulting from a t-test comparison of test and control waves (see the text for details). The circled point is defined as response latency and the boxed point is the time of peak response magnitude. (c) The difference wave calculated by subtracting control from the test average wave.

defined by significant p values. However, in some cases, the difference wave was continuously changing throughout the period defined by significant p values and a peak in the difference wave could not be defined. In these instances, the peak time and peak magnitude were measured from the difference wave at the time indicated by the most significant p value [the boxed value in Fig. 2(b)]. The response direction with respect to the stimulus direction was also noted for each average wave and classified as compensatory or "following." A response was considered to be compensatory if the response's direction was opposite to that of the stimulus direction, and conversely, a response was classified as "following" if the response was in the same direction as the pitch-shift stimulus.

TABLE I. Total number of "following" (FOL) and compensating (COMP) responses and nonresponses (NR) across three bi-tonal patterns (H-F, H-H, and H-R).

	H-F	H-H	H-R	TOTAL
FOL	3	0	4	7
COMP	15	18	15	48
NR	6	6	5	17
TOTAL	24	24	24	72

TABLE II. Total number of "following" (FOL) and compensating (COMP) responses and nonresponses (NR) across two timing conditions (100 and 250 ms).

	100	250	TOTAL
FOL	6	1	7
COMP	20	28	48
NR	10	7	17
TOTAL	36	36	72

For two of the disyllabic sequences there was a large, rapid drop in the F_0 trajectory (H-R, H-F). In most cases it was possible to measure a response just before or after the drop. However, in some cases, the response to the pitch-shift stimulus appeared to be a timing difference where the experimental average occurred either earlier or later than the control average (phase-shifted). In these cases, it was not possible to measure a change in magnitude and these were classified as "nonresponses" as a result of a phase-shift.

The statistical analysis of response magnitudes and latencies was done with repeated-measures ANOVAs. In cases where the averaged signals failed to differ significantly from control waves, neither latencies nor magnitudes were measured. These cases resulted in missing data for a subject for a condition. In order to meet the assumptions of a repeated-measures ANOVA, the missing data points were replaced with mean values calculated from the measured data from other subjects for that condition.

III. RESULTS

Out of a possible 72 averaged responses across subjects (12 averages per subject across 3 syllable conditions, 2 stimulus directions, and 2 onset conditions), there were 17 nonresponses, seven "following" responses, and 48 compensating responses. Tables I, II, and III provide a breakdown of response types across experimental conditions. It is noteworthy that no "following" responses were observed for the H-H productions and only one for the 250 ms timing condition. Compensating and nonresponses were rather evenly dispersed across bi-tonal patterns, stimulus timing, and stimulus direction. Eight nonresponses were due to an apparent phase shift in the observable response. Nine of the nonresponses were those in which the averaged waveform failed to differ significantly from the control waveform for a duration of at least 50 ms. The individual responses comprising the average (~20 responses per average) may have been a

TABLE III. Total number of "following" (FOL) and compensating (COMP) responses and nonresponses (NR) across stimulus direction (DOWN and UP).

	DOWN	UP	TOTAL
FOL	2	5	7
COMP	27	21	48
NR	7	10	17
TOTAL	36	36	72

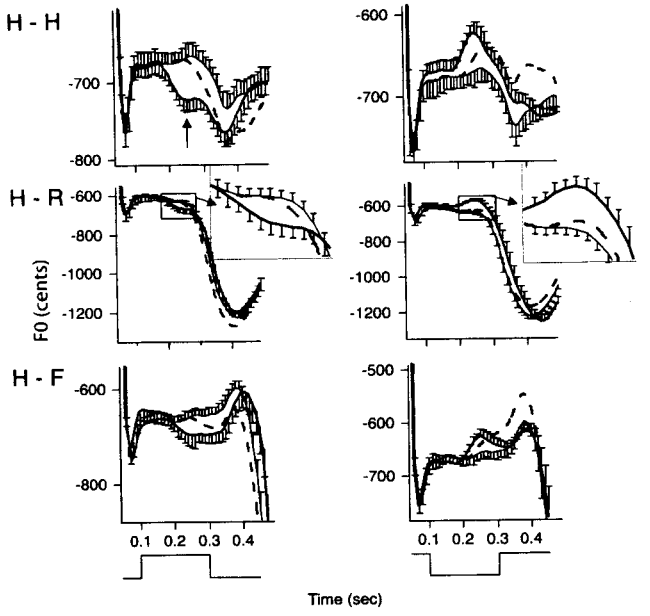

FIG. 3. Control (thin black line) and test average waves (thick black line) during H-H, H-R, and H-F sequences at the 100 ms stimulus timing condition. Heavy dashed lines are simulations produced by the model (see the text). The vertical arrow indicates time where the response magnitude was measured for this trace (see the text). Error bars represent the standard error of the mean for a single direction. The inset shows an expanded portion of average waves. Curves at the bottom indicate the time and direction of the stimulus. For all panels, the stimulus onset occurred approximately at 0.1 s. The x-axis (time) starts at 0.05 s, which is 0.05 s after vocalization onset. Note that the y-axis differs for each plot.

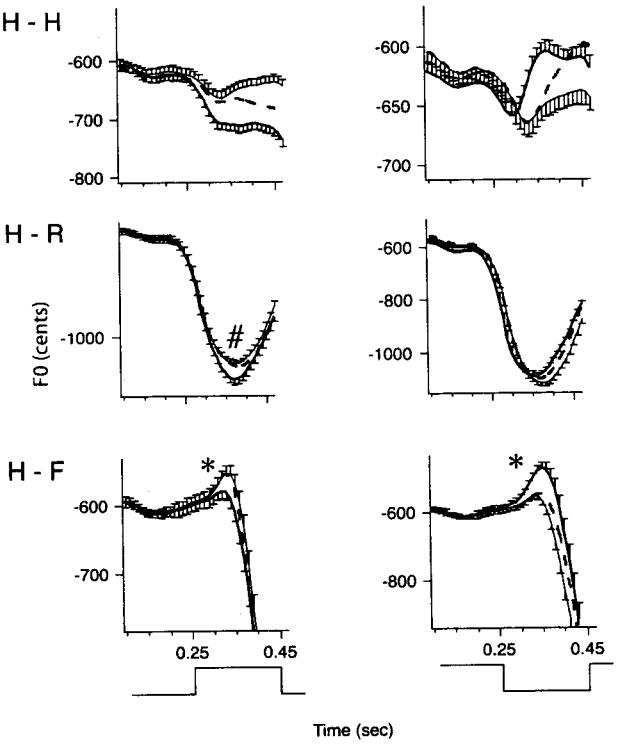

FIG. 4. Control (thin black line) and test average waves (thick black line) during H-H, H-R, and H-F sequences at the 250 ms stimulus timing condition. Error bars represent standard error of the mean for a single direction. H-R, "#" marks large difference between control and test waves mentioned in text. H-F, "*s" indicate rise in F_0 prior to major drop (see the text). Heavy dashed lines are simulations produced by model (see the text). Stimulus onset began at 0.25 s following vocal onset. In all illustrated examples, differences between control and test averages were statistically significant. The x-axis (time) starts at 0.1 s, which is 0.1 s after vocalization onset. Note that the y-axis differs for each plot. See the legend of Fig. 3 for further details. All traces for Fig. 3 and 4 were taken from the same subject.

mix of compensatory and "following" in direction. This may have resulted in the average being classified as a nonresponse due to cancelling of individual compensatory and following responses. However, it is difficult to make such determinations because of waveform variability. This is the reason for relying on the criteria of significant differences in *average* responses for determining whether responses occurred or did not.

There were a variety of responses types across subjects with the different phrases. For the H-H phrase, all responses occurred, by default, during steady F_0 productions. For the H-R phrase, some subjects produced a response prior to the drop in the F_0 trajectory, some during the drop, and some at the bottom of the trajectory. Most of the latter responses were observed with the 250 ms timing condition, however, a few were noted for the 100 ms timing condition as well. For the H-F phrase, most responses occurred prior to the drop in the F_0 trajectory, while a few occurred during the drop.

Figures 3 and 4 illustrate characteristic average responses to pitch-shifted stimuli (thick black lines with error bars) superimposed on average control curves (thin lines with error bars). Also shown are mathematical simulations (dashed lines with no error bars), which will be discussed subsequently. Figure 3 (H-H) displays representative data from one subject for the bi-tonal pattern (H-H) for increasing and decreasing pitch feedback under the 100 ms timing condition. The beginning of the traces on the left is 50 ms after vocal onset; vocal onset is not shown because of the large change in the cent magnitude with the onset of voicing. On the left side, an upward shift in pitch feedback shortly after vocal onset (indicated by the bottom trace representing the direction and timing of the stimulus) led to a decrease in the average response when compared with the control average. The response average drops below the control average at approximately 106 ms following the stimulus onset (latency) and reaches a peak depression (maximal separation between the control and response averages) of -59 cents at 165 ms (arrow). Following the peak response, the traces of the response and control averages converge following the stimulus offset. The traces on the right in Fig. 3 (H-H) represent the average response to a downward pitch-shift stimulus. The response latency occurs closer to 117 ms, and the response peak (38 cents) occurs at about 151 ms.

In Fig. 3 (H-R), the stimulus again begins 100 ms after onset of the "High" tone and ends at about the time the F_0 trajectory starts its descent toward the "Rising" tone. For both the increasing and decreasing stimuli, the response (thick lines with error bars) peaks prior to the beginning of the downward trajectory. The response to the downward stimulus approaches the trajectory of the control waves (thin lines with error bars), but never reaches it due to the time lag

TABLE IV. Mean latency (sd) in ms, magnitude (sd) in cents, and peak time (sd) in ms across three bi-tonal patterns.

PHRASE	LAT (sd)	MAG (sd)	PT (sd)
H-F	174 (79)	83 (50)	251 (73)
H-H	147 (41)	49 (20)	228 (45)
H-R	171 (70)	84 (45)	235 (70)

TABLE V. Mean latency (sd) in ms, magnitude (sd) in cents, and peak time (sd) in ms across timing conditions.

TIMING	LAT (sd)	MAG (sd)	PT (sd)
100	184 (77)	69 (42)	254 (76)
250	148 (50)	74 (44)	225 (49)

in the response trace. Such time lags were common in most subjects. The response to the upward stimulus peaks and then crosses the trajectory of the control wave as it seems to overshoot the control wave slightly. This behavior again could be interpreted as a time lag in the perturbed response. In both cases, responses occur prior to the end of the stimulus and are incorporated into the overall F_0 trajectory associated with the bi-tonal F_0 configuration.

For H-F phrases (Fig. 3), the F_0 remained at a higher level for a longer duration of time than was seen for the H-R phrases. Thus, the responses were seen during the steady phase of the F_0 prior to the "Falling" tone. The absolute response magnitude (59 cents) to the upward shift (left) is larger than that (41 cents) to the downward shift (right), as predicted for the rising F_0 trajectory prior to the drop in F_0. Following the early response, an apparent phase lag was seen during the "Falling" tone (Fig. 3 (H-F) left), and is similar to the phase-lag observed for the H-R phrases.

Figure 4 illustrates examples of responses to pitch-shift stimuli delivered 250 ms following vocal onset for the same subject as in Fig. 3. In these cases, the stimulus began during the first tone and terminated during the second tone. In the case of the H-H phrase (Fig. 4), responses are very similar to those seen in the 100 ms condition. That is, the overall F_0 trajectory was relatively flat through the entire phrase, and the responses compensated for the pitch-shifted auditory feedback. For both the H-R and H-F conditions, the stimulus was present during the time when there was a major reduction in F_0. There are three major observations for the responses in these conditions. First, sometimes a response appeared before the major drop in F_0 for the H-F phrases, as seen in Fig. 4 (H-F, left and right, indicated by "*"s). Because the high tone of the H-R phrase was shorter in duration than that in the H-F phrase, this early response prior to the drop in F_0 was typically not seen in the H-R phrase. Second, the onset of the response during the high tone of the phrase also was accompanied by a phase-shift, which became apparent by the end of the phrase in Fig. 4 (H-F, both left and right). Both phase leads and phase lags of response with respect to the control were observed. Third, there was frequently a large difference between the control and response waveforms that occurred near the bottom of the contour associated with the beginning of the final F_0 rise ["#," Fig. 4 (H-R)]. Since the H-F phrase did not have a transition and since vocalization frequently ceased at or near the low point of the "Falling" tone, it was not possible to measure a response at the bottom of the trajectory.

Quantitative measures in Tables IV, V, and VI provide means and S.D.'s of response latency, magnitude, and peak time across subjects and conditions. The latency values shown in these tables vary from 147 to 184 ms. However, when values are broken down by specific conditions (Table VII), latencies for two of the H-H and one H-F condition were less than 130 ms. Latency values in other conditions were close to 200 ms. The overall median latency was 143 ms. For a statistical analysis of latency measures, a square root transformation was done to achieve a normal distribution. A three-factor repeated-measures ANOVA was performed on latency with phrase type, direction, and stimulus timing as independent variables. No main effects were found. However, there was a significant interaction between phrase type and stimulus direction [$F(2,10)=5.09$, $p=0.03$]. The latencies for both the H-H and H-F phrases increased for the downward stimuli compared to the upward stimuli, whereas for the H-R phrase, latencies decreased for the downward stimuli. The shorter latency for the H-R phrases with downward stimuli may indicate that when a planned drop in F_0 coincides with the approximate time of a downward response, the response latency is reduced (for the H-R phrase, the drop in the F_0 trajectory occurs sooner than in the H-F phrase). Although there appeared to have been a dramatic decrease in latency comparing the 100 ms and 250 ms timing conditions for the H-F condition as a function of stimulus direction, there was no overall effect on latency. Nevertheless, this change is in the same direction as changes in the peak time measures (see below). A similar finding was reported previously by Hain *et al.* (2000) for sustained vowels.

A three-factor repeated measures ANOVA was performed on magnitude with phrase type, stimulus timing, and stimulus direction as independent variables. Significant main effects on response magnitude were found for phrase type [$F(2,10)=9.36$, $p=0.005$] and stimulus direction [$F(1,5)=12.7$, $p=0.016$] (Table VIII). Post hoc testing revealed that responses for the H-R and H-F phrases were significantly larger than those for the H-H phrase ($p=0.008$ and $p=0.020$; *post hoc* Sheffé). Response magnitudes were generally greater for downward than upward stimuli. A significant interaction was observed between phrase type and stimulus timing [$F(2,10)=5.19$, $p=0.028$], which was due to the much greater increase in response magnitude for the H-R and H-F phrases compared to the H-H phrase. A significant interaction was also observed between stimulus timing and stimulus direction [$F(1,5)=17.43$, $p=0.009$]. This effect may be due to the much greater increase in response magnitude between the upward and downward stimuli for

TABLE VI. Mean latency (sd) in ms, magnitude (sd) in cents, and peak time (sd) in ms by stimulus direction.

DIRECTION	LAT (sd)	MAG (sd)	PT (sd)
DOWN	165 (67)	85 (49)	237 (68)
UP	164 (65)	58 (29)	240 (60)

TABLE VII. Mean latency (sd) in ms across three phrase types, two timing conditions, and two stimulus directions.

PHRASE		H-H		H-R		H-F	
TIMING		100	250	100	250	100	250
DIR	UP	127 (38)	124 (25)	192 (46)	192 (74)	210 (91)	121 (21)
	DOWN	174 (55)	159 (36)	162 (93)	138 (51)	243 (100)	146 (55)

the 250 ms timing condition compared with the 100 ms timing condition, most apparent for the H-R and H-F phrases. When examined across both timing and direction conditions, response magnitudes for both H-R and H-F phrases were greatest for downward stimuli with the 250 ms timing condition. This observation does not seem to confirm our hypothesis that response magnitudes should be larger for upward directed stimuli when the F_0 trajectory is downwards. We note, however, that the exact timing of the stimulus with respect to the F_0 trajectories varied with changes of speaking rate, both within and across subjects. As a result, for both H-R and H-F, the 250 ms stimuli coincided sometimes with the downward movement of the F_0 trajectory, sometimes with the upward movement, and sometimes with both of them. This variability in timing may have precluded the identification of systematic changes in response magnitude according to our hypotheses. Nevertheless, the H-F response measures made at the time of the slight rise in F_0 preceding the large drop in F_0 [Figs. 3 (H-F) and 4 (H-F)], produced, as predicted, larger responses for downward stimuli compared with upward stimuli [$F(1,16)=9.06$, $p=0.008$].

A three-factor repeated measures ANOVA was performed also on peak time with phrase type, stimulus timing, and stimulus direction as independent variables (Table IX). There were no significant main effects. However, there was a significant interaction between phrase type and stimulus timing [$F(2,10)=5.05$, $p=0.03$]. For the H-R and H-F phrases, peak times decreased for the 250 ms timing condition compared with the 100 ms timing, while for the H-H phrase, peak times increased between these two timing conditions. Overall, mean peak times were 238 ms. Thus, there was a reduction in latency and peak time measures when the pitch-shift stimuli occurred closer in time to the drop in the F_0 trajectory.

IV. MODELING OF RESPONSES

We simulated our data using a previously published model of responses to pitch-shift stimuli for sustained vowel phonations (Hain *et al.*, 2000). Our purpose was to not to reproduce experimental data, but rather to compare timing and magnitude of responses for the present speech task to responses expected for non-speech tasks. Figure 5 shows the main features of the model and Figs. 3 and 4 contain simulations to be compared with experimental data. The model of pitch stabilization uses negative feedback and delays to simulate compensatory responses to perturbations in the pitch of auditory feedback. The purpose of this approach was to be able to compare timing and magnitude of responses here to those predicted by this model, which was previously optimized for simulation of a nonspeech task.

In producing the output shown in Figs. 3 and 4 the control F_0 was used as the desired F_0 signal. There were several differences between simulation and experimental responses that can be seen from an inspection of Figs. 3 and 4. Experimental responses frequently began earlier than the simulation and were often larger. This is consistent with the finding of earlier and larger magnitude responses for the present data than for previous studies of less behaviorally relevant productions. Experimental responses also tended to lag control and simulation traces for the 100 ms delay condition (Fig. 3), which was not consistent for the 250 ms delay (Fig. 4). This observation suggests that given enough time, perturbations may slightly slow down production of the next syllable.

V. DISCUSSION

We conducted the present experiment with three main questions in mind. First, would native speakers of Mandarin respond to pitch-shifted auditory feedback during natural production of bi-tonal sequences by changing their vocal pitch to compensate for the pitch-shift? The answer to this question is "yes"—most subjects responded to pitch-shift stimuli with a compensatory change in voice F_0 during the production of bi-tonal sequences, which consisted of various F_0 rises and falls. These results are compatible with those of previous studies (Donath *et al.*, 2002; Jones and Munhall, 2002; Natke *et al.*, 2003). Several "following" responses were also observed, which have been previously reported for sustained vowel productions.

Our second experimental question was as follows: Would the responses be fast enough to make a difference before the production of a tone is completed? The answer to this question seems to be "sometimes." The median response latency (143 ms) was shorter than the 200 ms latency found by Jones and Munhall (2002), which may relate to differences in methods. Many responses had latencies less than 130 ms, and some were close to 100 ms. These latencies

TABLE VIII. Mean response magnitude (sd) in cents across three phrase types, two timing conditions, and two stimulus directions.

PHRASE		H-H		H-R		H-F	
TIMING		100	250	100	250	100	250
DIR	UP	42 (15)	51 (30)	85 (43)	58 (27)	56 (26)	51 (24)
	DOWN	50 (14)	48 (20)	91 (65)	105 (20)	78 (43)	129 (51)

TABLE IX. Mean response peak time (sd) in ms across three phrase types, two timing conditions, and two stimulus directions.

PHRASE		H-H		H-R		H-F	
TIMING		100	250	100	250	100	250
DIR	UP	217 (63)	249 (30)	224 (56)	240 (88)	284 (63)	204 (46)
	DOWN	216 (67)	224 (38)	261 (79)	202 (51)	312 (130)	223 (33)

are shorter than those reported by previous studies on pitch-shift experiments with nonsense syllables in German and prolonged vowels in Mandarin (Donath *et al.*, 2002; Jones and Munhall, 2002; Natke *et al.*, 2003), but are comparable to the latencies previously reported with sustained vowels (Hain *et al.*, 2000; Larson *et al.*, 2001). However, several responses had longer latencies, similar to those reported by Jones and Munhall (2002), and are also similar to secondary responses previously reported during sustained vowels (Burnett *et al.*, 1998; Hain *et al.*, 2000). Yet, the median latency of ~143 ms is shorter than the syllable duration of the Mandarin syllables for similar CV structures found in Xu (1997): 186 ms for /ma/ with many tonal combinations (duration values unpublished before), or Xu (1999): 180 ms for /mao/, /mi/, /mo/, /na/, /mai/ and /tao/, also with many tonal combinations. That is, with a delay less than a syllable (143 ms), the compensation can begin. Nevertheless, the mean response latency (~164 ms) was comparable to average syllable length. Moreover, response peak times (~238 ms) also occur after the average syllable has completed. Thus, due to variability in response latency, some responses are fast enough to make a difference within a phrase, but on average the responses are quite late compared to the syllable duration.

Our third question was whether the latency and magnitude of the compensatory responses would be generally faster and larger than those observed in conditions involving only nonspeech. There were four observations that supported this hypothesis. First, the response magnitudes during the production of natural tone sequences found in the present study are indeed significantly larger than those reported for sustained vowels (Burnett *et al.*, 1998) and are slightly larger than those reported during German nonsense speech (Donath *et al.*, 2002) or in a previous investigation of Mandarin tones produced with prolonged vowels (Jones and Munhall, 2002). Second, responses during the dynamic H-R and H-F phrases were generally larger than in the static H-H phrase. This indicates that auditory feedback may be more important for dynamic control of F_0 than static control. Third, during the slight rise in F_0 prior to the drop in the H-F phrase, response magnitudes were larger for downward stimuli compared to upwards stimuli. Finally, response latency decreased with downward stimuli in the H-R phrase, indicating that the system can alter timing of responses in addition to the magnitude depending on the need for auditory feedback to help control voice F_0. Moreover, in a recent study, Natke *et al.* (2003) reported that responses were larger during singing than during speech, and presumably singing requires greater F_0 accuracy than speech. It is noteworthy that the response magnitudes reported in the present study are approximately the same as those reported during singing (Natke *et al.*, 2003), and suggest F_0 control in Mandarin is as sensitive to auditory feedback as is in singing. Task-dependent modulation of reflexes, or other types of stereotypic responses to stimulation has been observed in many systems and is generally interpreted to reflect dependence of accurate motor execution on sensory monitoring (Gracco and Abbs, 1989; Gracco and Abbs, 1985; Saltzman *et al.*, 1998; Shaiman, 1989; Shaiman and Gracco, 2002).

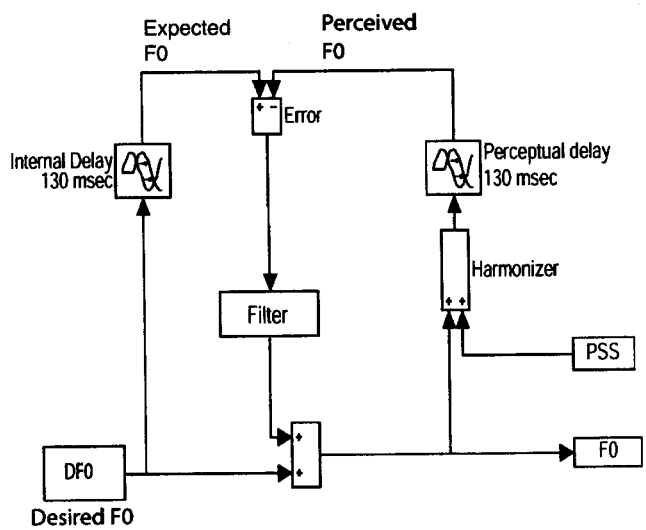

FIG. 5. Mathematical model of pitch stabilization. On the left side, *Desired* F_0 is input. Corrections are added at the summing junction at the center, bottom to produce F_0. Corrections are computed by comparing perceived F_0 (the upper right hand part of the diagram) with *Expected* F_0. *Perceived* F_0 is delayed by 130 ms with respect to F_0 reflecting delays in registration and production of sound. *Expected* F_0 is also delayed by 130 ms so that both signals are in the same time frame. The difference between *Expected* F_0 and *Perceived* F_0, *Error*, is filtered and used to adjust the F_0 signal.

An important experimental design feature of this study was the accurate timing of the stimuli with respect to the speech tokens. Care was taken to make sure stimuli were presented at the same approximate time in each of the phrases for all speakers by careful coaching of subjects, monitoring of signals produced by the subjects, and the use of a voice activated trigger circuit. However, we were unable to precisely control the timing of the MIDI program or the processing by the harmonizer, which together produced about 40 ms variability in timing. Because the primary goal, ascertaining whether production of F_0 contours during normal speech would be affected by perturbed auditory feedback, was supported by responses to pitch-shifted voice feedback within single or disyllabic phrases, the variability in stimulus timing does not appear to be a serious concern. However, stimulus timing variability may have contributed to the variability in the responses, and if future studies are able to deliver stimuli with greater control than was done in

the present study, results may more accurately reflect the degree to which auditory feedback is normally used to help regulate voice F_0 during speech.

With respect to our mathematical modeling, the model of Fig. 5, which serves well to simulate feedback driven modulation of sustained vowel phonations, accounts for general features of the responses during speech such as direction and approximate timing. A comparison between the simulations and experimental data in Figs. 3–4 illustrate that this model of vowel phonation fails to reproduce three aspects of speech. While it would be possible to adjust parameters of the model to fit the experimental data, we feel that it is more important to point out that these differences indicate that the auditory feedback stabilization system for speech is either separate from that used for vowel phonation, or that a common system can be rapidly reconfigured according to the demands of the task.

The first difference is that the model of vowel phonation fails to simulate the larger size of responses found under some speech conditions. In the context of the model, this would suggest that the gain of the feedback loop (incorporated in the "Filter" element in Fig. 5) is not fixed but rather can be varied, to account for this task-dependence. Response modulation has been found to occur in other systems as a function of motor activities such as walking, arm movements, cycling and grasping (Brooke *et al.*, 1991; De Serres *et al.*, 1995; Doemges and Rack, 1992; Dufresne *et al.*, 1980; Stein and Capaday, 1988; Zehr *et al.*, 2001).

The model also does not produce responses with latencies less than 130 ms, while experimentally such responses were noted. In the context of the model, reduced latencies could be obtained by reducing the matched delay elements. This again suggests that the auditory feedback stabilization mechanism can be modified substantially and rapidly depending on the specific task at hand.

Finally, the model does not produce enduring lags (delay) of perturbed vocalizations compared to control vocalizations, such as is seen in some traces of Fig. 3. Such lags presumably reflect intervention of other processes outside of the mechanisms that stabilize F_0. A possible explanation of these lags is that the rate of speech is slowed by mismatches between intended and perceived F_0. A speaker may simply be slowed down when a dissonance is detected. Another possibility is that the triggering of syllables is, to some extent, delayed by dissonance between intended and perceived F_0. This mechanism would predict an enduring lag between perturbed and control F_0 trajectories and would also account for the response durations that exceed a single syllable, as has been reported in nonsense syllables (Donath *et al.*, 2002). The observation that response magnitudes and timing varied as a function of the bi-tonal sequences supports the theory that this mechanism helps the speaker control voice F_0 across adjacent syllables. In tone languages this feature would be lexically relevant and in nontonal languages it would be important for the intonational aspects of speech production.

VI. CONCLUSION

We introduced pitch-shifted auditory feedback to native speakers of Mandarin while they were saying disyllabic sequences with different tonal combinations at a natural speaking rate. They reacted to the auditory feedback with compensatory pitch changes in most trials. The majority of the compensatory pitch changes occurred significantly sooner (143 ms) than the durations of typical Mandarin syllables (180 ms). In some conditions, latencies were short enough (<130 ms) for the response to correct for perturbations within the syllable, while in other cases latencies were too long. These findings, along with the observation that response magnitudes during tonal sequences involving dynamic F_0 trajectories were larger than sequences with a static F_0 trajectory, suggest the system may modulate the responses depending on the demand for auditory feedback to facilitate accurate and timely adjustments in F_0 control. Although the responses were both faster and larger than those during the production of sustained vowels (Burnett *et al.*, 1998; Hain *et al.*, 2000; Larson *et al.*, 2001), they are similar to those during singing (Natke *et al.*, 2003), suggesting that the production of tones requires a close control of voice F_0 similar to that used in singing. A simple mathematical model incorporating negative feedback was able to simulate the general features of the response patterns. This suggests that a control mechanism based on similar principles may be behind the observed compensatory responses. The modeling simulation also shows that both the timing and the magnitude of this control mechanism can be modified substantially and rapidly depending on the specific task at hand.

ACKNOWLEDGMENTS

This research was supported by NIH Grant No. DC006243-01A1 and NIH Grant No. DC03902. We gratefully acknowledge the helpful comments from two anonymous reviewers of a previous draft of the manuscript.

Abbs, J. H., and Gracco, V. L. (**1984**). "Control of complex motor gestures: Orofacial muscle responses to load perturbations of lip during speech," J. Neurophysiol. **51**, 705–723.

Alipour-Haghighi, F., Titze, I. R., and Durham, P. (**1987**). "Twitch response in the canine vocalis muscle," J. Speech Hear. Res. **30**, 290–294.

Brooke, J. D., Collins, D. F., Boucher, S., and McIlroy, W. E. (**1991**). "Modulation of human short latency reflexes between standing and walking," Brain Res. **548**, 172–178.

Burnett, T. A., Freedland, M. B., Larson, C. R., and Hain, T. C. (**1998**). "Voice f0 responses to manipulations in pitch feedback," J. Acoust. Soc. Am. **103**, 3153–3161.

Burnett, T. A., and Larson, C. R. (**2002**). "Early pitch shift response is active in both steady and dynamic voice pitch control," J. Acoust. Soc. Am. **112**, 1058–1063.

De Serres, S. J., Yang, J. F., and Patrick, S. K. (**1995**). "Mechanism for reflex reversal during walking in human tibialis anterior muscle revealed by single motor unit recording," J. Physiol. (London) **488**, 249–258.

Doemges, F., and Rack, P. M. H. (**1992**). "Task-dependent changes in the response of human wrist joints to mechanical disturbance," J. Physiol. (London) **447**, 575–585.

Donath, T. M., Natke, U., and Kalveram, K. T. (**2002**). "Effects of frequency-shifted auditory feedback on voice f_0 contours in syllables," J. Acoust. Soc. Am. **111**, 357–366.

Duanmu, S. (**1994**). "Syllabic weight and syllable durations: A correlation between phonology and phonetics," Phonology **11**, 1–24.

Dufresne, J. R., Soechting, J. F., and Terzuolo, C. A. (**1980**). "Modulation of the myotatic reflex gain in man during intentional movements," Brain Res. **193**, 67–84.

Gracco, V. L., and Abbs, H. H. (**1989**). "Sensorimotor characteristics of speech motor sequences," Exp. Brain Res. **75**, 586–598.

Gracco, V. L., and Abbs, J. H. (**1985**). "Dynamic control of the perioral system during speech: Kinematic analyses of autogenic and nonautogenic sensorimotor processes," J. Neurophysiol. **54**, 418–432.

Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S., and Kenney, M. K. (**2000**). "Instructing subjects to make a voluntary response reveals the presence of two components to the audio-vocal reflex," Exp. Brain Res. **130**, 133–141.

Jones, J. A., and Munhall, K. G. (**2000**). "Perceptual calibration of f0 production: Evidence from feedback perturbation," J. Acoust. Soc. Am. **108**, 1246–1251.

Jones, J. A., and Munhall, K. G. (**2002**). "The role of auditory feedback during phonation: Studies of Mandarin tone production," J. Phonetics **30**, 303–320.

Kelso, J. A. S., Tuller, B., Vatikiotis-Bateson, E., and Fowler, C. A. (**1984**). "Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures," J. Exp. Psychol. Hum. Percept. Perform. **10**, 812–832.

Kempster, G. B., Larson, C. R., and Kistler, M. K. (**1988**). "Effects of electrical stimulation of cricothyroid and thyroarytenoid muscles on voice fundamental frequency," J. Voice **2**, 221–229.

Larson, C. R., Burnett, T. A., Bauer, J. J., Kiran, S., and Hain, T. C. (**2001**). "Comparisons of voice f_0 responses to pitch-shift onset and offset conditions," J. Acoust. Soc. Am. **110**, 2845–2848.

Larson, C. R., Kempster, G. B., and Kistler, M. K. (**1987**). "Changes in voice fundamental frequency following discharge of single motor units in cricothyroid and thyroarytenoid muscles," J. Speech Hear. Res. **30**, 552–558.

Natke, U., Donath, T. M., and Kalveram, K. T. (**2003**). "Control of voice fundamental frequency in speaking versus singing," J. Acoust. Soc. Am. **113**, 1587–1593.

Natke, U., and Kalveram, K. T. (**2001**). "Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables," J. Speech Lang. Hear. Res. **44**, 577–584.

Saltzman, E., Löfqvist, A., Kay, B., Kinsella-Shaw, J., and Rubin, P. (**1998**). "Dynamics of intergestural timing: A perturbation study of lip-larynx coordination," Exp. Brain Res. **123**, 412–424.

Shaiman, S. (**1989**). "Kinematic and electromyographic responses to perturbation of the jaw," J. Acoust. Soc. Am. **86**, 78–88.

Shaiman, S., and Gracco, V. L. (**2002**). "Task-specific sensorimotor interactions in speech production," Exp. Brain Res. **146**, 411–418.

Stein, R. B., and Capaday, C. (**1988**). "The modulation of human reflexes during functional motor tasks," Trends Neurosci. **11**, 328–332.

Xu, Y. (**1997**). "Contextual tonal variations in mandarin," J. Phonetics **25**, 61–83.

Xu, Y. (**1999**). "Effects of tone and focus on the formation and alignment of f0 contours," J. Phonetics **27**, 55–105.

Xu, Y., and Wang, Q. E. (**2001**). "Pitch targets and their realization: Evidence from Mandarin Chinese," Speech Commun. **33**, 319–337.

Zehr, E. P., Hesketh, K. L., and Chua, R. (**2001**). "Differential regulation of cutaneous and *h*-reflexes during leg cycling in humans," J. Neurophysiol. **85**, 1178–1184.