A System for Online Dynamic Perturbation of Formant Trajectories and Results from Perturbations of the Mandarin Triphthong /iau/

Shanqing Cai†, Marc Boucek, Satrajit S. Ghosh, Frank H. Guenther, and Joseph S. Perkell

Speech Communication Group, Research Laboratory of Electronics, MIT, Cambridge, MA 02139, USA
†E-mail: cais@mit.edu

Abstract

Previous studies based on perturbation of formant frequencies in auditory feedback of speech (e.g., [1,2]) mainly used steady-state vowels. However, real speech is primarily time-varying. Currently little is known about the role of auditory feedback in the production of time-varying speech sounds such as diphthongs and triphthongs. We developed a system for online perturbation of such sounds. Using this system, we perturbed the formant trajectories of the triphthong /iau/ in a time-varying fashion as they were being produced by Mandarin speakers. Most subjects showed compensatory corrections to the formant trajectories in ways that reflected the time-varying nature of the perturbations. This result indicates that auditory feedback plays a role in the planning of the articulatory movements in time-varying sounds. In addition, the adaptations to these perturbations transferred to some related time-varying and steady-state vowels.

1. Introduction

Previous studies based on perturbation of vowel formant frequencies have shown that articulatory movements are planned and updated in ways that aim to maintain relatively stable auditory outcomes (e.g., [1-2]). These studies used real-time perturbations of vowel formant frequencies in speakers' auditory feedback. Speakers made compensatory adjustments to their productions in directions opposite to the perturbations, which counteracted the auditory shifts.

Such studies focused on monophthongs involving sustained values of the formant frequencies with little articulatory movement during the sound. This contrasts with the richness of dynamic articulator gestures in speech. For example, diphthongs are common in English and other languages. Triphthongs, which involve smooth movements of the articulators through the targets of three monophthongs, are abundant in languages such as Mandarin. In addition to these phonemes, supra-segmental elements such as the transitions between adjacent vowels and consonants also require movements of the articulators. Currently little is known about the role of auditory feedback in time-varying portions of speech. Studying the auditory-motor control of the time-varying sounds should shed light on the control of articulatory movements and the planning of movement sequences in speech.

Herein we present a system for real-time perturbation of formant trajectories in time-varying speech sounds. Using this system, we perturbed the F1 trajectories of the triphthong /iau/ produced by Mandarin speakers. /iau/ was chosen by virtue of its long duration and complex trajectory shape, as well as the abundance of similar time-varying vowels in Mandarin, which permitted examination of the spread of sensorimotor adaptation.

2. Materials and Methods

2.1. Subjects

The subjects of this study were 29 adult native speakers (14 females) of Mandarin Chinese with no strong regional accent, who reported no history of speech, hearing or neurological disorders. Mandarin was their first language and the primary language of instruction throughout their K-12 education. Pure-tone audiometry confirmed that all subjects had normal thresholds at 0.5, 1, and 2 kHz.

2.2. Experimental design

Similar to previous studies [1, 2], the experiment was arranged into *phases* (Fig. 1). Each phase contained a number of *blocks*. Each block contained one repetition of each of the stimulus utterances. The first half of a block contained 10 *training utterances* (Table 1, left), during which the subject heard auditory feedback through earphones. All the training utterances contained the triphthong /iau/ with the first (high-flat) tone. The second half was comprised of 10

test utterances (Table 1, right), which contained /iau/ and other vowels. When producing the test utterances, the earphones presented a speech-shaped masking noise (94 dBA SPL). The order of the utterances was randomized within each half-block.

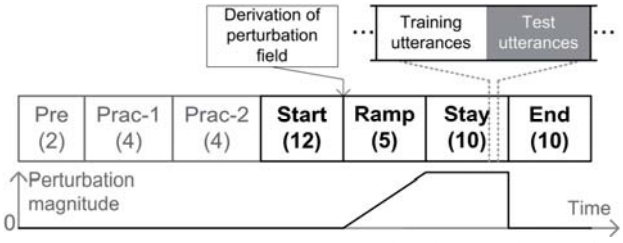

Figure 1: *Experimental protocol. The numbers of blocks in the phases are shown in the brackets. The time axis is not drawn to scale.*

In the *pre* phase, the subject was familiarized with the setup and the stimuli. In the *prac-1* and *prac-2* phases, the subject learned to produce the vowels within a range of sound level (78±4 dBA SPL) and a range of duration (350±48 ms). The subject was instructed to maintain the learned ranges of level and duration throughout the experiment. Hints were given on the screen whenever the ranges were exceeded. The *start* phase served as a no-shift baseline, on which basis a *perturbation field* was derived for the subject (see section 2.3). The perturbation is gradually ramped to maximum in the *ramp* phase, maintained through the *stay* phase, and removed at the beginning of the *end* phase (Fig. 1, bottom).

Table 1. *List of stimulus utterances*[1].
Carrier phrase: [ ]着 (/[ ] ʈʂɤ/)

Training		Test	
标 /piau$_{55}$/	漂 /ʈɕiau$_{55}$/	叼 /tiau$_{55}$/	夹 /ʈɕia$_{55}$/
飚 /piau$_{55}$/	浇 /ʈɕiau$_{55}$/	雕 /tiau$_{55}$/	包 /pau$_{55}$/
叼 /tiau$_{55}$/	挑 /tʰiau$_{55}$/	吊 /tiau$_{51}$/	乖 /kuai$_{55}$/
雕 /tiau$_{55}$/	消 /ɕiau$_{55}$/	揪 /ʈɕiou$_{55}$/	搭 /ta$_{55}$/
教 /ʈɕiau$_{55}$/	削 /ɕiau$_{55}$/	敲 /ʈɕʰiau$_{55}$/ (Repeated 2x)	

2.3. Apparatus

The apparatus designed for perturbation of the formant trajectories of time-varying oral vowels was adapted from a PC-based version of the system reported in [2]. Cepstral liftering and a formant tracker [3] were incorporated to improve formant estimation. As in [2], formant frequency shifting was based on pole-pair substitution in the z-plane. The total processing delay was 14 ms.

The subject was seated in front of a computer screen and wore a pair of insertion earphones. The voice was picked up by a microphone placed 10 cm away from the mouth. The auditory feedback through the earphones was amplified by 16.5 dB relative to the sound level at the microphone.

As exemplified in Fig. 2A, the triphthong /iau/ has a characteristic rising-falling F1 and falling F2. The formant trajectory in the F1-F2 plane has a bow shape (e.g., Fig. 2B). The triphthong /iau/ was detected automatically online based on the velocities of F1 and F2 (e.g., dashed lines in Fig. 2A).

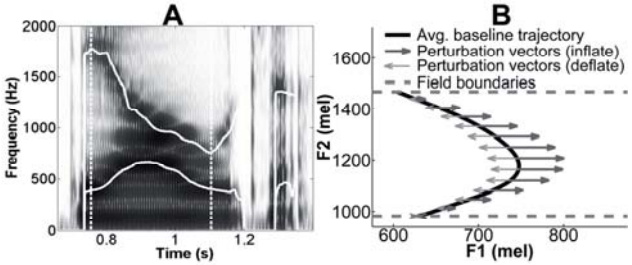

Figure 2: ***A.*** *Spectrogram of a typical training utterance with the triphthong /iau/ extracted automatically (between dashed lines).* ***B.*** *Inflate- and deflate-type perturbation fields.*

As shown in Fig. 2B, a perturbation field was derived based on the average baseline trajectory obtained during the start phase. The perturbation field spanned a region in the F1-F2 plane bounded by two constant-F2 lines, wherein the vowels were perturbed. The *perturbation vectors* (dark arrows in Fig. 2B) quantified the amounts and directions by which F1 and F2 were shifted. In the perturbation paradigms dubbed *inflate* / *deflate*, the perturbation vectors were parallel to the F1 axis, so that only F1 was perturbed. The magnitude of the field obeyed a quadratic function of F2. The magnitudes were zero at the boundaries of the field, and maximum at the center. The maximum magnitude of the vectors was determined as the range of F1 in the baseline trajectory multiplied by a fixed ratio (0.6 for inflate; 0.375 for deflate). As shown in Fig. 2A, the inflate perturbation led to increased F1 movement and exaggerated the curvature of the formant trajectory in the auditory feedback; the deflate perturbation (light arrows in Fig. 2B) led to decreased F1 movement and reduced curvature of the trajectory.

[1] The experiments in 7 of the 29 subjects used a slightly different test phrase list, in which three words /ti$_{55}$/, /tu$_{55}$/ and /tiou$_{55}$/ were used in place of the two /tiau$_{55}$/'s and one instance of /tɕʰiau$_{55}$/.

2.4. Data analysis

After the experiment, the recorded F1 and F2 tracks of the triphthongs were smoothed by 41-ms Hamming windows. The data from six subjects were rejected due to relatively large errors in the formant estimation. Recordings from the remaining 23 subjects (11 inflate, 12 deflate) were screened manually for gross formant estimation errors. Whenever possible, the errors were fixed by manually adjusting parameters of the formant tracker [3]. However, if an error could not be fixed, the utterance was discarded from further analysis.

3. Results

The response of a representative subject to an inflate perturbation is shown in Fig. 3A. Comparison between the average trajectory of /iau/ in the start phase (black) and the stay phase (gray) showed that the subject reduced the maximum F1 in the triphthong to counteract the effect of the perturbation (dashed gray). In the end phase, the subjects' production showed a pronounced aftereffect, as evident from the mean trajectory (light gray), which lay between the start- and stay-phase curves. There were decreases in F1 for all three components of /iau/. However, F1 decrease was the greatest near the center of the field, in the /a/ region, which led to a depressed curvature of the bow-shaped trajectory.

In this subject, the sensorimotor adaptation observed for the triphthong /iau/ generalized to the untrained vowels in the test utterances (Fig. 3B). There were decreases in F1 of the monophthong /a/, which was close to the second component of the triphthong. For the diphthongs /ia/ and /au/, which approximated the first and second halves of /iau/, there were decreases in the F1 ranges similar to the changes in the triphthong /iau/. More strikingly, the trajectory of the triphthong /uai/, which was roughly the temporal reverse of /iau/, also exhibited changes that closely resembled the adaptation for /iau/. However, the trajectory of the triphthong /iou/, which lay farthest from that of /iau/ in our test utterances, showed relatively minor changes.

Figure 3C shows another subject's response under the deflate paradigm. The subject dramatically increased the peak F1 and amplified the curvature of the trajectory of /iau/ in the stay phase to counteract the effect of the deflate perturbation. The F1 values at the onset and end of the triphthong were essentially unaltered. As in the previous example, there were pronounced aftereffects in the end phase. The adaptation for /iau/ also generalized to other vowels (Fig. 3D). With the exception of /iou/, there were marked increases in the F1 values near the /a/-region of the time-varying and steady-state vowels, which were very similar to the compensation for /iau/.

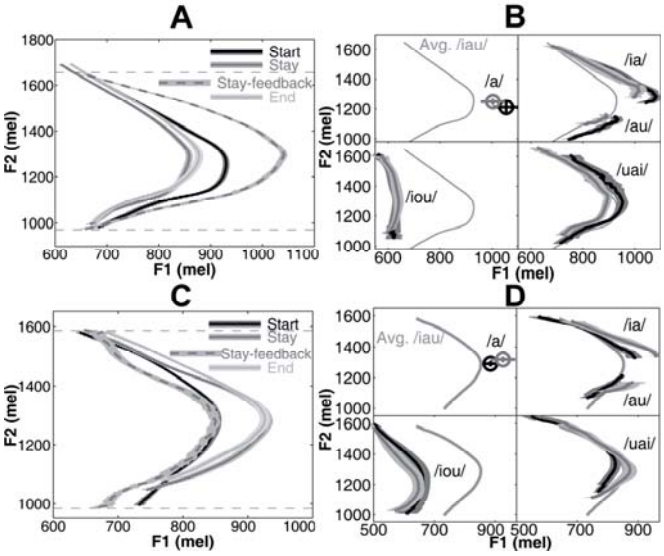

Figure 3: ***A.*** *Average formant trajectories from a representative inflate experiment. Shadings show ±2 SEM.* ***B.*** *Transfer of the adaptation to the test utterances. Average trajectory of /iau/ is plotted in each panel for reference.* ***C, D.*** *Examples from a representative deflate experiment.*

Nine of the 11 subjects in the inflate group showed compensatory responses similar to that shown in Fig. 3A. The other two subjects changed their /iau/ trajectories in the stay phase in the same direction as the auditory perturbations. Similarly, 9 of the 12 subjects in the deflate group showed compensatory responses, while 3 others didn't. Fig. 4 gives a population summary of the time course of the changes in the peak F1 of /iau/ in the training utterances. The F1 values shown in this figure were normalized by the start-phase mean of each subject, and then averaged across the subjects in each group. In both groups, there were significant changes in the peak F1 in the stay phases opposite to the F1 changes in the perturbations; these effects lasted into the end phase and gradually decayed toward baseline.

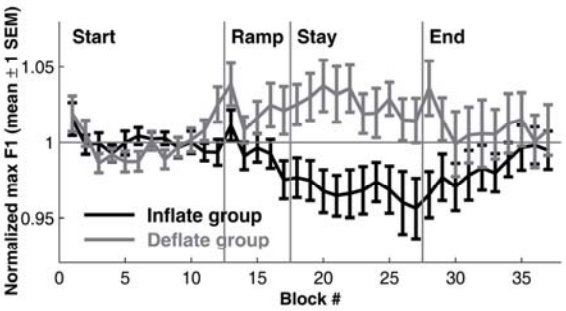

Figure 4: *Population average time courses of the maximum F1 in the triphthong*[2].

A three-way ANOVA (subject×phase×utterance, with phase containing two levels {start, stay}) indicated a significant main effect by phase on peak F1 of the triphthong /iau/ in the training utterances ($p<1\times10^{-10}$ for both groups). Post hoc Mann-Whitney test between the start and stay phases confirmed this significant changes in peak F1 ($p<1\times10^{-6}$ in both groups). Significant changes were also found for the F1 range during the triphthong ($p<1\times10^{-4}$ for both groups, post hoc Mann-Whitney test). By contrast, there were no significant change in the F1 values at the upper or lower boundaries of the perturbation field, except for a decrease in the F1 at the upper boundary in the deflate group only ($p<1\times10^{-5}$, post hoc Mann-Whitney test). These results indicate that the compensatory responses were achieved primarily from alterations of the trajectory shapes, instead of pure shifting of the trajectories in the F1-F2 plane.

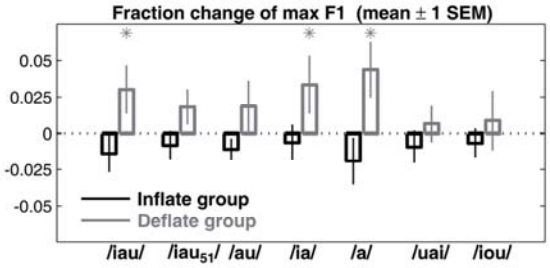

Figure 5: *Fraction change of maximum F1 of the test-utterance vowels in the stay phase relative to start phase. Population averages of the two groups are shown. Asterisks indicate significant changes (p<0.01, post hoc Mann-Whitney test).*

[2] The gradual decreases in F1 in the first 2-3 blocks of the start phase may reflect fatigue of the jaw lowering system; the sudden increases in F1 at the beginning of the ramp phase could be due to the short break given to each subject between the start and stay phases, which relieved the fatigue. But these observations do not alter the main conclusion drawn from this figure.

For the test utterances, at the population level, there were consistent trends for the peak F1s to show stay-phase increase or decrease in the deflate and inflate groups, respectively (Fig. 5). A three-way ANOVA (subject×phase×vowel) indicated significant main effects by phase ($p<1\times10^{-6}$) in both groups.

4. Discussion

In this study, we elicited adaptive changes in the formant trajectory of the triphthong /iau/ by imposing time-varying perturbations. In light of previous studies [1-2], this observation shows that auditory feedback plays roles not only in steady-state vowels, but also in the articulatory planning required for the production of vowels with time-varying formants. However, these compensatory responses to the time-varying perturbations (Fig. 4) were smaller compared to those seen previously for steady-state vowels. Possible explanations for this include the transient nature of the auditory error and certain somatosensory afferent inputs in time-varying vowels. But detailed discussion of this issue is beyond the scope of the current paper. Another interesting finding of this study was that the sensorimotor adaptation trained on /iau/ transferred to untrained vowels with trajectories overlapping with that of /iau/, regardless of the whether they are steady-state or time-varying. It can be inferred from this result that these two categories of vowels share some important aspects of the phonetic-to-articulatory mapping.

Acknowledgements

Supported by NIH grant DC0001925 and a graduate fellowship from MIT to the first author.

References

[1] J. F. Houde M. I. and Jordan. Sensorimotor adaptation of speech I. Compensation and adaptation. *J. Speech Lang. Hear. Res.*, 45:295-310, 2002.

[2] V. M. Villacorta, J. S. Perkell, and F. H. Guenther. Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *J. Acoust. Soc. Am.*, 122(4):2306-2319, 2007.

[3] K. Xia and C. Espy-Wilson. "A new strategy of formant tracking based on dynamic programming." In *ICSLP2000*, Beijing, China, October 2000.