# SIMULATING ONLINE COMPENSATION FOR PITCH-SHIFTED AUDITORY FEEDBACK WITH TARGET APPROXIMATION MODEL

*Hao Liu, Yi Xu*

Department of Speech, Hearing and Phonetic Sciences, University College London, UK
{h.liu.12, yi.xu}@ucl.ac.uk

## ABSTRACT

This study attempts to achieve modeling simulation of the well-known phenomenon of online compensation for pitch-shifted auditory feedback. We used the Target Approximation (TA) model as the underlying kinematic mechanism of pitch contour generation, and simulated feedback compensation through responsive perturbation of the height parameter of the TA model. Results show that both within-syllable and cross-syllable pitch compensation in disyllabic utterances can be replicated. Furthermore, our data analysis also revealed an over-rectification phenomenon. By adjusting the height parameter back and beyond its original value after the compensation, the over-rectification was also replicated, further improving the overall simulation results.

**Keywords:** speech production, pitch modeling, target approximation, speech motor control, online auditory feedback compensation

## 1. INTRODUCTION

In the past two decades, a number of studies have shown that auditory feedback plays an important role in controlling the fundamental frequency (F0) of voice. These studies investigated feedback response in various tasks, including singing [17], glissando [2], sustained vowels [1, 7, 11, 24], prolonged vowels [10], nonsense syllables [5, 18, 19] and normal speech [4, 28]. In the experiments of these studies, unexpected pitch shifts were applied *online* to the voice of the human subjects before being fed back to their ears. In response to the pitch shifts, subjects automatically made compensatory F0 adjustments (in the opposite direction of pitch-shift) with short latencies (100-150 ms on average [12, 28]). Meanwhile, other studies found similar feedback compensation in formants [8, 9, 14, 16, 22, 23, 25, 27]. Despite the extensive empirical studies, the underlying mechanism of this phenomenon has not yet been investigated sufficiently. A recent study [3] has computationally modeled such online auditory feedback compensation for formants with a modified DIVA model [6] and compared discrepancies during the modeling process between normal and stuttering speakers. To our knowledge, however, there is a lack of similar research on pitch control.

In the present study, we first conducted an empirical experiment to apply real-time pitch shift to auditory feedback to human subjects. We analyzed the collected behavioral data to verify previously reported feedback compensation. We then tried to simulate the behavioral data with a virtual agent built around an articulatory-based pitch controller — the Target Approximation (TA) model [21, 30, 31]. Our goal was to simulate feedback compensation in a way that is biomechanically plausible, so that it is also general enough to be extendable to other areas of motor control.

## 2. BEHAVIORAL DATA

### 2.1. Subjects

Eight subjects (six females and two males; age 22-27) speaking Beijing dialect of Mandarin Chinese participated in the experiment. All subjects passed hearing test and none of them reported history of neurological or speech disorders.

**Table 1:** Four disyllabic Chinese phrases used as stimuli. The first syllables are all in the High tone, whereas the second syllables are in the High, Rising, Low and Falling tones, respectively.

| Phrase | Pronunciation | Pattern |
|---|---|---|
| 妈妈 | /māmā/ | H-H |
| 妈麻 | /māmá/ | H-R |
| 妈马 | /māmǎ/ | H-L |
| 妈骂 | /māmà/ | H-F |

### 2.2. Procedure

Subjects were seated comfortably in a recording booth at UCL and asked to read aloud a full list of 300 Chinese disyllabic phrases displayed on a screen in front of them. The full phrase list consists of 4 phrases (Table 1) repeated 75 times (the order of the

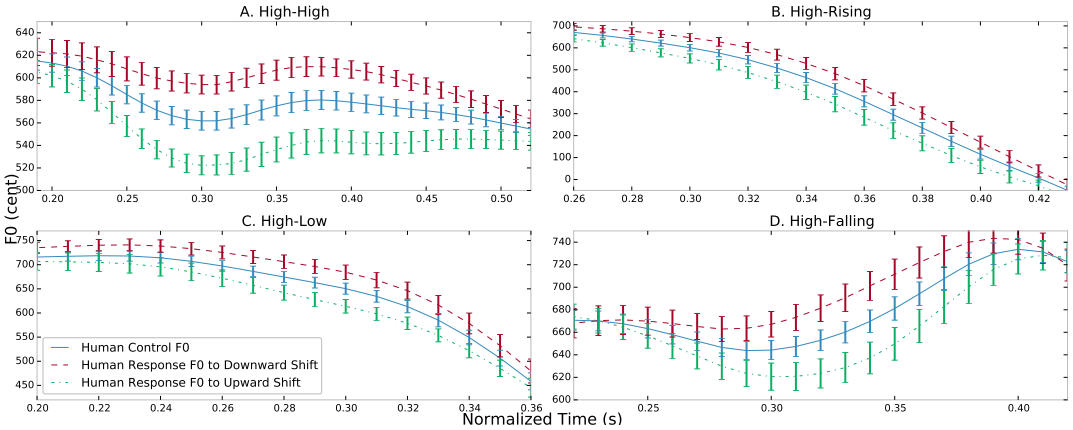


**Figure 1:** Productions of a female speaker for the four disyllabic Mandarin Chinese phrases with pitch shift starting from 100 ms after the detected vocalization onset. In each panel, the solid blue line denotes averaged control F0 signal (produced with no pitch shift in auditory feedback), the dashed red line denotes averaged response F0 signal with 200-cent downward pitch shift in the auditory feedback, and the dotted green line denotes averaged response F0 signal with 200-cent upward pitch shift. SEMs are indicated as error bars.

4 phrases was shuffled in each repetition).

Subjects were trained to maintain their speech rate at 0.5 second per phrase as steadily as possible. This is to make sure that the pitch shift applied during the experiment occur at roughly the same time relative to the whole utterance. Before the experiment, subjects were asked to practice reading the above-mentioned stimuli aloud at about 70 dB SPL (sound pressure level). Loudness was measured by a Brüel & Kjær 2203 sound level meter.

### 2.3. Apparatus and method

The speech signals were first captured by a Countryman ISOMAX headset microphone and transmitted to a homemade real-time pitch shifter, *FxTuner*, which applies a random pitch shift. The software relies on the PortAudio C/C++ library and the CoreAudio driver on MacOS for low-latency playback and applies Short-time Fourier Transformation (STFT) for fast and accurate pitch shifting. Processed speech signals were instantly delivered back to both ears via a Beyerdynamic DT231 PRO headphone, with an added masking pink noise at 40 dB.

*FxTuner* has an overall latency of around 12 ms for the whole process of voice "capture-manipulation-playback". This latency generally satisfies the requirement of "imperceptible" temporal distortion and caused little distraction to speakers. We also considered time-domain pitch manipulation techniques like TD-PSOLA [15] and WSOLA [26]. However, they generally require two pitch periods (20 ms at least for male voice) to achieve good performance, which is too slow for our purpose.

During the experiment, auditory feedback signals were randomly pitch-shifted upward or downward by 2 semitones (200 cents) for 200 ms, or left unchanged. The start of the pitch shift was either 100 ms or 250 ms after the detection of vocalization onset.

### 2.4. Behavioral results

The produced F0 signals were sampled at 100 Hz and transformed from Hertz to cent ($cents = 100 \times (39.86 \times \lg(f_0/195.997))$ where $f_0$ equals F0 in Hertz [12, 28]). Similar to the findings of previous studies, under the condition of pitch shift, not all productions were compensatory, as there were a small number of nonresponses and following responses (i.e., the reactive pitch change followed the direction of the feedback pitch shift). Because the handling of those cases is beyond the scope of this study, only compensatory responses to the stimuli were categorized, averaged and analyzed.

The results showed similar patterns as those reported in previous studies. A representative case is shown in Fig. 1, where the production data displayed are from a female subject with pitch shifted both upward and downward in the stimuli for the four bi-tonal patterns (H-H, H-R, H-L and H-F). For quantitative measurement, point-by-point serial t-tests were run between averaged control and response signals [28]. The robustness of the compensation is demonstrated by the means and standard errors plotted in the graph. These behavioral patterns were what would be modeled in the simulation task discussed in the next section.

## 3. SIMULATION

There was a previously published mathematical model of pitch stabilization using negative feedback and delays for sustained vowels [7], which was also later used for normal speech [28]. However, that model lacked a critical F0 production module. It simply used control F0 signal as input and filtered it afterwards in response to perturbations found in the feedback. Hence, the underlying mechanism of F0 production and how it reacts to pitch shifted feedback remains unclear.

### 3.1. Assumptions

Our basic assumption for the simulation is that feedback compensation has to happen as part of the basic F0 production mechanism. For this mechanism, we adopted the recently developed target approximation model as shown in Fig. 2.

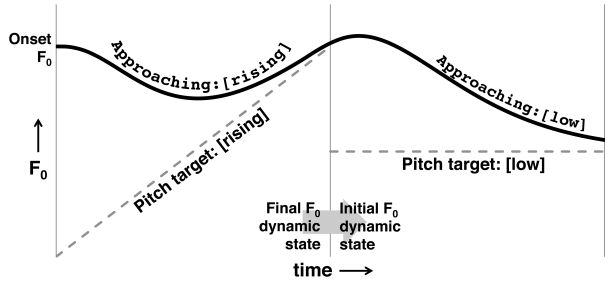


**Figure 2:** Target Approximation model.

The Target Approximation (TA) model assumes that continuous surface F0 contours are the results of successive, yet non-overlapping articulatory movements, each approaching an underlying target associated with a local host syllable [31]. A target can be either static or dynamic (Fig. 2), which can be represented by a simple linear equation:

$$(1) \quad x(t) = mt + b,$$

where $m$ and $b$ represent the spatial properties of the target in terms of target height and slope, respectively. And $t$ is time relative to the onset of the host syllable.

The quantitative implementation of TA model developed by Prom-on and Xu [21] is a third-order critically damped linear system as represented by the following equation

$$(2) \quad f_0(t) = x(t) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t},$$

where $f_0(t)$ is the complete form of the fundamental frequency in semitones, $x(t)$ is the forced response and the polynomial and the exponential are the natural response. $\lambda$ is the rate of target approximation, i.e., how rapidly the target is approached, which indicates the strength of target approximation movement. The transient coefficients $c_1$, $c_2$ and $c_3$ are jointly determined by the initial F0 dynamic state of the syllable, consisting of F0 level, velocity, and acceleration transferred from the offset of the preceding syllable:

$$(3) \quad c_1 = f_0(0) - b,$$

$$(4) \quad c_2 = f_0'(0) + c_1\lambda - m,$$

$$(5) \quad c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2.$$

At the end of the syllable, the final F0 dynamic state is transferred to the next syllable to become its initial state, which results in a smooth and continuous F0 trajectory across the syllable boundary (Fig. 2). TA is a feedforward model as there is no error checking during the approximation of the target. Thus a feedback behavior has to be simulated by an extra mechanism added on top of the TA-controlled F0 generation process.

Our specific assumption for the simulation is this: The surface compensatory responses observed in behavioral data is a result of temporary *adjustment* of underlying articulatory pitch targets during speech production in reaction to the pitch-shifted feedback. So there should be a momentary alternation of the originally planned target during the ongoing target approximation process. Following the behavioral data, the adjustment of pitch target should have a short latency and a weak amplitude. According to previous research, among the three TA target parameters, variation of target height affects surface contour formation the most [13]. So in the current simulation, target height was chosen as the model parameter to be adjusted.

### 3.2. Method

To model the behavioral data, we first created a virtual agent, built around the qTA model [21]. Before simulation, the agent was first trained using an exhaustive search to find optimal target parameters by fitting the averaged normal production of each tonal sequence by each human subject [29].

During simulation, three variables controlling the compensation were explored: timing of compensatory onset, duration of compensation and scale of compensation. These variables were generally set free within a relatively loose range when the virtual agent searched for the best fit to the behavioral data. For example, as previously reported, compensation normally starts 100-150 ms after vocalization

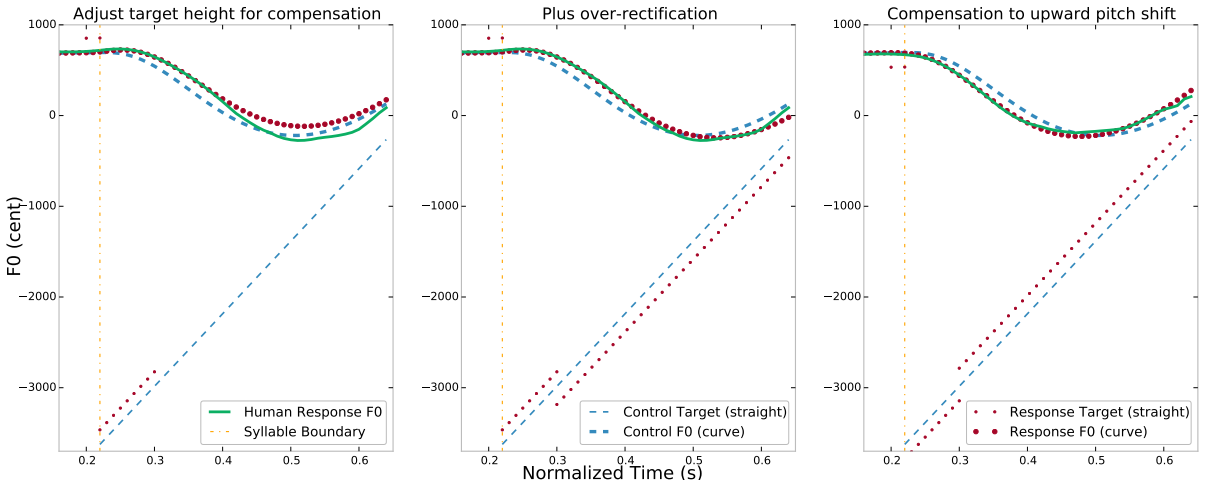


**Figure 3:** Simulated cross-syllable online compensation for the High-Rising phrase. The left panel displays an example when pitch target height was compensated (higher than normal) for 100 ms in the opposite direction of a downward pitch shift at 100 ms after detected vocalization onset. The middle panel displays a case improved by over-rectification after compensation. And the right panel displays a similar case of over-rectification in response to an upward pitch shift.

and only compensations occurring >60 ms after vocalization are considered as genuine. So, the values of the compensation timing variable were drawn from a discrete uniform distribution with 60 and 200 as bounds and 10 ms as intervals during optimization. The optimization process was automated by exhaustively exploring all possible combinations of the three variables to fit the natural pitch contours containing feedback compensation.

The initial results showed that this simulation strategy was not enough to obtain fully satisfactory results due to large discrepancies observed between the simulated and the natural contours in the *post-compensation* time interval. It seemed that after the compensation, the simulated F0 could not replicate the quick return in the natural contour. Therefore, we released the target height parameter in post-compensation production as a new variable to be explored. This change resulted in further improvement in the simulation.

### 3.3. Results

Representative outputs of the simulation are shown in Fig. 3. While compensatory adjustment of pitch target height alone did offer a better fit to the original than the no-adjustment control, it failed to simulate the *over-rectification* in post-compensation extensively observed in the natural contours (left panel). Adding post-compensation target adjustment further improved fitting (middle and right panels). Across subjects, the averaged root mean square error (RMSE) was 122.01 in cents with Pearson's $r = 0.96552$ for productions between natural and synthetic F0 without compensation; the averaged RMSE was 106.67 with $r = 0.96805$ with compensatory target height adjustment; and the averaged RMSE was 42.53 with $r = 0.98119$ with both compensatory and post-compensation adjustment.

## 4. DISCUSSION AND CONCLUSION

Online compensation of pitch-shifted auditory feedback plays an important role in both normal speech production and childhood speech acquisition [20]. Computational simulation could help to achieve an understanding of how such online compensation works. In this study we developed a virtual agent built around the TA model of dynamic pitch control in speech production [21]. The virtual agent simulates feedback compensation by finding optimal adjustments to the original TA-based pitch targets learned from normal production data. We found that the best simulation results were obtained when the agent applied both on-compensation target adjustment and post-compensation target over-rectification. These findings have demonstrated the effectiveness of this compositional approach in simulating detailed dynamic pitch control, which could be linked to neural control mechanisms in the brain in future research [20]. They have also provided further support for conceiving the basic human speech action as a dynamic process of target approximation [21, 31].

## 5. REFERENCES


[1] Bauer, J. J., Larson, C. R. 2003. Audio-vocal responses to repetitive pitch-shift stimulation during a sustained vocalization: Improvements in methodology for the pitch-shifting technique. *J. Acoust. Soc. Am.* 114, 1048–1054.

[2] Burnett, T. A., 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.

[3] Cai, S., Ghosh, S. S., Guenther, F. H., Perkell, J. S. 2011. Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing. *J. Neurosci.* 31, 16483–16490.

[4] Chen, S. H., Liu, H., Xu, Y., Larson, C. R. 2007. Voice F0 responses to pitch-shifted voice feedback during English speech. *J. Acoust. Soc. Am.* 121, 1157–1163.

[5] Donath, T. M., Natke, U., Kalveram, K. T. 2002. Effects of frequency-shifted auditory feedback on voice F0 contours in syllables. *J. Acoust. Soc. Am.* 111, 357–366.

[6] Guenther, F. H., Ghosh, S. S., Tourville, J. A. 2006. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language* 96, 280–301.

[7] Hain, T. C., Burnett, T. A., Kiran, S., Larson, C. R., Singh, S., 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.

[8] Houde, J. F., Jordan, M. I. 1998. Sensorimotor adaptation in speech production. *Science* 279, 1213–1216.

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

[10] Jones, J. A., Munhall, K. 2002. The role of auditory feedback during phonation: studies of Mandarin tone production. *Journal of Phonetics* 30, 303–320.

[11] Larson, C. R., Burnett, T. A., Bauer, J. J., Kiran, S., Hain, T. C. 2001. Comparison of voice F0 responses to pitch-shift onset and offset conditions. *J. Acoust. Soc. Am.* 110, 2845–2848.

[12] Liu, H., Larson, C. 2007. Effects of perturbation magnitude and voice F0 level on the pitch-shift reflex. *J. Acoust. Soc. Am.* 122, 3671–3677.

[13] Liu, H., Xu, Y. 2014. A simplified method of learning underlying articulatory pitch target. *Proc. of Speech Prosody* 1017–1021.

[14] MacDonald, E. N., Goldberg, R., Munhall, K. G. 2010. Compensations in response to real-time formant perturbations of different magnitudes. *J. Acoust. Soc. Am.* 127, 1059–1068.

[15] Moulines, E., Charpentier, F. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9, 453–467.

[16] Munhall, K. G., MacDonald, E. N., Byrne, S. K., Johnsrude, I. 2009. Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *J. Acoust. Soc. Am.* 125, 384–390.

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

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

[19] Natke, U., 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.

[20] Perkell, J. S. 2012. Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics* 25, 382–407.

[21] Prom-on, S., Xu, Y., Thipakorn, B. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *J. Acoust. Soc. Am.* 125, 405–424.

[22] Purcell, D. W., Munhall, K. G. 2006. Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *J. Acoust. Soc. Am.* 120, 966–977.

[23] Purcell, D. W., Munhall, K. G. 2006. Compensation following real-time manipulation of formants in isolated vowels. *J. Acoust. Soc. Am.* 119, 2288–2297.

[24] Sivasankar, M., Bauer, J. J., Babu, T., Larson, C. R. 2005. Voice responses to changes in pitch of voice or tone auditory feedback. *J. Acoust. Soc. Am.* 117, 850–857.

[25] Tourville, J. A., Reilly, K. J., Guenther, F. H. 2008. Neural mechanisms underlying auditory feedback control of speech. *Neuroimage* 39, 1429–1443.

[26] Verhelst, W., Roelands, M. 1993. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. *Proc. of ICASSP* volume 2. IEEE 554–557.

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

[28] Xu, Y., Larson, C. R., Bauer, J. J., Hain, T. C. 2004. Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *J. Acoust. Soc. Am.* 116, 1168–1178.

[29] Xu, Y., Prom-on, S. 2010–2012. PENTAtrainer1.praat. Available from: http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer1/.

[30] Xu, Y., Prom-on, S. 2014. Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication* 57, 181–208.

[31] Xu, Y., Wang, Q. E. 2001. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication* 33, 319–337.