PERCEPTION AND IMITATION OF PERIOD DOUBLING

Yaqian Huang

University of California San Diego; Acoustics Research Institute, Austrian Academy of Sciences
yqhuang1016@gmail.com

ABSTRACT

Period doubling consists of voicing periods alternating in amplitude and/or frequency and is often perceived as rough sounding and with an indeterminate pitch. Studies have found that pitch during period doubling is perceived as lower as the degree of amplitude/frequency modulation increased [1, 2]. However, it is unclear how listeners perceive period doubling when identifying linguistic tones. In an artificial language learning study, we test tonal identification and imitation using resynthesized period-doubled tones in native Mandarin and English speakers. We show that period doubling drives a low-tone percept, especially as the modulation, particularly that of frequency, increases. However, higher f0 (300Hz) led to more low-tone responses in amplitude-modulated tokens than lower f0 (200Hz). Period doubling is also imitated with lowered f0. Both Mandarin and English listeners behaved similarly, suggesting that period doubling is not perceived language-specifically. Period doubling is predicted to signal low tones, even when the f0 is high.

Keywords: f0, period doubling, pitch, tone, voice

1. INTRODUCTION

Typical modal voice possesses a single f0 as the primary correlate of pitch. However, when voicing is irregular such that multiple periods can be identified, determining the f0, and consequently the pitch, becomes problematic. This is the case with period doubling, a type of voice that carries at least two simultaneous periodicities, with an alternation in amplitude and/or frequency. This leads to an indeterminate pitch with low and rough quality [3, 4, 5]. It is considered to be a special case of "multiply pulsing", a subtype of creaky voice [3], commonly observed in ~25% of normal speakers' utterances [6].

How is pitch perceived during this voice given the presence of multiple frequencies and periods? Past pitch-matching studies in period doubling found that the perceived pitch became lower as the degree of amplitude and frequency modulation between the two alternating periods increased. The perceptual outcome also differed across f0s and modulation types: a lower f0 facilitated identification of a lower pitch, and the matched pitch dropped more quickly in frequency- than amplitude-modulated tokens [1, 2]. However, it is still unclear how period doubling is used to identify linguistic tones. For example, Mandarin tones are often realized with creaky voice, itself manifested as vocal fry and period doubling [4, 7]. But Huang [7] found that resynthesized tones with period doubling created by the "double pulsing" parameter in the Klatt synthesizer had a negative impact on tonal identification in Mandarin, even with the frequently-creaky dipping Tone 3. Thus, period doubling could hinder rather than facilitate tone identification. It might be perceived as roughness or as competing pitches that disrupt tone perception.

In this study, using an artificial language learning and shadowing paradigm with implicit categories of 'high' and 'low' tones, we test both English and Mandarin listeners' ability to perceive and imitate tonal stimuli manipulated with period doubling. We show that regardless of language background, higher modulation degrees and frequency modulation bias listeners to hear a low tone more frequently; and period doubling is often imitated with lowered f0 and creaky voice quality. Also contrary to past findings, a higher stimulus f0 leads to a low-tone percept in more amplitude-modulated tokens than a lower f0. We further discuss its implications for tonal perception.

2. METHODS

2.1. Stimuli

The stimuli consisted of resynthesized tokens of a vowel [a]. One pulse of the vowel was extracted from a region of stable formants. All stimuli contain a basis formed by duplicating and/or manipulating this extracted pulse according to the experimental condition (illustrated in Fig. 1). To reproduce the characteristics of period doubling, three experimental conditions were created based on the empirical ratios calculated from electroglottography in a scripted Mandarin corpus [7]: amplitude modulation, frequency modulation, and combined amplitude and frequency modulations of every other cycle. The resulting stimuli have alternating pulses of "long-short-long" periods and/or "high-low-high" amplitudes (as described in [8, 9]). All stimuli were 300 ms long and scaled to 70 dB. The following steps detail the process of stimuli creation, achieved using a custom PRAAT script [10].

- Non-modulated: duplicate and concatenate the extracted [a] pulse. (Fig. 1a)
- Amplitude modulation (*am*): the amplitude of the first pulse (a1) is retained, and that of the second pulse (a2) is reduced based on amplitude ratio a1/a2. (Fig. 1b)
- Frequency modulation (*fm*): the duration of the first pulse (d1) lengthens and that of the second pulse (d2) shortens based on frequency ratio d1/d2, while maintaining a fixed presumptive f0 given by 2/(d1+d2). (Fig. 1c)
- Combined amplitude and frequency modulation: the first pulse lengthens, and the second pulse both shrinks and lowers in amplitude. (Fig. 1d)

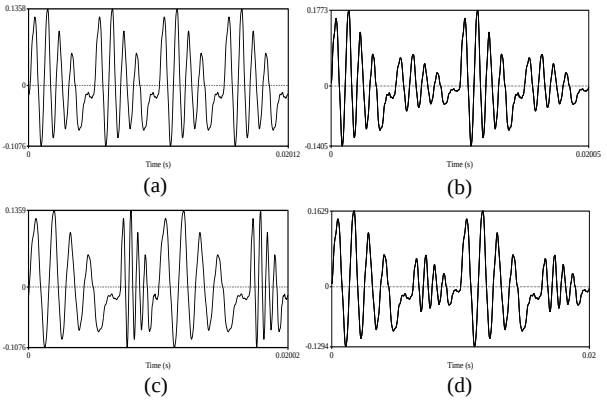

Figure 1: Resynthesized period-doubled pulses of 200 Hz: original unmodified token (a), amplitude-modulated at ratio 2.4 (b), frequency-modulated at 2.4 (c), and amplitude- (2.4) plus frequency-modulated (2.4) (d).

The range of ratios covered at least two standard deviations from the median of either distribution (*am*: 1.43+0.73*2=2.89; *fm*: 1.38+0.56*2=2.5). Table 1 shows the specifics adopted to create the stimuli. Non-modulated stimuli have a ratio of 1, meaning identical amplitudes and/or frequencies, expected to create a modal-sounding tone. Extreme values (4 for *am*; 3 for *fm*) were included to anchor the other end of the distributions to ensure period doubling. We expect to see that the perception changes along with the degrees of modulation from modal (a single pitch) to period doubling (multiple pitches). When the ratio increases, the modulation extent increases, and an increasingly stronger percept of the lower pitch (half of the original one) is expected, because the two alternating pulses become more and more distinct to facilitate the lower pitch percept cued by the lower f0, given by 1/(d1+d2).

	Ratio	Step	Total
Amplitude ratio: a1/a2	Range: (1, 3); Extreme: 4	0.2	11
Frequency ratio: d1/d2	Range: (1, 2.6); Extreme: 3	0.1	17

Table 1: Amplitude and frequency ratios used for modulation based on empirical data.

The pitch of the training stimuli was manipulated using the overlap-add method in PSOLA through the *Manipulation* function in PRAAT. Depending on the experimental condition, 40 training tokens were generated from a gaussian distribution around 200 Hz and 100 Hz (=200/2), or 300 Hz and 150 Hz (=300/2). The standard deviations were chosen considering the non-linearity of pitch perception after converting hertz to semitones to better simulate a comparable distance between different pitches by the human ear.

2.2. Participants and procedure

Thirty native Mandarin speakers (18F, mean age = 20.43), and thirty-one native English speakers (22F, mean age = 20.26) were recruited from the undergraduate population at an institution. No hearing or language disorders were reported. Recordings from twenty-five Mandarin and twenty-nine English speakers were analyzed for the shadowing experiment, due to equipment failure or participants' error during the experiment.

Both perception and shadowing experiments were implemented in PsychoPy [11] in a sound-attenuated booth. Stimuli were played from Focusrite Scarlett 8i6 pre-amplifier connected to a lab computer and participants used headphones throughout. To begin with, participants were told they were learning a novel language which has two tonal categories, represented iconically as ↑ and ↓. The participants went through a familiarization phase where they heard 40 modal tones of a single pitch between half of the f0 and the f0 of the experimental stimuli. They pressed an up or down arrow key upon hearing the sound to get familiarized with the corresponding category ↑ or ↓. Then they were tested in a training phase to categorize the aforementioned 40 tokens, with feedback provided. They were required to pass the training phase with a cumulative accuracy at the minimum of 75% after categorizing at least two rounds of the training tokens (80 tokens). Then they proceeded to the two repetition test blocks to categorize 380 [= (11 *am_degree* x 17 *fm_degree* + 3 extreme) x 2] resynthesized tokens of period doubling into the same categories. The two categories were always given by symbols without explicit explanation. Order of stimulus presentation was fully randomized across listeners and phases. The entire experiment was strictly timed, and participants were warned if they responded too slow (after 1.3 seconds upon hearing the audio). Two misses on the same token were counted as an invalid response. Between the two test blocks, participants were prompted to take a break. The entire perception experiment lasted approximately 20 minutes.

Immediately after the perception experiment, participants were asked to participate in the shadowing experiment. They were recorded using a desk-mounted microphone, with the headphones on. They were asked to reproduce the period-doubled tones (same test stimuli as those in the perception experiment) heard three times by imitating the pitch and voice of the stimuli. They were also allowed to play the audio as many times as needed and were encouraged to practice as many times as needed before they were ready to produce the sounds. Prior to the experiment, they were instructed to produce tone sweeps [12] to assess their vocal range to compare to the pitch imitation of period doubling.

2.3. Analysis

The mean f0s and voice quality correlates of the shadowing productions were extracted using VoiceSauce [13]. Each individual's vocal range was assessed by extracting the max and min f0s in their tone sweeps and coded as covariates included in the statistical models for predicting mean imitated f0s. Voice quality correlates include H1–H2, harmonics-to-noise ratio < 500 Hz (HNR), subharmonic-to-harmonic ratio (SHR), and strength of excitation (SoE). In general, a lower H1–H2 indicates a higher degree of glottal constriction; a lower HNR indicates a noisier quality; a higher SHR indicates stronger subharmonics; and a lower SoE indicates lower energy and possibly more constriction. All these measures in the explained direction would signal a creakier quality [14, 15, 16].

For both perception and production data, logistic and linear mixed-effects models were used to predict binomial categorization [↑ ('up') or ↓ ('down')], imitated f0 and voice quality correlates, given *manipulation type* (no, amplitude, frequency, combined modulation), *f0 condition* (200, 300 Hz), their interactions, *language* (Mandarin, English) and random intercepts of *subject* and *repetition* of the stimuli (first, second) if applicable. The baselines used for within-factor comparison were unmodulated, 200 Hz, and 'up' responses. For voice quality models, imitated f0 was included as a covariate. Further, we included the prior *categorization* as a predictor to probe whether the productions are correlated with the perceptual results.

Data were trimmed to remove outliers determined by log-transformed f0 or reaction time, or acoustic measures that are larger than 2.5 standard deviations from the mean (2.5% of the original data). Following Yuan and Liberman [17], for better interpretation, the imitated f0 was converted to semitones ($= 12 * log_2\,(f0/f0_{base})$), using a speaker-dependent f0 base: 5th percentile of all f0 values per speaker.

3. RESULTS

3.1. Perception during period doubling

No effect of language was found. Participants' categorization varied as a function of modulation type [$\chi^2(3)=1786.61$, $p<.001$], f0 [$\chi^2(1)=6.44$, $p<.05$], and their interactions [$\chi^2(3)=55.24$, $p<.001$]. Frequency modulation had a stronger effect ($\beta=5.00$, $p<.001$) than amplitude modulation ($\beta=2.21$, $p<.05$). The combined frequency and amplitude modulation had the strongest effect ($\beta=5.32$, $p<.001$), showing an additive effect. Pairwise comparisons confirmed this order of effect strength: combined modulation > frequency modulation > amplitude modulation > no modulation (Fig. 2). A linear trend can be observed in both amplitude and frequency modulation: higher modulation degrees lead to a larger proportion of low-tone responses (Fig. 3). Further, both Figs. 2 and 3 show the significant interaction between modulation type and f0 was driven by the increased proportion of 'down' responses for tokens with higher f0 (300 Hz), specifically for amplitude-modulated tokens.

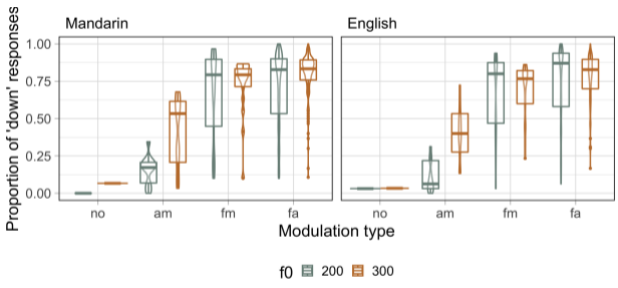

Figure 2: Proportion of 'down' responses as a function of modulation types in Mandarin and English listeners. 'no': unmodulated; 'am': amplitude modulation; 'fm': frequency modulation; 'fa': combined frequency and amplitude modulation.

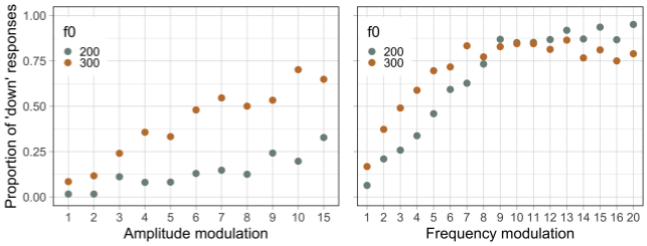

Figure 3: Proportion of 'down' responses vary by amplitude (left) and frequency (right) modulation degrees in purely amplitude- or frequency-modulated tokens.

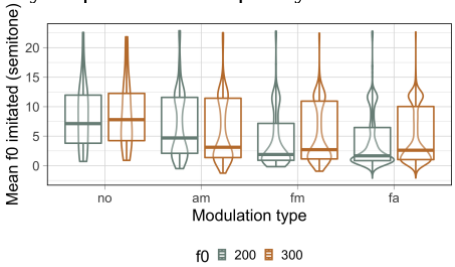

Figure 4: Imitated mean f0 (semitone) as a function of modulation types.

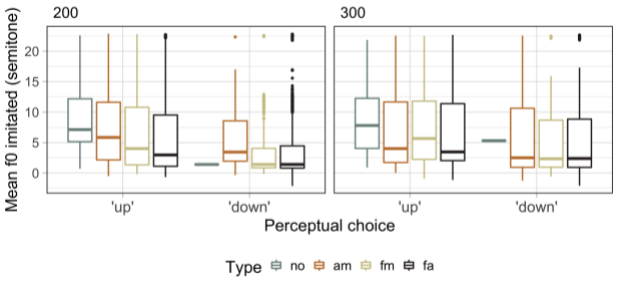

Figure 5: Imitated mean f0 (semitone) in different f0 conditions as a function of perceptual responses. There were only 49 unmodulated tokens ('no'), and those categorized as 'down' were very few.

3.2. Shadowing during period doubling

Modulation type and the interaction between modulation type and f0 condition were significant in predicting mean f0 (p<.001). In general, amplitude and frequency modulations were associated with lower imitated f0, and combined modulation had an additive effect. With this main lowering effect of modulation, imitation of amplitude-modulated tokens of 300 Hz tended to be even lower than 200 Hz, similar to the perceptual results (Fig. 4).

3.3. Perception-production link

The imitated f0 averaged across three repeated tokens produced by participants can be predicted using their own *categorization* response in perception. The main effect of *categorization* [F(1)=571.83, p<.001], and its interactions with modulation type [F(3)=5.62, p<.001] and f0 [F(1)=113.44, p<.001] were all significant. Specifically, when participants chose 'down' responses, their imitated f0 was also lower (β=-3.44, p<.05), showing a positive correlation between pitch perception and imitation. A lowering trend according to modulation types was largely observed within the 'up' responses for both stimulus f0s, except that amplitude-modulated 300 Hz tokens with 'up' responses had lower imitated f0 (Fig. 5). In terms of voice quality, when using perceptual response as a predictor, their imitations tended to have lower H1–H2, HNR, and SoE, but higher SHR, especially with frequency and combined modulation. This indicates that the imitated tones were produced constricted, noisier, and quieter, but with more subharmonics, as in the actual tokens heard.

4. DISCUSSION

This study adopted an artificial language learning paradigm to investigate how Mandarin and English listeners perceive and process (pseudo-)linguistic tones realized by period doubling. In both cases, we observe a general trend that the perceived pitch and imitated f0 are lower when the stimuli are modulated to have correlates of period doubling, and when those modulations increase in degree. This suggests that listeners tend to identify a lower pitch during period doubling, which was especially found for frequency modulations or when the modulation degree (of both frequency and amplitude) reaches a certain threshold. This threshold differs depending on the modulation type: amplitude modulation only has around 70% of the tokens identified as low tones even when its modulation degree reaches the extreme (the stronger cycle is 4 times louder than the weaker one).

Thus voicing with amplitude modulation can still signal a 'high' tone, even when the modulation is strong, and especially when the original f0 is *lower*. Both perception and shadowing results support this, though it seems counterintuitive as lower f0 is typically associated with low tones. But because pitch and tone perception is relative, it is possible that if the original unmodulated tokens with an already low f0 (200 Hz) are categorized as the 'high' tone baseline, the effect on the f0 induced by amplitude modulation would not be salient enough to signal pitch lowering for a 'low' tone category.

Yet frequency modulation biases listeners to hear nearly 100% of the tokens as low tones. Listeners are probably more sensitive to changes in period than amplitude of glottal pulses when detecting periodicity or extracting pitch of speech signals. This may be related to findings that listeners tend to be influenced by changes in the frequency rather than time domains. For example, temporal noise measures like jitter and shimmer are not perceptually relevant independently of spectral HNR [18, 14].

Based on the findings that period doubling leads to a low tone bias – regardless of the original f0 – we would predict that the presence of period doubling could be used to signal low tones in languages, even when the f0 of the original tone is high. It will also interfere with high-tone perception, at least with moderate to high modulation.

Speakers also imitated typical characteristics found for period-doubled voice, as well as other creaky voice subtypes. Thus, they not only imitated irregular voicing to match the rough quality, but they also could be reproducing period doubling, or using it to realize roughness in the stimuli. However, we cannot be entirely sure of the precise type of creaky voice produced – whether listeners were able to target the specific subtype of creaky voice. A study designed to test imitation of different creaky voice subtypes would be useful and generalizable.

Lastly, the similar behaviors between Mandarin and English speakers suggest that pitch perception during period doubling may not be language-specific and is likely not influenced by inherent knowledge of lexical tone.

5. REFERENCES

[1] Sun, X., Xu, Y. 2002. Perceived pitch of synthesized voice with alternate cycles. *Journal of Voice*, *16*(4), 443-459.

[2] Bergan, C. C., Titze, I. R., 2001. Perception of pitch and roughness in vocal signals with subharmonics. *Journal of Voice*, *15*(2), 165-175.

[3] Keating, P., Garellek, M., Kreiman, J. 2015. Acoustic properties of different kinds of creaky voice. *Proc. 18th ICPhS* Glasgow, 0821-1.

[4] Yu, K. M. 2010. Laryngealization and features for Chinese tonal recognition. In *Eleventh Annual Conference of the International Speech Communication Association*. Makuhari, 1529-1532.

[5] Schreibweiss-Merin, D., Terrio, L.M., 1986. Acoustic analysis of diplophonia: A case study. *Perceptual and motor skills*, *63*(2), 755-765.

[6] Klatt, D., Klatt, L. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am*. 87, 820-857.

[7] Huang, Y. 2022. Articulatory properties of period-doubled voice in Mandarin. *Proc. Speech Prosody 2022*, 545-549.

[8] Titze, I. R., 1994. Fluctuations and perturbations in vocal output. *Principles of voice production*, 209-306.

[9] Gerratt, B. R., Kreiman, J. 2001. Toward a taxonomy of nonmodal phonation. *Journal of Phonetics*, *29*(4), 365-381.

[10] Boersma, P., Weenink, D. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.24, retrieved 5 September 2021 from http://www.praat.org/

[11] Peirce, J. W. 2007. PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162* (1-2), 8-13.

[12] Keating, P., Kuo, G. 2012. Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, *132*(2), 1050-1060.

[13] Shue, Y. L., Keating, P., Vicenik, C., Yu, K. 2010. VoiceSauce: A program for voice analysis. *Energy*, *1*(H2), H1-A1.

[14] Garellek, M. 2019. The phonetics of voice. In: W. F. Katz & P. F. Assmann (eds), *The Routledge Handbook of Phonetics*. Routledge, 75-106.

[15] Garellek, M., Chai, Y., Huang, Y., Van Doren, M. 2021. Voicing of glottal consonants and non-modal vowels. *Journal of the International Phonetic Association*, 1-28.

[16] Kim, S., Matachana, C., Nyman, A., Yu, K. M. 2020. Creak in the phonetic space of low tones in Beijing Mandarin, Cantonese, and White Hmong. In *10th International Conference on Speech Prosody 2020*. Tokyo, 523-527.

[17] Yuan, J., Liberman, M. 2014. F0 declination in English and Mandarin broadcast news speech. *Speech Communication*, *65*, 67-74.

[18] Kreiman, J., Gerratt, B. R. 2005. Perception of aperiodicity in pathological voice. The *Journal of the Acoustical Society of America*, *117*, 2201–2211.