COMPENSATION FOR F_0 VARIATION WITH VOCAL EFFORT AND VOWEL HEIGHT IN CANTONESE TONE PERCEPTION

Wei Lai[1], Mark Liberman[1], Qianxin He[2]

[1]University of Pennsylvania, United States; [2]Jinan University, China
weilai@sas.upenn.edu; markyliberman@gmail.com; catherinehqx@163.com

ABSTRACT

This paper evaluates how listeners integrate vocal effort and vowel height, in addition to speaker gender [13, 14], in the perception of Cantonese level tones. 50 participants attended a word identification task, in which they heard /gʌ/ and /gu/ sounds with different F_0 height, voice gender and vocal effort, and identified them as either a high-tone word or a mid-tone word. The result showed that, with equivalent F_0, participants were more likely to hear a high tone with normal-effort stimuli than high-effort stimuli; the difference was not as robust between normal-effort and low-effort stimuli. Besides, /gʌ/ received more high-tone responses than high-vowel ones /gu/, everything else being equal. Lastly, stimuli manipulated from a high-tone syllable received more high-tone responses than those from a mid-tone one, indicating potential integration of acoustic properties of the base tone. The results suggest that listeners successfully integrate cues from multiple dimensions of phonetic variation in tone perception.

Keywords: F_0, lexical tone, speech perception, compensation, vocal effort, vowel height, Cantonese

1. INTRODUCTION

Fundamental frequency (F_0) conveys linguistic meanings through lexical tones and intonation. However, F_0 may vary with quite a few factors on separate dimensions, causing acoustic overlap between linguistic categories and raising ambiguity in comprehension. The present study focuses on three cues of F_0 variation, namely, speaker gender, vocal effort and vowel height, and investigates how they are integrated simultaneously in tone perception.

F_0 differs between genders. Adult males tend to have a low F_0 mean (100–125 Hz) and narrow F_0 range (70–200 Hz) whereas females tend to have a high F_0 mean (180–220 Hz) and wide F_0 range (140–400 Hz) [2], due to physiological [12, 21] and social-cultural [9, 17, 22] factors. It has been shown that listeners integrate the covariance between F_0 and gender in the perception of pitch height and tone categories [3, 13, 14, 15]. For example, listeners can reliably identify the location of an F_0 in an unknown speaker's F_0 range based on gender-related acoustic traits [3]; with a tone language, listeners tend to perceive more high tones with stimuli in a male voice and more mid/low tones with those in a female voice with equalized pitch [13, 14].

Articulatorily, F_0 varies considerably with the amount of physiological effort involved in vocalization, hereby termed "vocal effort"[1]. Adjustments of vocal effort are normally motivated to compensate for the intelligibility loss of speech from turbulent noises [20] or long-distance sound propagation [23]. A number of physiological changes are involved, including the subglottal pressure, vocal fold tension, jaw opening, and muscle activities in and around the larynx [4, 23]. Acoustically, greater vocal effort leads to higher values of intensity, F_0 and F1, as well as steeper spectral tilt [16, 23]. It is not clear yet how vocal effort may interfere with pitch perception. The closest literature we found investigated the impact of loudness instead of vocal effort: It was reported that louder musical tones were perceived to have lower pitch in the range of 50–500 Hz [8]; however, the loudness effect was not found with stimuli of synthesized vowels with a comparable pitch range [5].

Vowel height is a more nuanced cue of F_0 variation and presumably is less noticeable to listeners. The tendency of vowel-intrinsic F_0 variation has been long documented, namely, that higher vowels tend to have higher F_0 [6, 19]. This tendency seems to be a commonality across languages and language families, and independent from the language's vowel inventory [24]. Studies on pitch perception showed that listeners compensate for this vowel-intrinsic F_0 variation, such that /a/ was heard as 2.2 Hz higher than /u/, 0.8 Hz higher than /i/, and 0.2 Hz higher than /e/ [5].

The present study investigates perceptual compensation for F_0-covarying factors in the the context of Cantonese tone perception, by asking how listeners' tone perception is affected by the above three cues simultaneously. We hypothesize that listeners will show, in parallel, more high-tone responses with low-vowel stimuli than high-vowel ones, low-effort

stimuli than high-effort ones, and male-voiced stimuli than female-voiced ones, in their perception responses to the same set of tone stimuli with variation in all three dimensions.

2. METHOD

2.1. Participants

50 participants were recruited from Guangdong, China, by a native Cantonese speaker (the third author), to complete task online through Qualtrics. They are 19 males and 31 females, aged 18–40, all self-reported as native Cantonese speakers.

2.2. Stimuli

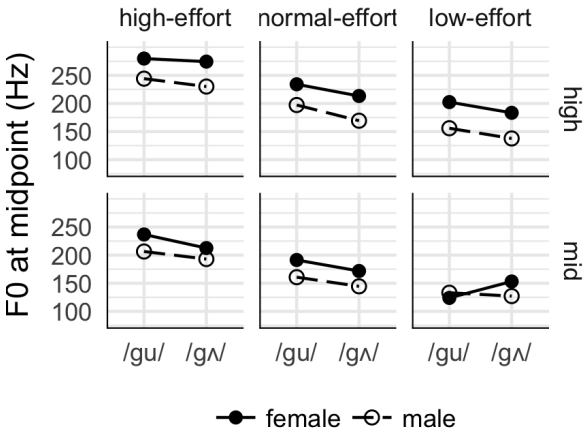
The stimuli were recorded from two native Cantonese speakers, a male and a female, in a professional sound booth in the University of Pennsylvania. Speakers produced a list of Cantonese monosyllabic words with level tones in three vocal effort levels, i.e., low, mid and high. They first produced the words with no instructions about the vocal effort, and those recordings are used as normal-effort stimuli. Next, the speakers were instructed to read the words aloud as if they were talking to someone far away. However, the elicited production did not differ much from their normal-effort speech, and was not used in the experiment. Speakers were then asked to put on a headphone and read the words with talker noises [25] played over their ears. This time, both of the two speakers raised their pitch and volume substantially, and those recordings were used as high-effort stimuli. At last, speakers were instructed to read the words in a very quiet voice to produce the low-effort stimuli. The above stimuli were all recorded with identical devices and parameter settings, with a constant mouth-to-microphone distance kept around 10 cm. The experimenter (the first author) was present throughout the whole procedure to ensure that the two speakers behaved in consistent ways, and that each word had several satisfying repetitions at each vocal effort level.

The word list contained 10 pairs of Cantonese syllables contrasted with high and mid tones, but only recordings of the following 2 pairs of words were used for stimuli construction. In all, the materials involved 4 word (2 vowel × 2 tone) × 3 vocal effort × 2 speaker = 24 tokens.

/gʌ55/	家	'family'	/gʌ33/	架	'shelf'
/gu55/	孤	'lonely'	/gu33/	固	'fixed'

Acoustic measurements of the 24 tokens showed that tokens articulated with different vocal effort levels differ in intensity and F_0 aspects. The mean intensity of tokens with high, normal and low vocal effort are respectively 78 dB, 71 dB and 62.5 dB. F_0 showed variation with vocal effort, speaker gender, tone and vowel in parallel (Fig. 1): F_0 is higher with higher-effort tokens than lower-effort ones, with male-voiced tokens than female-voiced ones, with high-vowel tokens (/gu/) than low-vowel ones (/gʌ/), and with high-tone tokens than mid-tone ones. The vowel-dependent F_0 appears to be less robust towards the lower end of the F_0 range [24].

Figure 1: The F_0 of high and mid tones in Hz conditioned by vowel identity and speaker gender



2.3. Manipulation

On the base of 100 Hz, we superimposed a 5-step pitch continuum of 11–15 st onto high-effort tokens, 7–15 st onto normal-effort tokens, and 7–11 st onto low-effort tokens, with the Linear Predictive Coding (LPC) algorithm in Praat. The pitch contour of the original production was maintained for each token, and the whole contour was raised or lowered until the midpoints matched to the desired pitch height.

Because the acoustic measurement shows no variation in intensity or duration between gender or vowels, we scaled all the high-effort tokens to 80 dB, normal-effort tokens to 70 dB and low-effort tokens to 60 dB, which is close to their group mean, and scaled all the stimuli to 0.7 seconds.

2.4. Procedure

A two-alternative forced-choice identification task was adopted. In each trial, participants heard either a /gʌ/ or a /gu/, and identified them as either a high-tone word or a mid-tone word with the same syllable from two options written in Chinese characters (i.e., 孤 and 固 for /gu/, 家 and 架 for /gʌ/).

All the trials were presented in two kinds of blocks. A *Softer* block contained low-effort and normal-effort items with pitch varying from 7–11 st, whereas a *Louder* block contained high-effort and normal-effort tokens with pitch from 11–15 st. Each type of block was repeated twice, and the participants completed the experiment either in an order of *Softer-Louder-Softer-Louder* or *Louder-Softer-Louder-Softer*. The order of stimuli was randomized within each block. In total, there were 2 gender×5 pitch step×2 vowel×2 block type×2 vocal effort×2 base tone×2 repetition = 320 tokens.

3. RESULTS

Two mixed-effects logistic regression models were conducted, one for each type of block (*Louder*, *Softer*), with Participant as the random factor, Response (mid:0, high:1) as the dependent variable, and GenderVoice (male:0, female:1), Pitch, VocalEffort (normal:0, high:1 in *Louder* blocks; normal:0, low:1 in *Softer* blocks), BaseTone (i.e., the token's original tone before manipulation, high:0, mid:1) and Vowel (/ʌ/:0, /u/:1) as fixed factors.

Table 1 shows the estimated coefficients of the fixed effects and their significance level. The coefficients stand for the logarithmic differences of high tone rates between conditions. As expected, a significant Pitch effect is reported by both of the models: One step of pitch would raise the logarithmic high tone probability by 0.38 in *Louder* blocks and 0.41 in *Softer* blocks. A significant vocal-effort effect is shown in *Louder* blocks, with a log high-tone rate 0.21 lower for high-effort tokens than mid-effort tokens. No such difference is found between normal-effort and low-effort stimuli in a *Softer* block. This is consistent with Fig. 2, which shows a split in high-tone rate by vocal effort levels in *Louder* blocks, but not in *Softer* ones.

Replicating [13, 14], the gender effect is shown significant with both block types. High tones were less often heard with female-voiced stimuli than male-voiced ones (by 0.38 in *Louder* and 0.41 in *Softer*, in log scale). Therefore, we expect to see speaker gender and vocal effort affect tone identification in parallel in a *Louder* block. This pattern is mostly clearly exhibited by stimuli manipulated from a high-tone syllable, as shown in Fig. 3. The responses are stratified by both gender voice and vocal effort: More high tones are identified with male-voiced stimuli than female-voiced ones, as shown across panels, and more are identified with normal-effort stimuli than high-effort ones within gender, as reflected within each panel.

Table 1: Estimated Coefficients of Fixed Effects

Factors	*Louder*	*Softer*
(Intercept)	-3.77***	-2.81***
GenderVoice	-0.65***	-0.38***
VocalEffort	-0.21***	0.001
Vowel	-0.54***	-0.50***
BaseTone	-0.44***	-0.53***
Pitch	0.38***	0.41***

Figure 2: Mean and standard error of the high tone rate by pitch step and vocal effort

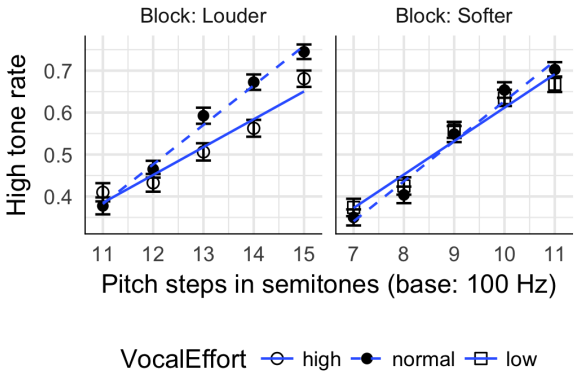
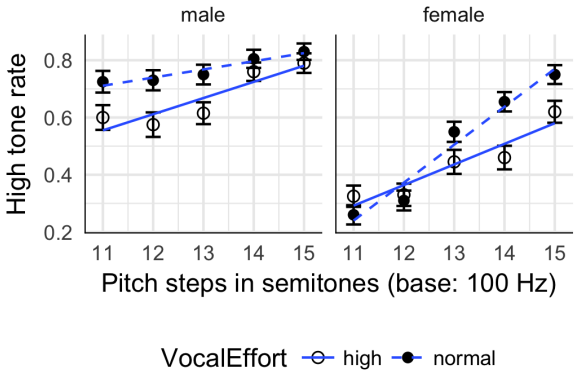


Figure 3: Mean and standard error of the high tone rate by gender and vocal effort with tokens with a high-tone base in *Louder* blocks



This 2-by-2 stratification becomes less clear after including stimuli with a mid-tone base, which indicates a potential interaction between vocal effort and stimulus base tone – another significant factor in the model that affects the perceived tone categories in both types of blocks. Fig. 4 shows the break-down of responses by the base tone of the stimuli. At each

Figure 4: Mean and standard error of the high tone rate by base tone and block type

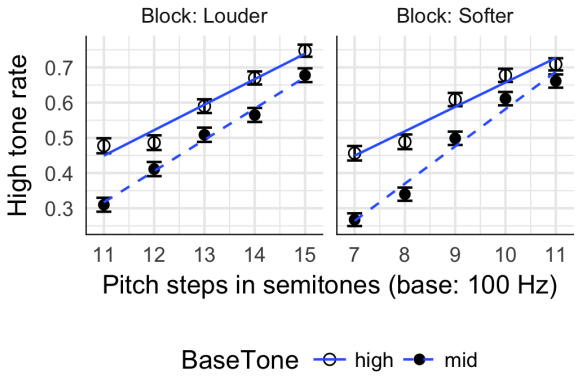
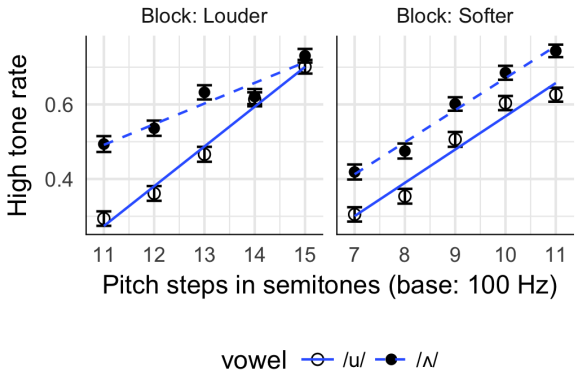


Figure 5: Mean and standard error of the high tone rate by vowel and block type



pitch step, tokens manipulated from a high-tone syllable have more high-tone responses than those from a mid-tone syllable, indicating that there may be other tone-relevant cues in the base apart from pitch height that listeners integrated in perception.

The last factor examined is vowel height. Fig. 5 shows the break-down of responses by vowels in each type of block, according to which, a /gʌ/ sound is more likely to be identified as a high-tone word than a /gu/ sound with equalized pitch. According to the model, the difference is as large as 0.54 in *Louder* blocks and 0.50 in *Softer* blocks in log odds.

4. DISCUSSION

Percepetual compensation is intensively studied at segmental level [1, 7, 18]. This study reports new findings at suprasegmental level by showing that listeners efficiently compensate for F_0 variation by cues of speaker gender, vowel height and vocal effort, indicating that linguistically naive listeners excel in integrating sophisticated language-specific phonetic variation to facilitate phonologization.

To the best of our knowledge, no previous studies have examined the integration of vocal effort on tone and pitch perception. The closest literature we find is concerned with the influence of loudness, which maps onto one dimension of vocal effort. The stimuli used in this paper were produced with different levels of vocal effort from the very beginning. Apart from amplitude, the stimuli also vary in formant frequencies and voice quality measurements according to post-hoc examination. Moreover,we found considerable variability between speakers in the acoustic correlates of vocal effort. Regardless, listeners still successfully compensate for F_0 variation with vocal effort, reflecting their sophisticated linguistic knowledge of phonetic co-variations.

It is also worth noting that tokens manipulated from a high-tone syllable are more likely to trigger a high tone response than those from a mid-tone syllable, regardless of the identical distribution of pitch and intensity cues they share after manipulation. This indicates potential integration of other acoustic cues intrinsic to tonal syllables in tone identification, cues that we don't have a good understanding of yet. Since we maintained the original pitch contour of the production for each token, one possibility is that listeners also attended to nuanced pitch fluctuation to perceive tones. Other possibilities include that listeners made use of voice-quality cues, which covaries with pitch height in production and becomes stored in the spectra, to infer pitch locations [10, 11]. Future studies should try to tease these possibilities apart.

5. CONCLUSION

In this paper, we reported a word identification experiment investigating whether listeners integrate vocal effort, vowel height and speaker gender in parallel in the perception of Cantonese level tones. The results show that listeners heard fewer high tone (and more mid tones) with higher-effort stimuli than lower-effort ones, higher-vowel stimuli than lower-vowel ones, and female-voiced stimuli than male-voiced ones, indicating that listeners are capable of simultaneously compensating for F_0 variation in multiple dimensions in tone perception.

6. REFERENCES

[1] Beddor, P. S., Krakow, R. A. 1999. Perception of coarticulatory nasalization by speakers of en-

glish and thai: Evidence for partial compensation. *The Journal of the Acoustical Society of America* 106(5), 2868–2887.

[2] Biemans, M. 2000. *Gender variation in voice quality*. Netherlands Graduate School of Linguistics.

[3] Bishop, J., Keating, P. 2012. Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. *The Journal of the Acoustical Society of America* 132(2), 1100–1112.

[4] Boone, D., McFarlane, S., Von Berg, S., Zraick, R. 2010. The voice and voice therapy. 8th.

[5] Chuang, C.-K., Wang, W. S.-Y. 1976. Influence of vowel height, intensity, and temporal order on pitch perception. *The Journal of the Acoustical Society of America* 60(S1), S92–S92.

[6] Crandall, I. B. 1925. The sounds of speech. *The bell system technical journal* 4(4), 586–639.

[7] Elman, J. L., McClelland, J. L. 1988. Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language* 27(2), 143–165.

[8] Fletcher, H. 1934. Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure. *The Journal of the Acoustical Society of America* 6(2), 59–69.

[9] Gussenhoven, C. 2002. Intonation and interpretation: Phonetics and phonology. *Speech Prosody 2002, International Conference*.

[10] Kuang, J., Guo, Y., Liberman, M. 2016. Voice quality as a pitch-range indicator. *Proceeding of Speech Prosody*.

[11] Kuang, J., Liberman, M. 2015. The effect of spectral slope on pitch perception. *INTERSPEECH* 354–358.

[12] Künzel, H. J. 1989. How well does average fundamental frequency correlate with speaker height and weight? *Phonetica* 46(1-3), 117–125.

[13] Lai, W. 2017. Auditory-visual integration of talker gender in cantonese tone perception. *Proc. Interspeech 2017* 664–668.

[14] Lai, W. 2018. Voice gender effect on tone categorization and pitch perception. *Proc. TAL2018, Sixth International Symposium on Tonal Aspects of Languages* 103–107.

[15] Lee, C.-Y. 2009. Identifying isolated, multispeaker mandarin tones from brief acoustic input: A perceptual and acoustic study. *The Journal of the Acoustical Society of America* 125(2), 1125–1137.

[16] Liénard, J.-S., Di Benedetto, M.-G. 1999. Effect of vocal effort on spectral properties of vowels. *The Journal of the Acoustical Society of America* 106(1), 411–422.

[17] Loveday, L. 1981. Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of english and japanese politeness formulae. *Language and Speech* 24(1), 71–89.

[18] Mann, V. A., Repp, B. H. 1981. Influence of preceding fricative on stop consonant perception. *The Journal of the Acoustical Society of America* 69(2), 548–558.

[19] Peterson, G. E., Barney, H. L. 1952. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 24, 175–184.

[20] Raitio, T., Suni, A., Vainio, M., Alku, P. 2014. Synthesis and perception of breathy, normal, and lombard speech in the presence of noise. *Computer Speech & Language* 28(2), 648–664.

[21] Rendall, D., Vokey, J. R., Nemeth, C., Ney, C. 2005. Reliable but weak voice-formant cues to body size in men but not women. *The Journal of the Acoustical Society of America* 117(4), 2372–2372.

[22] Szakay, A. 2006. Rhythm and pitch as markers of ethnicity in new zealand english. *Proceedings of the 11th Australasian International Conference on Speech Science & Technology', University of Auckland* 421–426.

[23] Traunmüller, H., Eriksson, A. 2000. Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America* 107(6), 3438–3451.

[24] Whalen, D. H., Levitt, A. G. 1995. The universality of intrinsic f0 of vowels. *Journal of phonetics* 23(3), 349–366.

[25] Youtube audio library: "crowding talking". http://https://www.youtube.com/watch?v=mLld3JVwxew.

[1] In the literature, the term "vocal effort" is sometimes used exclusively to refer to the vocal adjustment induced by changing the *talker-to-listener distance*, but not by changing noise levels. We don't make this distinction here.