Influence of spectral cues on the perception of pitch height

Jianjing Kuang, Mark Liberman

University of Pennsylvania
kuangj@sas.upenn.edu, markyliberman@gmail.com

ABSTRACT

This study aims to provide direct perceptual evidence of whether listeners integrate spectral cues in pitch-range perception. A force-choice pitch classification experiment with four spectral conditions was conducted to investigate whether spectral cues manipulation can affect pitch-height perception. The results show that the pitch classification function is significantly shifted under different spectral conditions. Listeners generally hear a higher pitch when the spectrum has higher high-frequency energy (i.e. tenser phonation). This study strongly supports the hypothesis that voice quality cues and F0 interact in pitch perception.

Keywords: pitch perception, voice quality, F0, spectral slope

1. INTRODUCTION

Pitch perception plays a crucial role in speech processing, as pitch conveys important linguistic information such as tone and intonation from a speaker. Although pitch is an auditory concept, in practice, it has been used interchangeably with fundamental frequency (F0), which appears to be the only acoustic correlate of pitch. Since F0 range differs across speakers, what is a low or high F0 varies by speaker, and phonetic categories (e.g. tonal categories) thus overlap in acoustic signals. In order to uncover the intended linguistic pitch by a speaker, listeners need to identify the pitch location within a speaker's pitch range.

Speaker normalization is certainly easier when listeners are previously exposed to a voice or when the context is available (e.g. [28]), but studies ([4][13]) have shown that listeners are able to identify the pitch location of very brief voice samples in an unknown speaker's range in the absence of any contextual cues. This suggests that listeners must use other signal-internal information that co-varies with F0 as cues to pitch range.

Both [13] and [19] speculated that voice quality could be such a cue. Indeed Lee [19] found that voice quality cues (H1-H2, H1-A3) were correlated with tone classification between high and low. However, he further noted that f0 was the only significant predictor for identification accuracy in the regression model. [4] replicated [13]'s experiment and found that acoustic measures of voice quality had only a very small effect on pitch location ratings. They suggested that voice quality only indirectly influences pitch perception, possibly through its information about sex.

Although pitch-location experiments did not find strong correlations between voice quality cues and pitch perception, the co-variation between F0 and voice quality has been found in pitch production studies. It has been proposed that pitch range is divided into three "registers" ([11][12][24][27]), basically three pitch sub-ranges, and each register is related to a certain type of phonation. Lowest pitch range is associated with vocal fry, and highest pitch range is associated with tense voice and falsetto.

Pitch-dependent voice qualities were less addressed in linguistic studies, but [17] suggested that they were important cues for the tones with extreme pitch targets. She found that voice quality cues in Black Miao, a five-level-tone language, significantly enhanced the contrasts between 55 (extreme high) and 44 (non-extreme high), between 11 (extreme low) and 22 (non-extreme low). In light of this, she proposed that the natural co-variation between creaky voice and low pitch is the reason why creaky has been found to facilitate the low tone perception in Mandarin [29] and Cantonese [30]. Similarly, tense and falsetto were attested to occur with some extreme high tones (e.g. Pakphanang Thai: [23]; Yueyang dialect: [21])

Studies from Automatic Speech Recognition also demonstrated the crucial role of voice quality cues in pitch classification. For example, [25] found that gross spectral measure was a better basis for tone classification of Mandarin tones than F0 estimates were. And the accuracy of tone recognition of this classifier was even better than that of a human listener. The surprising results suggested that spectral structure must carry important information to cue pitch ranges.

Taken together, it has been speculated from various perspectives that voice quality may play a role in pitch perception, but there is no direct evidence to support this claim yet. Thus the purpose of this study is to carry out an experiment that can directly test the effect of voice quality on pitch perception.

The voice quality cue will be examined in the present study is the spectral slope. It has been well established that the relative slope of the voice source spectrum is one of the most important acoustic correlates of voice quality (see [10] for review, especially Figure 11.12): A relatively steep spectral slope is associated with a breathier voice, and a flat spectral slope is associated with a tenser or creakier voice (Note that the latter also has pulse-to-pulse variability). The spectral tilt is usually measured as the amplitude of the fundamental (H1) relative to some higher-frequency components (e.g. H1-H2, H1-A1, H1-A2, and H1-A3. A1, A2, A3 are the amplitudes of the harmonic near the first, second and third formants). These measures have been found to be the reliable indicators of phonation contrasts across languages (e.g. Southern Yi: [18]; Green Mong: [2]; White Hmong: [8]; Takhian Thong Chong: [6]; Suai/Kuai: [1]; Javanese: [26]; Ju|'hoansi: [20]; Santa Ana Valle Zapotec: [7]; Mazatec: [5] [9]; Gujarati: [16], to name a few), and phonation classification in perceptual spaces (e.g. [15]) Therefore, the working hypothesis of the current study is that, if voice quality can affect pitch perception, manipulating the spectral slope of a voice should be able to shift listeners' perception of pitch height.

2. METHOD

2.1. Stimuli

Speech-like stimuli varying in pitch and spectral cues are synthesized. The stimuli are four sets of sine-wave overtones with two peaks, which were created by convolving a hamming window with a sawtooth whose baseline pitch value is always 120 Hz. The F0 of the first peak is always 169.34 Hz, and the second peak is a pitch continuum with 11 steps, at 153.06, 156.19, 159.38, 162.63, 165.96, 169.34, 172.80, 176.33, 179.93, 183.61, and 187.36 Hz respectively. This roughly covers the comfortable pitch range of a male speaker ([3]). And at step 6, peak 1 and peak 2 have the same F0 value. Pitch manipulation is shown in Figure 1.

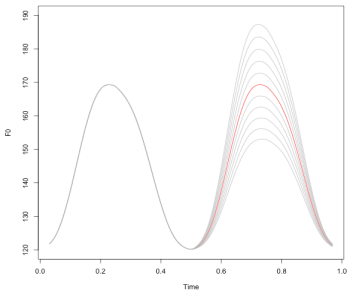
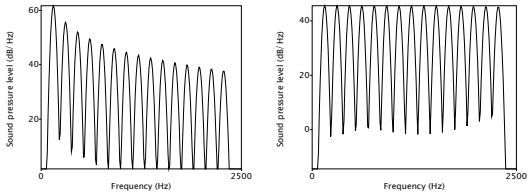


Figure 1: F0 manipulation: the first peak has a constant F0 value at 169.34 Hz, and the second peak is a continuum with 11 steps. Peak 1 and peak 2 are identical at step 6 (red/dark line for the second peak).

To manipulate voice quality cues, a tilted and a flat source spectrum were first created. In the tilted spectrum, overtone amplitude decreases as 1/N, to a point 15 dB down from the fundamental (Shown in Figure 2a). As can be seen here, as the result of the tilted slope, the strength of the first harmonic is relatively more prominent than that of higher-frequency harmonics. By contrast, in the flat spectrum, the overtone amplitude is kept constant, and thus the first harmonic is not prominent in the spectrum. Therefore, compared with the tilted spectrum, the flat spectrum, which has a greater energy in high-frequency harmonics, indicates a tenser voice ([10]).

Figure 2: Spectrum manipulation: tilted spectrum (a) vs. flat spectrum (b).



(a) (b)

The two types of source spectra were then applied to the two peaks of the complex tones, generating four spectral conditions (implicated phonation types are in the brackets, in relative terms):

- Set A: Both peaks have a tilted spectrum (i.e., breathier + breathier)
- Set B: Both peaks have a flat spectrum (i.e., tenser + tenser)
- Set C: The first peak has a tilted spectrum, and the second peak has a flat spectrum, with a 200ms transition in the middle (i.e., breathier + tenser)
- Set D: The first part is the flat spectrum, and the second part is the tilted spectrum, with a 200ms transition in the middle (i.e., tenser + breathier).

Therefore, there were 44 stimuli (11 F0 steps x 4 spectral conditions) in a total. All stimuli were 1s in duration.

2.2. Procedure

A forced-choice pitch classification task was used to test listeners' categorization of pitch values under different spectral conditions. Ten copies of each stimulus were presented in random order to each listener. For each trial, the listeners were asked to attend to pitch, and judge whether the second peak was higher or lower than the first peak by clicking on the corresponding buttons on the computer screen. All tests were executed in a sound booth with stimuli presented over Sennheiser 280 headphones.

2.3. Subjects

58 participants, between age 18 and 22, were recruited from the student population at the University of Pennsylvania. All of them reported to speak English as their primary language. Three of them failed to complete the task as instructed, and thus were excluded from the analysis. None of the participants reported to have hearing issues.

3. RESULTS AND DISCUSSIONS

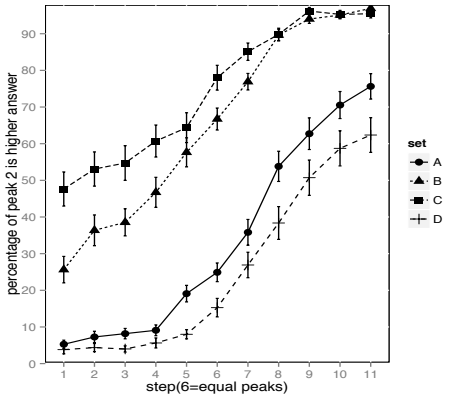
Figure 3 shows the proportion of "peak 2 is higher" responses across all listeners. The main effects of spectral conditions were evaluated using an MCMC generalized linear mixed-effects model (*mcmcglmm* package in R). F0 steps (1-11) and spectral conditions (A, B, C and D) were the fixed factors, and random intercepts and slope were included for subjects. Main effects of spectral conditions were summarized in Table 1. The results are reported as means of regression coefficients followed by 95% highest posterior density intervals in square brackets and associated p-values. As shown in Table 1, significant effects were found between every two spectral conditions, which means that pitch classification function is significantly shifted in each spectral condition. The proportion of "peak 2 is higher" responses was in the order of (see Figure 3) Set C (breathier + tenser) > Set B (tenser + tenser) > Set A (breathier + breathier) > Set D (tenser + breathier).

Overall, the perception of pitch height was strongly biased by the spectral cues. As can be seen in Figure 3, compared with set A, the pitch classification function for set C (breathier + tenser combination) was dominated by the "peak 2 is higher" responses, even when peak 2 was approximately 10 Hz lower than peak 1. By contrast, the pitch classification function of set D (tenser + breathier combination) was shifted in the opposite direction. In this condition, listeners hardly heard a higher peak 2, even when it was about 10 Hz higher than peak 1. In other words, when the second peak was tenser than the first peak, it tended to be perceived to have a higher pitch; and when the second peak was breathier than the first peak, it tended to be perceived to have a lower pitch. The co-variation between tense voice and high pitch has been well documented in pitch production studies (e.g. [27]), and the finding of the current experiment confirms that the co-variation is also true in the perception domain.

Table 1: Main effects of spectral conditions between every two conditions. Means of regression coefficients, followed by 95% highest posterior density intervals in square brackets and associated p-values.

	A	B	C
B	1.3[1.2,1.5] p<0.001		
C	1.7[1.6,1.8] p<0.001	0.4[0.3,0.6] p<0.001	
D	0.4[0.3,0.5] p<0.001	1.8[1.7,2.0] p<0.001	2.5[2.4,2.7] p<0.001

Figure 3: Pitch classification functions for all listeners. X-axis= F0 steps, y-axis= proportion of "peak 2 is higher" responses; line patterns are different spectral conditions. The error bars are 95% confidence intervals



Interestingly, the pitch classification functions for set A (breathier + breathier) and set B (tenser + tenser) were also significantly different, with set B more in favor of "peak 2 is higher". The reason why peak 2 still sounded higher could be that listeners still thought the second peak was tenser than the first one, even if the peaks had the identical spectrum condition. Perhaps just as listeners expected a F0 declination ([22]: when two stressed syllables sounded with the same pitch, the second one was actually lower), listeners may have also expected a declination in tenseness.

Post hoc analysis found that there was listener variability in pitch perception. As shown in Figure 4, seven listeners were found to only pay attention to the F0 dimension. For this group of listeners, pitch classification function had no shift between spectral conditions. On the other hand, eight listeners appeared to mostly pay attention to the spectral conditions of peak 2, as shown in Figure 5. For this group of listeners, F0 cues were even less important than spectral cues in judging pitch height. The classification function appears to have a two-way distinction. Majority listeners in our study, as shown in Figure 6, used both F0 and spectral cues in judging pitch heights.

Figure 4: Listeners who attended exclusively to F0.

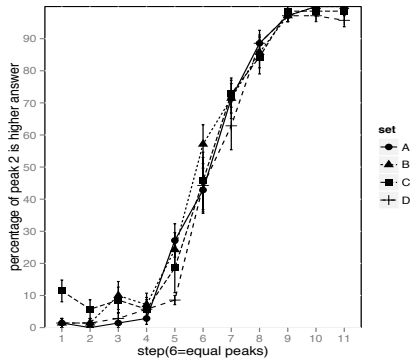


Figure 5: Listeners who mostly attended to spectral quality

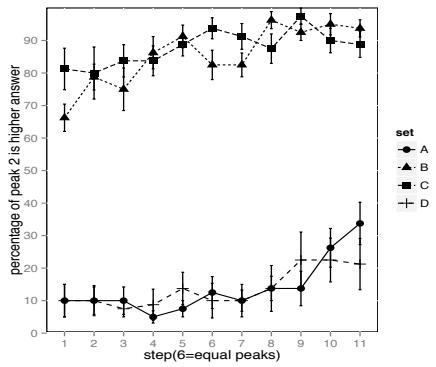
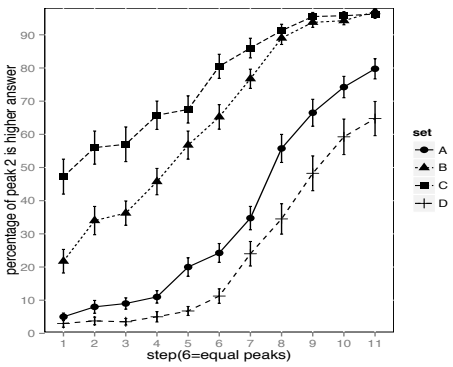


Figure 6: Listeners who attended to both cues



4. CONCLUSIONS

This study sought to examine whether voice quality cues are integrated in pitch perception. The major finding of the present study is that the pitch classification function can be significantly shifted by different spectral conditions, which means that spectral cues strongly influence on pitch perception. Moreover, listeners generally perceive a higher pitch when the higher-frequency components in the spectrum have higher energy (indicating a tenser voice quality ([10])). The direction of shift is consistent with the co-varying relationship between F0 and voice quality: high F0 is naturally produced with a tense voice ([27]). This study thus strongly supports the hypothesis that voice quality cues and F0 are integrated in pitch perception, and that voice quality is a direct indicator of pitch range.

The findings of this study have important implications for pitch-related linguistic studies: pitch is not just F0, either in production or in perception. As suggested in this study, pitch range is determined by both F0 and voice quality cues. So what is perceptually "higher" does not necessarily have a higher F0. This is very useful in speaker normalization, as voice quality can provide information about pitch location for a speaker: for example, a tense voice indicates that the speaker almost reaches his/her highest range (or speaker is talking at his/her high pitch). The ability to integrate F0 and voice quality is also very useful in processing multiple contrastive levels in languages (e.g. [17] [28]). The strong interaction between F0 and voice quality in production and perception should be taken into account when categorizing tones and prosodic structures.

The listener variability shown in Figure 4-6 is also striking. It appears that listeners have different strategies of interpreting "pitch": they can use either or both of the cues. Follow-up experiments should be conducted to explore what factors (e.g. language background, music training) may lead to this variation.

5. ACKNOWLEDGMENTS

This study is supported by a URF awarded by UPenn to the first author. We would like to thank Yong-Cheol Lee, Yixuan Guo, Jia Tian and Jingjing Tan for their assistance in conducting the experiment.

6. REFERENCES

[1] Abramson, A. S., Luangthongkum, T., Nye, P. W. 2004. Voice register in Suai (Kuai): An analysis of perceptual and acoustic data. *Phonetica* 61, 147-171.

[2] Andruski, J. E., Ratliff, M. 2000. Phonation types in production of phonological tone: the case of Green Mong. *Journal of the IPA* 30, 37-61.

[3] Baken, R. J., Orlikoff, R. F. 2000. *Clinical measurement of speech and voice.* Singular Publishing Group: San Diego.

[4] Bishop, J., Keating, P. 2012. Perception of pitch location within a speaker's range: fundamental frequency, voice quality and speaker sex. *J. Acoust. Soc. Am.* 132, 1100–1112.

[5] Blankenship, B. 2002. The timing of nonmodal phonation in vowels. *J. Phonetics* 30, 163-191.

[6] DiCanio, C. T. 2009. The phonetics of register in Takhian Thong Chong. *Journal of the IPA* 39, 162-188.

[7] Esposito, C. M. 2010. Variation in contrastive phonation in Santa Ana Del Valle Zapotec. *Journal of the IPA* 40, 181-198.

[8] Esposito, C. M. 2012. An acoustic and electroglottographic study of White Hmong phonation. *J. Phonetics* 40, 466-476.

[9] Garellek, M., and Keating, P. 2011. The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the IPA* 41, 185-205.

[10] Gobl, C., Ní Chasaide, A. 2012. Voice source variation. In: Hardcastle, W.J., Laver, J. (eds), *The Handbook of Phonetic Science*. Oxford: Blackwell, 378-423.

[11] Hollien, H. 1974. On vocal registers. J. *Phonetics* 2, 125-143.

[12] Hollien, H., Michel, J. F. 1968. Vocal fry as a phonational register. *J. Speech Lang. Hear. Res.* 11, 600-604.

[13] Honorof, D. N., Whalen, D. H. 2005. Perception of pitch location within a speaker's F0 range. *J. Acoust. Soc. Am.* 117, 2193–2200.

[14] Keating, P., Esposito, C., Garellek, M., Khan, S., Kuang, J. 2011. Phonation contrasts across languages. *Proc. 17th ICPhS* Hong Kong, 203–206.

[15] Kreiman, J., Gerratt, B.R. 2010. Perceptual sensitivity to first harmonic amplitude in the voice source. *J. Acoust. Soc. Am.* 128, 2085-2089.

[16] Khan, S. D. 2012. The phonetics of contrastive phonation in Gujarati. *J. Phonetics* 40, 780-795.

[17] Kuang, J. 2013. The tonal space of contrastive five level tones. *Phonetica* 70, 1-23.

[18] Kuang, J., Keating, P. 2014. Glottal articulations in tense vs. lax phonation contrasts. *J. Acoust. Soc. Am.* 136, 2784–2797.

[19] Lee, C.-Y. 2009. Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study. *J. Acoust. Soc. Am.* 125, 1125–1137.

[20] Miller, A. L. 2007. Guttural vowels and guttural co-articulation in Ju|'hoansi. *J. Phonetics* 35, 56-84

[21] Peng, J. G., Zhu, X. N. 2010. Falsetto in Yueyang dialect. *Journal of Contemporary Linguistics* 1, 24-32.

[22] Pierrehumbert, J. 1979. The Perception of Fundamental Frequency Declination, *J. Acoust. Soc. Am.* 66, 363-369.

[23] Rose, P. 1997. A seven-tone dialect in Southern Thai with super-High: Pakphanang tonal acoustics and physiological inferences. In: A. S. Abramson (eds), *Southeast Asian Linguistic Studies in Honour ofmVichin Panupong*. Bangkok: Chulalongkorn University Press, 191-208.

[24] Roubeau, B., Henrich, N., Castellengo, M. 2009. Laryngeal vibratory mechanisms: The notion of vocal register revisited. *Journal of Voice* 23, 425-438.

[25] Ryant, N., Slaney, M., Liberman, M., Shriberg, E., Yuan, J. 2014. Highly accurate mandarin tone classification in the absence of pitch information. *Proc. 7th Speech Prosody* Dublin, 673-677.

[26] Thurgood, E. 2004. Phonation types in Javanese. *Oceanic Linguistic*s 43, 277-295.

[27] Titze, I. R. 1988. A framework for the study of vocal registers. *Journal of Voice* 2, 183-194.

[28] Wong, P. C. M., Diehl, R. L. 2003. Perceptual normalization for inter-and intratalker variation in Cantonese level tones. *J. Speech Lang. Hear. Res.* 46, 413–421.

[29] Yang, R. X. 2011. The Phonation factor in the categorical perception of Mandarin tones. in *Proc. 17th ICPhS* Hong Kong, 2204-2207.

[30] Yu, K. M., Lam, H. W. 2014. The role of creaky voice in Cantonese tonal perception. *J. Acoust. Soc. Am.* 136, 1320–1333.