The effect of spectral slope on pitch perception

Jianjing Kuang [1], *Mark Libermann*[1]

[1] Department of Linguistics, University of Pennsylvania, U.S.A.

kuangjj@ling.upenn.edu, markyliberman@gmail.com

Abstract

This study aims to explore whether listeners integrate spectral cues in pitch-range perceptions. A forced-choice pitch classification experiment with four spectral conditions was conducted to investigate whether spectral cue manipulation can affect pitch-height perceptions. The participants in this experiment include tonal vs. non-tonal language speakers and musicians vs. non-musicians. The results show that the pitch classification function significantly shifted under different spectral conditions. Listeners generally hear higher pitches when the spectrum includes more high-frequency energy (i.e., tenser phonation). This study strongly supports the hypothesis that voice quality cues and F0 interact in pitch perceptions. Moreover, language experience and musical training can affect the magnitude of shifts.

Index Terms: pitch perception, voice quality, F0, spectral slope

1. Introduction

This study explores whether spectral cues affect pitch perceptions. As pitch conveys important linguistic information such as tone and intonation, pitch perception plays a crucial role in speech processing. Although pitch is an auditory concept, in practice, it has been considered interchangeable with fundamental frequency (F0), which appears to be the only acoustic correlate of pitch. Studies of pitch perception have primarily focused on F0 cues.

Recent findings suggest that pitch perceptions may involve other cues. For example, speech normalization studies ([1-3]) have shown that listeners are able to identify the pitch location of very brief voice samples of an unknown speaker's range in the absence of any contextual cues. This finding suggests that listeners must use other signal-internal information that co-varies with F0 as cues to pitch range.

Both [1] and [2] speculated that voice quality could be such a cue. Indeed, [2] found that voice quality cues (H1-H2, H1-A3) are correlated with high and low tone classification. However, the researcher found f0 to be the only significant predictor of identification accuracy in the regression model. [3] replicated [1]'s experiment and found that acoustic measures of voice quality have only a small effect on pitch location ratings. The authors suggested that voice quality only indirectly influences pitch perception, possibly through information related to sex.

Although pitch-location experiments did not find strong correlations between voice quality cues and pitch perception, co-variation between F0 and voice quality has been found in pitch production studies. Researchers have proposed that pitch range is divided into three "registers" ([4-7]), essentially three pitch subranges, and that each register is related to a certain type of phonation. The lowest pitch range is associated with vocal fry, and the highest pitch range is associated with tense voice and falsetto.

The spectral slope of the voice source spectrum is an important indicator of voice quality. It is well established (see [8] for a review, especially Figure 11.12) that a relatively steep spectral slope is associated with a breathier voice and that a flat spectral slope is associated with a tenser or creakier voice (note that the latter also includes pulse-to-pulse variability). The spectral tilt is typically measured as the amplitude of the fundamental (H1) relative to higher-frequency components (e.g., H1-H2, H1-A1, H1-A2, and H1-A3; A1, A2, and A3 are the amplitudes of the harmonic near the first, second and third formants, respectively). These measures have been found to be reliable indicators of phonation contrast across languages (e.g., Green Mong: [9, 10]; Mazatec: [11]; Suai/Kuai: [12]; Javanese: [13]; Ju|'hoansi: [14]; Takhian Thong Chong: [15]; Santa Ana Valle Zapotec: [16] [17]; Southern Yi: [18]; White Hmong: [19]; Gujarati: [20]) and of phonation classification in perceptual spaces (e.g., [21]).

Involving spectral cues in computational modeling can significantly improve tonal category classification. For example, in modeling the tonal space of Black Miao, a five-level tone language, [22] found that spectral cues significantly enhance the contrasts between 55 (extreme high) and 44 (non-extreme high) and between 11 (extreme low) and 22 (non-extreme low). In view of this result, the author proposed that the natural co-variation between tense and high F0 may facilitate perceptions of high pitch. In building an automated classifier for Mandarin tones, [23] found that a gross spectral measure is a better basis for the classification of Mandarin tones than F0 estimates are. And the accuracy of tone recognition of this classifier is even better than that of a human listener.

Researchers have speculated from various perspectives that voice quality may play a role in pitch perception, but this claim is not yet supported by direct evidence. Thus, the purpose of this study is to conduct an experiment that directly tests the effect of voice quality on pitch perception. Because spectral slope is known as one of the most important indicators of voice quality, the working hypothesis of the current study states that if voice quality can affect pitch perception, manipulating the spectral slope of a voice should be able to shift listener perceptions of pitch height. In addition, studies have shown that pitch perception can be influenced by language experience (e.g., [24, 25]) and by music training (e.g., [26, 27]). Therefore, to account for cross-listener variation, the experiment is replicated with tonal language speakers and musicians.

2. Method

2.1. Stimuli

Speech-like stimuli that vary 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 continuum approximates the comfortable pitch range of a male speaker ([3]). At step 6, peaks 1 and 2 have the same F0 value. A pitch manipulation graph is shown in Figure 1.

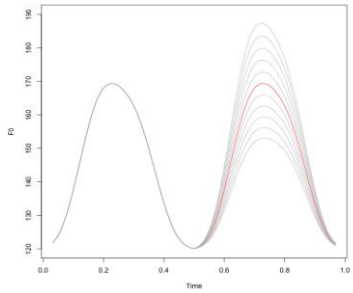

Figure 1: *F0 manipulation: the first peak has a constant F0 value at 169.34 Hz, and the second peak is a continuum with 11 steps. Peaks 1 and 2 are identical at step 6 (red/dark lines for the second peak).*

To manipulate voice quality cues, a tilted and flat source spectrum were first created. In the tilted spectrum, overtone amplitude levels decrease at 1/N to 15 dB below the fundamental (Figure 2a). As is shown here, as the result of the tilted slope, the strength of the first harmonic is more prominent than that of higher-frequency harmonics. By contrast, in the flat spectrum, the overtone amplitude remains constant, and thus, the first harmonic is not prominent in the spectrum. Therefore, compared with the tilted spectrum, the flat spectrum, which has more energy in high-frequency harmonics, indicates a tenser voice ([10]).

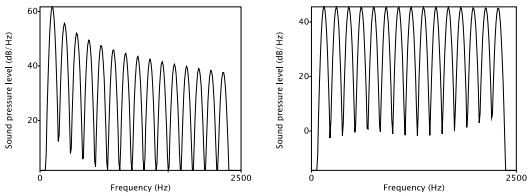

Figure 2: *Spectrum manipulation: tilted spectrum (a) vs. flat spectrum (b).*

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 presented in 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 has a flat spectrum, with a 200 ms transition in the middle (i.e., breathier + tenser)
- Set D: The first part constitutes the flat spectrum, and the second constitutes the tilted spectrum, with a 200 ms 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 1 s 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 focus on pitch and to evaluate 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 through Sennheiser 280 headphones.

2.3. Subjects

Fifty-eight English speakers between age 18 and 22 were recruited from the student population at the University of Pennsylvania, and 55 students completed the experimental task. Thirty-four Chinese speakers living in China between 20 and 40 years of age were recruited via the Qualtrics online survey system, and 21 of them completed the experiment. Listeners who reported having hearing disabilities were excluded from the analysis, leaving a total of 55 English speakers and 19 Chinese speakers in the sample. In addition, 5 musicians participated in the study via the Qualtrics survey system. All musicians have had professional musical instrument or voice training.

3. Results

3.1. English listeners

Figure 3 shows the proportion of "peak 2 is higher" responses for all English 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 used as fixed factors, and random intercepts and slopes were included as subjects. The main effects of the spectral conditions are 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, the results revealed significant effects between each pair of spectral conditions, demonstrating that pitch classification functions 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 spectral cues. As shown in Figure 3, compared with set A, the pitch classification function for set C (breathier + tenser combination) was dominated by "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) shifted in the opposite direction. In this condition, the listeners scarcely heard a higher peak 2, even when it was approximately 10 Hz higher than peak 1. In other words, when the second peak was tenser than the first, the second peak tended to be perceived as a higher pitch, and when the second peak was breathier than the first, the second tended to be perceived as a lower pitch. Co-variation between tense voice and high pitch has been well documented in pitch production studies (e.g., [6]), and the results of the current experiment confirm that co-variation also exists in the domain of perception.

	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

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

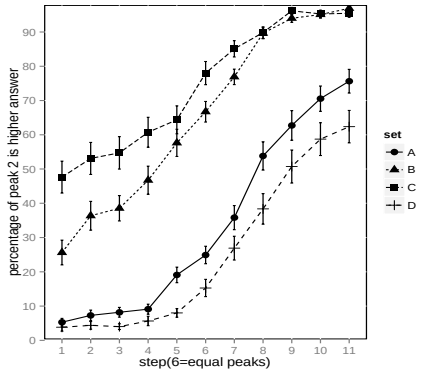

Figure 3: *Pitch classification functions for English listeners. X-axis=F0 steps, y-axis=proportion of "peak 2 is higher" responses; line patterns denote different* spectral *conditions. Error bars denote 95% confidence intervals.*

Interestingly, the pitch classification functions for set A (breathier + breathier) and set B (tenser + tenser) were also significantly different, with set B favoring "peak 2 is higher" results. Peak 2 may have sounded higher because listeners still believed the second peak to be tenser than the first peak, even when the peaks had identical spectrum conditions. Similar to how the listeners expected F0 decline ([28]: when two stressed syllables with the same pitch were given, the second was actually lower), they may have also expected a decline in tenseness.

3.2. Chinese listeners

The shifting effect shown by the experiment with English listeners is robust and prompts the question of whether it holds for speakers of tonal languages. Research has shown that the pitch processing of tonal language speakers differs from that of non-tonal language speakers (e.g., [26]). Hence, the same experiment was replicated with Chinese listeners. Figure 4 shows the proportion of "peak 2 is higher" responses given by all Chinese listeners. The same MCMC generalized linear mixed-effects model described in 3.1 was used to evaluate the main effect of spectral conditions. As in Table 1, significant shifts were found for every pair of spectral conditions. Therefore, the same shifting effect also applies to tonal language speakers. However, although Figure 4 shows a pattern similar to that shown in Figure 3, the spacing of classification functions for Figure 4 noticeably differs from that of Figure 3. The distance between sets A and B is smaller in Figure 4 than in Figure 3; furthermore, the differences between A and D and between B and C are more salient in Figure 4. Hence, the judgment strategy appears to differ slightly between the two language groups. The Chinese speakers paid more attention to differences between the two peaks, whereas the English speakers largely focused on the spectral quality of the second peak.

	A	B	C
B	0.9[0.6,1.2] P<0.001		
C	1.6[1.3,2.0] p<0.001	0.8[0.5,1.1] p<0.001	
D	-0.7[-1.0,-0.4] P<0.001	-1.6[-2.0,-1.3] p<0.001	-3.0[-3.7,-2.2] p<0.001

Table 2. *Main effects of spectral conditions for each pair of conditions. Means of regression coefficients followed by 95% highest posterior density intervals in square brackets and associated p-values.*

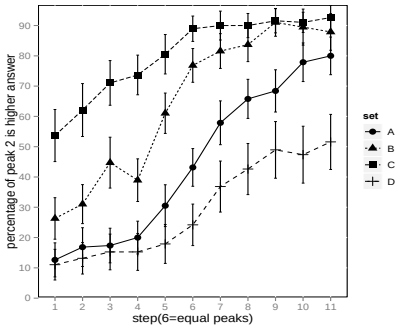

Figure 4: *Pitch classification functions for Chinese listeners. X-axis=F0 steps, y-axis=proportion of "peak 2 is higher" responses; line patterns denote different spectral conditions. Error bars denote 95% confidence intervals.*

3.3. Musicians

Although only a small number of musicians participated in this study, consistent patterns were observed for all individuals. As shown in Figure 5, the classification functions largely converge. Although the effect is weak and is not quite visible in Figure 5, the MCMC generalized linear mixed-effects model still shows significant shifts between the spectral conditions (Table 3) in a direction that is consistent with other participants in this study. Previous studies (e.g., [24][26]) have shown that musicians are generally more sensitive to fundamental frequencies, whereas non-musicians are less sensitive to fundamental frequencies and more often focus on

harmonic intervals in the spectrum. This sensitivity to fundamental frequencies likely exempts musicians from the influence of spectral slope.

	A	B	C
B	0.7 [0.5,0.8] P=0.012		
C	1.4[1.2,1.6] p<0.001	0.44[0.43,0.45] p<0.001	
D	-0.05[-0.5,0.1] p>0.5	-0.27[-0.28,-0.26] p<0.001	-1.5[-2.0,-1.0] p<0.001

Table 3. *Main effects of spectral conditions for every pair of conditions. Means of regression coefficients followed by 95% highest posterior density intervals in square brackets and associated p-values.*

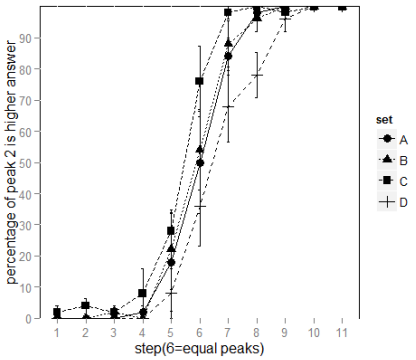

Figure 5: *Pitch classification functions for Musicians. X-axis=F0 steps, y-axis=proportion of "peak 2 is higher" responses; line patterns denote different spectral conditions. Error bars denote 95% confidence intervals.*

4. Discussion

This study sought to examine whether voice quality cues are integrated with pitch perceptions. The major finding of the present study is that the pitch classification function can shift significantly under various spectral conditions, which indicates that spectral cues strongly influence pitch perceptions. Moreover, listeners generally perceive higher pitches when higher-frequency components in the spectrum have more energy (indicating a tenser voice quality [8]). The direction of a shift is consistent with the co-varying relationship between F0 and voice quality: high F0 is naturally produced by a tense voice [6]. This study thus strongly supports the hypothesis that voice quality cues and F0 are integrated in pitch perceptions and that voice quality is a strong indicator of pitch range.

The findings of this study have important implications for pitch-related linguistic studies: pitch is not merely F0, either in production or in perception. As suggested in this study, pitch range is determined by both F0 and voice quality cues. Thus, what is perceptually "higher" does not necessarily have a higher F0. This finding is instrumental to speaker normalization, as voice quality can provide information on pitch location for a speaker; for example, a tense voice indicates that the speaker has nearly reached his/her highest range (or is speaking at his/her highest pitch). The capacity to integrate F0 and voice quality is also useful in processing multiple contrastive levels in languages (e.g., [22]). The strong interaction between F0 and voice quality in production and perception should thus be considered when categorizing tones and prosodic structures.

Our findings of listener variability between different language groups and between musicians and non-musicians are also interesting. Listeners appear to employ different strategies when interpreting "pitch," and F0 and spectral cues have different weights among listeners. Although spectral condition effects are consistent across all groups, tonal language experience and music training can affect the magnitude of shifts.

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] D. Honorof and D. Whalen, "Perception of pitch location within a speaker's F0 range," *J. Acoust. Soc. Am.*, vol. 117, pp. 2193-2200, 2005.

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

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

[4] H. Hollien and J. F. Michel, "Vocal fry as a phonational register," *Journal of Speech and Hearing Research* vol. 11, p. 600 1968.

[5] H. Hollien, "On Vocal registers," *Journal of Phonetics* vol. 2, pp. 125-143 1974.

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

[7] B. Roubeau, N. Henrich, and M. Castellengo, "Laryngeal Vibratory Mechanisms: The Notion of Vocal Register Revisited," *Journal of Voice*, vol. 23, pp. 425-438, 7// 2009.

[8] C. Gobl and A. Ní Chasaide, "Voice source variation," in *The Handbook of Phonetic Science*, W. J. Hardcastle and J. Laver, Eds., ed Oxford: Blackwell, 2012, pp. 378-423.

[9] J. E. Andruski, "Tone clarity in mixed pitch/phonation-type tones," *Journal of Phonetics*, vol. 34, pp. 388-404, 2006.

[10] J. E. Andruski and M. Ratliff, "Phonation types in production of phonological tone: the case of Green Mong," *Journal of the International Phonetic Association*, vol. 30, pp. 37-61, 2000.

[11] B. Blankenship, "The timing of nonmodal phonation in vowels," *Journal of Phonetics*, vol. 30, pp. 163-191, 2002.

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

[13] E. Thurgood, "Phonation types in Javanese," *Oceanic Linguistics* vol. 43, pp. 277–295, 2004.

[14] A. L. Miller, "Guttural vowels and guttural co-articulation in Ju$\vert$'hoansi," *Journal of Phonetics*, vol. 35, pp. 56-84, 2007.

[15] C. T. DiCanio, "The phonetics of register in Takhian Thong Chong," *Journal of the International Phonetic Association*, vol. 39, pp. 162-188, 2009.

[16] C. M. Esposito, "Variation in contrastive phonation in Santa Ana Del Valle Zapotec," *Journal of the International Phonetic Association*, vol. 40, pp. 181-198, 2010.

[17] M. Garellek and P. Keating, "The acoustic consequences of phonation and tone interactions in Jalapa Mazatec," *Journal of the International Phonetic Association*, vol. 41, pp. 185-205, 2011.

[18] J. kuang and P. Keating, "Glottal articulations in tense vs. lax phonation contrasts," *J. Acoust. Soc. Am.*, vol. 136, pp. 2784–2797, 2014.
[19] C. M. Esposito, "An acoustic and electroglottographic study of White Hmong phonation," *Journal of Phonetics*, vol. 40, pp. 466-476, 2012.
[20] S. u. D. Khan, "The phonetics of contrastive phonation in Gujarati," *Journal of Phonetics*, vol. 40, pp. 780-795, 2012.
[21] J. Kreiman, B. Gerratt, and N. Anto\~nanzas-Barroso, "Measures of the glottal source spectrum," *Journal of Speech, Language, and Hearing Research*, vol. 50, pp. 595-610, 2007.
[22] J. kuang, "The tonal space of contrastive five level tones," *Phonetica*, vol. 70, pp. 1-23, 2013.
[23] N. Ryant, M. Slaney, M. Liberman, E. Shriberg, and J. Yuan, "Highly Accurate Mandarin Tone Classification In The Absence of Pitch Information," in *International conference on Speech Prosody*, Dublin, 2014, pp. 673-677.
[24] D. R. Ladd, R. Turnbull, C. Browne, C. Caldwell-Harris, L. Ganushchak, K. Swoboda, *et al.*, "Patterns of individual differences in the perception of missing-fundamental tones," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 39, p. 1386, 2013.
[25] D. Schön, C. Magne, and M. Besson, "The music of speech: Music training facilitates pitch processing in both music and language," *Psychophysiology*, vol. 41, pp. 341-349, 2004.
[26] G. Peng, D. Deutsch, T. Henthorn, D. Su, and W. S. Wang, "Language experience influences nonlinguistic pitch perception," *Journal of Chinese Linguistics*, vol. 41, p. 448, 2013.
[27] A. Seither-Preisler, L. Johnson, K. Krumbholz, A. Nobbe, R. Patterson, S. Seither, *et al.*, "Tone sequences with conflicting fundamental pitch and timbre changes are heard differently by musicians and nonmusicians," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 33, p. 743, 2007.
[28] J. Pierrehumbert, "The perception of fundamental frequency declination," Journal of the Acoustical Society of America, vol. 66, pp. 363-369, 1979.