INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Pitch-range perception: the dynamic interaction between voice quality and fundamental frequency

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

[1] Department of Linguistics, University of Pennsylvania, U.S.A.
kuangj@sas.upenn.edu, markyliberman@gmail.com

Abstract

Effective pitch-range normalization is important to uncover intended linguistic pitch targets in continuous speech. Our previous study demonstrated that voice quality plays a role in pitch-range perception: "tense voice", implemented as stimuli with spectral balance tilted towards higher frequency, was perceived as higher in pitch. This psychoacoustic effect is consistent with the co-variation between pitch and tense voice in production. However, a spectral balance tilted towards higher frequency is also one of the properties of creaky voice, which is often associated with low pitch in production. Therefore, this raises the possibility that manipulating the f0 range of the stimuli or changing the sex of the speaker of the stimuli can reverse the direction of the shift. This current study replicates the previous experiment with the same forced-choice pitch classification experiment with four spectral conditions, but uses a female voice to create the stimuli. In addition, two f0 ranges are used in the current experiments, which resemble the lower range and the higher range of a female voice. Overall, the results show that spectral balance interacts with f0 range: the presence of voice quality cues affect the perception of pitch range; but, the spectrum with greater energy in the high-frequency range can be interpreted as either creaky or tense depending on the f0 range. This current study enriches our understanding of the interaction between voice quality and pitch.

Index Terms: pitch perception, voice quality, f0, spectral slope

1. Introduction

Effective pitch-range normalization is an important task in speech processing, because f0 range varies from speaker to speaker and from context to context, leading to overlapped signals for phonetic categories (e.g. tonal categories). So human listeners or a tone classifier need to uncover intended linguistic pitch targets. Studies [1-3] demonstrated that human listeners' ability of speech normalization is quite impressive: they are able to reliably identify the pitch location of very brief voice samples in an unknown speaker's range in the absence of any contextual cues. Voice quality has been speculated to contribute to such effectiveness. Indeed, f0 and voice quality are co-varying in pitch production studies (singing [4-7]; speech [8]): The lowest pitch range is associated with vocal fry or creaky voice, and the highest pitch range is associated with tense voice and falsetto. The question is then whether this co-variation also occurs in the perception domain.

Our previous studies [9,10] have created a paradigm to test the effect of voice quality cues on the pitch-classification function. Our manipulation has mainly focused on spectral balance, since it has been well established that spectral balance of the voice source spectrum is an important indicator of voice quality: a relatively steep spectral slope (i.e. less energy in the high-frequency region) is associated with a breathier voice and that a flat spectral slope (i.e. greater energy in the high-frequency region) is associated with a tenser or creakier voice.

Our previous studies have shown that: "tense voice", implemented as stimuli with spectral balance tilted towards higher frequency, was indeed perceived as higher in pitch. This is true regardless of the language background of the listeners. This psychoacoustic effect is consistent with the co-variation between pitch and tense voice in production. However, a spectral balance tilted towards higher frequency is also one of the characteristics of creaky voice/vocal fry, which is often associated with low pitch in production (see [11] for a review; cross-linguistic studies [12-22]). Therefore, this raises an ambiguity in interpreting the voice quality. As can bee seen in Figure 1, the relationship between f0 and voice quality is wedge-shaped: the mid range of a speaker's voice has the most relax voice (with greater H1-H2), and when pitch is moving towards the higher or lower limit the f0 range, the voice quality becomes tenser or creakier (with smaller H1-H2). (In fact, the low f0 can be also produced with a breathy voice, as indicated by the green dots in the Figure 1. We will come back to this point in discussion.)

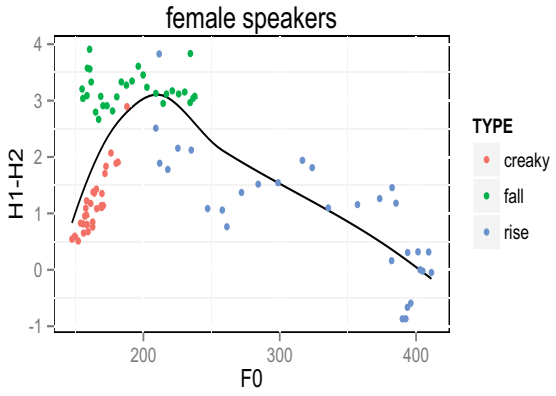


Figure 1: Co-variation between spectral slope and f0, taking from [8]; spectral slope is exemplified by H1-H2 here.

In our previous experiment, the stimuli were created from a male voice and the f0 range of the stimuli was set to 153.06 Hz - 187.36 Hz, which is the mid to high range for a male voice; however, this f0 range belongs to the low range for a female voice. In the male voice condition, a flat spectrum

http://dx.doi.org/10.21437/Interspeech.2016-1483

indicates a tense voice; but in the female voice condition, it is possible that the same spectral property can be interpreted as creaky voice. Because of the different interpretation of the voice quality cues, it is possible that the direction of the shift of the pitch-classification function can be reversed, i.e., stimuli with spectral balance tilted towards higher frequency are instead perceived as lower in pitch.

In sum, this study will continue to explore the interaction between f0 and voice quality in the perception domain. We will continue use the paradigm we used in the previous studies, but add two more manipulations: 1) use the same range of f0 contours as before, but with a female voice; 2) raise the range of f0 contours to the mid-high range of a female voice. The results will be compared with the ones based on male voice. We will test whether f0 range and speaker sex have any effects on the shift direction of the pitch perception.

2. Method

2.1. Stimuli

Same as the previous experiment, the goal was to create four sets of utterances with two f0 peaks, and the two f0 peaks vary in spectral conditions. In this current experiment, each peak was carried by three /ma/ syllables, so that the whole sequence had the prosodic pattern of a phrase like "phonetic condition" (i.e. both words have a weak-strong-weak stress pattern). The stimuli were resynthesized from the natural production of a female English speaker. The speaker was asked to produce tokens of /ma.ˈma.ma/ with the same intonation pattern as "two twenty".

For f0 manipulation, a Hann function (a cosine period with the peak in the middle) was used for each peak. The base was the same for both peaks; and the maximum value of the first peak was kept the constant for all stimuli, while the second peak was an 11-step continuum with 0.35-semitone interval. The f0 contours are shown in Figure 2. In this current study, two f0 ranges were used:

- "Female low" condition: the first peak was set to 169.34 Hz, and the range for the second peak was set to 153.06 Hz - 187.36 Hz. This setting is the same as our previous study based on a male voice [10].
- "Female high" condition: the first peak was set to 220 Hz, and the range for the second peak was set to 198.8 Hz - 243.4 Hz. This range resembles the mid to high range of a female voice.

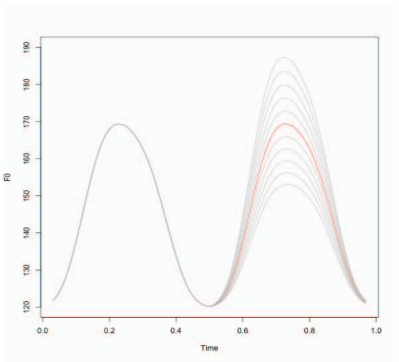
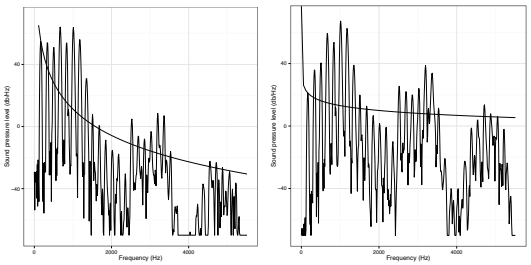


Figure 2: *f0 manipulation: the first peak has a constant f0 value, and the second peak is a continuum with 11 steps. Peaks 1 and 2 are identical at step 6 (red/dark lines for the second peak).*

To manipulate voice quality cues, two versions of spectral balance were created: one with relatively more high-frequency energy (i.e. tenser/creakier version) and one with relatively less high-frequency energy (i.e. breathier version). The breathier version was the original spectrum of the natural production, while the tenser/creakier version was modified so that the Fourier spectrum was 6 dB/octave greater than the breathier version. This modification corresponded to a differentiation operation of the Fourier spectrum, due to the derivative-differentiation property of Fourier transform, therefore, it is precisely equivalent to the spectral contrast in our previous study. The result of this spectral boost is depicted in Figure 3.



(a) (b)

Figure 3: *Spectral manipulation: (a) original; (b) boosted*

Finally, the modified parameters were combined and resynthesized into 22 tokens of different f0 peaks (11 steps) and spectral slope (2 values). These single peaks were 0.52 seconds in duration and then concatenated to create 4 sets of two-peak stimuli, labeled with letter A-D in the same way as previous (44 stimuli in total). Thus there were 4 different spectral conditions in the stimuli:

- Set A: Both peaks have the original spectrum (reference condition)
- Set B: Both peaks have the boosted spectrum (reference condition, tenser/creakier)
- Set C: The first peak has the original spectrum, and the second has the boosted spectrum, with a 200 ms transition in the middle (original + tenser/creakier)
- Set D: The first part has the boosted spectrum, and the second has the original spectrum, with a 200 ms transition in the middle (tenser/creakier + original)

Therefore, there were 44 stimuli (11 f0 steps x 4 spectral conditions) in a total for each experiment. All stimuli were 1.05 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. Five 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 "maMAma" word was higher or lower than the first one by clicking on the corresponding buttons on the

computer screen. To introduce the idea of pitch to an English speaker, we used the examples of English intonation. For example, the word "my name" is higher in "Anna may know my name?" than in "Anna may know my name." In the practice session, examples from set A were used to demonstrate the task. This was to make sure that listeners would attend to pitch difference but not other cues (e.g. intensity). The experiment was run with Qualtrics online survey system. The subjects were instructed to use headphones or earbuds to do the experiment.

2.3. Subjects

English speakers were recruited from the Amazon Mechanical Turk. There are 40 listeners in each experiment, and there are no overlapping between the two subject pools. All the subjects reported to have normal hearing and speaking.

3. Results

3.1. Experiment 1- "female low" condition

The first experiment replicates the previous experiment with a female voice. This condition has the same f0 range as our previous experiment in which the stimuli were resynthesized from a male voice. The f0 range (153 – 187 Hz) is a mid to high range for a male, but it belongs to a mid to low range for a female speaker. According to Figure 1, it is possible that the stimuli with boosted spectrum may be perceived as creaky voice instead of tense voice; and the different interpretations would have different effects on pitch perception. Specifically, our predictions for the current experiment are as follows:

Compared with set A and B, where the two peaks have identical spectral conditions,

1) If listeners interpret the stimuli with boosted spectrum as tense voice, then we would expect that set C (original + tenser) receives the most "peak 2 is higher" responses; set D (tenser + original), on the opposite, should receive the least "peak 2 is higher" responses -- same as the previous experiment, reproduced here as Figure 4.

2) However, if listeners interpret the stimuli with boosted spectrum as creaky voice, then we would expect a reversed pattern: set C (original + creakier) receives the least "peak 2 is higher", while set D (creakier + original) receives the most "peak 2 is higher".

The pitch-classification functions of the current experiment are presented in Figure 5. Overall, compared with set A and B, where the two peaks have identical spectral conditions, the pitch-classification functions for set C and set D are significantly shifted; however, the direction of shifts of set C and set D is reversed from Figure 4. In Figure 4, set C receives the most "peak 2 is higher" responses, while set D receives the least "peak 2 is higher" responses; but in Figure 5, set D now receives the most "peak 2 is higher" responses. This difference indicates that listeners in the current experiment generally interpret the stimuli with more energy in the high-frequency domain as creaky voice, and thus perceive the peak with a creakier voice as lower in pitch.

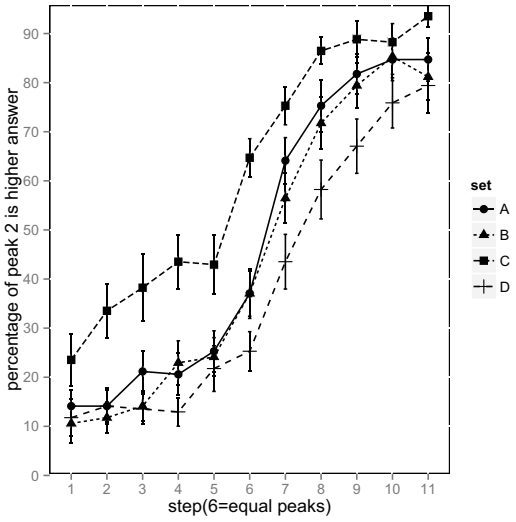


Figure 2: *From pervious study: Pitch-classification functions based on a male voice.*

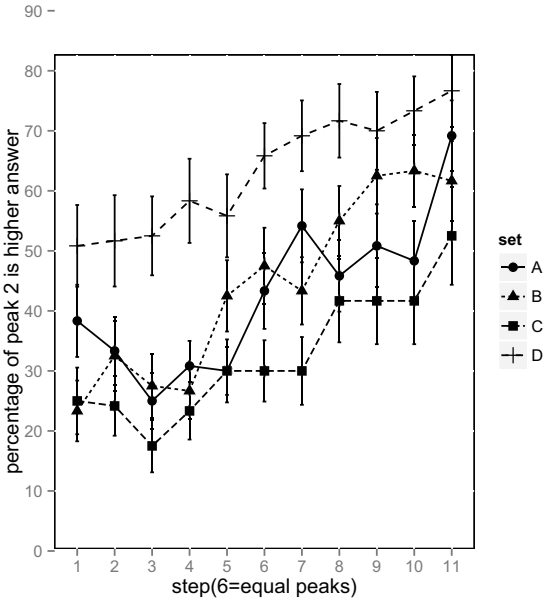


Figure 5*: Pitch classification functions for "female low" condition. X-axis=f0 steps, y-axis=proportion of "peak 2 is higher" responses; line patterns denote different spectral conditions. Error bars denote 95% confidence intervals.*

This experiment shows that listeners are more likely to interpret the spectral balance with more energy in high-frequency region as creaky voice, when the f0 range of the stimuli is in the lower side of the speaker's range. To further validate this result, in the second experiment, we will raise the f0 range of the stimuli to resemble the mid to high range of a female voice. We would expect that the pitch-classification functions should pattern like Figure 4: the spectral balance with more energy in high-frequency region is interpreted as

tense voice, and thus the peak with tense voice would sound higher to the listeners.

3.2. Experiment 2 – "female high" condition

The stimuli and procedure of experiment 2 are exactly the same, except that the f0 range is raised to 198 Hz to 243 Hz. Another 40 English speakers, who are not in the subject pool of the first experiment, participated in this experiment.

Figure 6 shows the pitch-classification functions of the "female high" condition. It is clearly that Figure 6 faithfully replicates the shift direction in Figure 4 (essentially the "male high" condition), with set C dominated by "peak 2 is higher" responses. This means that in the higher f0 range of a female voice, the boosted spectrum is interpreted as tense voice, and thus leads to the perception of a higher pitch.

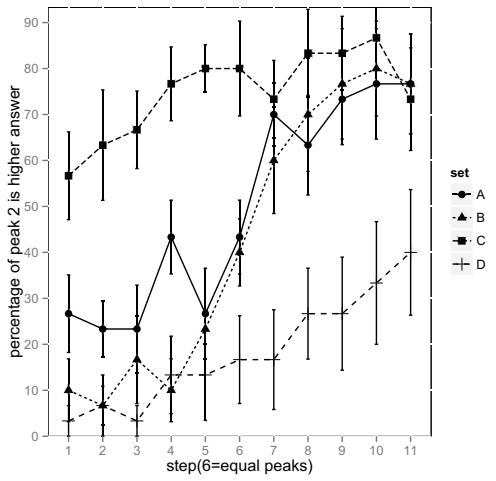


Figure 6: *Pitch classification functions for "female high" condition. X-axis=f0 steps, y-axis=proportion of "peak 2 is higher" responses; line patterns denote different spectral conditions. Error bars denote 95% confidence intervals.*

A further inspection of the individual patterns (figures not shown here due to space limitation) suggests that: little individual variation was found in the female high condition (can be evident by the nice separation of set C and D from set A and B in Figure 6); but quite a bit individual variation was found in the female low condition: some listeners still interpret the boosted spectrum as tense voice, and thus set C receives the most "peak2 is higher" responses.

4. Discussion

This study builds on our previous studies on the effects of voice quality on pitch perception. The results of this study continue demonstrating that voice quality can introduce significant shifts in the pitch-classification functions, and the direction of shifts is consistent with co-variation between pitch and voice quality in production. Both tense voice and creaky voice share the similar spectral balance, with more energy in the high-frequency domain, but they have distinctive f0 profiles. This study looks into how listeners interpret spectral balance in different f0 range, and how it can affect the perception of pitch height. We first resynthesized the stimuli with a female voice, thus the relative f0 range of the stimuli is changed from "male high" to "female low". We found this manipulation significantly change the direction of the shifts of the pitch-classification functions. Listeners interpret the boosted spectrum as tense voice when f0 stimuli are in the higher side of the speaker's range, while they are more likely to interpret the boosted spectrum as creaky voice when f0 stimuli are in the lower side of the speaker's range. And the different interpretation of voice quality in turn leads to opposite direction of shift of the pitch perception. Therefore, pitch perception is the result of the dynamic interaction between f0 and voice quality.

Moreover, our individual analysis suggests that some listeners still interpret the boosted spectrum as tense voice in the lower f0 range. This can be understood by the different variability of the realization of low f0. In production, as shown in Figure 1, low f0 can be produced by either breathy voice or creaky voice. People may use different strategies in perception as well. It should be noted that creaky voice is often cued by irregular pulse-to-pulse variation as well. Our previous study shows that adding Gaussian jitter to the signal can consistently lead to people to perceive a lower pitch. Perhaps jitter is a more reliable cue in perceiving creaky voice.

All in all, this study further enriches our understanding of the interaction between pitch and voice quality. Through a series of experiments, both the current and previous studies, we have gradually identified the acoustic cues beyond f0 that are crucial for pitch perception. We suggest that pitch analysis and synthesis should take voice quality cues into account.

.

5. 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," *J. of the Acoust. Soc. Am.,* 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, 2009.

[8] J. Kuang, The covariation between pitch and phonation: creaky voice in Mandarin tones. The 89th Annual Meeting of the Linguistic Society of America, 2015.

[9] J. Kuang and M. Liberman, "Influence of spectral cues on the perception of pitch height", Proceeding of ICPH 18, 2015.

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

[11] 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.

[12] 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.

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

[14] 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.

[15] E. Thurgood, "Phonation types in Javanese," *Oceanic Linguistics* vol. 43, pp. 277–295, 2004.
[16] A. L. Miller, "Guttural vowels and guttural co-articulation in Ju|'hoansi," *Journal of Phonetics,* vol. 35, pp. 56-84, 2007.
[17] C. T. DiCanio, "The phonetics of register in Takhian Thong Chong," *Journal of the International Phonetic Association,* vol. 39, pp. 162-188, 2009.
[18] 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.
[19] 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.
[20] J. Kuang and P. Keating, "Glottal articulations in tense vs. lax phonation contrasts," *J. Acoust. Soc. Am.,* vol. 136, pp. 2784–2797, 2014.
[21] C. M. Esposito, "An acoustic and electroglottographic study of White Hmong phonation," *Journal of Phonetics,* vol. 40, pp. 466-476, 2012.
[22] S. D. Khan, "The phonetics of contrastive phonation in Gujarati," *Journal of Phonetics,* vol. 40, pp. 780-795, 2012.