The effect of vocal fry on pitch perception

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

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

Abstract

Vocal fry is an aperiodic phonation that is naturally associated with low pitch production. This study aims to examine whether vocal fry can affect people's perception of pitch range. A two-alternative-forced-choice task is used to test whether vocal fry can facilitate the "low pitch" perception, and non-fry and fry sounds are played in pairs. The synthetic vocal fry stimuli vary in the f0 range and the ratio of pulse-to-pulse variability. The results show that people generally hear a lower pitch during vocal fry, and the magnitude of this effect depends on the f0 range as well as the ratio of fry. Heavy fry almost always sounds lower. The finding of this study further supports the hypothesis that voice quality is integrated in pitch perception.

Index Terms: speech perception, voice quality, vocal fry

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. [1]), but studies ([2][3]) 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 [3] and [4] speculated that voice quality could be such a cue.

Indeed, Kuang and Liberman [5] found that spectral slope, one of the most important voice quality cues, can significantly shift the perception of pitch height: Listeners tend to hear a higher pitch in the presence of the tense voice, which is naturally associated with a high F0 in pitch production [6-9]. Based on this previous study, we hypothesize that, vocal fry, the voice quality that is associated with the low end of the pitch scale, should bias the listeners to perceive a lower pitch.

The term "vocal fry" (or creak, creaky voice, laryngealization, glottalization [10-18] has been used to cover a broad range of phonations in literature (see [19] for a review). Vocal fry is usually defined as a train of discrete glottal pulses of very low frequency, with almost complete damping of the vocal tract between the pulses [16, 6]. And the singing literature [6-9] treats vocal fry as the low end of the pitch scale. The modal register for male speakers is approximately 86-170 Hz for males, and 175-240 for females. And the f0 of vocal fry is ranged from 31.6 to 69.1 Hz in [16], 10.9-52.1 Hz in [20], and 22-92 Hz in [21], 18-65 Hz in [22], 24-77 Hz in [23].

Vocal fry can be perfectly periodic, and the perception of "fry" is because the f0 is very low, listeners are able to hear the individual pulses. Vocal fry can also be period-doubled or tri-bled. Based on the properties of acoustic signals, vocal fry can be characterized into several categories (after [10, 19]):

- High degree of pulse-to-pulse irregularity, in both duration (jitter) and shimmer (amplitude).
- Fairly regular with strong secondary excitation peaks
- Fairly regular without strong secondary excitation peaks

Listeners are able to distinguish these different types of vocal fry [19]. Vocal fry commonly co-occurs with low tones in many languages, e.g. Mandarin and Cantonese. Figure 1 presents two examples of variations of vocal fry for Mandarin tone 3. [24] shows that vocal fry is not tied to certain tonal category, but co-varies with the f0 scale.

Since the periodic type of vocal fry is essentially low f0, this study will focus on the aperiodic type, as it is able to combine with different f0 ranges. By definition, this type of vocal fry has random period-to-period variability (jitter) (c.f. Figure 1 bottom). Therefore, we will use this parameter to synthesize vocal fry.

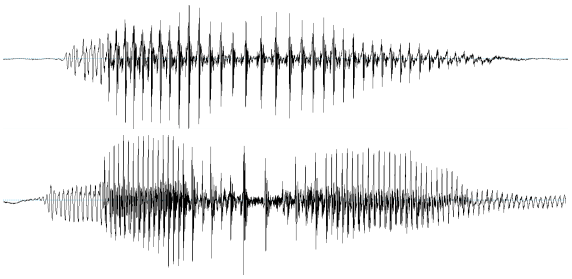

Figure 1: *Examples of vocal fry in Mandarin Tone 3. Top: example of regular pulses; Bottom: example of irregular oscillation.*

2. Method

2.1. Stimuli

A series of pulse trains with steady f0 were synthesized. Figure 2 and 3 lay out the schema of the stimuli. Three seven-stepped f0 continua were created, with the step interval at one semitone apart. So the highest and lowest steps were three semitones from the centers (i.e. the fourth step). The centers of these three f0 continua were at 50 Hz, 70 Hz and 90 Hz respectively. As reviewed before, 50 Hz is the average f0 range for pulse register across studies and across genders; 70 Hz is about the upper limit of pulse register, and [25] claims that people are able to hear individual pulses at this range; and 90 Hz is above the range of pulse register, but still at the low range of a male speaker.

To synthesize the "fry" version of the stimuli, random jitter was added into the regular buzz. We characterized jitter as the standard deviation of a presumed Gaussian distribution of periods. A normal sustained voice has a jitter of about 1.0%, and detectable jitter is about 2% (c.f. [26]). 8% and 15% were reported in natural vocal fry production [23], and a ratio above 20% was reported for the extreme cases [27]. Therefore, we set the ratio of jitter at 0% (i.e. regular buzz), 2%, 8%, 15% and 25%, to represent the different degrees of vocal fry attested in natural production.

Therefore, there were 105 stimuli in a total: 7 f0 steps (1-7) x 3 f0 ranges (50 Hz, 70 Hz, and 90 Hz) x 5 ratios of jitter. The duration for all stimuli was set to 1s.

2.2. Procedure

In a two-alternative-forced-choice pitch classification task, stimuli were presented in pairs. The stimuli were arranged into three blocks based on f0 ranges, and within each block: 1) for the first half, the first sound was always a regular buzz at step 4, and the second sound was the 35 sounds (7 steps x 5 ratios); 2) for the second half, the order of the stimuli was reversed (i.e. the first sound was the target, and the second sound was the reference). Repetition was set to 2, so there were 420 trials in a total. Randomization was done across all 420 trials.

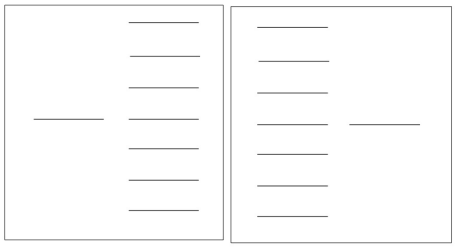

Figure 2: *Schema of F0 manipulation. AB order (left): the first sound is constant, and the second sound is a F0 continuum. At step four, the two sounds have the same F0. For the BA order, the constant sound is at the second position. The centers (step four) for the three pitch-range conditions were at 50 Hz, 70 Hz and 90 Hz.*

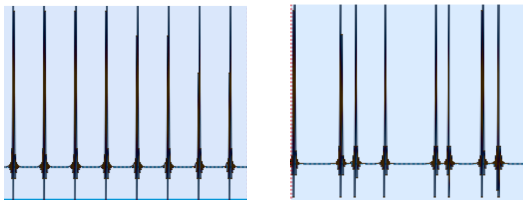

Figure 3: *Examples of jitter manipulation. Left: periodic condition (jitter ratio=0%); right: aperiodic condition (jitter ratio =25%).*

For each trial, the listeners were asked to attend to pitch, and judge whether the second sound is higher or lower than the first sound by clicking on the correspondent buttons on the computer screen. All testing took place in a sound booth with stimuli presented over Sennheiser 280 headphones.

2.3. Subjects

20 English speakers, aged between 18 and 22, were recruited from the student population at the University of Pennsylvania. None of them reported to have hearing or voice issue.

3. Results

The results are subset by f0 ranges. Figure 4- 6 present the percentage of "the target sound is higher", which averages across both AB and BA orders for all listeners. The right half of these figures (step 4-7) illustrates when the target stimulus has a higher f0 than the reference stimulus. As can be seen here, compared to the non-fry (i.e. 0% ratio) condition, listeners generally less likely to judge the target sound to have a higher pitch in the presence of fry. The main effects of ratio conditions were evaluated using an MCMC generalized linear

mixed-effects model (*mcmcglmm* package in R). F0 steps (1-7) and ratios conditions (0%, 2%, 8%, 15% and 25%) were used as fixed factors, and random intercepts and slopes were included as subjects. The main effects of the ratio are summarized in Table 1-3. The results are reported as means of regression coefficients followed by 95% highest posterior density intervals in square brackets and associated p-values.

For the 50 Hz condition (Figure 4), the 2% jitter does not make significant differences from the regular buzz (0% ratio). 8% and 15% jitter conditions are slightly different from the reference condition, but do not reach significance. Nonetheless, the 25% jitter condition has a significant shift -- the percentage of "the target is higher" does not pass the 50% threshold until the target stimulus is two semitones higher than the reference stimulus.

The classification functions for 70 Hz condition (Figure 5) is similar to the 50 Hz condition, but the magnitude of shift is greater for the fry stimuli. Particularly, the stimuli with 25% jitter never pass the 50% threshold, meaning that the sounds with heavy fry always sound lower to the listeners. And both 8% the 15% jitter conditions are significantly different from the reference condition, and the fry stimulus needs to be two semitones higher than the reference in order to sound higher. The 90 Hz condition (Figure 6) is very similar to the 70 Hz condition (Figure 5), although that the magnitude of the shift is even greater (shown in Table 2 and Table 3).

In addition, it is noticeable that the classification functions for the 50 Hz condition is less categorical than the ones for 70 Hz and 90 Hz conditions.

Table 1. *Main effects of the jitter ratio for 50 Hz condition. Summary of means of regression coefficients, 95% highest posterior density intervals, and p-values.*

	mean	*l-95% CI*	*u-95% CI*	*pMCMC*	
(Intercept)	-1.69	-4.07	-0.002	0.036	
2%	0.50	-0.79	2.23	0.454	
8%	0.36	-1.08	1.84	0.586	
15%	0.4	-0.94	2.14	0.474	
25%	-1.66	-3.81	-0.24	0.022	*

Table 2. *Main effects of the jitter ratio for 70 Hz condition. Summary of means of regression coefficients, 95% highest posterior density intervals, and p-values.*

	mean	*l-95% CI*	*u-95% CI*	*pMCMC*	
(Intercept)	-1.24	-2.73	0.26	0.10	.
2%	-0.27	-1.47	1.05	0.68	
8%	-1.69	-3.15	-0.42	0.00	**
15%	-2.63	-4.14	-1.14	<0.001	***
25%	-3.51	-5.19	-1.73	<0.001	***

Table 3. *Main effects of the jitter ratio for 90 Hz condition. Summary of means of regression coefficients, 95% highest posterior density intervals, and p-values.*

	mean	*l-95% CI*	*u-95% CI*	*pMCMC*	
(Intercept)	-3.95	-7.33	-0.68	0.01	**
2%	-0.98	-3.37	1.27	0.39	
8%	-4.15	-7.09	-1.67	<0.001	***
15%	-4.81	-7.92	-1.89	<0.001	***
25%	-10.15	-15.28	-6.34	<0.001	***

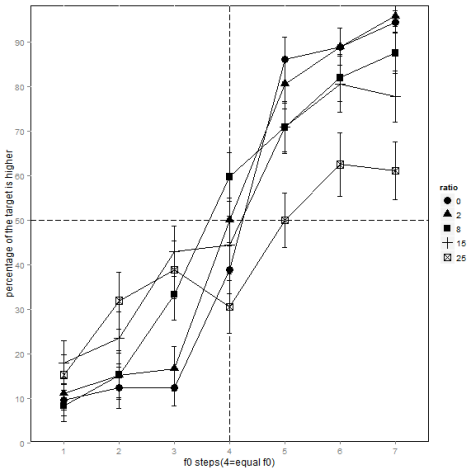

Figure 4: Pitch *classification function for 50 Hz stimuli. X-axis represents the seven f0 steps, y-axis represents the percentage of "the target sound is higher"; different lines represent the different ratios of jitter.*

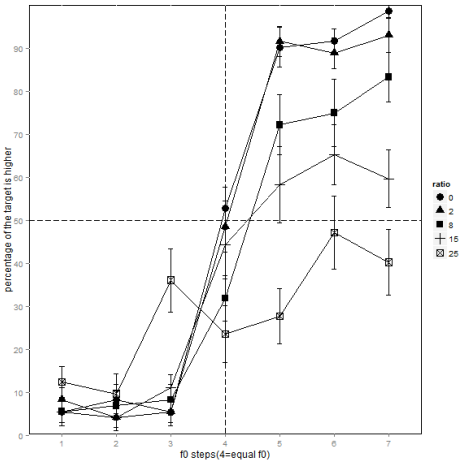

Figure 5: Pitch *classification function for 70 Hz stimuli. X-axis represents the seven f0 steps, y-axis represents the percentage of "the target sound is higher"; different lines represent the different ratios of jitter.*

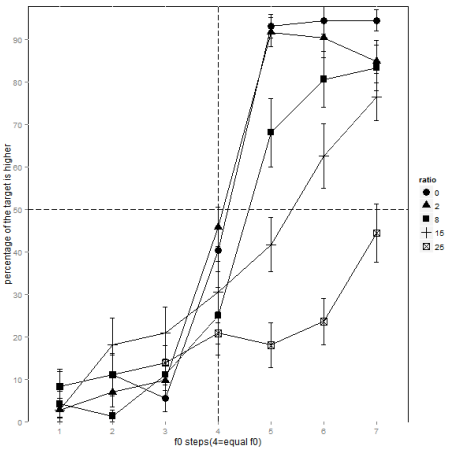

Figure 6: Pitch *classification function for 90 Hz stimuli. X-axis represents the seven f0 steps, y-axis represents the percentage of "the target sound is higher"; different lines represent the different ratios of jitter.*

4. General discussion

This study aims to examine whether vocal fry can affect people's perception of pitch range. A pitch classification task is used to test whether vocal fry can facilitate the "low pitch" perception. The results show that people generally hear a lower pitch during vocal fry, and the magnitude of the effect depends on the f0 range as well as the ratio of fry. Heavy fry almost always sounds lower, even if its f0 is several semi-tones higher than the non-fry stimulus in the same trial.

[5,24] proposed that since vocal fry is naturally associated with the low end of f0 scale, it is possible that people can utilize this voice quality to cue low pitch. The findings of this study support this hypothesis. It is possible that listeners treat vocal fry as extremely low pitch. Cross-linguistically, vocal fry has been commonly found to co-occur with low tones, and perception studies [28, 29] showed that the presence of vocal fry facilitated the perception of low tones. This study demonstrates that this effect is not language specific, but subject to universal psychoacoustic mechanisms. The fact that the aperiodic type of vocal fry can combine with different f0 range is especially useful, because speakers do not need to dip very far in f0 scale to sound low. Therefore, vocal fry can enhance the contrast between low and non-low tones in both production and perception.

It is also interesting that the magnitude of the shift for the 50 Hz condition is substantially smaller than higher-frequency conditions. By definition, vocal fry can simply be very low f0. As the frequency slow enough, people are able to hear individual pulses. So a 50 Hz regular buzz also sounds like vocal fry. This type of vocal fry is different from the aperiodic phonation discussed in this paper. However, it is possible that these two types of vocal fry are not very distinctive for very low f0 due to overall low continuity. Since listeners are able to hear individual pulses, and thus less sensitive to the periodicity, it requires a greater amount of jitter to let the listeners notice the irregularity. Finally, the magnitude of the shift also depends on the ratio of irregularity. Increasing aperiodicity can lead to more low pitch perception, which indicates that aperiodicity is a strong cue of low f0.

In sum, pitch perception is more than f0, and listeners are able to integrate voice quality cues in pitch range perception, and thus listeners can more successfully identify the linguistic categories, e.g. tones.

5. Acknowledgements

This study was supported by URF award from UPenn to the first author. Thank Jingjing Tan for the assistance of data collection.

6. References

[1] P. C. M. Wong and R. L. Diehl, "Perceptual normalization for inter-and intratalker variation in Cantonese level tones," *Journal of Speech, Language, and Hearing Research*, vol. 46, pp. 413–421, 2003.

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

[3] D. N. Honorof and D. H. Whalen, "Perception of pitch location within a speaker's F0 range," *The Journal of the Acoustical Society of America,* vol. 117, pp. 2193–2200, 2005.

[4] C.-Y. Lee, “Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study,” *The Journal of the Acoustical Society of America,* vol.125, pp. 1125–1137, 2009.
[5] J. Kuang and M. Liberman, “The effect of spectral slope on the perception of pitch height,” Proceeding of *INTERSPEECH* , 2015.
[6] H. Hollien, "On Vocal registers," *Journal of Phonetics,* vol. 2, pp. 125-143, 1974.
[7] H. Hollien and J. F. Michel, "Vocal fry as a phonational register," *Journal of Speech and Hearing Research*, vol. 11, p. 600, 1968.
[8] I. R. Titze, "A framework for the study of vocal registers," *Journal of Voice* vol. 2, pp. 183-194 1988.
[9] 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.
[10] T. Drugman, J. Kane, and C. Gobl, "Data-driven detection and analysis of the patterns of creaky voice," *Computer Speech & Language*, vol. 28, pp. 1233-1253, 2014.
[11] J. Slifka, "Some physiological correlates to regular and irregular phonation at the end of an utterance," *Journal of Voice*, vol. 20, pp. 171-186, 2006.
[12] S. Vishnubhotla and C. Y. Espy-Wilson, "Automatic detection of irregular phonation in continuous speech," in *INTERSPEECH*, 2006.
[13] T. Bőhm, Z. Both, and G. Németh, "Automatic classification of regular vs. irregular phonation types," in *Advances in nonlinear speech processing*, ed: Springer, 2010, pp. 43-50.
[14] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Glottalization of word-initial vowels as a function of prosodic structure," *Journal of Phonetics*, vol. 24, pp. 423-444, 1996.
[15] L. Redi and S. Shattuck-Hufnagel, "Variation in the realization of glottalization in normal speakers," *Journal of Phonetics*, vol. 29, pp. 407-429, 2001.
[16] H. Hollien and R. W. Wendahl, "Perceptual study of vocal fry," *The Journal of the Acoustical Society of America*, vol. 43, pp. 506-509, 1968.
[17] C. T. Ishi, K.-I. Sakakibara, H. Ishiguro, and N. Hagita, "A method for automatic detection of vocal fry," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, pp. 47-56, 2008.
[18] L. Wolk, N. B. Abdelli-Beruh, and D. Slavin, "Habitual use of vocal fry in young adult female speakers," *Journal of Voice*, vol. 26, pp. e111-e116, 2012.
[19] B. R. Gerratt and J. Kreiman, "Toward a taxonomy of nonmodal phonation," *Journal of Phonetics*, vol. 29, pp. 365-381, 2001.
[20] R. E. McGlone, "Air flow during vocal fry phonation," *Journal of Speech, Language, and Hearing Research*, vol. 10, pp. 299-304, 1967.
[21] T. Murry, "Subglottal pressure and airflow measures during vocal fry phonation," *Journal of Speech, Language, and Hearing Research*, vol. 14, pp. 544-551, 1971.
[22] R. E. McGlone and T. Shipp, "Some physiologic correlates of vocal-fry phonation," *Journal of Speech, Language, and Hearing Research*, vol. 14, pp. 769-775, 1971.
[23] M. Blomgren, Y. Chen, M. L. Ng, and H. R. Gilbert, "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *The Journal of the Acoustical Society of America*, vol. 103, pp. 2649-2658, 1998.
[24] J. Kuang, *Phonation in Tonal Contrast*, Ph.D., Linguistics, University of California Los Angeles, Los Angeles, CA, USA, 2013.
[25] I. R. Titze, *Principles of voice production*, 1994.
[26] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *The Journal of the Acoustical Society of America*, vol. 87, pp. 820-857, 1990.
[27] S. A. Cavallo, R. Baken, and S. Shaiman, "Frequency pertubation characteristics of pulse register phonation," *Journal of communication disorders*, vol. 17, pp. 231-243, 1984.
[28] R. X. Yang, "The Phonation factor in the categorical perception of Mandarin tones.," in *Proceedings of ICPhS XVII*, 2011, pp. 2204-2207
[29] K. M. Yu and H. W. Lam, "The role of creaky voice in Cantonese tonal perception," *The Journal of the Acoustical Society of America*, vol. 136, pp. 1320-1333, 2014.