PROBING VOICE QUALITY'S CONTRIBUTION TO TONE PERCEPTION: CHALLENGES FOR SYNTHESIS SOFTWARE

Kristine M. Yu

Department of Linguistics, University of Massachusetts Amherst, USA
krisyu@linguist.umass.edu

ABSTRACT

This paper illustrates capabilities and limitations of current voice synthesis software via a meta-analysis of acoustic properties of perceptual stimuli used to study the role of creaky and breathy voice quality in linguistic tone perception. The KlattGrid synthesizer in Praat allows the manipulation of parameters directly indexing sub-classes of creaky voice quality, but it's unclear that the parameters index acoustic properties of these sub-classes as they occur in natural speech. The UCLA voice synthesizer better enables targeting of particular individual spectral components than KlattGrid in the synthesis of breathy voice quality.

Keywords: synthesis, perception, voice quality, creak, tone

1. INTRODUCTION

Acoustic properties involved in voice quality affect listeners' perceptual judgments of a diverse range of properties of interest across different fields [18]. Yet, designing stimuli for perception experiments involving these properties remains challenging. As a case in point, [1] resorted to creating perceptual stimuli with vocal fry by asking speakers to mimic what they observed in video clips from the Internet. [1] had considered another option: to draw on speech samples with varying voice source characteristics from a corpus of natural speech. But they pointed out that it can be difficult to isolate particular properties of interest from other potential confounding acoustic properties in natural speech. And the authors of [1] were unaware of existing voice quality synthesis software that would offer fine control over specific parameters and thus potentially avoid such confounds. Recent perceptual studies on the role of creak and breathiness in linguistic tone perception have taken advantage of exactly this kind of synthesis software, e.g., [3, 5, 19, 20, 10], particularly the widely used synthesizer developed in [14] as implemented in Praat [2] by [25] (henceforth "KlattGrid"). This paper thus takes a closer look at the properties of stimuli generated by KlattGrid and also compares them with those of stimuli generated by another voice quality synthesizer, the UCLA voice synthesizer [15], as a case study for looking at the promise/challenge of synthesis tools for voice quality perception. I show that in some cases, KlattGrid synthesis parameters don't necessarily have a clear/straightforward relation to acoustic properties most relevant for actual human speech production and perception, and I enumerate desiderata for the further development of voice quality synthesis tools.

Work on the role of creak and breathiness in tonal perception offers a useful case study for thinking about what is needed from voice quality synthesis tools. First, the synthesizer needs to be able to manipulate f0 and other voice source parameters independently. This is because of two key research questions seeking to disentangle f0 from other acoustic properties in tone perception: (i) how is creak integrated with f0 in the perception of low tones [19, 20, 10], and (ii) assuming there are acoustically well-defined kinds of creak (including some with low f0, and others without), how these kinds interact with tone category perception [13, 4, 10, 27]. [8]'s early study of creak in Mandarin tone perception resynthesized ("biphasic") creak by halving f0—back then, tools were not available to separate creak from f0 manipulation.

Second, in order to assess the linguistic role of kinds of creak, the synthesizer needs to be able to be able to generate distinct exemplars of these different kinds. Few tone perception studies have distinguished kinds of creak. One kind of creak that could shed light on the integration of f0 and creak in tone perception is multiply/doubly pulsed (period-doubled) creak. Period-doubled creak consists of alternating longer and shorter pulses and/or higher and lower amplitudes, and importantly is not contingent on low f0 [7, 16, 13]. [26] used a library of period-doubled creak examples to show that listeners are sensitive to non f0-contingent creak in the contrast between T21 (often creaky) and T22 in Cantonese. However, as pointed out in [27], the use of natural productions of period-doubled creak there may have included stimuli with some mix of period-doubled creak and other kind(s) of creak. To circumvent confounds like this introduced by relying on naturally produced sitmuli, [10, 27] used KlattGrid to create different kinds of creak in studies of Mandarin and Cantonese tone perception. (The UCLA voice synthesizer currently isn't designed to tackle creak synthesis). Unlike [26], [27]'s (synthesized) period-doubled stimuli had little effect on inducing T21 responses in Cantonese. Similar results were found in [10] in Mandarin, for Tone 3. To probe these potentially conflicting results, we need to understand the property of period-doubled stimuli synthesized by KlattGrid.

Third, the synthesizer needs the extensibility to accommodate ongoing development and refinement of hypotheses about the acoustic parameters that underlie the role of phonation in tone perception. This is because development in our understanding of the phonetic spaces for voice quality is highly active. The development of [17]'s psychoacoustic model of voice quality has been grounded in perceptual experiments using the UCLA voice synthesizer to manipulate individual spectral components.

In the rest of the paper, I present analyses of the acoustic properties of stimuli generated by KlattGrid and the UCLA voice synthesizer—both synthesizers that allow the manipulation of f0 independently from other voice source parameters, using the methods described in §2. The first case study (§3.1) studies the acoustic properties of KlattGrid-generated period-doubled stimuli. The second (§3.2) compares acoustic properties of KlattGrid-generated stimuli to those of stimuli generated by the UCLA voice synthesizer in [5]'s Experiment 2 on the role of breathiness in Hmong tonal perception. Supplementary materials including code, additional detail on methods, and additional results can be found at the OSF repository at https://bit.ly/2HPXmvd.

2. METHODS

KlattGrid synthesis parameter settings [10, 27] manipulated particular KlattGrid parameters (indicated by bold-face font) to create different kinds of creak; parameter definitions from [14]. Period-doubled creak was synthesized by increasing **DI** (double pulsing: temporal offset and reduced amplitude of alternate periods), constricted/tense voice by decreasing **OQ** (open phase (or open quotient): ratio of open period to total period) and **TL** (spectral tilt: extra spectra tilt of the source (dB down at 3 kHz)), and irregular f0 by increasing **FL** (flutter: slowly varying statistical fluctuations to the fundamental period). [27] also synthesized breathy stimuli by increasing breathiness amplitude **AH** (amplitude of turbulent aspiration noise added to voicing), **TL**, and **OQ**. [3] synthesized breathy stimuli for Cham register perception by increasing **AH** and **TL**.

KlattGrid parameters used in [3, 10, 27] were individually manipulated to observe how they affected the acoustic parameters used to characterize the voice quality model in [17], using the settings shown in Table 1. These settings were chosen based on ranges and defaults in [3, 10, 27]. They were implemented together with two constant settings for formants (F1-F4) and f0 meant to model a man (M) and a woman (W), based on average values for [ɑ] from men and women in [9]. Bandwidths (B1-B4) were set to defaults from the original KLSYN88 code.

Table 1: Settings for generating continua for KlattGrid synthesis experiments. M=man, W = woman.

Parameter	Default	Range / Increment
TL (dB)	0	[0,50] / 1
OQ (%)	0.7	(0,1] / 0.05
FL (%)	0	[0,1] / 0.05
DI (%)	0	[0,1] / 0.05
F0 (Hz)	126 (M) 212 (W)	[40,500] / 20
AH (dB SPL)	0	[0,75] / 1

Acoustic voice quality measures Acoustic voice quality measures of the synthesized sounds were computed using VoiceSauce [22] under default settings. F0 was measured with Straight [11]; formants were estimated using Snack [23], and all parameters were computed except for epoch and excitation strength. Parameter values were extracted as means over 9 uniform time slices over the vowel, as well as mean values over the entire vowel. Harmonic amplitude measures reported are uncorrected for interactions with formants since vowel quality was constant and low across generated sound files. All statistical analyses were implemented in R [21].

Individual linear regression models were computed for each measured VoiceSauce parameter (estimated mean over the entire vowel) in [17]'s model, for each manipulated Klatt parameter. Results are summarized using linear regression coefficient plots [6], which visualize how much a unit of change in the manipulated synthesizer parameter affects acoustic voice quality parameters measured using VoiceSauce. Coefficient values are shown together with 95% CIs. The further to the left or right the coefficient is from the dashed line of 0, the bigger a change in the acoustic parameter effected by a change in the synthesis parameter.

Coefficient plots show results for selected acoustic measures: acoustic components in the psychoacoustic model of voice quality proposed in [17, Table 1]. These acoustic measures are grouped as follows: (i) harmonic source/spectral shape (H1-H2, H2-H4, H4-2kHz, 2kHz-5kHz), (ii) inharmonic source excitation (harmonic-to-noise ratios, e.g., HNR05/15/25/35 as measured using Voicesauce), (iii) time-varying source characteristics (f0 and amplitude parameters), and (iv) the vocal tract transfer function (formant frequencies, spectral zeroes, and bandwidths)). H1, H2, and H4 are the amplitudes of the first, second, and fourth harmonics, respectively; 2kHz and 5kHz indicate the harmonics closest to 2kHz and 5kHz, respectively. Uncorrected measures are indicated with a “u”, e.g., H1H2u for H1-H2, uncorrected.

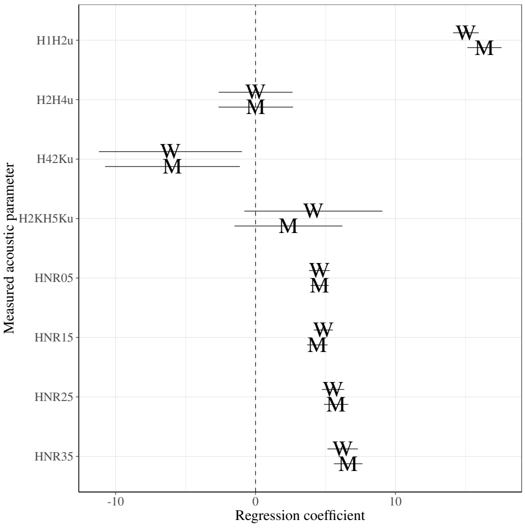
3. RESULTS AND DISCUSSION

3.1. Manipulation of KlattGrid DI parameter

The effect of varying the KlattGrid period-doubling **DI** parameter from 0.25 to 1 on harmonic and inharmonic source components is shown in Figure 1. Increasing **DI** resulted in increases in harmonic-to-noise ratios and large increases in (uncorrected) H1-H2. **DI** values below 0.25 were excluded because Straight-computed f0 values were halved with respect to input **F0** values for **DI**≥0.25; including **DI** values below 0.25 thus resulted in sharp discontinuities (i.e., nonlinearity) in acoustic measures as a function of **DI**.

In comparison, [13, Table 1] characterizes period-doubled (multiply pulsed) creak as having high noise (low HNRs) and low H1-H2 values, presumably relative to modal speech. If we take the **DI**=0 setting to be the baseline modal comparison (H1-H2 = 0, since **TL**=0, see Table 1), results show that setting **DI**>0 does result in lower H1-H2 values, since H1-H2u drops from 0 to -15dB with **DI**=0.25, but that H1-H2 steadily increases as **DI** does for **DI**≥0.25. **DI**>0 also results in higher HNR, so lower noise, rather than high noise. It is thus unclear that manipulating **DI** values produces acoustic consequences expected for period doubling, as described in [13]. What is clear is that since both HNR and H1-H2 are dependent on computed f0, the definition of period doubling in terms of these measures is critically dependent on how f0 is computed. How f0 should be computed is unclear. If f0 is computed using Praat or Snack instead of Straight, f0 is halved relative to the input **F0** value for non-zero **DI** values. If f0 is computed using [24]’s Subharmonic-to-Harmonic ratio, f0 for the man setting doubles as **DI** increases and is constant at 105 Hz for the woman setting for **DI**>0.25.

Figure 1: Linear regression coefficient plot for harmonic and inharmonic source component values as a function of **DI** in range [0.25,1] over [ɑ], M = man, W = woman.

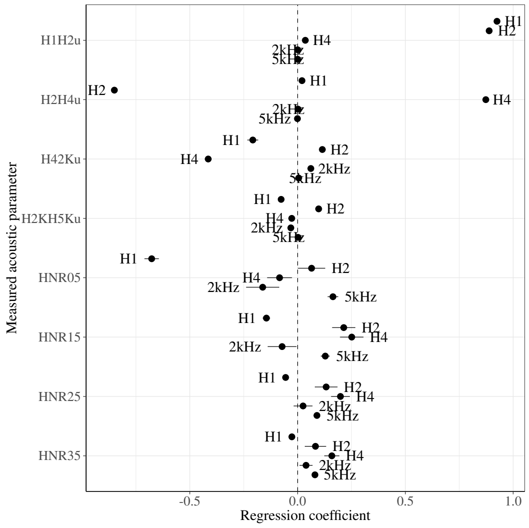


3.2. Comparison between KlattGrid and UCLA voice synthesizer in parameters for breathy voice quality synthesis

[5]’s Experiment 2 used the UCLA voice synthesizer to synthesize five different continua to investigate the perception of the breathy falling tone contrast in White Hmong. These continua individually varied the following harmonics: H1 to increase H1-H2; H2, to increase H1-H2 and decrease H2-H4; H4, to increase H2-H4 and decrease H4-2kHz; 2kHz, to increase H4-2kHz and decrease 2kHz-5kHz; and 5kHz, to increase 2kHz-5kHz. Figure 2 shows coefficient plots for these five different manipulations (labeled by the harmonic varied) for harmonic and inharmonic source components. Individual parameter manipulations primarily isolated the desired spectral components, e.g., the largest coefficient value for H1 manipulation is H1H2u; the largest coefficient values for the H2 manipulation are for H1H2u and H2H4u. However, Figures 2 also shows that a dB increase in H1 results in over a half decibel HNR05 decrease. Spectral manipulations can also have quite large effects on F1 and F2 (see repository).

Breathy stimuli in tone/register perception have been synthesized with KlattGrid by increasing **TL**, **OQ**, and **AH** [3, 27]. Thus, for comparison with the coefficient plots for the UCLA voice synthesizer, Figure 3 shows how changing **TL** affects acoustic measures. A change in **TL** targets H1H2u as well as H2H4u, and also H2KH5Ku and HNRs. Similarly, an increase in **OQ** results in large (and highly variable) changes in H1H2u, H2H4u, H42Ku and H2KH5Ku (see repository). The comparison verifies that the UCLA voice synthesizer better enables the targeting of individual spectral components identified in [17]'s psychoacoustic model of voice quality than the KlattGrid synthesizer.

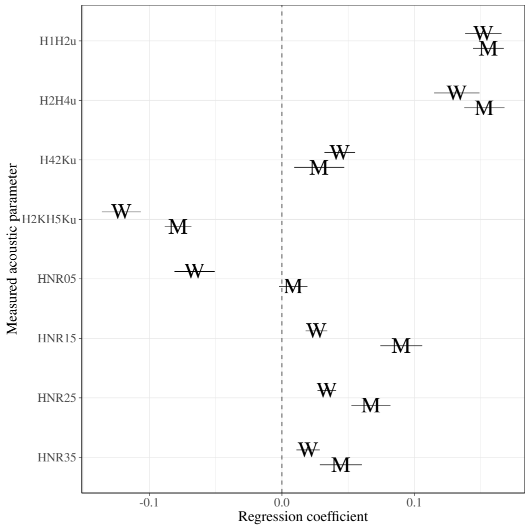
Figure 2: Linear regression coefficient plot for harmonic and inharmonic source components as a function of the five spectral parameters manipulated using the UCLA voice synthesizer in [5].



4. CONCLUSION

This paper has presented a small meta-analysis of recent work on the role of creaky and breathy voice quality in tonal perception to highlight what is still needed in developing voice quality synthesis tools. The synthesizer needs to be able to manipulate f0 and other voice source parameters independently. KlattGrid and the UCLA voice synthesizer offer this ability (see repository for validation). (The MATLAB-based synthesizer, Tandem-STRAIGHT [12], does as well, but isn't capable of direct manipulation of acoustic parameters.) The synthesizer also needs to be able to be able to generate distinct kinds of creak. §3.1 suggests that KlattGrid **DI** does not clearly effect acoustic qualities expected for period-doubled creak from [13]. Moreover, in an examination of naturally produced period-doubled samples, [16] found that "no consistent association exists between patterns of period and amplitude alternation". However, **DI** simultaneously sets both amplitude and period modulation: they can't be independently manipulated [14, 25]. **DI** does isolate a well-defined acoustic property, and perceptual experiments manipulating **DI** certainly inform us about the perception of period-doubled creak. But the relation between **DI** and naturally produced period-doubled creak isn't a transparent one.

Figure 3: Linear regression coefficient plot for harmonic and inharmonic source components as a function of KlattGrid **TL** over [ɑ].



What will it take for there to be a synthesis tool that can model acoustic properties of naturally produced period-doubled and other kinds of creaky (and breathy) voice quality? Further acoustic examination of libraries of naturally-produced creaky (and breathy) voice quality, especially since it may be that in defining kinds of creak where f0 is ill-defined, acoustic parameters beyond those dependent on f0—perhaps even parameters not yet identified in current models of voice quality like [17]—may be needed. Concomitantly, the final aforementioned desideratum: synthesis software with the extensibility to accommodate ongoing development and refinement of hypotheses about acoustic and perceptual properties of distinct kinds of creak. Ideally, this means voice synthesis software that is open source, cross-platform, and welcoming to user contributions.

5. ACKNOWLEDGEMENTS

Thank you to Yaqian Huang and Yubin Zhang for sharing their stimulating work and KlattGrid settings and to Marc Garellek for stimuli from [5].

6. REFERENCES

[1] Anderson, R. C., Klofstad, C. A., Mayvew, W. J., Venkatachalam, M. 2014. Vocal fry may undermine the success of young women in the labor market. *PLoS ONE* 9(5), 1–8.

[2] Boersma, P., Weenink, D. 2010. Praat: doing phonetics by computer (version 5.1.32) [computer program]. `http://www.praat.org`.

[3] Brunelle, M. 2012. Dialect experience and perceptual integrality in phonological registers: Fundamental frequency, voice quality and the first formant in Cham. *Journal of the Acoustical Society of America* 131, 3088—3102.

[4] Garellek, M. To appear. The phonetics of voice. In: Katz, W., Assmann, P., (eds), *The Routledge Handbook of Phonetics*. Routledge.

[5] Garellek, M., Keating, P., Esposito, C. M., Kreiman, J. 2013. Voice quality and tone identification in White Hmong. *The Journal of the Acoustical Society of America* 133(2), 1078–1089.

[6] Gelman, A., Hill, J. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

[7] Gerratt, B. R., Kreiman, J. Oct. 2001. Toward a taxonomy of nonmodal phonation. *Journal of Phonetics* 29(4), 365–381.

[8] Gårding, E., Kratochvil, P., Svantesson, J.-O. 1986. Tone 4 and Tone 3 discrimination in modern Standard Chinese. *Language and Speech* 29(3), 281–293.

[9] Hillenbrand, J., Getty, L. A., Clark, M. J., Wheeler, K. May 1995. Acoustic characteristics of american english vowels. *The Journal of the Acoustical Society of America* 97(5), 3099–3111.

[10] Huang, Y. 2019. Low f0 as a creak attribute in Mandarin tone perception. Conference presentation. The 93rd Annual Meeting of the Linguistic Society of America.

[11] Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A. 1999. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication* 27(3–4), 187–207.

[12] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. 2008. Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. *ICASSP 2008.* 3933–3936.

[13] Keating, P., Garellek, M., Kreiman, J. 2015. Acoustic properties of different kinds of creaky voice. *Proceedings of ICPhS 2015.*

[14] Klatt, D. H., Klatt, L. C. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America* 87(2), 820–857.

[15] Kreiman, J., Antonanzas-Barroso, N., Gerratt, B. R. 2016. The UCLA voice synthesizer, version 2.0. *The Journal of the Acoustical Society of America 140* 140, 2961.

[16] Kreiman, J., Gerratt, B. R. 2006. Spectral characteristics of period doubled phonation. *4th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan.*

[17] Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., Zhang, Z. 2014. Toward a unified theory of voice production and perception. *Loquens* 1(1), e009.

[18] Kreiman, J., Sidtis, D. 2011. *Foundations of Voice Studies: An Interdisciplinary approach to voice production and perception.* Malden, Massachusetts: Wiley-Blackwell.

[19] Kuang, J. 2017. Covariation between voice quality and pitch: Revisiting the case of mandarin creaky voice. *The Journal of the Acoustical Society of America* 142(3), 1693–1706.

[20] Kuang, J., Liberman, M. 2018. Integrating voice quality cues in the pitch perception of speech and non-speech utterances. *Frontiers in Psychology* 9, 2147.

[21] R Core Team, 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

[22] Shue, Y.-L., Keating, P., Vicenik, C., Yu, K. 2011. Voicesauce: a program for voice analysis. *Proceedings of ICPhS XVI.*

[23] Sjölander, K. 2004. Snack sound toolkit. http://www.speech.kth.se/snack.

[24] Sun, X. 2000. A pitch determination algorithm based on subharmonic-to-harmonic ratio. *Proceedings of the 6th International Conference on Spoken Language Processing* 676–679.

[25] Weenink, D. 2009. The KlattGrid speech synthesizer. *Proceedings of INTERSPEECH 2009* volume 10 2059–2062.

[26] Yu, K. M., Lam, H. W. 2014. The role of creaky voice in Cantonese tonal perception. *Journal of the Acoustical Society of America* 136(3), 1320–1333.

[27] Zhang, Y., Kirby, J. 2018. Weighting of f0 mean, f0 change and phonation cues in tone perception: The case of Cantonese tone 4/ tone 6. Manuscript.