# WHEN IS ENOUGH, ENOUGH? VOT JUDGMENTS VARY BY VOICING INTENSITY, ASPIRATION INTENSITY, VOICE QUALITY, AND RATE OF CHANGE

John Kingston and Amanda Rysling

University of Massachusetts, Amherst and University of California, Santa Cruz
jkingston@linguist.umass.edu, rysling@ucsc.edu

## ABSTRACT

Listeners categorized a short-long lag VOT /ba-pa/ continuum in eight conditions, which fully crossed (a) manipulating voicing intensity vs. voice quality, (b) whether time to ramp up from either 0 to peak voicing intensity or from lax to modal voice quality was short (10 ms) vs long (70 ms), (c) aspiration intensity (21 vs 33 dB). Listeners responded "pa" much more often when voicing reached peak intensity slowly, and measurably more often when voice quality became modal slowly. Apparently, voicing is not perceived until its intensity or quality crosses a threshold. Listeners' reliance on aspiration intensity was conditional: they only responded "pa" more with more intense aspiration if peak voicing intensity or modal voice were reached quickly.

**Keywords:** perception, VOT, aspiration and voicing intensity, voice quality

## 1. INTRODUCTION

Voice-onset time (VOT) is the most studied acoustic correlate of and perceptual cue to voicing contrasts between stops [1, 2]. In a typical study of its perception, stimuli are made by incrementally varying when a harmonically rich, periodic sound source, AKA "voicing," is turned on relative to when a high-pass filtered,[1] aperiodic sound source, AKA "noise," is abruptly turned on. The abrupt onset of the aperiodic source simulates the stop burst, and its continuation simulates any ensuing aspiration. What is typically reported is when, relative to the onset of that aperiodic source, listeners cross over from responding "voiced" to "voiceless," expressed as a VOT value. This category boundary has been found to vary as a function of other acoustic correlates, e.g., F1 onset frequency [3] and F0 [4, 5]. Even so, what the correlates of voicing are and how and why they serve as perceptual cues remain open questions.

Here, we take up these questions by manipulating voicing's intensity and quality independently of the timing of its onset. Repp [6] showed that listeners required a shorter lag in voice onset to respond "ta" to a stimulus when the intensity of the vowel was weaker relative to the intensity of the preceding aspiration. Using an aerodynamic model of stop consonant production, Ohala [7] showed that subglottal air pressure (Ps) is predicted to be lower during vowels following voiceless aspirated stops than those following voiced stops because glottal resistance is lower. Lower Ps would produce the less intense voice source observed following voiceless aspirated than voiced stops in English [8, 9, 10]. Lower Ps and weaker vowel intensity are expected after long- than short-lag stops, too, for the same reason: the glottal aperture is larger and resistance lower when the delay in voice onset is longer.

This expectation was confirmed: voice quality was laxer and breathier and intensity was lower for at least 25 ms following voiceless aspirated than voiceless unaspirated stops, presumably because the glottis did not close completely during early glottal cycles after voicing began, cf. [1, 11]. These observations led us to ask: do listeners use the laxer onset of voicing from coarticulation with an aspirated stop itself as a cue, or do they instead use the concomitant lesser intensity of voicing? Do a shorter VOT and weaker aspiration intensity suffice to elicit a "pa" response to a short-long lag VOT /ba-pa/ continuum when the vowel's initial voice quality was laxer, or do they only suffice if the vowel's initial voicing intensity were itself weaker?

## 2. METHOD

### 2.1. Stimuli

The Klatt synthesizer [12] was used to create eight short-to-long lag /ba-pa/ continua, which fully crossed three two-level factors (2x2x2): (a) dimension (voicing intensity, 0 to 72 dB, vs. voice quality, 0.9 to 0.5 open quotient and 18 to 0 dB spectral tilt), (b) ramp duration (from initial to final value in 10 vs. 70 ms), (c) aspiration intensity (21 vs. 33 db).[2] These manipulations produced steep or

gradual ramps in voicing intensity or voice quality, which occurred with either 21 dB or 33 dB of aspiration intensity. In all continua, VOT was 7, 18, 29, or 40 ms. The voiced portion of the vowel was constant in duration, cf. [13, 14, 15, 16]. Stimuli were presented synchronously to both headphone earpieces or 0.45 ms earlier to one earpiece; this tested the ultimately unsubstantiated hypothesis that intensity differences woould be more detectable at the midline [17], and so we do not return to it here.

The waveforms in Figure 1 show differences between the steep and gradual intensity ramps. Figure 2 shows that voicing reached peak intensity nearly as quickly for the gradual as the steep voice quality ramp. Figure 3 shows by standard measures that voice quality was laxer when the ramp was longer: the first harmonic, H1, was more intense relative to the second, H2, and the harmonic nearest F1, A1. The top of Table 1 shows that H1-H2 and H1-A1 were both greater when the voice quality ramp was gradual. The bottom shows H1-H2 didn't differ between 10- and 70-ms intensity ramps, and that H1-A1 did not differ nearly as much as it did between long and short voice quality ramps. All harmonics were at least 15 dB less intense when the intensity ramp was longer. Figure 4 shows aspiration intensities for the shortest and longest VOTs.

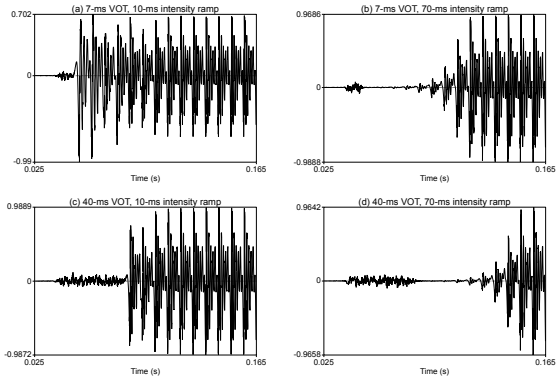


**Figure 1:** (a) 7-ms VOT, 10-ms intensity ramp, (b) 7-ms VOT, 70-ms intensity ramp, (c) 40-ms VOT, 10-ms intensity ramp, (b) 40-ms VOT, 70-ms intensity ramp. (33-dB aspiration intensity.)

## 2.2. Participants

34 listeners reported exposure only to English before age 6; 36 listeners reported early exposure to other languages (18 in all). 17 native and 16 non-native listeners responded to voicing intensity; 17 native and 20 non-native listeners to voice quality ramps.

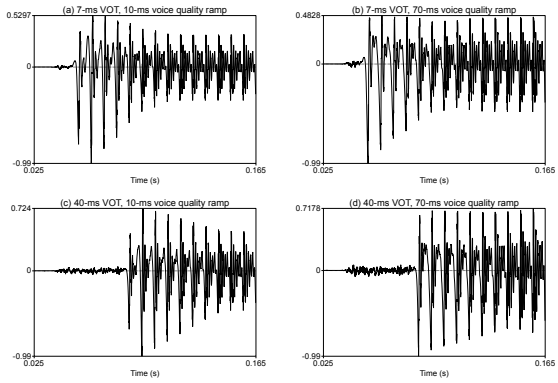


**Figure 2:** (a) 7-ms VOT, 10-ms voice quality ramp, (b) 7-ms VOT, 70-ms voice quality ramp, (c) 40-ms VOT, 10-ms voice quality ramp, (b) 40-ms VOT, 70-ms voice quality ramp. (33-dB aspiration intensity.)

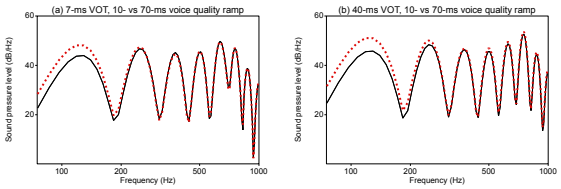


**Figure 3:** Spectra of 70 ms after voice onset; log 75-1000 Hz: (a) 7-ms VOT, (b) 40-ms VOT; (black solid) 10-ms voice quality ramp, (red dashed) 70-ms ramp.

| VOT (ms) | 7 | | 40 | |
|---|---|---|---|---|
| Ramp (ms) | 10 | 70 | 10 | 70 |
| H1-H2 | -2.7 | 0.8 | -2.5 | 1.2 |
| H1-A1 | -5.7 | -1.2 | -6.7 | -2.3 |
| H1-H2 | -2.7 | -2.6 | -2.5 | -2.5 |
| H1-A1 | -6.3 | -4.2 | -13.4 | -5.2 |

**Table 1:** H1-H2 and H1-A1 in dB, after 7- and 40-ms VOTs, for 10- and 70-ms ramps: (top) voice quality, (bottom) intensity.

## 2.3. Procedure

A training block and then 9 test blocks were presented twice. Participants were trained with correct-answer feedback to respond "ba" to the stimuli with the shortest VOT and "pa" to those with the longest. Each of the 2 VOTs was presented once during each training block with the 2 aspiration intensities, the 2 ramp durations, and the 3 apparent locations, for 48 training trials. The 2 intermediate VOTs in all their combinations with the other

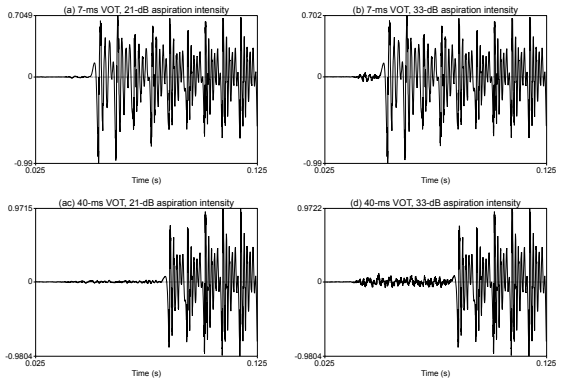


**Figure 4:** (a) 7-ms VOT, 21-dB aspiration intensity, (b) 7-ms VOT, 33-dB aspiration intensity, (c) 40-ms VOT, 21-dB aspiration intensity, (b) 40-ms VOT, 33-dB aspiration intensity. (10-ms intensity ramp.)

stimulus manipulations were added during each ensuing test block. Each stimulus was presented 18 times, for 864 total test trials. The order of stimulus presentation was randomized differently within each training and testing block for each participant.

Each trial begin with a display of a cross for 500 ms. The response options, "BA" and "PA," then appeared on opposite sides of the screen; which one appeared on the left vs right was counter-balanced between participants. The stimulus began when the response options appeared. Participants could respond while the stimulus was playing; afterward, they had up to 1500 ms more to respond. If they did not respond before that additional time had elapsed, they were encouraged to respond sooner. After the participant responded or the trial timed out, correct answer feedback was displayed for 500 ms on training trials. The interval before the next trial began was 750 ms in both training and test trials.

## 3. RESULTS

Figure 5 shows that, as expected, listeners responded "pa" more often as VOT lengthened. They also responded "pa" more when ramps were longer, but more so for voicing intensity than voice quality. Listeners only responded "pa" more often when aspiration was more intense if ramp duration was also short. When ramp duration was long, "pa" responses increased for less intense aspiration as well, enough to shrink aspiration intensity's effect when voicing intensity was manipulated. This shrinkage of aspiration intensity's effect across the two voicing intensity ramps occurred even though "pa" responses also increased *in general* for more intense aspiration. The difference in "pa" proportions between the stronger and weaker aspirations shrank for the long voice quality ramp.

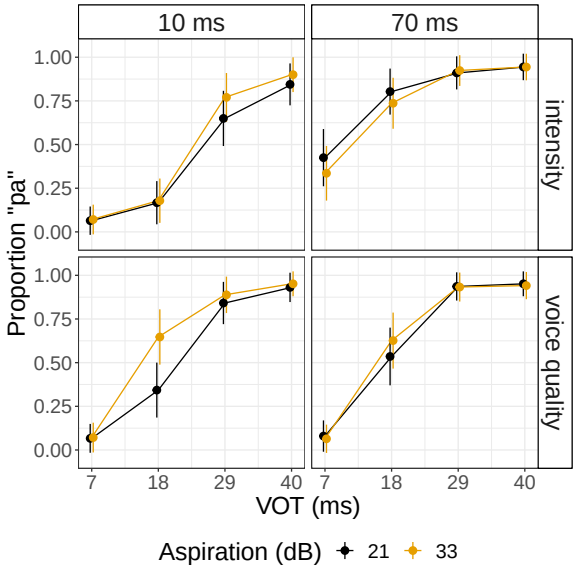


**Figure 5:** Mean proportions of "pa" responses (95% confidence intervals) by ramp duration, ramp dimension, VOT, and aspiration intensity

A Bayesian mixed effects logistic regression model was fitted to "pa" responses using the *brms* package [18]. Population-level effects included ramp dimension (voicing intensity = 0.5, voice quality = -0.5), ramp duration (10 ms = -0.5, 70 ms = 0.5), VOT (centered and scaled such that its standard deviation = 0.5) [19], and aspiration intensity (22 dB = -0.5, 33 dB = 0.5), and all interactions between them. A group-level effect of subject was included on the intercept and slopes of the within-subjects population-level effects (ramp duration, VOT, voicing intensity) and their interactions. Listener language exposure was not included; other analyses showed that the estimate's 0.95 credible interval (CI) included 0. Table 2 shows estimates of the included population-level effects.

The log Bayes factor (332.55) showed this model fit the data better than one without interactions between population-level effects. Chains did not diverge; autocorrelation within chains was modest; the model converged with `R-hat` values of 1.00 and ample effective sample sizes for all parameters; posterior predictive checks showed that replications closely matched the observed data.

The 95% CI for the intercept's positive estimate does not include 0, which shows a bias to respond "pa." No 95% CIs for the estimates of the individual population-level effects included 0; their signs show that listeners responded "pa" less often when the

| | Est | Err | Lwr | Upr |
|---|---|---|---|---|
| Intercept | 1.13 | 0.11 | 0.91 | 1.35 |
| Dim(ension) | -0.71 | 0.22 | -1.14 | -0.27 |
| Dur(ation) | 1.47 | 0.12 | 1.24 | 1.71 |
| VOT | 5.39 | 0.27 | 4.86 | 5.93 |
| Asp(iration) | 0.47 | 0.09 | 0.29 | 0.66 |
| Dim:Dur | 1.84 | 0.23 | 1.38 | 2.30 |
| Dim:VOT | -1.97 | 0.54 | -3.03 | -0.92 |
| Dim:Asp | -0.60 | 0.18 | -0.96 | -0.24 |
| VOT:Asp | 0.47 | 0.16 | 0.15 | 0.79 |
| Dur:Asp | -0.73 | 0.11 | -0.96 | -0.51 |
| Dim:Dur:VOT | -2.02 | 0.35 | -2.71 | -1.34 |
| Dim:Dur:Asp | 0.35 | 0.22 | -0.08 | 0.79 |
| Dim:VOT:Asp | 0.32 | 0.31 | -0.29 | 0.93 |
| Dur:VOT:Asp | -0.19 | 0.25 | -0.69 | 0.30 |
| Dim:Dur:VOT:Asp | 0.24 | 0.46 | -0.65 | 1.16 |

**Table 2:** Estimates of population-level effects, errors, and lower and upper bounds of 95% CIs.

manipulated dimension was intensity rather than voice quality (top vs. bottom panels in Figure 5), but more often when ramp duration was longer (right vs. left), VOT was longer (*x*-axis), and aspiration was more intense (colors). These effects are moderated by interactions with 95% CIs that do not include 0. When the manipulated dimension was voicing intensity rather than voice quality, the proportions of "pa" responses increased more for a longer ramp duration but less for a longer VOT or more intense aspiration (right vs. left in the top vs. bottom). When aspiration was more intense, the proportions of "pa" responses also increased more for longer VOTs but less when ramp duration was longer (left vs. right). The negative estimate for the three-way interaction of manipulated dimension, ramp duration, and VOT shows that "pa" proportions increased less for longer VOTs and the longer ramp duration when voicing intensity was manipulated rather than voice quality (top vs. bottom right).

## 4. DISCUSSION

The long, gradual ramp in voicing intensity dramatically increased the proportion of "pa" responses compared to the short, steep one; a finding that qualitatively and quantitatively resembles the effect of a slow, 60-ms long rise vs abrupt rise to peak voicing intensity in Darwin and Pearson's Experiment 1 [20]. Their /ba:pa/ category boundary was much earlier, 12.7 ms, when the rise was slow, and later, 30.6 ms, when it was abrupt, perhaps because voicing intensity was too weak relative to aspiration's intensity for listeners to use it when it rose slowly. Our category boundaries differed similarly between the long and short intensity ramp, 2.0 vs 25.5 ms. Both findings suggest that voice onset is not detected until its intensity reaches a criterial level. The proportion of "pa" responses also increased when the voice quality ramp was long and gradual rather than short and steep, but that increase is nearly entirely attributable to an aspiration intensity of 21 dB eliciting more "pa" responses when VOT was 18 ms; otherwise, "pa" responses increased equally and modestly for both aspiration intensities at VOTs of 29 and 40 ms when the voice quality ramp was longer. The effects of the duration of the voicing intensity ramp and aspiration intensity were more uniform across VOTs: "pa" responses increased for both aspiration intensities when the voicing intensity ramp was gradual.

But the potential three-way interaction was not supported: "pa" responses were no longer more numerous with more intense aspiration when the ramp duration was longer for either voicing intensity or voice quality. This collapse of the difference in voiceless responses between weaker and stronger aspiration levels with both longer ramps is surprising given Repp's original finding that listeners responded "ta" more often for a given voicing lag when aspiration was more intense [6]. The present results are also surprising given Darwin and Seton's finding that listeners responded "pa" more often when aspiration level was increased by 10 dB, and more intense aspiration increased "pa" responses more when the level of voicing was lower [21]. The surprise is most acute for the long intensity ramp because the voicing level remained low for so long, and voiceless endpoint responses might have been expected in the presence of only subtle voicing. But Darwin and Pearson did not find that category boundaries shifted to earlier VOT values when aspiration intensity was increased by 15 dB in stimuli in which voicing reached its peak intensity slowly [20]. There is thus some precedent for the finding that aspiration intensity is not determinative of voicelessness judgments when voicing intensity begins gradually. We can still conclude that those judgments depend on how late the periodic source was perceived to be too weak or too lax.

[1] To simulate acoustic coupling to the trachea while the glottis is open, the filter's lower cutoff is high enough that the aperiodic source does not excite the first formant.
[2] A ∼10 ms-long burst began with aspiration onset. Its intensity at successive 5-ms intervals was 0-45-35-0 dB.

## 5. REFERENCES


[1] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, pp. 384–422, 1964.

[2] A. S. Abramson and D. H. Whalen, "Voice onset time (VOT) at 50: Theoretical and practical issue in measuring voicing distinctions," *Journal of Phonetics*, vol. 63, pp. 75–86, 2017.

[3] K. R. Kluender, "Effects of first formant onset properties on vot judgments can be explained by auditory processes not specific to humans," *Journal of the Acoustical Society of America*, vol. 90, pp. 83–96, 1991.

[4] D. H. Whalen, A. Abramson, L. Lisker, and M. Mody, "Gradient effects of fundamental frequency on stop consonant voicing judgments," *Phonetica*, vol. 47, pp. 36–49, 1990.

[5] D. H. Whalen, A. S. Abramson, L. Lisker, and M. Mody, "F0 gives voicing information even with unambiguous voice onset times," *Journal of Acoustical Society of America*, vol. 47, pp. 36–49, 1993.

[6] B. H. Repp, "Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants," *Language and Speech*, vol. 22, pp. 173–189, 1979.

[7] J. J. Ohala, "A mathematical model of speech aerodynamics," *Annual Report of the Institute of Phonetics, University of Copenhagen*, vol. 8, pp. 11–28, 1974.

[8] A. S. House and G. Fairbanks, "The influence of consonant environment upon the secondary acoustical characteristics of vowels," *Journal of Acoustical Society of America*, vol. 25, pp. 105–113, 1953.

[9] P. Ladefoged and N. P. McKinney, "Loudness, sound pressure, and subglottal pressure in speech," *Journal of the Acoustical Society of America*, vol. 35, no. 4, pp. 454–460, 1963.

[10] I. Lehiste and G. E. Peterson, "Vowel amplitude and phonemic stress in American English," *Journal of the Acoustical Society of America*, vol. 31, no. 4, pp. 428–435, 1959.

[11] E. Fischer-Jørgensen and B. Hutters, "Aspirated stop consonants before low vowels, a problem of delimitation, – its causes and consequences," *Annual Report of the Institute of Phonetics, University of Copenhagen*, vol. 15, pp. 77–102, 1981.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," Retrieved: 21 November 2022, 2022.

[13] G. E. Peterson and I. Lehiste, "Duration of syllable nuclei in English," *Journal of the Acoustical Society of America*, vol. 32, pp. 693–703, 1960.

[14] A. J. Lotto, K. R. Kluender, and K. P. Green, "Spectral discontinuities and the vowel length effect," *Perception & Psychophysics*, vol. 58, no. 7, pp. 1005–1014, 1996.

[15] J. C. Toscano and B. McMurray, "Cue-integration and context effects in speech: Evidence against speaking-rate normalization," *Attention, Perception, & Psychophysics*, vol. 74, pp. 1284–1301, 2012.

[16] ——, "The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments," *Language, Cognition and Neuroscience*, vol. 30, no. 5, pp. 529–543, 2015.

[17] J. K. Maier, P. Hehrmann, N. S. Harper, G. M. Klump, D. Pressnitzer, and D. McAlpine, "Adaptive coding is constrained to midline locations in a spatial listening task," *Journal of Neurophysiology*, vol. 108, no. 7, pp. 1856–1868, 2012.

[18] P.-C. Bürkner, "brms: An R package for Bayesian multilevel models," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017.

[19] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su, "A weakly informative default prior distribution for logistic and other regression models," *The Annals of Applied Statistics*, vol. 2, no. 4, pp. 1360–1383, 2008.

[20] C. J. Darwin and M. Pearson, "What tells when voicing has started?" *Speech Communication*, vol. 1, no. 1, pp. 29–44, 1982.

[21] C. J. Darwin and J. Seton, "Perceptual cues to the onset of voiced aspiraion in aspirated initial stops," *Journal of the Acoustical Society of America*, vol. 74, no. 4, pp. 1126–1135, 1983.