SPEECH RATE EFFECTS IN FRENCH STOP VOICING PRODUCTION AND PERCEPTION*

Yoonjung Kang, Nicholas Haggarty

University of Toronto Scarborough, University of Toronto
yoonjung.kang@utoronnto.ca, n.haggarty@mail.utoronto.ca

ABSTRACT

Studies on speech rate variation have reported shortening of prevoicing and aspiration of stops in fast speech. Perception studies have also found that the boundary between short and long-lag VOT stops shifts to shorter VOTs when the speech rate increases in the carrier phrase. However, no previous study has explored how speech rate variation affects the perception of prevoiced stops. Our study aims to fill this gap by investigating the effects of speech rate variation on French stops, which contrast prevoiced and short-lag stops. We collected production and perception data from the same participants using the same speech materials. Thirty-one self-identified native speakers of French participated in an online experiment. The production results confirmed previous observations of shortening of prevoicing and aspiration in fast speech. The perception results showed an expected shift in perception for positive VOTs based on speech rate but not for negative VOTs.

Keywords: French, voicing, production, perception, speech rate

1. INTRODUCTION

Previous production studies have found that in fast speech compared to slow speech, the duration of prevoicing (negative VOTs) and aspiration (positive VOTs) shortens [1, 2]. Parallel to the variation in production, perception studies have found that the boundary between short and long-lag VOT stops shifts to shorter VOTs when the speech rate increases in preceding phrases and identical stimuli are more likely to be heard as the long-lag category (i.e., voiceless or aspirated) when embedded in fast speech [3-6]. To our knowledge, no previous study has examined if a similar perceptual adjustment is found for the perception of contrast between prevoiced and short-lag stops and, in particular if a rate-modulated boundary shift occurs in the negative VOT continuum. Given that languages can contrast multiple stop categories along the negative VOT dimension [7, 8], we should expect that listeners can develop sensitivity to not only the presence or absence of prevoicing but to the degree of prevoicing [9]. Thus, they should also be able to perform the type of rate normalization found in the positive VOT range. Our study fills this gap in the literature by investigating the effects of speech rate variation in French stops, which contrast prevoiced and short-lag stops [10, 11].

2. METHODS

2.1. Participants and experiment overview

Participants were recruited through prolific.co. The eligibility requirements were to reside in Canada or France and to consider French as their first and primary language. A total of 43 participants completed the study. Out of those, 11 were excluded for not meeting the eligibility requirement (for listing English as either their first language or their dominant language in the detailed background questionnaire). Out of the 32 participants who were retained for analysis, nine were residing in Canada, and the other 23 were residing in France. The experiment itself was built and hosted on gorilla.sc [12]. The full experiment included informed consent, a background questionnaire, a perception task, and a production task and took 28.2 minutes on average. The perception and the production tasks took 12.1 and 9.6 minutes on average, respectively.

2.2. Perception experiment

The perception task was an identification task, in which participants heard target words (*coût* vs. *gout*) embedded in a carrier phrase, “Mon mot favori est___.” and selected the word they heard. A male Canadian French speaker produced the speech materials. Two baseline stimuli were created by splicing together a token of carrier phrase (carrier), a voiceless stop closure from a voiceless stop production (clos), prevoicing of a voiced stop production (prev), an aspiration of a voiceless stop production (asp), and a post-stop vowel (v) of either a voiced or voiceless stop production. All acoustic analysis and manipulation were done in Praat [13]. (See Figure 1). The spliced baselines were subsequently manipulated to create two 18-step VOT continua of 10 ms intervals that range from -100 ms

to 70 ms. To create positive VOT tokens, the duration of prevoicing (prev) was reduced to 0 ms while the duration of aspiration (asp) was adjusted to match the intended VOT value. The voiceless closure that precedes the stop release was also manipulated to be equal to the mean of stop closure of all stop productions (voiced or voiceless) for the speaker so that the closure duration itself does not provide an independent cue for voicing. For the negative VOT continua, the aspiration (asp) portion was reduced to 0.3 ms to give a perception of stop release but not of aspiration, and the duration of prevoicing (prev) was adjusted to match the intended negative VOT value. The voiceless closure (clos) duration was also adjusted so that the total stop closure (clos + prev) was equal to the overall average closure duration.

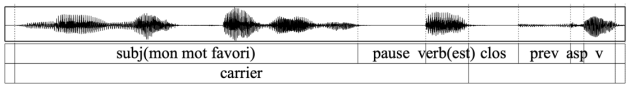

Figure 1: Segmentation of spliced baseline stimulus

The subparts of the carrier phrase (subj, pause, and verb) were also manipulated to the average value of all productions for the speaker to remove any potential durational cue to target voicing. This modified carrier was further manipulated to create fast and slow versions. The slow version was 1.5 times longer (108% of the overall average duration of the speaker's production) than the fast version (72%). These ratios were chosen to ensure that the two versions sound noticeably fast and slow without sounding unnatural. The speech rate of the carrier phrase, the VOT of the target word, and the base vowel following the target stop were orthogonally varied to create a total of 72 stimuli (18 VOT steps * 2 vowel bases (voiced vs. voiceless) * 2 speech rates). The stimuli were randomized and repeated twice for a total of 144 trials.

2.3. Production experiment

For the production task, participants heard one of two prompts from the model talker, which were the same as the fast and slow versions of the carrier phrase used in the perception task (without the target word). After hearing the prompt, participants were instructed to produce the full sentence including the target word displayed on the screen while trying to imitate the speech rate of the model talker's prompt as closely as possible. The recording was self-paced, and for each trial, participants began and ended the recording by clicking a button. In order to minimize data loss caused by participants stopping the recording too early and cutting off the end of the utterance, which is common in online production experiments [14], participants were asked to produce the sentence twice per prompt, and only the first production of each trial was analyzed.

Each target word was presented 20 times, ten times with a slow carrier prompt and ten times with a fast prompt. Each word was presented in its own block, within which the fast and slow trials were randomized. The order of the two blocks was randomized for each participant, with half of the participants starting with the voiced block and the other half starting with the voiceless block. Each participant was expected to produce a total of 40 tokens (2 target words * 2 speech rates * 10 repetitions). After excluding mispronunciations and other recording anomalies and omissions, 573 tokens for *coût* and 581 tokens for *gout* were retained for analysis. The production data were converted from .weba to mono .wav format for acoustic analysis. Each production was segmented for subj, pause, verb, clos, prev, asp, and v(owel) (see Figure 1) and measured for duration.

2.4. Statistical analysis

For the perception analysis, we built two separate mixed-effects logistic regression models using the *lme4* package [15] in R [16], one for the positive VOT values (>0 ms) and one for the negative VOT values (<0 ms). The response variable was the participants' RESPONSE (voiced = 0, voiceless = 1), and the fixed-effect predictors included VOT (ms) and BASE.VOWEL (voiced, voiceless), SPEECH.RATE (fast, slow), and DIALECT (Canadian, European), and full interaction terms. All fixed-effect predictors were centred by z-score transformation to promote model convergence. The random effects included a by-SUBJECT random intercept and by-SUBJECT slope adjustments to the two phonetic predictors, VOT and BASE.VOWEL, to account for individual variation in phonetic cue use. The full model was trimmed down using stepwise regression, as implemented in the *buildmer* [17] function in R [16].

We predict that VOT will exhibit a positive, statistically significant coefficient in the positive VOT model. However, the effect of prevoicing duration, as opposed to its mere presence or absence, on voicing perception is less clear. We also predict that BASE.VOWEL will have a significant effect, with voiced and voiceless vowels promoting voiced and voiceless percepts, respectively, due to the numerous secondary voicing cues present in post-stop vowels. We also included DIALECT and its interaction with other predictors to assess the differences in cue use between dialects. Finally, the primary factor of interest in our study is SPEECH.RATE. If both prevoicing and aspiration shorten in fast speech and

participants calibrate their perception of VOT duration cues depending on the ambient speech rate, we predict that listeners will be more willing to accept an ambiguous positive VOT token as voiceless (i.e., longer aspiration) and an ambiguous negative VOT token as voiced (i.e., long prevoicing) in fast speech.

For production analysis, we built mixed-effects linear regression models to compare the duration of each segment in slow vs. fast speech. The model included DURATION (ms) as the response variable and SPEECH.RATE (fast = -0.5, slow = 0.5), VOICING of the target stop (voiced = -0.5, voiceless = 0.5), and their interaction as fixed-effect predictors. A by-SUBJECT random intercept and by-SUBJECT slope adjustments to VOICING and SPEECH.RATE were also included. We predict that other things being equal, segments will be longer in slow than fast speech.

In the next section, we will first discuss the production results and establish the rate-conditioned change of prevoicing and aspiration in our pool of participants. We will then discuss the perception results and interpret them relative to the production variation.

3. RESULTS

3.1. Production results

The boxplots in Figure 2 summarize the distribution of the duration of each subpart by speech rate and stop voicing. For every segment, the duration is longer in the slow condition than in the fast condition, and the difference is statistically significant (subj: $b = 487.0$, $p < 0.001$; paus: $b = 95.5$, $p < 0.001$; verb: $b = 37.9$, $p < 0.001$; clos: $b = 63.0$, $p < 0.001$; prev: $b = 12.6$, $p = 0.021$; asp: $b = 16.2$, $p < 0.001$; vowel: $b = 10.0$, $p < 0.001$).

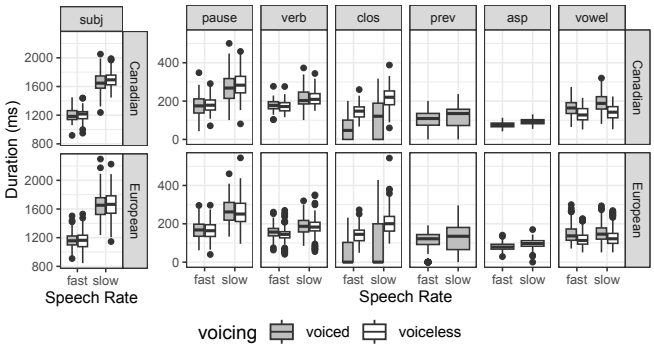

Figure 2: Boxplots of duration by segment, voicing, speech rate, and dialect

The mean duration of each segment by speech rate and the ratio of slow to fast duration are summarized in Table 1. For many of the segments, the slow-to-fast ratio is close to the model talker's ratio of 1.5, indicating that the task successfully elicited the intended speech rate in the participants' production. Although the ratios become smaller later in the sentence, and the ratios for prevoicing (prev), aspiration (asp) and the following vowel (vowel) are much smaller than 1.5., the rate difference was maintained until the end of the sentence.

	fast	slow	ratio
subj	1170.7 (106.8)	1663.1 (168.9)	1.42
pause	170.6 (45.6)	266.8 (71.8)	1.56
verb	157.3 (34.3)	195.4 (46.3)	1.24
clos	95.1 (73.4)	157.8 (104.3)	1.66
prev	104.9 (53.0)	118.5 (75.1)	1.13
asp	79.6 (19.3)	95.7 (21.2)	1.20
vowel	139.2 (45.9)	148.9 (50.1)	1.07

Table 1: Mean durations (standard deviation) in ms by speech rate and slow-to-fast duration ratios

For subj, pause, and verb, the voicing of the target stop does not affect their duration significantly (subj: $b = 6.9$, $p = 0.337$; paus: $b = -2.9$, $p = 0.528$; verb: $b = -6.6$, $p = 0.117$). For the post-stop vowel, the duration is shorter for the voiceless than the voiced stop condition (vowel: $b = -26.1$, $p < 0.001$). The (voiceless) closure duration excluding prevoicing (clos: $b = 99.2$, $p < 0.001$) is longer for the voiceless than the voiced stops, as expected. For the preceding vowel (verb: $b = 8.9$, $p = 0.007$), there's a significant interaction of stop voicing and rate, such that the effect of rate is stronger for the voiceless stop than for the voiced stop.

3.2. Perception results

Figure 3 (a) displays the proportion of voiceless responses by VOT, BASE.VOWEL, DIALECT, and SPEECH.RATE. The results align with the expectation as higher VOT values and a voiceless base vowel induced more voiceless responses. These phonetic predictors also interacted with each other, with VOT having a more gradient effect when the positive VOT values were combined with the voiced base vowel or the negative VOT values were combined with the voiceless base vowel which created incongruence in the acoustic cues. The two DIALECT groups had similar responses, although European speakers heard a higher proportion of voiceless stops than Canadian speakers, particularly when negative VOTs were combined with the voiceless base vowel. We can also see the effect of SPEECH.RATE, the main predictor of interest, is visible in the positive VOT ranges. The circles and solid lines, representing voiceless stop responses in the fast rate condition, were generally higher than the triangles and dotted lines, representing voiceless stop responses in the slow rate condition, except when voiceless responses were

close to 100% regardless of speech rate. The rate effect for the negative VOT range, on the other hand, was inconsistent.

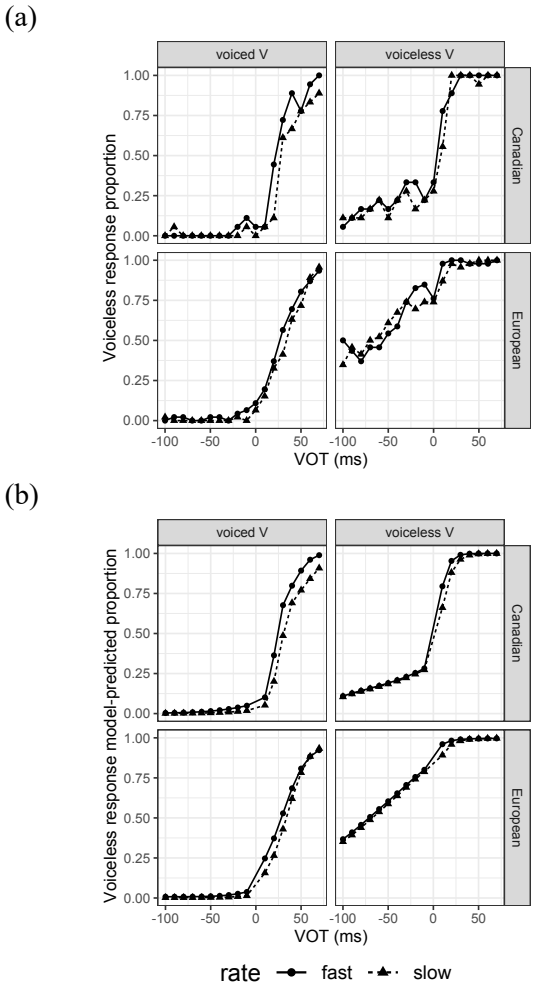

Figure 3: The average proportion of voiceless responses by VOT, BASE.VOWEL, DIALECT, and SPEECH.RATE: (a) Observed proportion; (b) Model predicted proportion

We now turn to the results of our statistical tests. First of all, in both the positive and the negative VOT models, by-PARTICIPANT random slopes for VOT and BASE.VOWEL were both retained as well as the random intercept, which means that participants varied in their phonetic cue use. By including these random slopes, our models test the phonetic predictors' effects at the group level while controlling for individual variation. Figure 3 (b) plots model-predicted voiceless response proportions.

We discuss the positive VOT model first. We found significant effects of the three linguistic predictors all in the expected direction (VOT: b = 3.16, $p < 0.001$; BASE.VOWEL: b = 2.78, $p < 0.001$; SPEECH.RATE: b = -0.43, $p < 0.001$). The main effect of DIALECT was not significant (b = - 0.24, p =0.641), but DIALECT interacted with VOT (b = - 0.71, p =0.018) and SPEECH.RATE (b = 0.31, p = 0.019) significantly. The three-way interaction of DIALECT x VOT x SPEECH.RATE was also significant (b = 0.26, p = 0.043). This means that the rate effect was stronger for Canadian speakers, and the dialect difference in speech rate effect was more pronounced in higher than lower VOT values.

For the negative VOT model, we also found a significant main effect of VOT (b = 1.08, p = 0.003), which means that not only the presence or absence of prevoicing but the duration of prevoicing can affect voicing perception. The main effect of BASE.VOWEL was also significant (b = 2.29, $p < 0.001$) in the expected direction. The interaction of VOT and BASE.VOWEL was significant (b = 0.40, p = 0.043), which means that the VOT duration effect was stronger when the base vowel comes from a voiced stop token. We found a significant effect of DIALECT (b = 1.08, p = 0.014) and a significant interaction of DIALECT and BASE.VOWEL (b = 0.81, p = 0.017). This means that the European speakers heard more voiceless stops than the Canadian speakers and this dialect difference is more pronounced when the base vowel is voiceless. Finally, we found a significant main effect of SPEECH.RATE (b = -0.34, p = 0.033) and a marginal interaction of SPEECH.RATE and BASE.VOWEL (b = 0.28, p = 0.082). This means that more voiceless stops were heard in fast speech than in slow speech, the opposite of the predicted pattern if the speech rate adjustment mirrors the shift in prevoicing duration in production, and this effect holds only when the base vowel is voiced and not when it is voiceless.

4. DISCUSSION

Our study examined if and how listeners adjust their perception of contrast between prevoiced and short-lag stops depending on the ambient speech rate. The results show that similar to the previous findings from the contrasts between short-lag and long-lag voiceless stops, listeners adjusted their expectation of VOT length and heard more voiceless stops (longer VOT category) in fast than in slow rate conditions. For the negative VOT values, fast speech induced more voiceless responses in the voiced base vowel condition. We interpret this marginal and unexpected effect as a lack of predicted effect. Despite the fact that listeners attend to the duration of prevoicing in perception, they were not more willing to accept a prevoiced stop as voiced in fast than in slow speech condition. We note that while both aspiration and prevoicing shortened in fast speech production, the effect was less consistent across speakers for prevoicing than for aspiration. This difference in production may explain the observed difference in perception between prevoicing and aspiration.

5. REFERENCES

[1] Kessinger, R., Blumstein, S. 1997. Effects of speaking rate on voice onset time in Thai, French and English. *Journal of Phonetics* 25, 143-168.

[2] Allen, J. S., Miller, J. L. 1999. Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *Journal of the Acoustical Society of America* 106 (4), 2031-2039.

[3] Summerfield, Q. 1981. Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception* 7(5), 1074.

[4] Kang, Y., Kung, K. Li, L. J., Ting, C., Yeung, J. 2018. Speech rate variation in English stop production and perception. *Toronto Working Papers in Linguistics 40*.

[5] Kidd, G. R. 1989. Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance* 15(4), 736.

[6] Miller, J. L., Green, K. P., Reeves, A. 1986. Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica* 43 (1-3), 106-115.

[7] Cho, T., Whalen, D. H., Docherty, G. 2019. Voice onset time and beyond: Exploring laryngeal contrast in 19 languages. *Journal of Phonetics* 72, 52-65.

[8] Atta, F., van de Weijer, J., Zhu, L. 2022. Saraiki. *Journal of the International Phonetic Association 52 (3),* 541-561.

[9] Hazan, V. L., Boulakia, G. 1993. Perception and production of a voicing contrast by French-English bilinguals. *Language and Speech* 36(1), 17-38.

[10] Abdelli-Beruh, N. B. 2004. The stop voicing contrast in French sentences: Contextual sensitivity of vowel duration, closure duration, voice onset time, stop release and closure voicing. *Phonetica* 61(4), 201-219

[11] Keating, P. 1984. Phonetic and phonological representation of stop consonant voicing. *Language* 60(2), 286-319.

[12] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.2.05, retrieved 5 January 2022 from http://www.praat.org/

[13] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2019. Gorilla in our midst: An online behavioral experiment builder. Behavior Research Methods.

[14] Sullivan, L. 2022. Participant error in online production data collection in Gorilla. Challenges for change: a crowd-sourced brainstorming session. A talk given at *LabPhon 18 Satellite Workshop.*

[15] Bates, D., Mächler M., Bolker B., Walker S., Christensen, B. R. H., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., Krivitsky, Pavel N. 2022. *Lme4:* Linear mixed-effects models using Eigen and S4. R package version 1.1-20.

[16] R Core Team. 2022. R: A Language and Environment for Statistical Computing. Vienna. R Foundation for Statistical Computing.

[17] Voeten, C. C. 2022. *Buildmer:* Stepwise Elimination and Term Reordering for Mixed-Effects Regression. R package version 2.4.

* This study was funded by the Social Sciences and Humanities Research Council of Canada (#435-2020-0209) and the University of Toronto Excellence Award.