Consonantal $F_0$ perturbation in American English involves multiple mechanisms

Yi Xu and Anqi Xu
Department of Speech, Hearing and Phonetic Sciences, University College London, London, United Kingdom

ABSTRACT:
In this study, we revisit consonantal perturbation of $F_0$ in English, taking into particular consideration the effect of alignment of $F_0$ contours to segments and the $F_0$ extraction method in the acoustic analysis. We recorded words differing in consonant voicing, manner of articulation, and position in syllable, spoken by native speakers of American English in both statements and questions. In the analysis, we compared methods of $F_0$ alignment and found that the highest $F_0$ consistency occurred when $F_0$ contours were time-normalized to the entire syllable. Applying this method, along with using syllables with nasal consonants as the baseline and a fine-detailed extraction procedure, we identified three distinct consonantal effects: a large but brief (10–40 ms) $F_0$ raising at voice onset regardless of consonant voicing, a smaller but longer-lasting $F_0$ raising effect by voiceless consonants throughout a large proportion of the following vowels, and a small lowering effect of around 6 Hz by voiced consonants, which was not found in previous studies. Additionally, a brief anticipatory effect was observed before a coda consonant. These effects are imposed on a continuously changing $F_0$ curve that is either rising-falling or falling-rising, depending on whether the carrier sentence is a statement or a question. © 2021 Acoustical Society of America. https://doi.org/10.1121/10.0004239

(Received 14 July 2020; revised 10 March 2021; accepted 22 March 2021; published online 29 April 2021)

[Editor: Zhaoyan Zhang]

I. INTRODUCTION
When a non-sonorant consonant occurs in a speech utterance, the vibration of the vocal folds is affected in two major ways. First, voicing may be interrupted, resulting in a break of otherwise continuous fundamental frequency ($F_0$) trajectory. This can be referred to as a horizontal disruption or voice break. Second, $F_0$ around the voice break may be raised or lowered because of the consonant. This is usually known as consonantal perturbation of $F_0$ (Hombert et al., 1979; Ohala, 1974). Other names include pitch skip (Haggard et al., 1970; Hanson, 2009), micro $F_0$ (Kohler, 1990), and CF0 (Kingston, 2007; Kirby and Ladd, 2016). We will refer to the raising and lowering effects as vertical perturbation in order to distinguish them from the effects of voice break. This distinction is necessary because research on the effects of consonants on $F_0$ over the past decades has focused predominantly on vertical perturbation, while the effects of voice break have received much less attention. As will be demonstrated, the assessment and interpretation of vertical perturbation is contingent on the treatment of voice break in $F_0$ measurement. In particular, full consideration of voice break may help answer four critical questions: (a) Are there both raising of $F_0$ by voiceless consonants and lowering of $F_0$ by voiced consonants? (b) Are there multiple mechanisms that jointly contribute to $F_0$ perturbation? (c) Are there both carryover and anticipatory $F_0$ perturbations? And (d) is $F_0$ perturbation affected by intonation?

A. Vertical perturbation and macro vs micro $F_0$
As early as in the middle of the last century, House and Fairbanks (1953) measured mean $F_0$ averaged across the entire vowel in English and found that it was higher after voiceless consonants than after voiced consonants. A similar finding was made by Lehiste and Peterson (1961) with peak $F_0$ as the measurement. Lea (1973) investigated the time course of the consonant perturbation and found that $F_0$ first rose after a voiceless consonant and then decreased throughout the vowel, while the opposite was true of voiced consonants. Hombert (1978) and Hombert et al. (1979) also reported a rise-fall dichotomy in the mean $F_0$ curves, as shown in Fig. 1, which has since been often cited as the prototypical dichotic consonantal perturbation of $F_0$. Later studies, however, started to show a more complex picture. Ohde (1984) and Silverman (1984) reported that $F_0$ fell after all obstruent consonants regardless of their voicing. Hanson (2009) applied an improved method to examine the time course of $F_0$ perturbation by including nasal consonants as the baseline. She found that $F_0$ was raised after voiceless consonants but not lowered after voiced ones. However, the rise-fall dichotomy remains a widely accepted notion, especially in its use as a key trigger for tonogenesis (Chen et al., 2017; Evans et al., 2018; Gao and Arai, 2019; Hill, 2019).

There has been less work on the anticipatory $F_0$ perturbation by consonants. Hombert et al. (1979) found no perturbation effect on the preceding vowels and Lehiste and Peterson (1961) reported that there was no consistent effect
for English. Kohler (1982), however, found that $F_0$ was lowered before voiced stops in contrast with voiceless stops when the sentence intonation is falling but not in sentences with either monotone or rising intonation. Silverman (1984) also reported a dichotomy in the preceding vowels according to consonant voicing.

As summarized above, there is still no clear consensus on vertical perturbation either as a carryover or anticipatory effect. In fact, two major issues remain unresolved. The first is the underlying cause of vertical perturbation. Two mechanisms have been proposed. The first is the aerodynamic hypothesis (Ladefoged, 1967), according to which the release of a voiceless stop is accompanied by a high rate of airflow across the glottis, which would increase the rate of vocal fold vibration. During a voiced consonant, on the other hand, the flow of air across the glottis is reduced, thus lowering pitch. The chief argument against this view is that the observed perturbatory effect lasts too long to be due to an aerodynamic effect. Lofqvist et al. (1995) have shown that the release of voiceless consonants is indeed accompanied by increased airflow, but only for a brief period of time, whereas vertical $F_0$ perturbation can last for at least 100 ms (Hombert et al., 1979).

An alternative hypothesis is that there is an adjustment of the tension of the vocal folds during the production of the consonant depending on voicing (Halle and Stevens, 1971). This is supported by electromyography (EMG) recordings that show higher cricothyroid (CT) activity during voiceless consonants than during voiced consonants (Dixit, 1975; Lofqvist et al., 1989). Also, significant voicing differences have been found in the vertical position of the larynx (Ewan and Krones, 1974) and the pharyngeal cavity (Bell-Berti, 1975; Westbury, 1983). The changes in the tension of the vocal folds would affect phonation threshold (Berry et al., 1996). In addition, the changes in laryngeal height would affect transglottal pressure (Hanson and Stevens, 2002). Both types of changes would help to stop voicing for voiceless consonants and sustain voicing for voiced consonants, but both of them would also affect $F_0$. The problem with this hypothesis is in fact part of the second unresolved issue about vertical perturbation: do voiced consonants actually lower $F_0$ or do they have no effects on $F_0$? So far there is no clear evidence that $F_0$ is lowered after voiced obstruents due to vocal folds slackening or larynx lowering. Hanson (2009) finds that $F_0$ following phonologically voiced stops in English is actually slightly higher than the nasal baseline. Kirby and Ladd (2016) reported that even for French and Italian voiced consonants (which are phonetically prevoiced consonants), there was only a marginal $F_0$ lowering after the oral closure according to the mean $F_0$ contours, and the effect was not statistically significant. These results have been further replicated in Kirby et al. (2020).

The above two possibilities have been considered as the only two alternative mechanisms so far. There is a third possibility that has not been contemplated before, however. That is, it is also possible that an aerodynamic effect and the effect of vocal fold tension both occur, but they differ in temporal scale. The aerodynamic effect may occur right after voice onset, but fade away quickly (Lofqvist et al., 1995), while the vocal fold tension effect may have a slow onset, but last longer (Hanson, 2009).

One of the reasons for the lack of consensus is that the observation of vertical perturbation may be affected by the method of its assessment. Silverman (1986) points out that the effect of consonantal perturbation cannot be properly understood unless the underlying intonation is well controlled. For example, if a consonant happens to occur in the course of a rising intonation, the $F_0$ rise after the consonant release may not be entirely due to the consonant. He further reports that, once the underlying intonation is taken into consideration, there is no more rise-fall dichotomy due to stop voicing in English because $F_0$ falls after both voiced and voiceless stops, except that the fall in the former is shallower than in the latter. Silverman’s argument is shadowed by the notion of macro versus micro $F_0$ (Kohler, 1982, 1990), the first of which refers to stress and intonation, and the second to segmental effects. Kohler (1982) reported that in German the $F_0$ divergence after voiced and voiceless consonants was large in rising or monotone contours but not in falling contours, while the effect of voicing of a following stop in $F_0$ was observable only in falling contours.

It is not always obvious what an underlying intonation looks like around a consonant, however. Although one could infer it from the $F_0$ trajectories before and after the consonant, it is also possible that a sharp pitch turn takes place right before, after, or even during the consonant. When that happens, the assessment of vertical perturbation becomes tricky. What is needed is a careful consideration of the relation between underlying intonation and voice break.

**B. Voice break and $F_0$-syllable alignment**

In a sentence consisting of only vowels and sonorant consonants, like the Mandarin phrase /hei1 ni2 li3 mao4/ (black woolen hat) in Fig. 2(a) (where the numbers indicate...
the high, rising, low, and falling tones, respectively), the $F_0$ trajectory would be largely smooth and continuous throughout the utterance. This is because the tension of the vocal folds, which is mainly responsible for $F_0$, cannot change instantaneously. A voluntary pitch change of just one semitone would take over 100 ms to complete on average (Xu and Sun, 2002). Once obstruent consonants occur in an utterance, continuous $F_0$ is interrupted by the voice breaks during the constriction and sometimes also during the release, as is the case with the Mandarin expression /shān1 qióng2 shuǐ3 jīn4/ (no way out) in Fig. 2(b). A question then arises as to whether the voice break also interrupts the continuous adjustment of vocal fold tension. This question might seem unwarranted, as how can there be a question then arises as to whether the voice break also interrupts the continuous adjustment of vocal fold tension. This question might seem unwarranted, as how can there be a question then arises as to whether the voice break also interrupts the continuous adjustment of vocal fold tension. This question might seem unwarranted, as how can there be a question then arises as to whether the voice break also interrupts the continuous adjustment of vocal fold tension. This question might seem unwarranted, as how can there be a

A further issue is how exactly $F_0$ contours should be aligned relative to the syllable. It has been shown that the $F_0$ contour of a syllable in English is a movement toward an underlying pitch target associated with lexical stress as well as other concurrent functions (Fry, 1958; Liu et al., 2013; Xu and Xu, 2005). It is further shown that such target approximation movement is synchronized with the syllable in English (Prom-on et al., 2009; Xu and Prom-on, 2014; Xu and Xu, 2005), just like in Mandarin (Xu, 1998, 1999), i.e., starting from the syllable onset and ending by syllable offset (Xu and Wang, 2001; Xu, 2020).

Assuming that the target approaching $F_0$ movement is indeed synchronized with the syllable in English, the full effect of voice break would be most clearly seen by using sonorant consonants like nasals as the reference, as they allow $F_0$ to be fully continuous with little vertical perturbation (Xu, 1999; Xu and Xu, 2005). Figure 3 is an illustration based on data from the present study. Here, the solid curve represents the $F_0$ contour of a syllable with a nasal onset, and the dotted and dashed curves represent those in syllables with voiced and voiceless initial stops, respectively. All the contours are aligned by the onset of the consonant closure on the left and by the offset of the vowel on the right. The time in between is normalized across all the contours. As can be seen, $F_0$ in both stops starts much later than in the nasal, but they also differ from each other in timing, because voiceless stops have longer voice onset time (VOT) than voiced consonants. What is important is that the estimated vertical perturbation would be different if the alignment of $F_0$ contours is changed. If the onset of the non-sonorant consonant contours is shifted leftward, the magnitude of the estimated perturbation would increase. Furthermore, if the onset of voiceless consonants is shifted leftward to align with the voiced consonants, the difference between them in perturbation would also increase. Therefore, how $F_0$ onsets are aligned to each other is a potential confound in the assessment of vertical perturbation.

In previous studies (Chen, 2011; Chen et al., 2017; Lea, 1973; Hombert, 1978; Jun, 1996; Ohde, 1984), including also those that have used nasal consonants as reference (Hanson, 2009; Kirby and Ladd, 2016; Kirby et al., 2020), $F_0$ contours have always been aligned at the onset of the vowel when estimating $F_0$ perturbation, as in Fig. 3(c). They differ only in terms of whether there are additional alignment points and whether time-normalization is applied. Some studies applied fixed time windows for the $F_0$ contours under comparison: 80 ms in Chen (2011), 100 ms in Jun (1996), and 150 ms in Hanson (2009). Instead of fixed time windows, Kirby and Ladd (2016) and Kirby et al. (2020) aligned the $F_0$ contours at vowel onset and offset, and then applied time-normalization across the vowel. The same method was also used by Gao and Arai (2019). By aligning $F_0$ contours at vowel onset, however, the potential effects of voice break on the assessment of vertical perturbation cannot be seen. Part of the goal of the present study is therefore to find this missing information by considering alternative alignments such as those shown in Figs. 3(a) and 3(b).

A further methodological issue is the quality of $F_0$ trajectory extraction. The finding of two different kinds of $F_0$ perturbation in the present study may help to explain the low consensus on the rise-fall dichotomy between voiced

![FIG. 2.](Color online) (a) Spectrogram of utterances consisting of only vowels and sonorants; (b) spectrogram of utterances consisting of vowels and consonants.)
and voiceless stops in previous studies. Those that do not catch the initial jumps (House and Fairbanks, 1953; Lehiste and Peterson, 1961; Lea, 1973; Hombert et al., 1979; Hanson, 2009) tend to report a simple voicing contrast with \( F_0 \) following voiceless stops being higher than the voiced stops. When the initial jumps are preserved, the \( F_0 \) falling after both types of consonants is observed (Ohde, 1984; Silverman, 1984; Hanson, 2009). In our statistical comparison of the initial jump of voiced and voiceless stops, the conventional way of \( F_0 \) processing that removes the abrupt \( F_0 \) shift with trimming and smoothing led to a statistically significant voicing contrast. However, when the initial jump was preserved, the \( F_0 \) following voiced and voiceless obstruent consonants was statistically indistinguishable.

C. The present study

The present study is designed to answer the four critical questions raised in Sec. I by assessing the size and manner of vertical perturbation based on direct comparisons of syllable-wise \( F_0 \) contours both before and after the consonant closure. The new approach takes a more careful consideration of alignment and time normalization than has been done before, based on a number of assumptions. First, as discussed in the above section, the adjustment of vocal fold tension should be continuous (rather than in a temporary halt) during the consonant closure. Second, each syllable should have a targeted pitch pattern or pitch target in English as one of its articulatory goals, and this pitch target is associated with word stress as well as other concurrent functions (Fry, 1958; Liu et al., 2013; Xu and Xu, 2005). Second, the \( F_0 \) movement toward the pitch targets is fully synchronized with the syllable in English (Prom-on, Xu and Thipakorn, 2009; Xu and Prom-on, 2014; Xu and Xu, 2005) as is in Mandarin (Xu, 1998, 1999).

Another major source of discrepancy in previous reports of perturbation is the technical precision in \( F_0 \) extraction. Earlier studies compared \( F_0 \) values at a few acoustic landmarks or averaged across a long interval (House and Fairbanks, 1953; Lehiste and Peterson 1961). Later experiments have often used autocorrelation with large smoothing windows to extract \( F_0 \) contours (Kingston, 2007; Kirby and Ladd, 2016). These methods are not highly sensitive to brief changes in fundamental frequency. As shown by Ohde (1984), brief pitch spikes can often be found at consonant offsets when \( F_0 \) is computed directly from vocal cycles. Those spikes are consistent with the \( F_0 \) falls at the voice onset reported by Silverman (1984). When using \( F_0 \) extraction algorithms with sizable smoothing windows, the spikes might be missed entirely, or smoothed into the following contour, creating the appearance of a long-lasting perturbation (see Fig. 1). In order to catch any consistent but brief perturbations, there is a need to extract \( F_0 \) directly from vocal cycles, as will be described in Sec. IID.

II. METHOD
A. Stimuli

The stimuli (Table I) were chosen to allow variation of a target consonant within a varying linguistic context. Target consonants were nasals, voiced and voiceless fricatives, stops and stop-sonorants, and voiceless affricates. These were embedded in CV syllables, CVC syllables with the first consonant as nasals, and CVCV syllables with the first consonant as either nasals or laterals. The target words were embedded in the carrier sentences “I should say W next time.” and “Should I say W next time?” The carries

<table>
<thead>
<tr>
<th>CV</th>
<th>CVC</th>
<th>CVCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiceless</td>
<td>Voiced</td>
<td>Voiceless</td>
</tr>
<tr>
<td>Nasal</td>
<td>say</td>
<td>they</td>
</tr>
<tr>
<td>Fricative</td>
<td>tray</td>
<td>day</td>
</tr>
<tr>
<td>Stop</td>
<td>tray</td>
<td>day</td>
</tr>
<tr>
<td>Stop sonorant</td>
<td>Che</td>
<td></td>
</tr>
<tr>
<td>Affricate</td>
<td>Che</td>
<td></td>
</tr>
</tbody>
</table>
were chosen to prevent the target consonants from being resyllabified with surrounding contexts (Xu, 1998).

B. Subjects

Subjects were four women and four men, all residents of New Haven, CT, and mostly students at Yale University. Their ages ranged from 20 to 54 years (from 20 to 24, excluding one subject), and all were native speakers of General American English. One subject, who had no difficulty with the task, had received six months of speech therapy as a young child, to treat a minor lisp. Otherwise, no speech or language disorders were reported.

C. Recording procedure

The recording was done in a soundproof studio at Haskins Laboratories, New Haven, CT. Subjects sat before a computer screen, on which one stimulus sentence appeared at a time. They read each sentence out loud into a head-mounted microphone and were recorded digitally onto the hard drive of an Apple Macintosh computer. Each sentence was presented five times. To elicit a narrow focus on the target word, we presented it in all capital letters and instructed subjects to emphasize it. Other intonational patterns, noticeable pauses, or voicing anomalies (most commonly creaky voice) rendered some tokens unusable. When this was noticed during the recording, the subject was asked to repeat the sentence. Some problems were not noticed, however, and occasionally both instances of a repeated token turned out to be usable, so the actual number of tokens was in some cases more or less than five.

D. Pitch extraction and processing

Phonetic data were extracted using a special version of ProsodyPro (Xu, 2013), a Praat (Boersma and Weenink, 2020) script for large-scale analysis of speech prosody. The script first used Praat’s To PointProcess function to mark all the vocal cycles. The marked cycles were then manually rectified before being converted to \( F_0 \) curves. Segment boundaries were manually labeled at the onset of consonant closure and at the onset of vowel formants in both the target word and part of the carrier (… say __ next…), as illustrated in Fig. 4.

In the case of the sentence “I should say name next time,” the boundary between [m] and [n] was not always easy to determine from the waveform or the spectrogram. Sometimes there was a faint burst that accompanied the labial release, and this was marked as the boundary, as shown in Fig. 5(a). Otherwise, the boundary was marked in the center of geminated nasal murmur [Fig. 5(b)].

Further analyses were performed using a custom-written version of ProsodyPro. The \( F_0 \) curves were trimmed with an algorithm described in Xu (1999), to remove sharp spikes. The vocal cycle next to a silent interval longer than 33 ms was exempted from this trimming to preserve the sharp spikes that consistently occur at voice onset and offset (based on the assumption that normal \( F_0 \) would not go below 30 Hz). The statistical analysis was conducted using linear mixed-effect models by lme4 (Bates et al., 2015) and emmeans (Lenth et al., 2020) for post hoc tests in the R (R Core Team, 2020). Random intercepts for SUBJECT and by-SUBJECT random slopes for fixed effects were then incorporated maximally (Barr et al., 2013). Subsequently, potential fixed effects were added. Only fixed effects that were judged to be superior to less specified models tested by likelihood-ratio tests were included in the model.

III. RESULTS

A. Graphical comparison of \( F_0 \) contours

Before deciding what measurements to take for statistical analysis, we first made direct comparisons of the \( F_0 \) contours to identify major differences between the conditions.
Figure 6 shows examples of mean $F_0$ contours by individual subjects, with Fig. 6(a) showing those of the target word /nay/ in a statement and Fig. 6(b) in a question. The vertical differences in $F_0$ are large, with female subjects tending to have higher fundamental frequencies. There are some differences in the location of the $F_0$ peaks. Regardless of the differences in the vertical level and the peak location, however, all speakers show similar general patterns.

Figure 7 shows mean $F_0$ contours with different ways of alignment and normalization. $F_0$ of CV syllables and parts of the carrier sentence in statements are aligned at vowel voice onset (a), syllable onset (b), syllable offset (c), and normalized across the entire syllable with alignment at both syllable edges (d). For display purposes only, each contour is an average across all repetitions by all subjects of the given stimulus. When averaging, each segment of each token is sampled at 20 even-spaced points. In the real-time plots, the mean time and $F_0$ of each of the points were averaged across repetitions and speakers. For the time-normalized plots, the mean time of each type of consonant was recalculated with reference to the mean time of nasals to align these points at both syllable onset and offset. The average plots in Figs. 7–9 reliably represent our data (see the supplementary material\(^2\) for individual plots for all participants).

In order to establish an appropriate reference level, we plotted $F_0$ curves using the syllable-wise alignment and conventional alignment methods employed in previous research. As can be seen in Fig. 7, methods of alignment and time-normalization both have clear consequences. When aligned at voice onset [Fig. 7(a)] following previous studies (Lea, 1973; Hombert, 1978; Ohde, 1984; Jun, 1996; Hanson, 2009; Chen, 2011), the $F_0$ curves of different consonants vary greatly both before and after the consonants. Aligning the $F_0$ contours at syllable onset [Fig. 7(b)] results in variations at the end of the syllable and the following contexts. When the $F_0$ contours are aligned at both vowel onset and offset [Fig. 7(c)], as done in Kirby and Ladd (2016), Kirby et al. (2020), and Gao and Arai (2019), the amount of cross-consonant $F_0$ difference is as large as in Fig. 7(a). Time normalizing $F_0$ curves between the onset and offset of the target syllable [Fig. 7(d)] seems to exhibit the least variable $F_0$ patterns across consonant types both within the target syllable and in the surrounding carrier sentences. In the following analysis, therefore, we will focus on comparing $F_0$ contours time-normalized with respect to the syllable.

Looking more closely at Fig. 7(d), we can see that, with the exception of voiced fricative, $F_0$ is first perturbed upward by non-sonorant consonants relative to the nasal baseline, although there are also apparent differences in voice onset time between various types of consonants. Afterward, for most of the consonant types, $F_0$ drops sharply toward the nasal baseline and starts to shadow its contour shape for the rest of the syllable. However, for voiceless stops, surprisingly, $F_0$ first rises rather than falls, and then also starts to shadow the nasal contour. Besides the initial drop or rise, there are also apparent differences between the consonant types in subsequent overall $F_0$ height, with voiceless consonants generally having higher $F_0$ than voiced consonants. These height differences, though gradually reducing over time, persist all the way to the end of the vowel.

Figure 8 displays $F_0$ contours in questions with various alignment and time-normalization schemes. Again, $F_0$ is perturbed upward after all non-nasal segments, although there is much variation in terms of perturbation size. After this initial jump, like in statements, $F_0$ quickly drops toward the nasal baseline and starts to shadow its shape for the rest of the syllable duration. Interestingly, voiceless stops again show the smallest perturbation/jump among the voiceless consonants. But unlike in statements, $F_0$ drops rather than rises after the initial jump. Presumably, the initial jump, though small in size, has raised $F_0$ much higher than the targeted low $F_0$ represented by the nasal contour. Also, like in statements, the overall $F_0$ height after the initial jump is higher in voiceless consonants than in voice consonants.
Figure 9 shows $F_0$ contours of CVC [Figs. 9(a) and 9(b)] and CVCV [Figs. 9(c) and 9(d)] syllables with part of the carrier sentences in statements and questions. In both cases, the target consonant is the second consonant in the sequences. These syllables enable the examination of anticipatory effects of obstruent consonants on the preceding $F_0$ within and across syllable boundaries. For CVC syllables in statements, as can be seen in Figs. 9(a) and 9(b), pre-closure $F_0$ of non-sonorant consonants inevitably drops sharply after reaching a peak. But before those drops, the overall $F_0$ height is raised in all cases relative to the nasal baseline. Interestingly, here the consonants seem to be grouped by voicing in statements. Similar overall raising of $F_0$ height by coda consonants is also seen in questions, except that there are no sharp drops before consonant closure. In contrast, for CVCV syllables, as shown in Figs. 9(c) and 9(d), the $F_0$ contours of vowels preceding the target consonants do not seem to diverge in both statements and questions. Instead, the lack of the anticipatory effect appears to parallel what we have seen in Figs. 7 and 8 for CV syllables, where the $F_0$ of vowels in the carrier words converges regardless of the upcoming consonants.

To summarize the graphical comparison, with $F_0$ contours of nasal consonants as the baseline, a number of initial observations can be made. First, non-sonorant initial consonants seem to exert two kinds of perturbations: (a) an abrupt initial jump in $F_0$ at voice onset, followed by either a sharp drop or rise (voiceless stop in statement), and (b) a sustained raising (voiceless consonant) or lowering of $F_0$ height throughout the rest of the syllable. Second, non-sonorant coda consonants also seem to exert two kinds of perturbations: (a) an abrupt drop in $F_0$ right before voice offset in statements, and (b) a raising of $F_0$ that extends back toward the midpoint of the vowel. Finally, aspiration, especially in stops, seems to reduce the magnitude of initial jump. This has led to a rise rather than a drop of $F_0$ immediately after
voice onset in a statement. In the next session, we will run statistical tests on the raw data to verify the visual observations.

B. Statistical analysis

The graphical comparison of $\text{F}_0$ contours shows initial indication of three different kinds of influences by initial consonants on $\text{F}_0$: (a) a voice break that interrupts continuous $\text{F}_0$, (b) a brief yet sometimes large jump relative to the nasal baseline, and (c) a long lasting raising or lowering effect, also relative to the nasal baseline. To closely examine these influences, closure duration, onset $\text{F}_0$, $\text{F}_0$ jump, $\text{F}_0$ elbow, elbow jump, and offset $\text{F}_0$ of all the repetitions by each speaker were measured and analysed, as illustrated in Fig. 10. For voiceless consonants, the closure duration equals VOT, while for voiced consonants, it is the time elapsed between the oral closure and the onset of the following vowel (thus disregarding any voicing during closure). Onset $\text{F}_0$ is the conventional way of observing initial consonantal perturbation, which is the first $\text{F}_0$ point at the onset of the vowel. $\text{F}_0$ jump is a new measurement not used in previous studies, which indicates the difference between onset $\text{F}_0$ and the $\text{F}_0$ of nasal baseline at the same relative time in normalized time, in the same intonation. Similar to $\text{F}_0$ jump, elbow jump is another new measurement that indicates the difference between $\text{F}_0$ elbow and the $\text{F}_0$ of nasal baseline in the same intonation at the same relative time in normalized time, where $\text{F}_0$ elbow is the $\text{F}_0$ turning point after the initial $\text{F}_0$ jump. Finally, offset $\text{F}_0$ is the $\text{F}_0$ at the end of the vowel preceding a target consonant, which evaluates whether the perturbation effects last until the end of the syllable.

1. Carryover effect

a. Consonant closure duration. As we can see from Figs. 7 and 8, there are noticeable differences in closure time between various classes of consonants, and the shape
of $F_0$ contours at the beginning of the following vowels are influenced by the duration of the closure. The longer the closure, the greater the magnitude of the initial $F_0$ perturbation, except for voiced stops. Table II lists means and standard deviations of closure duration of consonants in CV syllables separated by consonant types and intonation contexts. For the sake of data balance, statistical analysis was performed only on the stops, fricatives, and stop-sonorants that are minimal pairs. In a set of linear mixed models, CVOICE (voiced, voiceless), CMANNER (stop, fricative and stop-sonorant), INTONATION (statement, question), and their interaction were included as potential fixed effects. CVOICE improves the fit of the model ($\chi^2 = 24.077$, df = 1, $p < 0.001$): voiceless consonants tend to have longer closures than voiced consonants. CMANNER ($\chi^2 = 18.255$, df = 2, $p < 0.001$) also significantly predicts closure duration. The post hoc comparison showed that stop-sonorants have longer closures than fricatives ($p < 0.001$) and stops ($p = 0.046$). Meanwhile, closure duration of stops is longer than the fricatives ($p = 0.005$). INTONATION ($\chi^2 = 2.591$, df = 1, $p = 0.108$) does not significantly improve the model. The interaction between CVOICE and CMANNER ($\chi^2 = 10.861$, df = 2, $p = 0.004$) is significant. When the consonant is voiceless, the contrast in closure duration between

![FIG. 9. (Color online) Mean $F_0$ contours of vowels following target consonants in CVC syllables [(a) and (b)] and CVCV [(c) and (d)] and parts of carrier sentences. The time points of consonants are normalized with reference to the mean time points of nasals. Carrier sentence is declarative [left, (a) and (c)] or interrogative [right, (b) and (d)]. The vertical lines indicate the alignment points and the symbolic markers indicate segment boundaries. The consonants having the same manner of articulation are in paired colours with different grayscale values. The voiced consonants are darker than their voiceless counterparts.](image)

![FIG. 10. Illustration of onset $F_0$, $F_0$ jump, $F_0$ elbow, elbow jump, and offset $F_0$.](image)
The conventional way of only measuring onset does not consider closure duration, which may have implications. Here, we compare the onset and closure duration to understand their effects. Table III lists the percentages of phonetically voiced tokens among all phonological voiced consonants. As we can see from the table, there are individual differences in the production of voicing. Voicing is more likely to begin during the constriction for voiced fricatives and voiced stop sonorants compared with voiced stops. Most of the voiced stops are realized as voiceless unaspirated stops (72%), while the percentages of phonetically voiceless fricatives (33%) and stop sonorants (56%) are much lower. In addition, there are individual differences in voicing implementation. One of the speakers (F4) consistently devoiced all the voiced consonants, but the initial perturbation still differs substantially among all phonological voiced consonants. As we can see in Fig. 11, when F0 trimming and smoothing is applied, the onset F0 differs by a large amount after voiced stops and voiceless stops. However, when F0 is obtained without trimming and smoothing, the first few pitch values are very similar regardless of voicing feature.

The distributions of the onset F0 and F0 jump following voiced and voiceless stops obtained by different pitch processing methods are shown in Fig. 12. A clear distinction of voicing feature can be seen in the trimmed onset F0, while no such effect is observable in the untrimmed onset F0 and F0 jump. We ran statistical tests on the onset F0 and F0 jump obtained by the two methods to see whether the pitch extraction and processing method had a significant impact. The main effect of CVOICE is only significant in the model for the trimmed onset F0 (\(\chi^2 = 8.386, df = 1, p = 0.003\)) but not for either the untrimmed onset F0 (\(\chi^2 = 0.008, df = 1, p = 0.930\)) or the untrimmed F0 jump (\(\chi^2 = 0.799, df = 1, p = 0.371\)). The results indicate that the contrast between F0 following voiced and voiceless is exaggerated when trimming and smoothing are applied.

Following the new method, we further evaluated the initial perturbation of other consonant types by measuring both onset F0 and F0 jump, as summarized in Table II. As can be seen, the standard derivation (SD) of onset F0 (SD, 51) is larger than that of F0 jump (SD, 27) across different consonant types.

stops and fricatives is not significant (\(p = 0.895\)), but the contrast is significant in voiced consonants (\(p = 0.004\)).

The realisation of voicing in English consonants is influenced by linguistic contexts such as word position, adjacent consonants, and lexical tones (Davidson, 2016). Table III lists the percentages of phonetically voiced tokens among all phonological voiced consonants. As we can see from the table, there are individual differences in the production of voicing. Voicing is more likely to begin during the constriction for voiced fricatives and voiced stop sonorants compared with voiced stops. Most of the voiced stops are realized as voiceless unaspirated stops (72%), while the percentages of phonetically voiceless fricatives (33%) and stop sonorants (56%) are much lower. In addition, there are individual differences in voicing implementation. One of the speakers (F4) consistently devoiced all the voiced consonants, but the initial perturbation still differs substantially after voiced and voiceless consonants (see supplementary material for by-speaker plots). For four of the speakers (F2, F3, M3, and M4), F0 rises after voiceless stops, exhibiting a distinct pattern from other voiceless consonants (see supplementary material for by-speaker plots).

### b. Onset F0 and F0 jump

As shown in the previous section, closure duration varies with voicing. These variations may affect F0 at vowel onset, as seen in Figs. 7 and 8. The conventional way of only measuring onset F0 does not take closure duration into consideration, which may have potentially exaggerated or masked true vertical perturbation. Here, we compare the onset F0 of stop consonants measured by the conventional pitch-processing method based on autocorrelation with F0 trimming and smoothing and by our new method (i.e., without trimming and smoothing). As can be seen in Fig. 11, when F0 trimming and smoothing is applied, the onset F0 differs by a large amount after voiced stops and voiceless stops. However, when F0 is obtained without trimming and smoothing, the first few pitch values are very similar regardless of voicing feature.

The distributions of the onset F0 and F0 jump following voiced and voiceless stops obtained by different pitch processing methods are shown in Fig. 12. A clear distinction of voicing feature can be seen in the trimmed onset F0, while no such effect is observable in the untrimmed onset F0 and F0 jump. We ran statistical tests on the onset F0 and F0 jump obtained by the two methods to see whether the pitch extraction and processing method had a significant impact. The main effect of CVOICE is only significant in the model for the trimmed onset F0 (\(\chi^2 = 8.386, df = 1, p = 0.003\)) but not for either the untrimmed onset F0 (\(\chi^2 = 0.008, df = 1, p = 0.930\)) or the untrimmed F0 jump (\(\chi^2 = 0.799, df = 1, p = 0.371\)). The results indicate that the contrast between F0 following voiced and voiceless is exaggerated when trimming and smoothing are applied.

Following the new method, we further evaluated the initial perturbation of other consonant types by measuring both onset F0 and F0 jump, as summarized in Table II. As can be seen, the standard derivation (SD) of onset F0 (SD, 51) is larger than that of F0 jump (SD, 27) across different

---

**TABLE II. Means (standard deviations) of closure duration (ms), onset F0 (Hz), and F0 jump (Hz).**

<table>
<thead>
<tr>
<th>Consonant type</th>
<th>Statement</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Closure duration</td>
<td>Onset F0</td>
</tr>
<tr>
<td>Nasal</td>
<td>118 (21)</td>
<td>156 (43)</td>
</tr>
<tr>
<td>Voiced stop</td>
<td>122 (31)</td>
<td>174 (46)</td>
</tr>
<tr>
<td>Voiced fricative</td>
<td>102 (27)</td>
<td>157 (48)</td>
</tr>
<tr>
<td>Voiced stop-sonorant</td>
<td>134 (21)</td>
<td>163 (44)</td>
</tr>
<tr>
<td>Voiced consonant (excl. nasal)</td>
<td>119 (24)</td>
<td>165 (50)</td>
</tr>
<tr>
<td>Voiced fricative</td>
<td>175 (30)</td>
<td>177 (46)</td>
</tr>
<tr>
<td>Voiced stop-sonorant</td>
<td>172 (26)</td>
<td>209 (52)</td>
</tr>
<tr>
<td>Voiceless stop</td>
<td>189 (27)</td>
<td>192 (42)</td>
</tr>
<tr>
<td>Voiceless affricate</td>
<td>184 (29)</td>
<td>206 (47)</td>
</tr>
<tr>
<td>Voiceless consonant</td>
<td>179 (26)</td>
<td>196 (45)</td>
</tr>
</tbody>
</table>

---

**TABLE III. Percentages of phonetically voiced tokens in phonologically voiced stops, fricatives, and stop sonorants.**

<table>
<thead>
<tr>
<th>Consonant type</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>100</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>60</td>
<td>0</td>
<td>60</td>
<td>0</td>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td>Fricative</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>40</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Stop-sonorant</td>
<td>20</td>
<td>100</td>
<td>20</td>
<td>100</td>
<td>20</td>
<td>100</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>100</td>
<td>20</td>
<td>100</td>
<td>20</td>
<td>100</td>
<td>60</td>
</tr>
</tbody>
</table>

---

**FIG. 11.** (Color online) Schematic comparisons of F0 perturbation following voiced and voiceless obstruent consonants when applied with (solid) and without (dotted) trimming and smoothing pitch processing.
conditions. This is further confirmed in Fig. 13, where the boxplots show that $F_0$ jump is more consistent, i.e., with smaller variance than onset $F_0$ in both statements and questions, especially for voiceless consonants.

The main effect of CVOICE is significant in the model for onset $F_0$ ($\chi^2 = 10.491$, df = 1, $p = 0.001$) and for $F_0$ jump ($\chi^2 = 8.398$, df = 1, $p = 0.004$). Voiceless consonants show a greater onset $F_0$ as well as $F_0$ jump than voiced consonants. In contrast, CMANNER does not seem to have an impact on either onset $F_0$ ($\chi^2 = 4.268$, df = 2, $p = 0.118$) or $F_0$ jump ($\chi^2 = 5.016$, df = 2, $p = 0.081$). Further, INTONATION is non-significant for either onset $F_0$ ($\chi^2 = 2.664$, df = 1, $p = 0.103$) or $F_0$ jump ($\chi^2 = 1.751$, df = 1, $p = 0.186$).

The interaction between CVOICE and CMANNER is significant for both onset $F_0$ ($\chi^2 = 104.260$, df = 4, $p < 0.001$) and $F_0$ jump ($\chi^2 = 104.950$, df = 4, $p < 0.001$). As demonstrated in Fig. 14, the voicing contrast is more salient in fricatives (onset $F_0$: $p < 0.001$; $F_0$ jump: $p < 0.001$) and stop-sonorants (onset $F_0$: $p < 0.001$; $F_0$ jump: $p = 0.012$) than in stops (onset $F_0$: $p = 1.000$; $F_0$ jump: $p = 0.968$). It is worth noting that the interaction between CVOICE and INTONATION is significant in the model for onset $F_0$ ($\chi^2 = 8.136$, df = 2, $p = 0.017$), whereas $F_0$ jump is not affected by the interaction ($\chi^2 = 1.751$ df = 1, $p = 0.186$). As seen in Fig. 13, the onset $F_0$ of voiceless consonants is marginally higher in statements than questions ($p = 0.097$), but that of voiced stops is similar across intonation ($p = 0.786$).

For $F_0$ jump, which results from subtraction of the nasal baseline from onset $F_0$, the interference from the interaction between voicing and intonation is eliminated.

What remains unclear is whether the voicing contrast in the initial perturbation is due to $F_0$ raising by voiceless consonants or $F_0$ lowering by voiced consonants. We plotted a histogram of $F_0$ jump for all consonant types in Fig. 15. As can be seen, except for voiceless stops, nearly all the $F_0$ jumps of voiceless consonants are above zero, which suggests a significant $F_0$ raise relative to nasals. And, interestingly, $F_0$ jumps in voiced stops are also distributed largely above zero. In contrast, voiceless fricatives and voiceless stop-sonorants contain both negative and positive values. This indicates that voiced stops significantly raise $F_0$ at vowel onset relative to the nasal baseline, just like voiceless consonants, which is consistent with the findings of Ohde (1984) and Silverman (1984). In other words, instead of $F_0$ lowering versus $F_0$ raising, voiced and voiceless stops differ only in the magnitude of $F_0$ raising as far as $F_0$ jumps are concerned.

**c. $F_0$ elbow and elbow jump.** As can be seen in Figs. 7 and 8, the initial $F_0$ jump does not last long and the $F_0$ trajectories of different consonants gradually converge toward the nasal baseline after a sharp turn. The turning point ($F_0$ elbow) occurs around 41 ms (SD = 22) after vowel onset. However, it is not the case that an $F_0$ elbow occurs
after vowel onset in every utterance. The count and the height of $F_0$ elbow and elbow jump (the difference between $F_0$ elbow and the $F_0$ of nasal baseline in the same intonation at the same relative time point in normalized time, cf. Fig. 10) are summarized in Table IV. Figure 16 shows values of $F_0$ elbow and elbow jump in different voicing and intonation conditions. Like in the case of onset $F_0$ and $F_0$ jump, more variances can be seen in $F_0$ elbow (SD = 45) than in elbow jump (SD = 15). We fitted separate models for $F_0$ elbow and elbow jump with CVOICE (voiced, voiceless), CMANNER (stop, fricative, stop-sonorant), INTONATION (statement, question), and their interactions as potential fixed effects. The main effect of CVOICE is significant on $F_0$ elbow ($\chi^2 = 17.339$, df = 1, $p < 0.001$) and elbow jump ($\chi^2 = 9.270$, df = 1, $p = 0.002$): Voiceless consonants have higher $F_0$ elbow values than voiced consonants. CMANNER does not improve the fit of the model for either $F_0$ elbow ($\chi^2 = 0.442$, df = 2, $p = 0.801$) or elbow jump ($\chi^2 = 0.348$, df = 2, $p = 0.175$). $F_0$ elbow differs across intonation patterns ($\chi^2 = 6.406$, df = 1, $p = 0.011$): higher in declarative sentences than in interrogative sentences. In contrast, INTONATION does not significantly predict elbow jump ($\chi^2 = 1.074$, df = 1, $p = 0.3$). Similar to the results of onset $F_0$ and jump $F_0$ presented earlier, the interaction between CVOICE and INTONATION significantly improves the fit of the model for $F_0$ elbow ($\chi^2 = 6.806$, df = 1, $p = 0.009$) but not for elbow jump ($\chi^2 = 1.271$, df = 2, $p = 0.530$). The $F_0$ elbow of voiceless consonants has higher values in statements than in questions ($p = 0.002$), but not for voiced consonants ($p = 0.082$) (see Fig. 16).

FIG. 14. (Color online) Interaction between voicing and manner of articulation in onset $F_0$ (left, a) and $F_0$ jump (right, b). Nasals and affricates are excluded.

FIG. 15. (Color online) Histographic distributions of $F_0$ jump values by consonant type. The upper panel shows distributions of $F_0$ jump for voiced consonants and the lower panel for voiceless consonants. In each plot, the dashed vertical line marks the zero point on the x axis.
Figure 17 shows the values of elbow jump for each consonant type. Even after the abrupt initial $F_0$ jump, there are still clear differences between the $F_0$ values after voiced and voiceless consonants. Compared with the distribution of $F_0$ jump (Fig. 15), the raising effects by voiceless consonants have reduced while the lowering effects of voiced consonants have become more evident.

d. Offset $F_0$. As seen in Figs. 7 and 8, the differences in $F_0$ across consonant types do not end by the $F_0$ elbows but are sustained through the rest of the syllable. Remarkably, what can also be noticed is that the divergence in offset $F_0$ between voiced and voiceless consonants is not only due to the upward $F_0$ shifts following voiceless consonants but also due to the downward $F_0$ shifts following voiced consonants. Means and standard deviations of offset $F_0$ under different conditions are provided in Table V. Offset $F_0$ following voiced consonants is considerably lower than the nasal baseline, whereas it is close to the nasal baseline following voiceless consonants. We ran a series of linear mixed models to test whether the voicing contract remains statistically significant by the end of the syllable. CVOICE (voiced, voiceless), CMANNER (stop, fricative), INTONATION (statement, question), and their interaction as potential fixed effects. In CVC syllables, the main effect of CVOICE (voiced, voiceless) improves the fit of the model ($\chi^2 = 6.654$, df = 1, $p = 0.010$): The offset $F_0$ of vowels following voiceless consonants is higher than the ones following voiced consonants. However, neither CMANNER (stop, fricative, stop-sonorant: $\chi^2 = 3.365$, df = 2, $p = 0.186$) nor INTONATION (statement, question: $\chi^2 = 1.367$, df = 1, $p = 0.242$) shows significant effects on the offset $F_0$. The results, therefore, indicate that the $F_0$ height difference due to voicing lasts until the end of the syllable.

2. Anticipatory effect

a. Effect of syllable boundary. The consonantal perturbation may impact not only the $F_0$ of the following vowel but also the preceding vowel. As shown in Figs. 9(a) and 9(b), $F_0$ contours of vowels preceding the coda consonants in CVC syllables do not converge. In contrast, vowels before the target consonants in CV syllables have very close $F_0$ values (Figs. 7 and 8), which is similar to the first vowels in CVCV syllables where the second consonant is an obstruent, as shown in Figs. 9(c) and 9(d). The means and standard deviations of $F_0$ offset for vowels in CVC syllables, the first vowels in CV and CVCV syllables are listed in Table VI. We performed statistical analysis on the vowel offset $F_0$ with CVOICE (voiced, voiceless), CMANNER (stop, fricative), INTONATION (statement, question), and their interaction as potential fixed effects. In CVC syllables, the main effect of CVOICE ($\chi^2 = 10.018$, df = 1, $p = 0.002$) is significant. The $F_0$ at the vowel offset is higher when preceded by voiceless consonants than by voiced consonants. Neither

<table>
<thead>
<tr>
<th>Consonant type</th>
<th>Statement</th>
<th>Elbow jump</th>
<th>Question</th>
<th>Elbow jump</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiced stop</td>
<td>22(40)</td>
<td>161(42)</td>
<td>18(39)</td>
<td>139(35)</td>
</tr>
<tr>
<td>Voiced fricative</td>
<td>26(40)</td>
<td>161(41)</td>
<td>27(40)</td>
<td>144(41)</td>
</tr>
<tr>
<td>Voiced stop-sonorant</td>
<td>17(38)</td>
<td>167(39)</td>
<td>24(39)</td>
<td>150(45)</td>
</tr>
<tr>
<td>Voiceless stop</td>
<td>21(40)</td>
<td>188(50)</td>
<td>17(37)</td>
<td>157(37)</td>
</tr>
<tr>
<td>Voiceless fricative</td>
<td>21(39)</td>
<td>160(39)</td>
<td>16(40)</td>
<td>144(44)</td>
</tr>
<tr>
<td>Voiceless stop-sonorant</td>
<td>25(38)</td>
<td>184(43)</td>
<td>14(39)</td>
<td>163(43)</td>
</tr>
<tr>
<td>Voiceless affricate</td>
<td>29(38)</td>
<td>196(47)</td>
<td>13(40)</td>
<td>162(41)</td>
</tr>
<tr>
<td>Voiceless consonants</td>
<td>96(155)</td>
<td>183(46)</td>
<td>60(156)</td>
<td>156(41)</td>
</tr>
</tbody>
</table>

FIG. 16. (Color online) Boxplots of $F_0$ elbow (a) and elbow jump (b) separated by consonant voicing and intonation context. See Fig. 10 for definitions of $F_0$ elbow and elbow jump.

TABLE IV. The number of $F_0$ elbow/total available tokens and means (standard deviations) (in Hz) by intonational patterns and consonant types.
CMANNER ($\chi^2 = 1.172$, df = 1, $p = 0.279$) nor INTONATION ($\chi^2 = 1.061$, df = 1, $p = 0.303$) significantly predicts the offset $F_0$. The interaction CMANNER and INTONATION ($\chi^2 = 21.760$, df = 2, $p < 0.001$) is significant: the contrast between stops and fricatives is more pronounced in questions ($p < 0.001$) than in statements ($p = 0.095$). In short, voicing and manner of articulation of coda consonants influence the $F_0$ of vowels right before the closure and the effect interacts with sentence intonation.

When the syllable boundary is not a word boundary, as in the case of offset $F_0$ in the first vowel of the CVCV syllable, the main effects of CMANNER ($\chi^2 = 5.507$, df = 1, $p = 0.019$) and INTONATION ($\chi^2 = 5.905$, df = 1, $p = 0.015$) are significant, while the main effect of CVOICE ($\chi^2 = 0.227$, df = 1, $p = 0.634$) is not. No trace of $F_0$ differences at vowel offset before voiceless and voiced consonants was observed before syllable boundaries.

TABLE V. Means (standard deviations) of offset $F_0$ (Hz) following different types of consonants in declarative and interrogative carrier sentences.

<table>
<thead>
<tr>
<th>Consonant type</th>
<th>Statement</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasal</td>
<td>168(61)</td>
<td>181(51)</td>
</tr>
<tr>
<td>Voiced stop</td>
<td>164(55)</td>
<td>176(48)</td>
</tr>
<tr>
<td>Voiced fricative</td>
<td>169(59)</td>
<td>178(52)</td>
</tr>
<tr>
<td>Voiced stop-sonorant</td>
<td>161(56)</td>
<td>172(46)</td>
</tr>
<tr>
<td>Voiced consonants (excluding nasals)</td>
<td>164(56)</td>
<td>176(47)</td>
</tr>
<tr>
<td>Voiceless stop</td>
<td>168(60)</td>
<td>183(49)</td>
</tr>
<tr>
<td>Voiceless fricative</td>
<td>168(60)</td>
<td>182(52)</td>
</tr>
<tr>
<td>Voiceless stop-sonorant</td>
<td>168(59)</td>
<td>183(53)</td>
</tr>
<tr>
<td>Voiceless affricate</td>
<td>173(62)</td>
<td>184(53)</td>
</tr>
<tr>
<td>Voiceless consonants</td>
<td>169(60)</td>
<td>183(52)</td>
</tr>
</tbody>
</table>

For vowel $F_0$ offset preceding CV syllables, when the syllable boundary between the target consonant and the preceding vowel is also a word boundary, the main effect of CVOICE ($\chi^2 = 0.056$, df = 1, $p = 0.814$), CMANNER ($\chi^2 = 0.728$, df = 2, $p = 0.695$) and INTONATION ($\chi^2 = 0.779$, df = 1, $p = 0.378$) are not significant, and neither are the two-way interactions and three-way interactions. The anticipatory $F_0$ perturbation is also missing here, just like in CVCV syllables. If we combine the findings of offset $F_0$ in vowels before obstruent consonants in the CV, CVC, and CVCV syllables, it seems clear that anticipatory $F_0$ modulation at vowel offset is only present within a syllable.

b. Time course of anticipatory $F_0$ perturbation in CVC syllables. As seen in Figs. 9(a) and 9(b), in CVC syllables, $F_0$ contours vary visibly with different types of coda consonants. The differences are the greatest right before the consonant closure, which then gradually reduce leftward and eventually converge to the nasal baseline. Figure 18 plots the time course of the anticipatory $F_0$ perturbation effect in vowels preceding voiced and voiceless consonants in five in-syllable positions. We can see that $F_0$ is higher preceding voiceless consonants than preceding voiced consonants. The closer to the target consonant, the more prominent the contrast is. To examine the time course of the anticipatory effect, we fitted linear mixed models with TIME (five levels: onset, 1/4, 1/2, 3/4 of the vowel duration, and offset) being incorporated as a potential categorical fixed effect. In addition, CVOICE (voiced, voiceless), CMANNER (stop, fricative, stop-sonorant), INTONATION (statement, question), and their interactions are included as potential fixed effects. Detailed results of the linear mixed models can be found in FIG. 17. (Color online) Histogramic distributions of elbow jump values by consonant type. The upper panel shows distributions of elbow jump for voiced consonants and the lower panel for voiceless consonants. In each plot, the dashed vertical line marks the zero point on the x axis.
TABLE VI. Means (standard deviations) of offset $F_0$ (Hz) of vowels in CVC syllables, first vowels in CVCV syllables before syllable boundaries and first vowels in CV syllables before word boundaries in declarative and interrogative sentences.

<table>
<thead>
<tr>
<th>Consonant type</th>
<th>Statement</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CV</td>
<td>CVCV</td>
</tr>
<tr>
<td>Nasal</td>
<td>152(45)</td>
<td>175(53)</td>
</tr>
<tr>
<td>Voiced stop</td>
<td>152(42)</td>
<td>167(52)</td>
</tr>
<tr>
<td>Voiced fricative</td>
<td>148(43)</td>
<td>162(58)</td>
</tr>
<tr>
<td>Voiced stop-sonorant</td>
<td>151(45)</td>
<td>NA</td>
</tr>
<tr>
<td>Voiced consonants (excluding nasal)</td>
<td>150(43)</td>
<td>164(55)</td>
</tr>
<tr>
<td>Voiceless stop</td>
<td>147(44)</td>
<td>190(59)</td>
</tr>
<tr>
<td>Voiceless fricative</td>
<td>152(46)</td>
<td>182(52)</td>
</tr>
<tr>
<td>Voiceless stop-sonorant</td>
<td>149(42)</td>
<td>NA</td>
</tr>
<tr>
<td>Voiceless affricate</td>
<td>152(47)</td>
<td>NA</td>
</tr>
<tr>
<td>Voiceless consonants</td>
<td>150(44)</td>
<td>186(55)</td>
</tr>
</tbody>
</table>

Appendix A. The interaction between CVOICE and TIME is significant ($\chi^2 = 72.277$, df = 4, $p < 0.001$). Post hoc comparisons show that the difference in the $F_0$ of vowels before voiced and voiceless consonants is significant only at the very end of the syllable ($p < 0.001$), but not at the beginning ($p = 0.995$), 1/4 ($p = 0.990$), 1/2 ($p = 1.000$), or 3/4 ($p = 0.181$) of the vowel duration. Overall, the results indicate that there is an anticipatory $F_0$ perturbation effect that emerges from the very end of the vowel.

IV. DISCUSSION

The present study aims at achieving an accurate assessment of the nature and scope of the consonantal perturbation of $F_0$ by testing a number of methodological measures: (1) applying a nasal baseline as the reference; (2) using syllable-wise time-normalization to align $F_0$ contours in different syllable structures; (3) calculating $F_0$ cycle-by-cycle without smoothing with a large window; and (4) controlling underlying intonation in carriers spoken as either statements or questions. With these methods, we have found evidence that there are two rather different types of perturbations. One is a brief, yet sometimes large, $F_0$ jump at the vowel onset relative to the nasal baseline, and the other is a long-lasting raising or lowering of $F_0$ that persists all the way to the end of the syllable. In addition, we have also observed a brief anticipatory perturbation of $F_0$ before a coda consonant.

A. Large brief perturbations

From Figs. 7(d) to Fig. 8(d), we can see that the initial $F_0$ at vowel onset is in most cases well off the nasal baseline. We measured this initial deviation of $F_0$ in two different ways: onset $F_0$ (absolute $F_0$) and $F_0$ jump (relative to nasal baseline). Statistical results show a significant effect of consonant voicing on both onset $F_0$ and $F_0$ jump, but no effect of manner of consonant articulation. Onset $F_0$ is more variable than $F_0$ jump as a consequence of the impact of the interaction between consonant voicing and sentence intonation (see Fig. 13). The onset $F_0$ values of voiceless consonants are higher in statements than in questions. After this jump, in each case, $F_0$ quickly turns toward a trajectory that shadows the nasal baseline for the rest of the syllable.

Despite the shadowing, in most cases, the long-term trajectories stay away from the nasal baseline, with the general tendency of higher $F_0$ after voiceless consonants and lower $F_0$ after voiced consonants. Thus, the initial jumps seem to be rather different from the longer-lasting effects. Figures 7(d) and 8(d) further show that, surprisingly, $F_0$ jump is much smaller after voiceless stops than after other voiceless obstruents. In Fig. 7(d), after the release of a voiceless stop, $F_0$ even rises up to join the cluster of voiceless trajectories that are elevated well above the nasal baseline (which, as mentioned in Sec. III B 1 a, occurred in four of the eight speakers). This further implies that the initial jump is likely due to a different mechanism from the longer-term effects.

The first possibility is that the initial $F_0$ jump is due to an aerodynamic effect (Ladefoged, 1967). In that hypothesis, the buildup of oral pressure during a voiced stop reduces the pressure drop across the vocal cords, thus decreasing $F_0$ in the following vowel. In a voiceless stop, especially if it is aspirated, the high transglottal airflow at the release creates a boosted Bernoulli force, leading to increased $F_0$ in the following vowel (Hombert et al., 1979). However, the present data show that large $F_0$ jumps occur after the release of both voiced and voiceless obstruents. Moreover, at even greater odds with the aerodynamic hypothesis, voiceless stops show much smaller $F_0$ jumps than the other voiceless obstruents (Table II). This goes against the finding of Löfqvist et al.
(1995) that the level of airflow is greater after a voiceless stop than after a voiced stop.

Another possibility is that much of the $F_0$ jump could be due to a brief falsetto vibration (Xu, 2019). That is, the initial vibration at voice onset after an obstruent may involve only the outer (mucosal) layer of the vocal folds (Titze, 1994), which has a higher natural frequency than the main body of the vocal folds, due to its smaller mass (Miller et al., 2002). At the moment of voice onset, transglottal airflow is going through a sharp drop as the vocal folds are quickly being adducted for voicing. The addition process has to first involve the outer layers of the folds before engaging the main body, and a vibration involving only the outer layer would generate $F_0$ at the falsetto register rather than the chest register (Titze, 1994). Falsetto vibration has been suggested to happen at the end of utterance offsets, where $F_0$ is often observed to jump up abruptly in breach of the on-going downward intonation contour (Xu, 2019). This brief falsetto vibration hypothesis would predict that the level of $F_0$ jump is related to the speed of vocal fold adduction at voice onset, as falsetto vibration is more likely to happen when the adduction speed is relatively slow. This would be the case in voiceless fricatives which likely requires precise control of transglottal airflow. As shown in Table II, voiceless fricatives indeed have the largest $F_0$ jumps in both statements and questions. The brief falsetto vibration hypothesis would also predict that the magnitude of $F_0$ jump can vary positively with boundary strength. We analyzed the $F_0$ following the medial consonant in CVCV syllables (see Appendix B for the descriptive statistics and Appendix C for the results of the linear mixed models). Compared with the initial consonant at the word boundary in CV syllables, the closure duration of the medial consonant is much shorter and the magnitude of $F_0$ jump is also smaller in CVCV syllables.

The brevity of the initial $F_0$ jump makes it tricky to capture in $F_0$ analysis, however, as illustrated in Fig. 19. All the $F_0$ contours in the figure were generated by taking the inverse of every vocal period to obtain the raw $F_0$, and then applying a trimming algorithm (Xu, 1999) to prune very local spikes. They differ only in (a) whether the trimming is applied across silent intervals (edge-trimmed), and (b) whether a smoothing filter is applied after trimming. In Fig. 19(a), trimming was not applied across silent intervals longer than 33 ms (i.e., when $F_0$ would go below 30 Hz). With this method (which was used in the present study), the large $F_0$ jumps (relative to the nasals) as well as the sharp drops are clearly visible. In Fig. 19(b), trimming was again not applied across silent intervals, but a 70-ms triangular filter was applied to smooth the raw $F_0$. As a result, the initial jumps and the following drops are now much smaller. In Fig. 19(c), trimming was applied across silent intervals before smoothing. As can be seen, the large $F_0$ drops have now mostly disappeared, although the $F_0$ jumps are still clearly visible. With the new method, the large initial $F_0$ jumps can be found for all the speakers, despite some differences in magnitude (see supplementary material for by-speaker plots).

The finding of two different kinds of $F_0$ perturbation in the present study may help to explain the low consensus on the rise-fall dichotomy between voiced and voiceless stops in previous studies. Those that do not catch the initial jumps (House and Fairbanks, 1953; Lehiste and Peterson, 1961; Lea, 1973; Hombert et al., 1979) tend to report a simple voicing contrast with $F_0$ following voiceless stops being higher than the voiced stops. When the initial jumps are preserved, the $F_0$ fall after both types of consonants is observed (Ohde, 1984; Silverman, 1984; Hanson, 2009). In our statistical comparison of the initial jump of voiced and voiceless stops, the removal of the abrupt $F_0$ shift with trimming and smoothing led to a statistically significant voicing contrast. When the initial jump was preserved, however, the $F_0$ following voiced and voiceless obstruent consonants was statistically indistinguishable.

The present data also show that the brief perturbation lasts only around 41 ms (SD = 22), after which there is frequently a turning point where the initial perturbation fades away and the $F_0$ of all consonants starts to shadow the nasal baselines. At the $F_0$ turning point ($F_0$ elbow and elbow jump), voiceless consonants show higher absolute $F_0$ than voiced consonants, and the difference is more prominent in statements than in questions [Fig. 16(a)]. When measured in terms of elbow jump, which is relative to the nasal baseline, $F_0$ shows less variance and is not influenced by the sentence intonation [Fig. 16(b)]. Again, similar to the case of onset
$F_0$ versus $F_0$ jump, voicing contrast at the $F_0$ turning point, though large in magnitude, is masked by sentence intonation due to greater variability than elbow jump. The syllablewise alignment with the nasals eliminates the interference of intonation, which leads to higher consistency in $F_0$ jump and elbow jump.

**B. Sustained carryover perturbation**

After the $F_0$ turning point, a smaller upward perturbation is still evident when comparing voiceless consonants with voiced consonants. This effect has a magnitude of around 8 Hz, and it progressively diminishes till the end of the syllable. Furthermore, the distribution of this effect is different from that of the larger initial effect. While the former shows varying magnitudes after different obstruent consonants, the latter shows little differences in magnitude between consonants. This latter effect is consistent with the vocal fold tension mechanism proposed by Halle and Stevens (1971). That is, in a voiceless obstruent the vocal folds are stiffened to impede glottal vibration during the consonant closure, while in a voiced obstruent the vocal folds are slackened to facilitate glottal vibration. Previous studies, however, have not been able to find clear evidence of $F_0$ lowering in English voiced obstruents (Hanson, 2009).

In the present study, we observed an increasing downward perturbation after the initial perturbation. The lowering effect reaches around 13 Hz after stop-sonorants at the $F_0$ elbow. It then gradually declines to 5 Hz after voiced stops and 8 Hz after stop-sonorants compared with nasals at the syllable offset. No such perturbation is found after voiced fricatives. Unlike even the longer-lived upward perturbation, this effect shows no sign of abating for stop-sonorants even at the end of our measurement, which was on average 194 ms from the release of the target consonant. Not only is this consistent with Halle and Stevens (1971) hypothesis that the vocal folds are slackened to maintain voicing during a long oral closure when the transglottal pressure drop is quickly reduced below that of phonation threshold (Berry et al., 1996), but also it is first evidence that the voicing contrast is long lasting.

**C. Anticipatory perturbation by obstruent codas consonants**

As shown in Figs. 9(a) and 9(b), there are also two kinds of $F_0$ perturbations by coda consonants. Right before the closure of an obstruent coda, there is a very brief lowering of $F_0$, which is small in magnitude. Further back in time, there is a much greater perturbation: $F_0$ preceding voiceless coda consonants is higher than voiced coda. The raising effect starts to appear in the midpoint of the vowel toward the coda closure but does not reach statistical significance until the very last measurement point (Fig. 18). The $F_0$ contours in CVCV syllables before the second C and those before CV syllables, however, do not differ from one another. Thus, the anticipatory $F_0$ perturbation does not apply across syllable boundaries.

The anticipatory $F_0$ perturbation by coda consonants should be taken with caution, however, because they are potentially biased by difficulties in the alignment of obstruent and nasal contours. First, we marked the offsets of final obstruents at the resumption of voicing, if there was any voice break. The oral release, which often precedes the resumption of voicing, would be earlier when the coda is voiceless than when it is voiced. Second, there are significant differences in syllable duration due to the well-known pre-consonantal voicing effect in English (House and Fairbanks, 1953; House, 1961), which might have affected the phonetic implementation of the base $F_0$ contours. The average duration of target words is 380 ms with final nasals, 398 ms with final voiced stops, 408 ms with final voiceless stops, 411 ms with final voiced fricatives, and 442 ms with final voiceless fricatives. Since our method of measuring perturbation depends on the alignment of obstruent curves to nasals, errors in the placement of a syllable boundary in the nasal contour would result in misalignment to all corresponding obstruents, which would create gaps between the curves that are not due to actual perturbation but are measured as such. Looking from Figs. 9(a) and 9(b), however, even with adjustments in alignment, $F_0$ before voiceless consonant would still be higher in both statements and questions. Nevertheless, further studies are necessary to fully resolve this issue.

**V. CONCLUSION**

The present study is a further effort to improve the understanding of consonantal perturbation of $F_0$. Recent studies (Hanson, 2009; Kirby and Ladd, 2016; Kirby et al., 2020) have already shown reduced support for the simple rise-fall dichotomy of $F_0$ movement after voiced versus voiceless consonants (Hombert et al., 1979) illustrated in Fig. 1. These studies have demonstrated the importance of using $F_0$ of syllables with sonorant onsets as baseline when assessing the perturbation effect by obstruent consonants. The present study has explored further improvements of methodology by first using the entire syllable as the domain of $F_0$ alignment and time-normalization rather than the conventional alignment of $F_0$ contours at vowel voice onset. Furthermore, we tried to improve the precision of $F_0$ extraction by converting $F_0$ from individual vocal cycles without heavy smoothing. With these methods, we were able to observe, for the first time, three distinct kinds of vertical $F_0$ perturbations. The first is a large but brief raising effect immediately after most of the consonants, which we interpret as likely due to the vibration of only the outer layer of the vocal folds immediately after the consonant release. The second is a longer-sustained increase in $F_0$ both before and after voiceless consonants, which is likely due to an increase in the tension of the vocal folds to inhibit voicing during the voiceless consonant. The third is a sustained downward perturbation after voiced stops and stop-sonorant clusters, which is probably due to the slackening of the vocal folds for the sake of sustaining voicing during the stop closure.
The alignment method used in the present study is based on the assumption that underlying pitch targets associated with a syllable are synchronized with the entire syllable rather than with only the syllable rhyme (Xu and Liu, 2006; Xu, 2020). Based on this assumption, while voice breaks may mask continuous $F_0$ contours, they do not interrupt the underlying laryngeal movements that produce them. The assessment of the vertical $F_0$ perturbation by consonants should therefore treat voice breaks as internal to the syllable. The hypothetical nature of the synchronization assumption, however, means that the findings of the present study are also provisional and open to alternative interpretations.

ACKNOWLEDGMENTS

We would like to thank Andrew Wallace for helping to design the experimental stimuli, conducting the recording, performing the initial data processing, and contributing to an early version of the manuscript. The present work was supported by NIDCD (Grant No. R01 DC03902) and the Leverhulme Trust (RPG-2019-241).

APPENDIX A

Table VII shows statistical results of the anticipatory $F_0$ perturbation in CVC syllables.

APPENDIX B

Table VIII shows means of closure duration, $F_0$ onset, and $F_0$ jump in CVCV syllables.

APPENDIX C

See Table IX for statistical results of $F_0$ jump in CVCV syllables.

Although the same paper also included figures that show $F_0$ contours in syllables with voiced onset stops are similar to those in syllables with sonorant onset; this figure that gives the impression of a robust dichotomy is the most referred to.

1See supplementary material at https://doi.org/10.1121/10.0004239 for individual plots for all participants.

2In Hanson (2009), some of the initial jumps seem to be captured but others are not.


Lea, W. A. (1973). “Segmental and suprasegmental influences on fundamental frequency contours,” in *Consonant Types and Tone*, edited by L. M. Hyman (University of Southern California, Los Angeles, CA), pp. 15–70.


