JASA TUTORIAL

Manipulation of voice onset time in speech stimuli: A tutorial and flexible Praat script

Matthew B. Winn[a)]
Department of Speech-Language-Hearing Sciences, University of Minnesota, 164 Pillsbury Drive Southeast, Minneapolis, Minnesota 55455, USA

Voice onset time (VOT) is an acoustic property of stop consonants that is commonly manipulated in studies of phonetic perception. This paper contains a thorough description of the "progressive cutback and replacement" method of VOT manipulation, and comparison with other VOT manipulation techniques. Other acoustic properties that covary with VOT—such as fundamental frequency and formant transitions—are also discussed, along with considerations for testing VOT perception and its relationship to various other measures of auditory temporal or spectral processing. An implementation of the progressive cutback and replacement method in the Praat scripting language is presented, which is suitable for modifying natural speech for perceptual experiments involving VOT and/or related covarying F0 and intensity cues. Justifications are provided for the stimulus design choices and constraints implemented in the script. https://doi.org/10.1121/10.0000692

(Received 13 August 2019; revised 13 December 2019; accepted 22 January 2020; published online 6 February 2020)

[Editor: James F. Lynch] Pages: 852–866

TABLE OF CONTENTS

I. INTRODUCTION

Voice onset time (VOT) is perhaps the most commonly manipulated acoustic-phonetic speech cue in perceptual experiments. It is a well-documented, simple yet effective distinguisher of phonological voicing that emerges in a large number of languages (Lisker and Abramson, 1964; Cho *et al.*, 2019). VOT is an easily identifiable aspect of the acoustic signal, defined by the time elapsed between the release of stop consonant constriction (the "burst") and the onset of periodicity in the following voiced segment (see Fig. 1). The purpose of this article is to first describe VOT in sufficient detail to enable an experimenter to be informed about relevant acoustic properties, and then to introduce a freely available script to automate the creation of speech sounds that vary by VOT.

There is value in becoming familiar with the commonly used terms surrounding VOT and its accompanying properties in the domains of acoustics and articulation. Phonetic voicing refers to vocal fold vibration, which gives rise to periodicity in the waveform. Phonological voicing is more complex; it is an abstract categorical dimension that *usually* is accompanied by phonetic voicing, but not necessarily so

a)Electronic mail: mwinn@umn.edu

https://doi.org/10.1121/10.0000692

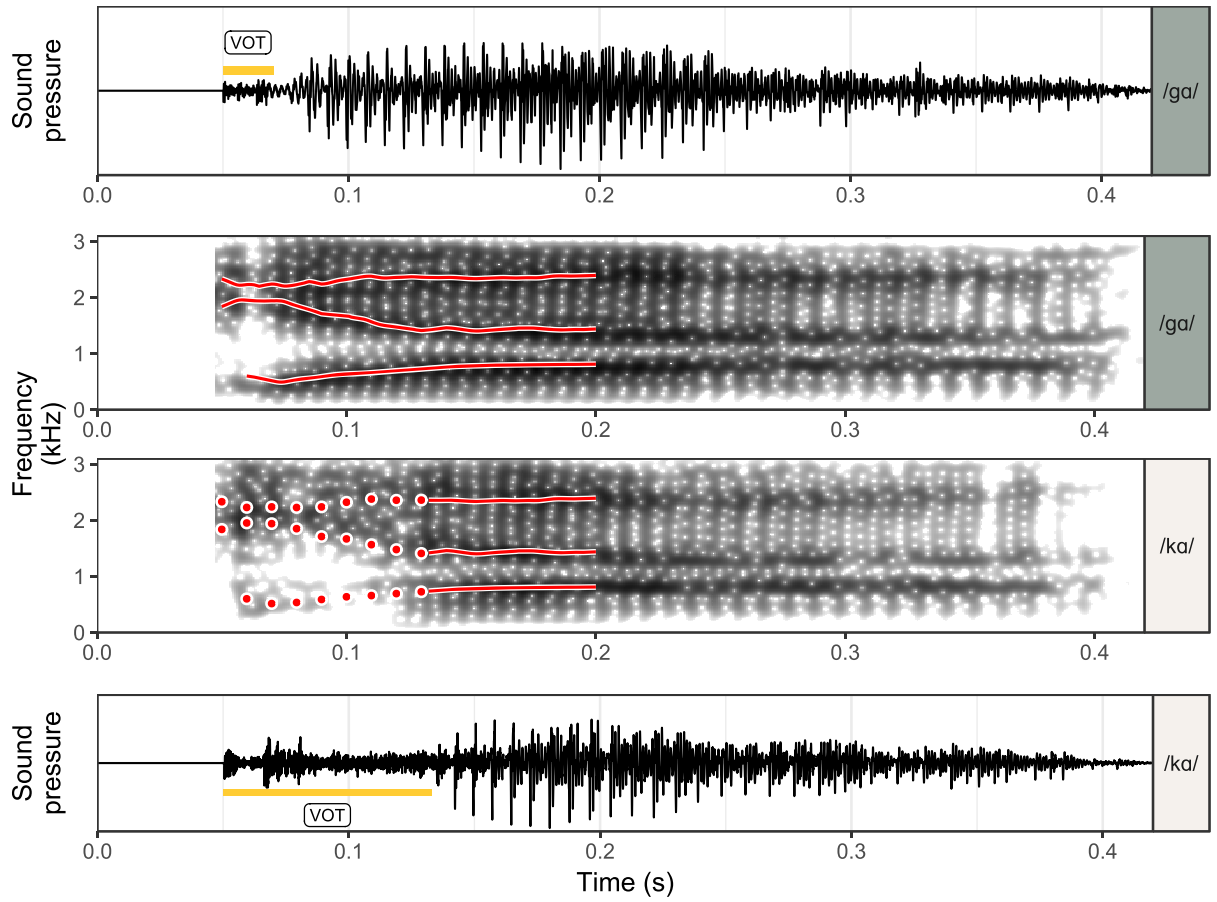

FIG. 1. (Color online) voiced (/gɑ/) and voiceless (/kɑ/) stop consonant waveforms (upper and lower panels) and spectrograms (middle panels). Voice onset time aspiration is marked with a yellow line and "VOT" label. Within the spectrograms, formant contours are indicated with banded lines during voiced sections (including the vowel onset in /gɑ/), and with dots during voiceless sections (for /kɑ/). The formant contours—both from the /gɑ/ onset—were estimated using the Burg algorithm in the Praat software.

in all languages and dialects (e.g., there are other articulatory gestures and acoustic properties that can signify phonological voicing). During the interval between the consonant release burst and the onset of the following vowel, there is aperiodic noise stemming from the burst itself, and perhaps aspiration as well, particularly in the case of long-lag VOT (greater than ≈30 ms). In English long-lag or aspirated stops are typically called "voiceless" stops for simplicity. VOT can be very short, where the consonant release burst is roughly simultaneous with the onset of phonetic voicing/periodicity; such consonants are described as "voiceless unaspirated" or "short-lag VOT" sounds. Alternatively, periodicity could begin *before* the burst; pre-burst periodicity is described as negative VOT and is usually called "prevoicing." Prevoiced stops are not used by all speakers of North American English, but they are common in some dialects and in many other languages of the world. In English, sounds in either the short-lag or prevoicing categories would typically be heard as "voiced." These same phonetic signatures also apply to affricates, where delayed voicing and aspiration would correspond to phonologically voiceless affricates, while short-lag or prevoicing would correspond to voiced affricates. There are also other stop sounds that are not native to North American English such as implosive and ejective stops (or, more generally, *non-pulmonic* stops), where the acoustics are more complex than what will be presented here.

The acoustics, production, and perception of VOT has been studied extensively, including descriptions of variation across languages (Lisker and Abramson, 1964; Cho and Ladefoged, 1999; Cho *et al.*, 2017), studies of second language perception (Flege, 1984), second language production (Flege and Hammond, 1982), individual differences (Allen *et al.*, 2003), phonetic cross-variation (Chodroff and Wilson, 2018), and the influence of the lexicon on perception of VOT (Ganong, 1980). Additionally, other language properties interact with VOT; there is measurable influence of semantics on VOT perception (Schertz and Hawthorne, 2018), and conversely, there is influence of VOT *on* lexical access (Andruski *et al.*, 1994). Furthermore, there are clinically minded studies that use VOT as a diagnostic marker for speech apraxia (Itoh *et al.*, 1982) and which use VOT perception as a proxy for functional auditory temporal processing (Elangovan and Stuart, 2005) among other topics.

Conspicuously missing from the acoustic phonetics literature is a dedicated tutorial on how to properly and efficiently construct stimuli that vary by VOT and other dimensions related to VOT. The current paper describes a "progressive cutback and replacement" approach whereby the onset of a word with a voiced stop sound is progressively deleted and replaced with a roughly equivalent amount of the onset from its voiceless-onset counterpart [cf. McMurray *et al.* (2008) to be discussed in further detail later]. This progressive cutback technique recognizes that the aperiodic segment in voiceless stops (such as English /p t k/) is not *added* (preappended) to the onset of the vowel, but instead is a devoicing of the vowel onset. Figure 1 illustrates this pattern in a clear way by aligning consonant release bursts in syllables with onset voiced-voiceless counterparts (/gɑ/ and /kɑ/). The spectrograms reveal that the spectral properties (formant transitions) in each syllable are most naturally aligned at the consonant release rather than at the onset of voicing.

For the rest of this article, the goals are (1) to comment on acoustic attributes that must be considered when varying VOT and (2) to introduce a freely available script for use with the open-source software Praat (Boersma and Weenink, 2019), to help an experimenter easily make stimuli that vary by VOT and other related dimensions in a controlled systematic way. The discussion of stimulus considerations is aimed at any experimenter who is interested in manipulating VOT or simply understanding VOT more thoroughly.

II. THE ACOUSTICS OF VOT AND RELATED CUES

When creating continua of sounds that vary by perceived voicing, there are numerous factors that can be manipulated, either in a covarying way or in an isolated way. In some cases, there are well established ranged for acoustic attributes that correspond to typical speech production. In other cases, the exact range of values and/or their impact on perception is unclear. In Secs. II A–II F, six different acoustic properties (VOT range, fundamental frequency, vowel duration, formant frequencies, aspiration intensity, and burst spectra) are discussed because they have known or potential relevance for creating VOT continua or perceiving stimuli. Each of these attributes is directly manipulable by the Praat script accompanying this paper, or can be manipulated separately by an experimenter who desires further investigation.

A. VOT ranges

The experimenter will need to consider reasonable and realistic ranges of VOT when planning stimulus creation. Detailed descriptive accounts of speech acoustics are available elsewhere, but in English it is common to work with VOT somewhere in range of 0 to 75 ms. Place of articulation (PoA) has systematic effects on VOT. On the basis of classical descriptions by Lisker and Abramson (1964), it has long been observed that VOTs are shorter for labial stops (e.g., 0–50 ms), intermediate for alveolar stops (e.g., 10–65 ms), and longest for velar stops (e.g., 20–80 ms). However, a more recent cross-linguistic study by Chodroff *et al.* (2019) challenges the long-held notion of this ordering of VOT based on place of articulation, showing a number of occasions where published studies have measured longer VOT for alveolar rather than velar stops. Another helpful description of the relationship between VOT and PoA is provided by Cho and Ladefoged (1999), who additionally discuss places of articulation that are used in languages other than English. In simple single-syllable recognition situations for English, one could expect the perceptual boundary between /b/ and /p/ to fall somewhere around 20–25 ms, the boundary between /d/ and /t/ to fall near 35 ms, and the boundary between /g/ and /k/ to fall near 40 ms.

The exact range of VOT will be different depending on the talker (Allen *et al.*, 2003). It has been thought that women generally have slightly longer VOT than men (Ryalls *et al.*, 1997; Whiteside and Irving, 1998), although Allen *et al.* (2003) challenge this finding. VOT is also affected by vowel context, with longer VOT before high vowels (Klatt, 1975) and at slower speaking rates (Miller and Volaitis, 1989). Figure 2 illustrates distributions of VOT values for stops in English (data from Chodroff and Wilson, 2018). Note how the distributions for short-lag stops (the voiced consonants /b/, /d/, and /g/) are rather compact compared to the widely variable distributions for the aspirated voiceless stops /p/, /t/, and /k/.

Based on the data from Chodroff and Wilson (2018) and Chodroff *et al.* (2019) and previous descriptive work as well as previous perceptual studies, the VOT continuum ranges in Table I are likely suitable for most perceptual experiments focused on English. These values extrapolate slightly from the means of production data to account for the full range of what might be observed in typical listening situations, but which do not include prevoicing.

B. Fundamental frequency

There is a systematic relationship between VOT and the fundamental frequency (F0) at the onset of the following sound. The reader is encouraged to seek out detailed explanations of the roots of this relationship for deeper understanding (e.g., Cho and Ladefoged, 1999; Hanson, 2009), but in short, F0 is higher after voiceless stops than after voiced stops (House and Fairbanks, 1953). The influence of voicing on F0 lasts for approximately 75 to 100 ms following the onset of the vowel (Hombert, 1975; Hanson, 2009). The range of F0 perturbation is dependent on context, and has not been conclusively identified in the literature, with estimates ranging from very modest effects of 10 Hz (House and Fairbanks, 1953) to more considerable effects around 75 Hz (Hanson, 2009). F0 perturbation appears to be proportional rather than absolute, with larger effects of voicing in high-F0 environments (Hanson, 2009). Moreover, in languages where F0 perturbation could interfere with F0-based coding of lexical tone, the effect of consonant voicing on F0 can be suppressed (Francis *et al.*, 2006; Kirby, 2018). So there is clearly some volitional control over F0 that might

https://doi.org/10.1121/10.0000692

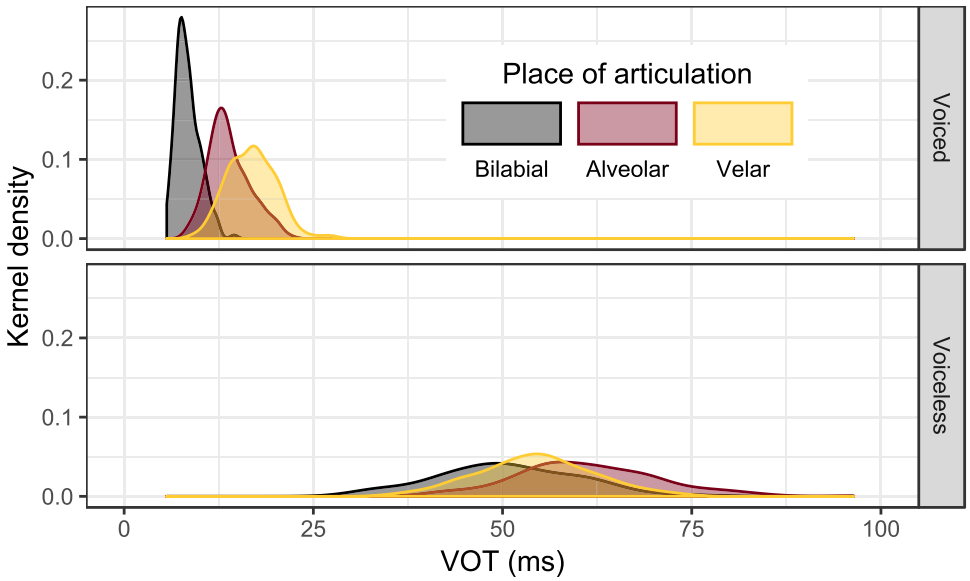

FIG. 2. (Color online) Distributions of voice onset times (VOTs) for stop consonants in English varying by place of articulation (color) and voicing (upper and lower panels). Data from Chodroff and Wilson (2018), and exclude voiced stops with pre-voicing.

relate to the efficiency of information transmission [see Kingston (1991) and Kingston and Diehl (1994) for further discussion of how this cue might be under the talker's control].

Listeners are sensitive to the covariation between VOT and F0. For example, F0 can play a meaningful role in perception of stop consonant voicing perception, especially over ranges of VOT that are ambiguous (Kapnoula *et al.*, 2017). Stop sounds whose F0 conflicts with the VOT (e.g., long VOT with low F0, or short VOT with high F0) are recognized more slowly than stimuli whose VOT and F0 cues are naturally complementary (Whalen *et al.*, 1993). When listening in degraded conditions of limited spectral bandwidth and in the presence of masking noise, F0 can play a very influential role in voicing decisions, to the extent of overriding the influence of VOT (Winn *et al.*, 2013). There is individual variability in the extent of integration of VOT and F0 cues (Kapnoula *et al.*, 2017).

In summary, the experimenter should recognize F0 as a cue that is related to VOT, and make a deliberate decision about whether it should covary naturally within a stimulus set, be controlled to be equalized across all members of a stimulus set, or be manipulated orthogonally across all levels of VOT within a stimulus set (i.e., to make a "matrix" of stimuli that vary in VOT and F0 separately). These options are all available in the Praat script described in Sec. IV.

C. Vowel duration

There tends to be an inverse relationship between VOT and duration of the following vowel (Summerfield, 1981). This relationship is reliable enough that it can be used by listeners in ordinary perception, as well as by computational models of voicing categorization (Toscano and McMurray, 2010). However, the degree of vowel shortening is not completely commensurate with changes in VOT. In other words, for every 1 ms of VOT, the vowel is shortened by less than 1 ms, and this trading relationship is far from consistent (Allen and Miller, 1999; Toscano and McMurray, 2010). Note, however, that the extent of the VOT-vowel duration relationship has been complicated by the fact that the onset of the vowel has been marked at the onset of voicing, whereas earlier in this article it was argued that the period of aspiration between a stop release burst and the onset of voicing could legitimately be described as a devoiced portion of the vowel. The experimenter should recognize the details involved in the trading relationship between VOT and duration of *periodicity* following aspiration.

TABLE I. Example VOT values to use as continuum endpoints.

	voiced	voiceless
/b p/	3 ms	60 ms
/d t/	10 ms	70 ms
/g k/	15 ms	70 ms

D. Formant frequencies

The formant frequencies at the onset of periodicity differ systematically with VOT, and these differences can play a meaningful perceptual role (Stevens and Klatt, 1974; Lisker, 1975; Kluender, 1991; Jiang *et al.*, 2006). Stop consonants have characteristic effects on adjacent vowel formants (due to the consonant gesture's effect on the shape of the vocal tract, and hence on its resonances) which are thought to be essential for the recognition of place of articulation. Given the nature of VOT as vowel-onset devoicing, it follows that by the time the vocal tract is excited by a periodic glottal sound source, much or all of the consonant release gesture will be completed, and hence much or all of the formant transition will have already elapsed. This means that the starting values of the formant frequencies in the voiced part of the vowel could be substantially different depending on VOT.

A stark example of the VOT-formant frequency covariance is observed for consonants in the context of low vowels

like /ɑ/. Figure 3 shows that the formant transition at the onset of /dɑ/ is clearly visible in the periodic portion, with onset formant frequencies at roughly 400, 1800, and 2350, respectively, each transitioning to a different frequency for the vowel portion. In contrast, for /tɑ/, by the time periodicity begins, the formant transitions are nearly complete, with starting frequencies of 700, 1300, and 2200, respectively. The difference of 300 Hz in F1 is not only a large portion of the possible range of F1 values—it also occupies a considerable amount of the perceptual frequency space. Since the cochlea analyzes and encodes sound in a logarithmic fashion, this 400–700 Hz frequency range occupies roughly 3 mm of cochlear space, which is roughly 10% of the tonotopic range of the basilar membrane in an adult human (Greenwood, 1990); for other talkers the frequency range of F1 formant transitions could encompass up to 20% of the basilar membrane's tonotopic range. Figure 3 illustrates how the first formant frequency changes systematically along with VOT in the context of the vowel /ɑ/. For this reason, the /ɑ/ vowel context could be a particularly unfortunate choice for experimenters hoping to isolate auditory processing of a purely temporal nature. Conversely, the /i/ context would be far less affected by VOT cutback, since (1) its formants are more stable across time in general (Hillenbrand *et al.*, 1995) and (2) the F1 of /i/ is already low, meaning that the *upward* F1 transition common to the other vowels would be minimized. F1 for /i/ simply remains at a low frequency regardless of the amount of vowel cutback, thus offering no covarying cue for VOT.

There are some good reasons for experimenters to be mindful of the relationship between VOT and formant frequencies. First, it could be tempting to make conclusions about perception of VOT (/aspiration duration) when a listener could have been perceiving F1 onset frequency instead (Stevens and Klatt, 1974; Kluender, 1991). In some previous studies where the vowel environment /ɑ/ was used (e.g., for /bɑ/-/pɑ/ sounds or /dɑ/-/tɑ/ sounds), it is impossible to disentangle the potential confound of formant cues that accompany changes in VOT. This is especially important in the context of using VOT to assess basic auditory perceptual skills like temporal resolution. Elangovan and Stuart (2005) found a relationship between VOT boundary and psychoacoustic temporal resolution, but Mori *et al.* (2015) found no such association. Steinschneider *et al.* (2003) identified representation of VOT in the brain, but found it to be tonotopically organized, suggesting a frequency-based rather than timing-based representation. It is likely that such results, or equivocal results from earlier studies, result from incomplete appreciation of the spectral covariance that occurs when VOT manipulations are imposed on stimuli that have low vowels.

The effect of formant frequencies could be further enhanced (and the effect of VOT weakened) in challenging situations involving background noise. Since formant frequencies (and any other periodic/harmonic sound) would be more robust to masking than the aperiodic aspiration noise, formants could become a preferred cue (Jiang *et al.*, 2006), particularly in the context of low vowels (Hillenbrand *et al.*,

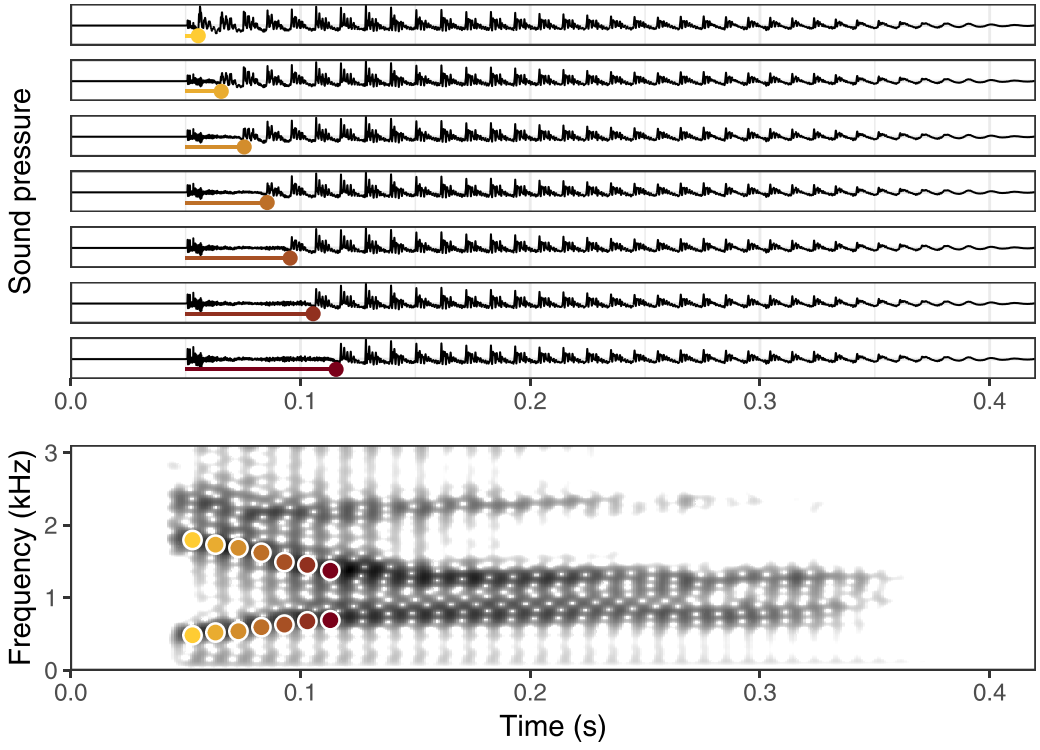

FIG. 3. (Color online) Upper panels: Waveforms showing changes in VOT accompanied by equivalent cutback in vowel onset in a continuum ranging from /dɑ/ to /tɑ/. Lower panel: spectrogram with dots showing changes in first and second formant frequency onsets resulting from each VOT in the upper panels. In this example, VOT and duration of vowel cutback have a 1:1 relationship for the purpose of easily aligning the visualization.

https://doi.org/10.1121/10.0000692

1984). The experimenter who wishes to avoid these complications of confounding F1 cues could consider using the /i/ vowel environment, where the first formant is less strongly perturbed by adjacent consonants, and where the formants tend to be more stable over time than any other English vowel (Hillenbrand *et al.*, 1995), yielding stimuli where F1 remains roughly at the same frequency before and after any de-voicing cutback into the vowel.

E. Aspiration intensity

Consonants with longer VOT tend to have higher subglottal pressure involved in sustaining closure before release of a voiceless stop. It should therefore follow that longer-lag VOT sounds should have aspiration noise that is louder than that of shorter-lag VOT segments. Aspiration intensity has a corresponding effect on perception of VOT, where lower-intensity aspiration lowers the probability of perceiving consonants as voiceless (Repp, 1979; Tamura *et al.*, 2019). VOT as an intensity-mediated cue accords with basic psychoacoustic models of loudness perception, whereby sensitivity to a sound is increased as its duration is increased; if the listener is essentially detecting the presence of aspiration noise, then a longer and quieter segment can be perceptually equivalent to a shorter segment with greater intensity (Hughes, 1946; Garner and Miller, 1947). Thus, the experimenter should be mindful about the decision to equalize sound-pressure intensity or perceived loudness across a range of VOT, or to alter intensity along with VOT in a way that reflects the natural covariation. The exact range of aspiration intensity that should correspond with a range of phonetic voicing has not been clearly established in the literature; experimenters are encouraged to explore this parameter space to determine reasonable ranges for VOT/intensity covariance.

F. Burst spectrum

Stop consonants with longer VOT tend to have release bursts that contain relatively higher amounts of high-frequency energy compared to the bursts of voiced stop consonants. Burst spectrum can therefore be used as a perceptual cue for voicing. Keating (1979) and Nittrouer (1999) found that substituting a voiceless consonant burst at the start of the aspiration in a VOT continuum between /d/ and /t/ was perceptually equivalent to adding 3 ms VOT. However, the effect of spectral shape appears to be complicated; Chodroff and Wilson (2014) found that higher-frequency spectral energy increases perceived category goodness for voiceless labial (/p/) and coronal (/t/) stops, but lower-frequency spectral energy does not improve category goodness for their voiced counterparts. They noted however, that the influence of the burst spectrum cue might have been elevated for the voiceless stops because their VOT values were less prototypical, allowing for greater impact of a secondary cue. Furthermore, they found no effect of burst spectrum for dorsal (i.e., velar) stops.

For the purpose of exploring the impact of burst spectrum on the perception of voicing, an experimenter has numerous choices. Previous work by Keating (1979), Nittrouer (1999), and Chodroff and Wilson (2014) manipulated the burst in a binary fashion, preappending the burst from the voiced or voiceless utterance onto the aspiration segment whose duration is further manipulated to create the VOT continuum. Alternatively, one could imagine gradually blending the bursts voiced and voiceless sounds to create a mixture that covaries with aspiration duration, or direct spectral manipulation such as pre-emphasis or de-emphasis. The exact type spectral manipulation is not clear, as the commonly used metric of spectral center of gravity (COG) [cf. Chodroff and Wilson (2014)] could arise from multiple variations in spectral shape. For example, a lowering of spectral COG could result from a lowering of a frequency peak, or the addition of an additional lower frequency peak. Chodroff and Wilson (2014) have illustrated reliably spectral differences across a collection of stop sounds, but it has not yet been established which aspects of the spectra drive the effect of spectral shape on voicing perception.

III. COMPARISON OF THE PROGRESSIVE-CUTBACK METHOD WITH OTHER VOT MANIPULATION METHODS

Progressive cutback and replacement of the vowel has been argued earlier in this paper to be a more "natural" way of understanding a continuous transition between voiced and voiceless stops. It has been used in a variety of studies using natural speech (Repp, 1979; Miller and Dexter, 1988; McMurray *et al.*, 2008; Winn *et al.*, 2013) and synthetic speech (Ganong, 1980; Whalen *et al.*, 1993; Tamura, 2019). Still, it is feasible to imagine other techniques. Iverson (2003) and Gordon Salant *et al.* (2006) created continua that varied by progressive addition of aspiration at the beginning of a voiced stop (Fig. 4, left column). Alternatively, there could be progressive deletion of aspiration at the onset of a voiceless stop (Fig. 4, center column). Andruski (1994) selectively deleted portions of the aspiration centered on the midpoint. In all of these cases, duration of aspiration (and hence VOT) is manipulated, and listeners are likely able to categorize the results into voiced and voiceless groups. However, there are several reasons to avoid these approaches. Acoustic inspection of voiced/voiceless cognate-pair words reveals that the formant structure remaining in the voiced stop with preappended aspiration (Fig. 4, left column) is not equivalent to that which would arise from a natural voiceless stop. This would essentially create a sequence of aspiration plus a fully intact voiced stop sound, which to the author's knowledge does not occur in any documented language. The compromised stimulus naturalness might cause complications when trying to draw conclusions about typical speech recognition. Specifically, one might expect that VOT boundaries would be at longer aspiration durations to overcome the conflicting cue of lower F1 onset frequency that would be present in an intact voiced-onset stop.

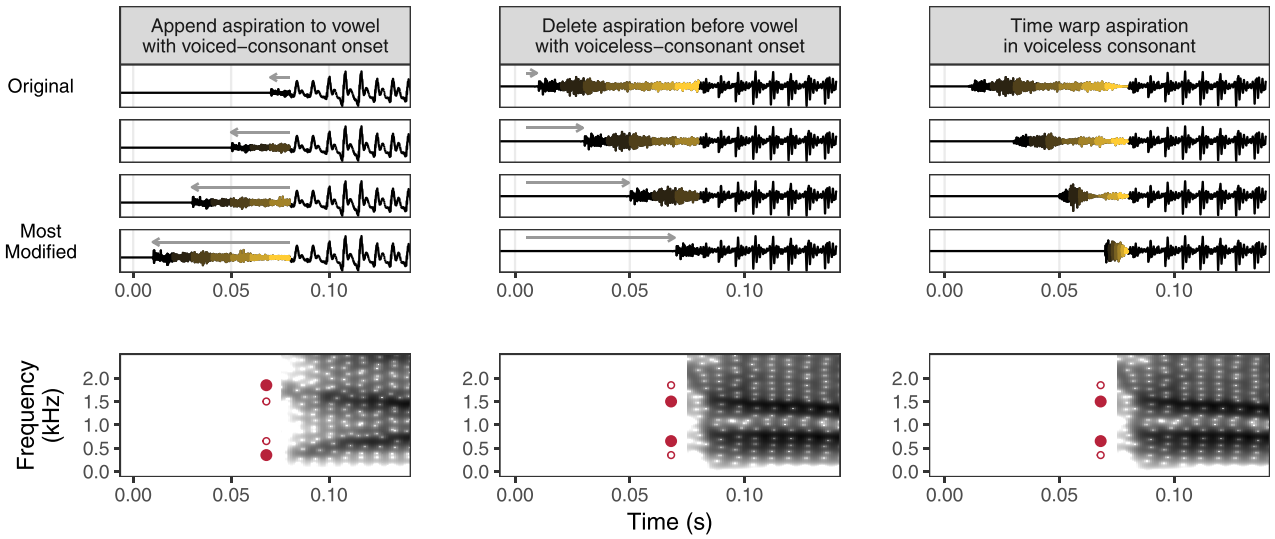

FIG. 4. (Color online) Illustration of three alternative methods for modifying VOT that are not recommended in the current paper. Left: progressively longer amounts of aspiration are added (preappended) to a vowel that was originally spoken after a voiced consonant (i.e., the vowel in /dɑ/). Center: aspiration for a voiceless consonant is incrementally deleted, starting at the boundary between the aspiration and the vowel, moving backwards. Right: the aspiration portion itself is time-compressed to change its duration while maintaining the range of spectral dynamics within the aspiration. Formant onset frequencies of the vowel are marked with filled points, with open points corresponding to formant frequencies for the voiceless endpoint (left column) or voiced endpoint (center and right columns).

The approach of selective *deletion* of the aspiration segment (so that the VOT of a voiceless stop becomes shorter) maintains signatures of the original long VOT, namely, the vowel will have truncated or absent formant transitions, an elevated pitch perturbation, and shortened duration. These are all attributes that will reduce the naturalness of the voiced end of the stimulus continuum. The common thread in the previous discussion has been that VOT is not simply duration of aspiration, but also the accompanying effects that the aspiration has on the surrounding acoustic properties.

In Fig. 4, each spectrogram (lower panels), marks formant onset frequencies of the vowel with filled or open points. In each of these examples, the lack of vowel cutback covariance with VOT results in mismatching acoustic cues at one continuum endpoint, whether inappropriate full formant transitions for long VOT in the left panel, or inappropriate lack of formant transitions for short VOT in the center and right panels. In the left column, the voiced endpoint would sound natural, but the voiceless end of the VOT continuum should have formants at the position of the open points. In the center and right columns, the voiceless endpoint would sound natural, but the voiced end of the VOT continuum should have formants at the position of the open points.

Another alternative approach is the use of *time warping* of the aspiration segment of a voiceless sound for the purpose of shortening the duration of the aspiration so that its duration becomes more like that of a voiced stop [cf. Schoonmaker-Gates (2015) and Schertz and Hawthorne (2018)]. Figure 4 (right panel) illustrates what this technique might look like. As opposed to the aforementioned selective deletion technique, this time-warping approach will at least preserve all of the spectral dynamics inherent in the voiceless stop, such as voiceless formant transitions. However, it still leaves the same limitations as the aspiration-deletion method described above, including mismatching onset formant frequencies, F0 and vowel duration that would be unnatural for a voiced stop. Specifically, because the vowel's formant transitions are complete before voicing begins, the vowel following a voiceless stop lacks the spectral dynamics (particularly the upward trajectory of F1) that would be characteristic of a vowel that would emerge after a voiced stop with short VOT.

While the techniques mentioned in this section are tempting because of their relative simplicity, they could result in one or more perceptually relevant acoustic attributes in a compromised or unnatural state. Specifically, the F1 cue will be mismatched to the VOT in a way that would bias the perception of voicing in the direction of the F1 (toward voiced for low F1, and toward voiceless for high-F1). In cases where the outcome measure is not the exact VOT boundary but relative *shift* in VOT perception, such techniques might not prevent a result that trends in the expected direction. For example, Schertz and Hawthorne (2018) identified effects of sentential context and language experience on VOT boundaries using a progressive addition technique (Fig. 4, left column). In their study, the effects were quantified not in absolute VOT but relative change of VOT boundary across conditions. Therefore, whatever intentional or unintentional effects the stimulus generation technique had on perception could reasonably be thought to be equivalent across all conditions. However, it is difficult to know whether results (however sensible) are affected in some undesirable way by reaction to stimuli that are not reflective of those that might arise in natural speech.

A. Controlling single acoustic cues

There are some situations where the progressive cutback method might not be the best option for an

https://doi.org/10.1121/10.0000692

experimenter. Specifically, if one is interested in the perception of a single isolated acoustic cue, then the covariance should be controlled to reduce confounds. There are solutions for experimenters in this situation. For isolating VOT, one could set the VOT-to-vowel-cutback ratio to zero (as is available in the current script at the end of the file, adjustable by direct coding), so that the vowel remains unchanged throughout the VOT continuum. One can also isolate F0 by imposing the same F0 contour across all members of a VOT continuum, or to manipulate F0 orthogonally, with both options available in the startup window in current script. To isolate the effect of F1, a vowel can be manipulated before VOT adjustment [e.g., using the methods described by Winn and Litovsky (2015)], or separate continua could be made by appending the aspiration onto the vowel extracted from a voiced-onset syllable, and then the vowel from a voiceless-onset syllable (in the current script, this can be done by intentionally selecting a voiceless-onset sound when prompted to select the voiced-onset sound). Cues in the burst and aspiration can be controlled even more directly by preparing a sound in advance of running the script, and choosing the option to use this pre-arranged object as the "VOT segment" that will be adjusted and appended to each continuum step. In each of these cases, the resulting stimuli could serve the needs of an experimenter who is seeking to isolate a particular acoustic attribute, albeit with some intentional manipulation of script options or, in the case of the vowel cutback ratio, the script text itself (easily accessible and clearly marked near the very bottom of the text).

IV. DESCRIPTION OF THE PRAAT SCRIPT

The script walks the user through a series of steps whereby pre-existing sounds (e.g., "deer" and "tier") are used to generate a continuum varying by VOT and other related properties. The user initiates a startup window where parameters are declared. Basic parameters include the range of VOT and the number of continuum steps. These are likely the most essential (and perhaps the *only*) parameters of interest to most readers of this paper. But there are other relevant factors such as starting F0 values for the voiced and voiceless continuum endpoints, whether F0 should change independently or in conjunction with VOT, and the duration (in milliseconds) over which any change in F0 will be imposed on the F0 contour in the syllable. In some cases, the default values of the script are motivated by generally realistic ranges of cues found in natural utterances. In other cases, the values are motivated by best guesses as to what will maintain natural sound quality; the experimenter is able to decide which of these parameters are deserving of extra scrutiny or control. In a case of exploration of a cue without extensive previous literature, the experimenter is encouraged to collect goodness or naturalness ratings along with identification labels.

When the input parameters are entered and the procedure is initiated, the user is prompted to select timing landmarks from existing sounds, and the script automatically generates a continuum of sounds that vary linearly between the user-defined input values, across a number of continuum steps. As the script progresses, useful information is printed to the information window regarding exact values of each varied parameter and a log of the user's selections and timing landmarks. Although the descriptions are written as if the user is working with single-syllable words, the script could in theory work for a consonant voicing contrast at the beginning of a multi-syllable utterance (e.g., "tessellated" and "desolated").

A. Choosing and preparing endpoint sounds

The experimenter should begin with clean noise-free recordings of a pair of words that are minimally contrastive, like deer/tier, big/pig, goat/coat, etc. Each word should be its own isolated Sound object in the Praat Objects list. The script will ensure that each sound has the same sampling frequency. Ideally the words will have intensities that are considered to be perceptually equivalent. If one of the words is dramatically louder than the other, then the respective components of that sound will remain unnaturally loud when combined with the other sound during the sound assembly process. Be cautious to not confuse RMS intensity equalization with equalization of perceived loudness; the inclusion of a lengthy aspiration portion will justifiably reduce overall RMS intensity of a signal, so equalization would result in the unnatural amplification of the syllable with voiceless onset. It is advisable to work with words that were recorded under similar conditions in the same recording session and apply the same amplification/attenuation to both words equally before initiating the VOT manipulation script. Alternatively, the experimenter could consider equalizing RMS intensity based on a steady-state portion of each vowel rather than the entire sound.

Because the manipulation of F0 involves pitch tracking, experimenters are encouraged to use stimuli that do not feature creaky voice (i.e., "glottal fry") at vowel onset, as pitch tracking would be poor, and the intended manipulations of F0 would probably not emerge cleanly. This will be discussed further in Sec. IV C 2.

B. Controlling the script

The script can be opened in Praat directly as a script file, or simply copied and pasted from a text editor into a new script window. Once run, the first thing that appears is a Form window where numerous parameters can be entered (all described below). The script text file itself can also be directly edited so that repeatedly-used parameters can be initiated as defaults.

Selecting stimulus parameters is by far the most time-consuming part of the process, because it should be driven by a thorough review of relevant literature and/or hypothesis-driven decisions. In the following, advice is offered to help with selecting input parameters to the script.

1. VOT steps and range

The user is encouraged to start with a relatively small number of steps (e.g., 6 to 8) upon first run, to make sure

that the range of acoustic parameters produce satisfactory results, before going on to make a large number of stimuli.

2. Method of selecting the aspiration

By default, the script asks the user to select the aspiration portion from the voiceless endpoint stimulus. A second option, to use a separate sound object as the source of the burst and aspiration, is available in a drop-down menu. That option could be useful when the experimenter desires an identical acoustic segment to be used across multiple continua, including those where the aspiration conflicts with the transitional information in the vowel. Alternatively, a modified (filtered, vocoded, distorted, etc.) aspiration could be used with otherwise intact vowel stimuli.

The script can extract the aspiration segment from the voiceless-onset word, or use a separate pre-arranged sound object containing the aspiration. If the aspiration segment provided is shorter than the maximum desired VOT, there are various ways that the script handles the request. If the user requests a VOT that is between 100% and 140% of the aspiration duration, the script will simply extract the last 40% plus 30 ms of the aspiration and duplicate it at the end of the sound object using 30 ms overlapping cross-fade (blend). If the user requests a VOT between 140% and 240% of the provided aspiration duration, the script will use the PSOLA algorithm to lengthen the aspiration sound. This algorithm is reserved for substantially long VOT requests because it sounds less natural than the output of the blending procedure. If the user requests a VOT that is greater than 240% of the provided aspiration duration, the script will exit and report a warning about an inappropriate request.

3. Prevoicing

The script can accommodate requests for pre-voicing, but uses a novel and unvalidated method, so users should inspect the procedure and its output to ensure it meets their personal standards. To create a segment that will serve as prevoicing, the onset of the vowel is extracted, its amplitude envelope is modified to approximate that of a prevoicing murmur, and then it is low-pass filtered and attenuated to sound more like a murmur. The F0 of the prevoicing is controlled to blend smoothly into the vowel onset. For demonstration purposes, this procedure will suffice. Experimenters who require more rigorous control of true prevoicing, are encouraged to simply preappend different amounts of natural prevoicing onto a short-lag VOT speech segment.

4. F0 steps and settings

For experimenters who are interested in manipulating VOT and F0 independently (as in a *cue-weighting* experiment), the user can also set an independent number of F0 steps and a range by which the F0 will change at the onset of voicing. If the number of F0 steps is greater than 1, then there will be a set of sounds that vary by F0 *for each* step of the VOT continuum. For example, a user can have a 7×6 continuum containing seven different VOT values, each created with six different F0 contours.

For experimenters who do not desire *any* meaningful role of F0 in their stimuli, the number of F0 steps can be declared as 1. In this case, it is advisable to set an F0 perturbation that is intermediate to what one might expect for both endpoints. For example, if the typical F0 perturbation is -10 Hz for voiced stops and $+40$ Hz for voiceless stops, the single perturbation could be a standard $+15$ Hz across all members of a continuum, to neutralize F0 perturbation as an informative cue.

The user is encouraged to set magnitudes of F0 perturbation relative to the original pitch contour of the sounds that are being manipulated during the procedure. This should maintain naturalness and coherence with the natural utterances. For example, minimum and maximum values of -5 and $+30$ would result in a continuum where the vowel has an onset F0 of 5 Hz below to 30 Hz above the original F0 starting value. For most use cases, the user is encouraged to leave the input value for "duration of F0 perturbation" at a value between 75 and 100 ms, as it would reflect the typical voicing-related F0 perturbation range (Hombert, 1975).

In special cases, the user might wish to set absolute starting F0 values that are independent of the original F0 contour. The use of this option would be less common—it would be for cases where very specific frequencies carry meaning for an experiment. For example, if an experimenter wishes to have a specific F0 of 100 Hz for the entire stimulus (as is common in some physiological testing paradigms), the absolute value of 100 Hz can be declared, along with a very long perturbation time of 600 ms, which is likely to last an entire syllable. The user who selects this option is encouraged to inspect the starting sound objects for their original F0 contours to see if the desired F0 starting values would be realistic for the voice being manipulated.

5. VOT and F0 covariance

In the case of having changed in both VOT and F0, there are two options to determine the relationship between VOT and F0 changes: they can be manipulated independently or in tandem. The former option results in a matrix of stimuli where each level of VOT and F0 is fully crossed with the other. For experimenters who are interested in natural covariation of VOT and F0, but who do not want the large number of stimuli that result from an orthogonal manipulation (e.g., 7 VOTs $\times$ 6 F0s), an option in the script is included to change both cues in a tandem fashion so that they change together as a single unit. Using this option would essentially create a "voicing" continuum rather than a purely "VOT" continuum. In the case that the user selects this tandem covariance option, the number of VOT steps is taken as the master number of output continuum output steps (overriding the declared number of F0 steps in the startup form). The range of F0 values would be divided by the number of VOT continuum steps; for each change in VOT, there would be an accompanying proportional change in F0.

https://doi.org/10.1121/10.0000692

6. Covariance of aspiration intensity

In light of the aforementioned effect of aspiration intensity on voicing judgments, the user might wish to covary intensity along with VOT. Registered as a Boolean (yes/no) value in the script startup window, the checkbox, when clicked, will attenuate the voiced (short-VOT) endpoint of the continuum by the value entered in the "aspiration intensity range" field, and gradually interpolate this attenuation level across the entire VOT continuum until the full-length (maximum) VOT, which remains unaltered. For example, if the attenuation is set to 6 dB, and there are seven steps in the continuum, the burst/aspiration would be lowered by 6 dB for step 1, by 5 dB for step 2, by 4 dB for step 3, and so on, until no attenuation for step 7. As mentioned previously in Sec. II, the exact range of intensity that should covary with VOT has not been clearly specified in the literature; the current script can serve as a tool to explore this parameter space.

7. Choosing landmarks

The user begins with a minimally contrastive pair of words that vary by the voicing of the initial consonant. For example, "deer" and "tier." Note the identification of landmarks in Fig. 5. It is important to properly mark the onset of the burst and aspiration, as well as the onset of periodicity. Marking the aspiration onset too early will result in systematic shortening of the actual VOT relative to the planned values. Marking the aspiration onset too late will cut off the burst, which could be relevant for perception. Marking the end of the aspiration too late will potentially introduce periodicity (rather than aspiration) into the VOT, effectively shortening it and lengthening the vowel. Marking the end of the aspiration too early will likely not have much significant effect, though it might result in the need to lengthen the extracted aspiration to meet the length demands of the maximum continuum output.

C. Assessing the results and handling pitfalls

When the script is finished, the Praat Objects window will contain the continuum steps. For continua less than ten steps with the "Covary VOT and F0 in tandem" option, the script will automatically concatenate and display the continuum. For longer continua or for orthogonally varied VOT and F0, the user can manually create this output by selecting the desired steps and choosing "Concatenate recoverably" from the "Combine" menu (see Fig. 6). It is simple to listen to the output sounds by highlighting any step or steps and pressing the "tab" key (on a Windows machine) or clicking the gray bar beneath the selection. If the user created a matrix of stimuli that varied orthogonally by VOT and F0, it is slightly more cumbersome to create this visualization, but still possible; on a Windows machine, hold the "control" key while selecting objects that have the same F0 value but vary incrementally by VOT.

1. Unexpected silence

If there is an unexpected silent gap in the output stimuli, it is likely that a landmark has been selected inappropriately, or that a pre-made aspiration segment contained silence at the end of the sound file. Pre-made aspiration segments, if used, will be treated as though the aspiration continues through to the very end of the signal.

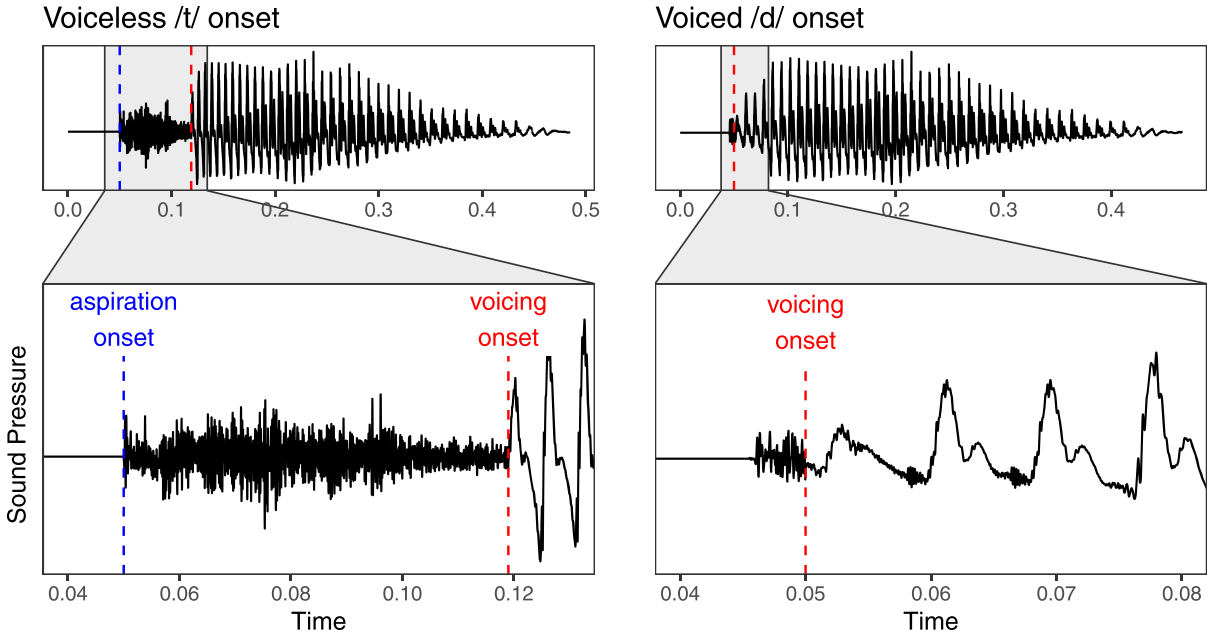

FIG. 5. (Color online) Left: Marking of timing landmarks in voiceless consonants, including burst/aspiration onset at the earliest signature of high-frequency energy (left panel, blue line), and onset of periodicity at the earliest signature of low-frequency periodicity/voicing (left panel, red line). Right: Marking of timing landmarks for the voiced consonant, including the onset of periodicity/voicing for the vowel onset (right panel, red line).

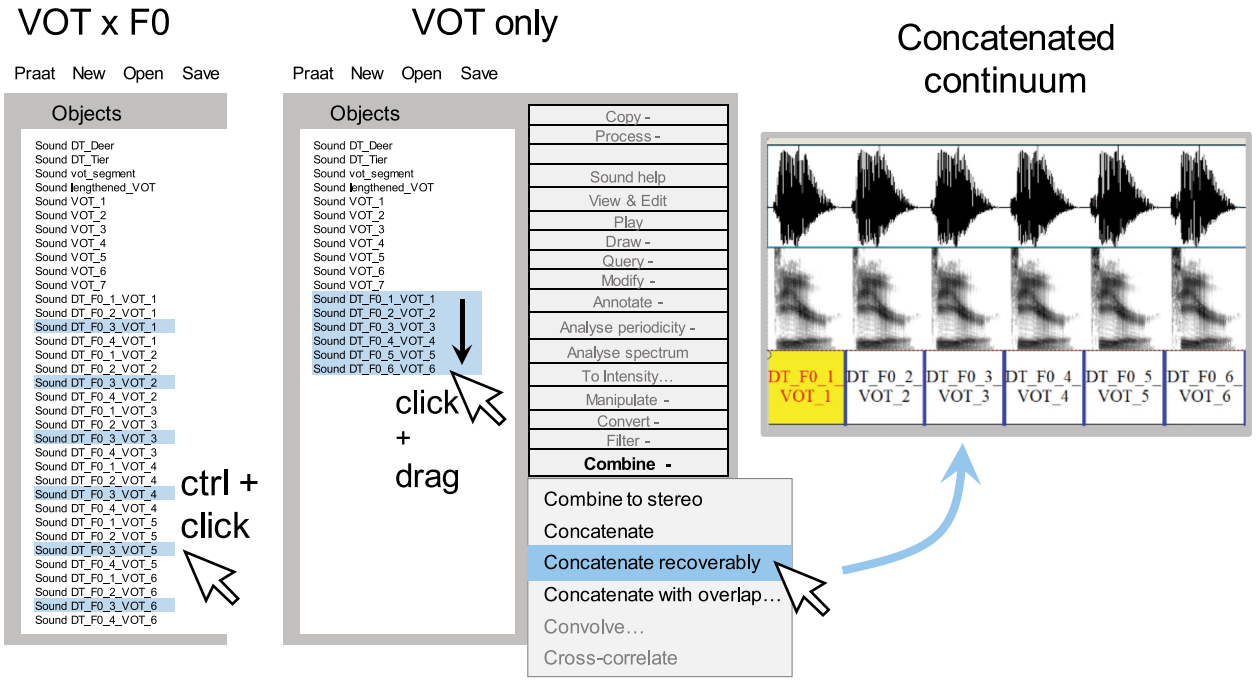

FIG. 6. (Color online) After the script is complete, selection of output objects varying by voice onset time. For VOT × F0 orthogonal (matrix) stimuli, different VOT steps with the same F0 step are selected by holding the "control" key and cling appropriate objects. For VOT continua with no crossed F0 variation, the VOT stimuli are selected by simply clicking and dragging along all of the continuum objects. After the objects are selected, the user clicks "Combine," then "Concatenate recoverably." The user is then able to click "View & Edit" on the resulting pair of objects including a Sound and TextGrid both named "chain." The continuum steps are annotated, enabling the user to listen to individual steps to inspect the output.

2. F0 contour did not change/did not correspond to the input values

In the case that F0 appears to be unchanged, it is likely due to poor pitch tracking of the original sound. To find out if this is the cause, view the original sound object using the Praat editor window by selecting the sound in the Objects window and clicking View & Edit. Pitch pulses are visible in both the Editor window and Manipulation window as vertical blue lines that align with pitch periods in the waveform envelope (see Fig. 7). In some cases, the pitch estimation is poor, or constrained by the user's input settings. Although the standard settings in the script are written to accommodate most voices, there are occasions where tracking breaks down. Sometimes pitch is unsalvageable in cases of creaky voice [Fig. 7(A), right]. There are also occasions where waveform periodicity appears very clearly to the user even if it is not tracked as such by Praat [Fig. 7(A), middle]. In such cases, the minimum and maximum pitch settings might not include the F0 of the voice, causing the F0 estimation to be half or twice the actual F0. In the case that Praat cannot detect any pitch within 25 ms of the user-defined vowel onset, there is an automatic procedure that is invoked where the user is presented with the manipulation object and is offered the opportunity to manually place pitch pulse markers before the script proceeds to the pitch manipulation [Fig. 7(B)].

3. Aspiration is too low or too high in intensity

For aspirations extracted directly from stimuli used for vowel cutback, this would be unusual, and would simply be

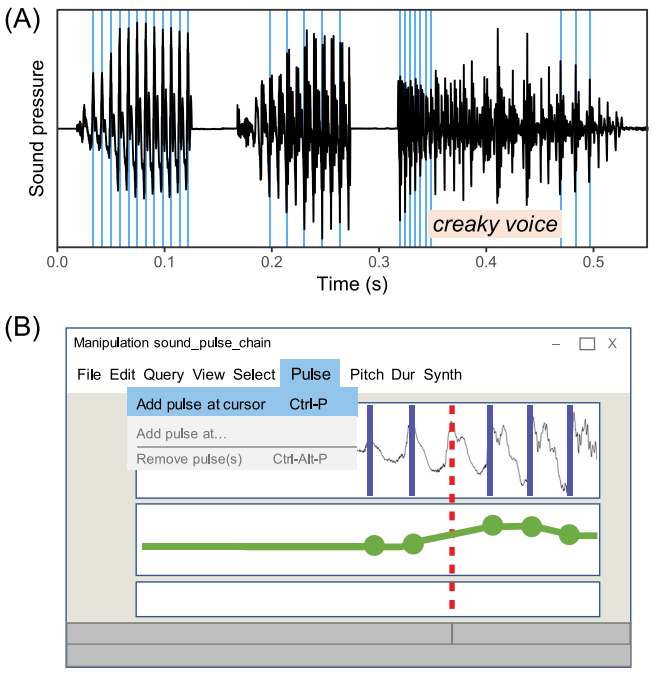

FIG. 7. (Color online) (A) Waveforms of three segments along with blue pulse markers (vertical lines) corresponding to period onsets identified by Praat. The left waveform has clean pitch period tracking. The center waveform has alternating pitch periods near the onset of the vowel that are missed. The right waveform has substantial loss of pitch tracking due to creaky voicing. (B) Simplified Praat manipulation window user interface, featuring an option for the user to add a pitch pulse marker at a selected timepoint in order to preserve capture relevant pitch periods when modifying F0.

https://doi.org/10.1121/10.0000692

representative of the original recordings. It is possible that the voiceless-onset word was not matched to the voiced-onset word in perceived loudness. Alternatively, this situation is more likely to arise in the case of a pre-made aspiration segment whose intensity was unrelated to the vowel used for cutback.

4. Documentation of script parameters at runtime

Each of the script parameters is displayed in the Praat info window upon completion of the script routine. The information includes the version of Praat used to run the script, the time landmarks identified by the user during segmentation, the timing properties of the VOT blending, and the vowel onset landmark. Also reported are the each VOT duration, each F0 onset frequency, the method of changing F0 (e.g., relative to the original pitch contour, or absolute), and the range of F0 analysis (e.g., from 65 Hz to 275 Hz). The info window also indicates the ratio of VOT-to-vowel cutback used in the procedure, as well as the absolute values (and relative values) of the aspiration intensities across the continuum. It also reports on other intensity manipulations such as envelope onset ramping for the vowel onset and the choice of whether RMS was equalized across the continuum or not.

5. Saving the output of the script

The output stimuli and various accompanying documentation file can optionally be saved. The user declares the "parent folder" path (which should already exist on your computer), into which a new sub-folder will be created, named by the "new folder name" field. Included in the output will be all the steps of the VOT(/F0) continuum, the original sounds used for landmark selection, an isolated sound file with just the burst/aspiration, a list of the sound files of the continuum, and a text file with details about the acoustic parameters and user-defined script options. Using these components, the full continuum could be regenerated by another user.

V. EXTRA DETAILS OF THE SCRIPT

Some advanced users might be interested in the details of the procedures operating on behind the scenes. The script includes in-line procedure documentation (denoted by #≫) and comments (denoted with #) in plain language that is intended to explain the rationale for the various stages of sound processing. In addition to the variables declared in the startup window, there are a number of additional variables used throughout the script which are set at the bottom of the text file. These include settings that are unlikely to change on different runs, such as the F0 analysis boundaries, the VOT-to-vowel cutback ratio, aspiration-vowel crossfade duration, naming schemes, etc. The advanced user is encouraged to browse the last procedure in the script, where these variables are declared, to change any values at her/his discretion. The script is intended to remain in perpetuity as a supplemental file to this article[1] and also in current form at Winn (2020).

A. Ordering of operations

The ordering of operations has some impact on the quality of stimulus generation. The script first matches the sampling frequencies of the user-selected minimal pair sounds, and then proceeds to calculate all of the continuum values across all varying parameters. The vowel is extracted from the voiced-onset sound and F0 is tracked at the onset. If no F0 is identified, the time of the F0 query is advanced until a stable F0 can be recovered. It is from this timepoint that the onset of F0 perturbation begins.

Figure 8 illustrates the sequence of steps done to modify the vowel before it is appended to the aspiration. The vowel from the voiced-onset syllable is progressively cut back from the timepoint identified by the user [Fig. 8(B)] with the caveat that the script will not cut away more than a user-defined maximum value, whose variable name is "max_cutback_dur." This cut-back segment is then padded with silence, since F0 tracking can be unreliable at the extreme leading and trailing edges of a sound. The sound is converted to a manipulation object, which enables dynamic manipulation of F0 and duration [Fig. 8(C)]. The onset F0 contour is eliminated from the onset F0-tracked timepoint through the duration of the F0 perturbation. At the onset, the new F0 onset value is superimposed on the pitch tier, which is then interpolated over the F0 perturbation time range using PSOLA.

After the F0 is modified, the padded silence is removed, and the intensity contour of the vowel onset is shaped so that it does not have a rapid onset that would cause unnaturalness [Fig. 8(D)]. The intensity contour begins at *half* the original intensity, and progressively rises to the full intensity at the end of the ramp time. The duration of this intensity contour ramp is half the cutback duration, resulting in a progressively smooth vowel onset for sounds with longer VOT (and conversely a more rapid onset for short-VOT sounds), consistent with observations of naturally spoken utterances. This procedure of intensity modification can be disabled by setting the variable "shape_vowel_envelope_onset" to 0 instead of 1, near the end of the script.

Following the (optional) modification of the intensity contour of the vowel, the aspiration is concatenated with the vowel with a specified amount of overlap-blending time [Fig. 8(E)]. This method is preferable over direct concatenation for two reasons. First, it frees the process from the need to concatenate only at zero crossings, which would place limitations on the minimal change in VOT that could be included in a continuum. Second, it creates a realistic output where the distinction between voiceless (aperiodic) and voiced (periodic) is more of a short transition rather than a discrete change.

The short but smooth transition from aspiration to vowel is accomplished using the "concatenate with overlap" function in Praat. This function is essentially a cross-fading

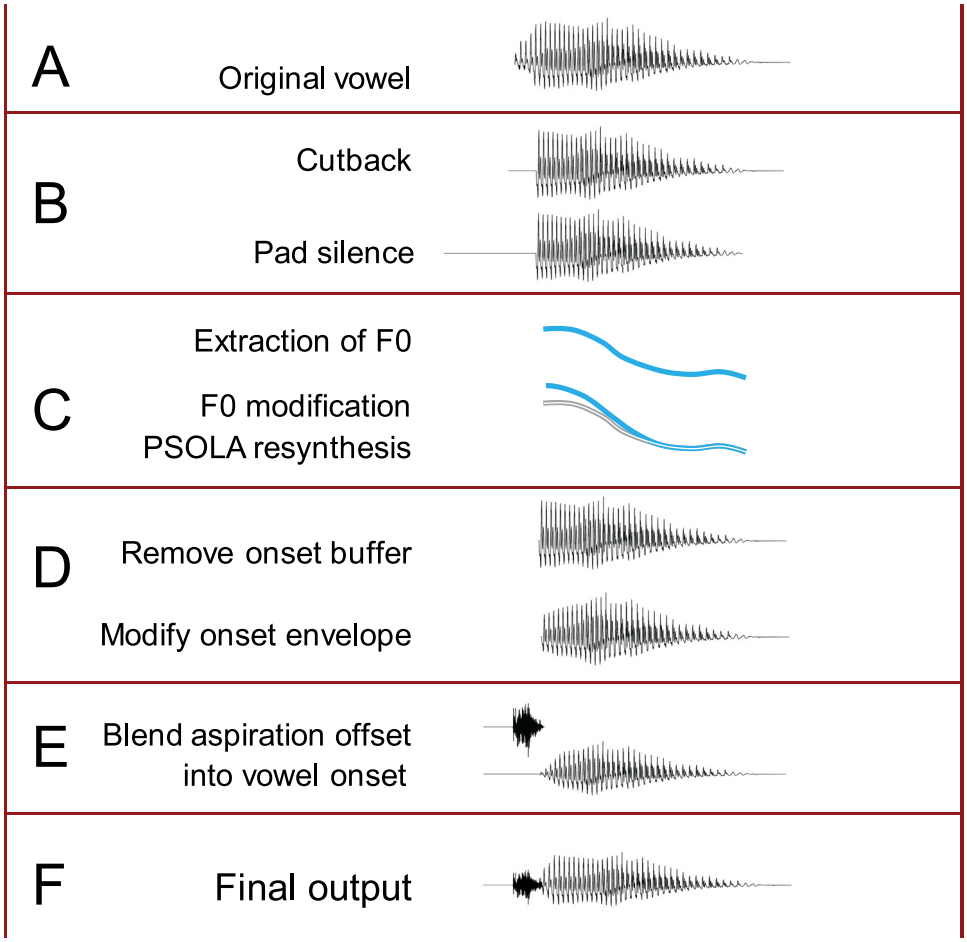

FIG. 8. (Color online) Sequence of integral processing steps involved in the script, including progressive vowel cutback, manipulation of F0 contour, intensity envelope modification, and combination of the aspiration and cutback vowel.

procedure where an overlap time of X seconds results in the final X/2 s of the onset sound tapering from full to zero amplitude, and the initial X/2 s of the following sound tapering from zero to full amplitude. These modified portions are added sample-by-sample so that they smoothly blend together without risk of introducing transient clicks at the boundary. In the script, this overlap blend time is declared in the variable "vot_crossfade_duration" near the end of the script, and can be modified by the user before initiating the script, if desires. The extracted duration of aspiration compensates for (half of) the blending time so that the output VOT should be exactly as the user intends. Lengthening this value would ensure a very smooth transition, and shortening the value would lock in extreme precision of aspiration/vowel boundary and risk perceptible discontinuity. Even with a blend time of 10 ms, the boundary is still identifiable. With a default value of 6 ms, there is sufficient blending so that there is not a perceptibly noticeable disjointed splicing of aspiration and vowel, but not so much that the estimation of VOT would be compromised by an unclear end of aspiration or onset of voicing.

After the component pieces are blended/concatenated [Fig. 8(F)], the resulting sound objects are renamed using a uniform scheme that should facilitate easy organization. The stimulus prefix (declared in the startup window) starts the object name. If the prefix is "DT," then the stimuli will be named "DT_F0_1_VOT_1" for the first step of F0 and the first step of VOT. In the case of continua that have more than nine steps, leading zeros are preappended so that the name would become "DT_F0_01_VOT_01."

B. Implementation rationale

The script described in this paper is implemented using the Praat software (Boersma and Weenink, 2019), which is an open-source freely available program. Though Praat can be limited in some respects compared to other general-purpose programming languages, the rationale is that (1) it is freely available, removing any financial barriers that might exist for other signal processing programs, (2) Praat is already commonly used in phonetics—particularly acoustic phonetics—and is also a flexible tool for many other basic audio manipulations including speech analysis and synthesis, and (3) PRAAT provides quick and responsive visualizations of sound waveforms and spectrograms, with easily accessible spectral analyses and queries for sound intensity, duration, and other properties relevant for speech specifically (e.g., formants). It is therefore an extremely valuable tool that is well suited for experimenters who work with speech sounds. The scripting language, while limited in some ways, uses functions that tend to be transparently named, and the "paste history" capability of PRAAT allows new scripts to be built up through the process of manually

executing commands through the GUI and then pasting the history of those commands in the scripting window. Therefore, the script is extensible for those users who wish to add other procedures in line with those already included in the current script.

VI. SUMMARY AND CONCLUSIONS

It is feasible to automate the creation of speech stimuli that vary by voice onset time (VOT) using natural speech, with a PRAAT script and modest background knowledge of covarying acoustic cues such as voice pitch (F0), first-formant (F1) transition, and aspiration intensity. This paper describes and justifies a technique for VOT continuum synthesis that is designed to make natural-sounding speech stimuli while also offering control over other factors relating to VOT/stop consonant voicing. The user begins with recordings of two isolated words that vary by voicing of an initial stop consonant, and should make deliberate decisions about other properties that might or might not be desirable to covary, such as F0 and aspiration intensity. The script outputs information that enables reproducibility by documenting the input parameters and user selections that led to the output sounds, as well as saving the original input sounds. The script is written for an open-source platform (PRAAT) that is already commonly used by many phoneticians and other speech scientists. The script is set up with reasonable defaults and options to help the novice user easily create a continuum without cumbersome preparation.

Standardization of a VOT creation process can aid in teaching demonstrations, the comparison of results across studies and encourage focused attention on acoustic factors that could be potentially important in numerous studies ranging from phonetic perception to basic auditory science.

ACKNOWLEDGMENTS

Daniel R. McCloy provided valuable input on earlier versions of this manuscript and the PRAAT script.

[1]See supplementary material at https://doi.org/10.1121/10.0000692 for the Praat script, example of output documentation, demonstration sounds, and an example of final stimulus output from the script.

Allen, J., and Miller, J. (**1999**). "Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words," J. Acoust. Soc. Am. **106**, 2031–2039.
Allen, J., Miller, J., and DeSteno, D. (**2003**). "Individual talker differences in voice-onset-time," J. Acoust. Soc. Am. **113**, 544–552.
Andruski, J., Blumstein, S., and Burton, M. (**1994**). "The effect of subphonetic differences on lexical access," Cognition **52**, 163–187.
Boersma, P., and Weenink, D. (**2019**). "Praat: Doing phonetics by computer" [computer program], version 6.0.56, http://www.praat.org/ (Last viewed June 25, 2019).
Cho, T., and Ladefoged, P. (**1999**). "Variation and universals in VOT: Evidence from 18 languages," J. Phonetics **27**, 207–229.
Cho, T., Whalen, D., and Docherty, G. (**2019**). "Voice onset time and beyond: Exploring laryngeal contrast in 19 languages," J. Phonetics **72**, 52–65.
Chodroff, E., Golden, A., and Wilson, C. (**2019**). "Covariation of stop voice onset time across languages: Evidence for a universal constraint on phonetic realization," J. Acoust. Soc. Am. **145**, EL109.
Chodroff, E., and Wilson, C. (**2014**). "Burst spectrum as a cue for the stop voicing contrast in American English," J. Acoust. Soc. Am. **136**, 2762–2772.
Chodroff, E., and Wilson, C. (**2018**). "Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice," in *Linguistics Vanguard* Vol. **4**, s2.
Elangovan, S., and Stuart, A. (**2005**). "Interactive effects of high-pass filtering and masking noise on word recognition," Ann. Otol. Rhinol. Laryngol. **114**, 867–878.
Flege, J. (**1984**). "The detection of French accent by American listeners," J. Acoust. Soc. Am. **76**, 692–707.
Flege, J., and Hammond, R. (**1982**). "Mimicry of non-distinctive phonetic differences between language varieties," Stud. Second Lang. Acq. **5**, 1–18.
Francis, A., Ciocca, V., Wong, V., and Chan, J. (**2006**). "Is fundamental frequency a cue to aspiration in initial stops?," J. Acoust. Soc. Am. **120**, 2884–2895.
Ganong, W. (**1980**). "Phonetic categorization in auditory word perception," J. Exp. Psych.: Hum. Percept. Perf. **6**, 110–115.
Garner, W., and Miller, G. (**1947**). "The masked threshold of pure tones as a function of duration," J. Exp. Psych. **37**, 293–303.
Gordon-Salant, S., Yeni-Komshian, G., Fitzgibbons, P., and Barrett, J. (**2006**). "Age-related differences in identification and discrimination of temporal cues in speech segments," J. Acoust. Soc. Am. **119**, 2455–2466.
Greenwood, D. (**1990**). "A cochlear frequency-position for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.
Hanson, H. (**2009**). "Effects of obstruent consonants on fundamental frequency at vowel onset in English," J. Acoust. Soc. Am. **125**, 425–441.
Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.
Hillenbrand, J., Ingrisano, D., Smith, B., and Flege, J. (**1984**). "Perception of the voiced–voiceless contrast in syllable-final stops," J. Acoust. Soc. Am. **76**, 18–26.
Hombert, J. (**1975**). "Towards a theory of tonogenesis: An empirical, physiologically and perceptually-based account of the development of tonal contrasts in language," doctoral dissertation, University of California, Berkeley, CA.
House, A., and Fairbanks, G. (**1953**). "The influence of consonant environment upon the secondary acoustical characteristics of vowels," J. Acoust. Soc. Am. **25**, 105–113.
Hughes, J. (**1946**). "The threshold of audition for short periods of stimulation," Philos. Trans. R. Soc. B **133**, 486–490.
Itoh, M., Sasanuma, S., Tatsumi, I. F., Murakami, S., Fukusako, Y., and Suzuki, T. (**1982**). "Voice onset time characteristics in apraxia of speech," Brain Lang. **17**, 193–210.
Iverson, P. (**2003**). "Evaluating the function of phonetic perceptual phenomena within speech recognition: An examination of the perception of /d/-/t/ by adult cochlear implant users," J. Acoust. Soc. Am. **113**, 1056–1064.
Jiang, J., Chen, M., and Alwan, A. (**2006**). "On the perception of voicing in syllable-initial plosives in noise," J. Acoust. Soc. Am. **119**, 1092–1105.
Kapnoula, E., Winn, M., Kong, E. J., Edwards, J., and McMurray, B. (**2017**). "Evaluating the sources and functions of gradiency in phoneme categorization: An individual differences approach," J. Exp. Psych.: Hum. Perc. Perf. **43**, 1594–1611.
Keating, P. A. (**1979**). "A phonetic study of a voicing contrast in Polish," Ph.D. dissertation, Brown University, Providence, RI.
Kingston, J. (**1991**). "Integrating articulations in the perception of vowel height," Phonetica **48**, 149–179.
Kingston, J., and Diehl, R. (**1994**). "Phonetic knowledge," Language **70**, 419–454.
Kirby, J. (**2018**). "Onset pitch perturbations and the cross-linguistic implementation of voicing: Evidence from tonal and non-tonal languages," J. Phonetics **71**, 326–354.
Klatt, D. (**1975**). "Voice onset time, frication, and aspiration in word-initial consonant clusters," J. Speech Hear. Res. **18**, 686–706.
Kluender, K. (**1991**). "Effects of first formant onset properties on voicing judgments result from processes not specific to humans," J. Acoust. Soc. Am. **90**, 83–96.
Lisker, L. (**1975**). "Is it VOT or a first-formant transition detector?," J. Acoust. Soc. Am. **57**, 1547–1551.

Lisker, L., and Abramson, A. (**1964**). "A cross-language study of voicing in stops: Acoustical measurements," Word **20**, 384–422.
McMurray, B., Aslin, R., Tanenhaus, M., Spivey, M., and Subik, D. (**2008**). "Gradient sensitivity to within-category variation in words and syllables," J. Exp. Psych.: Hum. Perc. Perf. **34**, 1609–1631.
Miller, J., and Dexter, E. (**1988**). "Effects of speaking rate and lexical status on phonetic perception," J. Exp. Psych.: Human Perc. Perf. **14**, 369–378.
Miller, J., and Volaitis, L. (**1989**). "Effect of speaking rate on the perceptual structure of a phonetic category," Percept. Psychophys. **46**, 505–512.
Mori, S., Oyama, K., Kikuchi, Y., Mitsudo, T., and Hirose, N. (**2015**). "Between-frequency and between-ear gap detections and their relation to perception of stop consonants," Ear Hear. **36**, 464–470.
Nittrouer, S. (**1999**). "Do temporal processing deficits cause phonological processing problems?," J. Speech Lang. Hear. Res. **42**, 925–942.
Repp, B. (**1979**). "Relative amplitude of aspiration noise as a voicing cue for syllable-initial stop consonants," Lang Speech **22**, 173–189.
Ryalls, J., Zipprer, A., and Baldauff, P. (**1997**). "A preliminary investigation of the effects of gender and race on voice onset time," J. Speech Lang. Hear. Res. **40**, 642–645.
Schertz, J., and Hawthorne, K. (**2018**). "The effect of sentential context on phonetic categorization is modulated by talker accent and exposure," J. Acoust. Soc. Am. **143**, EL231–EL236.
Schoonmaker-Gates, E. (**2015**). "On voice-onset time as a cue to foreign accent in Spanish: Native and nonnative perceptions," Hispania **98**, 779–791.
Steinschneider, M., Fishman, Y., and Arezzo, J. (**2003**). "Representation of the voice onset time (VOT) speech parameter in population responses within primary auditory cortex of the awake monkey," J. Acoust. Soc. Am. **114**, 307–321.
Stevens, S., and Klatt, D. (**1974**). "Role of formant transitions in the voiced-voiceless distinction for stops," J. Acoust. Soc. Am. **55**, 653–659.
Summerfield, Q. (**1981**). "Articulatory rate and perceptual constancy in phonetic perception," J. Exp. Psychol.: Hum. Percep. Perf. **7**(5), 1074–1095.
Tamura, S., Ito, K., Hirose, N., and Mori, S. (**2019**). "Effects of manipulating the amplitude of consonant noise portion on subcortical representation of voice onset time and voicing perception in stop consonants," J. Speech Lang. Hear. Res. **62**, 434–441.
Toscano, J., and McMurray, B. (**2010**). "Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics," Cogn. Sci. **34**, 434–464.
Whalen, D., Abramson, A., Lisker, L., and Mody, M. (**1993**). "F0 gives voicing information even with unambiguous voice onset times," J. Acoust. Soc. Am. **93**, 2152–2159.
Whiteside, S., and Irving, C. (**1998**). "Speakers' sex differences in voice onset time: A study of isolated word production," Percept. Mot. Skills **86**, 651–654.
Winn, M. (**2020**). "Praat script to manipulate VOT in natural speech" https://github.com/ListenLab/VOT (Last viewed January 31, 2020).
Winn, M., Chatterjee, M., and Idsardi, W. (**2013**). "Roles of voice onset time and F0 in stop consonant voicing perception: Effects of masking noise and low-pass filtering," J. Speech Lang. Hear. Res. **56**, 1097–1107.
Winn, M., and Litovsky, R. (**2015**). "Using speech sounds to test functional spectral resolution in listeners with cochlear implants," J. Acoust. Soc. Am. **137**, 1430–1442.