Obstruent voicing perception cues

Blake Rodgers
University of Wisconsin-Madison

1 Introduction

This study is an investigation into the features that influence coda obstruent perception in American English. The goal is to determine the relative importance of various acoustic cues in determining whether a listener perceives a voiced stop /d/ or voiceless stop /t/ in coda position. Vowel duration, consonant duration, and percent voicing during the hold of the stop (percent glottal pulsing) were systematically varied for a set of tokens. These tokens, or sample words, were then presented to a group of subjects to determine the effects of changing these three parameters on coda voicing perception. Vowel duration is expected to be the strongest factor of the three with longer durations shifting perception to a /d/ coda.

There is no clear picture from the literature of the relative strengths of the other two factors in this context, but more glottal pulsing and longer consonant are expected to bias perception in the direction of voicing. Natural tokens derived from actual speech were modified as necessary to achieve the desired characteristics. Another goal of the study is to determine the role of the burst in obstruent perception. To that end, two sets of tokens were produced that were identical to each other with the exception of the character of the stop release or burst. The two release types are 'd-burst' and 't-burst'. The expectation is that the 'd-burst' will bias the results in the direction of a /d/ perception, and 't-burst' will bias the results in the direction of a /t/ perception.

2 Background

The tokens used in obstruent perception studies can be broadly categorized as either (1) pure tone (sinusoid), (2) synthesized speech generated by Praat

LSO Working Papers in Linguistics 7: Proceedings of WIGL 2007, 135-143.

(Boersma & Weenink 2006), or other software, or (3) natural speech that has been manipulated. While investigating the role of F1 during transition and vowel duration, Nittrouer (2004) found that using natural tokens versus synthesized speech tokens led to very different results. She concluded that the natural tokens had some factors that were not simulated in the synthesized tokens. These factors will be referred to as hidden in the sense that they are neither easily measurable nor easily synthesizable. The decision was made in the present study to use natural tokens to capture as many of these hidden factors as possible and then vary vowel duration, percent glottal pulsing and consonant duration while keeping these hidden factors constant. Since the starting token in this study was 'bed', the set of cues for a given token can be thought of as (a) naturally occurring 'bed' features for that particular speaker, (b) vowel duration as specified, (c) percent glottal pulsing as specified, (d) consonant duration as specified, and (e) character of the release.

Many perception studies completely remove the burst from the tokens (Nittrouer 2004 and Thomas 2000, for example). While generating the tokens for this study, it became apparent that the release for a /t/ is quite different than that for a /d/. It seemed likely that the character of the stop release burst could change the perceptual outcome. For that reason, the decision was made to investigate the effect of the different releases.

Vowel duration, percent glottal pulsing and consonant duration were chosen because of their consistent high ranking in importance as acoustic cues in the perception literature (Moreton 2004, Nittrouer 2004, Thomas 2000 among many others). Goals of this study include determining a relative ranking of importance for these acoustic cues and also investigating the influence of the release burst.

3 Methodology

3.1 Subjects

The seven subjects who participated in this study were made up of three linguistics graduate students, two linguistics professors, and two undergraduate students, one of which had linguistics experience. All subjects participated on a voluntary basis, and all were native speakers of English.

3.2 Tokens

The 128 tokens used in this pilot study were derived from a single sample. The original sample was a female speaker from Watertown, Wisconsin saying the sentence "Please say bed for me". The word 'bed' was extracted from the carrier

sentence and then manipulated to produce a set of 64 tokens (4 vowel durations * 4 percent glottal pulsing durations * 4 consonant durations).

Table 1: Token Parameters

Vowel duration	80 msec	120 msec	160 msec	200 msec
Glottal pulsing	0 %	33%	67%	100%
Consonant duration	50 msec	83 msec	117 msec	150 msec

Praat software was used for all token manipulation. The base token was marked for start of vowel, end of vowel and end of consonant. The manipulation tier in Praat was used to manipulate the vowel and consonant duration to achieve the desired combinations. 'Bed' was used as a starting token because the final consonant in the original sample was fully voiced (100% glottal pulsing), and it was easiest to manipulate this final consonant to achieve the percent pulsing desired. The desired glottal pulsing percentage was achieved by zeroing the latter part of the consonant as necessary, being careful to start and end the window on a zero crossing. A 33% glottal pulsing ratio, for example, would be achieved with the first third of the consonant voiced normally, and the latter two thirds zeroed.

The values for these parameters were chosen based on observed variation in collected 'bed' and 'bet' tokens taken from the Wisconsin English Project (WEP) (Purnell, et. al., 2005). For vowel duration and consonant duration, the center of the range was deemed to be in the center of the observed variation, and the low and high limits were set somewhat beyond actual observed characteristics. This was done in an attempt to cover the full range of observed variation plus buffer zones at the low and high end. Percent glottal pulsing was varied from minimum to maximum (0% - 100%) because the full range was seen in the recorded samples from the WEP.

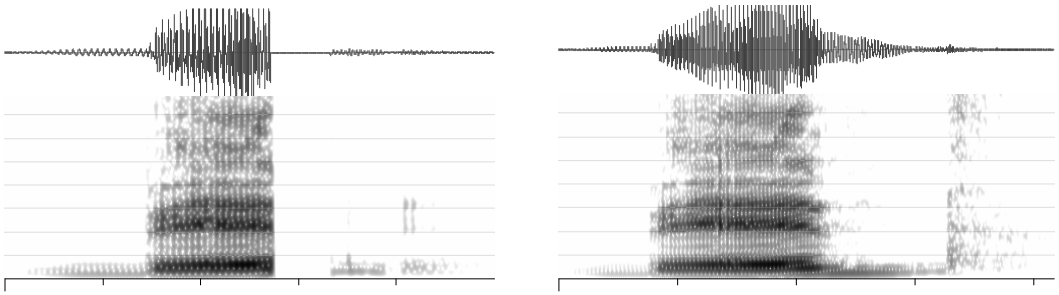

Fig. 1 Shortest token with 0% glottal pulsing and 'd-burst'; Longest token with 100% glottal pulsing and 't-burst'. Diagram produced in TF32 computer program (Milenkovic 2006).

A second starting token was created by removing the burst of the final obstruent (zeroing the 'd-burst' after release) and replacing it with a 't-burst' from

another speaker. This ‘t-burst’ was chosen because it showed distinctive aspiration. Another set of 64 tokens using the methodology described above was generated from this modified starting token for a total of 128. While preparing the data it was observed that the character of a ‘t-burst’ was very different from that of a ‘d-burst’. Figure 1 shows waveforms and spectrograms for two tokens at the extremes of the three acoustic cues.

The subjects sat in a sound isolation booth with headphones. The order of the tokens was randomized, and these were presented to the subjects in a simple forced choice format. They were seated at a computer that presented two boxes on the screen labeled ‘bit’ and ‘bid’. Subjects were instructed to click on the box that most closely matched their perception of the token, and to guess if the token was ambiguous or not clear. The boxes were reversed every 10 samples. The labels of the boxes on the screen were changed from ‘bet’ and ‘bed’ to ‘bit’ and ‘bid’, because early subjects reported that this was what they were hearing. The entire experiment took approximately ten minutes for each subject.

4 Results

Each of the following figures represents 448 trials (7 subjects * 64 tokens) summed over seven subjects. Figure 2 shows the results for ‘d-burst’ summed over consonant duration. The height on the z-axis (vertical axis in the figures) for any of the 16 points represents the number of ‘bit’ responses for a particular vowel duration and percent glottal pulsing summed over the four values of consonant duration. A 100% ‘bit’ response would translate to 28 on the plot (7 subjects * 4 consonant durations). There were 155 ‘bit’ responses out of 448. There are more ‘bit’ responses as glottal pulsing is reduced and vowel duration is decreased.

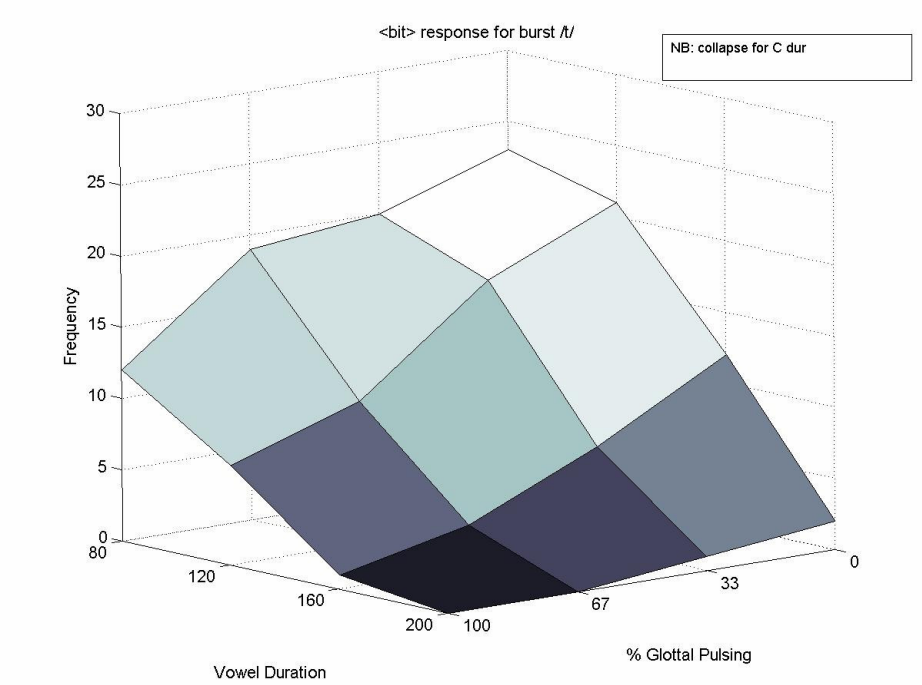

Fig. 2 Results for 'd-burst' summed over consonant duration.

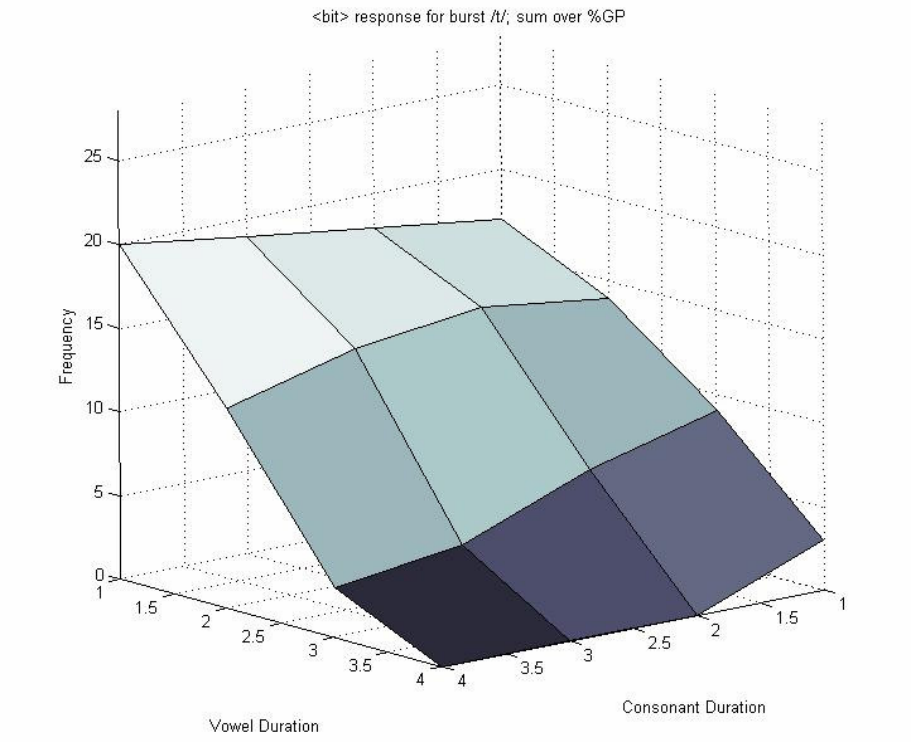

Fig. 3 Results for 'd-burst' percent glottal pulsing.

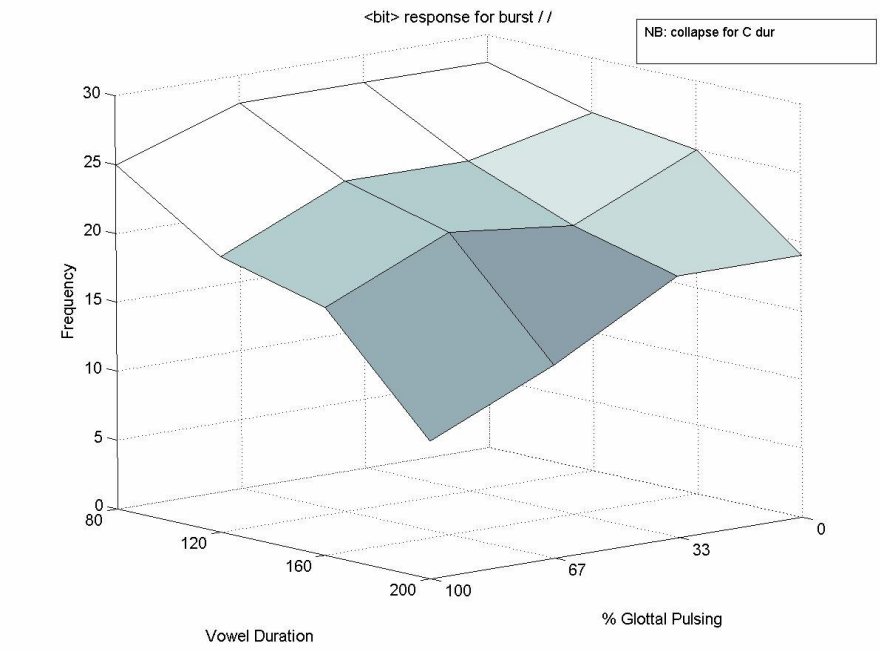

Fig. 4 Results for 't-burst' summed over consonant duration.

The picture is less clear for consonant duration in Figure 3, however. This is a plot of the same data summed over percent glottal pulsing. The slope across that dimension is not as steep, and it actually changes sign from short to long vowel duration. It is surprising that longer consonant duration seems to bias perception toward 'bid' at long vowel durations, and toward 'bit' at short vowel durations. Vowel duration is the strongest of the three factors influencing perception followed by percent glottal pulsing and consonant duration.

Figure 4 shows the results for the 't-burst' trials summed over consonant duration, and Figure 5 is the same data summed over percent glottal pulsing. Of particular note here is the bias in the direction of 'bit' responses. 351 out of 448 responses were 'bit' for the 't-burst' trials.

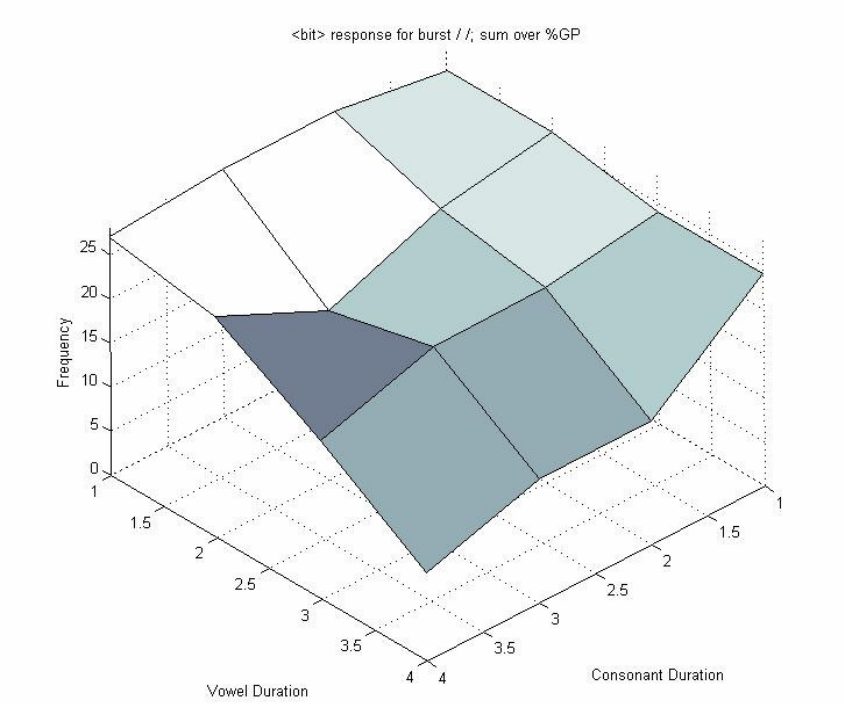

Fig. 5 Results for 't-burst' summed over percent glottal pulsing.

5 Discussion

The results show a clear picture of how these three factors influence the perception of the final stop. The slope of the lines across the vowel duration dimension is highest for all plots in this study. The slope of the lines across the percent glottal pulsing dimension are clear but not as steep as for vowel dimension, and consonant duration slopes do not show any clear picture when taken as a whole. Both the 't-burst' and 'd-burst' sets show the importance of vowel duration, and secondarily percent glottal pulsing. The strong bias of the 't-burst' data toward 'bit' responses makes it more difficult to see the relationships between the factors clearly, but the trends and the overall patterns are the same. The small number of participants in the study prevents drawing any definitive conclusions about the role of consonant duration, but it appears not to be a significant factor. With a larger sample it would be possible to determine in more detail the role of consonant duration and investigate the puzzle of its apparently contradictory behavior at different vowel durations.

6 Conclusion

The goals of the study were to determine the influence of vowel duration, percent glottal pulsing and consonant duration on final obstruent perception and to determine the effect of the final burst on perception. These objectives were achieved. Vowel duration is the strongest of the three features influencing coda stop voicing, followed by percent glottal pulsing and consonant duration. The data as presented here hint at a complicated set of trading relations between the various factors. The logical next step is to repeat the study with a refined token set and a larger number of participants and perhaps perform a regression analysis on the cues to quantify their contributions to the perception picture.

There are many avenues for further investigation. A production component could be added for each test subject to determine the character of their obstruent features. By recording subjects' speech for the tokens used and then comparing their perception crossover points (point where 50% of total is reached) with the midpoint between the features of the two recorded samples ('bit' and 'bid'), the relationship between production and perception can be directly studied. Certain models of human speech perception assume that production should match perceptual cues, and it would be possible using this framework to investigate that claim. It would be particularly interesting to perform such a study on various dialect groups around the country.

Another area of investigation is to compare results with natural tokens side by side with synthesized tokens. As Nittrouer (2004) found, such comparisons can yield radically different results. Most of the perception studies use synthesized speech tokens, and the question arises as to whether the results would be the same using natural speech tokens. Most of the perception studies also eliminate the final burst (Nittrouer 2004 and Thomas 2000, for example). Absence of burst appears to be the norm in running speech, so a logical next step would be to investigate these parameters without a final burst.

There are other cues that can be examined for their role in obstruent perception. Two candidates discussed in the literature are formant frequency drop in the vowel consonant transition (both F_1 and F_2) and differences in vowel character and contour (Thomas 2000 and Moreton 2004). The methodology here is extensible to handling multiple cues. One consideration in such a study would be the practical limit on the number of tokens that can be effectively presented to a subject in one sitting. Varying more than three or four cues in a particular token set would result in a prohibitively large number of samples. Alternately, separate token sets could be produced varying vowel duration and two other candidate cues, for example. The experiments could be conducted at separate sittings, and the results could be compared on the basis of the relative strength of a cue compared to its influence relative to vowel duration.

It is difficult to vary formant frequencies in natural speech tokens, and complex frequency manipulations might argue for the use of synthesized tokens. The factors inherent to natural speech that are not easily measured and characterized certainly warrant further investigation.

References

Boersma, Paul, and David Weenink. 2006. Praat: Doing Phonetics by Computer. (Version 4.4.30). [computer software]. Retrieved from http://www.praat.org/.

Milenkovic, Paul. 2006. Time Frequency analysis for 32-bit Windows [computer software]. Madison, WI: Department of Electrical Engineering, University of Wisconsin-Madison.

Moreton, Elliott. 2004. Realization of the English Postvocalic [voice] Contrast in F1 and F2. *Journal of Phonetics* 32: 1-33.

Nittrouer, Susan. 2004. The Role of Temporal and Dynamic Signal Components in the Perception of Syllable-Final Stop Voicing by Children and Adults. *Journal of the Acoustical Society of America* 115: 1777-90.

Purnell, Thomas C., Joseph C. Salmons, and Dilara Tepeli. 2005. German Substrate Effects in Wisconsin English: Evidence for Final Fortition. *American Speech* 80: 135-64.

Thomas, Erik R. 2000. Spectral differences in /ai/ offsets conditioned by voicing of the following consonant. *Journal of Phonetics* 28: 1-25.

University of Wisconsin-Madison
Department of Linguistics
1168 Van Hise Hall
1220 Linden Drive
Madison, WI 53706-1557

brodgers@wisc.edu