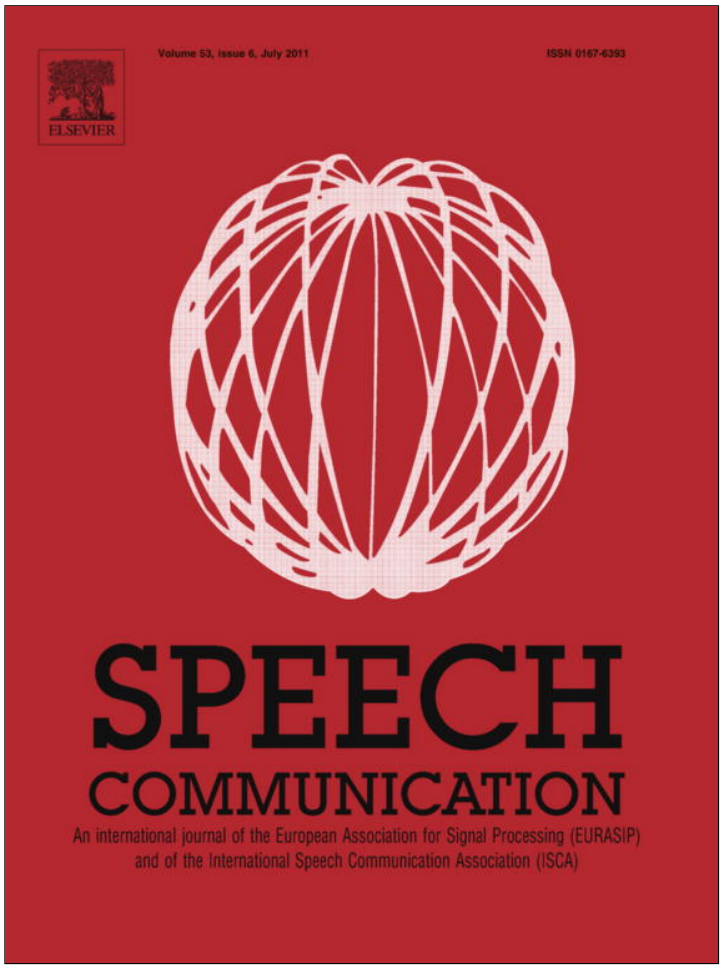

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 877–888
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A standard set of American-English voiced stop-consonant stimuli from morphed natural speech

Joseph D.W. Stephens *, Lori L. Holt

Psychology Department and Center for the Neural Basis of Cognition, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA

Received 29 August 2010; received in revised form 15 February 2011; accepted 23 February 2011
Available online 3 March 2011

Abstract

Linear predictive coding (LPC) analysis was used to create morphed natural tokens of English voiced stop consonants ranging from /b/ to /d/ and /d/ to /g/ in four vowel contexts (/i/, /æ/, /a/, /u/). Both vowel–consonant–vowel (VCV) and consonant–vowel (CV) stimuli were created. A total of 320 natural-sounding acoustic speech stimuli were created, comprising 16 stimulus series. A behavioral experiment demonstrated that the stimuli varied perceptually from /b/ to /d/ to /g/, and provided useful reference data for the ambiguity of each token. Acoustic analyses indicated that the stimuli compared favorably to standard characteristics of naturally-produced consonants, and that the LPC morphing procedure successfully modulated multiple acoustic parameters associated with place of articulation. The entire set of stimuli is freely available on the Internet (http://www.psy.cmu.edu/~lholt/php/StephensHoltStimuli.php) for use in research applications.

Keywords: Speech stimuli; Consonants; Linear predictive coding

1. Introduction

One of the mainstays of speech perception research is the use of series of speech stimuli that vary perceptually from one phonetic endpoint to another. Many experiments make use of perceptually ambiguous speech stimuli as a means of identifying factors that influence speech perception, including categorical perception (Liberman et al., 1957), lexicality (Ganong, 1980; Elman and McClelland, 1988), transitional probability (Pitt and McQueen, 1998), lexical neighborhood size and density (Newman et al., 1997), visual information (Massaro, 1987, 1998), native language experience (McCandliss et al., 2002), and adjacent phonetic or acoustic context (Mann, 1980; Lotto and Kluender, 1998; Holt, 2005). Relatively straightforward sound-editing methods can be used to create ambiguous sounds from natural speech tokens of unvoiced fricatives (e.g., /s/ and /ʃ/; McQueen, 1991) or phonemes that differ in voice-onset time (e.g., /t/ and /d/; Ganong, 1980). However, ambiguous sounds between natural tokens of voiced stop consonants (e.g., /b/, /d/, /g/) are more difficult to create using simple waveform mixing or editing techniques. Stop place of articulation is signaled by several acoustic cues including formant transitions, which contain complex temporal and frequency information. Speech synthesizers like those of Klatt (1980; Klatt and Klatt, 1990) allow for explicit control of the acoustic characteristics of speech, but compared to natural speech such synthetic speech can sound unnatural.

An alternative to the use of purely synthetic speech sounds is the application of digital signal processing techniques to "morph" between natural speech tokens and produce more realistic-sounding stimuli. Linear predictive coding (LPC) analysis (Atal and Hanauer, 1971; Markel

* Corresponding author. Present address: Department of Psychology, North Carolina A&T State University, 1601 E. Market St., Greensboro, NC 27411, USA. Tel.: +1 336 285 2266; fax: +1 336 334 7538.
E-mail addresses: jdstephe@ncat.edu (J.D.W. Stephens), lholt@andrew.cmu.edu (L.L. Holt).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.02.007

and Gray, 1976) can be used to model the vocal-tract filter properties of speech sounds and the resulting filters can be modified to produce acoustically and perceptually intermediate sounds. The interpolation of LPC-derived filters to create speech stimuli has been used successfully in speech perception experiments involving liquid consonants (e.g., McCandliss et al., 2002), and recent efforts have been made to enhance the ease and automaticity of speech-morphing procedures (Slaney et al., 1996; Pfitzinger, 2004). Nonetheless, the application of LPC methods can require significant time and effort, as well as trial and error, to successfully produce high-quality sounds that are free of noticeable acoustic artifacts. The purpose of the current project was to aid researchers by providing a useful set of LPC-morphed stimulus series, consisting of voiced stop-consonant utterances with high acoustic quality and natural characteristics. The series consist of vowel–consonant–vowel (VCV) and consonant–vowel (CV) utterances in which the consonants range from /b/ to /d/ and from /d/ to /g/, in four vowel contexts: /i/, /æ/, /a/, and /u/. The stimuli are freely available electronically (Appendix A).

It is important to note that because the stimuli presented here were created by interpolating between natural tokens, each stimulus in the set may vary from others across multiple acoustic dimensions. As a result, the stimuli may not be appropriate for use in experiments that require explicit control of individual acoustic dimensions. They will be most useful in situations where researchers wish to obtain straightforward measures of participants' consonant identification, discrimination, and/or intelligibility. For example, these stimuli have been used to examine visual influences on consonant identification at varying noise levels (Stephens and Holt, 2010). It should also be noted that, due to the resynthesis procedure, close listening may reveal some minor artifacts within certain stimuli. Decisions about how to use individual members or subsets of the stimulus set in particular applications should still be based upon careful examination and pilot testing whenever possible. Researchers may find it useful to truncate or otherwise further edit the stimuli to adapt them to individual circumstances.

The following text describes the methods used to create the stimuli, a behavioral experiment that evaluated perception of the stimuli by native speakers of American English, and several acoustic analyses of the stimuli. The experiment and analyses, respectively, demonstrated that the LPC-morphing procedure succeeded in producing sounds that were reliably perceived as members of English /b/, /d/, and /g/ categories, and that several perceptually relevant features of the stimuli were gradually adjusted across each series and were consistent with previously established observations of naturally-produced, voiced stop consonants in American English.

2. Morphing natural tokens

Eight 20-member series of acoustic stimuli were created with the morphing procedure. Each series consisted of VCV tokens in which the consonant ranged from /b/ to /d/ or from /d/ to /g/, in one of four vowel contexts: /i/, /æ/, /a/ or /u/. The stimuli were created by adjusting filter parameters of natural utterances derived from LPC analysis and applying the adjusted filters to source waveforms extracted from the natural tokens.

An adult, Midwestern American male speaker (J.D.W.S.) produced three repetitions of each of the 12 VCV combinations used in the stimulus set. The tokens were recorded digitally on a personal computer using Computer Speech Laboratory (CSL; Kay Elemetrics Corp., Lincoln Park, NJ) with 16-bit precision at a sampling rate of 11.025 kHz. The tokens were isolated and saved separately as monaural PCM .wav files. The tokens were then matched in RMS power prior to further processing.

The alignment of temporal characteristics of endpoint sounds is one of the primary challenges of sound morphing (particularly for automatic methods; Slaney et al., 1996; Pfitzinger, 2004). For the current stimuli, this problem was circumvented by maximizing the alignment of temporal characteristics between endpoints prior to morphing. Within each vowel context, the tokens for each consonant that were most compatible in pitch and temporal properties (i.e., speaking rate, burst length, and duration) were selected as series endpoints. The selected /b/, /d/ and /g/ tokens within each vowel context were then edited to produce further temporal alignment. The offsets and lengths of the initial and final vowels were aligned by deleting or duplicating pitch periods within the vocalic portions of the waveforms as necessary. The burst onsets of the consonants were aligned by shortening or lengthening the post-vocalic silences. In some cases it was necessary also to realign or shorten consonant bursts to maximize the compatibility of the endpoint tokens.[1]

An LPC analysis was performed on each of the edited natural endpoint tokens using the autocorrelation algorithm (Markel and Gray, 1976) implemented in the computer program Praat (version 4.3.19; Boersma, 2001) with the following parameters: prediction order 16; analysis width (Hamming window) 25 ms; time step 5 ms; pre-emphasis above 50 Hz. The resulting filter coefficients were saved in text files for editing. The /d/ tokens from each vowel context were inverse-filtered by their LPC coefficients to extract approximate voicing sources for each /d/ endpoint token. The resulting four source waves (one for each vowel context) were saved and used in the subsequent resynthesis of all stimuli within a corresponding vowel series.

To create series ranging perceptually between endpoint consonants, the LPC coefficients for each endpoint were read into a MatLab script (version 6.0.0.88; The MathWorks, Inc., Natick, MA) that computed the differences between the endpoint tokens' filter coefficients at every

[1] The original recorded tokens as well as the edited tokens used for the morphing procedure are also available online.

5-ms time step and recorded 20 new sets of coefficients by linearly interpolating between the two endpoints (thus, for each time step, 18 equally-spaced, intermediate LPC coefficients were computed). For any time step for which the endpoint filters possessed different numbers of coefficients, the lower number of coefficients was used in the output filter (as a result, the endpoint filters that were created as output were not necessarily exactly identical to the filters used as input). Thus, filters were created that ranged from /b/ to /d/ and from /d/ to /g/, in the /i/, /æ/, /a/, and /u/ VCV contexts. After each series of LPC filters was created, Praat was used to apply each of the filters to the source wave derived from the /d/ token with the corresponding vowel, so that all members of each VCV series were based on the same voicing source. The MatLab and Praat scripts used to create and apply the filters are provided in the online archive of stimuli.

Subsequent to resynthesis, all 160 VCV stimuli were RMS-matched. A 100-ms silent interval was added to the beginning of each waveform file and the durations of all stimuli were standardized to 830 ms by adding silence to the end of each file as necessary. An additional set of 160 CV stimuli were created by excising the initial vowel from each of the VCV stimuli. Each VCV was cut at the time point corresponding to the last zero-crossing prior to the consonant release in the relevant source wave. After editing, all 160 CV stimuli were RMS-matched. A 100-ms silent interval was added to the beginning of each waveform file and the durations of all stimuli were standardized to 515 ms by adding silence to the end of each file as necessary. In all, 320 stimuli were created (20-step morphing $\times$ 2 perceptual continua $\times$ 4 vowel contexts $\times$ 2 syllable types).

3. Behavioral experiment

The procedure described above defined sets of filters with parameters intermediate between two endpoint consonants. The usefulness of these sounds in speech research depends not on their filter parameters, but rather on their perceptual characteristics. Thus, to verify that the stimuli were perceived in the manner intended by the morphing process, a behavioral experiment was conducted in which participants identified the consonants in a 3-alternative, forced-choice task.

3.1. Method

3.1.1. Stimuli

The eight (two consonant series: /b/–/d/; /d/–/g/ $\times$ four vowels) 20-member series of VCV utterances and the eight 20-member series of CV utterances derived from the VCVs were used as stimuli in the experiment.

3.1.2. Participants

Thirteen participants were recruited from the Carnegie Mellon University community. One participant's data showed no discernible pattern in identification responses across consonant series; this participant's data were not included in group analyses. All participants were monolingual, native speakers of American English with no reported hearing impairment. Each participant received $7 after completing each session of the experiment (four sessions, $28 total). Each session lasted approximately 50 min.

3.1.3. Procedure

The two series ranging from /b/ to /d/ and /d/ to /g/ in each vowel context were combined to yield groups of 40 stimuli spanning the three response categories. Thus, each stimulus was presented within the context of the overall range of English voiced stop consonants. Each 40-member stimulus series was presented within a separate experimental trial block. Thus, there were eight blocks, one for each of the four vowels (/i/, /æ/, /a/, or /u/) within each utterance type (VCV or CV). These eight blocks were spread across four experimental sessions that were conducted on separate days. Each session separately tested perceptual identification for consonant stimuli within one of the four vowel contexts, with one block devoted to VCV stimuli and the other block devoted to the corresponding CV stimuli. The ordering of the four sessions for each participant was counterbalanced according to a Latin square design and the ordering of blocks within each session alternated from session to session. During each block, a set of 40 stimuli ranging from /b/ to /d/ to /g/ was repeated 10 times. For each repetition of the set, stimuli were presented in random order. Participants were instructed to listen to the stimuli and identify the consonants by pressing buttons on an electronic response box labeled "B," "D," and "G." Presentation of stimuli was controlled using Tucker-Davis Technologies System II hardware. After digital-to-analog conversion, stimuli were low-pass filtered at 4.8 kHz[2] and output diotically to Beyer DT-150 headphones at approximately 65–70 dB. Participants were seated in sound-attenuated booths during the experiment.

During data collection, it was discovered that a faulty switch on the output line to the headphones caused some stimuli to be presented through only the left channel. Although this problem was corrected, it affected a total of eleven experimental sessions spread across five participants: three sessions for two participants; two sessions for two participants; and one session for one participant. The other seven participants were not affected. In order to ensure that this malfunction did not influence the interpretation of the experimental results, statistical analyses were performed both with and without the data from the

[2] It will be noticed that based on the 11.025 kHz sampling rate, the stimuli contain frequencies up to the Nyquist frequency of 5.5125 kHz. The filtering of frequencies above 4.8 kHz is unlikely to have dramatically affected identification of these consonants; however, researchers attempting to replicate our findings might wish to apply the same low-pass filter to the stimuli.

affected sessions. The exclusion of the affected data did not substantially alter the results.

3.2. Results

Fig. 1 displays identification data averaged across 12 participants for each of the /b/–/d/–/g/ consonant series, separately for VCV (solid lines) and CV (dashed lines) stimuli. For each stimulus series, participants sorted stimuli into three distinct categories, with clear boundaries between stimuli labeled "B," "D," and "G." Thus, the LPC-morphing procedure succeeded in creating series that were perceived by listeners as shifting perceptually between good exemplars of the voiced stop-consonant categories.

One notable feature of the data is the apparent difference in responses due to presence or absence of an initial vowel. Across all four vowel contexts, participants responded with "D" more often to VCV stimuli than to CV stimuli. This pattern was examined using a 2 (initial vowel) × 40 (series number) repeated-measures, analysis of variance (ANOVA) on "D" responses, separately for each stimulus series. In each of the four vowel contexts, the effect of initial vowel was significant (for /i/, $F(1, 11) = 23.91$, $p < .001$, $\eta_p^2 = .685$; for /æ/, $F(1, 11) = 7.52$, $p = .019$, $\eta_p^2 = .406$; for /a/, $F(1, 11) = 9.23$, $p = .011$, $\eta_p^2 = .456$; for /u/, $F(1, 11) = 19.96$, $p = .001$, $\eta_p^2 = .645$). Consistent with the clear shifts in identification across each series, a multivariate effect of series number was significant in all four vowel contexts (for /i/, $F(39, 429) = 131.86$, $p < .001$, $\eta_p^2 = .923$; for /æ/, $F(39, 429) = 269.63$, $p < .001$, $\eta_p^2 = .961$; for /a/, $F(39, 429) = 132.23$, $p < .001$, $\eta_p^2 = .923$; for /u/, $F(39, 429) = 109.32$, $p < .001$, $\eta_p^2 = .909$). A significant interaction of initial vowel and series number was also found in each of the four vowel contexts (for /i/, $F(39, 429) = 6.13$, $p < .001$, $\eta_p^2 = .358$; for /æ/, $F(39, 429) = 8.17$, $p < .001$, $\eta_p^2 = .426$; for /a/, $F(39, 429) = 2.21$, $p < .001$, $\eta_p^2 = .167$; for /u/, $F(39, 429) = 6.32$, $p < .001$, $\eta_p^2 = .365$). These interactions indicate that the effect of initial vowel varied across the series; it can be seen from Fig. 1 that this effect tended to be greatest near the perceptual boundaries between consonant categories.

As stated above, all statistical analyses were repeated, excluding data from sessions in which participants erroneously received headphone output in only one ear. Excluding these data, the effect of initial vowel on "D" responses in the /æ/ series was no longer significant, $F(1, 10) = 4.43$, $p = .062$, $\eta_p^2 = .307$.

Inspection of data from individual participants indicated that, for some trial blocks, individuals' responses did not fit the overall pattern of categorization seen in the average data. That is, for the large majority of trial blocks run in the experiment (90 of 96) each of the three response options reached ceiling (i.e., 100%) in the participant's responses at some point along the stimulus series. However, in six of the trial blocks (three blocks for one participant and one block for each of three participants – three of these blocks were also affected by the output line

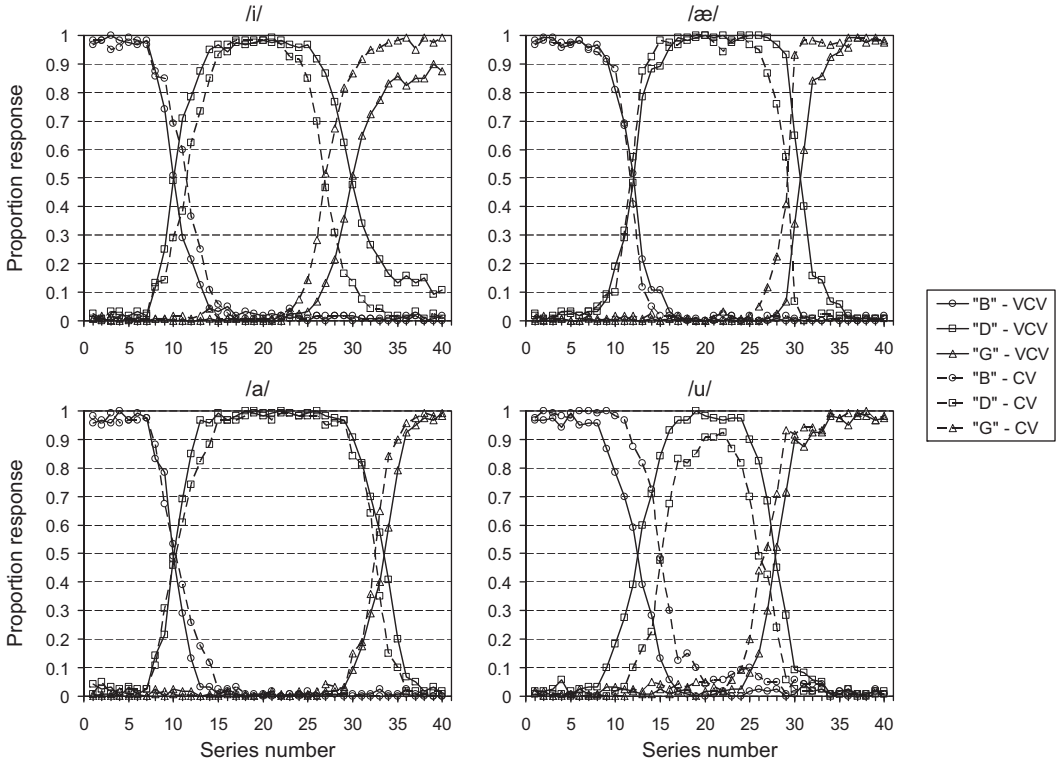

Fig. 1. Proportion "B" (circles), "D" (squares), and "G" (triangles) responses across the four stimulus series. Solid lines represent responses to VCV stimuli; dashed lines represent responses to CV stimuli.

problem described above), one of the response categories did not reach ceiling at any point along the stimulus series. In three of these blocks one response category did not exceed 50% for any stimulus. As a result, it could be the case that these few, anomalous blocks of data distorted the averages in a manner that caused spurious effects to be detected in the statistical analysis. The statistical analyses were therefore repeated excluding both data blocks (i.e., CV and VCV) from any sessions that contained an anomalous data block. All significant effects on "D" responses in the full data set were also significant when these anomalous data blocks were excluded.

3.3. Discussion

The LPC interpolation procedure was successful in creating stimuli that transitioned between good perceptual examples of American-English /b/, /d/, and /g/ consonant categories. Although the averaged data show a few cases in which one of the responses did not reach ceiling levels, individual participants nearly always used each of the three response labels at the maximum rate somewhere along each of the series. Interestingly, the perception of stimuli near consonant category boundaries was affected by whether the stimuli included the initial vowels produced in the original VCV utterances. Specifically, participants identified consonants as "D" more often when the initial vowel was present than when the initial vowel was absent. This effect was observed across all four series, although it was more pronounced for the /i/ and /u/ series than for the /a/ and /æ/ series, as evidenced by greater measures of effect size (i.e., η_p^2). With one exception (the /æ/ series) the effects of initial vowel on "D" responses remained significant in analyses that excluded potentially problematic behavioral data. Some possible sources for the effect of initial vowel are considered below, in Section 4.1.2.

The results of the identification task demonstrate the efficacy of the LPC-interpolation technique for creating stimuli that morph perceptually between natural tokens. For these stimuli to be used effectively in speech perception research, it will also be desirable to have information regarding their acoustic properties and the extent to which these acoustic characteristics are consistent with prior observations of the nature of voiced stop consonants. The following section describes a set of systematic acoustic measurements that were made of the LPC-interpolated stimuli.

4. Acoustic analyses

LPC interpolation and resynthesis were used to create these stimuli because of the ability of such a procedure to simultaneously vary numerous acoustic properties of speech sounds while retaining the sounds' natural quality. Thus, in addition to demonstrating the perceptual reliability of the stimuli, it is important to document their specific acoustic characteristics. Several acoustic analyses were performed to verify that the morphing procedure produced gradual changes across several acoustic dimensions, and that the acoustic properties of the resynthesized stimuli were consistent with previous research on natural productions of voiced stop consonants. Some light was also shed on the perceptual differences caused by the deletion of the initial vowel from the VCV stimuli. Specifically, several acoustic factors may have differentially affected consonant identification after the initial vowel across the various series. In some cases, the initial vowel had phonetic characteristics more consistent with /d/; in others, the formant frequencies of the initial vowel may have influenced consonant identification via spectral-contrast mechanisms; and in others, the gross acoustic characteristics of consonant onset were slightly atypical for the /d/ tokens. Detailed analyses are explained below.

4.1. Formant frequencies and locus equations

4.1.1. Measurements

One of the primary correlates of stop-consonant place of articulation is formant frequency trajectory, particularly for the second ($F2$) and third ($F3$) formants, although the formant patterns corresponding to each consonant category are highly dependent on vowel context (Öhman, 1966; Lindblom, 1963). Despite this dependence of formant frequencies on vowel context, $F2$ frequency of consonant onsets relative to neighboring vowels often exhibits a reliable linear relationship within each consonant (i.e., place of articulation) category, which can be modeled using linear regression to derive "locus equations" for consonant place (Lindblom, 1963; Sussman et al., 1991). Therefore, the relation of the formant frequencies of the current stimuli to stop consonant perception was also examined by computing locus equations for the three consonant categories.

An automatic procedure was used to measure formant frequencies of the current stimuli at specific time points. The following time points were marked separately for each of the VCV series: the midpoint of the initial vowel; the offset of the initial vowel (defined here as the offset of harmonic structure, e.g., Sussman et al., 1997); onset of the consonant (defined here as the onset of formants immediately after burst; e.g., Sussman et al., 1991); early in the second vowel (after consonant transition); and the midpoint of the second vowel. Using Praat (version 4.3.19; Boersma, 2001), formant trajectories were computed for all 160 VCV stimuli using the Burg algorithm with the following parameters: five formants; maximum formant frequency 5 kHz; window length 25.4 ms; and time step 6.35 ms. The formant values for $F1$, $F2$, and $F3$ at the identified time points were then recorded for each of the 160 stimuli. For a few of the stimuli near /ugu/, the formant analysis yielded anomalous values that were obviously inconsistent with neighboring series members. In these cases, the algorithmically-generated values were corrected by visual inspection of spectral plots (FFT, in Praat) of

the stimuli at the specified time points and measuring the formant center frequency.

Fig. 2 displays formant frequencies measured at the midpoint of the initial vowel, at the offset of initial vowel, at the onset of the consonant (i.e., the onset of formants after burst), early in the final vowel, and at the midpoint of the final vowel; the formant frequencies are plotted separately for each /b/–/d/ and /d/–/g/ series. The numerical values of the formant frequencies represented in Fig. 2 are also provided in tabular form in the online archive described in Appendix A. The measurements indicate that the interpolation of LPC filter coefficients led to gradual steps in formant frequencies across each series, although not every step was the same size. The primary shifts in formant frequencies across consonants in each series are seen in the frequencies of *F*2 and *F*3 at initial vowel offset and consonant onset. Vowel formant frequency characteristics remained stable across most series, with the exception of the /u/ series, for which *F*2 of the final vowel showed an apparent influence of the consonant.

Also included in Fig. 2 are corresponding formant frequencies from the temporally-aligned natural tokens that were used as inputs to the morphing procedure. It can be seen that, in general, the re-filtering of the /d/ source with the LPC filters of /b/ and /g/ resulted in formant frequencies that were very similar to the original /b/ and /g/ recordings. In two cases (/ibi/ and /aba/), the morphing procedure exaggerated some of the differences in the formant frequencies of the natural tokens; in one case (/ugu/) the morphing reduced some of the differences in formant frequencies of the natural tokens.

Using the data from the formant analyses, the acoustic correlates of place articulation in these stimuli were examined using locus equations. For each consonant category, the three most extreme series members were selected from each vowel context,[3] and locus equations were computed by fitting regression lines to the points defined by *F*2 center frequency at consonant onset and at the midpoint of the final vowel (e.g., Sussman et al., 1991). For stimuli representing /g/, separate lines are commonly used to fit locus equations for the more palatal [ɡ̟] and more velar [g] allophones that occur in front vowel and back vowel contexts, respectively (Sussman et al., 1991, 1997). In the current study, only two front (/i/, /æ/) and two back (/a/, /u/) vowel contexts were available, so the locus equations were allowed to overlap in order to reduce the influence of each individual vowel context on the regression. Thus the locus equation for front [ɡ̟] was estimated using /i/, /æ/, and /a/, and the locus equation for back [g] was estimated using /æ/, /a/, and /u/.

Fig. 3 displays locus equations for each consonant along with the measured points used to compute them. Also displayed are the locus equations for 15 male speakers of American English reported in two studies of naturally-produced voiced stop consonants (Sussman et al., 1991, 1997).[4] It can be seen that the locus equations for stimuli identified strongly as belonging to one of the consonant categories were generally similar to the locus equations for the corresponding consonants found by previous studies. Differences between the current locus equations and those of previous studies are due mainly to the influence of a few stimuli for which the relation between *F*2 at consonant onset and at vowel midpoint were slightly outside the typical range. Specifically, for stimuli identified as /ubu/ and /idi/, *F*2 frequencies for the consonant were somewhat higher relative to the vowel than in previous studies. For stimuli identified as /ugu/, consonant *F*2 was somewhat lower relative to vowel *F*2 than in previous studies.

4.1.2. Influence of initial vowel

The above analysis indicates that the formant frequencies of consonant onsets in the current stimuli compare favorably to previous measurements of naturally-produced, voiced stop consonants. However, the observed differences in perception of VCV and CV stimuli indicate that consonant perception is also influenced by information present in the initial vowel.

One possible explanation for how the initial vowel could affect consonant identification is that the formant transitions at the offset of the vowel carried information that was more consistent with a given consonant category than the formant transitions after consonant release. In this case, perhaps the greater proportion of "D" responses seen across each VCV stimulus series was due to the formant transitions in the initial vowel of any given stimulus carrying more /d/-like information than the formant transitions of the consonant alone. Comparison of the formant frequencies of the natural and morphed tokens in Fig. 2 suggests that, in general, the initial vowels of the morphed stimuli did not carry over residual /d/ qualities as a result of using a /d/ source in the morphing procedure. The one clear exception to this is the /udu/–/ugu/ series, in which the *F*2 offsets of the initial vowel in the morphed /g/ endpoints appear to have been higher in frequency than in the original recording (and thus more consistent with the *F*2 offset of the natural /d/ token). Thus, perhaps some of the initial vowel's influence in the /udu/–/ugu/ series may be the result of more /d/-like information in the initial vowel due to morphing.

Another possibility is that the articulatory properties of these VCV utterances generally result in more /d/-like

[3] Note that for /d/, since series numbers 20 and 21 had the same filter parameters, the three "most extreme" versions of /d/ used in the regressions for locus equations were series numbers 19, 20, and 22.

[4] The locus equations from Sussman et al. (1997) represent the measurements reported from initial consonants in CVC utterances. Although Sussman et al. also studied VCV utterances, the initial and final vowels were independently varied in that study to evaluate coarticulatory influences of the initial vowel.

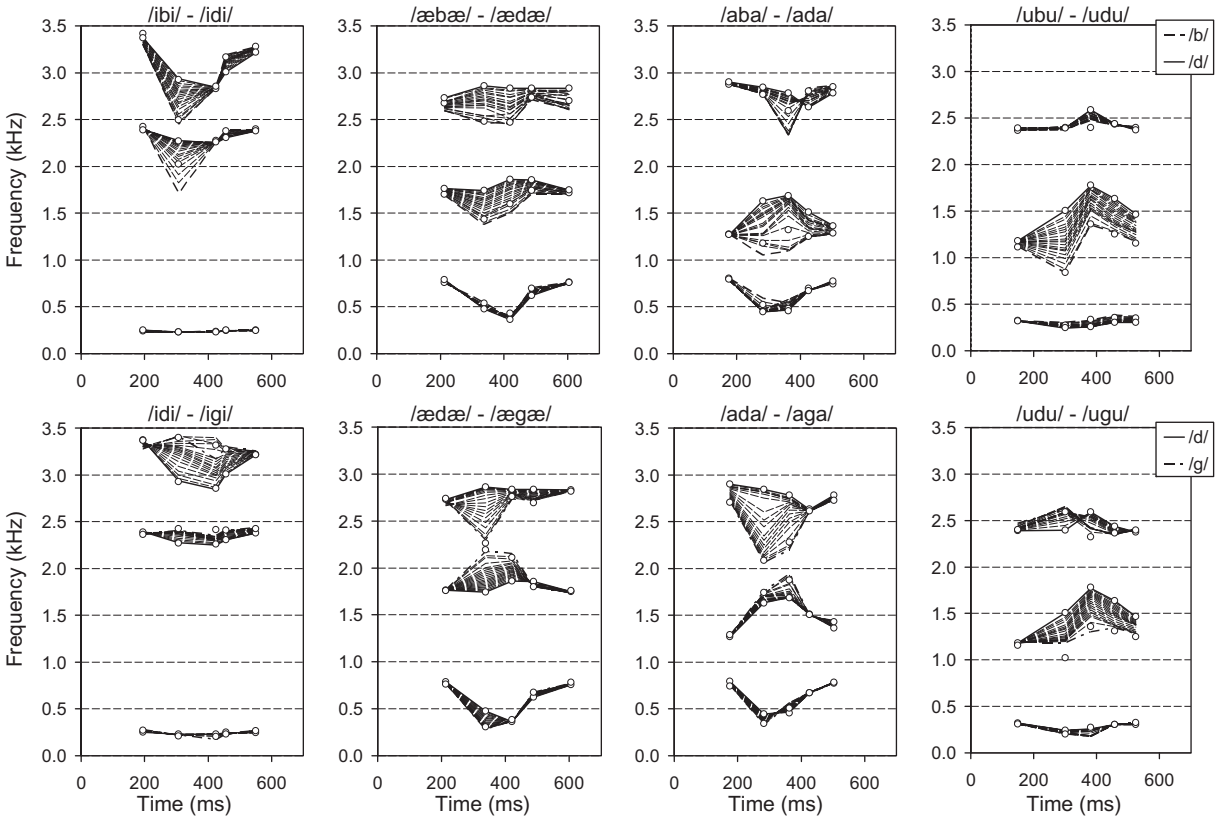

Fig. 2. Formant frequencies of F1, F2 and F3 measured at five time points in each VCV stimulus: the midpoint of the initial vowel; the endpoint of the initial vowel; the onset of the consonant; early in the final vowel; the midpoint of the final vowel. Solid lines indicate the /d/ endpoint of each series; dashed lines and dotted–dashed lines indicate the /b/ and /g/ endpoints of each series, respectively. For comparison, the open symbols represent formant frequencies measured from the aligned natural recordings prior to morphing.

information in the initial vowel. For example, it is evident in Fig. 2 that the many of the stimuli exhibit an asymmetry between formant frequencies at the offsets of the initial vowels and at the onsets of the consonants. These differences might influence the phonetic information available to the listener when the initial vowel is present. To evaluate the phonetic information present in the initial vowel, an additional set of locus equations was derived using data points defined by *F*2 frequency at the endpoint of the initial vowel and at the midpoint of the initial vowel. Table 1 compares these locus equations with the equations derived from measurements of *F*2 at consonant onset and at the midpoint of the final vowel. Table 1 also lists mean locus equation data from the male speakers in the study by Sussman et al. (1997), which compared syllable-initial and syllable-final consonants. Beyond the differences in slope

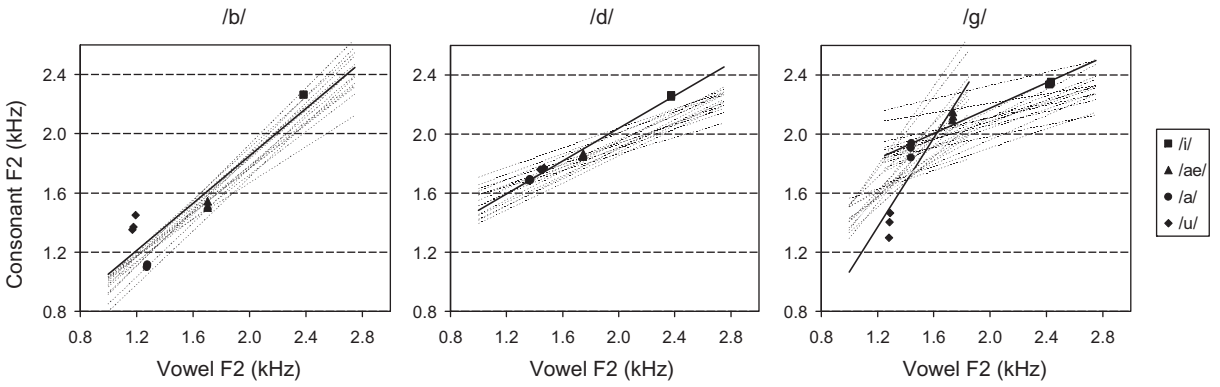

Fig. 3. Locus equations (solid lines) based on the three most extreme instances of each consonant category from each stimulus series, as indicated by responses in the perceptual experiment (symbols). The *y*-axis represents F2 frequency at consonant onset and the *x*-axis represents F2 frequency at the midpoint of the final vowel. Dotted lines depict locus equations for ten male speakers reported by Sussman et al. (1991, 1997).

Table 1
Locus equation slopes and *y*-intercepts for consonants at syllable onset and syllable offset. The data from Sussman et al. (1997) represent means of five male speakers.

	Syllable	/b/		/d/		Front [g̟]		Back [g]	
		Slope	*y*-Intercept	Slope	*y*-Intercept	Slope	*y*-Intercept	Slope	*y*-Intercept
Current stimuli	CV	.799	252	.553	932	.429	1318	1.51	−45
	VC	.683	194	.614	764	.601	1009	1.37	−23
Sussman et al. (1997)	CV	.734	290	.330	1280	.290	1453	1.07	316
	VC	.621	216	.351	1135	.436	1338	.857	303

between the current stimuli and previous measurements (discussed above in relation to Fig. 3), which suggest greater coarticulation with the vowel contexts, correspondences can be seen across studies with regard to the differences between CV and VC locus equations. In particular, for when analyzing VC utterances, the slopes of locus equations for /b/ and back [g] were reduced, whereas the slopes for /d/ and front [g̟] were increased, relative to CV utterances. Also, the differences in slope across all consonant categories were reduced among VC utterances than among CV utterances; similar observations have been made in other studies (Lindblom, 1963; Krull, 1988). Within each study, the general shift toward more intermediate slopes for VC offsets also resulted in slopes that were more similar to the slope for the /d/ category, particularly in the current stimulus set. Thus, the initial vowels might have shifted identification responses toward "D" by providing additional acoustic information that was *relatively* more consistent with the /d/ category than the information present from the onsets of the consonants onward (again, within the context of the *F*2 characteristics of the current stimulus set). However, this explanation does not offer clues as to why the initial vowel had a bigger effect in some vowel contexts than in others.

Another explanation for how the presence of initial vowels led to differences in "D" identification is that the formant frequencies of the initial vowels may have shifted perception of subsequent formant frequencies in a direction favorable to perception of /d/. Such a shift would be similar to phonetic context effects originating from contrastive auditory mechanisms (e.g., Lotto and Kluender, 1998; Holt, 1999). For instance, manipulating the *F*2 frequency of an initial vowel can cause shifts in the consonant identification of subsequent CV syllables (Holt and Lotto, 2002). Similar effects have been found using a variety of speech and non-speech stimuli (for a summary, see Lotto and Holt, 2006) and have been characterized as resulting from a perceptual mechanism that exaggerates differences between formant frequencies of temporally adjacent sounds. In the current stimulus set, this type of contrastive mechanism could shift perception toward /d/ when formant frequencies at the offset of an initial vowel are further from the /d/-endpoint than the formant frequencies of the subsequent consonant. For example, in the /ubu/–/udu/ series, *F*2 at consonant onset is higher for /d/ than for /b/ (see Fig. 2). Additionally, for each individual stimulus along the series, *F*2 at initial vowel offset is lower than *F*2 at consonant onset. If a contrastive auditory process is at work, the presence of the relatively low *F*2 information at initial vowel offset in VCV stimuli should shift perception of *F*2 at consonant onset toward higher perceived frequencies, and thus toward more "D" identifications.

Inspection of the formant data in Fig. 2 suggests that conditions for contrastive auditory mechanisms could exist in several of the series. As mentioned above, the low *F*2 offset in the /ubu/–/udu/ series could shift perception of *F*2 onset upward (i.e., toward /d/) for the consonant; the same pattern is seen in *F*2 for /ibi/–/igi/, /æbæ/–/ædæ/, /aba/–/ada/, and /udu/–/ugu/. Further, contrastive relationships favoring /d/ are present in *F*3 for /ibi/–/idi/, /idi/–/igi/, /ædæ/–/ægæ/, and /ada/–/aga/. Taken together, these relationships may offer at least a partial account for why the effect of the initial vowel was greater in some vowel contexts than others. For example, in the case of /ibi/–/idi/, both *F*2 and *F*3 perception may have been shifted toward /d/ due to spectral contrast caused by the initial vowel. For /idi/–/igi/, the consonants differed primarily in *F*3; thus the contrast in *F*3 alone might have strongly influenced identification. Similarly, for both the /ubu/–/udu/ and /udu/–/ugu/ series, *F*2 was the primary correlate of place articulation; thus the fact that *F*2 contrast favored /d/ in these series could have had a substantial impact (see also the analyses of gross spectral shape below for another potential factor influencing the perception of /udu/). On the other hand, within the /aba/–/ada/ and /ada/–/aga/ series, spectral contrast in *F*2 and *F*3 could have conflicting effects (i.e., toward /d/ in one formant and away from /d/ in the other formant) possibly explaining why the influence of initial vowel appears to have been smallest in the /a/ context.

4.2. Gross spectral shape

Another acoustic correlate of stop-consonant place of articulation is the shape of the "short-time" spectrum at consonant onset. Acoustic theory indicates that the gross spectral shape sampled over 10–20 ms following consonantal release should exhibit distinctive characteristics depending on place of articulation (Fant, 1960). These theoretically derived gross spectral shapes have been shown to be good approximations of natural consonant productions (Blumstein and Stevens, 1979), and to be relevant

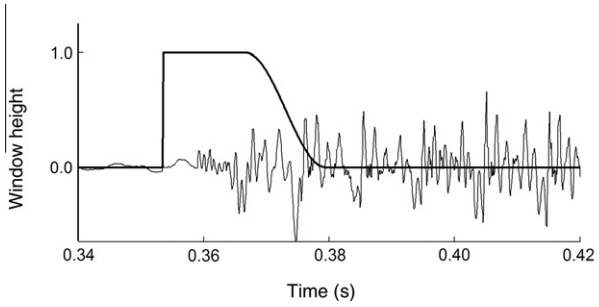

Fig. 4. Illustration of the window applied to each stimulus waveform in order to compute the short-time spectrum for consonant onset (see Blumstein and Stevens, 1979).

for perceptual identification of consonants (Stevens and Blumstein, 1978; Blumstein and Stevens, 1980). Therefore the extent to which the current stimuli agree with theoretical predictions concerning gross spectral shape at consonant onset was examined, particularly with regard to perceptual identification of the stimuli.

Following the method of Blumstein and Stevens (1979), the first difference of each stimulus waveform was computed (to pre-emphasize higher frequencies) and was multiplied by a modified raised cosine window over the first 26 ms after consonant release. A time point corresponding to consonant release was defined for the members of each stimulus series as the last zero-crossing prior to the onset of burst noise in the unfiltered source waveform used to generate that series. The shape of the window (see Fig. 4) was such that the earliest portion of the burst contributed most to the spectrum. The power spectral density was then computed using a 14-pole Burg algorithm (in MatLab).

Fig. 5 displays power spectral density plots for the consonant onsets of the three most extreme stimuli in each consonant category (i.e., the same items used in the locus equation analyses above). Blumstein and Stevens (1979) described the distinctive features of the onset spectra of consonants: "In the case of velar consonants, the theoretically predicted common attribute of the spectrum is a major spectral prominence in the midfrequency range; for alveolar consonants, the spectral energy is diffuse or distributed throughout the frequency range, but with greater spectral energy at higher frequencies; when the consonant constriction is made at the lips, the spectral energy is again diffuse, but the spectrum is weighted toward the lower frequencies" (p. 1002).

The overall shapes of the short-time spectra for each consonant in the current study are quite consistent with these expected shapes. For most of the consonant–vowel combinations, the spectra fit the templates defined by Blumstein and Stevens (1979) in their study of consonant productions: labials exhibited a diffuse-falling spectrum, alveolars exhibited a diffuse-rising spectrum, and velars exhibited a compact spectrum with one prominent peak.

The notable exception to the overall pattern was the group of /d/ consonants in the context of /u/. For these stimuli, the higher-frequency peaks are not greater in amplitude than the lower-frequency peaks, and these spectra would be rejected by Blumstein and Stevens' criteria for the diffuse-rising template. It is interesting to note that the effect of initial vowel on "D" responses was particularly strong for stimuli in the series with /u/ vowel context. For these stimuli, the dependence of perception on the initial vowel may have been partly due to the relative lack of gross spectral cues for /d/ at consonant onset.

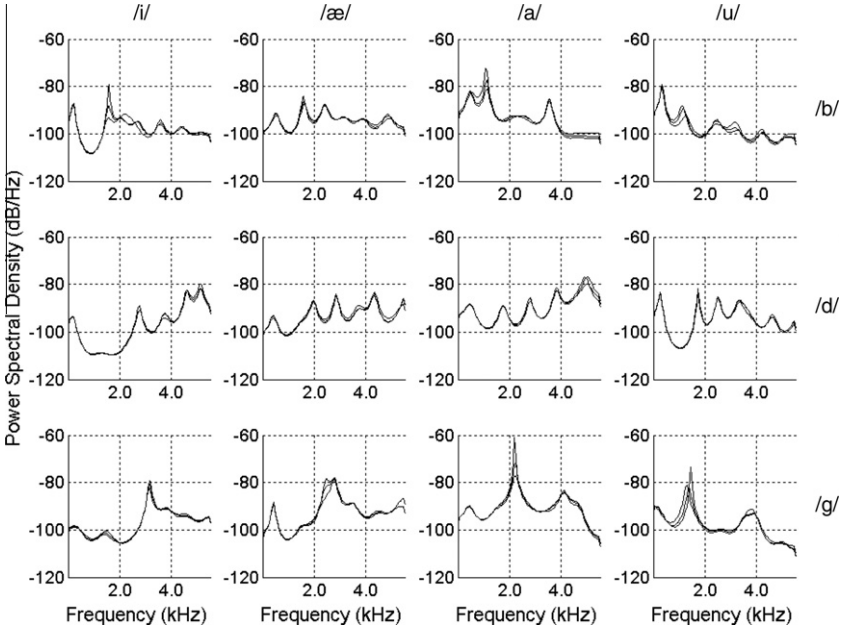

Fig. 5. Onset spectra for the three most extreme instances of each consonant category, for each of the four vowel contexts. Except for the /d/ stimuli with /u/ context, the spectra were consistent with the templates described by Blumstein and Stevens (1979).

4.3. Onset temporal characteristics

In addition to formant frequencies and onset spectra, place of articulation in stop consonants is related to certain temporal characteristics of consonant onsets. For instance, velar stops tend to have longer voice-onset times (VOTs) than labial stops (Lisker and Abramson, 1964; Kewley-Port, 1982; Stevens, 1998), and VOT can be used to predict differences in place of articulation in situations where formant frequencies are not very diagnostic (Engstrand et al., 2000). In the current study, the endpoints' burst onsets were aligned prior to morphing, and the same voicing source (from the /d/ endpoint) was used for all stimuli within each vowel context. These procedures eliminated any VOT differences that may have existed between the consonants in the original utterances. However, other temporal onset correlates of consonant place of articulation exist. For example, velar stops tend to exhibit a more gradual release of constriction than labials (Stevens, 1998). Thus, the amplitude of the acoustic waveform for velars tends to increase more slowly at consonant onset. Although this property of consonants is produced by different articulatory mechanisms than VOT, it has some of the same consequences: for instance, both longer VOT and more gradual release attenuate formant amplitudes at consonant onset. Differences in abruptness of consonant onset are also an interesting correlate of place of articulation regardless of VOT, so an attempt was made here to quantify these changes in abruptness across the LPC-interpolated stimuli.

It was possible to achieve a rough measure of how gradual the releases of the stops were by tracking the amplitudes of the waveforms over a brief period after consonant onset. The consonant release point for each vowel series was defined in the same manner as for the gross spectral shape measurements described above. For each of the 160 stimuli, the absolute value of its waveform was measured over the 25 ms period following the release point identified for the corresponding vowel series, and a simple linear regression was fit to the sample points within that window. The slope of the resulting line was taken as a measure of "onset velocity." Fig. 6 illustrates the methods used to obtain these values.

Fig. 7 displays the onset velocity measurements obtained from the consonants in each of the four vowel contexts. As expected, onset velocity generally showed an overall decrease from /b/ to /d/ to /g/, with some variations in each series. The primary exception to this overall pattern was the /a/ series between /d/ and /g/, in which onset velocity remained relatively constant. Each series fell within somewhat different ranges on the absolute scale of the onset velocity measure, but specific comparisons between the series are not warranted considering that different voicing sources were used in the synthesis of stimuli in the different vowel contexts. The measurement of onset velocity devised here was intended mainly as a relative measure to compare stimuli within the same vowel context.

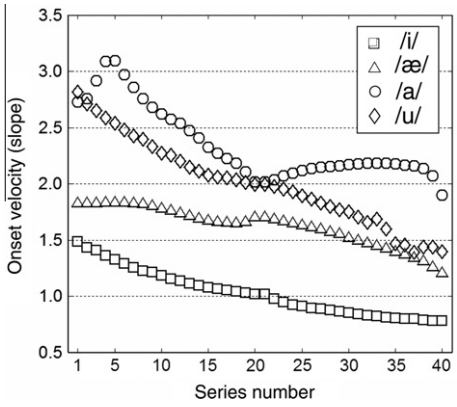

Fig. 7. Onset velocity measurements across each of the four stimulus series. Overall, onset velocity generally decreased from /b/ to /d/ to /g/, with the exception of the /a/ series between /d/ and /g/.

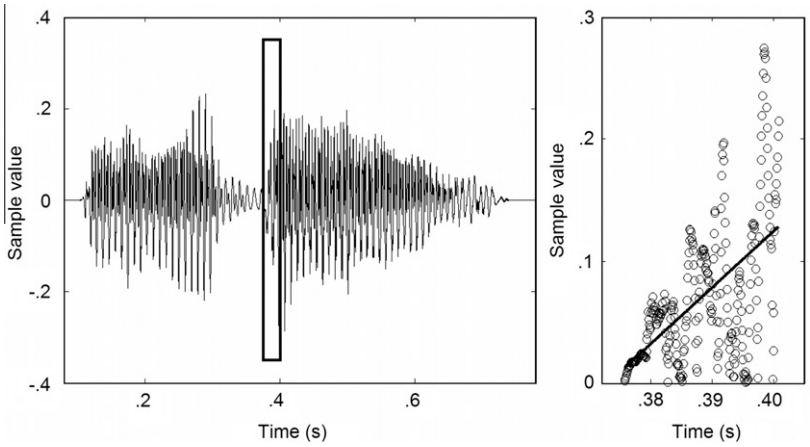

Fig. 6. Illustration of the procedure used to compute "onset velocity." A 25 ms segment of the waveform was sampled, starting at the time point corresponding to the last zero-crossing before release in the original source wave used in the LPC resynthesis. A linear regression was fit to the rectified sample values from this time window; the slope of the line was taken as the onset velocity.

4.4. Summary of acoustic analyses

Three general conclusions can be drawn from the acoustic analyses of the LPC-morphed stimuli described here:

(1) The LPC-morphing procedure produced gradual acoustic shifts along each series, across several acoustic parameters. Formant frequencies, onset velocities, and onset spectra were all incrementally modified via the LPC-interpolation procedure. A gradual manipulation of multiple acoustic dimensions was the objective of the current endeavor and accounts for the naturalistic quality of the stimuli. It should be noted that the stepwise changes in LPC filter parameters did not correspond to acoustic changes of equal magnitude. Thus, when using these stimuli, no assumptions should be made that steps along each series correspond to equal steps along any acoustic dimension. For reference, acoustic measurements of the formant frequencies of the stimuli are provided in the online archive (see Appendix A).

(2) The stimuli exhibited acoustic properties typical of voiced stop consonants. For most of the stimuli, all acoustic measurements fell within expected ranges based on previous research. The exceptions to this general observation were the points representing /ubu/, /idi/, and /ugu/ in locus equation space, and the gross spectral shape at consonant onset for /udu/, which did not exhibit as much energy at higher frequencies as expected from theory (Blumstein and Stevens, 1979). In addition, the onset velocity measurements did not monotonically decrease from /b/ to /d/ to /g/ across all four vowel contexts. Nonetheless, no stimulus exhibited gross acoustic abnormalities, and for those stimuli with slight anomalies in one acoustic dimension, normal characteristics were observed in the other acoustic dimensions studied. Further, the normal identification functions found in the perceptual experiment suggest that any perceptual difficulty that might be introduced by a minor irregularity in one acoustic property is made up for by other characteristics, and supports the notion that consonant perception is accomplished through the recognition of patterns across more than just one acoustic dimension (e.g., Engstrand et al., 2000).

(3) The effect of initial vowel observed in the perceptual experiment was probably due to acoustic information in the initial vowels that differed slightly from information in the consonant onsets. The locus equations computed from measurements of initial vowels revealed that they contained more /d/-like *F*2 information than the subsequent consonants. For many of the stimuli, formant frequencies in the initial vowels could also have shifted perception of consonant formants toward /d/ as a result of contrastive auditory effects.

5. General discussion

Interpolation of LPC coefficients for naturally-produced stop consonant utterances was used to create series of stimuli ranging perceptually from /b/ to /d/ and /d/ to /g/ in four vowel contexts and in both VCV and CV form. The purpose of this effort was to create natural-sounding, perceptually-ambiguous tokens that may be used by researchers for whom standard synthesis methods are either impractical or result in stimuli that are too acoustically impoverished for their purposes. The stimuli were shown to be reliably perceived by listeners as belonging to the three intended voiced stop-consonant categories in a 3-alternative forced-choice listening task. The VCV and CV stimuli were both found to be reliably identified by participants, although some differences in perception were observed as a function of the presence of the initial vowel. Acoustic analyses of the stimuli demonstrated that the LPC-interpolation procedure was successful in gradually shifting several acoustic properties of the original tokens and that the acoustic characteristics of the resulting stimuli were generally consistent with previous observations of naturally-produced utterances. These stimuli were created with the aim of providing researchers with a naturalistic consonant series varying along a place-of-articulation continuum. As such, the stimulus corpus is freely available for download, and readers interested in hearing or using the stimuli may see Appendix A for further information on accessing the materials via the Internet.

Acknowledgements

The authors thank Christi Adams Gomez and Tony Kelly for help with data collection and preparation of online materials, respectively. This work was supported by a National Research Service Award (1 F31 DC007284-01) from the National Institute on Deafness and Other Communication Disorders to J.D.W.S., by a grant from the National Science Foundation (BCS-0345773) to L.L.H., and by the Center for the Neural Basis of Cognition.

Appendix A. Instructions for downloading and use of stimuli

The full set of 320 stimuli is available on the Internet (with accompanying documentation) at http://www.psy.cmu.edu/~lholt/php/StephensHoltStimuli.php. The files named "VCV_stimuli.zip" and "CV_stimuli.zip" each contain eight 20-member morphed series, arranged hierarchically within directories. The directories corresponding to each vowel in "VCV_stimuli.zip" also include the original recorded utterances as well as the edited versions of the recordings and source waveforms used as input in the morphing procedure. The file named "all_stimuli.zip" contains all 320 stimuli and associated files within a single directory. Finally, the file named "formant_frequencies.pdf" contains the table of measured formant frequencies.

Researchers are advised to refer to this report and the online documentation for important information regarding the acoustic and perceptual properties of the stimuli. The stimuli may also be obtained via electronic mail (jdstephe@ncat.edu) or in CD format by postal mail addressed to Joseph Stephens, Psychology Department, 1601 E. Market St., Greensboro, NC 27411.

References

Atal, B.S., Hanauer, S.L., 1971. Speech analysis and synthesis by linear prediction of the speech wave. J. Acoust. Soc. Amer. 50, 637–655.

Blumstein, S.E., Stevens, K.N., 1979. Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants. J. Acoust. Soc. Amer. 66, 1001–1017.

Blumstein, S.E., Stevens, K.N., 1980. Perceptual invariance and onset spectra for stop consonants in different vowel environments. J. Acoust. Soc. Amer. 67, 648–662.

Boersma, P., 2001. PRAAT, a system for doing phonetics by computer. Glot Int. 5, 341–345.

Elman, J.L., McClelland, J.L., 1988. Cognitive penetration of the mechanisms of perception: compensation for coarticulation of lexically restored phonemes. J. Memory Lang. 27, 143–165.

Engstrand, O., Krull, D., Lindblom, B., 2000. Sorting stops by place in acoustic space. In: Proc. XIIIth Swedish Phonetics Conf. (FONETIK 2000), Skövde, Sweden, pp. 53–56.

Fant, G., 1960. Acoustic Theory of Speech Production Mouton. The Hague, The Netherlands.

Ganong III, W.F., 1980. Phonetic categorization in auditory word perception. J. Exp. Psychol. Human Percept. Perform. 6, 110–125.

Holt, L.L., 1999. Auditory Constraints on Speech Perception: An Examination of Spectral Contrast. Unpublished doctoral dissertation. Springer, Berlin.

Holt, L.L., 2005. Temporally non-adjacent non-linguistic sounds affect speech categorization. Psychol. Sci. 16, 305–312.

Holt, L.L., Lotto, A.J., 2002. Behavioral examinations of the neural mechanisms of speech context effects. Hear. Res. 167, 156–169.

Kewley-Port, D., 1982. Measurement of formant transitions in naturally produced stop consonant–vowel syllables. J. Acoust. Soc. Amer. 72, 379–389.

Klatt, D.H., 1980. Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Amer. 67, 971–995.

Klatt, D.H., Klatt, L.C., 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Amer. 87, 820–857.

Krull, D., 1988. Acoustic properties as predictors of perceptual responses: a study of Swedish voiced stops. In: Phonetic Experimental Research at the Institute of Linguistics, University of Stockholm (PERILUS), Vol. VII, pp. 66–70.

Liberman, A.M., Harris, K.S., Hoffman, H.S., Griffith, B.C., 1957. The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. 54, 358–368.

Lindblom, B., 1963. On Vowel Reduction, Report #29. The Royal Institute of Technology, Speech Transmission Laboratory, Stockholm, Sweden.

Lisker, L., Abramson, A.S., 1964. A cross-language study of voicing in initial stops: acoustical measurements. Word 20, 384–422.

Lotto, A.J., Holt, L.L., 2006. Putting phonetic context effects into context: a commentary on Fowler (2006). Percept. Psychophys. 68, 178–183.

Lotto, A.J., Kluender, K.R., 1998. General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. Percept. Psychophys. 60, 602–619.

Mann, V.A., 1980. Influence of preceding liquid on stop-consonant perception. Percept. Psychophys. 28, 407–412.

Markel, J.D., Gray Jr., A.H., 1976. Linear Prediction of Speech. Springer-Verlag, New York.

Massaro, D.W., 1987. Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Lawrence Erlbaum Associates, Hillsdale, NJ.

Massaro, D.W., 1998. Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. MIT, Cambridge, MA.

McCandliss, B.D., Fiez, J.A., Protopapas, A., Conway, M., McClelland, J.L., 2002. Success and failure in teaching the [r]-[l] contrast to Japanese adults: tests of a Hebbian model of plasticity and stabilization in spoken language perception. Cogn. Affect. Behav. Neurosci. 2, 89–108.

McQueen, J.M., 1991. The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. J. Exp. Psychol. Human Percept. Perform. 17, 433–443.

Newman, R.S., Sawusch, J.R., Luce, P.A., 1997. Lexical neighborhood effects in phonetic processing. J. Exp. Psychol. Human Percept. Perform. 23, 873–889.

Öhman, S.E.G., 1966. Coarticulation in VCV utterances: spectrographic measurements. J. Acoust. Soc. Amer. 39, 151–168.

Pfitzinger, H.R., 2004. Unsupervised morphing between utterances of any speakers. In: Proc. 10th Australian Internat. Conf. on Speech Science and Technology, Sydney, Australia, pp. 545–550.

Pitt, M.A., McQueen, J.M., 1998. Is compensation for coarticulation mediated by the lexicon? J. Memory Lang. 39, 347–370.

Slaney, M., Covell, M., Lassiter, B., 1996. Automatic audio morphing. In: Proc. 1996 IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP96), Atlanta, GA, pp. 1001–1004.

Stephens, J.D.W., Holt, L.L., 2010. Learning to use an artificial visual cue in speech identification. J. Acoust. Soc. Amer. 128, 2138–2149.

Stevens, K.N., 1998. Acoustic Phonetics. MIT, Cambridge, MA.

Stevens, K.N., Blumstein, S.E., 1978. Invariant cues for place of articulation in stop consonants. J. Acoust. Soc. Amer. 64, 1358–1368.

Sussman, H.M., McCaffrey, H.A., Matthews, S.A., 1991. An investigation of locus equations as a source of relational invariance for stop place categorization. J. Acoust. Soc. Amer. 90, 1309–1325.

Sussman, H.M., Bessell, N., Dalston, E., Majors, T., 1997. An investigation of stop place articulation as a function of syllable position: a locus equation perspective. J. Acoust. Soc. Amer. 101, 2826–2838.