Acoustic properties and perception of stop consonant release transients[a)]

656

Bruno H. Repp
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511-6695

Hwei-Bing Lin
Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06511-6695 and Department of Linguistics, University of Connecticut, Storrs, Connecticut 06268

(Received 12 February 1988; accepted for publication 27 July 1988)

This study focuses on the initial component of the stop consonant release burst, the release transient. In theory, the transient, because of its impulselike source, should contain much information about the vocal tract configuration at release, but it is usually weak in intensity and difficult to isolate from the accompanying frication in natural speech. For this investigation, a human talker produced isolated release transients of /b,d,g/ in nine vocalic contexts by whispering these syllables very quietly. He also produced the corresponding CV syllables with regular phonation for comparison. Spectral analyses showed the isolated transients to have a clearly defined formant structure, which was not seen in natural release bursts, whose spectra were dominated by the frication noise. The formant frequencies varied systematically with both consonant place of articulation and vocalic context. Perceptual experiments showed that listeners can identify both consonants and vowels from isolated transients, though not very accurately. Knowing one of the two segments in advance did not help, but when the transients were followed by a compatible synthetic, steady-state vowel, consonant identification improved somewhat. On the whole, isolated transients, despite their clear formant structure, provided only partial information for consonant identification, but no less so, it seems, than excerpted natural release bursts. The information conveyed by artificially isolated transients and by natural (frication-dominated) release bursts appears to be perceptually equivalent.

PACS numbers: 43.71.Es, 43.70.Fq, 43.70.Aj

INTRODUCTION

When the articulatory closure of an utterance-initial English stop consonant is released, a brief burst of noise is commonly generated before voicing starts. This *release burst* has been the focus of attention in a number of acoustic and perceptual studies.

Although the articulators not directly involved in the occlusion typically assume positions for the following vowel (or other segment) before the closure is released, the release burst provides, nevertheless, the most direct acoustic manifestation of consonant place of articulation. [The closure period itself is either silent or filled with low-amplitude voicing that conveys some, but not very salient, place of articulation information; see Barry (1984).] Following earlier acoustic analyses by Halle *et al.* (1957), and Zue (1976), among others, Stevens and Blumstein (1978) claimed that the spectrum computed at the stop consonant release contains invariant (i.e., context-independent) properties reflecting the different places of articulation, and they supported this claim with both acoustic analyses (Blumstein and Stevens, 1979) and perceptual data (Blumstein and Stevens, 1980). The onset spectra of English /b,d,g/ preceding various vowels were computed over a 26-ms time window that usually included the beginning of voicing as well as some formant movement. The typical labial, alveolar, and velar onset spectra (with high-frequency pre-emphasis) were characterized as diffuse falling or flat, diffuse rising, and compact, respectively. Visual classification of onset spectra according to these global characteristics was not perfect, however—about 85% correct (Blumstein and Stevens, 1979). Although these authors emphasized the relative sufficiency of static spectral properties for identifying place of articulation, more recent sudies have focused on their relative insufficiency and on the need to supplement them with information about dynamic spectral change (Kewley-Port, 1983; Lahiri *et al.*, 1984; Suomi, 1985). Even so, classification accuracy in these studies improved by only a few percentage points.

Several perceptual studies have investigated human listeners' ability to identify stop consonant place of articulation from the release burst, presented either in isolation or followed by some vocalic context with or without formant transitions. Some of these studies used synthetic speech in which the bursts had a single spectral peak and, therefore, may not have approximated the information content of natural release bursts (cf. Syrdal, 1983). Nevertheless, classic research at Haskins Laboratories (Liberman *et al.*, 1952) has shown not only that such synthetic noise bursts provide suf-

[a)] Portions of this article were presented at the 115th Meeting of the Acoustical Society of America in Seattle, WA [J. Acoust. Soc. Am. Suppl. 1 **83**, S67 (1988)].

 0001-4966/89/010379-18$00.80

ficient cues for some stop consonants preceding some steady-state vowels, but also that identical bursts often lead to different place of articulation percepts in the context of different vowels. The perceptual contribution of various synthetic release bursts in combination with vowels containing formant transitions was investigated later by Hoffman (1958). More recently, Blumstein and Stevens (1980) compared place of articulation perception in synthetic CV syllables with and without bursts intended to be optimal for /b,d,g/ and found that the burst enhanced identification accuracy, especially for /g/ and /d/. [However, Syrdal (1983) has pointed out that this comparison was confounded with differences in overall stimulus duration.] Their study included stimuli consisting only of the burst plus a single glottal pulse, whose identification was almost as good as that of full synthetic syllables (about 85% correct overall), except for /gi/, which required longer stimulus durations.

More directly relevant to the present investigation are perceptual studies that used bursts edited from natural speech. Some authors (Halle *et al.*, 1957; Malécot, 1958; Winitz *et al.*, 1972) employed released VC syllables, so that the burst occurred in utterance-final position and essentially in the context of a neutral following vowel. Listeners' identification of place of articulation from such release segments presented in isolation was in the vicinity of 80% correct. The releases were also found to provide important cues to place identification in full VC syllables (Malécot, 1958). Other researchers (Winitz *et al.*, 1972; Cole and Scott, 1974; Just *et al.*, 1978) presented listeners with natural-speech "bursts" that, according to a liberal definition, included both the release burst proper and all the following aspiration of initial voiceless aspirated stops. Intelligibility of these fairly long stimuli ranged from about 90% correct when cross spliced onto different vowels (Cole and Scott, 1974) to about 75% correct in isolation or with 100 ms of the original vowel following (Winitz *et al.*, 1972) to 83% correct in combination with steady-state vowels (Just *et al.*, 1978). When Just *et al.* (1978) replaced the aspiration with silence, listeners' performance fell to 61% correct. (See, also, Schatz, 1954.) Dorman *et al.* (1977) cross spliced bursts of voiceless unaspirated stops among different vocalic contexts and showed that the contribution of the burst to place of articulation perception was inversely related to the contribution of the vocalic formant transitions. They also prefixed VC syllables with initial bursts excerpted from CVC syllables and showed that, in general, the burst alone was not sufficient to permit very accurate initial consonant identification. They did not present the release bursts in isolation, however. This was done by Ohde and Sharf (1977) who obtained nearly perfect identification for stops from /i,u,ɔ/ contexts. A gating technique was employed by Tekieli and Cullinan (1979) with a variety of natural CV syllables. Identification of stop consonant place of articulation from the initial 10 ms was only about 60% correct. In a similar gating experiment employing only /b,d,g/ in five vocalic contexts, Kewley-Port *et al.* (1983) obtained a higher score of 83% correct for the initial 10-ms segments.

These various studies display considerable variation of identification scores reflecting different procedures and stimulus materials. Nevertheless, they show that stop consonant release bursts are perceptually important and contain considerable, though not entirely sufficient, place of articulation information even in isolation. They also contain information about the following vowel, however. Acoustic analyses (e.g., Zue, 1976) have shown that the energy maximum in release burst spectra varies systematically with the following vowel. Velar bursts, in addition, exhibit a duality of acoustic structure reflecting the relative shift in place of articulation preceding front and back vowels (see also Dorman *et al.*, 1977; Suomi, 1985). Winitz *et al.* (1972) showed that listeners could identify /i,a,u/ with good accuracy from the full aperiodic portion of voiceless aspirated natural stops. Ohde and Sharf (1977) also included voiceless unaspirated stops and found 73% correct /i,u,ɔ/ identification. Working with a larger sample of vowels following natural aspirated stops, Cullinan and Tekieli (1979) showed that the gated initial 10–30 ms of the syllables contained much information about the front–back vowel distinction, less about vowel height, and least about the tense–lax feature. Release bursts thus typically convey both consonant and vowel information due to anticipatory coarticulation of the vowel with the consonant.

In all of this research, release bursts were treated as unitary acoustic segments corresponding to the aperiodic portion of initial stop consonants. Fant (1960, 1973), however, had noted long ago that they consist of three distinguishable phases: an initial *transient*, a *fricative segment*, and an *aspirative segment*. The transient represents the response of the vocal tract to the impulse of the sudden pressure release; the fricative segment results from turbulence generated at the constriction while it is still narrow; and the aspirative segment reflects a glottal noise source that replaces the frication as the constriction is widened. These phases overlap in natural speech and cannot be easily separated in the waveform. Existing methods of acoustic analysis may also lack the required temporal resolution, though the new method of Wigner distribution analysis is promising in that regard (see Garudadri *et al.*, 1986; Wokurek *et al.*, 1987).

It is likely that, even within the short time span of a 10- to 30-ms release burst, there is considerable acoustic structure and dynamic spectral change. As the oral constriction widens rapidly, the release burst becomes increasingly "vocalic" and less "consonantal." The most consonantal part, then, should be the initial *release transient*, which represents the impulse response of the vocal tract to the sudden pressure release (the "explosion"). As Maeda (1987) has pointed out, this sound has a "coherent source" (i.e., its waveform and spectrum can be predicted precisely if the acoustic and aerodynamic conditions are known), as opposed to the stochastic noise sources of frication and aspiration. According to acoustic theory, the spectrum of the transient should contain peaks reflecting the vocal tract resonances immediately after the release. Since the excitation is so brief and uniform, the transient provides essentially an acoustic snapshot of the vocal tract. The sound will take a certain time to decay, however, and the initial widening of the constriction may be reflected in spectral changes during the decaying portion of the

transient. Nevertheless, if more extensive spectral change is needed for a listener to identify stop consonant place of articulation accurately (Kewley-Port, 1983; Kewley-Port *et al.*, 1983; Lahiri *et al.*, 1984), the transient by itself may not provide sufficient information. On the other hand, to the extent that static onset spectra have characteristic properties associated with different places of articulation (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979, 1980), release transients might contain these properties in purer form than do natural release bursts.

In natural speech, the transients are probably not only of relatively low amplitude but are almost immediately overlaid by the frication generated at the constriction, so that especially the decaying part of the transient is masked completely. Typical stop consonant onset spectra derived from natural speech (Blumstein and Stevens, 1979; Kewley-Port, 1983) show broad peaks that probably reflect this noise overlay, as well as spectral smearing due to dynamic change of formant frequencies. If dynamic change is minimized by choosing a shorter time window for conventional Fourier analysis (Kewley-Port, 1983), spectral resolution is sacrificed, and the peaks remain broad. In fact, release burst spectra have rarely been characterized in terms of formant frequencies in the literature, but rather in terms of global spectral properties. However, formant peaks may be expected to be more sharply defined in the vocal tract impulse response. This makes the acoustic and perceptual properties of release transients theoretically interesting. If transients could somehow be isolated, they might provide excellent information about the state of the vocal tract immediately after the release, especially with regard to the extent of its preparation for the following vowel (i.e., anticipatory coarticulation). Such information would be useful for acoustic and articulatory modeling of vocal tract dynamics, and the question about its potential perceptual value could be raised, regardless of whether the transient plays a prominent role in natural speech.

The purpose of the present study was to investigate these issues using stop consonant release transients that had been *produced in isolation*. This was accomplished by a human talker producing a variety of CV syllables in a very quiet whisper so that the frication and aspiration sources were attenuated to virtual silence and only the initial transient remained. The recorded transients were analyzed acoustically to describe the consonant and vowel information contained in them, and they were presented to listeners in isolation and in vocalic context to assess the perceptual usefulness of their acoustic structure. The isolated transients were assumed to be equivalent to those occurring in natural speech, though the latter may be weaker and masked by frication.

I. ACOUSTIC ANALYSES

A. Methods

1. Materials and recording procedures

The phonetic materials were a set of 27 CV syllables: the three voiceless unaspirated stop consonants /b,d,g/ preceding the nine vowels /i,e,ɛ,æ,ɑ,ɔ,o,u,ɝ/. The talker was the senior author, a native speaker of German from Vienna who has been virtually monolingual in English for the last 18 years but still speaks with an accent. It was not considered essential for this study that the vowels be prototypically American English. In fact, while the talker intended /i,ɛ,æ,ɑ,ɔ,u,ɝ/ to be American English in quality, he intended /e,o/ to be the German monophthongs of that name.

Each of the 27 syllables was whispered very quietly ten times in succession (e.g., /bibibi . . . /) without any frication or aspiration so that only audible release transients were emitted. Thus the speaker's glottis was open, as in regular whisper, but the airflow was very low, just sufficient to generate a certain minimal pressure behind the constriction, but not enough to cause audible noise following the transient. Each series of transients was preceded by a fully phonated syllable of the same type, and the articulatory movements of the whispered syllables were intended to be identical to those of the precursor syllables. The transients were picked up by a Sennheiser microphone placed about 10 cm from the talker's mouth and slightly off to the right side so as to avoid artifacts due to air puffs hitting the microphone. The recording took place in a sound-insulated booth. The tape recorder (Otari MX-5050) was located outside. The syllables including the vowel /ɔ/ were originally omitted and recorded at a later date.

Because of the close placement of the microphone and the high recording level appropriate for the faint transients, the precursor syllables were distorted and not usable for acoustic analysis. To have some natural productions available for comparison, the author recorded the same syllables with natural phonation at a later occasion. A more distant microphone placement (about 30 cm away) and lower recording level were used. The series of 27 syllables was produced five times.

2. Acoustic analysis methods

From the ten recorded tokens of each transient, five from the middle of the series were selected for analysis. They, and the phonated CV syllables, were sampled at a rate of 10 kHz and stored in separate computer files. Using a waveform editor, silence preceding the onset of energy was deleted, and the duration of all transient files was set at 25.6 ms, which accommodated the signal in all cases. Amplitude measurements and spectral analyses were conducted using programs of ILS (Interactive Laboratory System, distributed by Signal Technology, Inc.) and software developed at Haskins Laboratories. To describe the amplitude envelope of the transients, each file was subdivided into eight 3.2-ms sections whose root-mean-square (rms) amplitude was then calculated. For the purpose of spectral analysis, each transient file was divided into three overlapping 12.8-ms sections, whose Fourier spectra were calculated using a full Hamming window of the same width. (Other types of windows were tried and found to yield poorer spectral resolution; it seemed important not to give full weight to the abrupt initial spike.) The initial 25.6 ms of the phonated CV syllables were analyzed in a similar fashion.

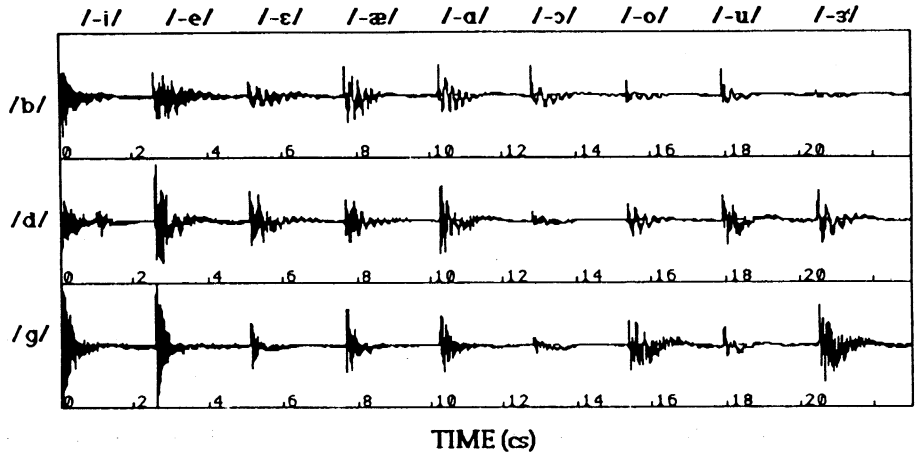

FIG. 1. Oscillograms of single tokens of isolated transients for the 27 consonant–vowel combinations. For purposes of this display, all transients for the same consonant have been concatenated. Each transient file is 25.6 ms in duration.

B. Results

1. Amplitude envelopes

a. Isolated transients. Figure 1 shows oscillograms of one token of each of the 27 CV transients. The selected tokens for each consonant place of articulation were concatenated for this display. The figure illustrates the sharp onset and rapid decay of each signal, as well as the absence of voicing or significant frication. Most transients seem to have a slowly varying waveform component of about 40 Hz, whose origin is not clear. Vowel or consonant differences visible in this figure should not be interpreted, as the selected tokens may not be representative.

The amplitude envelope of each transient signal was quantized in terms of eight decibel values, representing successive 3.2-ms time frames. These data were submitted to a three-way analysis of variance, with the factors consonant (three levels), vowel (nine levels), and time frame (eight levels), with tokens nested within the CV combinations as the random variable. From the initial analysis, it emerged that /-ɔ/ tokens which had been recorded separately, were about 10 dB higher in absolute level than the other signals; therefore, the analysis was repeated after attenuating these signals by 10 dB. (This correction was also applied in Fig. 1.) All main effects and interactions in the analysis of variance were highly significant ($p<0.0001$). Since it was thought that some of the interactions may have derived from a floor effect in late time frames (all signals falling to a common decibel baseline reflecting ambient noise), the analysis was repeated for the initial three time frames only. However, the results were quite similar. These statistical findings imply that there were reliable differences among consonants, vowels, and CV combinations with regard to overall amplitude as well as to their amplitude contour over time (i.e., the decay of the transient).

Differences in overall amplitude must be regarded with caution because the transients were produced in a fixed sequence, so that talker fatigue or changes in the mouth-to-microphone distance could have had a systematic effect on relative signal amplitudes. Nevertheless, the differences follow a reasonable pattern. This pattern is shown in Fig. 2 as a function of consonant and vowel. The intensity of labial release transients was, on the average, 3 dB lower than that of alveolar and velar transients. However, this difference derived almost entirely from rounded back vowel contexts (/o/, /u/, and /ɜ/), where labials were as much as 10 dB lower. In the context of front vowels, labial transients were as intense as alveolar and velar ones. As to the effect of vowel context itself, average transient intensity decreased from front to back vowels (excluding /ɜ/), but there were large differences among consonants in the context of /o/, /u/, and /ɜ/. In terms of overall amplitude, /bɜ/ was at the low end of the /b-/ series, while /dɜ/ and /gɜ/ were the strongest signals among the /d-/ and /g-/ syllables. An unusually strong signal was also associated with /go/.

The average transient amplitude envelopes for the three consonant places of articulation are shown in Fig. 3. As expected, all transients had their highest amplitude right at onset. The average amplitude decrease from the first to the last 3.2-ms time frame was 28 dB. The slope of the amplitude envelope (in decibels) was almost linear except for a leveling off between time frames 4 and 5. The average signal thus decayed exponentially within about 25 ms; the step in the

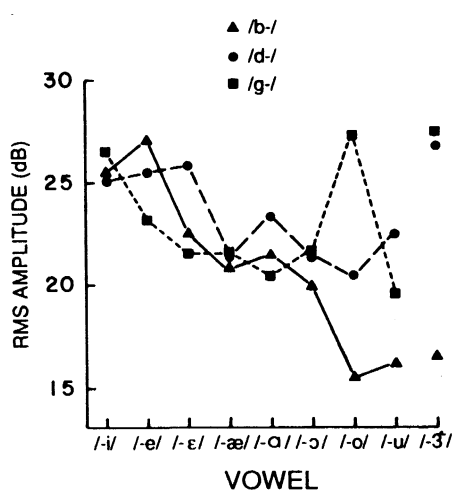

FIG. 2. Transient amplitude (in dB, *re*: an arbitrary reference) as a function of consonant and vowel, averaged over tokens and time (25.6 ms).

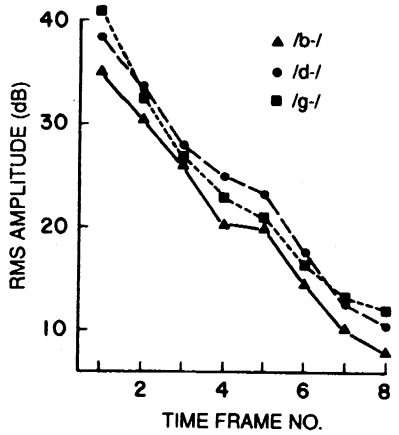

FIG. 3. Transient amplitude as a function of consonant and time (eight 3.2-ms time frames), averaged over vowels and tokens.

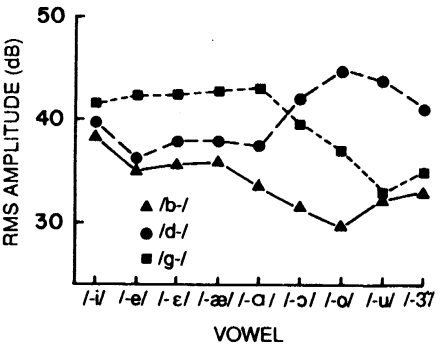

FIG. 5. Average onset amplitude (first 9.6 ms only) of CV syllable release bursts as a function of consonant and vowel, averaged over tokens.

middle may have to do with the slowly varying waveform component (cf. Fig. 1), which crossed the zero line at about that point. Velar releases showed a somewhat steeper initial decay than did alveolars and labials, and (not shown in Fig. 3) high vowels showed a steeper initial decay than low vowels. These interactions were rather small and were statistically significant mainly because of the very small error variance.

b. Natural release bursts. Figure 4 shows the waveforms of the concatenated onsets (the initial 25.6 ms) of full CV syllables. The boundaries between the selected tokens are marked by vertical lines. Each onset includes the release burst and the onset of voicing, except for a few velar tokens whose bursts exceeded 25.6 ms in duration. Multiple release bursts were common for alveolars and velars in back vowel contexts and were invariably present for velars in front vowel contexts. Only the latter bursts bear some resemblance in waveform to isolated transients (cf. Fig. 1), but their amplitude is much higher, due to the higher air pressure in production. (This is not evident in comparing Figs. 1 and 4 because of the different recording levels.)

The amplitude data for the initial 25.6 ms of the full CV syllables were analyzed statistically in the same way as the transient data. In a first analysis, including all eight time frames, all effects were significant at $p < 0.0001$. In a second analysis, including only the first three time frames, the vowel by time frame interaction was nonsignificant, and the vowel main effect was only marginally significant ($p = 0.026$); all other effects remained highly significant.

The corresponding data are plotted in Figs. 5 and 6. The absolute amplitudes in Figs. 5 and 6 should not be compared with those in Figs. 2 and 3 because of the different recording levels. Figure 5 shows release burst amplitudes averaged over the first three time frames only, so as not to include the beginning of voicing. (Even so, some labial tokens may have included the beginning of the first glottal pulse.) It is evident that labials had lower amplitudes than alveolars and velars (by 6 dB, on the average), though they did not differ much from alveolar bursts preceding front vowels and from velar bursts preceding /u/ and /ɜ/. For front vowels, velar release bursts were more intense than alveolar ones; for back vowels, the situation was reversed.

Figure 6 shows that the amplitude of labial release bursts decreased sharply over the initial two time frames, whereas that of alveolar and velar bursts decreased more gradually over the initial four or five time frames, probably due to the stronger fricative noise component and the presence of multiple bursts. Note that multiple bursts never oc-

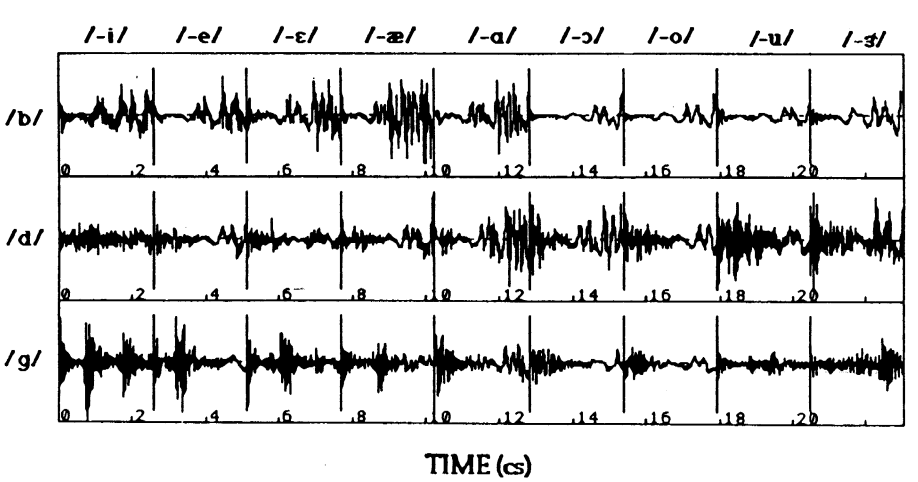

FIG. 4. Oscillograms of single tokens of release bursts (the initial 25.6 ms) of natural consonant–vowel syllables. For the purpose of this display, all waveforms for the same consonant have been concatenated; vertical lines mark the boundaries.

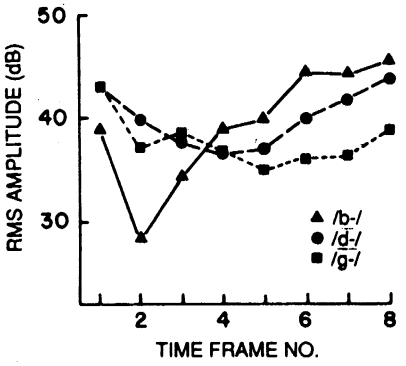

FIG. 6. CV release burst amplitude as a function of consonant and time (eight 3.2-ms time frames), averaged over vowels and tokens.

curred in the isolated transients; their initial amplitude decrease, therefore, was much steeper than that of CV syllable release bursts, except for labials (cf. Fig. 3). Following these decreases, the amplitudes of the CV syllable onsets increased towards the vowel, most rapidly for labials and most slowly for velars. These differences, and the corresponding amplitude differences among the three places of articulation in time frames 5–8, reflect the different average voice onset times (VOTs) of the tokens. The VOT (the time from energy onset to the onset of the first detectable glottal pulse) of each token was measured in oscillograms. Average values were 10 ms for labials, 16 ms for alveolars, and 23 ms for velars. These values are in good agreement with the literature (e.g., Zlatin and Koenigsknecht, 1976; Dorman *et al.*, 1977; Ohde, 1984) and with the amplitude envelopes plotted in Fig. 6.

2. Spectral properties

a. Isolated transients. A display of transient spectra is provided in Fig. 7. Each spectrum represents the linear average (in dB) of the five token spectra obtained from the initial 12.8 ms using a full Hamming window over that interval. The five token spectra were highly similar. Although better spectral resolution could have been obtained by applying a 25.6-ms Hamming window to the complete transient, the shorter window was used for comparison with the CV release burst spectra, where exclusion of the onset of voicing was desirable.

What is most notable about these spectra is the clear definition (i.e., relatively narrow bandwidth) of formant peaks. All spectra have a fixed peak at the very low end of the frequency scale, which presumably reflects the unexplained, slowly varying component in the waveforms (cf. Fig. 1). All spectra also have a clear first formant ($F1$) between 300 and 600 Hz, which probably indicates that the speaker's glottis was rather narrow. The $F1$ frequency varies inversely with vowel height, as it does in real vowels. Most spectra have a prominent second formant ($F2$), between 800 and 2000 Hz, whose frequency decreases with increasing vowel backness, as it should. There are also clear effects of consonant place of articulation on $F2$, which will be discussed below. Labials in back vowel contexts (lower left-hand panels in Fig. 7) exhib-

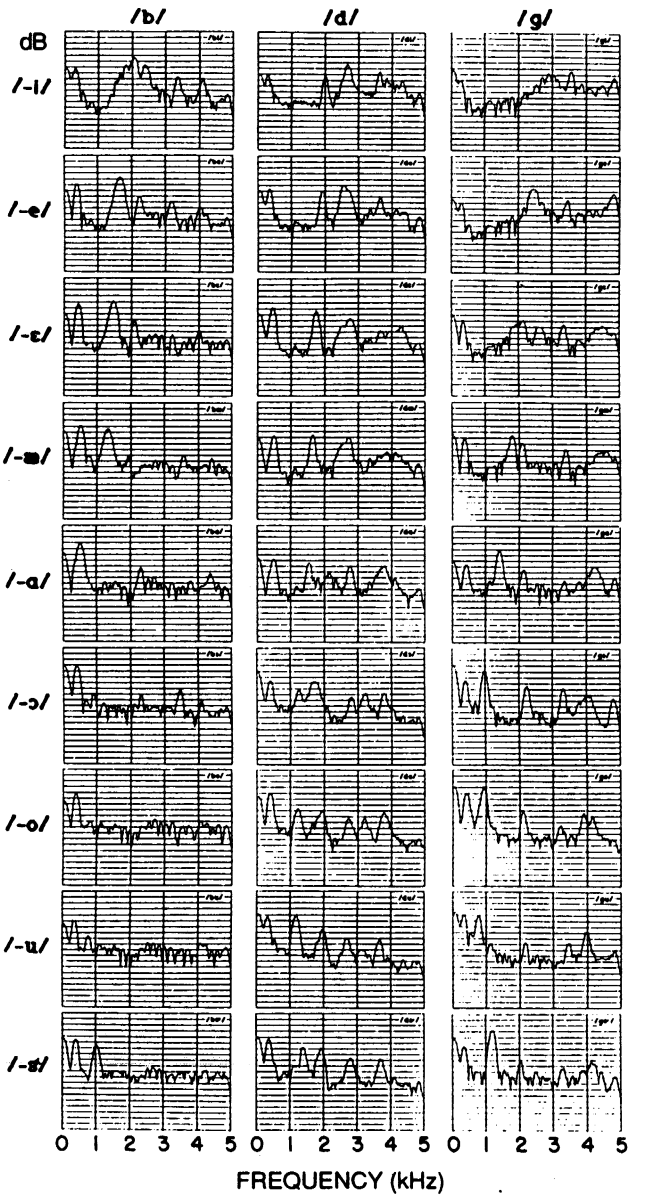

FIG. 7. Average transient onset spectra, computed from a full Hamming window over the first 12.8 ms, averaged over tokens. All spectra are amplitude normalized and include high-frequency pre-emphasis of about 6 dB per octave above 1 kHz (less below). Each horizontal division corresponds to 5 dB.

it a very weak $F2$, and, for velars in high front vowel contexts (upper right-hand panels), $F2$ appears to be entirely absent (or else abnormally high and merged with $F3$). The pattern of the third formant ($F3$), ranging from 1700 to 3000 Hz, is more complicated because it differs for front and back vowels. In front vowel contexts, $F3$ increases with vowel height for all three places of consonant articulation; in alveolars, $F3$ seems to merge with $F4$ into a single peak. In back vowel contexts, $F3$ is relatively fixed for all three consonant places of articulation, but it disappears for labials in high back vowel contexts. In front vowel contexts, labials have lower $F3$ frequencies than alveolars and velars, whereas, in back vowel contexts, alveolars exhibit a lower $F3$ than do labials and

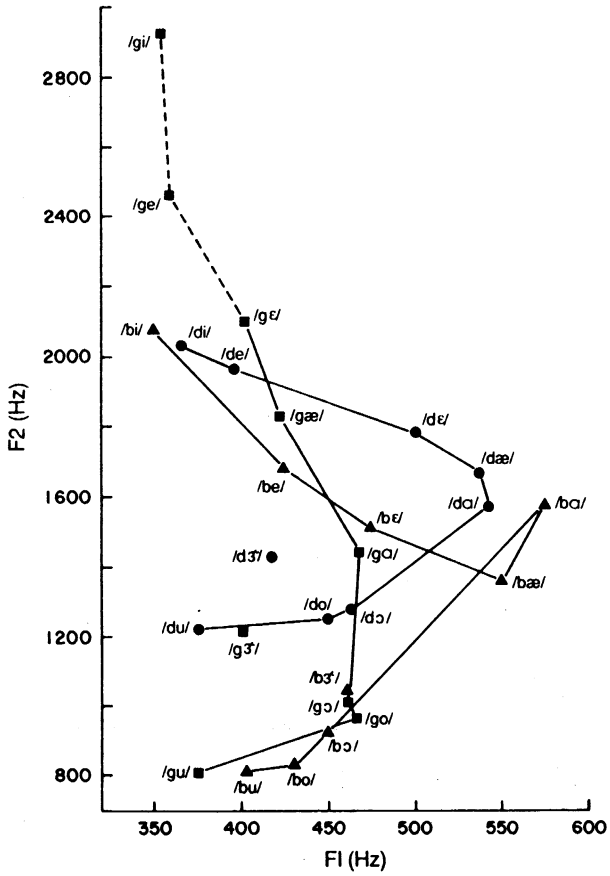

FIG. 8. First and second formant frequencies in transient onset spectra (cf. Fig. 7) as a function of consonant.

velars. A relatively constant fourth formant ($F4$) is visible in all spectra, except for labials in high back vowel contexts. Its frequency is much lower in alveolars (around 2800 Hz) than in labials and velars (3200–3500 Hz). A relatively invariant fifth formant ($F5$) can also be found in most spectra, except for labials in high back vowel contexts and for velars in front vowel contexts. Its frequency is again lower for alveolars (around 3800 Hz) than for labials and velars (around 4000 Hz). Alveolars in back vowel contexts show an additional peak between $F4$ and $F5$, at about 3200 Hz.

To clarify the pattern of the lower spectral peaks, the average frequencies of $F1$ and $F2$ for the 27 consonant–vowel combinations are plotted in Fig. 8. (For /gi/ and /ge/, the second spectral peak is probably $F3$, as indicated by the dashed line.) The points for each consonant place of articulation have been connected, proceeding from /-i/ to /-u/, excluding /-ɜ/. These lines describe a "vowel triangle" in $F1$–$F2$ space for each consonant place of articulation and thus portray the joint influence of consonant place of articulation and vocalic context on $F1$ and $F2$. It is evident that the principal difference between labials and alveolars is in $F2$ frequency, which is higher for alveolars, resulting in an upward shift of the vowel triangle. Exceptions are /bi/ and /di/, which differ very little, and /bɑ/ and /dɑ/, which seem to be distinguished by $F1$ rather than $F2$ frequency. The vowel triangle for velars is rather different in shape, in accordance with the allophonic change in place of occlusion between front and back vowel contexts. Velars preceding low back vowels resemble labials in these contexts, but, starting already with /gɑ/, front vowel height has a much more dramatic effect on $F2$ frequency than it does for labials and alveolars. This effect is enhanced by the fact that what is plotted as $F2$ for /gi/ and /ge/ is probably $F3$; $F2$, expected to be around 2000 Hz (Fant, 1973), appeared to be absent in these spectra (see Fig. 7).

It is evident that most syllables can be distinguished by their $F1$ and $F2$ frequencies alone. Those syllables that are very similar in their $F1$ and $F2$ frequencies can be distinguished by higher frequency spectral properties, as a glance back at Fig. 7 will confirm. Thus /bi/ and /di/ transients differ in the amplitude ratio of $F2/F3$ and $F2/F4$, with the alveolar spectrum having stronger high-frequency components, and labials differ from velars in high back vowel contexts in that they lack higher formants.

The $F2$ variation among alveolar transients as a function of vocalic context is larger than the variation of $F2$ onset frequencies following stop releases reported in the literature (Stevens *et al.*, 1966; Fant, 1973; Kewley-Port, 1983). This may imply that a rapid change in $F2$ occurs between the transient and the onset of voicing in natural utterances. Another possibility is that it reflects articulatory adjustments made by the present speaker in order to produce audible isolated transients without accompanying frication, which was more difficult to achieve for alveolars than for labials and velars.

To show the influence of vocalic context more clearly, Fig. 9 replots the same data as Fig. 8, but with connected points representing the same vowel context. The resulting small triangles outline the large vowel triangle in $F1$–$F2$ space. There is great variability in the shapes of the small triangles, signifying that the relationship among the three places of consonant articulation is not constant but varies from vowel to vowel. A comparison of the average transient formant frequencies for each vowel context (excluding the "German" vowels /e/ and /o/) with the $F1$ and $F2$ frequency norms for full vowels produced by male speakers of American English (Peterson and Barney, 1952) is instructive, assuming that the present speaker's vocal tract size did not deviate very much from the average. In most cases, the transient $F1$ frequencies are lower than those of full vowels, obviously because the vocal tract was less open at consonant release. It is well known that vowels following stop consonant releases have a rising $F1$ transition due to the progressive opening of the vocal tract, except for very high vowels. Indeed, it is surprising that so much variation in $F1$ was obtained in transients with such constricted articulations. The $F1$ frequencies of /i/ and /u/, however, are higher than those of full vowels; this may indicate that the very high tongue position for these vowels was not fully established at release. The $F2$ frequencies are generally in the same range as those of full vowels, except for /ɔ/, which is a little too high, and for /ɑ/, which is much too high. This latter difference may represent an influence of the talker's native German.

Another question of interest is to what extent the formant frequencies changed while the transient decayed. The

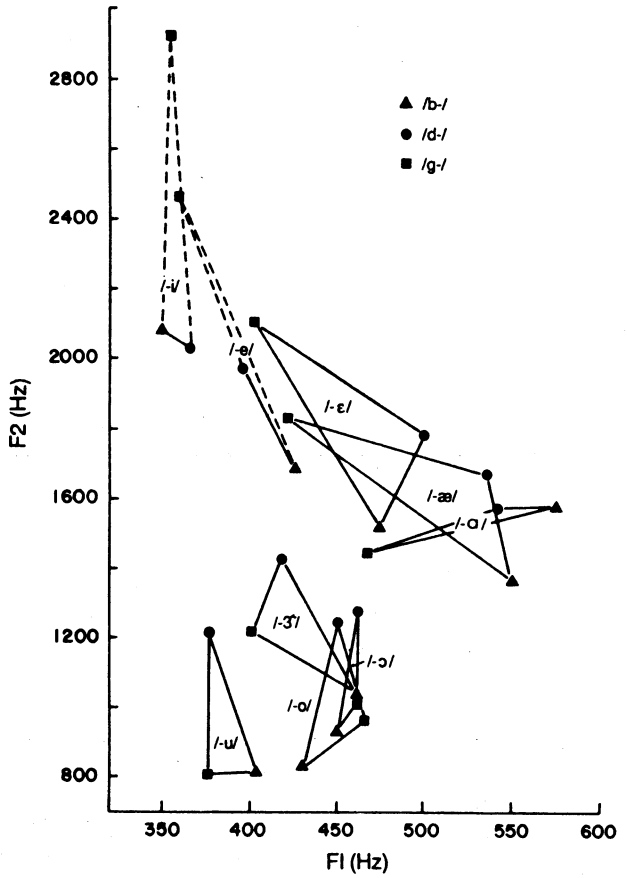

FIG. 9. First and second formant frequencies in transient onset spectra as a function of vowel.

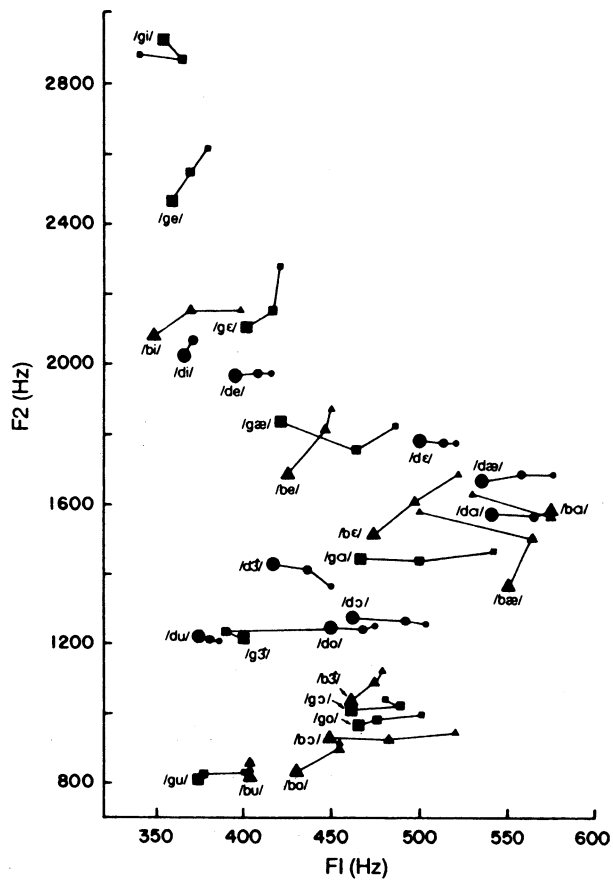

FIG. 10. First and second formant frequencies in transients as a function of time (three overlapping 12.8-ms time frames) for all consonant–vowel combinations. The largest symbol represents the first time frame, and the smallest represents the last frame.

opening gesture is very rapid, and some information about the changing vocal tract resonances may have been available in the very brief signal. Figure 10 replots the data once more in the $F1$–$F2$ plane, but now with two additional data points of diminishing size connected to the large symbol representing each transient's onset frequencies. These two additional data points derive from two additional 12.8-ms Hamming windows overlapping by 6.4 ms. It is evident that, in nearly all instances, there were increases in $F1$ frequency in the course of the transient by as much as 70 Hz for syllables such as /gɑ/ and /bɔ/. This reflects the rapid opening of the vocal tract. Two syllables, /bɑ/ and /bæ/, are atypical in that $F1$ falls at the end. Changes in $F2$ are much less pronounced and occur only for a few syllables (/be/, /bɛ/, /bæ/, /bɜ/, /ge/, /gɛ/), with frequency rising in each instance.

b. Natural release bursts. The average 12.8-ms onset spectra of the 27 full CV syllables are shown in Fig. 11. Comparison with the transient spectra in Fig. 7 reveals a number of differences, as well as some similarities. The most striking difference is the general absence of clearly defined formants in the CV onset spectra. The peaks that do appear in the spectra are typically low and broad. The first formant is generally absent, and so is the fixed low-frequency artifact found in the transients, although labial CV release bursts do show an energy concentration in the low frequencies. The absence of a clear $F1$ reflects the relative predominance of frication noise and the relatively open glottis at CV release. Careful comparison of corresponding spectra in Figs. 7 and 11 reveals that the higher formants are usually present at similar frequencies in the CV onset spectra, but only in very weak form. Similarities between the two sets of spectra are evident in the global properties of the spectra—their slope and general energy distribution—, especially for labials and velars. Alveolars in back vowel contexts, however, show a high-frequency predominance in CV release bursts, presumably due to frication, which is not present in the transients.

Since the formants were so poorly defined in the CV release burst spectra, it proved very difficult to obtain reliable estimates of formant frequencies that could be compared to those of the transients. Nevertheless, peaks corresponding to $F2$ can often be detected in the average spectra (Fig. 11), and they show a pattern similar to that of the $F2$ peaks in isolated transients (Fig. 7), except for an apparently greater constancy of $F2$ in alevolars, which is in agreement with $F2$ measurements at voicing onset (Stevens *et al.*, 1966; Fant, 1973; Kewley-Port, 1983). These $F2$ resonances may derive from the initial transient in natural speech, though they may also have been excited by the frication generated at a widening constriction (Soli, 1981).

As a measure of global spectral similarity, we computed

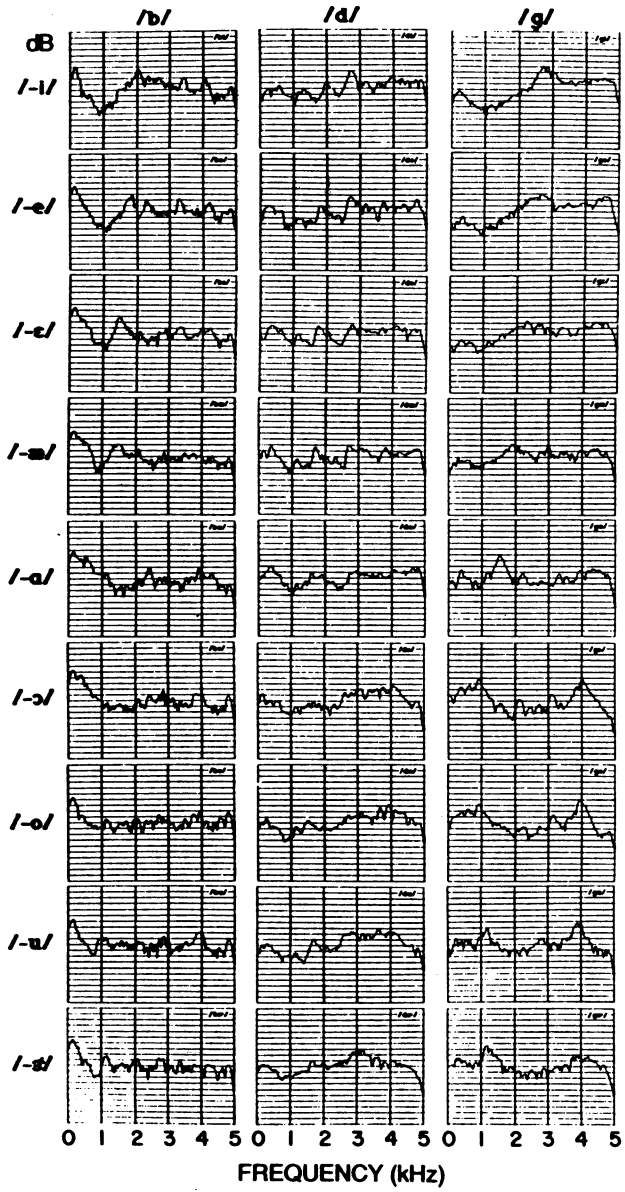

FIG. 11. Average CV syllable release burst onset spectra, based on a full Hamming window over the first 12.8 ms and averaged over tokens. All spectra are amplitude normalized and include high-frequency pre-emphasis. Each horizontal division corresponds to 5 dB.

the linear correlations between corresponding transient and CV onset spectra, each represented as 256-dB values across the frequency range. These correlations are shown in Table I. For each pair of spectra, a second correlation was computed after omitting the low-frequency region of the spectrum (below 1000 Hz for front vowels and /ɑ/, below 700 Hz for other vowels), so as to exclude $F1$ and the artifactual low-frequency peaks. It can be seen that the highest correlations for all three consonants were obtained in the /i/ context. Lower but still positive correlations were obtained in other front vowel contexts. In back vowel contexts, however, only labials (whole spectrum) and velars showed substantial correlations; for alveolars, and for labials after exclusion of the low-frequency region, there was little similarity between the transient and CV onset spectra in back vowel contexts. This confirms the impression gained from the informal visual comparison of Figs. 7 and 11.

TABLE I. Linear correlations among transient and CV onset spectra ($N = 256$). Correlations with low-frequency components omitted are shown in parentheses.

	/b-/	/d-/	/g-/
/-i/	0.81 (0.82)	0.58 (0.73)	0.73 (0.82)
/-e/	0.58 (0.49)	0.44 (0.42)	0.61 (0.65)
/-ɛ/	0.61 (0.54)	0.46 (0.52)	0.44 (0.67)
/-æ/	0.62 (0.42)	0.42 (0.34)	0.22 (0.48)
/-ɑ/	0.70 (0.32)	0.20 (0.07)	0.50 (0.56)
/-ɔ/	0.72 (0.15)	0.05 (0.05)	0.49 (0.51)
/-o/	0.51 (0.20)	−0.08 (0.00)	0.62 (0.70)
/-u/	0.44 (0.02)	−0.23 (−0.23)	0.24 (0.39)
/-ɜ/	0.54 (0.17)	0.03 (0.07)	0.58 (0.66)

C. Discussion

It is not easy to locate displays of release burst spectra in the literature that can be compared directly to Fig. 11. Stevens and Blumstein (1978) and Blumstein and Stevens (1979), for example, show selected CV onset spectra obtained from a window twice as wide as the present one, so that the onset of voicing was usually included. (They also used a half-Hamming window and LPC analysis, with resulting smoothing of the spectrum.) Their spectra often show clearly defined formant peaks, including $F1$, which presumably derive from the voiced portion of the signal. Nevertheless, their characterization of the global spectral shapes as diffuse falling (or flat) for labials, diffuse rising for alveolars, and compact for velars, is generally supported by the present data for CV release bursts. The absence of clear formant peaks in most of the present spectra is probably due to the predominance of the fricative component in the release bursts.

The transient spectra (Fig. 7) conform much less well to Stevens and Blumstein's classifications, due to their pronounced formant structure. There is nothing diffuse about the alveolar spectra, for example, and, in back vowel contexts, they are falling rather than rising. This confirms the conjecture that the natural release burst spectrum is dominated by the frication generated at the constriction, especially for alveolars. For labials and velars, especially in front vowel contexts, the transient seems to exert a greater influence on the release burst spectrum, probably due both to its greater strength when the front cavity is small (cf. Fig. 2) and to the weaker frication associated with velar and especially labial places of articulation. In that connection, it is of interest that both Repp (1986) and Maeda (1987) noted the natural occurrence of release transients without accompanying fricative noise and without cessation of voicing in utterances of /mi/ and /ni/. Experiments with edited nasal consonant–vowel stimuli also confirmed the perceptual importance of these transients (Repp, 1986). Maeda (1987) pointed out that the high tongue position and spread lips for high front vowels tend to increase the contact area of the

articulators, which, in turn, results in a stronger release transient. He also discussed other articulatory parameters that are likely to influence the relative strengths of release transient and fricative noise: the velocity of the opening gesture, the width of the glottal opening, and the subglottal air pressure.

Clearly, the transient spectra reflected not only the consonant place of articulation but also the vocalic context. Since the talker tried to produce each transient with the same articulatory position as that of the corresponding natural CV syllable, the finding indicates that coarticulatory effects are pronounced already at the moment of the release. The formant peaks are probably continuous with those observed at the onset of voicing in natural CV syllables. Unfortunately, we were unable to obtain reliable formant onset frequencies for many of our CV syllables even at voicing onset. However, Fant (1973) provides a table of formant values (not including $F1$) "at instant of release" for /b,d,g/ in various Swedish vowel contexts. The $F2$ values reported there are quite compatible with those observed in the present transients, except for /gi/ and /ge/, in which Fant did find an $F2$ separate from $F3$ around 2000 Hz. The formant frequencies we did manage to extract from our own CV tokens tend to confirm the conclusion that the initial acoustic snapshot of the vocal tract offered by the transient is continuous with the later manifestation of vocal tract resonance frequencies at voicing onset.

II. PERCEPTUAL EXPERIMENT 1

The very clear definition of the vocal tract resonances in the spectra of isolated release transients suggests that these brief signals should be perceptually informative, perhaps even more so than the spectrally more diffuse, natural release bursts. On the other hand, they do not sound very speechlike, so that a listener asked to identify consonants and/or vowels from them would have to employ a process of inference. This is equally true, however, of natural release bursts presented in isolation (cf. Schatz's, 1954, comments). Of particular interest from a perceptual viewpoint is the fact that consonant and vowel information are physically merged in these brief, almost static, signals. That is, the auditory percept corresponding to a transient has a single timbre; any disentanglement of "consonantal" and "vocalic" qualities would have to be a cognitive act on the part of the listener.

In the first perceptual experiment, we tested subjects' ability to identify either the consonant or the vowel from isolated transients. Because of the unwieldy number of 27 response alternatives, a CV syllable identification condition was not included. However, to examine whether listeners could cognitively separate consonant from vowel information, identification of each segment was tested in two conditions: with the other segment known (fixed) or unknown (randomly varying). If such a cognitive separation were possible, then identification accuracy should be higher in the fixed than in the random conditions.

A. Methods

1. Subjects

Eight paid volunteers, five women and three men, served as subjects. They were recruited by notices on Yale campus; most of them were college students. They were all native speakers of American English and reported having normal hearing.

2. Stimulus tapes

Six test tapes were recorded. The first two (consonant identification with vowel fixed) each contained nine blocks of 45 isolated transient stimuli each. Each block contained the five tokens of the three consonants in a fixed vowel context, repeated three times in random order with interstimulus intervals (ISIs) of 2.5 s. The two tapes differed only in the order of vowels across blocks: /i,e,ɛ,æ,ɝ,ɑ,ɔ,o,u/ on tape 1 and the reverse order on tape 2. (In all perception tests and their analyses, the vowel /ɝ/ was placed between /æ/ and /ɑ/.)

Tapes 3 and 4 (vowel identification with consonant fixed) each contained three blocks of 135 stimuli. Each block contained the five tokens of the nine vowels in a fixed consonant context, repeated three times in random order with ISIs of 4 s. (A longer ISI was used because of the larger number of response alternatives.) The order of consonants across blocks was /b,d,g/ on tape 3 and the reverse on tape 4.

Tapes 5 and 6 each contained three blocks of 135 stimuli. Each block contained the five tokens of all 27 CV combinations in random order. The ISIs were 2.5 s on tape 5 (consonant identification) and 4 s on tape 6 (vowel identification).

Each subject listened to four tapes (1 or 2, 3 or 4, and 5 and 6) and thus gave 15 responses to each consonant in each vowel context and to each vowel in each consonant context, in both the fixed and random conditions. Each subject had a different schedule: Half did consonant before vowel identification (on separate days), while half did the opposite; half received one order in the fixed conditions, while half received the other. For all subjects, the random conditions followed the fixed conditions to counteract any confounding of practice effects with the hypothesized advantage in the fixed conditions.

3. Procedure

The subjects were tested individually in a quiet room, and the stimuli were played back from a high-quality tape recorder to TDH-39 earphones at a comfortable intensity. Each subject received written instructions that explained in detail how the stimuli had been generated. A complete set of printed answer sheets was provided. Each line listed all the response alternatives, and the subjects entered their response by circling one alternative. To make the vowel sounds unambiguous, the following spelling was employed on the answer sheets: -EDE, -ADE, -ED, -AD, -URD, -OD, -AWD, -ODE, -UDE. The instructions mentioned that the final D was not pronounced. The nine alternatives just listed were used in the vowel identification condition with randomly

varying consonants (tape 6). When the consonant was fixed (tapes 3 and 4), the dashes were replaced by the consonant symbol appropriate for each block. In the consonant identification task with vowels varying randomly (tape 5), the alternatives were B-, D-, G-; when vowels were fixed (tapes 1 and 2), the dashes were replaced with the vowel spelling appropriate for each block. At the beginning of the first session, one 135-item block of stimuli from tape 5 was presented to familiarize the subject with the sound of the stimuli; no responses were required at that stage.

B. Results and discussion

1. Consonant identification

Overall performance was 51% correct, which is clearly above chance (33%), but not very impressive. Individual subjects differed substantially in their accuracy, however. The best subject achieved 71% correct, the worst 27% correct (though he was not guessing randomly). A three-way repeated-measures analysis of variance with factors condition, consonant, and vowel revealed a marginally significant difference between the fixed and random conditions [$F(1,7) = 6.88$, $p = 0.0343$]: When the vowel was fixed, performance was about 3% higher, on the average, than when the vowel varied randomly; six of the eight subjects showed a difference in this direction. The vowel main effect was not significant, but its interaction with condition was highly significant [$F(8,56) = 5.59$, $p < 0.0001$]. This interaction is shown in Fig. 12. It is evident that holding the vowel constant led to a large improvement in consonant identification for the front vowel series /i,e,ɛ,æ/, but to a decrement for the back vowel series /ɑ,ɔ,o,u/, with little change in the /ɜ/ context.

Figure 13 breaks the results down further by individual consonants. Here, it can be seen that alveolar consonant identification improved across the board when the vowel was fixed, though more so in front vowel contexts, whereas labial and velar consonant identification improved in front vowel contexts but was impaired in back vowel contexts. Especially striking is the large decrement in scores for /bo/ and /bu/, which received mostly D- responses when the vowel was fixed. In the statistical analysis, the only significant effect involving consonants was the consonant by vowel interaction [$F(16,112) = 2.01$, $p = 0.0182$], a small effect. The complete confusion matrices are presented in Table II.

At least part of the response pattern can probably be explained by the physical merging of consonant and vowel information. By virtue of their articulatory and spectral properties, certain vowels have affinities with certain consonants. Thus the rounded back vowels /u/ and /o/ are most similar to labial stops, while high front vowels such as /i/ and /e/ are most similar to alveolar stops. When the vowel context is not known, response biases towards alveolars in front vowel contexts and towards labials in rounded back contexts may be expected. When the vowel is known, however, listeners presumably would distribute their responses more evenly among the three consonantal alternatives. The data (Fig. 13) are not totally incongruent with these expectations, but they do not fit them very well, especially as far as alveolar responses are concerned. The pattern of confusions, too, does not show any very convincing trends (Table II). Part of the reason for this lies in the large individual differences among the subjects. Several subjects did show a marked decrease in alveolar responses and an increase in labial and velar responses from front to back vowels, especially in the variable vowel condition. Other subjects, however, showed the opposite. The strangest pattern was exhibited by the subject with the lowest overall score, who systematically misidentified many of the stimuli. For exam-

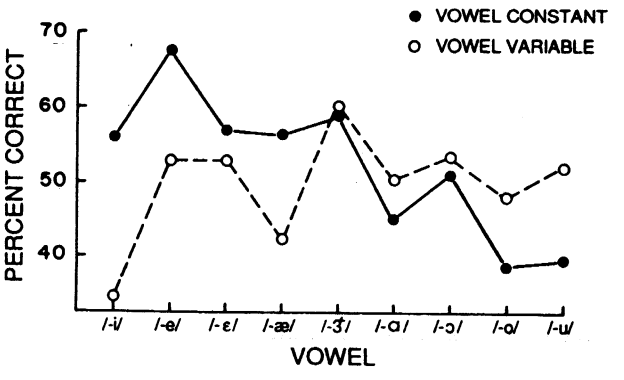

FIG. 12. Average consonant identification scores as a function of vowel and condition (experiment 1).

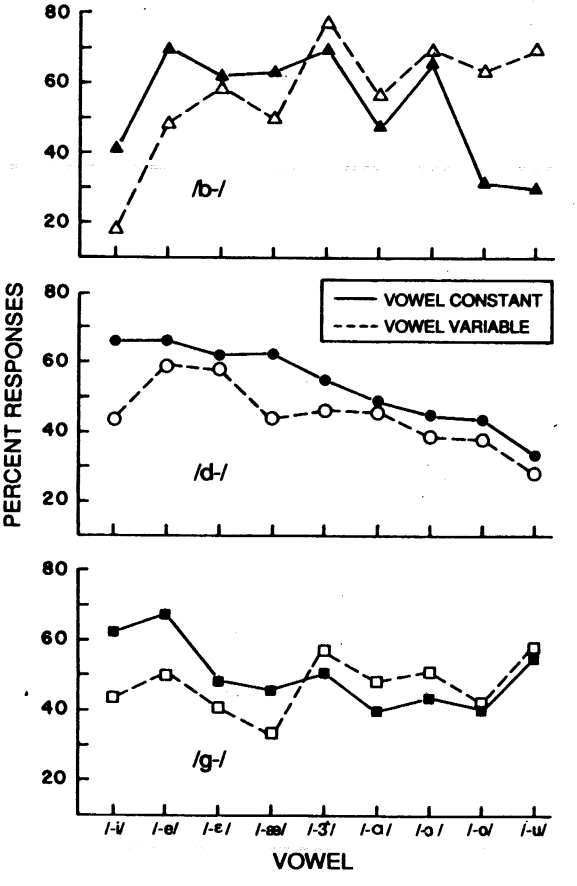

FIG. 13. Consonant identification scores as a function of vowel and condition, separately for each consonant (experiment 1).

TABLE II. Consonant confusion matrices in two experimental conditions and nine vowel contexts (response percentages).

Stimulus	Response: Vowel fixed B-	D-	G-	Vowel varying B-	D-	G-
/bi/	40.8	45.0	14.2	18.3	44.2	37.5
/di/	13.3	65.8	20.8	15.0	43.3	41.7
/gi/	10.8	27.5	61.7	15.0	41.7	43.3
/be/	70.0	25.0	5.0	49.2	37.5	13.3
/de/	23.3	65.8	10.8	20.0	59.2	20.8
/ge/	0.8	31.2	67.5	15.8	34.2	50.0
/bɛ/	61.7	25.8	12.5	59.2	27.5	13.3
/dɛ/	17.5	61.7	20.8	16.7	58.3	25.0
/gɛ/	15.8	35.8	48.3	19.2	40.0	40.8
/bæ/	63.3	15.8	20.8	50.0	15.8	34.2
/dæ/	16.7	62.5	20.8	30.0	44.2	25.8
/gæ/	13.3	40.8	45.8	10.8	55.8	33.3
/bɜ/	70.0	20.8	9.2	77.5	15.8	6.7
/dɜ/	10.8	55.0	34.2	24.2	45.8	30.0
/gɜ/	14.2	35.0	50.8	20.8	22.5	56.7
/bɑ/	47.5	20.0	32.5	56.7	15.0	28.3
/dɑ/	10.8	49.2	40.0	18.3	45.8	35.8
/gɑ/	6.7	53.3	40.0	10.0	41.7	48.3
/bɔ/	65.8	21.7	12.5	70.0	18.3	11.7
/dɔ/	12.5	45.0	42.5	18.3	39.2	42.5
/gɔ/	26.7	30.0	43.3	26.7	22.5	50.8
/bo/	30.8	55.0	14.2	64.2	31.7	4.2
/do/	17.5	44.2	38.3	32.5	38.3	29.2
/go/	32.5	26.7	40.8	28.3	30.0	41.7
/bu/	30.0	51.7	18.3	70.0	25.0	5.0
/du/	30.0	34.2	35.8	27.5	28.3	44.2
/gu/	17.5	27.5	55.0	25.0	16.7	58.3

ple, he responded consistently with G- to /bɑ/ and dɑ/, and with D- to /gɑ/. It seems that he was able to distinguish the timbres of the stimuli well but drew wrong inferences about the phonetic categories associated with them.

2. Vowel identification

Considering as correct responses only those that corresponded most closely to the stimuli, overall accuracy was 28% correct—not an impressive score, but clearly above chance (11% correct). Again, there were large individual differences, with scores ranging from 14%–41% correct. There was a positive correlation of 0.79 ($p < 0.01$) between the consonant and vowel identification scores across subjects.

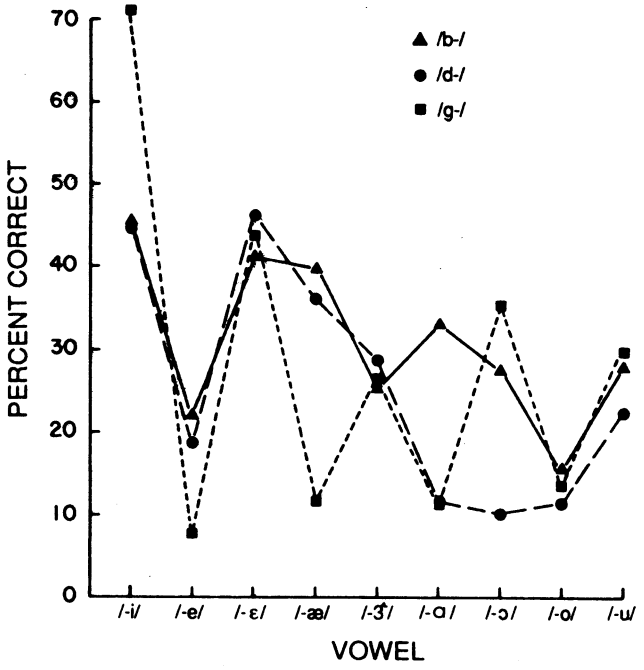

FIG. 14. Vowel identification scores as a function of consonant (experiment 1).

Vowel identification was actually better than suggested by these low scores, since confusions occurred mostly with adjacent response alternatives (see the confusion matrices in Tables III and IV). If the responses are scored so that "adjacent" confusions are considered correct responses, then the average score increases to 61% correct (chance = 33%). By that criterion, the transients actually contained more reliable vowel than consonant information.

Further analyses were conducted on correct responses under the narrower definition, since some of the response patterns of interest concerned shifts between adjacent alternatives. A three-way repeated-measures analysis of variance yielded three significant effects. First, there was a main effect of vowel [$F(8,56) = 5.61$, $p < 0.0001$]: The highest scores were obtained for /i/ (54% correct) and /ɛ/ (45% correct), and the lowest were obtained—perhaps not surprisingly—for /e/, /ɑ/, and /o/, which were the talker's "German" vowels. In the case of /e/ and /o/, a bias against the diphthongal responses, -ADE and -ODE, also may have played a role. The second significant effect was the condition by vowel interaction [$F(8,56) = 3.10$, $p = 0.0058$]: Although vowel identification was *not* significantly more accurate overall when the consonant was known, some vowels (such as /i/, /ɛ/, and /ɔ/) did benefit, while others (/ɑ/, /u/) suffered. This may have been due mainly to a more even distribution of responses among the nine alternatives in the variable consonant condition. The third, and most interesting, effect was the consonant by vowel interaction [$F(16,112) = 5.41$, $p < 0.0001$]: Since consonant and vowel jointly affected the transient spectrum, listeners apparently could not avoid interpreting the effects of consonant place of articulation as changes in vowel color.

The pattern of scores corresponding to this last interaction is graphed in Fig. 14. It is evident that identification of

some vowels (/ɛ/, /ɝ/, /o/) was unaffected by the nature of the consonant, while that of others was strongly affected. These effects can be interpreted with reference to the acoustic data plotted in Fig. 9, and to the complete confusion matrices shown in Tables III and IV. Thus identification of /i/ was enhanced in /gi/, obviously due to the high second peak (really, $F3$) in the spectrum. Identification of /e/, on the other hand, was poorest in /ge/, but for the same reason: The high second peak ($F3$) frequency led to an increased number of incorrect /i/ responses. Even though correct /ɛ/ identification was not much affected, incorrect /i/ responses were more frequent with /gɛ/, too, due to the higher $F2$ and lower $F1$ in these transients. For /æ/, scores are much lower for velars than for labials and aveolars. Figure 9 and the tables again reveal why: /gæ/ transients had a much lower $F1$ than /bæ/ and /dæ/ transients, which led to an increase in incorrect /ɛ/ responses. The central vowel /ɝ/ seemed to be relatively immune to consonant effects. Identification of /ɑ/ was best in /bɑ/, while /dɑ/ and /gɑ/ evoked incorrect /æ/ and /ɛ/ responses, in agreement with their lower $F1$; the increase in /ɔ/ responses to /bɑ/ seen in the tables is somewhat surprising, however. Scores for /ɔ/ were lowest in /dɔ/, whose $F2$ was raised, and which received a larger number of incorrect front vowel (/æ/, /ɛ/, even /ɝ/) responses. For /o/, Fig. 14 reveals no strong consonant effect, but the confusion matrices show a difference in response distributions for /go/ vs /bo/ and /do/; the former received mostly /ɑ/ and /ɔ/ responses, whereas the latter elicited also /ɛ/ and /ɝ/ responses. Finally, /u/ was identified more poorly in /du/ than in /bu/ and /gu/, due to its elevated $F2$, which brought it into the central vowel region; indeed, /ɝ/ responses were most common to /du/.

There was a fourth, marginally significant effect in the analysis of variance, the triple interaction among condition,

TABLE III. Vowel confusion matrix for the constant consonant condition (raw response frequencies, maximum = 120).

Stimulus	Response -EDE	-ADE	-ED	-AD	-URD	-OD	-AWD	-ODE	-UDE
/bi/	67	2	9	1	15	5	1	7	13
/di/	62	6	16	3	6	5	5	5	12
/gi/	87	10	13	2	2	0	0	4	2
/bɛ/	25	32	27	12	3	10	3	7	1
/dɛ/	32	16	23	9	9	6	4	11	10
/gɛ/	71	15	22	1	3	2	0	3	3
/bɛ/	8	21	49	13	10	4	11	1	3
/dɛ/	3	22	59	9	9	8	5	4	1
/gɛ/	12	17	66	9	6	2	3	3	2
/bæ/	2	5	30	46	5	10	14	2	6
/dæ/	0	20	42	44	3	3	7	1	0
/gæ/	7	9	69	12	11	4	4	2	2
/bɝ/	2	2	16	9	33	23	16	11	8
/dɝ/	2	4	28	5	32	17	10	11	11
/gɝ/	0	4	13	11	35	27	16	10	4
/bɑ/	1	2	14	18	6	37	25	13	4
/dɑ/	2	11	31	47	9	10	4	6	0
/gɑ/	3	11	44	19	21	15	5	2	0
/bɔ/	2	0	5	2	13	33	33	24	8
/dɔ/	1	6	26	10	16	27	18	9	7
/gɔ/	0	2	3	3	11	45	48	8	0
/bo/	2	6	17	11	26	21	11	17	9
/do/	1	5	20	6	25	25	12	15	11
/go/	0	2	8	3	9	59	27	12	0
/bu/	0	1	8	2	18	18	22	29	22
/du/	0	3	5	1	22	18	19	32	20
/gu/	0	0	0	0	9	12	15	45	39

TABLE IV. Vowel confusion matrix for the variable consonant condition (raw response frequencies, maximum = 120).

Stimulus	Response -EDE	-ADE	-ED	-AD	-URD	-OD	-AWD	-ODE	-UDE
/bi/	42	19	22	8	6	6	5	2	10
/di/	45	21	25	7	3	5	2	8	4
/gi/	84	6	6	5	1	4	4	5	5
/bɛ/	16	21	48	4	13	3	7	2	6
/dɛ/	21	29	33	1	11	5	3	8	9
/gɛ/	77	3	18	3	4	6	1	5	3
/bɛ/	2	16	51	20	7	8	4	8	4
/dɛ/	1	12	52	21	17	3	3	8	3
/gɛ/	33	18	39	8	6	2	3	6	5
/bæ/	0	3	26	50	11	18	7	3	2
/dæ/	7	6	34	43	7	10	7	2	4
/gæ/	6	17	54	16	7	13	1	3	3
/bɝ/	0	5	10	11	28	41	12	9	4
/dɝ/	2	2	20	5	36	15	8	13	19
/gɝ/	1	3	8	15	29	25	12	11	16
/bɑ/	1	2	12	9	4	43	37	9	3
/dɑ/	1	6	40	35	10	18	6	2	2
/gɑ/	0	7	41	26	20	12	4	4	6
/bɔ/	0	1	5	5	12	37	33	21	6
/dɔ/	5	7	19	18	26	24	6	10	5
/gɔ/	0	2	6	4	12	43	37	9	7
/bo/	1	4	18	2	17	24	17	20	17
/do/	2	2	14	7	35	25	14	12	9
/go/	0	0	6	4	8	49	27	21	5
/bu/	3	1	9	1	9	16	10	26	45
/du/	0	4	10	3	36	10	7	16	34
/gu/	0	1	6	1	8	16	16	40	32

consonant, and vowel [$F(16,112) = 1.80$, $p = 0.0393$], which suggests some changes in response distributions across conditions. On the whole, however, there was no evidence that knowledge of the consonant enabled the listeners to "factor out" the consonantal influences and recover the vowel information in purer form.

III. PERCEPTUAL EXPERIMENT 2

Experiment 1 showed that, even though isolated release transients contain precise information about vocal tract resonances, listeners are not particularly good at interpreting this information in terms of either consonant place of articulation or vowel configuration. Also, they do not seem to be able to separate consonant and vowel information cognitively, given that this information is coded by the same physical parameters. The purpose of experiment 2 was to selectively enhance either the consonantal or the vocalic aspect of the transient, to see whether this physical manipulation improves listeners' utilization of vocal tract resonance information in identification of phonetic segments.

The first part of the experiment was merely a demonstration and was not actually run with a group of subjects. This part concerned the physical enhancement of the vocalic aspect, which was achieved by iterating each release transient ten times. This resulted in vowel-like, steady-state signals with a fundamental frequency of approximately 40 Hz (due to the 25.6-ms duration of each transient "period"). These vowels had a rather strange voice quality, but their timbre was clearly that expected from the formant frequencies observed in the transient. The stimuli were not actually presented to subjects because they seemed too difficult to identify in terms of conventional vowel categories. It will be recalled (cf. Fig. 9) that $F1$ was relatively low in all low vowels because of the narrow opening during transient production. All of those vowels, therefore, had a more or less

neutral quality and were difficult to associate with phonetic symbols. Moreover, since the stimuli were steady-state vocoids, listeners would not have had the opportunity to think of them as onsets that were "going somewhere," as with the isolated transients of experiment 1.

The second part of the experiment concerned the enhancement of the consonantal aspect in the release transients. This was achieved by following each transient with a steady-state vowel corresponding to its intended vocalic context. The transient thus assumed the status of a consonantal cue, and additional information about spectral change between consonantal release and vocalic steady state became available to listeners' perceptual systems. The question was how effective a cue the transient would be, given that other components of natural stop consonant releases (frication, aspiration, and vocalic formant transitions) were lacking. In addition, because the relative amplitude of stop consonant release bursts has been shown to be a secondary cue to place of articulation (Ohde and Stevens, 1983; Repp, 1984), this parameter was also given attention.

A. Methods

1. Subjects

Eleven subjects participated: Eight were paid volunteers recruited from the Yale community and three were unpaid graduate students doing research at Haskins Laboratories. All were native speakers of English and reported having normal hearing.

2. Stimuli

Each transient was immediately followed by a steady-state synthetic vowel; that is, the "voice onset time" was 25.6 ms. Since, at the time of this experiment, the full CV syllables (see above) had not yet been recorded and analyzed, estimates of the formant frequencies for the nine vowels were obtained as follows. Those for /i,e,ɛ,æ/ were taken from a study by Repp and Williams (1987), in which the first author (BHR) had produced these vowels in isolation. The formant frequencies of /i,ɛ,æ/ were compared to those reported by Peterson and Barney (1952), and the average percentage deviation was computed for each formant, as a correction for BHR's (apparently somewhat larger) vocal tract size relative to the average American male. These percentage factors were then applied to the Peterson and Barney data for /ɑ,ɔ,u,ɝ/ to yield estimates of BHR's productions of these vowels. Finally, the non-English vowel /o/ was synthesized using formant frequencies extracted from an actual vowel produced by BHR for this purpose. All resulting nine vowels sounded like the intended vowels to BHR. Their formant frequencies are listed in Table V.

Each of the vowels was completely steady state and of constant amplitude, except for the last 70 ms, over which the amplitude was made to fall linearly by 20 dB. The total duration was 250 ms, and the fundamental frequency fell linearly from 100 to 80 Hz. The serial configuration of the Klatt (1980) synthesizer was used to generate the waveforms, which were appended to the transients by a stimulus sequencing program during recording of the test tapes.

TABLE V. Formant frequencies (in Hz) and rms amplitudes (in dB) of the nine synthetic vowels used in experiment 2.

Vowel	*F*1	*F*2	*F*3	Amp
/i/	240	2165	2931	43.9
/e/	326	2025	2531	43.8
/ɛ/	456	1745	2403	44.4
/æ/	592	1619	2358	47.5
/ɝ/	433	1266	1639	44.3
/ɑ/	645	1022	2367	48.9
/ɔ/	503	788	2338	49.0
/o/	360	790	2428	43.4
/u/	265	816	2173	40.6

Even though pilot observations suggested that the relative amplitudes of transient and vowel did not matter much, we nevertheless decided to vary this parameter. Keeping the transient amplitudes constant, we presented the vowels not only at, but also 10 and 20 dB below, their original amplitudes. The amplitude adjustment was made in the synthesis parameters (i.e., in the amplitude of the simulated glottal source). Because the synthesizer emulates the human vocal tract, low vowels had higher amplitudes than high vowels, as indicated in the last column of Table V. Although these dB values have an arbitrary reference, they are directly comparable to the rms amplitude values for the transients, as shown in Figs. 2 and 3. In terms of average peak (onset) amplitude, the transients were roughly 7 dB below, 3 dB above, and 13 dB above the vowel, respectively, in the three relative amplitude conditions. In terms of average amplitude, they were at least 15 dB below these values. The precise amplitude relationship varied from token to token as a function of random variability, consonant place of articulation, and vocalic context; these relationships (except for the token variability) can be puzzled out from Fig. 2 in conjunction with Table V, if necessary.

The experimental tape contained a total of three (consonants) × nine (vowels) × five (tokens) × three (vowel amplitudes) = 405 stimuli, arranged randomly in 9 blocks of 45, with ISIs of 2.5 s.

3. Procedure

The procedure was similar to that of experiment 1, but the subjects were simply asked to write down B, D, or G for each stimulus heard. Pilot observations suggested that a "no consonant" category was not necessary, even though the consonant heard was sometimes ambiguous.

B. Results and discussion

Overall accuracy in this experiment was 66.4% correct—not as high as expected, but better than the 52.4% correct obtained in the fixed-vowel condition of experiment 1, where the vocalic context was known but physically absent. The improvement was significant in an ANOVA comparing the two experiments, ($F1,17) = 6.84, p = 0.0181$. Individual subjects varied considerably in their accuracy, from 47.6%–83.5% correct. (Author BHR, who listened to the

tape as a pilot subject, averaged 83.5% correct for these stimuli derived from his own productions, compared to 62% correct for the isolated transients in experiment 1; thus he did no better than the best regular subject.)

The relative amplitudes of transient and vowel had no overall effect on identification accuracy: Scores were 65.7%, 66.8%, and 66.8% correct, respectively, in the three amplitude conditions. The error pattern did seem to change somewhat, however, across these conditions since there was a significant triple interaction between consonant, vowel, and vowel amplitude in the statistical analysis, $F(32,320) = 1.86$, $p = 0.0041$. In that analysis, there was also a significant main effect of consonant, $F(2,20) = 8.84$, $p = 0.0018$, due to highest average scores for labials and lowest scores for alveolars, and a highly significant consonant by vowel interaction, $F(16,160) = 17.29$, $p < 0.0001$. The vowel main effect was not significant. Thus, while overall consonant intelligibility did not vary significantly across vowels, the intelligibility of individual consonants was highly vowel dependent. Moreover, the nature of that dependency was not the same as for isolated transients since it interacted strongly with the between-group factor in the statistical comparison of experiments 1 (fixed-vowel condition) and 2, $F(16,272) = 9.39$, $p < 0.0001$. Consonant and vowel main effects also interacted with the between-group factor, suggesting that the results followed different patterns in the two experiments.

Figure 15 compares these patterns. Each panel shows

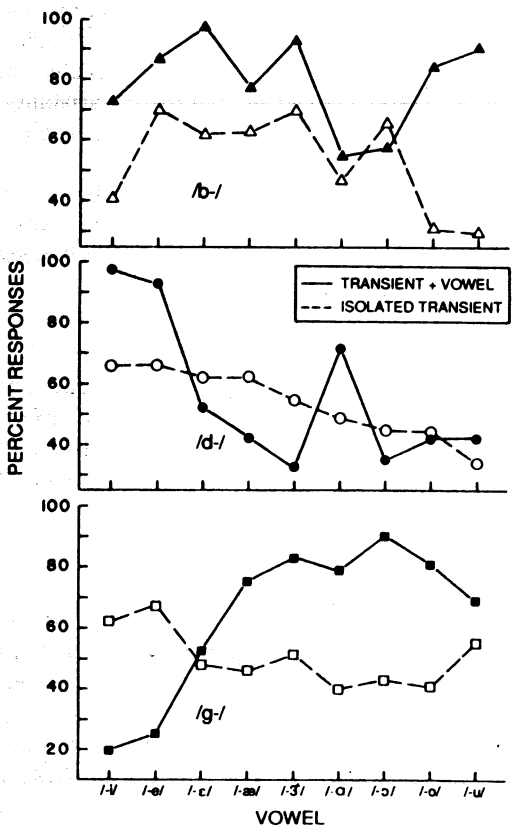

FIG. 15. Consonant identification scores as a function of vowel and condition, separately for each consonant (experiment 2).

percent correct scores for one consonant as a function of vocalic context. It can be seen that /b/ was identified best in the physical context of /ɛ/, /ɜ/, /o/, and /u/; for the latter two especially, addition of the synthetic vowel led to a substantial improvement in intelligibility over the isolated transients. For two contexts, /ɑ/ and /ɔ/, addition of the steady-state vowel did not help /b/ identification, which may indicate that formant transitions are perceptually important in these contexts. Alveolar consonants were identified very well in /i/ and /e/ contexts, and moderately well in /ɑ/ context, better than isolated transients. For most other contexts, however, addition of the synthetic vowel *decreased* identification scores. The likely explanation here is the absence of the frication component in the release bursts, which is probably strongest and perceptually most salient for alveolar consonants. Velar consonant scores were almost complementary to the alveolar ones: Identification was good and

TABLE VI. Confusion matrix (response percentages) for transient-plus-vowel stimuli (experiment 2).

Stimulus	Response /B/	/D/	/G/
/bi/	72.7	27.3	0.0
/di/	0.6	97.6	1.8
/gi/	1.2	78.8	20.0
/be/	86.7	12.7	0.6
/de/	3.0	92.8	4.2
/ge/	2.4	72.2	25.4
/bɛ/	97.6	0.6	1.8
/dɛ/	21.2	52.1	26.7
/gɛ/	13.3	34.5	52.2
/bæ/	77.6	4.8	17.6
/dæ/	23.6	42.5	33.9
/gæ/	13.9	10.9	75.2
/bɜ/	93.4	3.6	3.0
/dɜ/	15.2	32.7	52.1
/gɜ/	13.9	3.6	82.5
/bɑ/	55.1	27.9	17.0
/dɑ/	16.4	71.5	12.1
/gɑ/	11.5	9.7	79.3
/bɔ/	57.8	25.5	26.7
/dɔ/	12.7	35.2	52.1
/gɔ/	5.5	4.2	90.3
/bo/	84.9	10.9	4.2
/do/	24.2	41.9	33.9
/go/	11.5	7.9	80.6
/bu/	91.0	4.8	4.2
/du/	29.7	42.3	28.0
/gu/	27.3	3.6	69.1

improved substantially over isolated transients in all back vowel contexts, but scores were very low and much poorer than for isolated transients in front vowel contexts (/i/, /e/), where they were invariably misidentified as /d/. The poor identifiability of velar stops preceding high front vowels has been repeatedly commented on in the literature (e.g., Liberman *et al.*, 1952; Dorman *et al.*, 1977; Blumstein and Stevens, 1980; Kewley-Port *et al.*, 1983), and even natural utterances of /gi/ are often misidentified (Dorman *et al.*, 1977). The complete confusion matrix is shown in Table VI, averaged over the three attenuation conditions.

These results may be compared directly with those of Dorman *et al.* (1977) who, in one of their several experimental conditions, prefixed nine different vowels (naturally produced VC syllables) with natural release bursts of unaspirated stops excerpted from CVC syllables containing the same vowels. The pattern of their results (for the better of two talkers) is quite similar to the present one: Labial bursts were moderately effective before all vowels, most so before /ɛ,u/, and least so before /i,ɔ/. Alveolar bursts were effective before high front vowels and possibly /ɜ/, but not elsewhere. Velar bursts were most effective before high back vowels. Average performance levels were also comparable. Thus it appears that the present isolated transients conveyed about as much information as do excerpted natural release bursts.

IV. GENERAL DISCUSSION

Our acoustic analyses have shown that isolated transients contain information about both the consonant and the vowel of CV syllables, and that this information is contained in a very distinct formant structure. The perceptual experiments demonstrated that listeners can utilize this information to some extent: Both consonant and vowel identification from isolated transients were clearly better than chance. However, the isolated transients were not sufficient for unambiguous identification of either segment. Among the reasons for this must have been the brevity of the stimuli, their nonspeech timbres, the absence of other cues present in natural speech, and the merging of consonant and vowel information due to coproduction.

Even when listeners knew the vocalic context, they were not really able to separate the consonantal (place of articulation) information from the vocalic influences, though there was a slight improvement. Conversely, knowing the consonant did not help vowel identification at all. When an explicit vocalic context was provided, consonant identification improved but was still far from perfect. The improvement probably reflected the availability to perception of the spectral relationship between the transient and the vowel, which listeners could not simply infer by imagining the vowel. Nevertheless, it must be concluded that transients alone are not sufficient to reliably cue perception of stop consonant place of articulation. Additional cues that are normally present in natural speech were missing, most notably the frication immediately following the release and the formant transitions in the vowel; the transitions especially not only provide additional cues but also make the release burst cohere better with the following periodic signal. Because of the similarity of the present results to those obtained by Dorman *et al.* (1977) with natural release bursts preceding steady-state vowels, it seems that the absence of vocalic formant transitions was the primary factor depressing subjects' identification scores. Nevertheless, the fact that similar identification accuracy was achieved with transients containing a clear formant structure and with natural release bursts dominated by spectrally diffuse frication suggests that there is a perceptual trade-off between precise information about vocal tract resonances and more global information about the nature and location of the fricative noise source. In other words, both kinds of information contribute to consonant identification, and one can substitute for the other, but neither is sufficient. Since the frication seems to be largely responsible for generating the invariant spectral shapes described by Blumstein and Stevens (1979), it also seems that these shapes are neither necessary nor sufficient for place of articulation identification, not even when they are combined with a following vowel. Apparently, only the simultaneous presence of all major cues enables listeners to arrive at unambiguous percepts. A very similar point was made recently by Lisker (1987) with regard to voicing perception.

An interesting secondary finding is the absence of any effect of the amplitude relationship between transient and vowel. Since such amplitude effects have been demonstrated with both synthetic (Ohde and Stevens, 1983) and natural release bursts (Repp, 1984), the results imply that it is the relative amplitude of the fricative component, not of the transient, that causes the effect. The information conveyed by the transient appears to be purely qualitative.

The present results leave open the possibility that perception of stop consonant place of articulation will improve when the initial transient of natural release bursts is artificially enhanced. This may be tested in future experiments using synthetic speech or high-resolution spectral analysis and waveform manipulation. It seems unlikely, however, since the signals giving rise to the most accurate perception are usually those most similar to natural speech. If transients are naturally weak, this is probably what listeners expect to hear.

ACKNOWLEDGMENTS

This research was supported by NICHD Grant HD-01994 to Haskins Laboratories. We are grateful to Richard McGowan and Kenneth Stevens for discussions and advice, and to Leigh Lisker for comments on an earlier version of the manuscript.

Barry, W. J. (**1984**). "Place-of-articulation information in the closure voicing of plosives," J. Acoust. Soc. Am. **76**, 1245–1247.

Blumstein, S. E., and Stevens, K. N. (**1979**). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. **66**, 1001–1017.

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

Cole, R. A., and Scott, B. (**1974**). "The phantom in the phoneme: Invariant cues for stop consonants," Percept. Psychophys. **15**, 101–107.

Cullinan, W. L., and Tekieli, M. E. (1979). "Perception of vowel features in temporally-segmented noise portions of stop-consonant CV syllables," J. Speech Hear. Res. 22, 122–131.

Dorman, M. F., Studdert-Kennedy, M., and Raphael, L. J. (1977). "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues," Percept. Psychophys. 22, 109–122.

Fant, G. (1960). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Fant, G. (1973). "Stops in CV-syllables," in *Speech Sounds and Features* (MIT, Cambridge, MA), pp. 110–139.

Garudadri, H., Benguerel, A.-P., Gilbert, J. H. V., and Beddoes, M. P. (1986). "Application of smoothed Wigner distribution (WD) to speech signals," J. Acoust. Soc. Am. Suppl. 1 79, S94.

Halle, M., Hughes, G. W., and Radley, J.-P. A. (1957). "Acoustic properties of stop consonants," J. Acoust. Soc. Am. 29, 107–116.

Hoffman, H. S. (1958). "Study of some cues in the perception of the voiced stop consonants," J. Acoust. Soc. Am. 30, 1035–1041.

Just, M. A., Suslick, R. L., Michaels, S., and Shockey, L. (1978). "Acoustic cues and psychological processes in the perception of natural stop consonants," Percept. Psychophys. 24, 327–336.

Kewley-Port, D. (1983). "Measurement of formant transitions in naturally produced stop consonant–vowel syllables," J. Acoust. Soc. Am. 72, 379–389.

Kewley-Port, D., Pisoni, D. B., and Studdert-Kennedy, M. (1983). "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," J. Acoust Soc. Am. 73, 1779–1793.

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

Lahiri, A., Gewirth, L., and Blumstein, S. E. (1984). "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study," J. Acoust. Soc. Am. 76, 391–404.

Liberman, A. M., Delattre, P., and Cooper, F. S. (1952). "The role of selected stimulus-variables in the perception of the unvoiced stop consonants," Am. J. Psychol. 65, 497–516.

Lisker, L. (1987). "Orchestrating acoustic cues to linguistic effect," in *Proceedings of the Eleventh International Congress of Phonetic Sciences* (Academy of Sciences of the Estonian S.S.R., Tallinn, Estonia, USSR), Vol. 6, pp. 66–67.

Maeda, S. (1987). "On the generation of sound in stop consonants," Speech Communication Group Working Papers, Vol. V, Research Laboratory of Electronics, MIT, pp. 1–14.

Malécot, A. (1958). "The role of releases in the identification of released final stops," Language 34, 370–380.

Ohde, R. N. (1984). "Fundamental frequency as an acoustic correlate of stop consonant voicing," J. Acoust. Soc. Am. 75, 224–230.

Ohde, R. N., and Sharf, D. J. (1977). "Order effect of acoustic segments of VC and CV syllables on stop and vowel identification," J. Speech Hear. Res. 20, 543–554.

Ohde, R. N., and Stevens, K. N. (1983). "Effect of burst amplitude on the perception of stop consonant place of articulation," J. Acoust. Soc. Am. 74, 706–714.

Peterson, G. E., and Barney, H. L. (1952). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. 24, 175–184.

Repp, B. H. (1984). "Closure duration and release burst amplitude cues to stop consonant manner and place of articulation," Lang. Speech 27, 245–254.

Repp, B. H. (1986). "Perception of the [m]–[n] distinction in CV syllables," J. Acoust. Soc. Am. 79, 1987–1999.

Repp, B. H., and Williams, D. R. (1987). "Categorical tendencies in imitating self-produced isolated vowels," Speech Commun. 6, 1–14.

Schatz, C. D. (1954). "The role of context in the perception of stops," Language 30, 47–56.

Soli, S. D. (1981). "Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation," J. Acoust. Soc. Am. 70, 976–984.

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

Stevens, K. N., House, A. S., and Paul, A. P. (1966). "Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation," J. Acoust. Soc. Am. 40, 123–132.

Suomi, K. (1985). "The vowel-dependence of gross spectral cues to place of articulation of stop consonants in CV syllables," J. Phon. 13, 267–285.

Syrdal, A. K. (1983). "Perception of consonant place of articulation," in *Speech and Language: Advances in Basic Research and Practice, Vol. 9*, edited by N. J. Lass (Academic, New York), pp. 313–349.

Tekieli, M. E., and Cullinan, W. L. (1979). "The perception of temporally segmented vowels and consonant–vowel syllables," J. Speech Hear. Res. 22, 103–121.

Winitz, H., Scheib, M. E., and Reeds, J. A. (1972). "Identification of stops and vowels for the burst portion of /p,t,k/ isolated from conversational speech," J. Acoust. Soc. Am. 51, 1309–1317.

Wokurek, W., Kubin, G., and Hlawatsch, F. (1987). "Wigner distribution—a new method for high-resolution time-frequency analysis of speech signals," in *Proceedings of the Eleventh International Congress of Phonetic Sciences* (Academy of Sciences of the Estonian S.S.R., Tallinn, Estonia, USSR), Vol. 1, pp. 44–47.

Zlatin, M. A., and Koenigsknecht, R. A. (1976). "Development of the voicing contrast: A comparison of voice onset time in stop perception and production," J. Speech Hear. Res. 19, 93–111.

Zue, V. W. (1976). "Acoustic characteristics of stop consonants: A controlled study," Tech. Rep. 523, Lincoln Laboratory, MIT, Lexington, MA.