Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants

Diane Kewley-Port[a] and David B. Pisoni
Speech Research Laboratory, Department of Psychology, Indiana University, Bloomington, Indiana 47405

Michael Studdert-Kennedy
Queens College and Graduate Center of The City University of New York, New York, New York 10036 and Haskins Laboratories, New Haven, Connecticut 06510

(Received 21 June 1982; accepted for publication 10 February 1983)

Two recent accounts of the acoustic cues which specify place of articulation in syllable-initial stop consonants claim that they are located in the initial portions of the CV waveform and are context-free. Stevens and Blumstein [J. Acoust. Soc. Am. 64, 1358–1368 (1978)] have described the perceptually relevant spectral properties of these cues as static, while Kewley-Port [J. Acoust. Soc. Am. 73, 322–335 (1983)] describes these cues as dynamic. Three perceptual experiments were conducted to test predictions derived from these accounts. Experiment 1 confirmed that acoustic cues for place of articulation are located in the initial 20–40 ms of natural stop-vowel syllables. Next, short synthetic CV's modeled after natural syllables were generated using either a digital, parallel-resonance synthesizer in experiment 2 or linear prediction synthesis in experiment 3. One set of synthetic stimuli preserved the static spectral properties proposed by Stevens and Blumstein. Another set of synthetic stimuli preserved the dynamic properties suggested by Kewley-Port. Listeners in both experiments identified place of articulation significantly better from stimuli which preserved dynamic acoustic properties than from those based on static onset spectra. Evidently, the dynamic structure of the initial stop-vowel articulatory gesture can be preserved in context-free acoustic cues which listeners use to identify place of articulation.

PACS numbers: 43.70.Ve, 43.70.Dn

INTRODUCTION

The search for acoustic cues to place of articulation in initial stop consonants has focused on two different sources of information in the stop-vowel syllable: context-free and context-dependent. The context-free source has usually been taken to be the stop release burst, represented as a single, *static* spectral section and described in the terminology of distinctive feature theory (Halle *et al.*, 1957; Fant, 1960; Stevens, 1975; Stevens and Blumstein, 1978). The context-dependent source has usually been taken to be the formant transitions following the release burst, represented as *dynamic* variations in the distribution of energy within restricted regions of the spectrum and described with an emphasis on their origins in the stop articulatory gesture (Liberman *et al.*, 1954; Liberman *et al.*, 1967; Liberman and Studdert-Kennedy, 1978). Finally, some investigators have studied the combined role of release bursts and formant transitions, stressing variation in the effective perceptual weights of the two cues as a function of vowel context (Cooper *et al.*, 1952; Fischer-Jørgensen, 1972; Dorman *et al.*, 1977).

The research reported here attempts to combine a view of stop-consonant place information as invariant, context-free, and largely located in the release burst with a view of the information as dynamic rather than static. The hypothesis of invariance derives, in part, from the acoustic theory of speech production (Fant, 1960; Stevens and Blumstein, 1978), according to which invariant acoustic properties, associated with particular places of articulation, should lie in the brief initial portions of a CV waveform. The hypothesis that dynamic rather than static properties of the release play a role in perception derives from previous work of Kewley-Port (1980, 1983), and is intended as an explicit alternative to the hypothesis of Stevens and Blumstein (1978, 1981) that static properties of the release determine perception of place of articulation in stop consonants.

Stevens and Blumstein have suggested that the auditory system integrates spectral energy over approximately 20 ms at the onset of a release burst. The resulting static spectrum is said to contain the gross spectral properties that distinguish among places of articulation. Blumstein and Stevens (1979) constructed templates to characterize these spectral properties for each place of articulation. They visually matched the templates to a single, static spectrum positioned at the burst onset for a large corpus of stop-vowel syllables and obtained 84% correct place identification. By contrast, Kewley-Port's approach (1980, 1983) was based on a different model of the auditory system, one which suggests that short-term spectra are rapidly updated to preserve temporally changing spectral information (Schroeder *et al.*, 1979). Thus she proposed that running spectra based on linear prediction analysis might provide a more appropriate visual representation of the rapidly changing auditory information. Kewley-Port (1983) defined three time-varying visual features to distinguish place of articulation in the running spectra and showed that judges could use these features to identify place of articulation 88% correctly from running spectral displays of a corpus of CV syllables.

[a] Current address: Bell Laboratories, Room 2C-529, 600 Mountain Avenue, Murray Hill, NJ 07974.

 0001-4966/83/051779-15$00.80

A more detailed look at the specific spectral and temporal properties of the proposed place cues reveals theoretically important similarities and differences between the two approaches. The invariant acoustic cues for place according to Stevens and Blumstein (1981) may be distinguished on the basis of *spectral* properties alone. There is no apparent temporal dimension in their account because spectral energy is integrated over a fixed 25.6-ms window. The gross shapes of the onset spectra are defined as diffuse-falling for bilabials, diffuse-rising for alveolars, and compact for velars. In contrast, the three time-varying features proposed by Kewley-Port (1980, 1983) have both spectral and temporal dimensions. The first feature is the spectral tilt of the burst observed in approximately the first 5 ms of the stop-vowel waveform. Tilt of burst is defined as rising for alveolars, as flat or falling for bilabials, and as varying for velars. The description of the *spectral* properties of the burst for bilabials and alveolars is quite similar for Kewley-Port and for Stevens and Blumstein. However, two important differences in the actual burst spectrum examined by these investigators should be noted. First, Stevens and Blumstein integrate energy over a longer time window. Thus their onset spectrum usually contains some vowel transition information, whereas the spectrum calculated by Kewley-Port over a 5-ms window contains only burst energy. Second, Stevens and Blumstein's templates exclude spectral energy below 1000 Hz whereas Kewley-Port's tilt feature includes this frequency region. Kewley-Port's tilt feature, therefore, incorporates important vocal tract source information. (Specifically, a voiced versus voiceless source is represented by the presence versus absence of an $F1$ spectral peak.) Thus a "rising" spectrum for Stevens and Blumstein is not necessarily rising for Kewley-Port.

The role of the *temporal* dimension in the description of the burst spectrum in Stevens and Blumstein's approach, emphasizing a static description, and Kewley-Port's approach, emphasizing a time-varying description, should be clarified. In both approaches, the burst itself must be located in time prior to determining the spectral tilt. By both Stevens and Blumstein's (Stevens, 1980; Stevens and Blumstein, 1981) and Kewley-Port's (1983) account, the burst is signaled acoustically by an abrupt change in energy over time. Thus the role of the temporal dimension is actually the same in both approaches. The choice of associating—or not associating—this dynamic property of detecting the burst with a tilt of burst feature, however, relates directly to the different models of the auditory system assumed by the investigators.

The second time-varying feature is the presence or absence of a single mid-frequency peak extending in time for at least 20 ms. While the spectral quality of the extended mid-frequency peaks is similar to that of Stevens and Blumstein's compact onset spectra, again there are significant differences between the two approaches. Spectrally, Blumstein and Stevens' (1979) velar template specifies many narrow frequency regions in which a compact peak may occur. Thus their velar template can accept onset spectra with multiple peaks which would not be judged as velar in Kewley-Port's system. More importantly, Kewley-Port's feature emphasizes the *persistence* of the mid-frequency peaks for 20 ms. Thus a high-energy but brief peak, which often occurs for alveolars (Blumstein and Stevens, 1979; Kewley-Port, 1983), will be distinguished in Kewley-Port's system from mid-frequency peaks persisting in time. Only the extended mid-frequency peaks are the predicted acoustic correlates of velar stops (Fant, 1968, p. 223). In fact, Blumstein and Stevens (1980, p. 661) have stated that perhaps "a longer time is necessary to build up a representation of the 'compact' onset spectrum in the auditory system." The studies reported here investigate the necessity of longer duration waveforms for the perception of velars in some detail.

The third time-varying feature described by Kewley-Port (1983) is late onset of an $F1$ peak relative to the burst. This is essentially a measure of voice onset time (VOT), previously shown to correlate with place of articulation (Lisker and Abramson, 1964). In the running spectral analysis, a late $F1$ onset is viewed as a secondary feature specifying velar place of articulation in voiced stops. This feature is based on the assumption that the time of the change from frication to voicing plays a role in place identification. The feature has no counterpart in Stevens and Blumstein's account, because the fixed time window they employ integrates energy over both the frication and voiced portions of the CV waveform.

To examine the extent to which static and dynamic acoustic properties are perceptual cues for place of articulation, three experiments were conducted. In experiment 1, we investigated how much information is needed from the initial portions of natural CV waveforms to identify place of articulation accurately. More specifically, we sought to determine whether place of articulation could be identified accurately from 20-ms waveform segments, as Stevens and Blumstein have claimed. Experiments 2 and 3 were designed to examine whether static or dynamic acoustic properties are used by listeners to identify place in short CV waveforms. These experiments compared the perception of natural speech segments with two sets of synthetic stimuli. One set of synthetic stimuli modeled the Stevens and Blumstein onset spectra; the other was patterned after the time-varying features proposed by Kewley-Port. Experiment 2 used a digital, parallel-resonance synthesizer; experiment 3 replicated the results using linear prediction synthesis. The overall goal of this research was to determine whether the acoustic structure that supports the perception of place of articulation in English stop consonants is more properly described as static or dynamic in nature.

I. EXPERIMENT 1: IDENTIFICATION OF TRUNCATED NATURAL CV'S

The purpose of experiment 1 was to determine the duration of the initial portion of a stop-vowel syllable necessary to identify place of articulation accurately. A related study directed at a somewhat different problem was carried out recently by Tekieli and Cullinan (1979). They measured the minimum duration necessary for listeners to identify consonants and vowels correctly from the electronically gated, initial portions of CV syllables, spoken by a single talker. The authors did not score identification of place of articulation separately from voicing, but their results indicate that

place of articulation, averaged over the six English stops, was identified correctly about 98% of the time from 30-ms waveform durations (cf. Winitz *et al.*, 1972).

Experiment 1 of the present study was nearly completed when Blumstein and Stevens (1980) reported a series of experiments examining place identification in short initial portions of *synthetic* CV stimuli. They concluded that "information with regard to place of articulation for a voiced stop consonant resides in the initial 10–20 ms of a consonant-vowel syllable (p. 660)." Unfortunately, the stimuli and procedures employed in their study make it difficult to generalize the results to the identification of naturally produced CV's. The bursts in their synthetic stimuli were acoustically impoverished since they were generated with only one formant. Furthermore, Blumstein and Stevens always presented their results in terms of the duration of the *voiced* portion of the stimuli. Since the stimuli resembling natural CV's contained aperiodic bursts preceding the voiced portions, the shortest, so-called 10-ms stimuli actually varied in duration from 20 ms for /ba/ to 35 ms for /gi/. Therefore, a reliable estimate of the duration necessary to identify place of articulation in natural CV's cannot be determined from the results of the Blumstein and Stevens study.

In experiment 1, then, we sought to determine directly how place information actually is distributed over the early portions of natural stop-consonant-vowel syllables. We used naturally spoken CV syllables obtained from two male talkers consisting of /b, d, g/ before five different vowels. The short aperiodic and following waveform segments were edited from the syllable and measured digitally by computer. Naive listeners were required to identify place of articulation in the truncated CV's at various waveform durations.

A. Method

1. Stimuli: CV syllables

A set of 30 CV syllables spoken by two male talkers (RP and TF) was chosen from a larger set of utterances used in an earlier experiment (Kewley-Port, 1983). The syllables, one from each talker, consisted of all combinations of initial /b, d, g/ and the vowels /i, e, ɑ, o, u/. These syllables were read in the carrier sentence, "Teddy said CV" from randomized lists in a sound attenuated room and recorded on an Ampex AG-500 tape recorder. The sentences were low-pass filtered at 4.9 kHz and digitized with a 12 bit A/D converter at a 10-kHz sample rate using a PDP 11/05 computer. They were then edited so that only the target CV was permanently stored on disk.

Before the set of stop-vowel syllables was edited further for this experiment, we checked that listeners could correctly identify the consonants in the full syllables. A computer program was used to randomize and output the 30 full syllables through a 12 bit D/A converter for recording on audio tape. The tape consisted of ten blocks of the 30 CV's for a total of 300 trials. Six naive subjects listened to the tape over headphones in a quiet room. All subjects were paid for their services. Subjects were given instructions to write down the letter which corresponded to the consonant they heard at the beginning of each syllable. The response set, therefore, was the *open* set of all English consonants. Results showed that subjects correctly identified the stop consonants in the full syllables at a level of 99.8% correct, with no consonant responses other than b, d, or g. Evidently, all 30 original CV syllables were good exemplars of the stop consonant the talkers had intended to produce.

2. Stimuli: Waveform editing

Each of the 30 original CV's was then digitally edited in order to retain only the initial portions of the waveforms. Five different cuts were made at zero crossings to produce five truncated tokens from each original CV syllable. For /d/ and /g/, the first cut was made just before the first voicing pulse. This aperiodic portion of the waveform, containing the stop release burst and aspiration, will be referred to as the burst. Its mean duration was 14 ms for /d/ and 21 ms for /g/. The second cut included the burst and the first pitch pulse. For /b/, it was not always possible to obtain a burst-only waveform portion because voicing was occasionally continuous from the carrier phrase, "Teddy said," into the voiced stop syllable. Thus the first waveform cut for /b/ included the burst and the first pitch pulse following the burst with a mean duration of 13 ms. The next cut included the burst and two pitch pulses. The remaining cuts for all stops were made so that the waveform segments included the burst plus three, five, or seven pitch pulses. The total number of test stimuli produced by this editing procedure was 150, with durations ranging from 6 to 111 ms in length.

After the data were collected and analyzed for this experiment, the initial pattern of results suggested that one stimulus should be reexamined to determine if a waveform editing error had occurred. This stimulus was the burst plus one pitch pulse /dɑ/ from speaker RP. Examination of the waveform on the CRT revealed that the last digitized point deviated extensively from zero. Although the resulting click was not easy to hear in the 19-ms stimulus, it was nonetheless perceptible and did appear to interfere with the subjects' correct identification of the stimulus as alveolar.

Audio tapes for the experiment were produced by a computer program that selected the digital waveforms on disk and then output the stimuli through the D/A converter. Identification tapes consisted of six blocks of all 150 truncated stimuli. Stimuli were randomized within blocks, with three seconds between stimuli and seven seconds after each block of 50 stimuli. Two tapes were made for the familiarization task preceding consonant identification. The first tape contained a subset of 60 of the 150 truncated stops, 30 from each talker, in an ordered sequence from /b/ to /d/ to /g/. The second tape contained 25 additional truncated stops selected from each talker in a random sequence.

3. Procedure

The experimental session began with a brief familiarization task, followed by the identification test. The identification test was given on two days. Day 1 included the short familiarization tasks plus the forced-choice identification test for the first three blocks of test trials. On day 2 just the

remaining three blocks of identification were presented. Responses were always recorded by hand on prepared answer forms. The responses were: b, d, g, p, t, or k. Although all stimuli were edited from voiced stop consonants, pilot work had indicated that naive subjects were more comfortable identifying the shortest waveforms with the voiceless stop responses p, t, and k.

Subjects listened to the stimuli through TDH-39 earphones in a quiet testing room. Audio tapes were played back on an Ampex AG-500 tape recorder. A comfortable listening level for the brief stimuli was selected and a single repeated stimulus recorded on each tape was used to calibrate the listening level for all tapes. Separate written instructions were given for each task.

4. Subjects

Subjects were contacted through a laboratory subject pool and were paid for their participation. Subjects were phonetically naive and had no known history of a hearing or speech disorder at the time of testing, as assessed by a pretest questionnaire. Ten subjects participated. One of them omitted so many responses on the first day that she was asked not to return for the second day of testing. Thus, results were analyzed from nine subjects, providing a total of 54 data points for each truncated stimulus.

B. Results

Responses were scored as correct when place of articulation was correctly identified regardless of the voicing feature. Collapsing over all responses and stimuli, subjects identified place of articulation correctly on 93.2% of all trials. This high level of performance is close to that reported by the Tekieli and Cullinan study (1979). Relatively little effect of learning could be observed from day 1 to day 2, a change in the mean percent correct from 92.8% to 93.7%. Percent identification of the stimuli was the same for the two talkers.

Figure 1 shows identification accuracy averaged across vowels for each stop and each talker as a function of the number of pitch pulses in the stimuli. The functions are very similar for both talkers and therefore provide an internal replication of the basic results.

Identification performance for /b/ starts with an average of 90% correct identification for the burst plus one pitch pulse stimulus and rises to nearly 100% correct with the next pitch pulse. The identification functions for /d/ are similar to those for /b/, but they rise somewhat more gradually to 100% correct. The identification performance for /g/ differs from both /b/ and /d/. For the burst-only segment, identification is not very accurate with performance at about 70% correct. Furthermore, /g/ identification functions never reach the 100% correct level even for the longest stimuli (cf. Winitz *et al.*, 1972), as do the /b/ and /d/ functions.

To examine in more detail the relations between the *durations* of the truncated stops and the correct identification of place, the results are plotted separately for all 30 CV's in Fig. 2. Identification performance for all vowel contexts of /b/, shown in Fig. 2, is similar to the average functions shown in Fig. 1. Individual functions for /d/ in Fig. 2 are

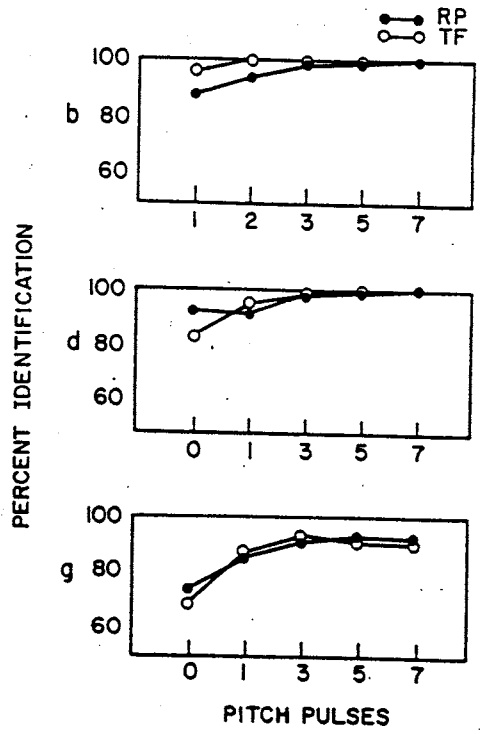

FIG. 1. Percent correct consonant identification for truncated waveforms as a function of the number of pitch pulses. Each panel presents the data by consonant averaged over five vowels shown separately for talkers RP and TF.

also quite similar to the average /d/ functions with two exceptions. First, the /do/ burst was identified correctly only 44% of the time for speaker TF. The short 6-ms duration alone cannot explain this poor level of identification, since there were two /b/'s whose duration was also 6 ms that were identified 86% correctly. The second exception was the /dɑ/ stimulus from talker RP which elicited the only nonmonotonic identification function in the experiment. Apparently, the waveform editing error described above reduced the correct identification of the alveolar stop.

The results for /g/ demonstrated substantially more vowel dependency effects than did those for /b/ or /d/. The most unusual identification functions were obtained for the /gi/ stimuli. Identification was poor at all waveform durations and was less than 50% for the longest (93 ms) stimulus. The identification of /ge/ was also quite poor for the burst-only segments, but increased with three pitch pulses to better than 95% correct. The identification functions for the back vowels with /g/ are quite similar to those for /b/ and /d/. A /g/ before front vowels, with a more palatal place of articulation, was more difficult for subjects to identify than /g/ before back vowels, with a velar place of articulation. In fact, place of articulation errors strongly favored alveolar responses, as predicted from previous research (Fant, 1973; Kewley-Port, 1983).

To summarize, naive subjects were able to identify place of articulation correctly from the initial portions of natural CV syllables. The shortest waveform (burst-only for /d/ and /g/ or burst plus one pitch pulse for /b/) was identified with greater than 80% accuracy for 24 out of the 30

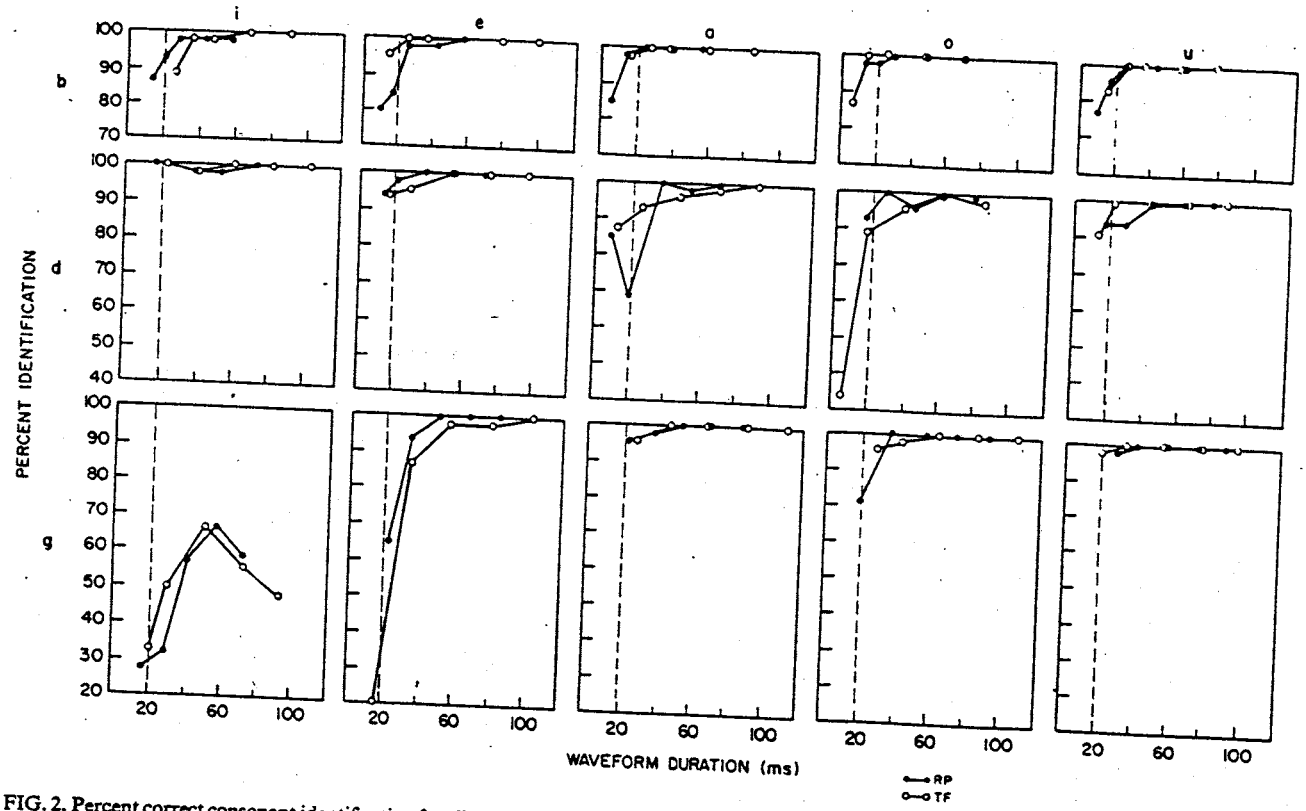

FIG. 2. Percent correct consonant identification for all truncated stop-vowel stimuli as a function of signal duration plotted separately for talkers RP and TF.

CV's examined. The identification functions for /gi/ differed substantially from all others; identification performance was less than 60% correct even for the longest duration stimuli.

C. Discussion

The results from experiment 1 can be used to evaluate both the Blumstein and Stevens' (1979) static onset spectra hypothesis and Kewley-Port's (1983) proposed time-varying features. Blumstein and Stevens tested their onset spectra hypothesis experimentally using visual templates designed to fit spectral sections of the first 25.6 ms of a stop waveform. After windowing, the resulting spectra had an effective duration of 20 ms. To facilitate comparisons, a dashed vertical line has been drawn in each panel in Fig. 2 at 20 ms. The intersection of an identification function with the dashed line indicates the predicted response accuracy for identification of place from the first 20 ms of a test stimulus. The 20-ms identification values were estimated separately for all 30 CV's (smoothing over the bad /dɑ/ stimulus for RP). These values, averaged across vowels and talkers for each consonant, are shown in the first column of Table I.

As shown in the table, the first 20 ms of a stop waveform contain sufficient place information for /b/ (96% correct) and /d/ (94% correct), but not for /g/ (73% correct). Errors for /g/ were not uniformly distributed but occurred mostly for the syllables /gi/ and /ge/; that is, for /g/ before front vowels. Thus the results for /b/ and /d/ appear to be consistent with Stevens and Blumstein's hypothesis of a fixed time window. The results for /g/ are clearly not compatible with this hypothesis.

Consider now the alternative hypothesis that dynamic changes in the distribution of spectral energy over time specify differences in place of articulation. The most important feature for distinguishing /b/ from /d/ in Kewley-Port's three-feature system is the tilt of the spectrum at burst onset, similar to the formulation of Stevens and Blumstein. The Kewley-Port feature system also requires that extended mid-frequency peaks be absent from the running spectra of /b/ or /d/. Since information about spectral tilt of the burst and absence of the mid-frequency peaks resides in the earliest portions of the stop waveform, this feature system implies that identification of /b/ and /d/ should be quite good for very short truncated stimuli. The high level of identification performance observed for the first 20 ms of /b/ and /d/ as shown in Table I is consistent with this hypothesis.

On the other hand, consider the time-varying features for /g/. The definitions of both late $F1$ onset and extended mid-frequency peaks imply that more than 20 ms of a stop waveform is needed to identify /g/ correctly. The identification functions shown in Fig. 2 indicate that relatively high levels of identification of /g/ were achieved at durations of 40 to 50 ms for all vowels except /i/. In order to quantify the

TABLE I. Percent correct identification estimated at 20- and 40-ms stimulus durations averaged across vowels and talkers.

Consonant	Duration 20 ms	40 ms
b	96	99
d	94	98
g	73	90

identification performance of /g/ at longer waveform durations, 40-ms identification values were located on Fig. 2. Forty-millisecond values were chosen because the duration of spectral information displayed in the eight frames in Kewley-Port's (1983) running spectra experiment was 40 ms. The 40-ms identification values for each consonant were averaged over vowels and talkers and are presented in Table I. These results indicate substantial improvement in identification of /g/ from the 20-ms point at 73% correct to the 40-ms point at 90% correct. In the case of /b/ and /d/, performance was close to the 100% asymptote level.

To summarize, the results from naive listeners' identification of truncated CV syllables showed that /b/ and /d/ could be identified at 94% or better from the information contained in the first 20 ms of the waveform. We conclude that either the Stevens and Blumstein's (1978) onset spectrum or the running spectrum of Kewley-Port (1983) is adequate to specify invariant place cues for /b/ and /d/. Velar stops, on the other hand, are poorly identified (73%) from the first 20 ms of waveform and evidently require longer waveform durations for accurate identification. From this result we conclude that the time-varying spectral features proposed by Kewley-Port are likely to prove more successful overall at specifying reliable information for place of articulation in CV syllables. By corollary, Stevens and Blumstein's *fixed* 20-ms integration time is less likely to capture the acoustic cues for specifying velar place, and their later hypothesis (1980) of "a longer time" for identification of velars is more plausible. However, this modified hypothesis for velars is difficult to reconcile with the proposal that onset spectra are detected by innate property-detecting mechanisms in the human auditory system (Stevens and Blumstein, 1978, p. 1367; Blumstein and Stevens, 1979). These property detectors are intended to specify place of articulation as a phonetic feature within distinctive feature theory (Chomsky and Halle, 1968). They are therefore insensitive to details of acoustic structure or phonetic context, being designed to specify place of articulation not only for initial stops, but for final stops and for nasal consonants as well. The role of a *fixed* integration window in Stevens and Blumstein's innate property detector theory is to eliminate the temporal dimension, thereby making the onset spectra more abstract and potential candidates as general place of articulation feature detectors. The results of the present experiment, and those of Blumstein and Stevens (1980) themselves, suggest that the temporal dimension should not be disregarded and that static onset spectra alone are not sufficient to specify cues for place of articulation in stop consonants (see also the recent findings of Blumstein *et al.*, 1982; Walley and Carrell, 1983).

II. EXPERIMENT 2: IDENTIFICATION OF SHORT SYNTHETIC CV's

The purpose of experiment 2 was to conduct a more direct test of the inferences drawn from experiment 1 by perceptual study of *synthetic* CV syllables constructed to distinguish between the rival static and dynamic hypotheses. The idea behind the experiment was, in fact, alluded to earlier by Stevens and Blumstein themselves:

> A stronger test of (the) theory would be to determine whether perception of place of articulation depends on attributes of the gross shape of the spectrum at onset, independent of fine details such as burst characteristics and formant onset frequencies (1978, p. 1367).

Since naive subjects were reasonably successful in identifying place in the truncated natural speech stimuli of experiment 1, we used a similar design here. Two sets of truncated synthetic stimuli were constructed to test the static and dynamic hypotheses. Subjects first participated in an identification task using truncated natural speech CV's. Subjects who could identify place better than chance for the natural speech stimuli were then required to identify stimuli from both synthetic stimulus sets.

In addition to an identification response for each stimulus, we also gathered a confidence rating to indicate whether a subject thought his response was a guess, was surely correct, or was somewhere between the two. We were interested in whether the subjects' confidence ratings would indicate perceptual differences between the synthesized stimuli and the natural stimuli not revealed by the identification responses alone.

A. Stimuli: Synthesis parameters

The success of this experiment depended on clearly stated and executed principles for synthesizing the two stimulus sets, one designed to represent the Stevens and Blumstein static onset spectra (hereafter called S + B) and the other to represent time-varying features in running spectra (hereafter called RS). Each synthesized syllable was modeled after the appropriate natural syllable by *visual* spectral matching techniques. Spectral analysis of the natural stimuli was carried out by linear prediction analysis as implemented in the SPECTRUM program (Kewley-Port, 1979). The S + B and RS stimuli were synthesized on the KLATT digital synthesizer (Klatt, 1980) as implemented in the KLTEXC program (Kewley-Port, 1978).

In addition to stimulus set type, three independent variables—consonant type, vowel type and duration—were manipulated in experiment 2. The consonants /b,d,g/ were each paired with /i,a,u/ producing nine CV syllables, each at three waveform durations: 20, 30, and 40 ms. These durations spanned the 20-ms duration at which /b/ and /d/ were accurately identified in experiment 1 and the 40-ms duration at which /g/ (except for /gi/) was accurately identified.

The natural stimuli consisted of the nine base syllables spoken by talker RP. The SPECTRUM program was used to calculate the onset spectrum of each syllable by a procedure identical to that used by Blumstein and Stevens (1979), with the exception of substituting a one-half Hamming window for a one-half Kaiser window. All onset spectra were visually examined on a CRT graphics display. Only natural CV's whose spectra clearly fit the overall template descriptions of diffuse-rising, diffuse-falling, or compact, *and* which appeared to meet all the template rules described by Blumstein and Stevens (1979), were accepted. In addition, the onset spectra were required to include at least a small $F1$ peak, so that the resulting synthesized waveforms would all have a

voiced component. The running spectral display for each syllable was also examined to see that it contained good exemplars of the time-varying features. If all criteria were not met, another utterance spoken by talker RP in the same recording session was selected. In the end, five of the nine natural CV's, /bi,di,bɑ,bu,gu/, were taken from the stimuli used in experiment 1. The natural syllables were then edited digitally at zero crossings to approximate the 20-, 30-, and 40-ms waveform durations as closely as possible. The average durations proved to be 21, 30, and 39 ms, respectively.

For the synthetic stimuli, the overall strategy was to keep as many of the synthesis parameters as possible the same between the S + B and RS sets, while incorporating differences in the static versus dynamic acoustic properties. To accomplish this, the KLATT synthesizer was configured as a parallel formant synthesizer with six formants (Klatt, 1980). Glottal resonance characteristics were shaped for talker RP's /i/ and then kept constant. The fundamental frequency was set to a constant 100 Hz. Synthesis parameters always terminated exactly at the 20-, 30-, or 40-ms durations.

The synthesis of the S + B stimuli was accomplished in several steps. With the KLATT synthesizer, it was possible to generate a steady-state stimulus such that its spectrum at any point matched the overall shape of the calculated 25.6-ms onset spectrum of the original natural CV. We were concerned, however, that a rising $F1$ transition might be a necessary cue for perceiving the stimuli as stop consonants (Delattre *et al.*, 1955; Stevens and House, 1956; Blumstein and Stevens, 1980). However, Kewley-Port (1982) observed that a rising $F1$ transition cannot always be measured in stop-vowel syllables. In particular, $F1$ transitions measured for the vowels /i/ and /u/ in this experiment were nearly flat. Since it was not clear what kind of $F1$ transitions should be used in the short synthetic S + B stimuli, we carried out a pilot experiment using the S + B stimuli synthesized with and without $F1$ transitions to determine if listeners would judge stimuli with $F1$ transitions as more stop-like than stimuli with $F1$ steady states (see Kewley-Port, 1980 for a more detailed description of this study). Since the outcome showed a slight 7% advantage for stimuli containing $F1$ transitions, we used the $F1$ transition values measured from the natural CV's as synthesis parameters for the S + B stimuli. All higher formants were steady-state.

Figure 3 shows the match between the onset spectra of the final synthesized S + B stimulus for /gɑ/ and the natural /gɑ/ stimulus. A good spectral match was obtained by adjusting only the bandwidth and amplitude parameters for all spectral peaks calculated from the original linear prediction onset spectrum. The amplitude of the voicing source (AV) was always set to its maximum value.

Synthesis parameters for the RS stimuli were derived from previous studies of RP's stop-vowel syllables. The running spectra, as shown in Fig. 4, were produced by calculating the linear prediction smoothed spectra at 5-ms intervals following the procedures described in Kewley-Port (1983). The first frame shows the stop release burst whose onset was always positioned at the center of the 20-ms Hamming window. The other sources of parameter information were the

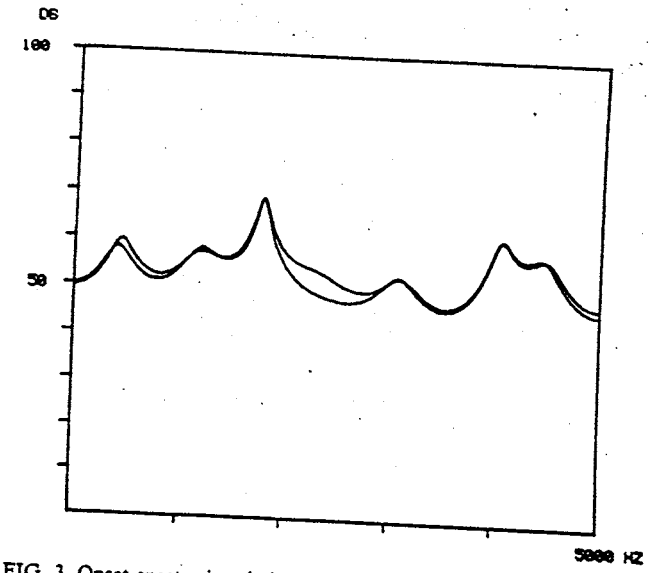

FIG. 3. Onset spectra in relative dB for the first 25.6 ms of the natural speech /gɑ/ and the S + B synthetic /gɑ/.

average formant transitions and VOT values calculated for five repetitions of each of RP's syllables (Kewley-Port, 1982, Appendix). The synthesis procedures always started by carefully matching the burst frame because it contains the spectral tilt information in the running spectra analysis. After the burst frame, the average VOT values were approximated as

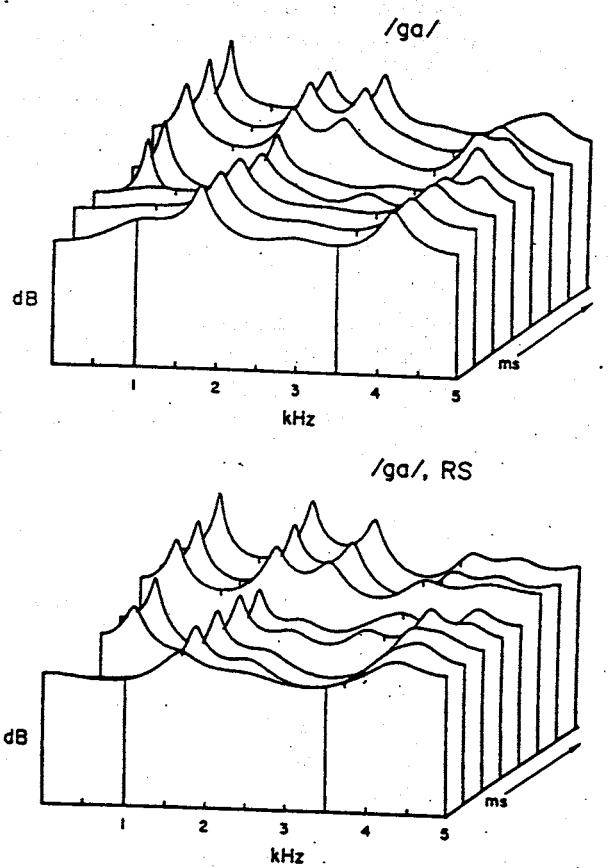

FIG. 4. Running spectral displays of the first 40 ms of the natural speech /gɑ/ (top panel) and the RS synthetic speech /gɑ/ (bottom panel).

/b/ = 0 ms, /d/ = 10 ms, and /g/ = 20 ms using the aspiration source in the synthesizer. The spectral shape of these and succeeding frames was determined by inserting the average values of the transition parameters (frequency and duration) after the release burst. The amplitude of the voicing source always started at 5 dB below the maximum, and increased to the maximum in 5 ms. In addition, careful spectral matching of the voiceless frames for /g/ was often needed in order to preserve the mid-frequency peaks following the burst frame. Figure 4 shows the running spectra for the natural /gɑ/ and the final synthesized RS /gɑ/. Thus the overall procedure for generating the RS stimuli was not based on frame-by-frame spectral matching of the natural and synthetic stimuli. While the burst frames were spectrally matched on one frame for /b/ and /d/ and on the four voiceless frames for /g/, the remaining synthesis parameters were determined essentially by rule using average values of the formant transitions.

Thus whatever merits or weaknesses there were in the synthesis of the S + B and RS stimuli, the procedures for both sets were developed on the basis of visual spectral matching alone. Furthermore, we note that the spectral information contributed by the following vowel context was equally represented in both the RS and S + B 20-ms long stimuli.

B. Procedure

The testing procedures were controlled on-line by a PDP-11/05 computer. Stimuli were output at a 10-kHz sampling rate through 12-bit D/A converters, low-pass filtered at 4.8 kHz and then presented over TDH-39 earphones. The output amplitude was calibrated across sessions at a comfortable listening level. Button press responses were collected from up to six subjects at a time and stored on disk. The basic design of the present experiment was intended to follow closely that of experiment 1, allowing for the appropriate changes from audio tape recordings and written responses to an on-line computer controlled perceptual experiment.

The testing procedures were spread out over two days. Day 1 served to screen subjects audiometrically and to train and test them with the natural speech stimuli. On day 2, subjects identified only the synthetic stimuli. Subjects were excluded from the experiment after day 1 if they identified the natural stimuli essentially at chance, that is to say, if any of the nine natural speech syllables, averaged over all durations, was identified correctly less than 40% of the time.[1]

Responses were collected on a response box with feedback lights. Because voicing was not an experimental variable, the consonant response buttons were labeled "B/P" for bilabial, "D/T" for alveolar, and "G/K" for velar. Three additional buttons were labeled with the confidence rating responses as "very sure," "sure" or "guess." Each trial was signaled by a cue light, and both a consonant response and a confidence rating were collected for each stimulus on each trial in the experiment.

On day 1, each subject was screened by an audiometric test for the octave frequencies from 500 to 8000 Hz at a sound pressure level of 20 dB (ANSI-1969) using a Grason–Stadler model 1701 audiometer. Two familiarization tasks were then conducted using the 27 truncated natural speech stimuli. The first, called "cued familiarization," required listening to all 27 natural stimuli once while a light signaled the correct place response. In the second familiarization task, two repetitions of each natural stimulus were presented to the subjects randomly for identification with feedback. Subjects were then instructed on how to carry out the identification task and enter their confidence ratings. For the natural set of stimuli, subjects heard five blocks of two repetitions of each stimulus for a total of ten trials per stimulus. Stimulus presentation was paced to the slowest subject's responses. No feedback was given in this test.

Subjects returning on day 2 were not told about differences in the nature of the stimuli to be presented. Instructions merely stated, "You will be listening to additional short consonant stimuli one at a time." No familiarization trials were presented on day 2 and subjects began the identification test directly, having previously listened to only the natural speech stimuli on day 1. The 27 S + B stimuli and the 27 RS stimuli were fully randomized within one stimulus block of 54 trials. Subjects listened to ten blocks of stimuli. Each subject therefore provided a total of ten responses to each stimulus. Subjects were obtained through a laboratory subject pool and were paid for their services. None of the subjects who participated in experiment 1 was contacted for this experiment.

C. Results

Of the 21 subjects tested, one subject did not pass the audiometric screening test and was dropped from further testing. Ten subjects did not achieve the 40% correct level of performance with the natural stimuli on day 1 and were asked not to return for testing on day 2.[2] Thus data were collected from all three sets of stimuli for ten subjects, resulting in 100 data points per stimulus.

To assess the contributions of the variables to identification of place of articulation, a four-way analysis of variance over stimulus type, consonant, vowel, and stimulus duration was conducted. When appropriate, one-way analyses of variance were calculated using a Scheffé post-hoc analysis at the $p < 0.05$ level. Probability levels greater than 0.05 were considered nonsignificant.

A summary of the results obtained for the three stimulus sets is shown in Table II. The first row presents the percent correct consonant identification for each stimulus set

TABLE II. Percentages of response measures collected in experiment 2 presented for each stimulus type averaged over vowel and consonant type, and stimulus duration.

Response measures	Natural speech	Synthetic, RS	Synthetic, S + B
Correct consonant identification	94	78	68
Guess rating (−)	4	3	11
Sure rating (+)	21	21	26
Very sure rating (+ +)	75	76	63

averaged over the consonant and vowel types and stimulus duration. Subjects showed high levels of accuracy in identifying place from very brief initial portions of natural stop consonants; overall percent correct was 94%. Performance levels for both sets of synthetic stimuli fell below that of the natural speech stimuli. Consonants were identified better with the RS stimuli (78% correct) than with the S + B stimuli (68% correct). The post-hoc analysis showed that the stimulus types—natural, RS, and S + B—were significantly different [$F(1,2) = 49.74, p < 0.001$]. Consonants were identified more accurately for the natural speech stimuli than for the RS synthetic stimuli, and more accurately for the RS stimuli than for the S + B synthetic stimuli.

The next three rows in Table II show the percentage of each confidence rating response obtained for each stimulus set. These results show that subjects were equally confident of their responses to the natural and RS stimuli, but were less confident of their responses to the S + B stimuli. Evidently, there were aspects of the natural and RS stimuli that subjects judged as similar to each other but dissimilar to the S + B stimuli.

The main effects of each of the experimental variables—consonant type, vowel type, and stimulus duration—on percent correct consonant identification for each stimulus set are shown in Fig. 5. Performance decreased from the natural to the RS and then again from the RS to the S + B stimuli across all variable types but one [$F(2,8) = 78.28$, $p < 0.001$]. For /b/ syllables, the S + B set was better than the RS set.

The effects of each experimental variable were then examined in more detail. No differences were observed in identification performance among /b/, /d/, and /g/ [$F(2,8) = 1.05$, NS] nor among the three waveform durations [$F(2,8) = 0.88$, NS]. For the vowels, differences in performance level were significant [$F(2,8) = 67.63, p < 0.001$]: consonants were identified better before the vowel /ɑ/ (93% correct) than before /u/ (80% correct) or before /i/ (68% correct). Vowel type formed three distinct groups in the post-hoc analysis [$F(1,2) = 42.26, p < 0.001$]. Vowel context, therefore, had an important effect on the identification of consonants. It is interesting that the vowel /ɑ/, often used in synthetic speech perception studies of stop-vowel syllables, provided the most reliable context for stop identification.

The identification results are broken down separately by individual stimuli in the nine panels of Fig. 6. This figure illustrates the major sources of the variation observed in the average identification functions shown in Fig. 5.

The ordered performance levels of the natural, RS and S + B stimuli were preserved for all conditions except for the consonant /b/ and for the syllable /gi/. Figure 6 shows that consonant identification was better for the S + B compared to the RS sets for /bi/ and /bu/. (/ba/ was identified perfectly for all sets.) A post-hoc spectral analysis of the RS /bi/ and /bu/ stimuli revealed an additional variable previously overlooked in our earlier analyses. Although the spectral tilt and shape of the burst frames were carefully matched, the relative levels of signal energy in the burst frame differed between the natural and RS stimuli: The bursts for RS /bi/ and /bu/ were 6 dB higher than the burst for their natural counterparts. The higher amplitude bursts resulted from a synthesis rule implementing a constant voicing source for all synthesized syllables. This rule, intended to keep the source constant across place, was arbitrary and evidently resulted in an artifact in these stimuli.

Further study of burst energy for five repetitions of talker RP's stops before eight different vowels revealed systematic differences in amplitude across different places of articulation. These differences were not represented in either the RS or S + B sets of stimuli. We do not, of course, know for certain that place of articulation in RS /bi/ and /bu/ was misidentified because of the unnaturally loud bursts. Nevertheless, this possibility served as part of the motivation for the next experiment.

The effect of vowel context was not uniform across the consonants (see Fig. 6). In fact, consonant and vowel types interact [$F(4,24) = 26.22, p < 0.001$], as do consonant type, vowel type, and stimulus set [stimulus × consonant, $F(4,24) = 36.63$, $p < 0.001$; stimulus × vowel, $F(4,24) = 7.07, p < 0.001$; stimulus × consonant × vowel, $F(8,32) = 10.77, p < 0.001$]. Evidently, individual consonant-vowel syllables contributed differentially to the overall effects of the ordered difference in performance levels between the natural, RS, and S + B stimuli.

As shown in Fig. 6, variation in waveform durations

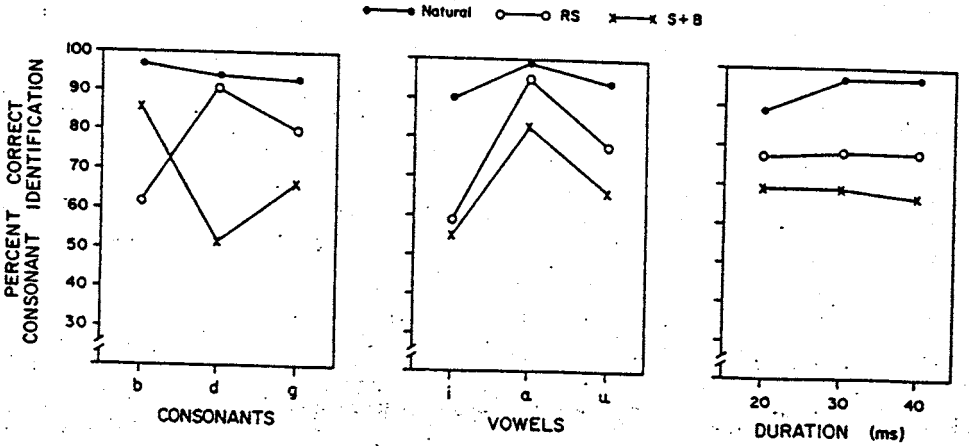

FIG. 5. Percent correct consonant identification shown separately for each stimulus type, natural, RS, and S + B. Each panel displays results for one independent variable averaged over the other two experimental variables.

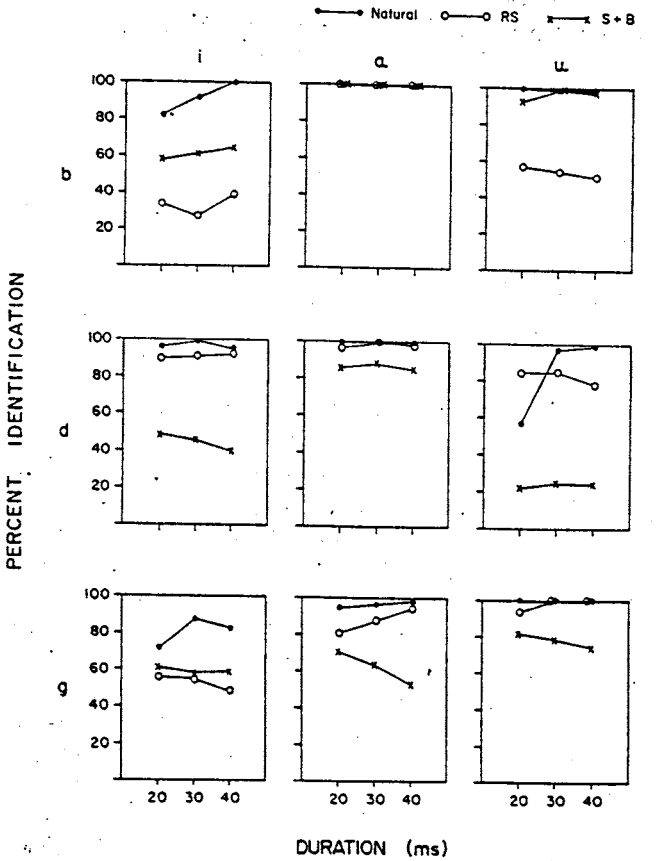

FIG. 6. Percent correct consonant identification plotted separately for each experimental variable, stimulus type, vowel context, stimulus duration, and consonant.

had a very small effect on identification performance. From experiment 1, we had expected that several of the natural stimuli (especially /gi/) would show an improvement in consonant identification with stimulus duration. This result was, in fact, observed since /bi/, /du/, and /gi/ all showed increased identification performance with longer duration waveforms. For the synthetic RS and S + B stimuli, however, consonant identification did not improve with longer stimulus durations.

Finally, if the confidence rating scale was reliable, higher confidence ratings should correlate with an increase in correct responses. We examined this relation by calculating the conditional probability of obtaining a correct response C, for each rating category R_j, as $P(C/R_j)$. Subjects' confidence ratings were indeed highly correlated with their ability to identify place of articulation correctly in the three sets of stimuli (see Kewley-Port, 1980 for more details of this analysis). We also calculated the conditional probabilities of obtaining a rating R_j for only the correct responses. These conditional probabilities confirmed the rank order effects displayed in Table II across the three stimulus sets. For example, subjects were equally confident of their correct responses to the natural and RS stimuli and were overall less confident of their responses to the S + B stimuli.

D. Discussion

The results of experiment 2 showed that listeners correctly identified place on 68% of the S + B stimulus trials. Of the nine S + B syllable types examined, only three /bɑ/, /bu/, and /dɑ/, were identified better than 85% correct. The remaining six CV's averaged only 55% correct identification. Evidently, subjects cannot reliably identify place of articulation from information contained in only the overall gross shape of the onset spectra of short CV waveforms.

This conclusion is strengthened by the comparison between the S + B stimuli and the RS stimuli. Since the RS stimuli were modeled after the same natural speech stimuli as the S + B stimuli, their *onset spectra* should be good exemplars of the onset spectra for the three places of articulation. A post-hoc comparison of the onset spectra for the RS stimuli and the S + B stimuli showed that the RS onset spectra were, indeed, good place exemplars—in fact, for half of the comparisons the onset spectra for the RS and S + B stimuli were virtually indistinguishable. Thus both the static onset spectra properties *and* the temporally varying spectral properties are present in the RS stimuli. Since subjects correctly identified place on 78% of the RS trials, they gained 10% in identification performance when the dynamically changing place information was present. Ten percent is not a large overall increase in performance, but the sources of the increase, as shown in Fig. 6, bring the differences between the two stimulus sets into sharper focus. For the three S + B syllables with high levels of identification (/bɑ,bu,dɑ/), two of the RS stimuli, /bɑ/ and /dɑ/, also had high levels of identification. But for the six S + B stimuli with poor levels of identification, four of the RS stimuli showed substantially increased identification (/di,du,gɑ,gu/). In fact, the average identification scores for these six stimuli increased from 55% for the S + B set to 75% for the RS set, a gain of 20%.

Moreover, if we omit RS /bi/ and /bu/ (for which a synthesis error was made) the RS stimuli were identified correctly on 87% of all trials. For half of the CV's, the identification functions for the natural speech and RS stimuli were virtually indistinguishable (see Fig. 6). Since the synthesis procedures for the RS stimuli were primarily rule-governed rather than based on frame-by-frame spectral matching, the difference of 7% between 87% for the RS stimuli and 94% for natural stimuli suggests that the synthesis rules did a reasonably good job of capturing the dynamic acoustic information that specifies place of articulation in these syllables.

Finally, the confidence ratings (see Table II) establish that subjects were equally confident about identification of the natural and RS stimuli, but were less confident of their responses to the S + B stimuli. The confidence ratings do not appear to reflect the "naturalness" of the stimuli since in debriefing subjects never reported that the synthetic stimuli heard on day 2 differed qualitatively from the natural stimuli on day 1. Rather, it appears that the fine temporal detail of the natural and RS stimuli provide salient properties which subjects can use to identify place of articulation accurately and confidently. Overall, the poor performance of the S + B stimuli, particularly for /d/ and /g/, strongly suggests that the acoustic properties specified by onset spectra are *not* sufficient cues to place of articulation.

III. EXPERIMENT 3: IDENTIFICATION OF SHORT LPC SYNTHESIZED CV'S

The previous experiment tested the adequacy of onset spectra as acoustic cues to place of articulation in stop-vowel syllables. While the outcome was reasonably clear, the experiment could be criticized on two counts. First, the visual matching procedure employed in synthesis might be subject to experimenter bias. Second, unnatural relative levels of energy present in the bursts of all the RS and S + B synthetic stimuli might have influenced the results. The present experiment was designed to replicate the previous experiment using procedures designed to counter these criticisms.

Two sets of stimuli were synthesized from linear prediction coefficients derived from the same natural stimuli as used in experiment 2. One set was synthesized from the coefficients specifying only the onset spectra as defined by Stevens and Blumstein: Spectral information was constant throughout the stimulus. These stimuli will be referred to as the onset spectra (OS) stimuli. The other set was synthesized with the linear prediction coefficients used to produce running spectra (see Fig. 4). These coefficients were updated every 5 ms. These synthesized stimuli will be referred to as the time-varying (T-V) stimuli.

In linear prediction synthesis, fundamental frequency and overall amplitude are controlled independently of the analysis coefficients. For both the OS and T-V stimuli, the fundamental frequency was set to 100 Hz. The relative amplitude of both sets was adjusted so that the energy present in the first 20 ms of each syllable in the three stimulus sets was equal. Thus these stimuli were synthesized by computer algorithms quite independently of the experimenter except for matching rms energy in the first 20 ms. Furthermore, any effect that burst energy might have as an acoustic cue to place of articulation would be held constant across all stimulus types.

In order to test the major difference between Stevens and Blumstein's (1978) onset spectra account and Kewley-Port's (1983) time-varying feature analysis, we will focus our analyses on the 20-ms stimuli. The onset spectra, burst energy, and vowel context information represented spectrally was the *same* for each 20-ms CV in the natural, OS, and T-V sets. (In contrast, the 40-ms natural and T-V sets preserved more vowel context information than the 40-ms OS set.) Thus the primary difference between the two 20-ms synthetic stimulus sets was the static or dynamic quality of the spectral information used to specify place of articulation.

A. Stimuli

Since there was no overall improvement in consonant identification over the three stimulus durations examined in experiment 2, only 20- and 40-ms stimuli were used in this experiment. The natural CV syllables from experiment 2 were also used here.

The linear prediction coefficients previously calculated for stimuli in experiment 2 were used again in this experiment for synthesizing the OS and T-V stimuli. A program (MODSYN) was written by the first author for synthesizing waveforms from a set of reflection coefficients. Arbitrary pitch and amplitude parameters could also be specified. The TWOMUL algorithm of Markel and Gray (1976, Sec. 5.5.2) was implemented with a pulse generator as the pitch source or a pseudo-Gaussian random number generator as the frication source to synthesize the waveforms. An OS stimulus was first synthesized from the coefficients for the onset spectra calculated in experiment 2 along with a gain estimate and a pitch of 100 Hz. After synthesis, the energy in the first 20 ms of the stimulus was calculated. If this energy did not match the calculated energy in the first 20 ms of the natural syllable within 1 dB, the input gain was adjusted appropriately and the OS stimulus was resynthesized. The summed energy in the first 20 ms of the natural syllables relative to a 12 bit waveform was measured as: /bi/ = 77 dB, /bɑ/ = 78 dB, /bu/ = 72 dB, /di/ = 68 dB, /dɑ/ = 72 dB, /du/ = 67 dB, /gi/ = 61 dB, /gɑ/ = 61 dB, and /gu/ = 63 dB. The T-V stimuli were synthesized from the coefficients of the running spectra at a 5-ms frame rate as calculated in Kewley-Port (1983) using a relative gain value estimated algorithmically for each frame. The VOT values used in experiment 2 were also employed here. Voiceless frames used the frication source while voiced frames were synthesized with a pitch of 100 Hz. When necessary, the first 20 ms of energy in each T-V stimulus was set to match the natural stimulus by adjusting the relative gain for all frames by a fixed amount.

B. Procedure

The procedure of experiment 3 was identical to that of experiment 2 except that one-third fewer stimuli were presented because the 30-ms duration was dropped. Both identification and confidence rating responses were collected over two days of testing. Subjects were obtained through a laboratory subject pool and paid for their services. None of the subjects who participated in the first two experiments was contacted for this study.

C. Results and discussion

All seventeen subjects participating in the experiment passed an audiometric screening test. One failed to return on the second day, and six others did not achieve the 40% correct level of performance on the natural speech stimuli. Thus ten subjects participated in both days of testing, resulting in 100 data points per stimulus. The same statistical analyses performed in the previous experiment were used here.

Table III summarizes the results. The findings are simi-

TABLE III. Percentage of response measures collected in experiment 3 for each stimulus type averaged over vowel and consonant type, and stimulus duration.

Response measures	Natural speech	Synthetic time-varying (T − V)	Synthetic onset spectra (OS)
Correct consonant identification	94	87	59
Guess rating (−)	9	5	30
Sure rating (+)	23	28	37
Very sure rating (++)	68	67	33

ment 1, Kewley-Port's (1983) hypothesis of time-varying features predicts that identification of velar place of articulation should improve to relatively high levels with an increase in duration from 20 to 40 ms, while performance for /b/ and /d/ should show little change. Furthermore, since 20-ms velars were poorly identified in this experiment, Stevens and Blumstein would presumably predict that velar identification should improve at a 40-ms duration, as the compact shape of the velar onset spectra becomes more salient. These predictions were tested by analysis of variance. In a three-way analysis of place of articulation (/b/ vs /d/), stimulus duration and stimulus type (natural versus T-V), there was a small, though significant, difference of 4% in identification between the natural and T-V stimuli [$F(1,3) = 6.84$, $p = 0.01$], but no significant difference for place of articulation [$F(1,3) = 1.61$, NS] or for duration [$F(1,3) = 2.70$, NS] (see Fig. 9). In a separate analysis of stimulus duration by stimulus type for /g/, there was a significant improvement in identification for the longer stimuli [$F(1,2) = 7.75$, $p < 0.006$], but no difference between the natural and T-V stimuli [$F(1,2) = 3.99$, NS]. These results are entirely consistent with the time-varying features hypothesis. By contrast, the prediction derived from Stevens and Blumstein's approach that improved velar identification would result with increased duration was not supported for the onset spectra stimuli, since the correct identification for OS /g/ stimuli fell slightly from the 20- to 40-ms durations (Fig. 9). We conclude that a static representation of onset spectra in stop consonants is clearly less successful in characterizing the acoustic cues to place of articulation than a time-varying, running spectrum.

IV. GENERAL DISCUSSION

The three experiments presented here have addressed some of the claims made previously by Stevens and Blumstein and Kewley-Port concerning the acoustic cues which specify place of articulation in initial stop-vowel syllables. Both approaches have stressed that the cues are found in the brief initial portions of the stop-vowel waveform, and that the cues are context-free. Experiment 1 supported the claim that the cues are found in the initial 20- to 40-ms of a natural stop-vowel syllable and was in general agreement with the earlier study conducted by Blumstein and Stevens (1980) using synthetic speech. The claim that the acoustic cues are invariant with respect to the following vowel context was not addressed here, but has been supported in previous studies (Stevens and Blumstein, 1978; Blumstein and Stevens, 1979; Kewley-Port, 1983).

However, Stevens and Blumstein and Kewley-Port's hypotheses diverge in describing the acoustic cues to place of articulation as static versus dynamic. Experiments 2 and 3 directly tested the Stevens and Blumstein claim that the static cues are specified as integrated onset spectra. This claim was shown to be false because alveolar stops were not identified above chance from onset spectra, while velars were identified rather poorly. In contrast, results of all three experiments showed that place of articulation was specified by the dynamic spectral properties located in the onset of the stop-vowel syllable. In two separate synthesis experiments, brief, temporally changing stimuli were perceived similarly to their natural speech counterparts: Identification of place of articulation was relatively high. Subject's confidence ratings were also comparable for the natural and synthetic items. We conclude that a dynamic representation of acoustic cues to place of articulation is better than a static representation.

On the other hand, this research cannot be considered to be a perceptual test of Kewley-Port's three time-varying features. While experiment 2 used these features as the basis of synthesis procedures, the results were not clear-cut, perhaps due to the unnatural relative energies of the stimuli. We showed, however, that the results of the perception studies were consistent with the definition of the time-varying features as based in the acoustic theory of speech production (Fant, 1960, 1973), and with specific predictions derived from an earlier analysis experiment (Kewley-Port, 1983).

Stevens and Blumstein, however, have not viewed the dichotomy between static and dynamic place cues in the same manner in which we have discussed it. For Stevens and Blumstein (1978, 1981) the static onset spectra are the "primary," invariant correlates of place of articulation. The dynamic cues are the "secondary" correlates found in the context-dependent formant transitions (Blumstein and Stevens, 1979, p. 1015). Formant transitions are said merely to "provide the acoustic material that links the transient events at the onset to the slowly varying spectral characteristics of the vowel" (Blumstein and Stevens, 1980, p. 660). This point of view is essentially identical to that of Cole and Scott (1974). However, the theoretical arguments of Fant (1968, 1973) and the results of the running spectral analysis presented earlier by Kewley-Port (1983) demonstrate that the distinction between primary, static properties and secondary, dynamic properties is arbitrary and empirically unjustified (cf. also the perceptual study by Walley and Carrell, 1983). The primary acoustic correlates of place of articulation appear to be the time-varying spectral properties that reflect the movements of articulatory release—whether the rapid release typical of labial and alveolar stops or the somewhat slower release typical of velars (Fant, 1968, p. 223).

More generally, the present experiments demonstrate that to identify the context-free with the static, and the dynamic with the contextually-dependent, is false. If the spectral correlates of a phonetic segment derive from an articulatory gesture, and if the essence of gesture is structural change, then the spectral correlates must reflect that change. We propose that a characteristic pattern of change may be invariant across a variety of contexts. Dynamic articulation does not necessarily imply a lack of invariance in the acoustic signal.

Finally, if the Stevens and Blumstein hypothesis of static cues followed by dynamic ones may be called a *serial* model of place perception, might not the present research suggest a *parallel* model? That is, perhaps "primary" static onset spectra and "secondary" dynamic cues are perceived independently, but simultaneously. However, this seems artificial and unnecessarily mechanistic. Surely it is simpler and, in some sense, more natural to acknowledge that the unitary, dynamic gestures of stop articulation are reflected in the co-

herent, time-varying properties of running spectra. If these coherent spectral changes are precisely what the listener perceives, we then have a naturally unified account of both perception and production.

ACKNOWLEDGMENTS

We are grateful to Katherine S. Harris, Dennis H. Klatt, and Lawrence J. Raphael for their comments on earlier versions of this work. Some of these experiments were presented before the Society at the 100th meeting in Los Angeles, CA in 1980 and the 102nd meeting in Miami Beach, FL in 1981. The first two experiments were part of a doctoral thesis submitted to the Graduate Center of The City University of New York by the first author. This research was supported by the National Institutes of Health, Research Grant NS-12179 to Indiana University in Bloomington, IN, and, in part, by the National Institute of Child Health and Development, Research Grant HD-01994 to Haskins Laboratories, New Haven, CT.

[1]An earlier pilot study for this experiment revealed that some subjects completely misidentified some truncated syllables. For example, /gɑ/'s were identified as alveolars. Subjects in experiment 1 did not respond this way. Apparently, naive subjects differ in their ability to correctly extract place of articulation information when stimulus duration is always shorter than 40 ms.

[2]See footnote 1.

[3]Note that Kewley-Port's (1983) time-varying features would not have classified all the OS stimuli as having the same place of articulation as the original natural stimuli. This is due to the differences between time-varying features and the template analysis mentioned in the introduction. In particular, different identifications of place would be made for the OS /dɑ/, /du/, and /gɑ/ stimuli. Thus the time-varying feature analysis of the OS stimuli does not lead to any clear predictions for the outcome of this experiment.

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.

Blumstein, S. E., Isaacs, E., and Mertus, J. (1982). "The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants," J. Acoust. Soc. Am. 72, 43–50.

Chomsky, N., and Halle, M. (1968). *The Sound Pattern of English* (Harper and Row, New York).

Cole, R. A., and Scott, B. (1974). "Toward a theory of speech perception," Psychol. Rev. 81, 348–374.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (1952). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. 24, 597–606.

Delattre, P., Liberman, A. M., and Cooper, F. S. (1955). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. 27, 769–773.

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. (1968). "Analysis and synthesis of speech processes," in *Manual of Phonetics*, edited by B. Malmberg (North-Holland, Amsterdam), pp. 173–277.

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

Fischer-Jørgensen, E. (1972). "Tape cutting experiments with Danish stop consonants in initial position," *Annual Report VII* (Institute of Phonetics, University of Copenhagen), pp. 104–175.

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

Kewley-Port, D. (1978). "KLTEXC: Executive Program to implement the KLATT software speech synthesizer," *Research in Speech Perception Progress Report No. 4* (Dept. Psychol., Indiana University, Bloomington, IN), pp. 235–246.

Kewley-Port, D. (1979). "SPECTRUM: A program for analyzing the spectral properties of speech," *Research in Speech Perception Progress Report No. 5* (Dept. Psychol., Indiana University, Bloomington, IN), pp. 475–492.

Kewley-Port, D. (1980). "Representations of spectral change as cues to place of articulation in stop consonants," *Research in Speech Perception Technical Report No. 3* (Dept. of Psychol., Indiana University, Bloomington, IN).

Kewley-Port, D. (1982). "Measurements of formant transitions in naturally produced stop consonant-vowel syllables," J. Acoust. Soc. Am. 72, 379–389.

Kewley-Port, D. (1983). "Time-varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am. 73, 322–335.

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

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). "Perception of the speech code," Psychol. Rev. 74, 431–461.

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954). "The role of consonant-vowel transitions in the perception of the stop and nasal consonants," Psychol. Monogr. 68 (8, Whole No. 379), 1–13.

Liberman, A. M., and Studdert-Kennedy, M. (1978). "Phonetic Perception," in *Handbook of Sensory Physiology: Perception*, edited by R. Held, H. Liebowitz, and H. L. Teuber (Springer-Verlag, New York), pp. 143–179.

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

Markel, J. D., and Gray, A. H. (1976). *Linear Prediction of Speech* (Springer-Verlag, New York).

Schroeder, M. R., Atal, B. S., and Hall, J. L. (1979). "Optimizing digital speech coders by exploiting masking properties of the human ear," J. Acoust. Soc. Am. 66, 1647–1652.

Stevens, K. N. (1975). "The potential role of property detectors in the perception of consonants," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic, New York), pp. 303–330.

Stevens, K. N. (1980). "Acoustic correlates of some phonetic categories," J. Acoust. Soc. Am. 68, 836–842.

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., and Blumstein, S. E. (1981). "The search for invariant acoustic correlates of phonetic features," in *Perspectives on the Study of Speech*, edited by P. D. Eimas and J. Miller, (Erlbaum, Hillsdale, NJ), pp. 1–38.

Stevens, K. N., and House, A. S. (1956). "Studies of formant transitions using a vocal tract analog," J. Acoust. Soc. Am. 28, 578–585.

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.

Walley, A. C., and Carrell, T. D. (1983). "Onset spectra and formant transitions in the adult's and child's perception of place of articulation in stop consonants," J. Acoust. Soc. Am. 73, 1011–1022.

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.