Perceiving vowels in the presence of another sound: Constraints on formant perception

C. J. Darwin
Laboratory of Experimental Psychology, University of Sussex, Brighton, BN1 9QG, England

(Received 9 February 1984; accepted for publication 15 June 1984)

Speech is normally heard against a background of other sounds, yet our ability to isolate perceptually the speech of a particular talker is poorly understood. The experiments reported here illustrate two different ways in which a listener may decide whether a tone at a harmonic of a vowel's fundamental forms part of the vowel. First, a tone that starts or stops at a different time from a vowel is less likely to be heard as part of that vowel than if it is simultaneous with it; moreover, this effect occurs regardless of whether the tone has been added to a normal vowel, or to a vowel that has already been reduced in energy at the tone's frequency. Second, energy added simultaneously with a vowel, at a harmonic frequency near to the vowel's first formant, may or may not be fully incorporated into the vowel percept, depending on its relation to the first formant: When the additional tone is just below the vowel's first formant frequency, it is less likely to be incorporated than energy that is added at a frequency just above the first formant. Both experiments show that formants may only be estimated after properties of the sound wave have been grouped into different apparent sound sources. The first result illustrates a general auditory mechanism for performing perceptual grouping, while the second result illustrates a mechanism that may use a more specific constraint on vocal-tract transfer functions.

PACS numbers: 43.70.Dn, 43.66.Jh

INTRODUCTION

It is a curious fact that most experiments on the perception of speech have used only speech as a stimulus. The fact is curious since in the normal course of events we hear speech in the presence of other sounds and other voices. Studying the perception of the speech of a single speaker in isolation has told us a great deal about which acoustic consequences of a speaker's articulation are necessary or sufficient for the appropriate percept, but we know very little about how these acoustic properties are actually extracted from the raw sound wave. The problem is severe, since for any speech recorded outside of the soundproof, anechoic chamber, properties of the recorded soundwave will not correspond in a simple way with properties of the spoken soundwave. Other sound sources and echoes will also contribute to the recording.

A conceptual distinction needs to be made (cf. Bregman, 1978) between properties of the raw soundwave and derived properties that may be *source specific*. The large literature on speech perception has addressed itself to the problem of what source-specific properties are perceptually salient but has almost entirely ignored the problem of how such properties could be extracted from the raw acoustic signal. Yet the problem lies at the very heart of successful speech recognition and has radical theoretical implications. What constraints need to be applied by a process that can separate the speech of a single speaker from other sounds? Are they speech specific or can constraints that apply to sounds in general do the trick? In order to answer these questions, experiments need to be considered on the perception of speech in the presence of other structured sounds.

A. Perceptual grouping of speech sounds

Most experiments on the perception of speech in the presence of other sounds have used one of two types of additional sound—either random noise (Miller and Nicely, 1955; Pickett, 1957) or an additional formant (Darwin, 1981) or voice (Brokx and Nooteboom, 1982; Darwin, 1981; Scheffers, 1983). With random noise, the main perceptual problem is the detection of structure, whereas with an additional formant or voice, the main problem is to group evident structure appropriately (cf. Bregman, 1978; McAdams, 1980; McAdams and Bregman, 1979).

In performing appropriate grouping of formants, it is clear that a common harmonic spacing is influential. Sentences synthesized on a different fundamental frequency from an interfering passage of speech are more intelligible than those of the same fundamental (Brokx and Nooteboom, 1982). The same is true for pairs of simultaneous isolated vowel sounds (Scheffers, 1983); the intelligibility of the vowels is slightly higher when they are synthesized on different fundamentals (80%) than when they are synthesized on the same fundamental (68%). Similarly, when four formants may be grouped in two alternative ways to give both a three-formant syllable and a separate single formant, listeners tend to group together the three formants that share a common fundamental (Darwin, 1981, experiment IV).

The size of the grouping effect by a common fundamental is not large in some experiments, and clearly can have no effect in voiceless speech, or for formants in which the individual harmonics are too close together, relative to the critical bandwidth, to be resolved (but see Bregman *et al.*, 1983). It is likely that other factors also play a role. In Scheffers'

 0001-4966/84/121636-12$00.80

experiment, for example, listeners could identify 68% of vowels presented in pairs on the same fundamental frequency. Grouping by fundamental is of no use in separating the different vowels in this experiment, yet listeners still performed substantially above chance. Constraints other than fundamental frequency must have been used to separate out those formants that could form one of the nine different vowels used in the experiment.

Another factor that could be important in grouping together different formants in running speech is onset time. A formant that starts at a different time from others is less likely to contribute to the phonetic quality of a syllable than if it starts at the same time as the others (Darwin, 1981 experiment IV; cf. Bregman and Pinker, 1978; Dannenbring and Bregman, 1978; Rasch, 1978). A difference in onset time is not, of itself, sufficient reason to separate one formant from the remainder: The first formant of an aspirated stop typically starts later than the higher formants, yet it is well integrated into the vowel. Rather, a difference in onset time provides the potential for a separate perceptual group.

B. Estimating *F*1 frequency

The experiments described in this paper form part of a series (see also Darwin, 1983; Darwin and Sutherland, 1984); examining the perception of sounds that differ in their first formant frequency, in the presence of extra energy at one of the harmonic frequencies close to the first formant. The first formant is of particular interest in the context of the extraction of speech features, since its value almost always has to be inferred from the raw spectral data presented by the ear to the brain. In the first formant region, the individual harmonics of voiced speech are generally spaced by substantially more than the critical bandwidth as recently described (Moore and Glasberg, 1983) and so are resolved as separate peaks. In general, there is no actual peak present at the frequency corresponding to the first formant frequency. Yet it is clear that the percept of vowel quality is not given by the frequency of the most intense harmonic (Karnickaya *et al.*, 1975; Carlson *et al.*, 1975; Assmann and Nearey, 1983); rather, some smoothing function is applied to the spectral peaks or some weighted average formed of them in order to estimate the formant frequency. The need to derive the formant frequency raises the following question: Which frequencies should be included in the estimation procedure? Can some of the frequency components present at a particular time be excluded from the estimation procedure by virtue of, for example, their having a different starting time from the other frequencies present in the vowel? Our recent experiments have indeed shown that they can.

C. Perceptual grouping of first formant harmonics

When extra energy is added to a vowel at a harmonic frequency close to the first formant, the vowel quality changes, indicating that the extra energy is being perceived as part of the vowel and is contributing to the estimated frequency of the first formant. However, if this extra energy is made to start earlier than the main vowel, listeners can perceptually subtract it out from the vowel, yielding the original vowel quality (Darwin, 1983). Such perceptual subtraction of a leading tone could be due to two types of mechanism: peripheral adaptation at the frequency of the leading tone (cf. Summerfield *et al.*, 1981, 1984), or perceptual grouping on the basis of onset-time differences. Darwin and Sutherland (1984) have argued that perceptual grouping plays a substantial role in the effect on the following grounds. First, an additional (500 Hz) tone can be separated from a *short* vowel when it starts at the same time as the vowel but continues after it; with longer vowels, the perceptual separation produced by a leading tone is unchanged, but that produced by a lagging tone is weaker (presumably because the listener has already decided on the vowel quality by the time the tone is heard sticking out at the end of the vowel). Second, the perceptual separation of a leading tone (at 500 Hz) from the following vowel is reduced if the leading portion of the tone is accompanied by a second (1000 Hz) tone that starts at the same time as the leading tone, but which stops as the vowel starts; the 1000-Hz tone tends to form a new perceptual group with the leading portion of the 500-Hz tone, leaving the remainder of the 500-Hz tone free to be incorporated again into the vowel. This result cannot be explained by adaptation.

In the experiments reported here, we pursue two questions relating to the perceptual separation of a tone from a vowel. First, does the perceptual separation due to timing differences only hold when the sound that results from subtracting out the extra energy has a more "normal" vowel spectrum? Second, what are the limits on the amount of energy that may be added at a harmonic frequency and that will still be incorporated into the vowel percept?

I. EXPERIMENT 1

We have shown previously that the change in vowel quality, produced by adding extra energy to a harmonic near the first formant frequency, can be reduced or abolished when the energy starts (Darwin, 1983) at a different time from the main vowel. It is not clear to what extent our results reflect a tendency on the part of subjects to prefer simple vowels (as produced by a formant synthesizer with conventional buzz excitation) over those whose harmonics have been modified in intensity. The previous experiments showed that subjects can use onset and offset time to separate a normally synthesized vowel from an additional tone. But what if a vowel sound is used that gives a normal spectrum *before* the additional tone has been subtracted; will perceptual separation occur, giving an abnormal vowel percept?

In order to assess whether perceptual separation of the tone and vowel is influenced by the resulting vowel spectrum, experiment 1 compares the case where energy is added to a harmonic of a normal vowel, with the case where energy is added to a vowel that has already been *depleted* in energy at that frequency. If a subjective preference for a particular type of vowel spectrum is influencing our results, then we would expect to find that temporal offsets had less of an effect on the separation of energy added to a depleted vowel than on the separation of energy added to a normal vowel. In the former case, *retaining* the extra energy as part of the

vowel percept gives a normal vowel spectrum, while in the latter, *separating* the energy gives the normal spectrum.

A. Method

Perceived vowel quality is estimated by measuring the phoneme boundary between /I/ and /ε/ along a continuum of sounds differing in their first formant. We refer to the first formant frequency that was used to synthesize a sound as its *nominal F* 1. When we add energy to a harmonic of a vowel, its *nominal F* 1 stays constant by definition. Phoneme boundaries for the various conditions are measured in terms of this nominal *F* 1. So, if the addition of energy to a vowel has no perceptual effect on its quality, then the phoneme boundary will stay at the same nominal *F* 1 value. However, if adding energy produces the percept of a higher *F* 1 frequency, then the phoneme boundary will appear at a *lower* nominal *F* 1 value along the continuum. Conversely, if adding energy produces the percept of a lower *F* 1 frequency, then the phoneme boundary will appear at a *higher* nominal *F* 1 value.

1. Synthesis techniques

The basic continuum of sounds was produced by Klatt's software parallel-formant speech synthesis program (Klatt, 1980). The additional 500-Hz tones were produced for each member of a continuum by digitally filtering each member of the original vowel continuum (with an appropriate duration) twice through a 101-coefficient Finite Impulse Response filter, attenuating harmonics other than 500 Hz by at least 56 dB, and leaving the 500-Hz component unchanged in intensity. To produce continua with different levels of additional energy at 500 Hz, the tones so produced were digitally added to or subtracted from the vowels of the original continuum. The filtering introduced a time delay of 10 ms, so, before adding in the filtered tone, the original vowel was also shifted by 10 ms. These digital signal processing operations were all performed with ILS software on the laboratory's VAX 11/780 computer.

2. Stimulus continua

There were 11 different continua, the original continuum and ten others derived from it. Five were made by adding to each member of the original continuum various 500-Hz tones which could either be simultaneous with the vowel or start or stop at different times from it. Each tone raised by 6 dB the level of the 500-Hz component in the vowel to which it was added.

Five more continua were made in a similar way, but this time the tone was added to a vowel continuum which had already been reduced in energy at 500 Hz. The effect of adding in the extra tone was to bring the energy at 500 Hz back to its original level during the vowel.

a. Original vowel. The original vowel continuum had seven members whose first formants differed in equal steps between 375 and 500 Hz. The vowels were synthesized on a constant fundamental frequency of 125 Hz, so that the first formant fell between the third and fourth harmonics. The first formant bandwidth parameter of the Klatt synthesizer was fixed at 70 Hz. The second through fifth formant frequencies were fixed at 2300, 2900, 3800, and 4600 Hz. The vowels were 56 ms long, including a 16-ms rise and fall time, so that the steady state was 24 ms (or three pitch pulses). Short (56 ms) vowels are used in the experiment in order to give offset-time effects a chance to appear.

b. Original vowel plus tone. Five more continua (each with seven members) were constructed from the original continuum by adding a 500-Hz tone to each member. Each 500-Hz tone had the same amplitude and phase as the 500-Hz component of the vowel to which it was added. So each vowel in the new continua had 6 dB more energy at 500 Hz than the vowel in the original continuum from which it was derived. In one continuum, the extra tone was simultaneous with the original vowel; we will refer to this as the *augmented* vowel continuum. The remaining four continua were constructed in a manner similar to the augmented but with different durations and temporal alignments of the added tone. The tone was always present during the vowel; it could either (i) start 32 or 240 ms before the vowel but stop at the same time as the vowel, or (ii) start with the vowel but stop 32 or 240 ms after it. The spectra of the resulting vowels were checked to ensure that the phase relations between the additional tone and the vowel stayed constant to produce the required (6 dB) increase in level. Schematic waveform envelopes and spectra for a sound from near the middle of the original vowel continuum (nominal *F* 1 450 Hz), and the corresponding sound with extra energy starting 32 ms before the vowel are shown on the left of Fig. 1.

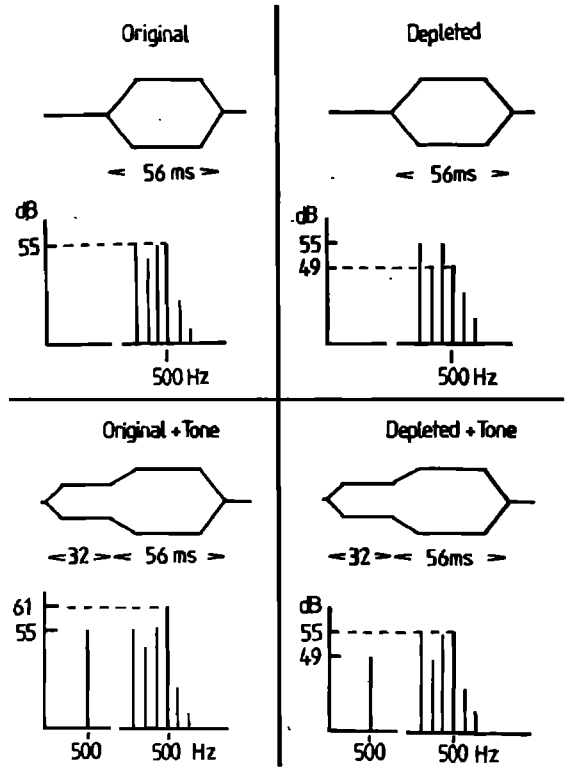

FIG. 1. Spectra and schematic waveform envelopes for stimuli near the middle of four continua used in experiment 1. In the bottom-left panel a tone at 500 Hz has been added to the original vowel (top-left). In the bottom-right panel a 500-Hz tone has been added to a vowel that has already been depleted in energy at 500 Hz (top-right), so that adding in the tone restores the 500-Hz component to its level in the original vowel.

c. Depleted vowel. The *depleted* continuum was constructed from the original by reducing the level of the 500-Hz component of each member by 6 dB. The depletion was accomplished by subtracting from each member of the original continuum one-half of its 500-Hz component.

d. Depleted vowel plus tone. Four more continua were constructed from the depleted continuum by adding various 500-Hz tones. The intensity of the tone added to each sound was appropriate to bring the total level at 500 Hz back to its original value. The tones had the same onset and offset times relative to the main vowel as in the conditions described in Sec. I B. Note that the simultaneous condition here is identical to the original vowel; the original vowel continuum was used in order to reduce the number of conditions to be taken by the subjects.

Schematic waveform envelopes and spectra for a sound from near the middle of the depleted vowel continuum, and the corresponding sound with extra energy starting 32 ms before the vowel are shown on the right of Fig. 1.

3. Procedure

Twelve student subjects, all native speakers of British English and without hearing problems, were tested individually in a soundproof booth over Sennheiser HD-414 headphones. The level of the member of the basic continuum with the lowest first formant was approximately 58 dB(A).

The sounds were produced on line by the laboratory's VAX 11/780 computer (via a microprocessor controlled peripheral—the DEC LPA-11K—at a sampling frequency of 10 kHz, low-pass filtered at 4.5 kHz and 48 dB/oct). Subjects responded on a conventional terminal keyboard with either of the two identification responses ("I" for /I/ and "E" for /ɛ/). If they were not sure of a sound's category on first hearing, they could press "R" to repeat it. Following each key press, the appropriate next sound was played after a pause of 1 s. If the key press occurred while the current sound was being played, the 1-s pause started at the end of that sound. The terminal screen showed the allowed response keys and the current trial number, as well as castigating the use of other keys.

The main experiment was preceded by a demonstration of the basic continuum, the augmented continuum, and the depleted continuum, followed by a practice identification session using ten successive random orderings of those three continua. After the practice session, subjects were told that they might now hear tones mixed in with the vowels, but they were to ignore them and simply report the vowel that they heard. For the main experiment, each subject heard ten successive random sequences of the 77 items (11 continua × 7 steps) and was free to take a rest at any time.

B. Results

Identification functions pooled across the 12 subjects for the 11 continua are shown in Figs. 2 and 3. They are expressed as the percentage of /I/ responses given to each sound, where each sound is referred to by its *nominal* first formant frequency. If additional energy made no difference to vowel quality, the phoneme boundary would be at the same nominal first formant frequency. There are, however, clear differences between the conditions.

In order to pool across subjects without confounding between-subject variability in boundary position with the slope of the individual identification functions, the curves of the individual subjects were aligned around each curve's 50% boundary before averaging. The resulting averaged curve was then plotted at the mean boundary. The slope of each plotted curve thus gives the average slope of the individual identification functions. The slopes do not change substantially across conditions.

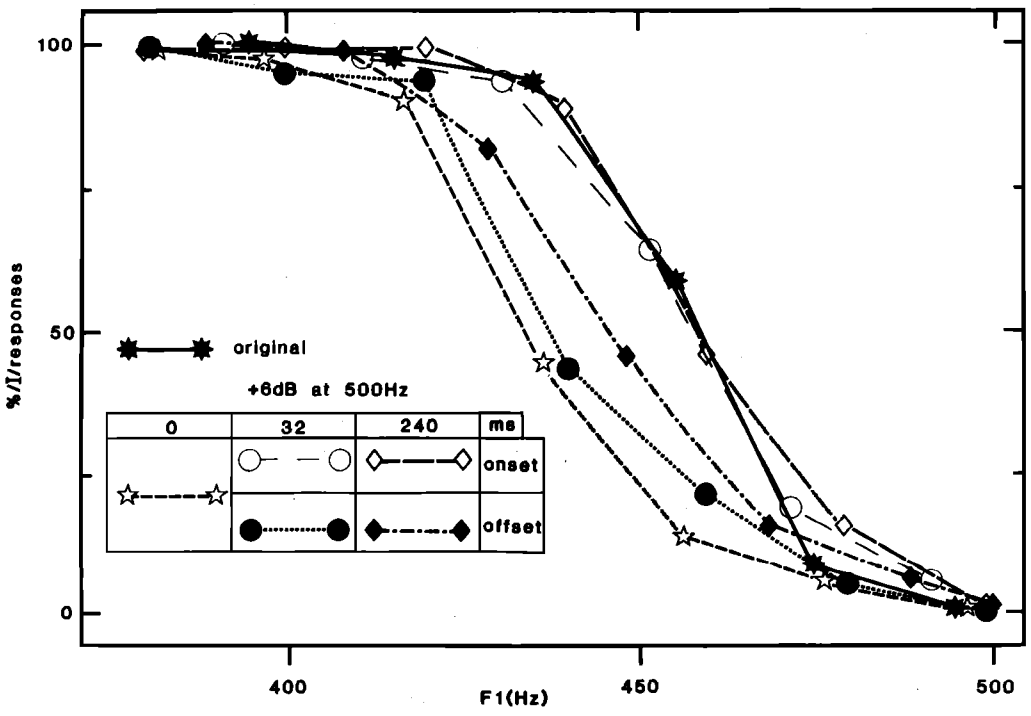

FIG. 2. Mean identification functions from experiment 1 for 12 subjects for $F1$ continua differing in the onset (open symbols) and offset times (filled symbols) of + 6 dB of energy at 500 Hz added to the original vowel continuum.

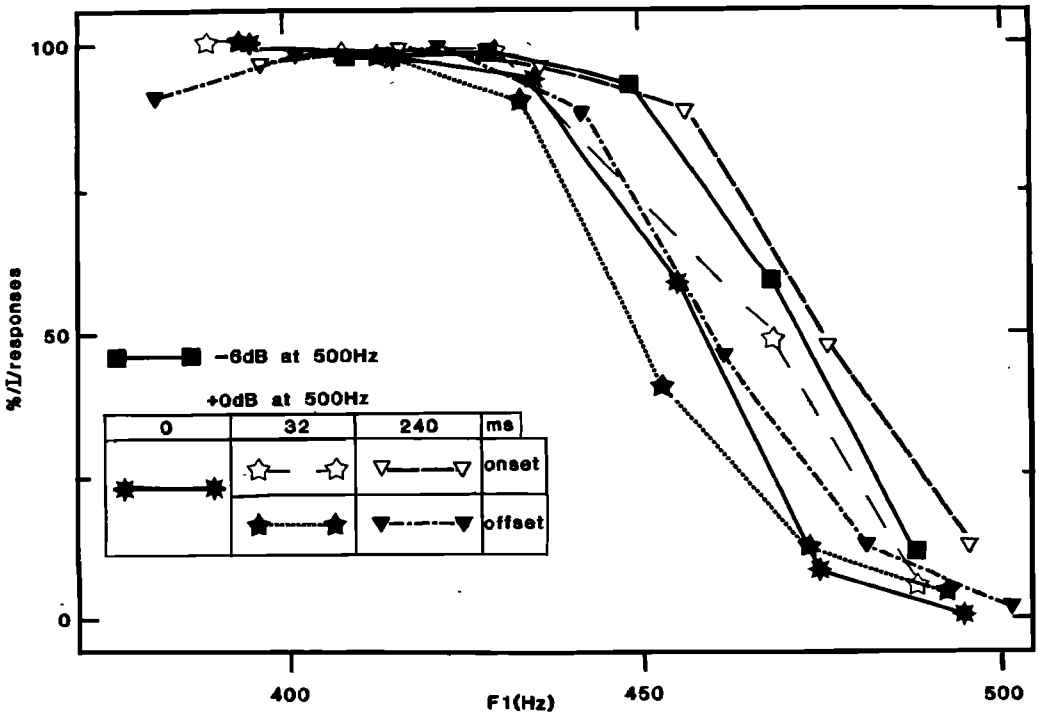

FIG. 3. Mean identification functions from experiment 1 for 12 subjects for $F1$ continua differing in the onset (open symbols) and offset times (filled symbols) of + 6 dB of energy at 500 Hz added to a vowel continuum that has been depleted by 6 dB at 500 Hz.

The phoneme boundaries for each subject in each condition were found by probit analysis, and the average phoneme boundaries with their standard errors are shown in Fig. 4 for all 11 conditions. The variability of the phoneme boundary across subjects increases in the conditions with added tone compared with the original condition. The standard deviation of particular condition's boundary across the 12 subjects ranged from 9.7 Hz for the original condition to 18.7 Hz for one of the long offset-time conditions.

1. Onset differences

First, this experiment confirms previous findings that a harmonic that starts before the main vowel makes less of a contribution to vowel quality than one that starts at the same time as the remaining harmonics. In Fig. 4(a), the original vowel continuum gives a phoneme boundary at the value shown by the filled triangle and the horizontal line. Adding energy at 500 Hz that starts and stops at the same time as the main vowel augments the energy at 500 Hz by 6 dB. This augmented continuum gives a phoneme boundary with a nominally lower $F1$ value [$t(11) = 6.76; p < 0.001$] showing that the extra energy is making a contribution to the quality of the vowel. But when the additional energy starts 32 or 240 ms earlier than the main vowel (solid lines), it ceases to make any contribution to the vowel quality—the phoneme boundary moves back from the augmented value [$t(11) = 5.71$, $p < 0.001$ ms; $t(11) = 5.95, p < 0.001$ for 240 ms] to regain its original position.

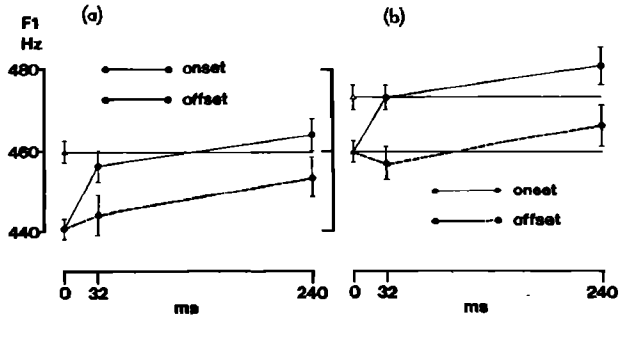

FIG. 4. Mean phoneme boundaries with standard error across 12 subjects from experiment 1. In (a) extra energy has been added to the original vowel continuum, starting or stopping at different times from the vowel. The filled triangle and horizontal line indicate the boundary value for the original continuum. In (b) energy has been added to a vowel continuum depleted in energy by 6 dB relative to the original. The depleted continuum's boundary is shown with an open triangle.

Second, the present experiment shows a similar result when the removal of extra energy leads away from, rather than back towards, the originally synthesized vowel. Figure 4(b) shows results from the case where 6 dB of energy is first removed from the 500-Hz harmonic, giving a vowel continuum with a boundary at a *higher* nominal frequency than the original [$t(11) = 5.01, p < 0.001$]. Adding back in the missing 6 dB to this depleted continuum, of course, returns us to the original vowel quality. But now, if the tone corresponding to the missing 6 dB starts earlier than the main vowel (solid lines), the vowel quality moves *away* from the original [$t(11) = 6.40, p < 0.001$] to that appropriate to the condition depleted by 6 dB. Thus vowel quality changes with onset time in the same direction and by approximately the same amount whether the perceptual removal of the tone moves the vowel's spectrum towards or away from the originally synthesized vowel. It is thus likely that the shift with onset time *towards* the original vowel quality, that we have previously found and replicate here, does not reflect simply a preference for the original vowel.

2. Offset differences

First, this experiment confirms that a harmonic that stops at a different time from the remainder of a vowel makes less of a contribution to its quality than if it were simulta-

neous. In Fig. 4(a), there is a significant shift in the phoneme boundary away from the augmented condition towards the original continuum's boundary when the extra energy stops 240 ms after the vowel [$t(11) = 3.1, p < 0.01$], but not when it stops 32 ms after [$t(11) < 1.0$]. There is a similar trend in Fig. 4(b). Combining the data from the two halves of Fig. 4 in an analysis of variance showed that the offset effect at 240 ms is significant overall [$F(1,11) = 7.2, p = 0.02$], and there is no difference in its size between the two halves of the figure.

3. Comparison of onset and offset effects

An obvious feature of Fig. 4 is that there is an almost constant change in vowel quality between the 32- and 240-ms conditions irrespective of the original vowel quality (normal or depleted) and onset versus offset. Analysis of variance on the relevant eight conditions confirms this with a significant effect of time [$F(1,11) = 7.16, p = 0.02$] that does not interact with either onset/offset or vowel quality.

On the other hand, there *is* a clear difference between onset and offset conditions when we compare the boundary shift between 0 and 32 ms [$F(1,11) = 22.2, p < 0.002$]. A 32-ms onset-time difference has a larger effect compared with the simultaneous case than does a 32-ms offset-time difference. The normal and depleted vowels are similar in this respect.

C. Discussion

This experiment has confirmed our previous findings and produced a clear answer to the new question that it addressed. It confirms our previous findings that a harmonic that starts or stops at a different time from the rest of a vowel can be perceptually segregated from that vowel. Since offset-time differences are effective in producing some perceptual segregation, our previous claim (Darwin and Sutherland, 1984) is confirmed that the perceptual separation is not simply due to adaptationlike mechanisms. Rather, we must appeal to some mechanism such as perceptual grouping to explain the results. A tone that starts or stops at a different time from the main vowel forms its own perceptual stream and so makes a reduced contribution to the vowel quality. Such streaming is clearly present at the smallest onset-time difference of 32 ms, but longer offset-time differences are required to give a clear effect. The difference between onset and offset effects might perhaps be reduced if even shorter vowels than those used here (56 ms) were used.

The main question that the experiment addressed was whether the perceptual separation of a tone from an accompanying vowel was influenced by whether the vowel spectrum was more normal before or after the tone had been perceptually separated from it. The results clearly show that the perceptual segregation of a synchronous tone is *not* influenced by whether the spectrum that remains after the tone has segregated is a normal vowel, or one depleted in energy at the tone's frequency. The effect cannot, therefore, be explained as being due to a preference for a particular type of spectral profile. Rather, the effect appears to be due to general auditory grouping mechanisms that use onset and offset asynchronies as an indication that different sound sources may be present (cf. Bregman and Pinker, 1978; Dannenbring and Bregman, 1978).

II. EXPERIMENT 2

The finding that a difference in onset time can serve to segregate perceptually two sound sources raises a paradox for the perception of normal speech. We have shown that a harmonic that starts at a different time from the rest of a vowel makes little contribution to the vowel's phonetic quality. But in normal speech the harmonics of the voice become audible and inaudible at different times as a formant peak sweeps across the harmonics of a varying fundamental. How then do we ever manage to group them together? The answer is not entirely clear, but one way out of the apparent paradox is to note that, in our experiment, the tone that is potentially separable on the basis of time differences, is *not* a speechlike sound. Speech does not consist of pure tones alternating with vowels. The *potential* separation afforded by the difference in onset time becomes an *actual* separation because the separable sounds could not be from the same speech source. What then determines whether sounds *are* from the same speech source? It could be simply that the sounds share a common pitch (cf. Darwin and Bethell-Fox, 1977), or it may be a more complex attribute concerning possible articulatory maneuvers. The question is an empirical one and is under investigation.

According to the above view, then, onset time differences *allow* perceptual segregation to occur, but they are not *sufficient* for it to occur. Are they *necessary*? Can a harmonic that is strictly simultaneous with the rest of a vowel make a reduced contribution to it simply by virtue of the fact that to include it would give a spectrum that was not speechlike?

In the next experiment, we investigate the effect on vowel quality of adding to the members of an /I/–/ɛ/ continuum different amounts of extra energy at one of two harmonic frequencies, 375 and 500 Hz. The effect of the added energy on the perceived first formant frequency is assessed by measuring the shift in the phoneme boundary produced by the added energy. The obtained shift is then compared with the shift expected according to two different ways of estimating the first formant frequency from the spectrum. If the perceptual shift is the same as that estimated from the spectrum, then we can be confident that *all* the extra energy is being incorporated into the vowel percept.

A. Method

The stimuli and measurement techniques were essentially similar to those used in the first experiment. Different amounts of energy at one of two different harmonic frequencies (375 and 500 Hz) were added to an original vowel continuum and the consequent changes in the /I/–/ɛ/ phoneme boundary were measured.

1. Stimuli

Eleven different vowel continua were used in the experiment. Ten of them differed in the amount of extra energy that had been added to the original vowel continuum at either 375 or 500 Hz (the third and fourth harmonics of the

125-Hz fundamental). They were synthesized using the methods described for experiment 1.

The original continuum varied in first formant frequency with nine values of $F1$ between 375 and 542 Hz in roughly 21-Hz steps. (Five more values of the first formant were used in deriving other continua; they had frequencies of 312, 333, 354, 563, and 584 Hz.)

The other ten continua were derived from the original continuum by adding to each member of the continuum extra energy at either 375 or 500 Hz. The tones that provided the additional energy were produced as in experiment 1 by filtering the appropriate member of the original continuum to isolate a particular harmonic and then adding it back with variable intensity into its parent sound to produce different amounts of gain. The added energy increased the level of the appropriate harmonic in each member of a continuum by 3, 6, 9, 12, or 15 dB.

Each continuum had nine members, but the range of first formant values in the original continuum used to produce each of the derived continua varied so that the phoneme boundary could be estimated efficiently. Those conditions with first formant ranges other than 375–542 Hz were as follows: with additional energy at 375 Hz: + 9 dB, 396–563 Hz; + 12 dB and + 15 dB, 417–584 Hz; with additional energy at 500 Hz: + 3 dB, 354–521 Hz; + 6 dB, 333–500 Hz; + 9 dB, + 12 dB and + 15 dB, 312–479 Hz. Since sounds from the different continua were randomized together in the experiment, range effects (see, e.g., Brady and Darwin, 1978) could not have influenced the results.

B. Procedure

The subjects' task in this experiment was simply to label each sound they heard as an /I/ or an /ɛ/. Seventeen subjects (the data of five were subsequently rejected) were tested individually as in experiment 1. In the main experiment, each subject heard ten successive random sequences of the 99 items (11 conditions × 9 continuum steps) and was free to take a rest at any time.

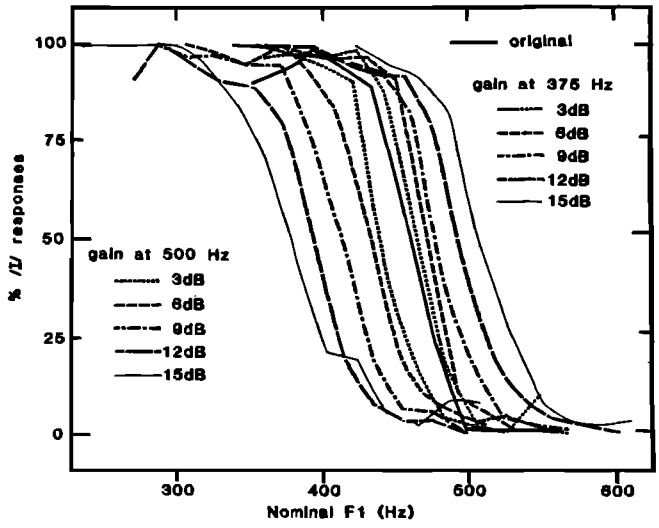

FIG. 5. Mean identification functions from experiment 2 for 12 subjects for $F1$ continua differing in the amount of extra energy added at either 375 or 500 Hz.

C. Results

The data of five subjects were discarded since they did not allow the phoneme boundary of each of the 11 conditions to be estimated reliably. The problem conditions were usually those where a high level of energy at 375 Hz had been added. One subject's boundaries in two conditions fell too far outside the range of sounds used to be estimated accurately; the other four gave markedly nonmonotonic identification functions in one or more condition. Identification functions pooled across the remaining 12 subjects for the 11 continua are shown in Fig. 5. The curves are plotted as in experiment

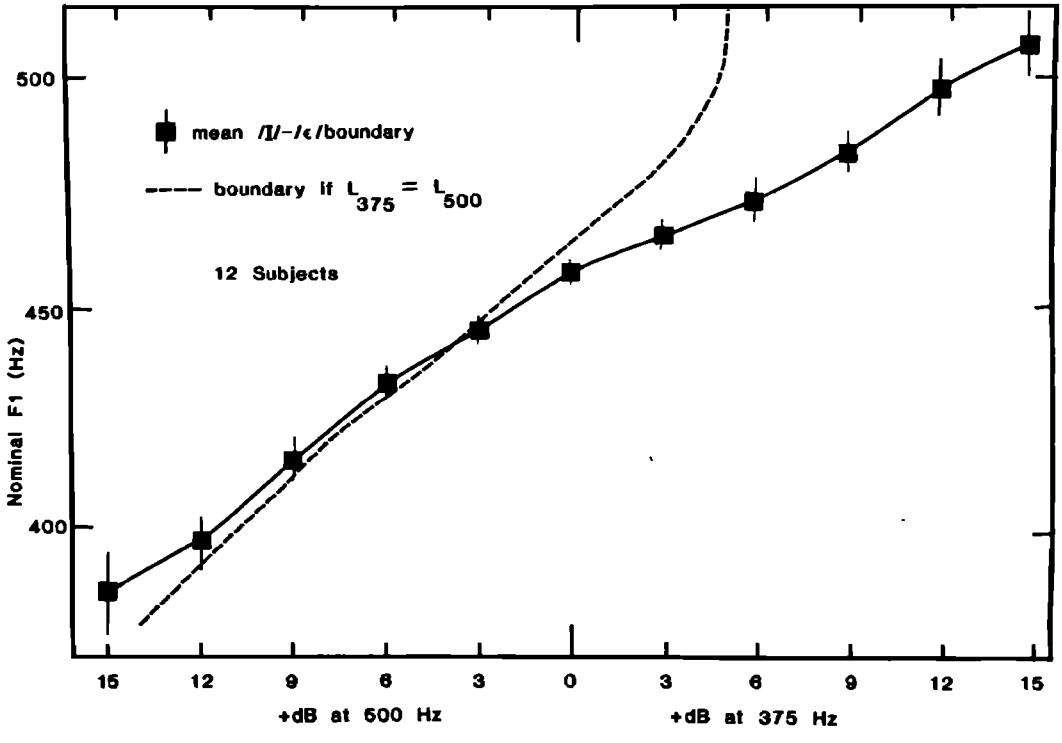

FIG. 6. Mean phoneme boundaries and standard error across 12 subjects for vowel continua with different levels of additional simultaneous energy at 375 or 500 Hz in experiment 2. The dashed line shows the nominal $F1$ values corresponding to equal levels of 375 and 500 Hz in the different conditions.

1; their slopes are the average slopes of the 12 subjects and do not change substantially across the conditions.

As before, the phoneme boundaries for individual subjects were estimated by probit analysis and are shown together with their standard errors across subjects in Fig. 6. The phoneme boundaries are plotted in terms of the *nominal* first formant frequency used to synthesize the sound from the *original* continuum from which the boundary sound for each condition was derived. If the additional energy has no effect on vowel quality, the results will appear as a horizontal line. If the addition of energy gives the percept of a lower $F1$, then the phoneme boundary will move to a higher nominal F value. Conversely, a higher $F1$ percept will give a boundary at a lower nominal $F1$. We would expect and indeed find that increasing the level of the 375-Hz component of the vowel leads to higher nominal $F1$ values, whereas increasing the level of the 500-Hz component leads to lower nominal $F1$ values. It is clear that there is a marked shift in the boundary as energy is added at either 375 or 500 Hz, reflecting the fact that the extra energy is being treated to *some* extent as part of the vowel.

The shift in the nominal $F1$ boundary produced by changing the amount of extra energy at 375 or 500 Hz is approximately linear in Fig. 6. This simple result might lead one to suppose that all the additional energy was being incorporated into the vowel. But, if we look at the spectra of sounds that fall on the phoneme boundary in the various conditions, we see a more complicated picture emerging.

Eleven new sounds were produced that lay exactly on the average boundaries of the 11 different continua. The boundary stimulus for a particular condition was made by first synthesizing a new token along the original continuum with the appropriate nominal first formant value, and then adding to it the appropriate amount of energy at either 375 or 500 Hz for that condition. The conditions in which extra energy has been added at 500 Hz (Fig. 7, right-hand panel) give boundary sounds whose spectra look broadly similar;

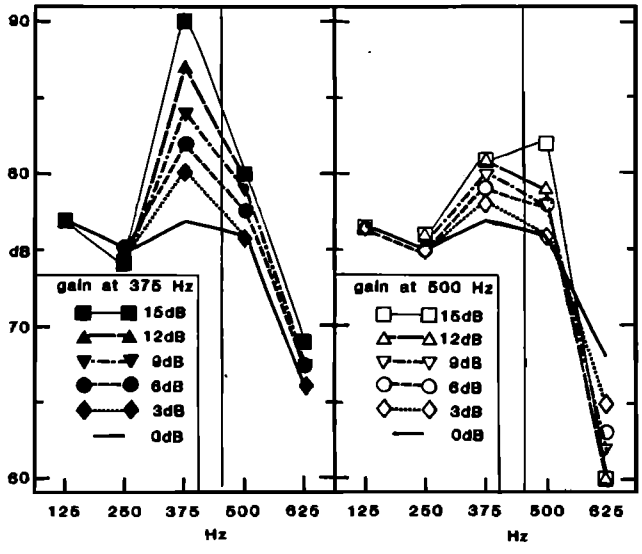

FIG. 7. Levels (in arbitrary dB units) of the first five harmonics of sounds synthesized to be at the phoneme boundary for each condition in experiment 2. The conditions differ in the amount of energy added to either the 375- or the 500-Hz component.

the level of the formant rises with increased additional energy, but the relative levels of harmonics around the formant peak are similar across conditions. By contrast, the conditions in which extra energy has been added at 375 Hz (Fig. 7, left-hand panel) give more heterogeneous boundary sounds. Their spectra have very different envelopes, with an increasingly prominent peak at 375 Hz. Why should these very different spectra all be heard as similar? Subjects were, after all, free to label all the sounds from a continuum with high levels of additional energy at 375 Hz as/I/. The answer appears to be that not all the extra energy at 375 Hz is being incorporated perceptually into the vowel, perhaps because of subjects' knowledge about vocal tract transfer function constraints.

The *actual* extent to which extra energy is being perceived as part of the vowel has been estimated in two ways. The first way uses only local information around the formant peak, while the second uses information from the whole spectrum. Both methods allow the same conclusion: Subjects are not using all of the additional 375-Hz energy when estimating the first formant frequency.

1. Weighting two local harmonic levels

The phoneme boundary in the *original* continuum happens to occur at a first formant frequency for which the 375- and 500-Hz components have almost identical levels. If we assume that only these two harmonics are used in computing the position of the first formant, then we can estimate to what extent the extra tone is being incorporated into the vowel by comparing the observed first formant boundaries with those necessary to give equal intensity at 375 and 500 Hz. The dashed line in Fig. 6 shows the expected values. It follows the listeners' values closely when energy is added at 500 Hz, but deviates markedly when 375-Hz energy is added.

To help in understanding possible reasons for the difference between energy added at 375 and at 500 Hz, the continuous lines in Fig. 8 show how the level of those two harmonics in the *original* vowel continuum would change as a function of the nominal $F1$ frequency. The phoneme boundary is marked by a vertical line. If subjects were estimating the first formant simply by comparing the amplitudes of those two frequencies, then it is clear that adding energy at 500 Hz can readily be compensated for by moving to a lower nominal $F1$ value. Lower $F1$ sounds in the original continuum have up to 10 dB less energy at 500 Hz than does the boundary sound. The filled circles indicate the levels of the 500-Hz component at the perceptual boundary for the five continua that have additional energy at 500 Hz. They stay close to the level of the corresponding 375-Hz component. But adding energy at 375 Hz cannot be compensated for by moving the nominal $F1$ to higher frequencies than the original boundary value, since the level of the 375-Hz component only drops by about 3 dB.

If subjects were adopting the simple strategy of weighting the two harmonics closest to the peak in order to estimate the first formant, then they would not show phoneme boundaries for any continuum that had 6 dB or more added to the 375-Hz tone. Yet all of the 17 subjects tested gave clear boundaries when up to 9 dB was added at 375 Hz and 13 out

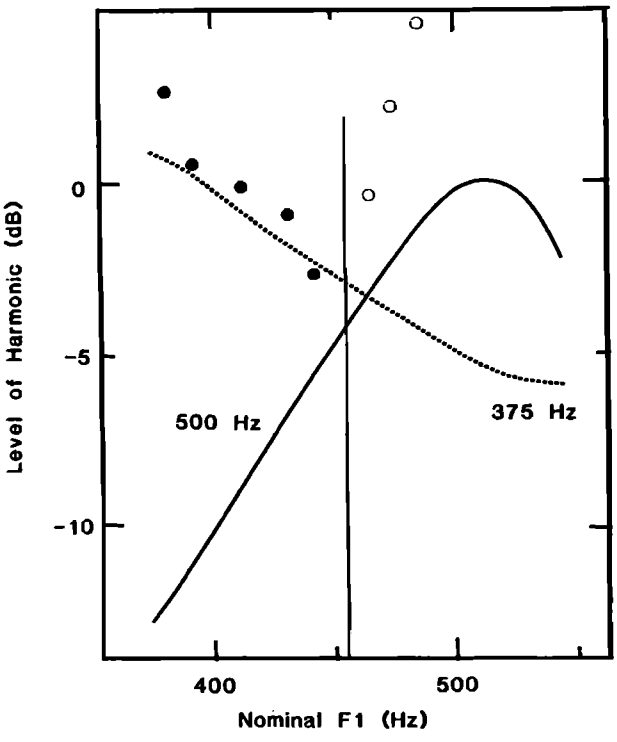

FIG. 8. The continuous lines indicate the relative levels of the 375- and 500-Hz components of the original vowel continuum as a function of the synthesized nominal $F1$ value. The filled circles show the levels of the 500-Hz component for boundary sounds in continua that have had that component increased by 3, 6, 9, 12, or 15 dB. The unfilled circles show the corresponding levels of the 375-Hz component from continua with that component boosted by 3, 6, and 9 dB. The 12- and 15-dB points lie off the graph.

of the total of 17 subjects tested gave estimable boundaries out to 15 dB. The open circles in Fig. 8 show the level of the 375-Hz component at the perceptual boundary for the first three conditions with added energy at 375 Hz. The points are all substantially above the 500-Hz component curve. The remaining two points lie off the graph. It is clear then that if subjects were adopting a local strategy for estimating formant frequency, some of the energy at 375 Hz would have to be discounted.

2. LPC estimation of first formant frequency

The second estimate uses an operation that takes into account the whole spectrum rather than just the two harmonics closest to the first formant peak. Linear predictive analysis (LPC) finds the best fitting all-pole spectrum to a given sound. It provides a good estimate of formant frequencies for simple (non-nasal) vowel sounds spoken in isolation. For each subject's data in each condition, a vowel token was synthesized that lay on the probit-estimated boundary. A 12-coefficient LPC analysis (with 95% pre-emphasis, using ILS software) was made of each of these sounds and the first formant frequency estimated from the resulting poles. If the extra energy were being entirely incorporated into the vowel and listeners were using a method equivalent to LPC analysis to estimate $F1$, we would expect that the LPC-estimated first formant of sounds synthesized to be on the perceptual boundary would remain approximately constant across conditions. On the other hand, if only part of the extra energy is being incorporated into the vowel, we would expect the LPC-estimated first formant to vary across the different conditions. Specifically, if the 375-Hz component is not being fully incorporated into the vowel, we would expect the first formant of a boundary sound, as estimated by LPC analysis, to be at a lower frequency than the LPC-estimated first formant from the original condition.

The mean LPC-estimated first formant values of boundary sounds from the 11 continua are shown in Fig. 9. For the original continuum, the mean LPC estimate of the first formant of each subject's boundary sound is exactly the synthesized value (456 Hz). The boundary sounds from the conditions in which extra energy has been added at 500 Hz give mean estimated $F1$ boundaries that are not significantly different (across subjects) from 456 Hz, except for the 15-dB condition [$t(11) = 2.8, p < 0.02$]. If listeners were using LPC analysis to estimate $F1$ frequency, then they would also be incorporating all the additional energy at 500 Hz into the vowel, except in the 15-dB condition. With extra energy added at 375 Hz, though, a very different picture emerges. Now the mean LPC-estimated first formant frequencies of the

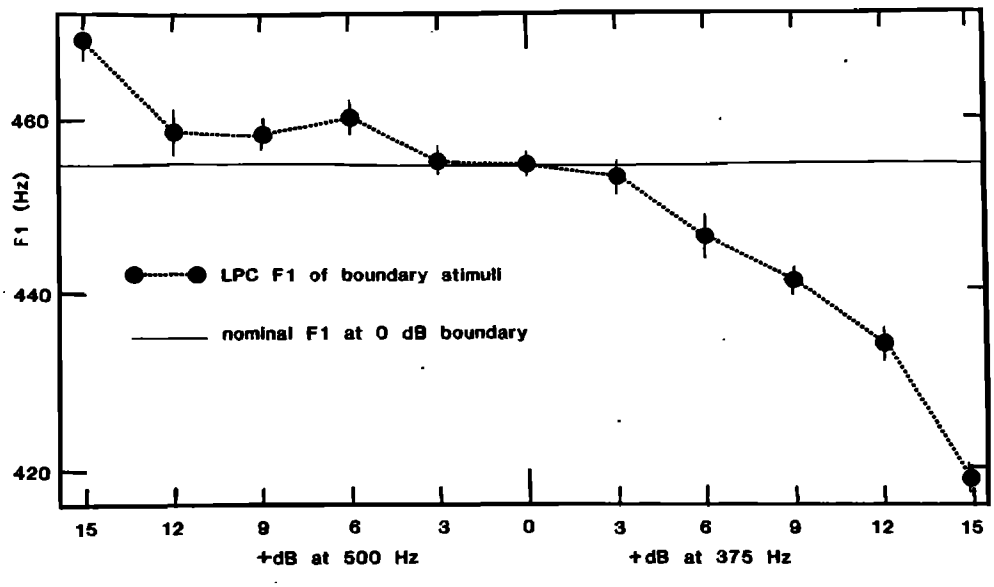

FIG. 9. Means (and their standard errors) of first formant frequencies estimated by LPC analysis from sounds synthesized to be at the perceptual boundary between /I/ and /ɛ/ in experiment 2. The 11 conditions differ in the amounts of energy added to the 375- or 500-Hz components. The horizontal line lies at the $F1$ phoneme boundary for the original condition with no extra energy.

boundary sounds fall significantly below 456 Hz in all but the 3-dB condition. If the human listener were using a method equivalent to LPC analysis, he would first have had to discount some of the 375-Hz energy in the four most intense conditions.

It is, of course, a matter of empirical test what precise method is being used by listeners to estimate the first formant, but it is hard to see how any simple weighting function, when applied to the spectra of Fig. 7 could yield a constant estimate of the first formant frequency. Both methods that we have used lead to the same conclusion: Some of the 375-Hz energy must have been discounted by the listener.

D. Discussion

Experiment 2 has set some limits to the mechanisms underlying first formant frequency estimation. On the one hand, it is clear that extra energy added to the simple spectrum that results from a formant synthesizer excited by a conventional glottal source *can* be incorporated into a vowel percept, changing the perceived formant frequency. We have shown that up to 12 dB of harmonic energy at 500 Hz can be incorporated into a vowel whose first formant lies *below* 500 Hz. On the other hand, additional energy at 375 Hz is only partially incorporated into the formant frequency estimate. What might be the reason for the asymmetry? Is it due to some psycho-acoustic factor such as masking, or must it be attributed to a higher-level constraint?

Upward spread of masking from an intense 375-Hz component could not be responsible for the effect since masking would make the 500-Hz frequency component less audible, and so would tend to produce a lower first formant frequency; the result that we find is that the perceived first formant is higher in frequency than would be expected from the levels of the harmonics.

A possible higher-level reason for the difference between the 375- and 500-Hz conditions in this experiment lies in the shape of the vocal tract transfer function. The function must pass through a gain of 0 dB at 0 Hz, constraining the amplitude changes that can be achieved in harmonics that lie below the first formant peak more than those that lie above it. The resulting constraint on harmonic intensities around a first formant peak has already been discussed and is illustrated in Fig. 8. It is arguable whether such a constraint should be described as specific or not, since many other acoustic systems have transfer functions that are similarly constrained. Constraints that are more specific to speech, such as those on average formant spacing and on formant bandwidths, may prove necessary to explain thoroughly the effect that we have found, but this testing must await further empirical data.

The actual spectrum of a vowel is, or course, a combination of the vocal tract transfer function with the source spectrum. So it is possible that additional low-frequency energy could be regarded as coming from a prominent peak in the source spectrum rather than from the vocal tract transfer function itself. Subjects could be removing part of the 375-Hz energy from the formant frequency calculation in one of two ways. First, they could be interpreting the extra energy at 375 Hz as arising from a different voice quality, rather than from a different vocal tract configuration. Second, they could be rejecting the extra energy completely from the vowel percept, hearing it as a separate tone.

It is likely that some additional energy was rejected completely from the vowel percept by the subjects, since many were aware of extra tones present along with the speech. We have not yet been able to develop a paradigm that would allow us to measure the loudness of the perceived extra tones.

III. GENERAL DISCUSSION

The experiments described here have made two points: (i) A difference in onset or offset time can be used to segregate energy at a harmonic frequency from a vowel regardless of whether the segregation leads towards or away from a normal spectral envelope; (ii) additional energy at a harmonic frequency that is slightly above the first formant frequency is incorporated into a vowel percept at higher energy levels than is energy at a harmonic frequency slightly below the first formant frequency.

The first result, we would maintain, reflects *general auditory* mechanisms of perceptual grouping, while the second reflects the operation of a higher-level constraint that may be speech specific.

Similar results to those found in experiment 1 have been reported in the perception of nonspeech sounds by Dannenbring and Bregman (1978) and by Bregman and Pinker (1978). Dannenbring and Bregman found that a harmonic that started or stopped at a slightly different time from the rest of a harmonic complex was more likely to form a perceptual group with another tone of similar frequency that alternated with the complex than one that was simultaneous. Similarly, Bregman and Pinker showed that the rated richness of a harmonic complex was reduced when one of its components had been segregated out in this manner.

The results of our experiment 2, however, show that energy at a harmonic frequency may still be excluded from a vowel percept when the harmonic is simultaneous with the vowel. There are limits on what energy can be included in formant frequency estimation. Such limits probably reflect a listener's knowledge of constraints on vocal-tract transfer functions.

A. Levels of description in speech perception

The speech perception literature consists, for the most part, of descriptions of the relationship between properties of entities such as formant, burst, voice-onset time, and silence, on the one hand, and linguistic units such as phoneme or distinctive feature, on the other. The experiments that have established such relationships have used the speech of a single talker, natural or synthetic, so that there is no experimental distinction between properties of the sound wave and properties that are specific to a particular sound source.

The main theoretical thrust of the experiments reported here and of previous experiments in a similar vein (Darwin, 1981, 1983; Darwin and Bethell-Fox, 1977; Darwin and Sutherland, 1984) is to establish a case for two different levels of description for auditory information. An initial level de-

scribes sound in terms of properties that are evident in the waveform. Such properties may be used subsequently to establish more abstract properties that could be due to a single sound source. The more abstract properties are those that should make contact with stored phonetic knowledge on the attributes of phonetic categories.

The earlier level of description must capture explicitly both those properties of the sound that can be used in assigning features to different putative sources, as well as properties that will subsequently be used to identify a sound's category. It could be viewed as the auditory equivalent of Marr's (1982) primal sketch.

Marr identifies four different representational stages in the process of identifying three-dimensional objects in vision. The first stage is the image, which simply represents the intensity of light at each point; the final stage is a model-based description of 3-D objects. Between these stages lie two more, the primal sketch and the 2 1/2-D sketch. The primal sketch makes explicit important information about the two-dimensional image such as intensity changes and their geometrical organization. It deals in such primitives as edge segments, terminations, discontinuities, and virtual lines and is computed from the image by using general constraints on the way that surfaces and edges of objects structure light. The 2 1/2-D sketch interprets the primal sketch in terms of explicit surfaces with specific orientations relative to the viewer.

Each level of representation makes explicit different types of property—intensity, edges, surface orientation or shape, and different constraints on the nature of the physical world are exploited in moving from one level to the next. Information about the raw intensity levels present in the image representation never makes contact with information about the three-dimensional properties of particular objects, since a particular property of a specific object can result in many very different values in the intensity of light in the image depending on such factors as the level and direction of illumination, the object's distance from the viewer, occlusion by other objects, and so on.

For the same reasons it is a mistake to allow raw spectral information from the soundwave, or even properties extracted directly from the soundwave, to make immediate contact with stored knowledge on the properties of phonetic categories. Knowledge about the properties of phonetic categories must be represented by properties of the sound produced by a single (though not necessarily any particular) speaker. Yet properties that are apparent in the raw waveform are not specific to a single speaker or sound source; they are properties that are due to whatever sound sources are present at the time. For example, the silence necessary to cue an intervocalic stop consonant is silence of a single sound source; there may be no actual silence present in the waveform (see Darwin and Bethell-Fox, 1977). Raw spectral features are also influenced by other phonetically irrelevant factors such as the transfer function between the talker and the listener (Fant, 1980).

The lower level of auditory analysis should capture, for example, information about spectral peaks, local direction of movement in amplitude and frequency of energy regions, time of onset of energy in different regions, and so on. Such a description should then serve as a rich data base for the operation of processes that can identify appropriate, more abstract structures. General auditory and more specialized phonetic knowledge can be brought to bear to interpret the data in terms of specific sound sources. It is only after such processes have worked over the initial representation that we may sensibly talk of such source-specific cues to phone identity as formants, silence, voice-onset time, and the like. The experiments described here provide some experimental evidence for the distinction between the two levels and suggest two different types of knowledge that might mediate between them.

ACKNOWLEDGMENTS

The research was supported by grant GR/A 83977 from the UK SERC to Professor N. S. Sutherland and, subsequently, by grant GR/C 6009.9 to the author. John Doyle and Ian Winter helped to run the experiments while employed on Sutherland's grant. Arthur G. Samuel's comments on an earlier draft helped substantially to clarify my arguments and presentation.

Assmann, P. F., and Nearey, T. M. (1983). "Perception of height differences in vowels," J. Acoust. Soc. Am. Suppl 1 74, S89.

Brady, S. A., and Darwin, C J. (1978). "Range effect in the perception of voicing," J. Acoust. Soc. Am. 63, 1556–1558.

Bregman, A. S. (1978). "The formation of auditory streams," in *Attention and Performance VII*, edited by J. Requin (Erlbaum, Hillsdale, NJ), pp. 63–76.

Bregman, A. S., Abramson, J., and Darwin, C. J. (1983). "Effect of amplitude modulation upon fusion of spectral components," J. Acoust. Soc. Am. Suppl. 1 74, S9.

Bregman, A. S., and Pinker, S. (1978). "Auditory streaming and the building of timbre," Can. J. Psychol. 32, 19–31.

Brokx, J.P.L., and Nooteboom, S. G. (1982). "Intonation and the perceptual separation of simultaneous voices," J. Phonet. 10, 23–36.

Carlson, R., Fant, G., and Granstrom, B. (1975). "Two-formant models, pitch and vowel perception," in *Auditory Analysis and Perception of Speech*," edited by G. Fant and M. A. A. Tatham (Academic, London), pp. 55–82.

Dannenbring, G. L., and Bregman, A. S. (1978). "Streaming vs fusion of sinusoidal components of complex tones," Percept. Psychophys. 24, 369–376.

Darwin, C. J. (1981). "Perceptual grouping of speech components differing in fundamental frequency and onset time," Q. J. Exp. Psychol. 33A, 185–207.

Darwin, C. J. (1983). "Auditory processing and speech perception," in *Attention and Performance X*, edited by D. G. Bouwhuis (Erlbaum, Hillsdale, NJ).

Darwin, C. J., and Bethell-Fox, C. E. (1977). "Pitch continuity and speech source attribution," J. Exp. Psychol.: Hum. Percept. Perf. 3, 665–672.

Darwin, C. J., and Sutherland, N. S. (1984). "Grouping frequency components of vowels: when is a harmonic not a harmonic?" Q. J. Exp. Psychol. 36A, 193–208.

Fant, G. (1980). "Perspectives in speech research," Speech Transmission Laboratory, Stockholm, QPSR 1980/2–3, 1–16.

Karnickaya, E. G., Mushnikov, V. N., Slepokurova, N. A., and Zhukov, S. Ja. (1975). "Auditory processing of steady-state vowels," in *Auditory Analysis and Perception of Speech*," edited by G. Fant and M. A. A. Tatham (Academic, London), pp. 37–53.

Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer,"

J. Acoust. Soc. Am. 67, 971–995.
Marr, D. (1982). *Vision* (Freeman, San Francisco).
McAdams, S. (1980). "Spectral fusion and the creation of auditory images," in *Music, Mind and Brain: the Neuropsychology of Music*, edited by M. Clynes (Plenum, New York).
McAdams, S., and Bregman, A. S. (1979). "Hearing musical streams," Comp. Music J. 3(4), 26–43, 60.
Miller, G. A., and Nicely, P.E. (1955). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. 27, 338–352.
Moore, B. C. J., and Glasberg, B. R. (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. 74, 750–753.
Pickett, J. M. (1957). "Perception of vowels heard in noise of various spectra," J. Acoust. Soc. Am. 29, 613–620.
Rasch, R. A. (1978). "Perception of simultaneous notes such as in polyphonic music," Acustica 40, 21–33.
Scheffers, M. T. M. (1983). "Sifting vowels: auditory pitch analysis and sound segregation," Doctoral dissertation, Groningen University.
Summerfield, A. Q., Foster, J., Gray, S., and Haggard, M. P. (1981). "Perceiving vowels from 'flat spectra'," J. Acoust. Soc. Am. Suppl. 1 69, S116.
Summerfield, A. Q., Haggard, M. P., Foster, J., and Gray, S. (1984). "Perceiving vowels from uniform spectra: phonetic exploration of an auditory after-effect," Percept. Psychophys. 35, 203–213.