Perception & Psychophysics
1989, 45 (6), 529-536

Auditory enhancement and the perception of concurrent vowels

QUENTIN SUMMERFIELD and PETER F. ASSMANN
University of Nottingham, Nottingham, England

Listeners identified both constituents of *double vowels* created by summing the waveforms of pairs of synthetic vowels with the same duration and fundamental frequency. Accuracy of identification was significantly above chance. Effects of introducing such double vowels by visual or acoustical precursor stimuli were examined. Precursors specified the identity of one of the two constituent vowels. Performance was scored as the accuracy with which the other vowel was identified. Visual precursors were standard English spellings of one member of the vowel pair; acoustical precursors were 1-sec segments of one member of the vowel pair. Neither visual precursors nor contralateral acoustical precursors improved performance over the condition with no precursor. Thus, knowledge of the identity of one of the constituents of a double vowel does not help listeners to identify the other constituent. A significant improvement in performance did occur with ipsilateral acoustical precursors, consistent with earlier demonstrations that frequency components which undergo changes in spectral amplitude achieve enhanced auditory prominence relative to unchanging components. This outcome demonstrates the joint but independent operation of auditory and perceptual processes underlying the ability of listeners to understand speech despite adversely peaked frequency responses in communication channels.

The experiment described in this paper addresses two distinct but related issues. First, how do listeners identify the constituents of pairs of concurrent vowels that possess the same fundamental frequency? Second, can auditory and perceptual mechanisms that are sensitive to changes in spectral amplitude help listeners to discount coloration from communication channels with uneven frequency responses? The following paragraphs establish the background to these questions, explain the link between them.

The frequency response of the communication channel between a speaker's mouth and a listener's ear is rarely flat or time-invariant. Earphones and loudspeakers can impose gross peaks and troughs, while sound paths in reverberant rooms introduce a fine-grained pattern of resonances and antiresonances due to cancellation and reinforcement of wavefronts. Figure 1 illustrates the problem for the listener. It shows the spectrum of a neutral vowel presented against a low level of background noise through two communication channels that have different, uneven, frequency responses. In neither case does the spectrum of the vowel correspond to its anechoic form. Rather, it is colored by the frequency response of the communication channel.

Some of the data reported in this paper were presented at the winter meeting of the Experimental Psychology Society, London, January 1986, and to the NATO ARW on "The Psychophysics of Speech Perception," Utrecht, July 1986; they are summarized in Summerfield and Assmann (1987). We thank Mark Haggard and an anonymous reviewer for constructive criticisms of a draft of this paper. Correspondence may be addressed to Quentin Summerfield, MRC Institute of Hearing Research, University Park, Nottingham NG7 2RD, United Kingdom.

It would help listeners to discount the coloration if they were selectively sensitive to patterns of spectral amplitude change. As shown on the right of Figure 1, the spectral changes that occur at the onset of the vowel bear a more obvious relation to its formant structure than do either of the steady-state spectra. A relevant form of auditory sensitivity to patterns of spectral-amplitude change has been studied by Summerfield, Haggard, Foster, and Gray (1984; cf. Summerfield, Sidwell, & Nelson, 1987). We shall describe it and then outline its possible relevance for discounting spectral coloration.

Summerfield et al. (1987) synthesized a harmonic complex in which all harmonics were equal in intensity, except for three pairs which were set to a lower level. Each pair straddled the center frequency of one of the first three formants of a vowel, creating the *spectral complement* of that vowel. This signal was presented for a few hundred milliseconds and then the levels of the lowered harmonics were restored, thereby filling the valleys. The resulting signal with a uniform spectrum sounded like the vowel whose complement had preceded it, despite the absence of peaks and valleys in its spectral envelope. Through omission of energy at different frequencies in the complement, percepts of a range of vowels could be produced, and so the accuracy with which the appropriate vowel was identified could be used to measure parameters of the effect. In what follows, we shall refer to stimuli playing the roles of vowel complements as *precursors* and stimuli playing the roles of uniform spectra as *test stimuli*. The spectral relationship between precursors and test stimuli that generates this *flat spectrum vowels* (FSV) effect is illustrated schematically in line 1 of Figure 2.

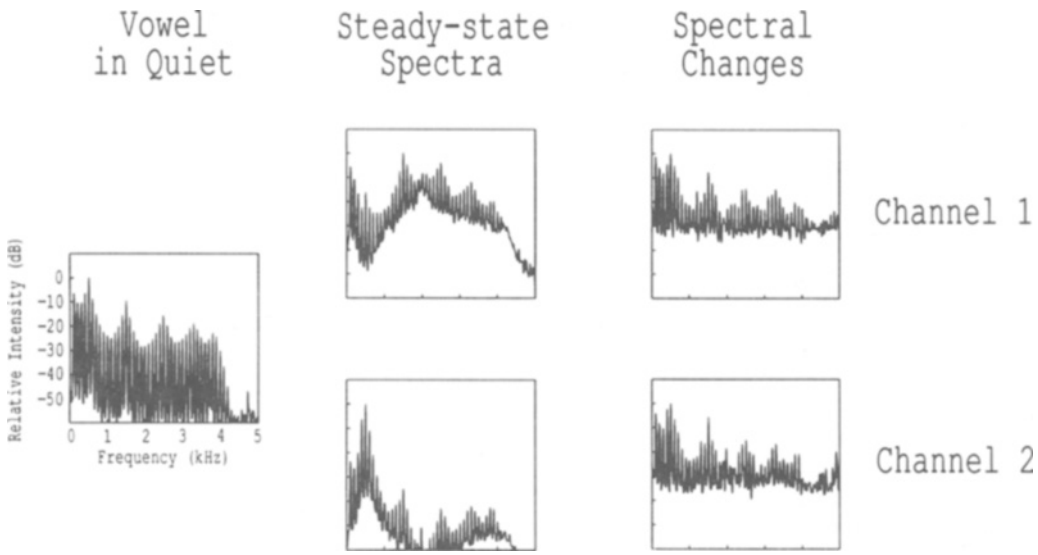

Figure 1. Left: The amplitude spectrum of a neutral vowel presented in quiet through a communication channel with a flat frequency response. Center: Amplitude spectra of the same vowel presented against a low level of background noise through two communication channels with different, uneven frequency responses. Right: The changes in spectral amplitude that occur at the onset of the vowel in the two channels. The changes in spectral amplitude are more similar to each other, and to the spectrum of the original vowel, than are either of the steady-state spectra in the center panel.

In the FSV effect, there is an enhancement of the auditory prominence of frequency components that undergo an increment in spectral amplitude, relative to the pre-existing, unchanging components. The incremented components fill spectral valleys. However, this is not a necessary condition for enhancement to occur. Summerfield et al. (1987) synthesized a test stimulus in which three pairs of adjacent harmonics were raised from their levels in a uniform spectrum to define an approximation to the formant pattern of a vowel. This test stimulus was identified more accurately when introduced by a precursor that possessed a uniform spectrum than when presented in isolation. The spectral relationship between precursors and test stimuli that gives rise to this *peaked spectrum vowels* (PSV) effect is schematized in lines 2 and 3 of Figure 2.

The PSV effect is more akin to the naturally occurring situation in which a vowel is added to a broad-band background sound, than is the FSV effect. Line 4 of Figure 2 illustrates a further development of this analogy. The precursor now has an uneven spectral envelope, that of an

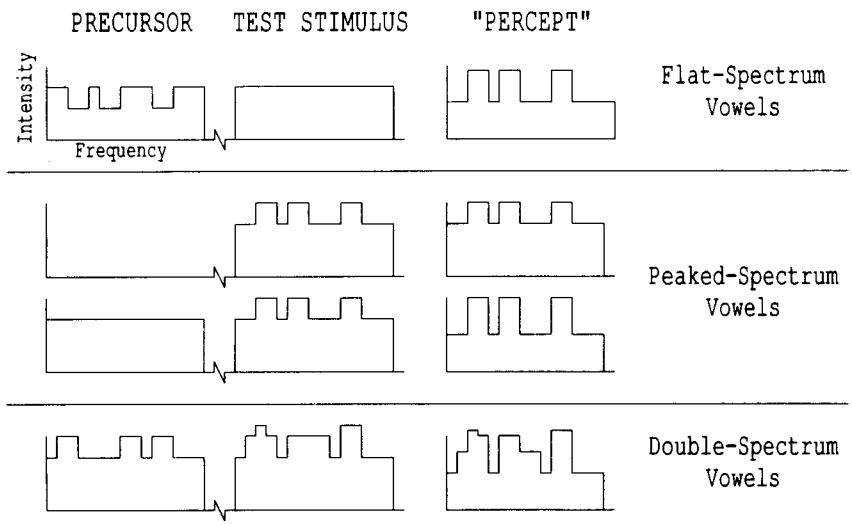

Figure 2. Left and center: Schematic illustrations of precursors and test stimuli giving rise to the flat spectrum vowels effect (line 1), the peaked spectrum vowels effect (lines 2 and 3), and the double spectrum vowels effect (line 4). Right: Schematic spectra of the subjective impressions given by the test stimuli.

approximation to a vowel, and the test stimulus is created by means of raising the levels of an additional three pairs of harmonics corresponding to the formants of a second vowel. This condition provides a powerful test of the idea that auditory sensitivity to patterns of spectral-amplitude change can help discount the spectral coloration of communication channels. If this idea is correct, the sequence of precursor and test stimulus should sound like a sequence of two vowels: one defined by the precursor, the other by the increments in spectral amplitude at the onset of the test stimulus. This outcome would confirm that the spectrum of a background sound need not color the interpretation of the spectrum of an added signal.

In a pilot experiment, we confirmed that this *double spectrum vowels* (DSV) effect occurs. Precursors were approximations to the five vowels /i/, /a/, /u/, /ɔ/, and /ɜ/ created by synthesizing a signal with a uniform spectrum and then raising the levels of three pairs of harmonics by 9 dB. Five test stimuli were generated from each precursor by the addition of the waveform of the precursor to the waveform of one of the five vowels. In one condition, 200-msec test stimuli were presented in isolation; in another condition, the same stimuli were introduced by 1-sec precursors. In each case, performance was scored as the accuracy with which the vowel *not* defined by the precursor was identified. With isolated test stimuli, performance was 71% correct; when precursors introduced the test stimuli, performance improved significantly to 97% correct.[1] Moreover, as predicted, the sequences of precursors and test stimuli generally sounded like sequences of two vowels, rather than one long vowel followed by a concurrent pair of shorter vowels.

The first aim of the experiment reported below was to consolidate this result and to establish that it can occur with stimuli that possess more natural spectral envelopes than those schematized in Figure 2. The second aim was to use the effect to test an account of the way listeners identify the members of pairs of concurrently presented vowels. The rationale for this test and its relevance to the DSV effect are described in the next section.

POSSIBLE BASES OF THE DSV EFFECT

The FSV effect (Figure 2, line 1) is one of a family of auditory aftereffects (e.g., see Cardozo, 1967; Viemeister, 1980; Wilson, 1980) whose results can be summarized by saying that when the spectral valleys in a signal with a complex spectral envelope are filled, the resulting signal with a uniform spectrum may be perceived to possess a spectral envelope complementary to that which has preceded it. Mechanisms that could give rise to these effects have been discussed by Darwin (1984a, 1984b), Darwin and Gardner (1987), Summerfield and Assmann (1987), Summerfield et al. (1984, 1987), Viemeister (1980), and Viemeister and Bacon (1982). The parameters of the effects are broadly consistent with an origin in two processes: peripheral adaptation and source segregation.

In invoking adaptation as part of the explanation, Summerfield et al. (1987) argued in three stages. First, the effective auditory level of the components of the precursor declines over its duration because of adaptation. Second, while adaptation reduces the auditory response to existing components, it does not change the size of the increase in the auditory response produced by energy added to existing components (e.g., see Smith, 1979; Smith, Brachman, & Frisina, 1985). Hence, the components whose levels are raised to define the test stimulus produce a greater auditory response than the unchanging components. Third, the resulting difference in effective auditory level may be increased further by a reduction, due to adaptation, of lateral suppression of the incremented components by the unchanging components along the lines suggested by Viemeister and Bacon (1982).

Darwin (e.g., 1984a) demonstrated that these effects of adaptation are likely to be supplemented by perceptual processes of source segregation. The energy that defines the precursor and the energy added to define the test stimulus necessarily start at different times. As a result, they may be perceived to originate from different sources, and they are segregated perceptually by a process of spectral subtraction.

The combined result of adaptation and source segregation is that the added energy in the FSV and PSV effects is heard to stand out from the background of the unchanging components. Both adaptation and source segregation are likely also to contribute to the DSV effect. For the reasons that follow, a third process may also contribute. We shall refer to it as *categorization-guided segregation*. In the stimuli that give rise to the DSV effect, the test stimulus is effectively a pair of concurrent vowels, with no onset or offset asynchrony and identical fundamental-frequency and amplitude contours. Scheffers (1983) and Zwicker (1984) have shown that listeners can identify both constituents of such *double vowels* with an accuracy significantly above chance. In the DSV effect, the precursor indicates the phonemic identity of one of the two vowels that compose the test stimulus. Possibly, this knowledge aids listeners in identifying the added vowel by indicating the spectral pattern that should be subtracted from the spectral pattern of the test stimulus to reveal the added vowel.

According to these hypotheses, source segregation and categorization-guided segregation differ in the way in which the spectrum to be subtracted is established. In source segregation, the spectrum to be subtracted is the (auditory) spectrum of the precursor. In categorization-guided segregation, the spectrum to be subtracted is derived from the listener's stored knowledge of the spectrum of the vowel whose category is exemplified by the precursor. The distinction is pertinent because a process analogous to categorization-guided segregation has been proposed by Zwicker (1984) as a means by which listeners may identify the constituents of a double vowel that is presented in isolation. In this account, the auditory spectrum of the double vowel is compared with single-vowel

templates. The best match establishes the identity of the first vowel. Its spectrum (or selected properties of its spectrum) are then subtracted from the spectrum of the double vowel to reveal the second vowel.

An alternative strategy that listeners might use to identify the constituents of isolated double vowels was discussed by Scheffers (1983). He suggested that listeners might carry out independent comparisons of the auditory spectrum of the double vowel with templates representing the spectra of single vowels. The two best-matching templates would determine the responses. Scheffers favored this strategy of *simultaneous independent comparisons* for two reasons. First, it is simpler than categorization-guided segregation: the template-matching operation is performed twice and the intervening stage of spectral subtraction is dispensed with. Second, listeners generally perceive double vowels as a "dominant" vowel whose quality is colored by that of a second "background" vowel, rather than as two separate vowels. The aim of the experiments described below was to contribute evidence that might help to distinguish these alternatives. Specifically, we sought evidence for categorization-guided segregation, by determining whether it makes a contribution to the DSV effect that can be distinguished from the contributions of adaptation and source segregation.

The test stimuli consisted of double vowels presented monaurally. In one condition they were presented in isolation. In other conditions, they were introduced by precursors that specified the identity of one of their two constituents. Acoustical precursors were segments of one of the two constituents, presented ipsilaterally or contralaterally. Visual precursors were displays of the orthographic identity of one of the two constituents. We measured the accuracy with which listeners identified the constituent of the double vowel that was not defined by the precursor, and compared performance when double vowels were presented in isolation with performance when double vowels were introduced by precursors.

The following considerations guided the choice of these conditions:

1. As a result of the DSV effect, performance is predicted to be better with ipsilateral precursors than with no precursors.

2. The FSV and PSV effects do not occur with contralateral precursors (Summerfield et al., 1984, 1987), because peripheral adaptation and source segregation are monaural phenomena. Thus, if categorization-guided segregation contributes to the DSV effect, performance will be better with contralateral precursors than with no precursors; and the contribution of adaptation and source segregation to the DSV effect (controlling for the effect of categorization-guided segregation) can be estimated from the difference between the ipsilateral and contralateral conditions.

3. Categorization-guided segregation requires a knowledge of the phonemic identity of one of the vowels in the test stimulus. A visual precursor, like an acoustical precursor, should provide this knowledge. Therefore, performance should be better with a visual precursor than with no precursor.

Table 1
Frequencies in Hz of the First Three Formants Used to Define the Five Single Vowels in the Cascade and Six-Harmonic Stimulus Sets

	Vowel				
Formant	/i/	/a/	/u/	/ɔ/	/ɜ/
F1	250	650	250	350	450
F2	2,250	950	850	750	1,250
F3	3,050	2,950	1,950	2,850	2,650

Note—The cascade vowels also included fourth and fifth formants set to 3300 Hz and 3850 Hz, respectively. The 3-dB bandwidths of the formants in the cascade stimuli were 90, 110, 170, 250, and 300 Hz.

METHODS

Acoustical Stimulus Components: Cascade Vowels and Six-Harmonic Vowels

The acoustical components of the stimuli were constructed from two types of vowel-like sounds, distinguished by gross differences in their spectral structure and method of synthesis. These components were steady-state approximations to five of the vowels of British English, /i/, /a/, /u/, /ɔ/, and /ɜ/, synthesized with constant formant frequencies, a constant fundamental frequency of 100 Hz, and constant overall amplitude. They were generated digitally (10,000 samples/s, 12-bit amplitude quantization). The frequencies of the formants used to define the vowels are listed in Table 1. In the *cascade* stimuli, the vowels had magnitude and phase spectra determined by the model for formant synthesis described by Klatt (1980). In the *six-harmonic* stimuli, each vowel was composed of only six equal-amplitude harmonics. These harmonics formed three adjacent pairs straddling the formant frequencies listed in Table 1. The phases of the harmonics were chosen randomly. Figure 3 shows line spectra of the vowel /ɜ/ from the two stimulus sets.

The reason for using the cascade vowels was to confirm that the DSV effect improves the intelligibility of an added vowel when both the precursor and the added vowel possess a natural spectral envelope. The reason for including the six-harmonic vowels was to de-

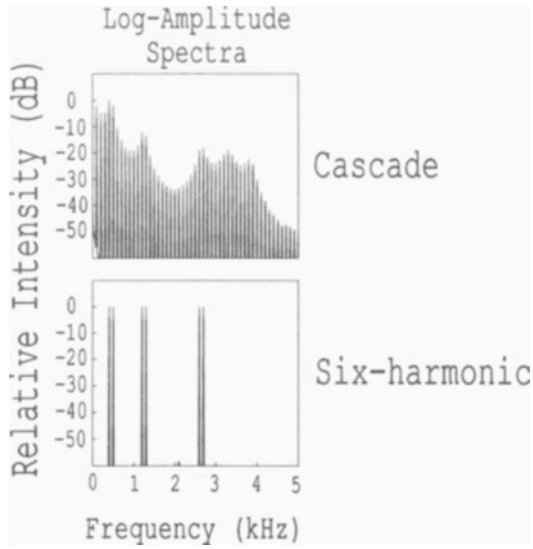

Figure 3. Line spectra of the single vowel /ɜ/ from the two stimulus sets.

Table 2
Conditions of the Experiment

Condition	Precursor	Test Stimulus	Ear
1. Single vowels	-	single vowel	left
	-	-	right
2. Double vowels	-	double vowel	left
	-	-	right
3. Ipsilateral precursor	single vowel	double vowel	left
	-	-	right
4. Contralateral precursor	-	double vowel	left
	single vowel	-	right
5. Visual precursor	orthographic representation	double vowel	left
		-	right

termine whether the use of precursors and test stimuli with grossly unnatural, and thus unfamiliar, spectral envelopes reduces the ability of listeners to identify the constituents of double vowels, and to use prior knowledge to identify an added vowel.

Conditions

Each stimulus set was used in five conditions, which are summarized in Table 2.

The stimuli for Condition 1 were 200-msec segments of each of the five single vowels.

For Condition 2, double vowels were created by forming all possible pairings of the five single vowels. This was done by summing corresponding samples in the sequence of samples defining each single vowel. Nominally, 25 pairwise combinations of five vowels can be created, made up of 5 identical pairings and 20 nonidentical pairings. However, since it is not meaningful to distinguish the order of items within a pair, the 20 nonidentical pairings form only 10 distinct combinations. Two instances of each nonidentical pairing were created, along with one instance of each identical pairing, giving a total of 25 stimuli.

In Condition 3, each double vowel from Condition 2 was introduced by a 1-sec precursor whose spectral structure was identical to that of one of the two vowels making up the pair. For each double vowel whose constituents were identical, one such stimulus was created. For each double vowel whose constituents were not identical, two stimuli with different precursors were created. This gave a total of 25 stimuli.

Condition 4 was the same as Condition 3, except that precursors were presented to the ear contralateral to the test stimuli.

In Condition 5, the acoustical precursor vowel in Condition 3 was replaced by a visual display of its orthographic representation ("EE," "AH," "OO," "OR," or "ER") on a VDU screen.[2]

Onsets and offsets of the acoustical components of the stimuli were shaped by a 10.7-msec Kaiser function (Kaiser, 1966) resulting in durations between the −6 dB points of 188.6 msec (200 msec between the 0-V points) for the single vowels and double vowels and 988.6 msec (1,000 msec between the 0-V points) for the precursors. The offsets of the acoustical precursors and the onsets of the test stimuli were aligned at the 0-V points.

Subjects

Six adult listeners, including the authors, took part. All had pure-tone thresholds within 15 dB of the ANSI standard (ANSI, 1969) in both ears. All were native speakers of British English except P.F.A., who is a native speaker of Canadian English. The cascade vowels were presented before the six-harmonic vowels to 3 subjects. The order was reversed for the other 3 subjects.

Procedure

The acoustical components of the stimuli were presented on-line (DEC PDP-11/60, LPA-11K), low-pass filtered at 4.25 kHz (KEMO VBF/8, −135 dB/octave), attenuated and presented through Sennheisser HD414 headphones. Presentation was monaural to the subjects' left ears in all conditions except Condition 4, in which the precursors were presented to the right ear and the test stimuli to the left ear. The acoustical components of the stimuli were presented at a quiet, comfortable level. Peak presentation levels were measured for the single and double vowels (Bruel and Kjaer artificial ear, type 4153, with flat plate adaptor, type DB0843, half-inch microphone, type 4134, and sound-level meter, type 2235 on its "fast" setting). For the cascade stimuli, levels ranged from 55.1 to 68.8 dB(A) for single vowels and from 59.6 to 73.3 dB(A) for double vowels. For the six-harmonic stimuli, the six harmonics were presented at the same level as the intense harmonics in the first-formant region in the cascade vowels. The presentation level of the single vowels ranged from 56.0 to 61.1 dB(A) and the double vowels from 60.8 to 67.1 dB(A).

The stimuli from each major condition were randomized together in 10 blocks of trials. Each block contained 1 instance of each of the 105 stimuli.

The subjects sat in a sound-attenuated room, facing the VDU. Prior to each trial, a 2-sec prompt on the VDU told the subject which condition to expect. In Condition 5, the prompt also indicated the orthographic identity of one of the two vowels in the test stimulus. The subjects responded by pressing keys on the VDU labelled with the orthographic representations of the five vowels. They made two responses on each trial. In Condition 1, they were instructed to identify the vowel defined by the test stimulus and to press the key corresponding to this vowel twice. In Condition 2, they were instructed to identify both vowels defined by the test stimulus. In Conditions 3, 4, and 5, they were instructed to use their first response to identify the vowel specified by the auditory or orthographic precursor, and their second response to identify the additional vowel present in the test stimulus.

At the start of the experiment, the subjects received practice sequences containing 20 replications of each of the five single vowels with feedback. The sequences were presented until the subject could identify the single vowels with an accuracy greater than 95% correct. For different subjects, this required between one and three sequences. The subjects then responded to a second practice sequence which was equivalent to a single block of trials from the main experiment, with no feedback. They then received the 10 blocks of the main experiment, again with no feedback. The trials were self-paced, with each stimulus occurring 3 sec after the second response to the previous trial.

RESULTS

Overall Pattern

The data from each major condition were scored separately in the ways described below to obtain proportions

of correct responses, which were averaged over the five vowels. These results have been plotted as percentages in Figure 4. Parametric statistical tests were performed on the proportions following a root-arcsin transform (Winer, 1971). Initially, tests of linear and quadratic trend over blocks were applied to the data from each condition. There were no significant linear components, and only one significant quadratic component; there was a peak in performance across Blocks 5-8 for the six-harmonic double vowels. [$F(1,5) = 7.60$, $p < .05$]. Thus, performance was sufficiently stable not to have vitiated other results. Further analyses were intended to reveal the nature of the differences among conditions.

In Condition 1 (single vowels), performance was scored as the number of vowels identified correctly. Mean accuracy was 99.0% (cascade vowels) and 98.3% (six-harmonic vowels). In Condition 2 (double vowels), performance was scored in two ways. The mean accuracy with which both vowels were identified correctly was 49.2% (cascade vowels) and 54.2% (six-harmonic vowels). Every subject performed above the chance level of 6.7%, assuming that chance performance would involve picking responses randomly from the 5 identical pairings and the 10 nonidentical pairings. Thus, despite differences in the language, in other details of the stimuli, and in the procedures for presenting them, the experiment replicated the results of Scheffers (1983) and Zwicker (1984) in showing that listeners perform at levels above chance when identifying the constituents of double vowels.

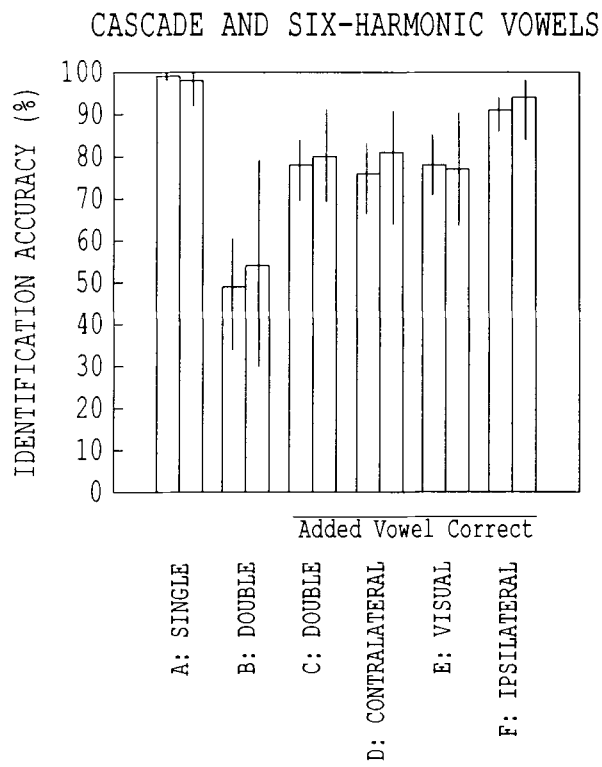

Figure 4. The percentage of correct responses for each condition averaged over subjects, blocks, and vowels. The left-hand member of each pair of bars indicates results obtained with cascade vowels. The right-hand member indicates results obtained with six-harmonic vowels. A: single vowels; B: double vowels, both correct; C: double vowels, added vowel correct; D: contralateral precursor; E: visual precursor; F: ipsilateral precursor. Vertical lines indicate the range of scores over subjects.

Condition 2 was then rescored to allow it to be compared with the three conditions with precursors. For each pair of vowels, we determined the accuracy of identification of the vowel corresponding to the *added* vowel (the vowel not defined by the precursor) in the conditions with precursors. For example, to enable responses to the double vowel /a,i/ to be compared with responses in conditions where the precursor was /a/ and the added vowel /i/, we scored the response to /a,i/ as correct if either response given was /i/. Then, to enable responses to the pair /a,i/ to be compared with responses in conditions where the precursor was /i/ and the added vowel /a/, we scored the response to /a,i/ as correct if either response given was /a/. Mean accuracy in Condition 2 was 78.0% (cascade vowels) and 80.2% (six-harmonic vowels).

In Conditions 3, 4, and 5, which involved precursors, performance was scored as the accuracy with which the added vowel was identified. Mean accuracies were (listing the result for the cascade vowels first in each case): Condition 3 (ipsilateral precursor), 90.9% and 93.7%; Condition 4 (contralateral precursor), 76.3% and 80.6%; Condition 5 (orthographic precursor), 77.9% and 76.8%.

To compare the double-vowel condition with the three conditions with precursors, one-way ANOVAs were run with the data pooled over blocks. The effect of condition was significant in each case [$F(3,12) = 41.02$, $p < .001$, for the cascade vowels, and $F(3,12) = 37.33$, $p < .001$, for the six-harmonic vowels]. Post hoc tests using the criteria recommended by Scheffé (Winer, 1971) confirmed that the added vowel was identified more accurately with the ipsilateral precursor both for the cascade vowels, [$F(3,12) = 143.0$, $p < .001$] and for the six-harmonic vowels [$F(3,12) = 127.3$, $p < .001$] than in the other three conditions. These did not differ from one another [$F(2,12) = 2.28$ for the cascade vowels and $F(2,12) = 3.53$ for the six-harmonic vowels].

Thus, of the three ways of presenting prior information about one of the two vowels in a double vowel, only one led to any advantage in identifying the other vowel: it is advantageous to hear one of the two vowels in the same ear as the double vowel, but neither hearing it in the contralateral ear, nor reading its orthographic representation, provides any advantage. This outcome is unlikely to have been due to inaccurate identification of the precursor vowels. The ipsilateral, contralateral, and visual precursors were themselves identified with accuracies of 98%, 98%, and 95% for the cascade vowels, and 97%, 96%, and 95% for the six-harmonic vowels.[3]

Pattern for Individual Double Vowels

As shown by Scheffers (1983), it is very much easier to identify some nonidentical double vowels than others.

With the cascade vowels, performance ranged from 16% correct (/a,ɜ/) to 82% correct (/a,i/). With the six-harmonic vowels, performance ranged from 33% correct (/a,ɜ/) to 80% correct (/i,i/). Assmann and Summerfield (1989) have demonstrated that the accuracy with which the constituents of double vowels are identified is predicted by the extent to which the lowest three formants of each are resolved in the auditory excitation pattern of the double vowel.

The major feature of the overall pattern of results shown in Figure 4, better performance with ipsilateral precursors than in other conditions involving double vowels, was produced with the majority of individual double vowels. For 20 of the 25 cascade double vowels, and for 21 of 25 six-harmonic double vowels, performance with ipsilateral precursors was as good as, or better than, performance in any other condition involving double vowels.

GENERAL DISCUSSION

Effects of Precursors

The accuracy with which listeners identify either constituent of a double vowel is increased if the other member is presented as an acoustical precursor in the same ear, but not if it is presented to the contralateral ear, nor if its identity is conveyed by a visual display of its orthographic representation. In other words, listeners cannot generally take advantage of a prior specification of the phonemic identity of one constituent of a double vowel to recover the other constituent. This outcome seems to be incompatible with the idea that listeners use the strategy of categorization-guided segregation to identify the constituents of double vowels.

Elsewhere (Assmann & Summerfield, 1989) we have implemented the alternative strategy of simultaneous independent comparisons as a computational procedure. We have established the extent to which it can predict the pattern of listeners' identification responses to double vowels in the conditions of the present experiments without precursors. The model predicts the probability with which each pair of identification responses will be made to each isolated double vowel. The product-moment correlation between the predictions of the most successful form of the model and the performance of the listeners was 0.94 for both the cascade and six-harmonic vowels. This moderately high correlation, together with the absence of any benefit from contralateral or visual precursors in the present experiments, suggests that listeners' performance in identifying the constituents of isolated double vowels that possess the same fundamental frequency is described adequately by the strategy of simultaneous independent comparisons.[4]

The result that prior knowledge of the acoustical structure of one constituent of a double vowel does not help a listener to identify the other constituent is counterintuitive. Such knowledge might be expected to be beneficial, given that many engineering solutions to the problem of retrieving signals from noise benefit from, and may require, a prior sample of the noise (see, e.g., Lim, 1983). The present result may be specific to the stimuli and tasks described here, within which there were no cues that would lead listeners to assign the two constituents of the double vowel to different sources. However, a similar result has been described before. Haggard (1974) required listeners to identify words whose spectral components had been transposed to higher or lower frequencies randomly from trial to trial. An introductory phrase processed with the same distortion as was the test word improved performance, but an unprocessed phrase that told the listener which transposition to expect, but did not exemplify it, did not help. This result, and the failure of listeners to benefit from visual precursors in the present experiments, suggests that there are limits to the ability of listeners to use prior knowledge of acoustical distortions to discount their effects.

Similarity of Results for Different Vowel Types

Accurate identification of the constituents of double vowels was found for stimuli synthesized with natural spectral envelopes and with simplified envelopes. The cascade vowels possessed natural spectral envelopes while the six-harmonic vowels preserved the frequency locations of formant peaks, but neither their relative amplitudes nor other details of spectral shape. This outcome is compatible with demonstrations by Carlson, Granstrom, and Klatt (1979), who showed that judgments of the phonetic similarity of vowels are influenced by the frequency locations of formant peaks, but are affected only a little by the bandwidths of formants and the overall tilt of the spectrum.

There are several advantages in attaching more importance to formant-peak frequencies than to other aspects of spectral shape. Peaks are intrinsically resistant to masking by background noises. Their frequencies are affected only a little by changes in overall spectral tilt resulting, for example, from differences in source spectrum or ambient absorption, or by other aspects of the frequency response of the communication channel, or by the frequency selectivity of the listener. The robustness of formant peaks, combines with the enhancement of increments in spectral amplitude shown here by the DSV effect, to help listeners to discount spectral coloration from noisy communication channels with uneven frequency responses.

SUMMARY AND CONCLUSIONS

Listeners identified the constituents of double vowels created by summing the waveforms of synthetic vowels with the same fundamental frequency and their pitch pulses in synchrony. Performance in identifying both constituents was significantly above chance, replicating reports by Scheffers (1983) and Zwicker (1984).

Double vowels were presented both in isolation and in conditions in which they were introduced by precursors. Precursors specified either the acoustical or the phonemic identity of one of the constituents of the double vowel. Performance was compared across conditions by means

of an assessment of the accuracy with which the other constituent was identified. Compared with the condition in which double vowels were presented in isolation, neither an orthographic precursor nor an acoustical precursor in the contralateral ear produced an improvement in performance. This outcome seems to be incompatible with the idea that the constituents of double vowels are identified by first establishing the vowel whose spectrum best matches that of the double vowel, and then subtracting this spectrum to reveal the other vowel. It is more likely that listeners establish the two vowels whose spectra independently best match the spectrum of the double vowel.

Performance improved when an acoustical precursor was presented to the same ear as was the double vowel. This result illustrates the operation of auditory and perceptual processes that highlight changes in spectral amplitude. These processes should contribute to the ability of listeners to maintain perceptual constancy for speech sounds despite the uneven frequency responses of many communication channels.

These results were obtained both with vowels possessing natural spectral envelopes, and with stimuli possessing simplified spectral envelopes that preserved the frequency locations of spectral peaks but not other details of spectral shape. One advantage of the strategy of identifying vowels primarily from spectral peaks is that peaks are more resistant to interfering noise, including the sound of competing talkers, than are other details of spectral shape.

REFERENCES

ANSI (1969). *American national standards for audiometers* (ANSI S3.6-1969). New York: American National Standards Institute.

ASSMANN, P. F., & SUMMERFIELD, A. Q. (1989). Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency. *Journal of the Acoustical Society of America*, **85**, 327-338.

CARDOZO, B. L. (1967). *Ohm's Law and masking* (IPO Annual Progress Report No. 2, pp. 59-64). Eindhoven, The Netherlands: Institute for Perception Research.

CARLSON, R., GRANSTROM, B., & KLATT, D. H. (1979). *Vowel perception: The relative perceptual salience of selected acoustic manipulations* (Quarterly Progress Report on Speech Research, STL-QPSR 3-4/1979, pp. 73-83). Stockholm, Sweden: Speech Transmission Laboratory, Royal Institute of Technology.

DARWIN, C. J. (1984a). Auditory processing and speech perception. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and performance X: Control of language processes* (pp. 197-209). London: Erlbaum.

DARWIN, C. J. (1984b). Perceiving vowels in the presence of another sound: Constraints on formant perception. *Journal of the Acoustical Society of America*, **76**, 1636-1647.

DARWIN, C. J., & GARDNER, R. B. (1987). Perceptual separation of speech from concurrent sounds. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception* (pp. 112-124). Dordrecht, The Netherlands: Martinus Nijhoff.

HAGGARD, M. P. (1974). Selectivity for distortions and words in speech perception. *British Journal of Psychology*, **65**, 69-83.

KAISER, J. F. (1966). Digital filters. In F. F. Kuo & J. F. Kaiser (Eds.), *Systems analysis by digital computer* (chap. 7). New York: Wiley.

KLATT, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, **67**, 971-995.

LIM, J. S. (1983). *Speech enhancement*. Englewood Cliffs, NJ: Prentice-Hall.

SCHEFFERS, M. T. M. (1983). *Sifting vowels: Auditory pitch analysis and sound segregation*. Unpublished doctoral dissertation, University of Groningen, Groningen, The Netherlands.

SMITH, R. L. (1979). Adaptation, saturation, and physiological masking in single auditory-nerve fibers. *Journal of the Acoustical Society of America*, **65**, 166-178.

SMITH, R. L., BRACHMAN, M. L., & FRISINA, R. D. (1985). Sensitivity of auditory-nerve fibers to changes in intensity: A dichotomy between decrements and increments. *Journal of the Acoustical Society of America*, **78**, 1310-1316.

SUMMERFIELD, A. Q., & ASSMANN, P. F. (1987). Auditory enhancement in speech perception. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception* (pp. 140-150). Dordrecht, The Netherlands: Martinus Nijhoff.

SUMMERFIELD, A. Q., HAGGARD, M. P., FOSTER, J. R., & GRAY, S. (1984). Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect. *Perception & Psychophysics*, **35**, 203-213.

SUMMERFIELD, A. Q., SIDWELL, A., & NELSON, A. (1987). Auditory enhancement of changes in spectral amplitude. *Journal of the Acoustical Society of America*, **81**, 700-708.

VIEMEISTER, N. F. (1980). Adaptation of masking. In G. van den Brink & F. A. Bilsen (Eds.), *Psychophysical, physiological, and behavioral studies in hearing* (pp. 190-198). Delft, The Netherlands: Delft University Press.

VIEMEISTER, N. F., & BACON, S. (1982). Forward masking by enhanced components in harmonic complexes. *Journal of the Acoustical Society of America*, **71**, 1502-1507.

WILSON, J. P. (1970). An auditory after-image. In R. Plomp & G. F. Smoorenburg (Eds.), *Frequency analysis and periodicity detection in hearing* (pp. 303-315). Leiden, The Netherlands: A. W. Sijthoff.

WINER, B. J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.

ZWICKER, U. T. (1984). Auditory recognition of diotic and dichotic vowel pairs. *Speech Communication*, **3**, 265-277.

NOTES

1. The conditions of the pilot experiment were the same as those of the experiment described in this paper, and the data were scored in the same ways. Mean accuracy (5 subjects) in the six conditions shown in Figure 4 was: A: single vowels, 97%; B: double vowels (both correct), 33%; C: double vowels (added vowel correct), 71%; D: contralateral precursor, 68%; E: visual precursor, 71%; F: ipsilateral precursor, 86%.

2. The grapheme sequence *OR* provides an acceptable representation of the vowel /ɔ/ because postvocalic *r* is not pronounced in most dialects of British English.

3. The less than perfect accuracy with which listeners responded to the visual precursors reflects an occasional failure to make responses in the correct order, rather than any limitation of the VDU display or in the subjects' abilities in interpreting it.

4. This conclusion may apply only to the case where cues for source segregation are absent. It is likely that the strategy of multiple independent comparisons is supplemented by processes of spectral subtraction when cues such as differences in fundamental frequency or time of onset allow concurrently presented vowels to be assigned to different sources.

(Manuscript received February 19, 1988;
revision accepted for publication November 17, 1988.)