Perception & Psychophysics
1993, 53 (2), 125-133

The perceptual segregation of simultaneous vowels with harmonic, shifted, or random components

MAGDALENE H. CHALIKIA and ALBERT S. BREGMAN
McGill University, Montreal, Quebec, Canada

This experiment was an investigation of the ability of listeners to identify the constituents of double vowels (pairs of synthetic vowels, presented concurrently and binaurally). Three variables were manipulated: (1) the size of the difference in F0 between the constituents (0, ½, and 6 semitones); (2) the frequency relations among the sinusoids making up the constituents: harmonic, shifted (spaced equally in frequency but not integer multiples of the F0), and random; and (3) the relationship between the F0 contours imposed on the constituents: steady state, gliding in parallel, or gliding in opposite directions. It was assumed that, in the case of the gliding contours, the harmonics of each vowel would "trace out" their spectral envelope and potentially improve the definition of the formant locations. It was also assumed that the application of different F0 contours would introduce differences in the direction of harmonic movement (common fate), thus aiding the perceptual segregation of the two vowels. The major findings were the following: (1) For harmonic constituents, a difference in F0 leads to improved identification performance. Neither tracing nor common-fate differences add to the effect of pitch differences. (2) For shifted constituents, a difference between the spacing of the constituents also leads to improved performance. Formant tracing and common fate contribute some further improvement. (3) For random constituents, tracing does not contribute, but common fate does.

In most listening situations, we rarely hear a single sound in complete isolation. Several sound sources are often active at the same time, producing a complex pattern of vibrations on our eardrums. The auditory system is, therefore, faced with the problem of distinguishing the different sets of components that correspond to separate sound sources. Otherwise, it would not be possible to understand, for example, what one speaker is saying in the presence of competing speakers or background noises.

Different experiments have studied the ability to selectively attend to one speech signal in a mixture of continuous speech signals (Broadbent, 1952; Brokx & Nooteboom, 1982; Cherry, 1953; Darwin, 1981). These studies have suggested that perceptual separation can improve when the signals have different pitches. Scheffers (1983) investigated the effects of fundamental frequency (F0) differences on the identification of two simultaneous steady-state synthetic vowels and confirmed the hypothesis that the two vowels can be more easily separated when the F0s differ by more than 1-2 semitones. Listeners' ability to identify both vowels improved by about 18% with different F0s. However, separability, which was attributed by Scheffers to the limited frequency selectivity of the ear, seemed to reach a maximum around 2-3 semitones.

Subsequent studies with superimposed vowels (Assman & Summerfield, 1990; McAdams, 1989; Summerfield & Assman, 1991; Zwicker, 1984) have also shown that it is easier to identify the steady-state components of such double vowels when their F0s are different. Scheffers (1983) and Assman and Summerfield (1990) have developed models capable of identifying the components of double vowels at a level of success approximating that of human listeners. However, neither model has been able to reproduce the gradual improvement in performance found with increasing F0 separation, a difficulty overcome by a recent model (Meddis & Hewitt, 1992). Meddis and Hewitt's model segregates sounds in a way similar to the autocorrelation method used by other researchers (Weintraub, 1985, 1987). The output of each filter in an initial bandpass filtering stage (simulating the characteristics of the human auditory periphery) is autocorrelated to extract pitch and timbre information. The pooled autocorrelation function based on all channels is used to derive a pitch estimate for one of the component vowels from a signal composed of two vowels. Frequency-selective channels are segregated into two mutually exclusive sets, if two pitches are found. Each vowel is then identified through some template-matching procedure based on pooled periodicity profiles summed across channels belonging to a set of channels. This procedure does not require an accurate pitch estimation, but the separation of channels into two subsets.

This research was supported by the National Sciences and Engineering Research Council of Canada, through a grant awarded to A. S. Bregman. The valuable contributions of P. Ahad, B. Roberts, S. McAdams, C. Darwin, and Q. Summerfield are gratefully acknowledged. M. H. Chalikia is now in the Department of Psychology, Moorhead State University, Moorhead, MN 56563. A. S. Bregman is in the Department of Psychology, McGill University, 1205 Dr. Penfield Ave., Montreal, Quebec H3A 1B1, Canada. Correspondence may be sent to either author.

This separation requires only one pitch estimate. The model seems to work well with harmonically structured components of a double vowel with two different steady pitches.

In a recent study (Chalikia & Bregman, 1989), pairs of simultaneous vowels were used for which the F0s were steady, gliding in parallel, or gliding in opposite directions (crossed glides) so that the pitch contours crossed one another midway through the stimulus duration. Also, the maximum F0 separation between vowels varied from 0 to 12 semitones. The results confirmed previous findings with the steady-state (SS) vowels. That is, with a ½-semitone F0 difference, vowel identification performance improved considerably. In fact, the greatest increment appeared in the first ½ semitone, a frequency ratio of 1.03. This finding is in agreement with those of other researchers (Darwin & Gardner, 1986; Moore, Glasberg, & Peters, 1985, 1986; Moore, Peters, & Glasberg, 1985) showing that a partial of a harmonic series that is mistuned by 3% or more starts to no longer fuse with the other harmonics. Chalikia and Bregman (1989) mistuned a group of harmonics (those of one vowel relative to the harmonics of the other), rather than a single harmonic; however, the effects were similar. Performance remained steady until the octave separation, where there was a decrease (most likely due to the great amount of overlap between the harmonics of the two vowels). Although there was a tendency for parallel glides (PGs) to give lower identification scores than crossed glides (CGs), the only significant difference between the two glide conditions was at the octave, where there was a decrease in performance for PGs (as with the SSs) where a harmonic relation existed between the components of both vowels, but not for the CGs where the relation did not exist. However, for F0 separations of 3, 6, and 12 semitones, the CGs produced better separation than did the SSs. Therefore, it appeared that the effect of CGs was additional to that produced by an F0 separation.

A "spectral-peak-picker" mechanism, suggested by Chalikia and Bregman (1989), could account for the fact that the constituents of vowel pairs often could be identified well, even when they had the same F0. Such a mechanism could parse the general spectrum on the basis of spectral peaks. Peaks would contribute to the identification of each vowel on the basis of some kind of spectral pattern matching (Klatt, 1980; Scheffers, 1983). Several researchers have stressed the importance of spectral peaks in vowel identification (Carlson, Granstrom, & Fant, 1970; Chistovich, 1971; Joos, 1948). Assman and Summerfield (1989) have shown that identification performance can be predicted by assuming that listeners simultaneously match templates representing all available responses ("simultaneous independent comparisons"), and select the two whose formants more closely match the candidate formants. When the spectral envelopes of the vowels are very dissimilar, the mixture can be decomposed easily. However, when the spectral envelopes are similar, the task of separation becomes difficult.

When an F0 difference is introduced, additional cues are available that contribute to the decomposition of the overall spectrum and the subsequent identification of the two vowels. Increased accuracy of identification could occur under either of two conditions:

1. Performance could improve if there were additional evidence to allow the two sources (vowels) to be distinguished. A difference in F0 could provide one cue, that is, membership in a particular harmonic series (regardless of whether or not the two vowels were SSs or glides). A different direction of F0 change could provide another cue, called "common fate." Common fate would result in the reinforcement of harmonicity binding the components of a vowel together, and would separate them from the components of another vowel if they moved in a different direction (as in the case of CGs). Common fate would predict that CGs should be better than PGs, as well as better than SSs. Only CGs would provide differential grouping of the partials, since a separate motion trajectory would group the partials of each vowel. Common fate could explain the superiority of the CGs over SSs (or PGs at the octave) in the Chalikia and Bregman (1989) study. McAdams (1989) has found that modulating harmonics make a vowel in a complex of three vowels more prominent than nonmodulating ones.

2. Performance could also improve if the candidate formants were defined more accurately. Increasing the overall number of harmonics could do this. An F0 difference would in effect double the number of harmonics as compared with the case of no separation. A minimum separation of ½ semitone is required for the effect to be evident. It is possible that F0 difference benefits were not found at smaller separations because there were strong interactions between corresponding harmonics in the two vowels, which grossly distorted the overall spectral envelope in the region of F1. Another cue that could contribute to the definition of formant peaks is "formant tracing." It has been argued that the presence of gliding pitch contours would contribute to the parsing of the spectrum because of formant tracing (McAdams & Rodet, 1988). As an F0 (and all its harmonics) glided in frequency, the amplitude envelope of the vowel would be "traced out" by the changes of the amplitudes of the harmonics. In this manner, the two vowel spectra would be better defined and, therefore, more separable. However, recent work (Marin & McAdams, 1991) did not support the hypothesis that vowel separation may be due to spectral envelope tracing. The researchers uncoupled envelope tracing from frequency modulation by keeping the amplitudes of modulating harmonics fixed. They found that frequency modulation contributed to the perceived prominence of a vowel, even without spectral tracing. This effect may have been due, therefore, to common fate or to an increased sensitivity of the auditory system to spectral regions in which frequencies are changing.

The results of Chalikia and Bregman (1989) showing a superiority of CGs over SSs (and PGs at an octave pitch separation) are consistent with either an explanation in terms of the improvement of the definition of formants or an explanation based on common fate. However, other studies in which one harmonic (Gardner & Darwin, 1986) or a formant (Gardner, Gaskill, & Darwin, 1989) in a vowel-like spectrum is frequency modulated differently (or incoherently) from the remainder of the spectrum have shown that the modulation has no independent effect (in addition to harmonicity) on the perceptual grouping of harmonics or formants. The latter studies consider the effect of F0 differences to be the most important cue. Nevertheless, Chalikia and Bregman have shown that common fate improves the identifiability of a vowel and its perceptual separation of another vowel, compared with SSs. Also, McAdams (1989) and Marin and McAdams (1991) have shown that modulated vowels are judged to be more prominent than unmodulated vowels, even though identifiability per se was not tested. It is possible that the auditory system may be able to take better advantage of incoherent modulation patterns when the separate patterns are defined by more than a single harmonic or a few harmonics.

One question that may be asked, regarding the Chalikia and Bregman (1989) results, is whether an improvement in identification scores can imply that better segregation of the vowels has occurred. If two vowels are detected, then there has to have been segregation at some level, at least at the level of recognition. A more crucial point at issue may be not whether there has been segregation, but what kind. Is it the kind of segregation that occurs when two vowel schemas both receive adequate stimulation and are activated—a process that has been referred to (Bregman, 1990) as schema-based segregation—or is it primitive segregation, based on general acoustic cues to the presence of two signals? It may be impossible, with vowels, to have a test of source segregation that is independent of recognition. However, it may be possible to distinguish the two types of segregation (primitive vs. schema driven). For example, PGs could help schema-driven segregation by better defining the spectral peaks. Any improvement of CGs over PGs could be attributed to primitive segregation, because CGs do not improve the information available to the speech schemas but do supply common fate, a cue not limited to speech signals, and therefore possibly used by primitive processes of segregation.

One issue that the above experiments do not address is the question of whether it is important not only that the components of one vowel undergo similar (coherent) frequency changes, but that they should be harmonically related in order for their grouping to occur. The effects of F0 differences with the SS pairs seem to indicate that the existence of harmonic relations is important. McAdams (1989) and Marin and McAdams (1991) found no difference between coherent and incoherent vibratos on vowels separated by 5 semitones. It was concluded that harmonicity may be a constraining factor on the grouping power of coherent modulation and that harmonicity is probably a stronger grouping cue than frequency-modulation coherence. On the other hand, Bregman, Levitan, and Liao (1990) have shown that harmonicity has little or no effect on perceptual fusion due to coherent amplitude modulation.

McAdams (1984, Appendix F) has reported that coherent frequency changes can make partials fuse (i.e., become part of the same subset and thus heard as separate from others) in the absence of good harmonic relations. That is, in a series made up of partials "stretched" on log-frequency coordinates, in which constant frequency ratios have been maintained among the frequency components, listeners tended to hear more sources when vibratos on even and odd harmonics were incoherent compared with when they were coherent. In contrast, Bregman and Doehring (1984) have found that it is not sufficient for partials to glide in parallel in order for fusion to occur; they must also maintain simple harmonic relations.

The present experiment was designed to explore the contribution of harmonicity to the grouping of components that have SS or gliding (coherent or incoherent) pitch contours. In addition to vowels with harmonic partials, two different sets of vowels with inharmonic partials were also used. One set was made up of shifted partials, that is, partials that are spaced equally in frequency but are not integer multiples of a common F0. The other set was made up of random partials.

METHOD

Subjects

Twelve paid volunteer males and females were used from the McGill University population.

Apparatus

All the stimuli were synthesized on a Compaq 386/20 PC using the MITSYN signal-processing software (Henke, 1987). The signals were played via a 16-bit DAC at a sampling rate of 16 kHz. Following antialias filtering at 4.5 kHz, the stimuli were amplified by a Pioneer (SA8500II) amplifier and presented binaurally over Sennheiser 414 headphones at about 80 dBA as measured by a sound-level meter (General Radio 1551-C) at A weighting.

Stimuli

Five vowel sounds were synthesized using a serial three-formant method. The glottal pulse was created by additive synthesis of 40 components in sine phase, with an intensity drop-off of −12 dB/octave, modified by a subsequent radiation characteristic imposed by a first-order difference filter, yielding a net spectral slope of −6 dB/octave. This method allowed the independent specification of each component's frequency and amplitude. Thus it was possible to create different spectra with harmonic, shifted, or random component sequences (see below) and still conform to the intensity drop-off of −12 dB/octave. The formants were imposed by a series of three formant filters. The vowels were /i/, /ɛ/, /ɑ/, /u/, and /ɔ/, with a duration of 1 sec each, including 200 msec rise/fall. Each had an F0 at 140 Hz.

The formant frequencies were set equal to the ones found in Peterson and Barney (1952) for a male voice and are shown in Table 1. The formant frequencies remained constant at these values, regardless of any changes in the F0s. The bandwidths for the formant filters were set at 100, 120, and 140 Hz for F1, F2, and F3, respectively.

The vowels were presented as nonidentical pairs. All possible combinations of the five vowels, taken two at a time, resulted in 10 pairs. Further pairs were created by altering the F0 of each vowel (see below) so that the F0s of the two vowels were separated by

Table 1
Formant Frequencies (in hertz) of the Vowels

Formant	i	ɛ	ɑ	u	ɔ
F1	270	530	730	300	570
F2	2290	1840	1090	870	840
F3	3010	2480	2440	2240	2410

Table 2
Values of the F0s for the Different Semitone Frequency Separations

Middle Value	High F0	Low F0	F0 Separation
	Steady States and Crossed Glides		
140.00	140.00	140.00	0.00
140.00	142.10	137.90	0.50
140.00	166.51	117.70	6.00
	Parallel Glides		
140.00	140.00	140.00	0.00
140.00	144.20	135.90	0.50
140.00	197.90	98.90	6.00

Note—Middle value, high F0, and low F0 are in hertz; F0 separation is in semitones.

½ semitone and 6 semitones, with the two F0s placed symmetrically around 140 Hz (on log-frequency coordinates). The frequency values corresponding to these separations are shown in Table 2.

The case where a vowel had an F0 of 140 Hz served as a reference (standard) stimulus on the basis of which the other stimuli were synthesized. The pitch contour (change in F0 over time) used for the synthesis was SS or gliding. In the SS case, both constituents of the double vowel were SS. In the gliding case, both constituents had changing F0s, but there were two conditions. In the CG condition, one vowel glided down and the other glided up in frequency, that is, their pitch contours were incoherent. In the PG condition, both vowels glided up, that is, their pitch contours were coherent (see Figure 1). Except for the standard case, where both vowels in the pair were SS with an F0 at 140 Hz, in all other cases their F0s belonged in a "high" or "low" frequency range (above or below the original F0 of the standard). A pair always had a high and a low component (with the relations between the components being one of SSs, PGs, or CGs).

For example, in the case where the F0 difference was ½ semitone, for SSs the high vowel had an F0 of 142.1 Hz and the low vowel had an F0 of 137.93 Hz, yielding a ratio of 1.03. For CGs, the F0 of the low vowel started at 137.93 Hz and ended at 142.1 Hz, and the F0 of the high vowel started at 142.1 Hz and ended at 137.93 Hz. In this case, the terms "high" and "low" defined the points at which the glides started. Both glides swept through the same range of frequencies, one gliding up and the other gliding down. In the PGs, the F0s of the two glides maintained a constant semitone separation as they glided upward. For example, in the case of ½-semitone separation (as above), the high glide glided up starting at 140 Hz and ending at 144.2 Hz, and the low glide glided up starting at 135.9 Hz and ending at 140 Hz. All glides were linear on log-frequency coordinates.

The F0 separations listed in Table 2 indicate constant frequency separations for the SSs, but only maximum separations for the CGs. That is, for both SSs and PGs, a difference of, say, ½ semitone between the F0s in the pair refers to a constant frequency separation of that magnitude maintained throughout the duration of the signal. For the CGs, the same difference refers to the maximum frequency separation obtained only at the beginning and the end points and to less than that separation at all the points in between. Therefore, the given nominal frequency separation of the CGs overestimates their F0 separation and makes it harder for them to be more segregated than the "F0-matched" SS or PG condition.

Table 3 shows the 10 vowel pairs that were used, as well as the pitch of each constituent. Chalikia and Bregman (1989, Experiment 2) had created an additional set of vowel pairs, opposite in assignment of F0 to vowel, and had found no difference between the two sets. In other words, given a vowel pair and a pitch difference, it did not really matter which was the high vowel and which was the low one. Therefore, only one high/low ordering of each vowel pair was used, as shown in Table 3, and each vowel occurred twice in a high position and twice in a low position.

All the vowel pairs were synthesized using the three types of pitch contours for all the F0 separations. The total number of pairs was 90 (10 pairs at each of the pitch separations—0, ½, and 6 semitones—for SSs, PGs, and CGs). When the F0 difference was zero, all contours degraded to SS and were thus equal to what was called the "standard" stimulus for each vowel.

In the *harmonic* set, all 40 partials of a vowel were harmonic, that is, they were all integral multiples of F0. Two more sets of 90 pairs each were created, where each vowel had shifted or random components.

In the *shifted* set, each vowel was created as follows. Each of the 40 partials was created by adding 35 Hz to the harmonic values. That is, any two adjacent partials would have a constant frequency separation of F0, but the components would form an inharmonic series whose lowest frequency was 175 Hz. The amplitude levels of the partials were controlled by another vector of values so that they conformed to the requirement of an intensity drop-off of −12 dB/octave (subsequently modified by the difference filter to yield a net spectral slope of −6 dB/octave). These amplitude and frequency values were input to the formant filters.

In the *random* set, the frequency of each partial was derived by taking the corresponding harmonic value and displacing it by a random amount constrained to be not less than plus or minus the value of F0 × 0.45. The vector values are shown in Table 4. In this manner, even though each vowel would be composed of nonharmonic partials, the two vowels were excited by patterns of frequency components with an average density equivalent to different F0s. The amplitude levels of the components were adjusted in the manner described for the shifted set.

The total number of pairs used was 270. All vowels were equated for total *RMS* amplitude. There were four independent variables, vowel pair (10 vowel pairs), pitch contour (SS, CG, or PG), F0 separation (0, ½, and 6 semitones), and harmonicity (harmonic, shifted, or random components). The dependent variable was the identification score.

Procedure

The experiment started with a training session. The listeners were each seated individually in a test chamber. At first, they were presented with the individual vowels so that they could familiarize themselves with the identification of synthetic vowels. One hundred and

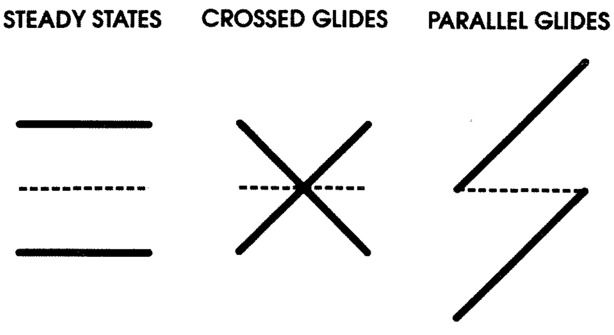

Figure 1. Illustration of the pitch differences for SSs, CGs, and PGs. Dashed lines represent the SS 140-Hz F0.

Table 3
The 10 Vowel Pairs Used and Their Pitch

Pitch of Constituent	
Low	High
ɛ	i
i	ɑ
u	i
i	ɔ
ɑ	ɛ
ɛ	u
ɔ	ɛ
ɑ	u
ɔ	ɑ
u	ɔ

Table 4
Random Vector Values Used to Generate the Stimuli for the Random Set

1.00, 1.56, 3.29, 4.11, 5.41, 5.61, 6.70, 8.35, 8.70,
9.62, 11.12, 12.34, 13.33, 14.30, 14.83, 16.30, 17.17, 17.99,
19.17, 20.27, 20.86, 22.03, 23.03, 23.74, 25.56, 27.31, 27.67,
29.39, 30.39, 30.77, 31.87, 32.76, 33.84, 34.93, 36.42, 36.89,
38.24, 38.68, 40.10, 40.73

five vowels were presented binaurally in a random order (45 SS, 5 for each harmonicity and pitch-contour condition; 30 PG, excluding the 0 F0 separation; 30 CG), and the subjects had to make an identification response for each vowel by pressing one of five keys on a keyboard. Phonetic symbols and words were used to indicate the possible choices. The words were "heed" for /i/, "head" for /ɛ/, "hard" for /ɑ/, "who'd" for /u/, and "hoard" or "hawed" (they were both presented all the time) for /ɔ/. There was no feedback to the subjects. The listeners proceeded with the second part of the training if they had correctly identified each vowel (out of the 35) in 2 out of 3 judgments. All listeners identified all vowels correctly. In the second part of the training, all vowel pairs were presented, once each. The listeners were told that pairs of vowels would be presented and that they had to identify *both* constituents. The subjects were not told that only nonidentical stimulus pairs would be presented. The listeners were allowed breaks after each set of 90 pairs. Actual testing started on a different day. The listeners were given three testing sessions of about 1 h each on 3 different days. One session presented the 90 harmonic pairs, one the 90 shifted pairs, and one the 90 random pairs. The listeners were assigned to the testing sessions in a counterbalanced fashion. All listeners heard all 270 pairs. In each session, six sets of 90 trials were presented in a random order. Breaks were given after trials 90, 240, and 390. The first set of responses for any session was considered as training and was not included in the analysis.

RESULTS

Scoring and Analysis

On each trial, each subject received a score of 0, 1, or 2, depending on whether none, one, or both of the vowels in the pair had been identified correctly. The scores were then converted to percentage correct, on the basis of five replications, and collapsed over the different vowel pairs. These averaged values were used in the analysis.

The questions of interest that the analyses were intended to answer were the following: (1) Does a difference in F0 between the constituents of a vowel pair contribute to their separation? (2) Does the relationship between the pitch contours imposed on the constituents (SS, PG, CG) contribute to their separation? (3) Do the frequency relations among the sinusoids making up the constituents (harmonic, shifted, random) affect their separation? Figures 2, 3, and 4 show the mean results.

Overall, the F0 separation variable contributed to the identification of the vowels in the SSs ($p < .0001$), the PGs ($p < .0001$), and the CGs ($p < .0001$). Moreover, in the case of the PGs and CGs, the 6-semitone separation produced better results than did the ½-semitone separation ($p < .05$, and $p < .01$, respectively). This effect was also evident in the harmonic ($p < .0001$), shifted ($p < .0001$), and random ($p < .01$) pairs. The harmonic and shifted pairs produced better identification scores (at ½ and 6 semitones) than did the random pairs ($p < .01$). The contribution of the pitch-contour variable was evident in the 6-semitone separation, where the CGs produced better results than did either the PGs or the SSs ($p < .01$) and the PGs gave better results than did the SSs ($p < .05$). Separate two-way repeated measures analyses of variance for each harmonicity condition gave the following results.

In the harmonic set (Figure 2), only F0 separation gave a significant result [$F(2,22) = 21.85, p < .00001$]. The ½- and 6-semitone separations produced better results than did the 0-semitone separation ($p < .01$, Newman-Keuls). There was no effect of pitch contour.

In the shifted set (Figure 3), pitch contour [$F(2,22) = 4.090, p < .03$], "F0 separation" (there were no actual F0s for shifted partials) [$F(2,22) = 31.79, p < .00001$], and their interaction [$F(4,44) = 3.17, p < .02$] gave significant effects. Tests of simple effects showed that F0 separation contributed to the identification of the two vowels in the SSs [$F(2,22) = 27.75, p < .00001$], the PGs [$F(2,22) = 21.68, p < .0001$], and the CGs [$F(2,22) = 33.61, p < .0001$]. In all three cases, ½ semitone and 6 semitones produced better results than did 0 semitone ($p < .01$). In the PGs and CGs, 6 semitones produced better results than did ½ semitone ($p < .05$ and $p < .01$, respectively). The pitch-contour variable produced significant effects only in the 6-semitone separation [$F(2,22) = 11.50, p < .003$], where the CGs facilitated more correct identifications compared with the SSs ($p < .01$) or the PGs ($p < .05$). Also, the PGs gave better results than did the SSs ($p < .05$).

In the random set (Figure 4), the analysis gave a significant result for "F0 separation" (there were no actual F0s for random partials) [$F(2,22) = 5.86, p < .009$], a marginally significant result for pitch contour [$F(2,22) = 3.26, p < .06$], and a significant result for their interaction [$F(4,44) = 5.82, p < .0008$]. The F0 separation variable produced significant effects in the CGs only [$F(2,12) = 10.29, p < .001$]. Newman-Keuls tests indicated that a 6-semitone separation produced a better result than did either a ½- or a 0-semitone separation ($p < .01$). Also, at 6 semitones, CGs gave better results than did either the PGs or the SSs ($p < .05$). However, the PGs were not significantly different from the SSs ($p > .05$).

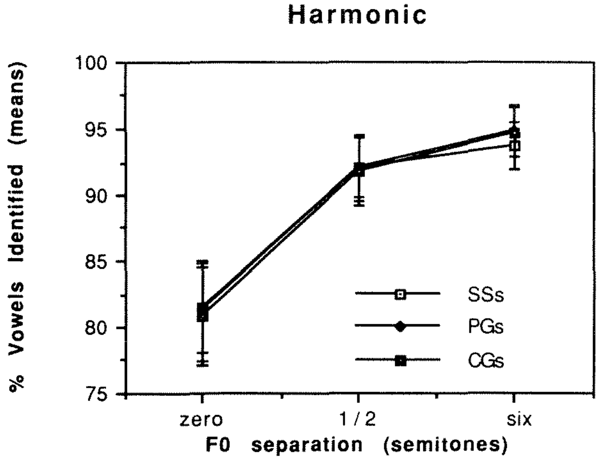

Figure 2. Harmonic set: Mean scores for each of the F0 separations, for each type of pitch contour.

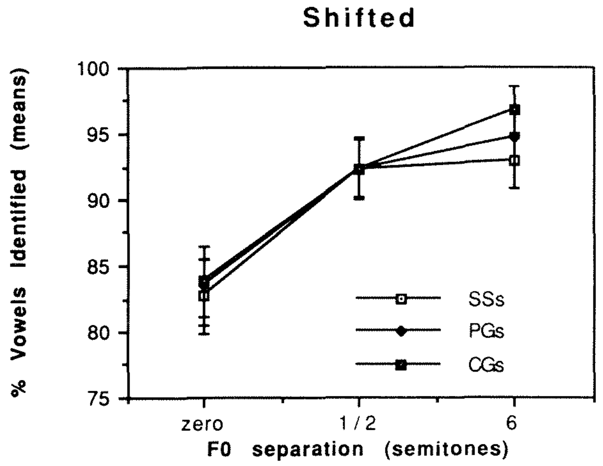

Figure 3. Shifted set: Mean scores for each of the F0 separations, for each type of pitch contour.

One result that was rather striking was the high performance of the listeners with vowels of the same pitch for all harmonicity sets (80%-85%). According to the scoring method used, the listeners received credit for getting any vowel correct. Some other authors (e.g., Assman & Summerfield, 1989, 1990) have scored similar data in terms of the percentage of trials in which both vowels were identified correctly, thus obtaining a more conservative measure. We thought it possible that the latter approach might be more sensitive by helping to avoid ceiling effects that could perhaps account for the absence of a pitch-contour effect in the harmonic set, in contrast with the findings of our earlier study (Chalikia & Bregman, 1989). We therefore rescored the data, giving the listeners credit only if they got both vowels correct (scores of 0 or 2), and repeated the analyses. The results are shown in Figures 5, 6, and 7.

Although scores were shifted downward by the more stringent criterion, especially for the 0 F0 separation, statistical analyses as well as visual inspection show that the pattern of results was unchanged. Moreover, the following improvements in performance were found. In the shifted set (Figure 6), a 6-semitone separation produced better results than did ½ semitone with SSs also ($p <$ .01, Newman-Keuls). In the random set (Figure 7), F0 separations of ½ and 6 semitones produced better results than did 0 semitone in the PGs as well ($p < .05$, Newman-Keuls). Also, 6 semitones improved performance over ½ semitone ($p < .01$) in the CGs. In general, then, with either scoring criterion, the results (concerning the effects of the independent variables) were comparable.

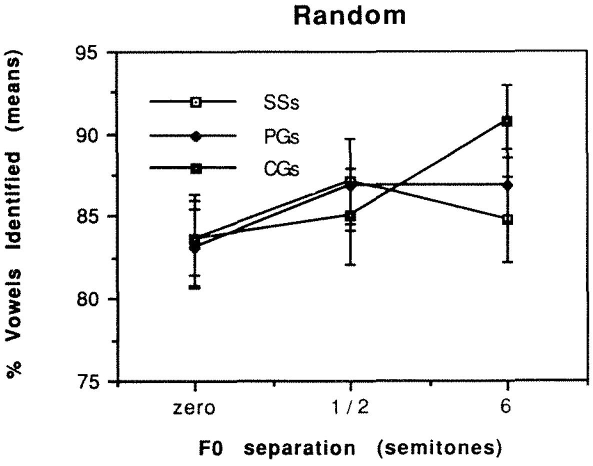

Figure 4. Random set: Mean scores for each of the F0 separations, for each type of pitch contour.

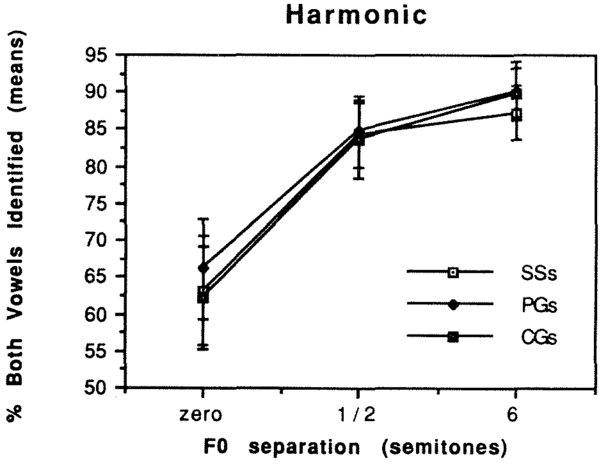

Figure 5. Harmonic set: Mean scores for each of the F0 separations, for each type of pitch contour. Criterion: both vowels correct.

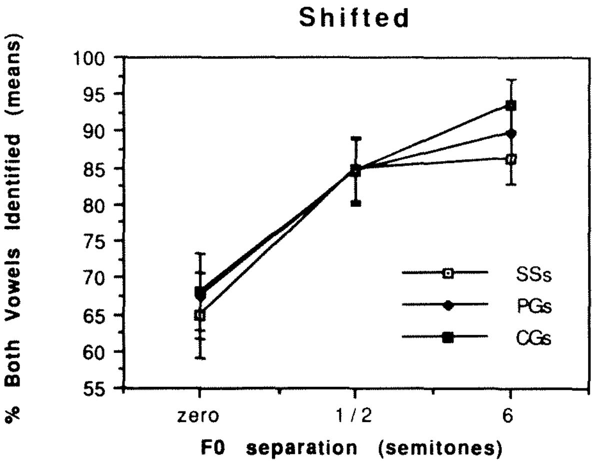

Figure 6. Shifted set: Mean scores for each of the F0 separations, for each type of pitch contour. Criterion: both vowels correct.

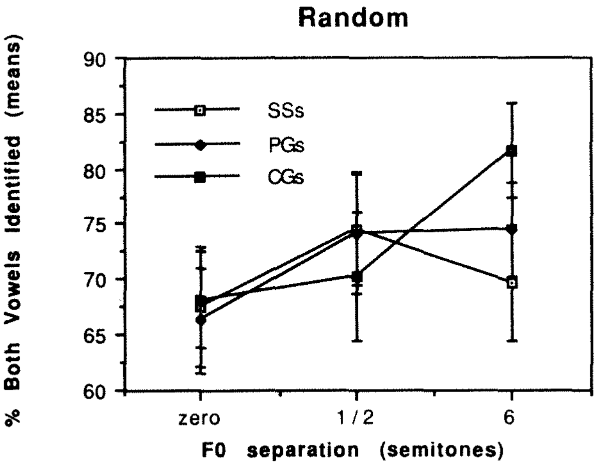

Figure 7. Random set: Mean scores for each of the F0 separations, for each type of pitch contour. Criterion: both vowels correct.

DISCUSSION

The results, for the harmonic condition, indicate that F0 differences facilitated vowel identification in the SS stimuli and agree with previous findings (Assman & Summerfield, 1990; Chalikia & Bregman, 1989; Halikia & Bregman, 1984a, 1984b; Scheffers, 1983; Zwicker, 1984). This effect was also evident with the PGs and the CGs, where, in addition to a difference of ½ or 6 semitones being better than 0 semitone, the 6-semitone difference was better than the ½-semitone difference.

The greater increment that resulted from separating the two F0s appeared in the first ½ semitone, a frequency ratio of about 1.03. Similar results were found in Chalikia and Bregman (1989). These data relate to findings (Darwin & Gardner, 1986; Moore, 1987; Moore, Glasberg, & Peters, 1985, 1986; Moore, Peters, & Glasberg, 1985) that show that a partial that is part of a harmonic series but is gradually being mistuned starts to segregate from the other harmonics at about a 3% separation.

A new finding, in the present study, was that the effect of F0 differences is observed not only with harmonic components, but also with shifted and random components. This is interesting because the partials in the last two cases are inharmonic. In the harmonic set, the F0 cue provided (on the basis of the presence of different periodicities) enough information to promote the grouping of components belonging to one F0 and their segregation from the group of components belonging to the other F0. Recent evidence (Darwin & Culling, 1990) suggests that the improvement in vowel identification with increasing pitch difference (up to about 2 semitones or more, if there are no inconsistencies between the lower and higher formant regions) is due almost entirely to changes in the first formant region. Other studies (Palmer, 1990) have suggested that the distribution of synchronized activity across the population of nerve fibers, or from computations based on intervals between discharges, allows the identification of the two F0s in a double vowel. The shifted components all shared a constant frequency separation (spacing) that was equal to that of the components in the harmonic set. Amplitude modulation with a period that is the reciprocal of the frequency spacing of the components is only found in the outputs of higher frequency auditory filters that do not resolve individual harmonics. This information is evidently used in combination with information from the resolved components (even though it is not clear how the nervous system combines these kinds of information). It is possible that, for the component differences used in this study, across-vowel component misalignment information from the resolved region could have contributed to the segregation of both harmonic and shifted components without the need to estimate explicitly the F0s of the two constituents (Summerfield & Assman, 1991).

Performance in both the harmonic and shifted sets was better (at ½ and 6 semitones) than in the random set, where the components were not only inharmonic but did not share a common spacing. The presence of a (relatively small) F0 difference effect in the random set is interesting. What does an F0 separation mean in this case? As mentioned earlier, each vowel was made up of inharmonic partials. In a vowel pair, any two corresponding components between the vowels (e.g., the second, etc.) would be a given distance apart (½ or 6 semitones). However, within each vowel, there was no apparent cue that would contribute to the grouping of its components (such as the cues available in the harmonic and shifted sets). The change from 0 to ½ semitone seems to account for the whole F0 separation effect (about a 4% change in performance) in both the SSs and the PGs. Perhaps this was due to the fact that there are only half as many partials in the 0-semitone condition (i.e., 40 partials total instead of 80) compared with the others. Therefore, the doubling of the number of partials at ½ semitone may contribute to the improved definition of the spectrum of the mixture of the two vowels and to their subsequent recognition. This doubling of partials is also true for the harmonic and shifted sets, where the effect is even stronger. It may be the case that the large change between zero and the lowest nonzero separation that is found in all experiments on F0 separation of vowels (about 10% change in this paper; 30% in Chalikia & Bregman, 1989) may be partly due to this doubling of the number of partials. Benefits from this doubling of partials are not found when the difference in F0 is as small as ¼ semitone (e.g., Scheffers, 1983), presumably because, in that case, there are strong interactions between corresponding harmonics in the two vowels, which probably distort the overall spectrum envelope in the region of F1. If this notion is plausible, one could predict an improvement in performance, even when both constituents have the same F0, simply by halving the F0 (Q. Summerfield, personal communication). Summerfield and Assman (1991) have found that the accuracy of identifying both constituents of double vowels rose by about 20%, comparing the case where both F0s were 200 Hz and the case where both were 100 Hz. Reducing the F0s to 50 Hz, however, did not lead to any further improvement, so the benefits may be limited to F0s above 100 Hz.

It is clear that the use of PGs and CGs had an effect that was additional to the effect of F0 separation, since in both cases an effect was produced that was greater than that of SSs at 6 semitones. Similar results were observed in Chalikia and Bregman (1989). In that study, for SSs, performance remained steady after a ½-semitone separation and did not improve by the use of larger F0 separations. However, with CGs, there was a further improvement in performance (of about 10%) between ½ semitone and 3 semitones that remained at that level for larger F0 separations. The present findings differ, in some respects, from those of the earlier study.

One of the surprising observations was that the results in the harmonic set did not replicate previous findings (Chalikia & Bregman, 1989) regarding the glide effect. In the present study, the only effect was that of F0 differences. In the 1989 study, there was a general advantage of CGs over SSs for differences of 3, 6, and 12 semitones and an advantage of CGs over PGs at the octave separation. McAdams (1989), using a mixture of three vowels frequency modulated (coherently and incoherently) with F0s 5 semitones apart, had found that the modulation, in general, made the vowels more prominent. No additional effects were found by comparing coherent and incoherent modulation on judged prominence. One would expect incoherent FM to contribute to vowel segregation. In coherent gliding (e.g., of a vowel's harmonics), the components retain the same simple frequency-ratio relationships to one another at every instant in time and thus are expected to be grouped together. When the gliding is incoherent (e.g., as in the case of two vowels with crossed pitch contours), only the components within each vowel retain their ratio relations. The components across two vowels are decorrelated, and the vowels are expected to segregate. Gardner et al. (1989) have pointed out that the absence of the effects of incoherence of frequency modulation between McAdams's vowels could be due to the fact that the vowels also had large F0 differences (5 semitones). It is possible that the frequency-modulation effect cannot exert an independent influence over a maximum F0 effect. However, this explanation cannot account for the fact that there was a difference in perceived prominence between unmodulated and modulated vowels in the McAdams study. However, since perceived prominence may not be the same as vowel identification, such differences in procedure among studies may make the comparison of results difficult. It is not clear why our 1989 study showed a glide effect over and above that of F0 differences for the harmonic vowels.

On the other hand, the pitch-contour effect is evident in the shifted set and seems to be adding to grouping over and above the effect of F0 differences. With both kinds of glides, the 6-semitone results are better than the ½- and 0-semitone results. Also, in the 6-semitone case, CGs are better than PGs and PGs are better than SSs. CGs are better than PGs because of the decorrelation of the components across the two vowels, as stated above. It apparently becomes more helpful to unify a spectrum by giving it a distinct pattern of movement when it is not already being unified by the property of harmonicity. Grouping, based on common fate, accounts better for the results than does an envelope-tracing explanation. However, since PGs are better than SSs, some other cue must contribute to the untangling of the constituents. Common fate should fuse the two vowels *together* (since the components across vowels retain their ratio relations as well). Perhaps parallel gliding may contribute, under some conditions, to envelope tracing, and hence to recognition, rather than to segregation per se. It is also possible that frequencies that are changing are a better stimulus to the auditory processes that detect formant peaks for some reason other than formant tracing. In general, changing signals act as better stimuli for sensory systems. Common fate may contribute to segregation only when there are separate fates, that is, more than one subset of motions. This may occur because the default state of the auditory-scene analysis system is fusion. Fusion occurs unless there is specific evidence for segregation.

The pitch-contour effect was also evident in the random set. There, at a 6-semitone F0 separation, CGs were better than PGs and SSs. However, PGs were not significantly better than SSs. These findings seem to suggest that, as in the shifted set, incoherent (contrary-direction) motion of two sets of components is the necessary cue for segregation.

The major conclusions that can be drawn from the present findings, then, are the following:

1. A difference in F0 is a reliable cue that aids the identification of the constituents in a double vowel. Its effect may operate largely by increasing the number of components in the stimulus.

2. The existence of two different rates of amplitude modulation can help in segregation in vowel pairs formed of shifted harmonics.

3. Common fate can play a role, but its effects are shown only when harmonicity is reduced or absent.

4. Tracing does not play a consistent role, but may contribute to the recognition of the constituents under some conditions.

REFERENCES

Assman, P. F., & Summerfield, Q. (1989). Modeling the perception of concurrent vowels: Vowels with the same F0. *Journal of the Acoustical Society of America*, **85**, 327-338.

Assman, P. F., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, **88**, 680-697.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: Bradford.

Bregman, A. S., & Doehring, P. (1984). Fusion of simultaneous tonal glides: The role of parallelness and simple frequency relations. *Perception & Psychophysics*, **36**, 251-256.

Bregman, A. S., Levitan, R., & Liao, C. (1990). Fusion of auditory components: Effects of the frequency of amplitude modulation. *Perception & Psychophysics*, **47**, 68-73.

Broadbent, D. E. (1952). Failures of attention in selective listening. *Journal of Experimental Psychology*, **44**, 428-433.

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics*, **10**, 23-26.

Carlson, R., Granstrom, B., & Fant, G. (1970). Some studies concerning the perception of isolated vowels (Speech Transmission Laboratory Q: Prog. Status Rep. STL-QPSR 2- 3/1970, pp. 19-35). Stockholm, Sweden: Royal Institute of Technology.

Chalikia, M. H., & Bregman, A. S. (1989). The perceptual segregation of simultaneous auditory signals: Pulse train segregation and vowel segregation. *Perception & Psychophysics*, **46**, 487-496.

Cherry, E. C. (1953). Some experiments on the recognition of speech with one and two ears. *Journal of the Acoustical Society of America*, **25**, 975-979.

Chistovich, L. (1971). Auditory processing of speech-evidences from psychoacoustics and neurophysiology. In *Proceedings of the 7th International Congress on Acoustics* (Vol. 1, pp. 27-42). Budapest.

Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Quarterly Journal of Experimental Psychology*, **33A**, 185-208.

Darwin, C. J., & Culling, J. F. (1990). Speech perception seen through the ear. *Speech Communication*, **9**, 469-475.

Darwin, C. J., & Gardner, R. B. (1986). Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality. *Journal of the Acoustical Society of America*, **79**, 838-845.

Gardner, R. B., & Darwin, C. J. (1986). Grouping of vowel harmonics by frequency modulation: Absence of effects on phonemic categorization. *Perception & Psychophysics*, **40**, 183-187.

Gardner, R. B., Gaskill, S. A., & Darwin, C. J. (1989). Perceptual grouping of formants with static and dynamic differences in fundamental frequency. *Journal of the Acoustical Society of America*, **85**, 1329-1337.

Halikia, M. H., & Bregman, A. S. (1984a) Perceptual segregation of simultaneous vowels presented as steady states and as glides. *Canadian Psychology*, **25**, 210.

Halikia, M. H., & Bregman, A. S. (1984b). Perceptual segregation of simultaneous vowels presented as steady states and as parallel and crossing glides. *Journal of the Acoustical Society of America*, **75**, S83.

Henke, W. L. (1987). *MITSYN: A coherent family of command level utilities for time signal processing* [Computer program]. Available from W. L. Henke, 133 Bright Street, Belmont, MA 02178.

Joos, M. (1948). Acoustic phonetics. *Language*, 24 (Suppl.), 1-126.

Klatt, D. H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. In R. Cole (Ed.), *Perception and production of fluent speech* (pp. 243-288). Hillsdale, NJ: Erlbaum.

Marin, C. M. H., & McAdams, S. (1991). Segregation of concurrent sounds: II. Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width. *Journal of the Acoustical Society of America*, **89**, 341-351.

McAdams, S. (1984). *Spectral fusion, spectral parsing and the formation of auditory images*. Unpublished doctoral dissertation, Stanford University.

McAdams, S. (1989). Segregation of concurrent sounds: I. Effects of frequency modulation coherence. *Journal of the Acoustical Society of America*, **86**, 2148-2159.

McAdams, S., & Rodet, X. (1988). The role of FM-induced AM in dynamic spectral profile analysis. In H. Duifhuis, J. Horst, & H. Wit (Eds.), *Basic issues in hearing* (pp. 359-369). London: Academic Press.

Meddis, R., & Hewitt, M. J. (1992). Modeling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, **91**, 233-245.

Moore, B. C. J. (1987). The perception of inharmonic complex tones. In W. A. Yost & C. S. Watson (Eds.), *Auditory processing of complex sounds* (pp. 180-189). Hillsdale, NJ: Erlbaum.

Moore, B. C. J., Glasberg, B. R., & Peters, R. W. (1985). Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America*, **77**, 1853-1860.

Moore, B. C. J., Glasberg, B. R., & Peters, R. W. (1986). Thresholds for hearing mistuned partials as separate tones in harmonic complexes. *Journal of the Acoustical Society of America*, **24**, 175-184.

Moore, B. C. J., Peters, R. W., & Glasberg, B. R. (1985). Thresholds for the detection of inharmonicity in complex tones. *Journal of the Acoustical Society of America*, **77**, 1861-1867.

Palmer, A. R. (1990). The representation of the spectra and fundamental frequencies of steady-state single- and double-vowel sounds in the temporal discharge patterns of guinea pig cochlear-nerve fibers. *Journal of the Acoustical Society of America*, **88**, 1412-1426.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.

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

Summerfield, Q., & Assman, P. F. (1991). Perception of concurrent vowels: Effects of harmonic misalignment and pitch-pulse asynchrony. *Journal of the Acoustical Society of America*, **89**, 1364-1377.

Weintraub, M. (1985). A theory and computational model of monaural auditory sound separation. Unpublished doctoral dissertation, Stanford University.

Weintraub, M. (1987). Sound separation and auditory perceptual organization. In M. E. H. Schouten (Ed.), *The psychophysics of speech perception* (NATO ASI Series, pp. 125-134). Dortrecht: Nijhoff.

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

(Manuscript received April 23, 1990;
revision accepted for publication August 13, 1992.)