Segregation of concurrent sounds. I: Effects of frequency modulation coherence[a)]

Stephen McAdams[b)]
Institut de Recherche et Coordination Acoustique/Musique (IRCAM), 31, rue Saint-Merri, F-75004 Paris, France

(Received 30 March 1989; accepted for publication 13 July 1989)

Frequency modulation coherence was investigated as a possible cue for the perceptual segregation of concurrent sound sources. Synthesized chords of 2-s duration and comprising six permutations of three sung vowels (/a/, /i/, /o/) at three fundamental frequencies (130.8, 174.6, and 233.1 Hz) were constructed. In one condition, no vowels were modulated, and, in a second, all three were modulated coherently such that the ratio relations among all frequency components were maintained. In a third group of conditions, one vowel was modulated, while the other two remained steady. In a fourth group, one vowel was modulated independently of the two other vowels, which were modulated coherently with one another. Subjects were asked to judge the perceived prominence of each of the three vowels in each chord. Judged prominence increased significantly when the target vowel was modulated compared to when it was not, with the greatest increase being found for higher fundamental frequencies. The increase in prominence with modulation was unaffected by whether the target was modulated coherently or not with nontarget vowels. The modulation and pitch position of nontarget vowels had no effect on target vowel prominence. These results are discussed in terms of possible concurrent auditory grouping principles.

PACS numbers: 43.66.Mk, 43.66.Lj, 43.71.Es, 43.66.Jh[WAY]

INTRODUCTION

An important experimental problem for hearing science is to determine the acoustic cues used in perceptual segregation of different sound sources. A situation often confronting the human auditory system is to listen to several sound sources at once and to try to extract the meaning of one of them. As it performs its limited frequency and temporal analysis, the system must decide which components belong to which source. Any two frequency components may or may not derive from the same source. Decoding a speech signal in the presence of other signals, for example, involves selecting among the components that are present and grouping some of them to define a voice. If all of the sounds present in the environment were to fuse together into a whole, the qualities of the original vocal source would be lost. If they could be perceptually segregated, however, the vocal source could then be independently recognized and understood. Two questions are implied here: (1) What cues are used by the auditory system to segregate concurrent sound sources, and (2) what is the relation between grouping processes and the derivation of source qualities such as vowel timbre? This article will consider this relation and some of the cues that are candidates for concurrent grouping, focusing experimentally on frequency modulation coherence.

A. Perceptual grouping

Bregman (1978) has proposed that sound qualities (such as musical timbre and vowel timbre) are group properties that emerge only when the appropriate components are perceptually fused together. This hypothesis is supported by evidence for auditory grouping mechanisms that evaluate onset and offset synchrony, harmonicity, or spatial location (Darwin and Bethell-Fox, 1977; Bregman and Pinker, 1978; Dannenbring and Bregman, 1978; Darwin, 1979, 1981, 1983, 1984; Darwin and Sutherland, 1984; McAdams, 1984a). For example, a mechanism that restricts the frequencies contributing to a certain virtual pitch to those that have harmonic relations appears to have a frequency deviation tolerance on the order of 3%–4% (cf. Duifhuis *et al.*, 1982; Scheffers, 1983b; Moore *et al.*, 1985a; Hartmann *et al.*, 1986). When fusion occurs in these cases, it seems that the acoustic elements tend to lose their individual identities as such and give rise to an emergent group quality. Although Moore *et al.* (1985b) have suggested that a mistuned, separately heard component may still affect virtual pitch, Moore *et al.* (1985a) found that the contribution of the component to the virtual pitch begins to decline rapidly as its frequency deviates progressively from harmonicity by more than 3%. At 8% its contribution is negligible.

Differences in fundamental frequency (F_0) of two simultaneously presented sounds have also been shown to enhance segregation and identification. Stumpf (1890) claimed that two harmonic sounds tend to fuse when close in pitch and lose their characteristic qualities, although they may be perceptually segregated and recognized individually

[a)] Portions of this article were presented at the 104th Meeting of the Acoustical Society of America in Orlando, FL [J. Acoust. Soc. Am. Suppl. 1 72, S11 (1982)] and in the author's unpublished Ph.D. disseration (Chap. 5, 1984b).

[b)] Present address: Laboratoire de Psychologie Expérimentale (CNRS URA 316), 28 rue Serpente, F-75006 Paris, France.

 0001-4966/89/122148-12$00.80

when different in pitch. Scheffers (1983a) found that source separability (judged by vowel identification) improved up *to a difference of 6% in* F_0, but did not improve beyond that. Brokx and Nooteboom (1982) found improvement (judged by errors in reproduction of vocal utterances) up to an 18% difference. Darwin (1981), Scheffers (1983a), and Weintraub (1985) have proposed that pitch (or harmonicity) may be used in these cases. According to Scheffers, the listener can apparently group those formants within which the harmonics belong to the same F_0 or decide that separate formants with harmonics not related to the same F_0 belong to another vowel.

Other research has found that changes in the onset synchrony or static mistuning of a partial can have a strong effect on perceived phoneme boundaries, this effect being correlated with perceptual grouping (Darwin, 1981; Darwin and Sutherland, 1984; Darwin and Gardner, 1986). In the context of the present study, another possible cue for source segregation will be considered: frequency modulation coherence among components.

B. Frequency modulation coherence

All natural, sustained-vibration sounds contain small-bandwidth random fluctuations in the frequencies of their components. These have been found for voice (Lieberman, 1961; Flanagan, 1972; Kersta *et al.*, 1960) and musical instruments [Cardozo and van Noorden, 1968; Grey and Moorer, 1977; MacIntyre *et al.*, 1981, 1982; McAdams, 1984b (Appendix B)].

There has not been much research directed toward determining the relative coherence of modulations among partials. However, studies by Charbonneau (1981) have shown that, when the slightly different modulation functions of partials in instrument tones are all replaced by the same modulation function (e.g., the one extracted from the fundamental frequency), the difference is undetectable by human listeners. One may conclude that the auditory system is insensitive to the small degree of incoherence that *is* present in the natural jitter of harmonics in instrument tones.

Brokx and Nooteboom (1982) found that subjects were better able to understand and reproduce a speech stream heard in the presence of another, competing speech stream when real monotone voices were used than when the voices were resynthesized with perfectly steady frequencies. They found no difference in performance between reproductions of real voices spoken either in monotone fashion or with normal intonation. It is possible that, among other dynamic factors, the natural jitter present in these voices aids in fusing the image and in distinguishing it from the competing speech stream.

Preliminary work [McAdams 1982, 1984b (Chaps. 2–3, Appendix F)] has suggested that applying different modulation waveforms coherently to separate subgroups of components that are embedded in a complex spectrum results in changes in the number of reported sources and in the noticeable pitches and timbres present in a complex tone. "Coherence" is achieved by maintaining constant frequency ratios among the frequency components. Both harmonic and inharmonic stimuli are perceived as more fused when subjected to coherent modulation, which provides evidence that ratio-preserving FM may be one of those "circumstances which assist us first in separating the musical tones arising from different sources, and secondly, *in keeping together* [i.e., fusing into a single unified image] *the partial tones of each separate source*" [italics added] (Helmholtz 1877/1885, p. 59).

There is additional evidence of the contribution of frequency modulation incoherence to source separation. Rasch (1978) presented listeners with two simultaneous harmonic complexes with different F_0's. The level of the higher complex was adjusted to determine the threshold at which it was masked by the lower complex. The lower complex was never modulated. When the higher complex had a 5-Hz, 4% vibrato imposed on it, its masked threshold was 17.5 dB *below* the threshold obtained when it was not modulated. Thus the difference in frequency modulation behavior helped separate the two sources. This type of paradigm most likely touches upon aspects of source separation that are more closely tied to peripheral auditory mechanisms than to more central organization processes but indicates that dynamic stimuli are less easily masked than steady ones.

The research cited above suggests that frequency modulation coherence may play a role in the perceptual segregation of sound sources. To test this hypothesis, stimuli were constructed with three simultaneous sung vowels at different pitches in a chord. In some cases, the frequency components of single vowels were frequency modulated against a background (made up of the other two vowels) that was itself either steady or modulated. In other cases, either no vowels were modulated or all vowels were modulated coherently (i.e., with identical modulation waveforms maintaining ratio relations among all frequencies). The separation of the F_0's of the sources as well as their harmonic coincidence (spectral overlap) was held constant. Subjects were to judge the prominence of each vowel in the complex under these various conditions. One would expect that, when no vowels were modulated, it would be difficult to separate them and that the judged prominence would be low. In this condition the effect of masking and spectral overlap between the vowels would be the limiting factor in source identification. Similarly, when all vowels were modulated coherently in frequency, it should also be difficult to separate them. Finally, one would expect that, when a vowel was modulated independently of the others, its components would be more easily separated from the background into a distinct source image which would subsequently be judged as more prominent.

I. PRETEST

Before sense could be made out of judgments of the prominence of synthetic vowels embedded in a complex spectrum, it had to be ascertained that the component vowels were identifiable in isolation. This pretest also allowed subjects to have prior experience with the synthetic stimuli that were to be identified under more difficult circumstances later.

A. Stimuli

Stimuli were synthesized at a sampling rate of 16 kHz in 32-bit floating point format on a DEC10 mainframe computer and then stored in 16-bit integer format. They were transferred to a PDP 11/34 minicomputer with Tim Orr 16-bit DACs and an 8-kHz, −96-dB/oct, low-pass filter for the experiment. Tones were 2 s in duration with 150-ms linear attack and decay ramps. The three vowels /a/,/i/, and /o/ were used. Since the subjects were to be drawn from a multilingual pool, it was felt that these vowels were the closest to being common across languages. They are also quite common in the Western classical singing repetoire. In addition, these vowels are well separated in the classical "vowel space" that plots first-formant frequency (related to the closedness or openness of the mouth) versus second-formant frequency (related to the position of the tongue controlling the size of the mouth cavity).[1] The vowels were derived from a male singing voice and synthesized by the computer program CHANT according to a time-domain formant-wavefunction synthesis algorithm developed by Rodet (1980). This synthesis method sums waveforms that are equivalent to formant impulse responses and is thus closer to parallel than to serial formant synthesizers. The technique is, in effect, a precursor to what is currently referred to as wavelet synthesis. The formant frequencies, bandwidths, and relative amplitudes are shown in Table I. Each vowel was synthesized at three pitches (fundamental frequencies)—C_3 (130.8 Hz), F_3(174.6 Hz), and Bb_3(233.1 Hz)—without modifying formant frequencies as a function of pitch.

Each of the stimuli was synthesized both with and without subaudio frequency modulation. Due to the synthesis algorithm, any modulation in frequency was coupled to a modulation in amplitude of the frequency components such that a constant resonance structure was maintained throughout. The modulation waveform was composed of both vibrato (periodic) and jitter (aperiodic) components. The vibrato component was sinusoidal with a frequency of 5.1 or 6.3 Hz and an amplitude yielding a peak-to-peak frequency excursion of 3% of a given partial's frequency (2.1% total rms excursion). The jitter component had a total rms excursion of 1.6% of the center frequency and had a spectral content composed of a 30-Hz low-pass noise band plus a 150-Hz low-pass noise band at 40 dB below the level of the 30-Hz band. The jitter was chosen to resemble normal human vocal jitter in trained professional singers (Rodet, 1982). The compound modulation (vibrato plus jitter) was scaled over time, beginning with no modulation for the first 300 ms, followed by a linear growth to maximum modulation width at 700 ms, and a constant modulation width for the remainder of the stimulus duration. In preliminary explorations, the scaling over time of the vibrato was found to enhance the effect of vibrato on vowel prominence.

The levels of the stimuli were adjusted for equal loudness by the experimenter. The presence of modulation had very little effect on the perceived loudness,[2] modulated vowels were attenuated 0.4 dB on the average in relation to the unmodulated vowels. However, adjustments in level on the order of 2 dB were sometimes necessary to equalize the loudness at different pitches of a given vowel. On the average (across pitches), /a/ stimuli were attenuated 3.8 dB and /i/ stimuli 4.4 dB relative to /o/ stimuli (see Table I). The /o/ stimuli were presented at an average sound pressure level of 75 dBA at the earphone as determined with a flat-plate coupler connected to a Bruel & Kjaer 2209 sound level meter.

TABLE I. Parameters for vowel synthesis with the singing voice synthesis program CHANT. The level value in parentheses is the measured formant peak level in the synthesized vowels.

Formant frequency (Hz)	Bandwidth (Hz)	Level (dB *re*: F_1)
Vowel /a/ (av. rms level = −3.8 dB *re*: /o/)		
600	78	0.0 (0.0)
1050	88	−6.2 (−8.1)
2400	123	−12.0 (−18.1)
2700	128	−11.0 (−17.7)
3100	138	−23.8 (−34.3)
Vowel /i/ (av. rms level = −4.4 dB *re*: /o/)		
238	73	0.0 (0.0)
1741	108	−19.6 (−23.5)
2450	123	−16.5 (−24.2)
2900	132	−19.6 (−26.3)
4000	150	−31.7 (−39.1)
Vowel /o/		
360	51	0.0 (0.0)
750	61	−11.5 (−13.5)
2400	168	−29.3 (−43.6)
2675	184	−26.4 (−42.4)
2950	198	−35.4 (−54.3)

B. Method and results

Stimuli were presented diotically, routed from the DACs through a Neeve professional mixing console to a Revox A740 stereo power amplifier, and then to AKG K242 earphones. The experiment took place in an acoustically treated sound studio. Each of the 18 stimuli (three vowels × three pitches × two modulation conditions: with and without) was presented five times in a randomized block to each subject. The subject's task was to identify the sound as /a/, /i/, or /o/ and to flip one of the three switches accordingly. All subjects identified all vowels with perfect accuracy regardless of pitch and presence or absence of modulation. Some subjects felt the sounds labeled as /o/ were closer to /u/. They were told that these sounds were to be considered as /o/ in the main experiment.

II. MAIN EXPERIMENT

A. Stimuli

The stimuli from the pretest were combined by a digital mixing program (in 32-bit floating point format) to form chords of three different simultaneous vowels at the three different pitches; the chord always consisted of the pitches C_3, F_3, and Bb_3 with some permutation of the three vowels /a/, /i/, and /o/. Six permutations result as illustrated in Fig. 1.

These permutations were included to test the effects of fundamental frequency on each vowel's perceptual promi-

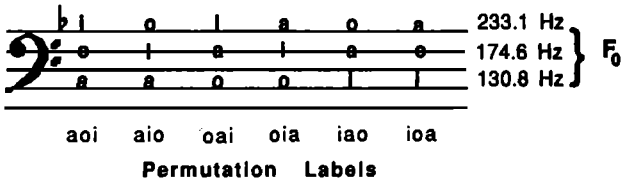

FIG. 1. Vowel by pitch permutations.

nence: When unmodulated, the spectral form of the vowel is less well defined at higher F_0's since the formants are not as well filled out by the components. They were also designed to examine the potential effects of masking among the different vowels. The pitch interval of a perfect fourth (ratio 4:3) was chosen in order to achieve a compromise between minimizing the coincidence of partials among the simultaneous sources and minimizing the dissonance and roughness due to first- and second-order beats (cf. Plomp, 1976). This interval (five semitones) is also beyond the one-to-three semitone pitch separation range found to be a limiting factor in voice separation by Brokx and Nooteboom (1982) and Scheffers (1983a).

The interest here was in the salience or prominence of the vowel sounds under various conditions of (1) modulation of a given vowel (the "figure") and (2) modulation of the two other vowels (the "ground"). A figure condition is defined as the imposing of an independent modulation function on the frequency components of a given vowel. Four conditions were used: /a/ modulated independently (labeled *Afig*), /o/ modulated independently (*Ofig*), /i/ modulated independently (*Ifig*), and no vowel modulated independently of the others (*Nofig*). The "ground" comprised the vowels that are not chosen as figure. Two ground conditions were used with each figure condition: ground unmodulated or steady (*Gsteady*), and ground modulated or given vibrato and jitter (*Gmod*). In the *Gmod* condition, the ground vowels were modulated identically. All of the modulation possibilities for a permuted chord are listed in Table II. In each cell, the modulation specifications for each vowel (regardless of pitch position) are shown.

Mod1 and Mod2 in Table II represent two independent modulation functions. Mod1 was as described for the pretest with a 5.1-Hz, 2.1% rms vibrato and a 1.6% rms jitter. Mod2 had a 6.3-Hz, 2.1% rms vibrato and an independent 1.6% rms jitter. The two jitter waveforms had similar statistical characteristics (spectrum and amplitude probability density function), but were uncorrelated. With ground = *Gsteady* and figure = *Nofig*, no vowels were modulated. For the other *Gsteady* conditions, only one vowel was modulated with Mod1. In the *Nofig/Gmod* conditions, all vowels were modulated coherently with Mod2, thus maintaining the frequency ratios among all of the partials in the complex. In the other *Gmod* conditions, the vowel chosen as figure was modulated with Mod1 independently of the other two, which were modulated with Mod2. The two ground conditions were included to investigate the following: (1) comparison between conditions where no vowels were modulated and where all were modulated coherently (*Nofig/Gsteady* vs *Nofig/Gmod*): Is a target source perceived differently in a coherently moving ground (*Gmod*) compared to a nonmoving ground (*Gsteady*)? (2) comparisons between conditions where a vowel figure was modulated against a steady ground and those where it was modulated independently of a moving ground (*Afig/Gsteady*, *Ifig/Gsteady*, *Ofig/Gsteady* vs *Afig/Gmod*, *Ifig/Gmod*, *Ofig/Gmod*, respectively): Does the modulation state of the ground (*Gsteady* or *Gmod*) affect the prominence of the figure?

One would expect that it would be difficult to hear out any vowels that are spectrally obscured in conditions *Nofig/Gsteady* and *Nofig/Gmod*, and that judgments of salience or prominence of these vowels would be low. Any difference between these two conditions would most likely be attributable to the reduction in ambiguity of the spectral forms provided by the coupled amplitude and frequency modulations. For conditions *Afig*, *Ofig*, or *Ifig*, a significant increase in the salience judgment for that particular vowel was expected, whereas less increase was expected for the two vowels that made up the ground.

B. Method

Stimuli were presented at approximately 75 dBA in the same situation described for the pretest. Ten subjects with four different native tongues (English, French, Finnish, and Rumanian) were run in the experiment and were paid for their participation. All subjects were fluent in at least two languages with English as either a first or second language. Experimental instructions were given in either French (five

TABLE II. Modulation state of the vowels under different figure–ground combinations. Mod1 is composed of a 5.1 Hz, 2.1% rms vibrato and a 1.6% rms jitter. Mod2 is composed of a 6.3 Hz, 2.1% rms vibrato and an independent 1.6% rms jitter.

		Vowel modulated independently (figure)			
		Nofig	*Afig*	*Ifig*	*Ofig*
Modulation of other vowels (ground)	*Gsteady*	/a/ = none /i/ = none /o/ = none	/a/ = Mod1 /i/ = none /o/ = none	/a/ = none /i/ = Mod1 /o/ = none	/a/ = none /i/ = none /o/ = Mod1
	Gmod	/a/ = Mod2 /i/ = Mod2 /o/ = Mod2	/a/ = Mod1 /i/ = Mod2 /o/ = Mod2	/a/ = Mod2 /i/ = Mod1 /o/ = Mod2	/a/ = Mod2 /i/ = Mod2 /o/ = Mod1

subjects) or English (five subjects) according to the wish of the subject. Five subjects were highly trained musicians and five subjects reported having no formal training in music, though one considered himself an accomplished amateur pianist. All subjects reported having no hearing problems. Each of the 48 stimuli (six permutations × four figure conditions × two ground conditions) was presented five times in a randomized block design. Each diotic stimulus was presented once before any stimulus was repeated.

A trial consisted of a single chord presented repeatedly. The subjects were allowed to listen to the chord as long as necessary to make the judgments. The subject was informed that a complex tone would be heard with three pitches in a chord and that any or all or none of the sounds at these pitches might be the vowels /a/, /i/, or /o/ as heard in the pretest. The task was to adjust a sliding potentiometer to indicate on a linear scale the degree of salience or prominence of a given vowel, or the certainty that the given vowel was present. It was hypothesized that perceptual salience or prominence would increase if stimulus parameters, such as independent modulation, favored segregation of the vowel. For each stimulus, three judgments were made on three separate potentiometers—one each for /a/, /i/, and /o/. The top position indicated that the vowel was "very prominent"

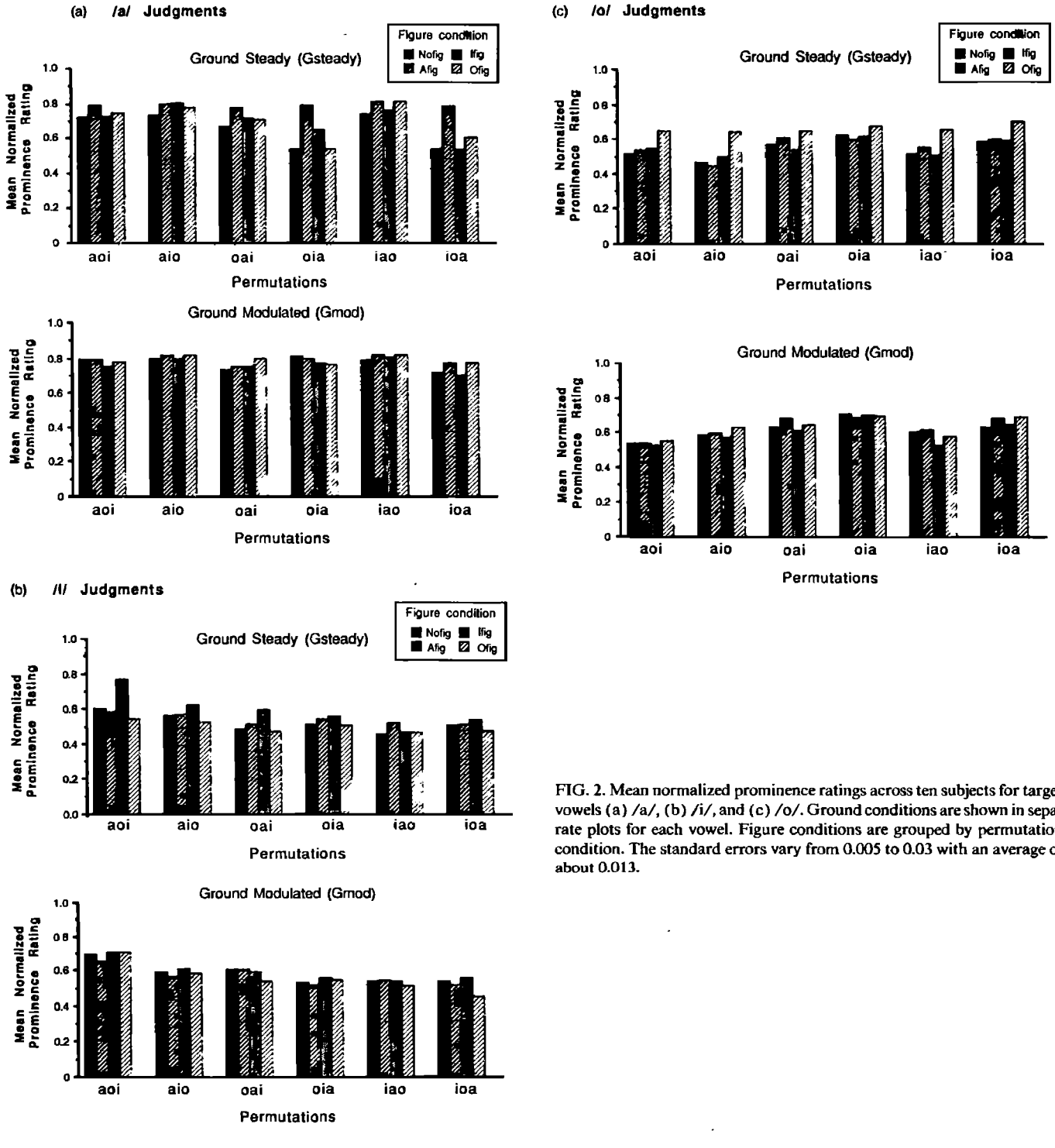

FIG. 2. Mean normalized prominence ratings across ten subjects for target vowels (a) /a/, (b) /i/, and (c) /o/. Ground conditions are shown in separate plots for each vowel. Figure conditions are grouped by permutation condition. The standard errors vary from 0.005 to 0.03 with an average of about 0.013.

or that the subject was "perfectly certain" that the vowel was present. This position was coded with a value of 1.0 in the *data. The bottom* position, encoded as 0.0, indicated "not at all prominent," or "perfectly certain" that the vowel was not present. The subjects were advised to use the following strategy in order to make the judgments quickly: (a) Focus on one vowel at a time and try to hear out that vowel at the different pitches (the clarity of these pitches depended strongly on the modulation context); (2) judge that vowel's prominence; and then (3) focus on the next vowel, and so on. This procedure ensured that the subject was listening for, and trying to hear, the vowel currently being judged. Once all three judgments were made, a switch was closed by the subject. At this point, the stimulus presentation ceased and the positions of the sliders were registered. The same switch was opened again for presentation of the next trial. The experimental session lasted 100–150 min depending on the self-pacing of the subject. Subjects were allowed to take breaks as they wished between trials in order to avoid fatigue and flagging concentration.

C. Results

The values of the five judgments for each stimulus were averaged for each subject and these mean prominence ratings were used as data for further analysis. The intrasubject variability, while relatively high, is consistent across subjects. The average standard deviation is about 0.18 and the average standard error is about 0.013. However, subjects used the range of the potentiometers quite differently: The range of values used for a given vowel by a given subject varied from 0.11 to 1.0. Accordingly, the data for each subject were normalized with respect to the mean and standard deviation over all of their judgments, i.e., over /a/, /o/, and /i/ judgments for all conditions. (This transformation assumes, of course, that all subjects were using a linear scale.) Then, the individual normalized data were scaled and translated so that the data in their final form for all subjects fell between values of 0.0 and 1.0. This tranform operation preserves the pattern of ratio relations among data for any given subject, but reduces the standard deviations within cells across subjects by approximately a factor of 2. It also allows more sensitive comparisons across conditions to be made and allows patterns in the data common to all subjects to emerge.

The means of the normalized data are plotted in Fig. 2.[3] It is important to remember in examining and interpreting these results that the data represent judgments made on a complex sound with all three vowels present while focusing on a single vowel at a time (the "target" vowel).

1. Analysis of variance

To test the main factors of the experimental design, three-way analyses of variance (figure × permutation × ground) were performed on the normalized data for each target vowel separately. The results are listed in Table III. The main effects of figure modulation state (***Nofig, Afig, Ifig, Ofig***), permutation (***aoi, aio, oai, oia, iao, ioa***), and

TABLE III. ANOVA tables for three-factor analyses of variance on /a/, /i/, and /o/ judgments.

Source	*df*	Sum of squares	Mean square	*F* ratio	*p* value
/a/ judgments					
Figure mod state (F)	3	0.406	0.135	16.107	0.0001
Permutation (P)	5	0.899	0.180	21.393	0.0001
F×P	15	0.303	0.020	2.401	0.0024
Ground mod state (G)	1	0.567	0.567	67.462	0.0001
F×G	3	0.237	0.079	9.394	0.0001
P×G	5	0.340	0.068	8.088	0.0001
F×P×G	15	0.184	0.012	1.459	n.s.
Error	432	3.629	0.008		
/i/ judgments					
Figure mod state (F)	3	0.264	0.088	8.737	0.0001
Permutation (P)	5	1.267	0.253	25.171	0.0001
F×P	15	0.154	0.010	1.022	n.s.
Ground mod state (G)	1	0.166	0.166	16.447	0.0001
F×G	3	0.079	0.026	2.616	0.0506
P×G	5	0.080	0.016	1.590	n.s.
F×P×G	15	0.147	0.010	0.976	n.s.
Error	432	4.348	0.010		
/o/ judgments					
Figure mod state (F)	3	0.398	0.133	12.683	0.0001
Permutation (P)	5	0.892	0.178	17.035	0.0001
F×P	15	0.152	0.010	0.967	n.s.
Ground mod state (G)	1	0.202	0.202	19.244	0.0001
F×G	3	0.219	0.073	6.957	0.0001
P×G	5	0.136	0.027	2.594	0.0251
F×P×G	15	0.035	0.002	0.221	n.s.
Error	432	4.524	0.010		

ground modulation state (*Gsteady, Gmod*) are highly significant for all three vowels. For figure modulation, prominence ratings tend to increase when the target vowel is the figure compared to when it is not, but this is only true in *Gsteady* conditions. Prominence ratings are generally higher for a given target vowel in all *Gmod* conditions, though this is not the case in comparisons between *Gsteady* and *Gmod* conditions where the target vowel is the figure. Such conditions tend to have similar prominence ratings. The two effects just mentioned give rise to the significant ground main effect and the figure × ground interactions for all three target vowels which suggest that prominence ratings increase whenever the target is modulated, whether it is in the figure or in the ground. As for the permutation main effect, the differences among the chord permutations are similar for /a/ and /o/, but quite different for /i/. For both of the former vowels, there is a slight tendency for decreased prominence ratings at higher F_0's of the target vowel than at lower ones. For /i/ judgments, the reverse tendency is seen. These trends are complicated in all cases by some unsystematic effects of the permutation of the nontarget vowels.

Although the three-way interaction was not significant for any of the three vowels, two-way interaction effects varied considerably. All three two-way interactions were significant for /a/ judgments, the figure × ground interaction was just barely significant for /i/, and the figure × ground and permutation × ground interactions were significant for /o/. Most of these effects seem to be primarily due to two factors: (1) Prominence judgments increase independently of whether the target vowel is the figure or part of the ground, and (2) prominence judgments change with the pitch of the target vowel in a way that interacts with its modulation state. These results thus suggest a regrouping of the data by modulation state and pitch of the target vowel. Before performing this regrouping, however, it should be ascertained for the groups of conditions which are to be regrouped that (1) there exist no differences among the conditions with the same modulation state, and (2) there exist no differences among different nontarget vowel permutations.

2. Comparisons among conditions with the same modulation state of the target vowel

In examining the data collapsed across permutations for each vowel (Fig. 3), it appears that increased prominence due to modulation of the target vowel is independent of whether the vowel is the "figure" or part of the "ground." Vowel prominence judgments are similar among figure–ground combinations in which the target vowel is not modulated (e.g., among *Nofig, Ifig,* and *Ofig* under *Gsteady* conditions for target vowel /a/). Judgments are also similar among combinations where the target vowel is modulated, regardless of whether it is figure or part of the ground (e.g., compare among all *Gmod* conditions and *Afig/Gsteady* for target vowel /a/). These relations are the basis for the significant figure × ground interactions for all target vowels. The marginal statistical significance of this interaction for /i/ judgments is due to a smaller overall increase in prominence ratings with the addition of modulation.

To test for equality among conditions where the target

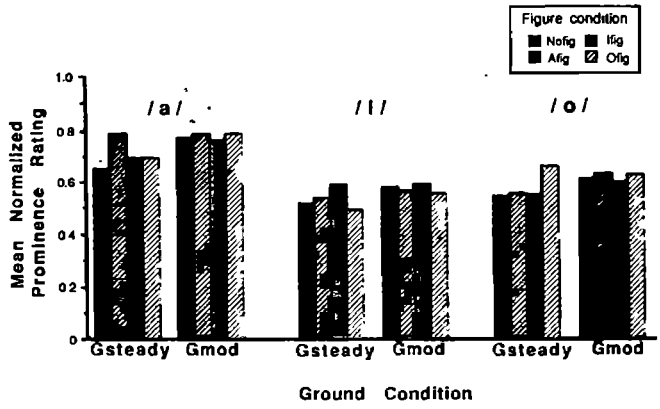

FIG. 3. Mean normalized prominence ratings across permutation conditions and subjects for target vowel /a/, /i/, and /o/. The figure conditions are grouped by ground condition.

was unmodulated and among those where it was modulated, a series of Scheffé *post hoc* comparisons were performed (Ferguson, 1971, pp. 270–271). None of the comparisons were significant at the 0.05 level, indicating that there is no effect of the modulation state of nontarget vowels on prominence ratings of the target vowel. This test also indicates that there is no effect of the frequency of modulation of the target vowel; if such were the case, we would have obtained significant comparisons between, for example, *Afig* and the other figure conditions (*Nofig, Ifig,* and *Ofig*) under *Gmod* for vowel /a/.

3. Comparisons among permutation conditions of nontarget vowels for a given pitch position of the target vowel

Another question one might ask is whether the permutation of the two nontarget vowels has any effect on the prominence of the target vowel. A difference in judged prominence of the target vowel when the pitch positions of the nontargets is inverted might indicate that masking is playing a role since one permutation may mask certain features of the attended vowel's spectrum more than the reverse. In fact, evidence for this effect is rather sparse across conditions. Scheffé *post hoc* comparisons were performed between permutation conditions within a given target vowel's pitch position across figure and ground conditions (e.g., *aoi* vs *aio* for /a/ at C_3, *oai* vs *oia* for /o/ at C_3, *iao* vs *ioa* for /i/ at C_3, etc.). Only two comparisons were found to be statistically significant: the Bb_3 comparison for /i/ [$F'(5,474) = 37.61, p < 0.05$ for $k = 6$] and the F_3 comparison for /o/ [$F'(5,474) = 28.03$, $p < 0.05$ for $k = 6$]. For these two cases, no individual comparisons within figure and ground conditions was statistically significant. We may thus conclude that the effect of permuation of nontarget vowels for a given pitch position of the target vowel is of minor significance.

4. Regrouping the data and reanalysis

The relative insignificance of these two types of comparison (effect of nontarget modulation states and of nontar-

TABLE IV. Permutations and modulation conditions to be regrouped for each target vowel. (Σ_{Gmod} = *Nofig/Gmod* + *Afig/Gmod* + *Ifig/Gmod* + *Ofig/Gmod* = all ground-modulated conditions. NS = *Nofig/Gsteady*. AS = *Afig/Gsteady*. IS = *Ifig/Gsteady*. OS = *Ofig/Gsteady*.)

		Target vowel /a/	/i/	/o/
Pitch Position	C_3	aoi + aio	iao + ioa	oai + oia
	F_3	oai + iao	aio + oia	aoi + ioa
	Bb_3	oia + ioa	aoi + oai	aio + iao
Modulation State	*Unmod*	NS + IS + OS	NS + AS + OS	NS + AS + IS
	Mod	AS + Σ_{Gmod}	IS + Σ_{Gmod}	OS + Σ_{Gmod}

get permutations on prominence ratings of the target vowel) allows a collection of the data into categories among which more meaningful comparisons can be made. These new categories are related to the parameters of the target vowel within a given stimulus. Accordingly, data are collected into target modulation state (*Unmod* = unmodulated, *Mod* = modulated) by target pitch position (C_3, F_3, Bb_3). These new labels will be used to avoid confusion with the original stimulus conditions in subsequent discussion. Since the regrouping under modulation state and pitch position is with respect to the target vowel, a given combination is a collection of data from different stimulus conditions for each target vowel. The conditions included in each new cell are listed for each vowel in Table IV. In Table V are shown the collected means ($\bar{x}$), standard deviations (s.d.), and standard errors (s.e.) derived from the stimulus combinations listed in Table IV. For example, the cell /a/ *Unmod-*C_3 represent six stimuli: two permuations (*aoi, aio*) under three modulation conditions (*Nofig/Gsteady, Ifig/Gsteady, Ofig/Gsteady*); the cell /i/ *Mod-*Bb_3 represents ten stimuli: two permutations (*aoi, oai*) under five modulation conditions (*Ifig/Gsteady, Nofig/Gmod, Afig/Gmod, Ifig/Gmod, Ofig/Gmod*). These data are plotted in Fig. 4 to compare among pitch positions and modulation states.

TABLE V. Means, standard deviations, and standard errors for data pooled according to target vowel's pitch position and modulation state. For *Unmod*, $n = 60$; *for Mod*, $n = 100$.

		/a/ *Unmod*	/a/ *Mod*	/i/ *Unmod*	/i/ *Mod*	/o/ *Unmod*	/o/ *Mod*
C_3	$\bar{x}$	0.75	0.79	0.49	0.52	0.59	0.67
	s.d.	0.077	0.059	0.086	0.085	0.100	0.101
	s.e.	0.010	0.006	0.011	0.008	0.013	0.010
F3	$\bar{x}$	0.73	0.78	0.53	0.57	0.56	0.61
	s.d.	0.103	0.082	0.108	0.088	0.104	0.108
	s.e.	0.013	0.008	0.014	0.009	0.013	0.011
Bb_3	$\bar{x}$	0.56	0.77	0.53	0.64	0.50	0.60
	s.d.	0.139	0.102	0.108	0.139	0.101	0.116
	s.e.	0.018	0.010	0.014	0.014	0.013	0.012

The pooled data for each target vowel were submitted to separate two-way analyses of variance with factors pitch and modulation state of the target vowel. The results are listed in Table VI. Both main effects are highly significant statistically for all three target vowels. The pitch × modulation interaction is significant for /a/ and /i/ judgments.

5. Effect of the modulation state of the target vowel

The significant main effect of modulation state reflects the fact that in every case the mean normalized prominence rating of a modulated target vowel was greater than that for a similar unmodulated target. The effect of modulating the target vowel was to increase significantly its prominence in a complex spectral background whether that background was steady or modulated. This increase also occurred independently of whether the other vowels were modulated coherently or incoherently with the target vowel.

6. Effects of pitch position of the target vowel

It is clear from studying Fig. 4 that the relative prominence of a target vowel at the different pitches changes with the vowel's modulation state, which is reflected in the significant main effect for pitch. It is also clear that the vowels are affected differently by the pitch position in which they are placed, which is in turn reflected by the significant interaction between pitch and modulation state. For vowels /a/ and /o/, prominence always decreases with increasing fundamental frequency, but this decrease is less when the vowel is being modulated. For these vowels, one effect of modulation is to reduce the differences due to pitch position. For the vowel /i/, however, prominence tends to increase with increasing fundamental frequency. An effect of modulation for this vowel is to increase the difference due to pitch position. For all three vowels, the greatest change in prominence with modulation occurs with the Bb_3 position (the highest pitch, where the partials have the greatest spread).

The spread of spectral energy (and thus presumed maskability) is quite different for each vowel, with /a/ having the least spread followed by /o/ and then /i/ whose energy is very widely distributed across the 4-kHz range. It should be recalled that these vowels were matched for equal loudness in isolation. Therefore, it is likely that the major differences between the three vowels can be attributed to masking effects.

7. Summary of main results

The results of this experiment may be summarized as follows. (1) The judged prominence of a target vowel increases significantly when it is modulated compared to when it is not modulated. (2) The judged prominence of a target vowel at a given pitch is unaffected by the pitch position and modulation state of the other two vowels in the chord, regardless of the modulation state of the target. (3) The degree of increase in judged prominence between an unmodulated target vowel and a modulated target is a function of the vowel's fundamental frequency (or pitch position in the chord). For all vowels, the greatest increase is found for the highest position (Bb_3).

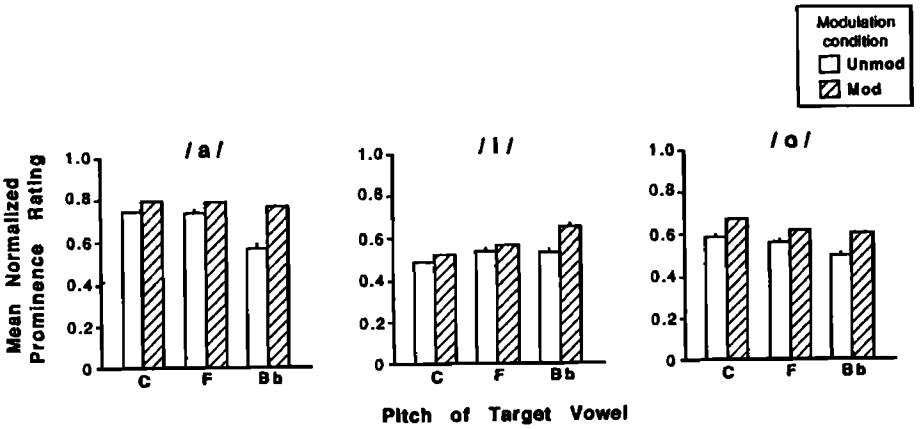

FIG. 4. Mean normalized prominence ratings collected under pitch and modulation state of the target vowel. Modulation conditions are grouped by pitch position. The vertical bars represent one standard error of the mean (for those cases where it is big enough to be seen).

III. DISCUSSION AND CONCLUSIONS

Some of these results are surprising, given the *a priori* assumptions, and several issues are brought into question.

A. Modulation of target vowels and coherence with nontargets

The data indicate quite clearly that, when a target vowel is modulated, its judged prominence increases significantly compared to when it is not modulated, contrary to an interpretation accorded to these results by Gardner and Darwin (1986, p. 183). One might conclude from this that some cue or cues associated with the coherent frequency modulation of a harmonic series can be used for grouping decisions. It is certainly true that coherence is maintained among the harmonics of a given vowel when it is modulated. However, the relative coherence of modulation of one vowel with another had little or no effect. This result (though derived from a very different task) is concordant with that of Gardner and Darwin (1986) who found that modulating a single harmonic incoherently with respect to the rest of the harmonic complex forming a vowel spectrum did not prevent that partial from contributing to the vowel identity (as measured by a change in phoneme boundary between two vowels). If the partial did not fuse with the complex due to its incoherent modulation, one would expect that it would be less likely to contribute to the spectral envelope used to identify the vowel. In the present study, if coherent frequency modulation alone were responsible for grouping, we would have expected vowels modulated coherently with respect to one another to be more difficult to segregate perceptually, resulting in lower prominence judgments. The fact that coherent modulation across source subgroups does *not* reduce separation with these stimuli suggests to the contrary that some other factor associated with frequency modulation is playing a strong role. The simple presence of modulation would thus seem to increase the subjective prominence of embedded vowels.

A remark on the size of the effect of modulation is pertinent at this point. While the effect of modulation is highly significant statistically, the difference in the means of pooled data between modulated and unmodulated target vowels is relatively small. The mean difference between *Unmod* and *Mod* conditions in Table V is 0.08. The range of means across subjects on normalized data is 0.45–0.82. The difference between *Unmod* and *Mod* conditions is thus approximately 20% of the total range of means for all vowels. This small size is somewhat surprising given that these stimuli have repeatedly proven very effective in demonstrating the modulation effect in public lectures, where people are nearly unanimous in reporting the emergence with modulation of a vowel that could not be heard when unmodulated, in spite of the deleterious effects that room reverberation should cause for a temporal cue like frequency modulation. Aside from the difference in presentation through loudspeakers as opposed to headphones, it seems likely that the experimental situation may be partially responsible. The subject is allowed to listen to many repetitions of the stimulus, is requested to

TABLE VI. ANOVA tables for two-factor analyses of variance on /a/, /i/, and /o/ judgments.

Source	*df*	Sum of squares	Mean squares	*F* ratio	*p* value
/a/ judgments					
Pitch (P)	2	1.004	0.502	57.271	0.0001
Mod state (M)	1	1.092	1.092	124.570	0.0001
P×M	2	0.623	0.311	35.513	0.0001
Error	474	4.156	0.009		
/i/ judgments					
Pitch (P)	2	0.555	0.277	25.140	0.0001
Mod state (M)	1	0.388	0.388	35.146	0.0001
P×M	2	0.156	0.078	7.069	0.0009
Error	474	5.229	0.011		
/o/ judgments					
Pitch (P)	2	0.506	0.253	22.470	0.0001
Mod state (M)	1	0.674	0.674	59.819	0.0001
P×M	2	0.043	0.021	1.899	n.s.
Error	474	5.339	0.011		

focus successively on the three vowels, and is subjected to 240 such trials—engendering a rather heavy bias toward an analytic listening independent of local stimulus context. This bias may also have contributed to the lack of effect of modulation coherence among vowels. Another reason may be the rather large individual differences in relative prominence ratings across subjects which would tend to reduce the size of a group effect, since the data of some subjects show some of the effects and not others and the individual effects obtained vary considerably across subjects.

B. Pitch position and modulation state of nontarget vowels

Neither of the parameters pitch position nor modulation state of nontarget vowels had any systematic effect on the judged prominence of a given target vowel. If, for example, /a/ was positioned at Bb_3 and was not modulated, its judged prominence was unaffected by whether or not /i/ and /o/ were modulated, or whether /i/ was at C_3 and /o/ at F_3 or vice versa. Three things are implied here. (1) Masking effects specifically due to the pitch arrangement of the nontargets are very slight; i.e., masking would seem to be relatively homogeneous for the vowel-pitch combinations used in this study. (2) Pitch separation may have reached its maximum effect as a cue for source distinction at the intervals used here. This hypothesis is supported by the results of Brokx and Nooteboom (1982) and Scheffers (1983a) which suggest that the maximum pitch separation effect for two voicelike sources is reached by about one to three semitones. A five-semitone interval is used in the present study. (3) No additional increase in prominence (or possibly release from masking) occurs due to the modulation of nontarget vowels. If such were the case, modulating a nontarget would increase the prominence of the target vowel. This result is somewhat surprising since one might imagine *a priori* that modulation of a nontarget, resulting in its emergence as a more clear source image, would simplify the situation allowing the target vowel to be more easily extracted. Obviously, the evidence from the present study indicates the contrary. Alternatively, these results could be explained by the interaction of two counteractive effects.[4] Modulation of nontarget vowels might reduce masking, but it also makes the nontargets more salient, and so the target is relatively less salient. The decreased masking might then be offset by the reduced salience. Further studies, varying the potential masking relations among the several simultaneous sources, are needed to determine the reason for this result.

In summary, the judged prominence of a given target vowel would seem to be independent of the specific pitch or modulation state of the other vowels, as well as of the prominence judgments made on those vowels. The presence or absence of modulation on masking (nontarget) vowels would not appear to have any effect on their masking potential as judged by these prominence ratings.

C. Informal reports on the perceived pitch of the vowels

There is another cue that may play a role in source perception when it is coupled with frequency modulation: the harmonicity of the frequency components of each vowel. Of interest here is the possibility that the harmonicity of a subset of partials may be a cue for separating it from the rest of the spectrum. There is some suggestion in the musical work of Chowning (1980) and McNabb (1981) that coherent, subaudio frequency modulation increases the apparent fusion and naturalness of synthetic voice and instrument sounds. It seems possible that for a sensory system which most often processes dynamic, rather than steady-state, signals, a coherently modulated harmonic series may be somehow less ambiguously harmonic (or less ambiguously a harmonic *group*) than a steady-state harmonic series. This possibility was lent informal support in the present experiment by reports of some musical subjects that there was a vast difference between the various stimuli with respect to the pitches perceived. They were told to expect three pitches (there were, in fact, three harmonic series), but reported sometimes hearing four to six pitches. In verifying this with different sets of highly trained musical subjects and a subset of the stimuli, it appeared that pitch ambiguity occurred most often where two or three of the vowels were steady. When all vowels were modulated, subjects reported that the three pitches were more clear and unequivocal.

Modern pitch theories would all have us believe (and our ears usually agree in the right context) that the most unambiguous or unequivocal pitch sensation in complex tones is perceived with a harmonic series. If FM coupled with a harmonic series reduced the ambiguity as to whether any subgroup was or was not harmonic (perhaps by reducing the possibility of analytic listening), we would expect a decrease in ambiguity of pitch perception. Several authors have noted that the virtual pitches of certain stimuli with little or no energy at the F_0 are better heard when the complex is modulated coherently than when it is stable (Thurlow and Small, 1955; Plomp, 1976). That such appears to be the case in the present study as well supports the hypothesis that perception of at least some (not specifically vocal) source qualities is dependent on the properties of the ensemble of elements collected as a group. These phenomena, while admittedly informally reported, may suggest that the effect of modulation is not limited to increasing the prominence of vowel phonemes, but can have an influence on pitch prominence as well.

D. Summary

In summary, we may conclude that factors associated with frequency modulation serve to increase the perceived prominence of vowels embedded among other vowels, but that this is independent of the coherence of the modulation of the different vowels. It is an open question of whether the harmonicity of the vowel spectra plays a role, though informal judgments on the pitch content of the multiple source stimuli indicate that dynamic harmonic spectra give less equivocal pitch percepts when several harmonic series are present than is the case for unmodulated spectra with identical frequency content. A system biased toward the processing of dynamic stimulus structures may consider a coherently modulated harmonic series as less ambiguously harmonic than one that is unmodulated. Aside from dynamic harmonicity as a possible grouping cue that may allow the vowel

identification *after* grouping, it seems equally possible that, for the stimuli reported in this study, modulation provides some as yet unspecified cues to special speech phoneme recognition mechanisms independently of grouping criteria. While much research by Darwin has shown that some grouping cues can influence phonetic quality (cf. Darwin, in press), other research, primarily on duplex perception, has demonstrated that separate sources can contribute to the same phonetic quality (cf. Mattingly and Liberman, 1988). The data on this question remain, on the whole, ambiguous.

It is difficult to determine from this experimental design the extent to which each of the factors of coherent FM, spectral envelope, and harmonicity contribute separately to source image formation and separation. For the stimuli used in this experiment, the subgroups of components on which all three cues converge are the three individual vowel spectra. For example, even when all three vowels are coherently modulated together, the whole ensemble is inharmonic and *may be* decomposed into three harmonic series. Also, there are incompatibilities in the amplitude modulation patterns of adjacent partials belonging to separate vowels that arise as each partial follows its own spectal envelope. Within each vowel there is harmonicity and coherence of frequency modulation under a single, constant, familiar spectral structure. The coupling of frequency and amplitude modulation in the tracing of a spectral envelope has been shown to be useful to subjects in the discrimination and identification of multiformant stimuli (McAdams and Rodet, 1988).

One useful test would be to uncouple the spectral envelope tracing from the frequency modulaton, by using an additive synthesis algorithm in which FM and AM can be controlled separately. In such a case, particularly with coherent modulation of all three vowels, one would expect vowel prominence to be degraded since the tracing information believed to be responsible for the increase in prominence would be perturbed, while at the same time the relative amplitudes of the harmonics would remain the same. This research will be reported in a subsequent paper (Marin and McAdams, in preparation).

ACKNOWLEDGMENTS

The basic ideas for this experiment derived from musical demonstrations by John Chowning and discussions with David Wessel. Bill Hartmann provided much advice, critical analysis, and encouragement; he also designed the response boxes. Bennett Smith wrote many of the experimental routines. Xavier Rodet helped with the singing vowel synthesis techniques. Dominique Lépine advised me on statistical issues. Laurent Demany, Earl Schubert, Cécile Marin, Chris Darwin, Marie-Claire Botte, Brian Moore, and Deb Fantini gave helpful comments on earlier versions of the manuscript. The preparation of this article was made possible in part by a research grant from the Fyssen Foundation, Paris, France.

[1]/a/ is an open-middle vowel; /o/ is a medium-back vowel; /i/ is a closed-front vowel. The choice to study sung as opposed to spoken vowels was an arbitrary one: The intial interests of this study were oriented toward computer-synthesized music.

[2]Loudness in the context of multiple sources will, of course, be changed by partial masking. Though the loudness matching procedure may appear superfluous at this point, an analysis of the perceptual data in the main experiment was performed, which, in principle, tested for mutual partial masking.

[3]A comparison by unpaired, two-tailed t tests of the 48 means for each vowel judgment between musically trained and untrained subjects yielded only two statistically significant comparisons, which can be attributed to chance. Therefore, it is concluded that there is no difference between the two groups. A possible criticism of the method would question whether subjects could make unbiased judgments, if they guessed that /a/, /i/, and /o/ were actually always present. Since one subject was the experimenter, who knew that this was the case, we can examine his data in relation to the means and standard deviations of the group. Only 30 of the 144 judgments had absolute z scores greater than 1. No absolute z scores were greater than 2. Of these 30, one-half were positive and the other half were negative, indicating there was no tendency for the prominence judgments to be higher as a result of the prior knowledge of the stimulus set.

[4]This explanation was suggested by Brian Moore.

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

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

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

Cardozo, B. L., and van Noorden, L. P. A. S. (**1968**). "Imperfect periodicity in the bowed string," IPO Prog. Rep. **3**, 13–15.

Charbonneau, G. R. (**1981**). "Timbre and the perceptual effects of three types of data reduction," Comp. Mus. J. **5**, 10–19.

Chowning, J. M. (**1980**). "Computer synthesis of the singing voice," in *Sound Generation in Winds, Strings, Computers* (Royal Swedish Academy of Music, Stockholm), pp. 4–13.

Dannenbring, G. L., and Bregman, A. S. (**1978**). "Stream segregation and the illusion of overlap," J. Exp. Psychol.: Hum. Percept. Perf. **2**, 544–555.

Darwin, C. J. (**1979**). "Perceptual grouping of speech components," in *Hearing Mechanisms and Speech*, edited by O. Creutzfeld, H. Scheich, and C. Schreiner (Springer, Berlin), pp. 333–340.

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

Darwin, C. J. (**1983**). "Auditory processing and speech perception," in *Attention and Performance X*, edited by H. Bouma and H. Bouwhuis (Erlbaum, Hillsdale, NJ), pp. 197–209.

Darwin, C. J. (**1984**). "Perceiving vowels in the presence of another sound: Constraints on formant perception," J. Acoust. Soc. Am. **76**, 1636–1647.

Darwin, C. J. (in press). "The relationship between speech perception and the perception of other sounds," in *Modularity and Motor Theory*, edited by I. G. Mattingly and M. G. Studdert-Kennedey (Erlbaum, Hillsdale, NJ).

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

Darwin, C. J., and Gardner, R. B. (**1986**). "Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality," J. Acoust. Soc. Am. **79**, 838–845.

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

Duifhuis, H., Willems, L. F., and Sluyter, R. J. (**1982**). "Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception," J. Acoust. Soc. Am. **71**, 1568–1580.

Ferguson, G. A. (**1971**). *Statistical Analysis in Psychology and Education* (McGraw-Hill, New York).

Flanagan, J. L. (**1972**). *Speech, Analysis, Synthesis and Perception* (Springer, Berlin).

Gardner, R. B., and Darwin, C. J. (**1986**). "Grouping of vowel harmonics by frequency modulation: Absence of effects on phonemic categorization," Percept. Psychophys. **40**, 183–187.

Grey, J. M., and Moorer, J. A. (**1977**). "Perceptual evaluations of synthesized musical instrument tones," J. Acoust. Soc. Am. **62**, 454–462.

Hartmann, W. M., McAdams, S., and Smith, B. K. (**1986**). "Matching the pitch of a mistuned harmonic in a complex sound," IRCAM Annu. Rep. 1986, 54–63.

Helmholtz, H. L. F., von (**1877**). *On the Sensations of Tone*, translated by A. J. Ellis (**1885**) (Dover, New York, 1954).

Kersta, L. G., Bricker, P. D., and David, E. E. (**1960**). "Human or machine? A study of voice naturalness," J. Acoust. Soc. Am. **32**, 1502 (abs).

Lieberman, P. (1961). "Perturbations in vocal pitch," J. Acoust. Soc. Am. 33, 597–603.
MacIntyre, M. E., Schumacher, R. T., and Woodhouse, J. (1981). "Aperiodicity in bowed string motion," Acustica 49, 13–32.
MacIntyre, M. E., Schumacher, R. T., and Woodhouse, J. (1982). "Aperiodicty in bowed string motion: On the differential-slipping mechanism," Acustica 50, 294–295.
Marin, C. M. H., and McAdams, S. (1989). "Segregation of concurrent sounds, II. Lack of effect of spectral envelope tracing" (in preparation).
Mattingly, I. G., and Lieberman, A. M. (1988). "Specialized perceiving systems for speech and other biologically significant sounds," in *Functions of the Auditory System*, edited by G. W. Edelman, W. E. Gall, and W. M. Cowan (Wiley, New York).
McAdams, S. (1982). "Spectral fusion and the creation of auditory images," in *Music, Mind and Brain: The Neuropsychology of Music*, edited by M. Clynes (Plenum, New York), pp. 279–298.
McAdams, S. (1984a). "The auditory image: A metaphor for musical and psychological research on auditory organization," in *Cognitive Processes in the Perception of Art*, edited by W. R. Crozier and A. J. Chapman (North-Holland, Amsterdam), pp. 289–323.
McAdams, S. (1984b). "Spectral fusion, spectral parsing and the formation of auditory images," unpublished Ph.D. dissertation (Stanford University, Stanford, CA).
McAdams, S., and Rodet, X. (1988). "The role of FM-induced AM in dynamic spectral profile analysis," in *Basic Issues in Hearing*, edited by H. Duifhuis, J. W. Horst, and H. P. Wit (Academic, London), pp. 359–369.
McNabb, M. M. (1981). "Dreamsong: The composition," Comp. Mus. J. 5, 36–53.
Moore, B. C. J., Glasberg, B. R., and Peters, R. W. (1985a). "Relative dominance of individual partials in determining the pitch of complex tones," J. Acoust. Soc. Am. 77, 1853–1860.
Moore, B. C. J., Peters, R. W., and Glasberg, B. R. (1985b). "Thresholds for the detection of inharmonicity in complex tones," J. Acoust. Soc. Am. 77, 1861–1867.
Plomp, R. (1976). *Aspects of Tone Sensation* (Academic, London).
Rasch, R. (1978). "The perception of simultaneous notes such as in polyphonic music," Acustica 40, 21–33.
Rodet, X. (1980). "Time-domain formant-wave-function synthesis," in *Spoken Language Generation and Understanding*, edited by J. C. Sinon (Reidel, Dordrecht, The Netherlands), pp. 429–441.
Rodet, X. (1982). IRCAM, Paris (unpublished data).
Scheffers, M. T. M. (1983a). "Sifting vowels: Auditory pitch analysis and sound segregation," unpublished doctoral dissertation (University of Groningen, The Netherlands).
Scheffers, M. T. M. (1983b). "Simulation of the auditory analysis of pitch: An elaboration on the DWS pitch meter," J. Acoust. Soc. Am. 74, 1716–1725.
Stumpf, C. (1890). *Tonpsychologie* (Hirzel-Verlag, Leipzig; reissued 1965, KnufBonset, Hilversum-Amsterdam) [cited in Brokx and Nooteboom (1982)].
Thurlow, W. R., and Small, A. M. (1955). "Pitch perception for certain periodic auditory stimuli," J. Acoust. Soc. Am. 27, 132–137.
Weintraub, M. (1985). "A theory and computational model of monaural auditory sound separation," unpublished Ph.D. dissertation (Stanford University, Stanford, CA).