Segregation of concurrent sounds. II: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width[a)]

Cécile M. H. Marin
Institut de Recherche et Coordination Acoustique/Musique (IRCAM), 31, rue Saint-Merri, F-75004 Paris, France

Stephen McAdams[b)]
Laboratoire de Psychologie Expérimentale (CNRS URA 316), Université René Descartes, 28, rue Serpente, F-75006 Paris, France and IRCAM, 31, rue Saint-Merri, F-75004 Paris, France

(Received 16 January 1990; accepted for publication 31 July 1990)

The research presented here concerns the simultaneous grouping of the components of a vocal sound source. McAdams [J. Acoust. Soc. Am. 86, 2148–2159 (1989)] found that when three simultaneous vowels at different pitches were presented with subaudio frequency modulation, subjects judged them as being more prominent that when no vibrato was present. In a normal voice, when the harmonics of a vowel undergo frequency modulation they also undergo an amplitude modulation that traces the spectral envelope. Hypothetically, this spectral tracing could be one of the criteria used by the ear to group components of each vowel, which may help explain the lack of effect of frequency modulation coherence among different vowels in the previous study. In this experiment, two types of vowel synthesis were used in which the component amplitudes of each vowel either remained constant with frequency modulation or traced the spectral envelope. The stimuli for the experiment were chords of three different vowels at pitch intervals of five semitones (ratio 1.33). All the vowels of a given stimulus were produced by the same synthesis method. The subjects' task involved rating the prominence of each vowel in the stimulus. It was assumed that subjects would judge this prominence to be lower when they were not able to distinguish the vowel from the background sound. Also included as stimulus parameters were the different permutations of the three vowels at three pitches and a number of modulation conditions in which vowels were unmodulated, modulated alone, and modulated either coherently with, or independently of, the other vowels. Spectral tracing did not result in increased ratings of vowel prominence compared to stimuli where no spectral tracing was present. It would therefore seem that it has no effect on grouping components of sound sources. Modulated vowels received higher prominence ratings than unmodulated vowels. Vowels modulated alone were judged to be more prominent than vowels modulated with other vowels. There was, however, no significant difference between coherent and independent modulation of the three vowels. Differences among modulation conditions were more marked when the modulation width was 6% than when it was 3%.

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

INTRODUCTION

In real life, we rarely hear a single sound source in complete isolation. We often listen to a voice speaking amid other voices, or an instrument playing in the midst of an ensemble. How do our ears distinguish the components of that voice from those of all the other voices and of noise? How does it group components into distinct sources? From research done on the responses of auditory-nerve fibers, we know that the ear analyzes complex sounds into narrow frequency bands and that it encodes the temporal behavior of the signal in those bands. This analysis is transmitted through frequency-specific channels to the brain. How does the brain, based on this analysis, manage to interpret sound sources as being distinct? To distinguish perceptually a sound object from others in the environment, the ear has to group various components that belong to a single source. Proposed simultaneous grouping cues include the common spatial origin of a sound's components, their shared harmonicity or periodicity, their coherent amplitude and frequency modulation behavior, and the coherent behavior of resonance structures (see McAdams, 1984a, for a review). Two of these cues will be considered in this paper: frequency modulation coherence and resonance structure behavior.

Subaudio frequency modulation, applied to the components of a sound in such a way as to maintain the frequency ratios among them ("coherent" modulation), has been shown to contribute to the ability of a listener to segregate

[a)] This study was realized in partial fulfillment of the requirements for C. M. H. Marin's DEA diploma at the Université Paris III (Marin, 1987). The original thesis in French may be obtained by writing to C. M. H. Marin at IRCAM.

[b)] Requests for reprints should be addressed to S. McAdams at the Laboratoire de Psychologie Expérimentale.

 0001-4966/91/010341-11$00 80

perceptually the sound from a background (Chowning, 1980; McAdams, 1984a,b, 1989; Chalikia and Bregman, 1989). In McAdams (1989), three vowels at fundamental frequencies (F_0) separated by five semitones were presented simultaneously to listeners. Several modulation configurations were used: (1) no vowels modulated, (2) a single vowel modulated and two vowels steady, (3) a single vowel modulated against a background of two coherently modulated vowels, and (4) all three vowels modulated coherently. Subjects were asked to rate the perceived prominence of each vowel in the mixture. It was assumed that if the components of a vowel were not grouped by the ear, subjects would not be able to perceive the target vowel easily and would thus give it a low prominence rating. If, on the other hand, the vowel was clearly distinguished from the other background sounds, subjects would easily perceive it and rate its prominence higher. The hypothesis was that a vowel modulated independently of other vowels would be more easily separated and thus receive higher prominence ratings than when it was not modulated or was modulated coherently with other vowels. The results showed that a vowel that was modulated was judged to be more prominent than a vowel that was not modulated. However, this increase in prominence was the same in conditions where the vowel was modulated either independently of, or coherently with, other vowels. Thus, while modulation of a harmonic sound increased its perceived prominence, the coherence of modulation on vowels at different F_0's separated by a ratio of 1.33 had no effect.

Huggins (1952, 1953) has suggested that the auditory system encodes aspects of the structure of a physical source. In the case of resonant sources, one aspect of this structure would be closely related to the spectral envelope. There is, in addition, an important possible interaction between frequency modulation and resonance structure in the perception of sound sources. Formant placement for a given vowel changes systematically across pitch registers in singing and formant relations evolve in ongoing speech. However, the spectral envelope tends to change relatively slowly with respect to the rate of frequency change of jitter or vibrato on the fundamental. With frequency modulation, each modulated component traces the spectral envelope of the vowel to which it belongs, thus possibly providing additional information about the resonance structure embedded in a multisource complex. The coupling between amplitude modulation and frequency modulation as a function of the resonance structure may help define the spectral envelope (by "tracing" the envelope) and thus help with its identification (McAdams, 1984a). Vibrato-induced spectral envelope tracing has been shown to facilitate the discrimination and identification of resonance structures (McAdams and Rodet, 1988). Spectral tracing is a feature of the formant-wave-function synthesis algorithm (Rodet, 1980) used in McAdams (1989), who hypothesized that this property of a resonance structure may help the auditory system identify the vowel and consequently result in an increase in its judged prominence.

It seems logical, then, that fixed resonance structure (as encoded through spectral envelope tracing) might be a cue for grouping. Features of the spectral envelope are certainly a crucial part of the information from which vowel identity is derived. Vowel prominence judgments are most likely closely tied to the ability of a listener to extract a spectral envelope from the complex spectrum and identify it as such. Factors that may impede this extraction would include: (1) the lack of definition of the spectral envelope by the frequency components composing the vowel, as is the case, for example, with a higher F_0, (2) the inability to extract the spectral envelope when the components that define it are grouped with other components, which thus give rise to a different spectral envelope, or (3) the masking of features essential for vowel recognition and identification, such as the lower two or three formant peaks (Carlson *et al.*, 1975; Karnickaya *et al.*, 1975).

With the synthesis algorithm used by McAdams (1989), each group of components representing a vowel was modulated under a constant spectral envelope, i.e., the resonance structure was unchanging. With no modulation, the nature of the resonance structure may have been ambiguous depending on the number of partials contained in each formant band. As modulation was added, each partial's frequency-amplitude motion potentially provided information about the slope of the spectral envelope in that region. When these components were taken as an ensemble, this slope information may have greatly reduced the ambiguity of identity. We would expect a reduction in ambiguity to be accompanied by an increase in prominence judgments.

This hypothesis was supported in the 1989 data by the large increase in prominence with modulation for the highest pitch of each vowel. With no modulation and a higher fundamental, there were fewer components within the formant passbands and the spectral form was thus less well defined. With modulation, this structure would be more clearly defined by the coupled frequency-amplitude motions. For adjacent partials belonging to separate vowels, these motions might be incompatible and indicate separate formant structures. In essence, each subgroup of partials traced its own spectral envelope. But this increased definition would only be possible if the auditory system could succeed in extracting the envelope information from the unresolved adjacent partials. Such would need to be the case for the relatively dense spectra employed in that experiment. The extent to which these envelopes could then be separated would influence the judged prominence of each of them.

The possible role of spectral envelope tracing in the presence of frequency modulation on vowel spectra was tested in the following experiment. Chords of three vowels in various pitch permutations were used as in the 1989 study. Two new conditions were also used: one in which the amplitudes of the spectral components varied as a function of the vowel spectral envelope, and one in which they remained at the amplitudes assigned to the components in steady-state vowels. These two cases are illustrated in Fig. 1. Our hypothesis was that, in addition to the positive effect of frequency modulation on judged vowel prominence, we should see greater prominence judgments for stimuli in which the spectral envelope was traced compared to those for which it was not traced. Note that, in the latter case, the spectral envelope as a whole is modulated in frequency.

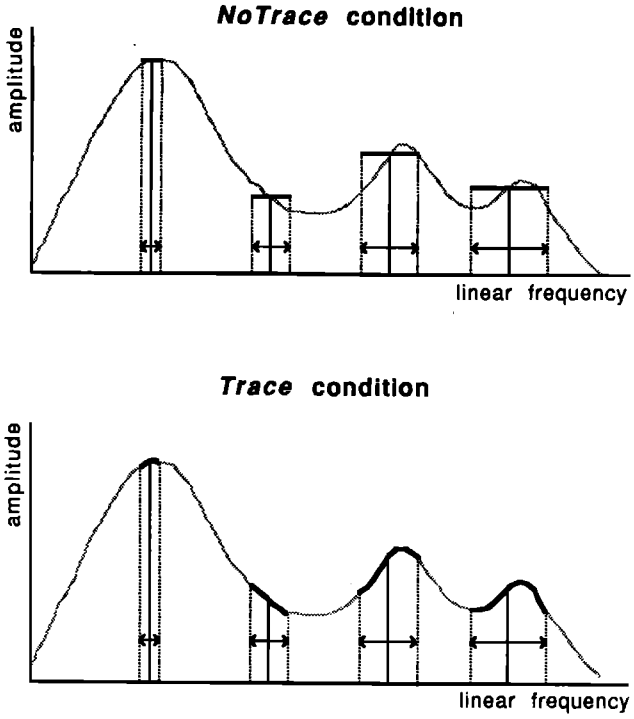

FIG. 1. Illustration of two kinds of vowel synthesis. Amplitudes of frequency components are initially chosen according to a spectral envelope for each vowel. In the *NoTrace* condition, these amplitudes remain constant as the frequencies are modulated. In the *Trace* condition, the amplitudes modulate with the frequency according to the spectral envelope.

TABLE I. Parameters of the spectral envelopes of vowels. With the additive synthesis algorithm used, the level values specified as parameters are those actually obtained in the synthesized signal.

	Formant frequency (Hz)	Bandwidth (Hz)	Level (dB *re*: F_1)
vowel /a/	600	78	0
	1050	88	−6
	2400	123	−12
	2700	128	−11
	3100	138	−24
vowel /i/	238	73	0
	1741	107	−20
	2450	123	−16
	2900	132	−20
	4000	150	−32
vowel /o/	360	51	0
	750	61	−11
	2400	168	−29
	2675	183	−26
	2950	198	−35

I. METHOD

A. Stimuli

The vowels /a/, /i/, and /o/ were used. They were presented at one of three fundamental frequencies (C_3–130.8 Hz, F_3–174.6 Hz, and Bb_3–233.1 Hz). Their durations were 2 s with 150-ms raised cosine attack and decay ramps. When vibrato was present on a given vowel, the vibrato width remained at zero for the first 300 ms, and increased linearly to its maximum value over the next 400 ms. The vibrato function consisted of a sinusoidal waveform. Two maximum peak-to-peak vibrato widths were used, 3% and 6%, and were presented to separate groups of subjects. The vibrato frequency chosen for a given vowel depended on the experimental condition, but was always chosen from among three values: 5.1, 5.7, and 6.3 Hz. (In the Appendix, an experiment is described which demonstrated that a concurrent pair of harmonic series modulated with the same vibrato rate can be discriminated from the same series modulated with independent rates. The same pitch separations and vibrato rates were used in the Appendix and the main study.) All vowels were synthesized at a sampling rate of 16 kHz in 24-bit floating-point format and then stored on disk in 16-bit integer format.

To analyze the role of the spectral envelope tracing by the harmonics of a vibrato vowel, two blocks of stimuli were prepared (see Fig. 1). The first block (*NoTrace*) was composed of vowels in which the amplitudes of the harmonic components remained constant when modulated in frequency. They did not trace the spectral envelope of the vowel. The second block (*Trace*) was composed of vowels with harmonics that traced the spectral envelope of the vowel when modulated.

The vowels used in McAdams (1989) experiment had been synthesized with a formant-wave-function synthesis algorithm (Rodet, 1980). The latter method does not permit independent behavior of amplitude and frequency of harmonics when the frequency is modulated. Therefore, the vowels in the present study were synthesized via an additive (Fourier) synthesis algorithm on an FPS-100 array processor connected to a VAX 11/780 computer. The center frequencies, bandwidths, and relative levels of the formants of the different vowels are listed in Table I. The spectral envelope functions derived from these parameters were stored in a table. To synthesize the vowels in the *Trace* block, the instantaneous value of the amplitude was given as a function of instantaneous frequency, according to the function in the table, for each harmonic at each sample. For the vowels in the *NoTrace* block, the amplitude value of each harmonic was determined beforehand according to the function table. Each harmonic then kept this amplitude value for the entire steady-state duration of the vowel.

Loudness matching was performed by six subjects on each vowel in isolation, at all pitches, with and without vibrato in both tracing and nontracing conditions. The vibrato had a rate of 5.1 Hz and a width of 3%. Each condition was judged at least two times by each subject.[1] The means of these judgments were used to equalize the loudnesses of individual vowel stimuli before mixing them into chords. There was no effect of envelope tracing and vibrato presence and very little effect of pitch on loudness matches. Vowels /a/ and /i/ tended to be adjusted on the order of 3 dB below the rms level of /o/ vowels.

As in the McAdams (1989) study, each experimental stimulus consisted of a chord composed of one each of the three vowels /a/, /i/, and /o/ at the three fundamental frequencies. Permuting the pitch positions of the three vowels

gives six chords that are each notated in order of increasing F_0 for the vowels: *aoi, aio, oai, oia, iao, ioa.*

For each permutation, six different modulation conditions were used: *Nomod*—no vowel was modulated; *Amod*—vowel /a/ was modulated alone at a rate of 5.1 Hz, while the other two vowels remained unmodulated; *Imod*—vowel /i/ was modulated alone at 5.1 Hz; *Omod*—vowel /o/ was modulated alone at 5.1 Hz; *Cohmod*—all three vowels were modulated coherently at 5.1 Hz; *Sepmod*—all three vowels were separately modulated at different vibrato rates (5.1, 5.7, 6.3 Hz). In the latter condition, the vibrato rates were randomly assigned to each vowel within different permutation and tracing combinations.

To obtain the different permutation and modulation combinations for each level of the tracing and vibrato width conditions, individually synthesized vowels were combined by a digital mixing program (in 32-bit format) to form the chords and stored in 16-bit format. These stimuli were then transferred to a PDP 11/34 minicomputer and presented through Tim Orr 16-bit DACS and a 6.4-kHz, −96-dB/oct, low-pass filter to Beyer DT-48 headphones. They were presented diotically at a level of approximately 75 dBA as measured at the earphones with a flat-plate coupler connected to a Bruel and Kjaer 2209 sound level meter.

B. Procedure

The experiment was conducted in a Soluna SN1 double-walled sound isolation chamber. At the outset, subjects received a recognition test of the individual synthesized vowels, in order to ascertain that they were able to recognize the timbre of each vowel in isolation. They were asked to identify all vowels used in the two spectral envelope tracing conditions, both with and without vibrato (at a rate of 5.1 Hz), and at all pitches. Each vowel stimulus was presented three times. To continue the experiment, the subject was required to obtain a global correct identification score of 95%, with no more than one mistake made for any given stimulus configuration.

Twenty subjects participated in the main experiment. Ten subjects were presented stimuli with a peak–peak vibrato width of 3% and ten others received the 6% vibrato width stimuli. The main experiment was divided into two sessions with the *Trace* and *NoTrace* conditions presented in separate sessions on different days. The order of the two conditions was evenly divided between the subjects within each of the 3% and 6% groups. Stimuli consisted of six permutations × six modulation conditions × five repetitions for each session. These 180 stimuli were presented in block randomized order with each stimulus being heard before a repetition was presented.

Subjects were informed that they were to judge the perceptual prominence of the vowels /a/, /i/, and /o/ within a complex stimulus. A slider provided for entering the judgments was labeled "very prominent" at the top and "not at all prominent" at the bottom. The experimenter indicated to the subjects that if the vowel was very clear or prominent and they were certain of its presence, then the slider was to be positioned at the top. If it was clearly not present, the slider was to be positioned at the bottom. If the impression of prominence or presence was in between, the slider was to be positioned accordingly. This demonstration should have induced the subjects to use the slider according to a linear scale of prominence. Subjects were told prior to the experiment that the three vowels might not necessarily be present in each stimulus. On each trial, subjects heard a 2-s complex sound repeated twice with an interval of 1 s between the two sounds. They were to judge the prominence of the vowel/a/ and position the slider. Then upon pressing a button, the same sound was heard twice a second time, after which they were to judge the prominence of the vowel /o/. Following this judgment a third presentation was provided for a judgment of the prominence of the vowel /i/. At each presentation, the target vowel to be judged was indicated on the screen of a computer terminal. Upon pressing the button for each judgment, the slider position was recorded and coded with a value between 0 (not at all prominent) and 100 (very prominent). At the end of three such judgments,the experimental program proceeded to the next stimulus configuration. Two breaks were introduced during each session, which made for periods of roughly 25–30 min each.

In each session, a practice sequence was presented during which the subject was to rate the presence of each vowel in the six modulation conditions of one permutation configuration, according to the tracing block being tested. The practice sequence was repeated if either subject or experimenter felt that the task had not been understood.

In summary, two groups of ten subjects each received either the 3% or 6% vibrato width. Within each group, all subjects received the *Trace* and *NoTrace* conditions in two separate sessions. Within each session, 180 stimuli were presented comprising six permutations each in six modulation configurations, with each such combination being repeated five times.

II. RESULTS

The values of the five prominence ratings for each stimulus configuration were averaged for each subject. The ranges of slider positions used across all conditions within a given vowel varied from 10–100 across subjects with an average range of 87. The mean prominence ratings were thus normalized with respect to the mean and standard deviation over all judgments for each subject and these values were then scaled and translated in order to fall within 0–100 across all subjects. The mean normalized data are presented in Figs. 2–4 for judgments on vowels /a/, /i/, and /o/, respectively.[2] For these data, the standard deviations and standard errors of the repetitions for each stimulus configuration and each subject were calculated. The global average of the standard deviations was 14 and the average of the standard errors was 6.4.

The mean normalized prominence ratings for each target vowel were submitted to independent four-way analyses of variance: subjects (10) within vibrato width (2) × tracing (2) × permutation (6) × modulation condition (6).

A. Effects of spectral envelope tracing

The main effect of spectral tracing was not significant for any of the target vowels. This factor was, however, in-

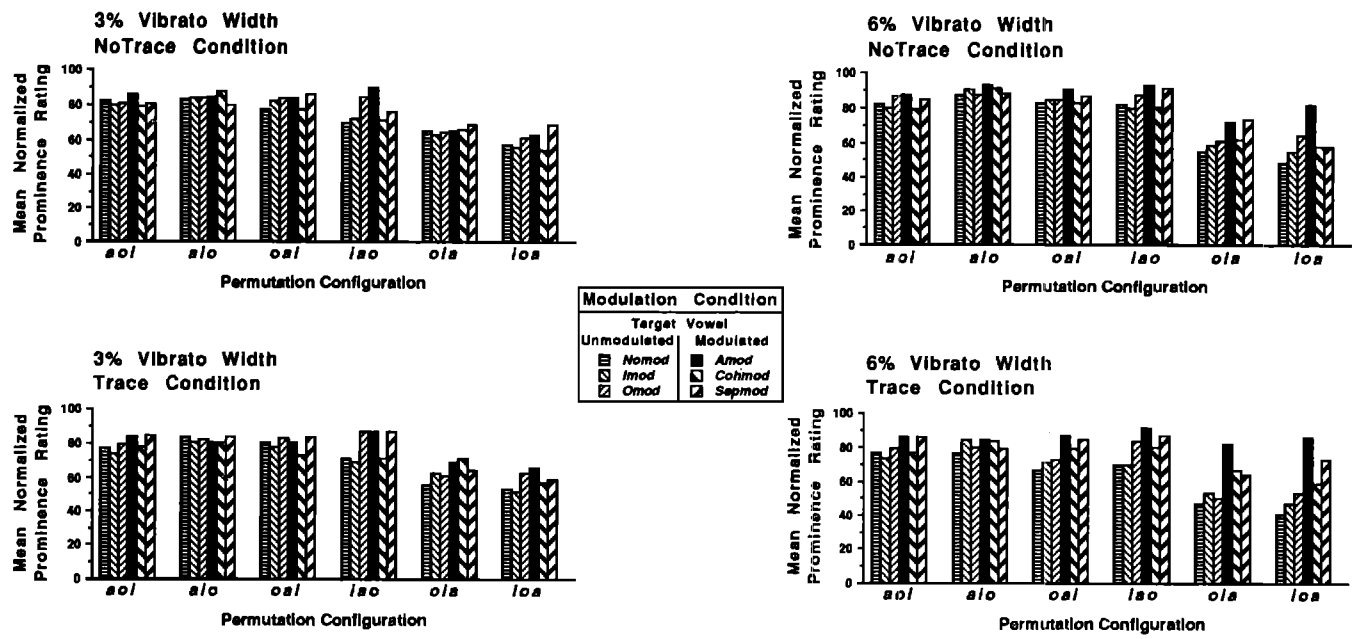

FIG. 2. Mean normalized prominence ratings on vowel /a/ across ten subjects within each of the vibrato width conditions. Vibrato width × tracing conditions are shown in separate plots. Modulation conditions are grouped by permutation configuration. Permutation configurations are arranged from left to right in order of increasing F_0 of the target vowel. Modulation conditions are arranged with target unmodulated conditions to the left, and target modulated conditions to the right within each permutation configuration.

volved in significant interactions with permutation and modulation conditions.[3]

B. Effects of modulation condition

The modulation condition main effect was significant at the 0.001 level for all three vowels [/a/: $F(5,90) = 17.92$; /i/:$F(5,90) = 9.97$; /o/:$F(5,90) = 14.09$]. For /a/, ratings in *Amod* and *Sepmod* conditions were higher than the others. For /i/ and /o/, ratings for conditions where the target vowel was modulated alone (*Imod* or *Omod*) had the highest ratings. Modulation condition interacted strongly with vibrato width for all three vowels [/a/: $F(5,90) = 3.39$, $p < 0.01$; /i/: $F(5,90) = 5.66$, $p < 0.001$; /o/:

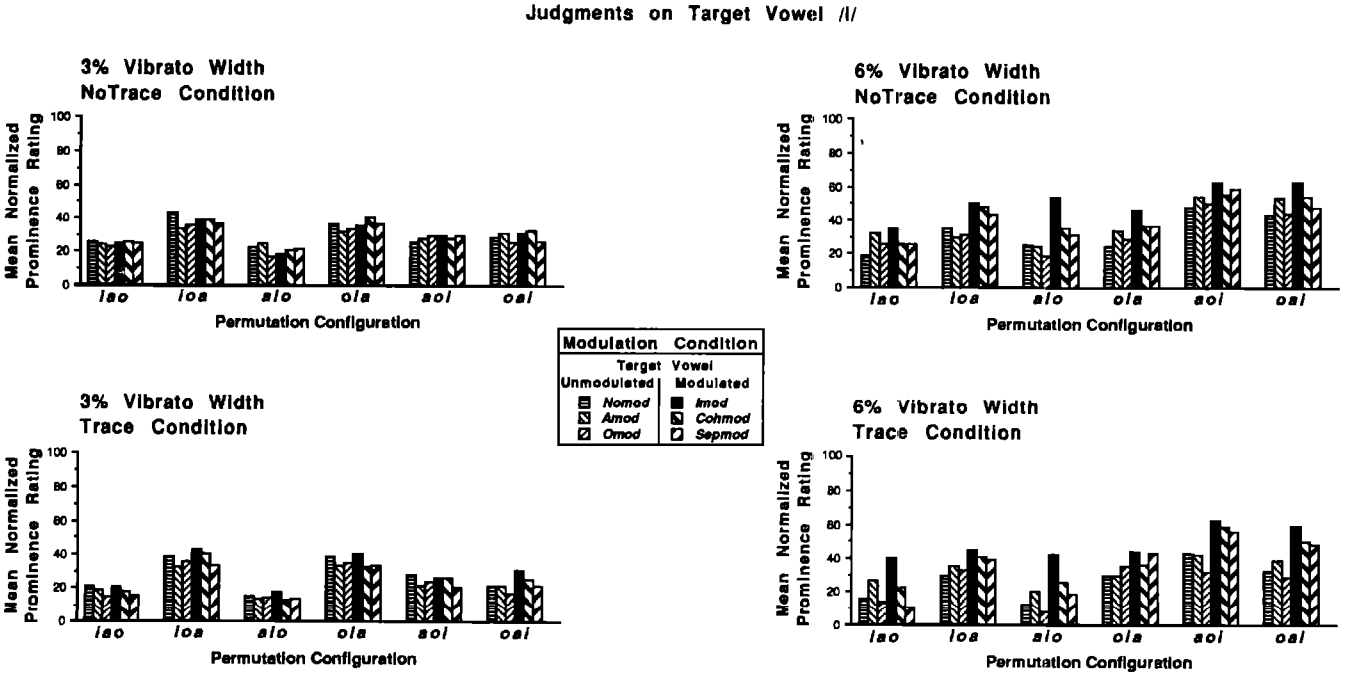

FIG. 3. Mean normalized prominence ratings on vowel /i/ as in Fig. 2.

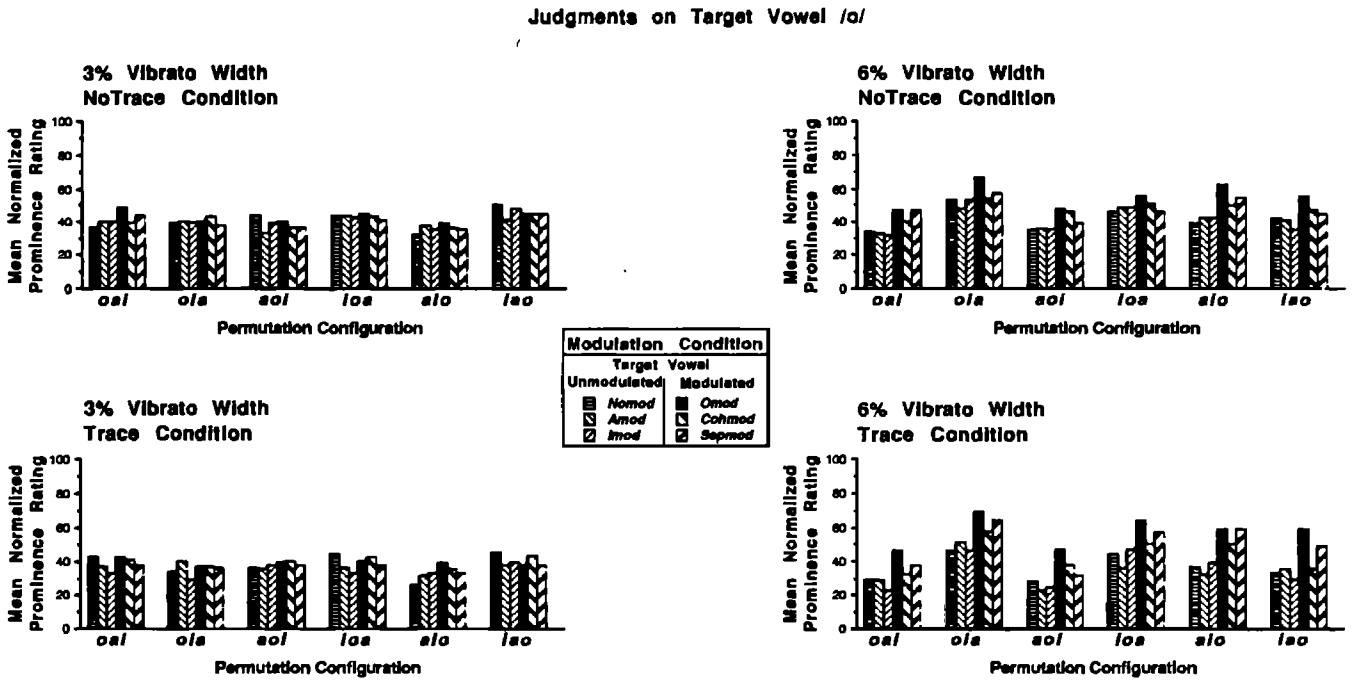

FIG. 4. Mean normalized prominence ratings on vowel /o/ as in Fig. 2.

$F(5,90) = 8.43, p < 0.001$]. In all cases, modulation conditions where the target vowel was modulated had greater prominence ratings at the 6% than at the 3% vibrato width. This increase with vibrato width was particularly strong when the target vowel was modulated alone. That this effect was not due simply to the fact that prominence judgments in the 6% subject group were generally higher than those in the 3% group, but was due to the effect of vibrato width itself, was verified by comparing the *Nomod* condition across the two groups, since these stimuli were identical. There was no statistically significant difference between these vibrato width groups for any of the vowels.

C. Effects of permutation configuration

The permutation main effect was significant for vowels /a/ and /i/ [/a/: $F(5,90) = 14.97$, $p < 0.001$; /i/: $F(5,90) = 7.22, p < 0.001$]. For /a/, permutations in which the target vowel was at the highest F_0 (*oia,ioa*) had lower prominence ratings than when it was situated at the two lower F_0's. For /i/, the permutations *aio* and *iao* had lower prominence ratings than the others. Permutation interacted significantly with vibrato width for /i/ and /o/ judgments [/i/: $F(5,90) = 3.61$, $p < 0.05$; /o/:$F(5,90) = 2.71$, $p < 0.05$). For both of these vowels, this interaction effect did not appear to be systematically related to the F_0 or modulation state of the target vowel.

D. Interactions between permutation configuration and modulation condition

The effects of permutation and modulation interacted significantly for vowels /a/ and /i/ [/a/: $F(25,450) = 5.75$, $p < 0.001$; /i/: $F = 1.52$, $p \approx 0.05$], though the significance level of 0.05 was just barely attained in vowel /i/ judgments. For both vowels, the variation was unsystematic, though /a/ judgments at the highest F_0 varied more with modulation condition than was the case with stimuli where /a/ was at lower F_0's.

E. Effects of vibrato width

From this initial analysis of the results, several interaction effects have been shown to exist among tracing, permutation, and modulation conditions which are dependent on the vibrato width in spite of the lack of statistical significance of the main effect in this ANOVA design. To investigate these effects more closely, individual analyses of variance were done on the separate vibrato width groups, since some of the effects of interest that were apparent in the data at the 6% vibrato width seemed to be obscured in the global analysis where 3% and 6% groups were mixed. Table II summarizes the results of these analyses. In most cases, the effect at 6% is much stronger than the effect at 3%. It is important to note that the tracing effect is significant at 6% for vowels /a/ and /i/ and approaches significance for vowel/o/. In all such cases, prominence ratings in *NoTrace* conditions are higher than those in *Trace* conditions. This result is contrary to one of the main initial hypotheses of the experiment. The modulation effect is highly significant for all vowels at 6%, but only for vowel /a/ at 3%. The permutation effect is highly significant for all target vowels at 6% and for vowels /a/ and/i/ at 3%. In these separate ANOVAs, no interaction effects were found to be statistically significant. Therefore, a linear model comprising only the main effects was used to make a number of *post hoc* contrasts.

F. *Post hoc* contrasts

Contrasts were performed to test for specific differences among permutation configurations and among modulation conditions. McAdams (1989) had only found differences

TABLE II. Results of separate three-way analyses of variance for each target vowel at vibrato widths of 3% and 6%. In each cell is shown the probability that the null hypothesis was true.

Source	df	/a/ 3%	/a/ 6%	/i/ 3%	/i/ 6%	/o/ 3%	/o/ 6%
Tracing	(1,708)	n.s.	0.001	n.s.[a]	0.027	n.s.	n.s.[a]
Permutation	(5,708)	<0.001	<0.001	<0.001	<0.001	n.s.	<0.001
Modulation	(5,708)	0.003	<0.001	n.s.	<0.001	n.s.	<0.001

[a] These effects approached significance ($p < 0.06$).

due to pitch position and modulation state of the target vowel. Particularly striking in those data was the lack of difference between conditions where the target vowel was modulated independently and those where it was modulated coherently with the other vowels. A new condition included in the present experiment was one where all three vowels were modulated independently of one another.

The permutation contrasts indicated that the relative pitch position of both nontarget and target vowels had an influence on target vowel prominence judgments.[4] These results contrast with those of McAdams (1989) where only the pitch of the target was important.

For modulation contrasts, we were interested in differences (1) among modulation conditions where the target vowel was not modulated (e.g., among *Nomod, Imod, Omod* for /a/), and (2) among modulation conditions where the target vowel *was* modulated (e.g., among *Amod, Cohmod, Sepmod* for /a/). In general, comparisons among unmodulated target vowel conditions were not statistically significant, indicating that the modulation state of nontarget vowels did not affect prominence judgments. The results of two groups of orthogonal contrasts on conditions in which the target was modulated are summarized in Table III. Comparisons among modulated target vowel conditions depended strongly on vibrato width. Only two comparisons were significant at the 3% vibrato width for /a/. At the 6% vibrato width, the comparisons between vowel modulated alone and either *Cohmod* or *Sepmod* conditions were highly significant for all three vowels. The *Cohmod–Sepmod* comparison was not significant. These results indicate that the greatest prominence is attained when the target vowel is modulated alone, compared to when it is modulated at the same time as the other vowels, whether this modulation be coherent or not. They also confirm McAdams' (1989) finding that the coherence of modulation among vowels has no effect on prominence ratings.

These comparisons suggest that the data for conditions where the target vowel is not modulated can be regrouped (into *Unmod*), as can conditions where all three vowels are modulated (into *Allmod*). The second group of contrasts [see Table III(b)] compares these two regrouped conditions with the one in which the target vowel is modulated alone (*Vmod*). Here, the difference in the pattern of results at 3% and 6% vibrato widths is also strongly apparent, as can be seen in the presentation of regrouped data in Fig. 5. Comparisons are rarely significant at the 3% width. Only the *Unmod–Vmod* comparison just attains the criterion significance level for the vowel /a/. All comparisons at the 6% vibrato width are highly significant for all three vowels. The ordering of the means is always *Unmod* < *Allmod* < *Vmod*. Taken together, these results support those of McAdams (1989), which demonstrated that the presence of a certain amount of frequency modulation on a vowel embedded among other vowels increases its perceptual prominence. In the present experiment, this effect is strongest if the target vowel is the only one modulated. Concurrent modulation of the other vowels reduces the target vowel's prominence, though with a 6% vibrato width this prominence is still greater than when the vowel is not at all modulated.

III. DISCUSSION

The present results are qualitatively consistent with those obtained by McAdams (1989), who used similar kinds of stimuli but which were produced with a different synthesis algorithm. The common points include the importance of

TABLE III. Results of *post hoc* orthogonal contrasts for each target vowel at vibrato widths of 3% and 6%: (a) among conditions within which the target vowel is modulated and (b) among various grouped conditions. In each cell is shown the probability that the null hypothesis was true for $F(1,708)$. *Vmod* indicates the condition where only the target vowel is modulated (*Amod, Imod,* or *Omod,* accordingly). *Unmod* indicates a grouping of all conditions in which the target vowel is not modulated. *Allmod* indicates a grouping of *Cohmod* and *Sepmod*.

Contrast	/a/ 3%	/a/ 6%	/i/ 3%	/i/ 6%	/o/ 3%	/o/ 6%
(a) *Vmod* vs *Cohmod*	0.007	<0.001	n.s.	0.013	n.s.	<0.001
Vmod vs *Sepmod*	n.s.	0.015	n.s.	<0.001	n.s.	0.010
Cohmod vs *Sepmod*	0.040	n.s.	n.s.	n.s.	n.s.	n.s.
(b) *Unmod* vs *Vmod*	0.004	<0.001	n.s.	<0.001	n.s.	<0.001
Unmod vs *Allmod*	n.s.	<0.001	n.s.	<0.001	n.s.	0.001
Vmod vs *Allmod*	n.s.	<0.001	n.s.	0.001	n.s.	<0.001

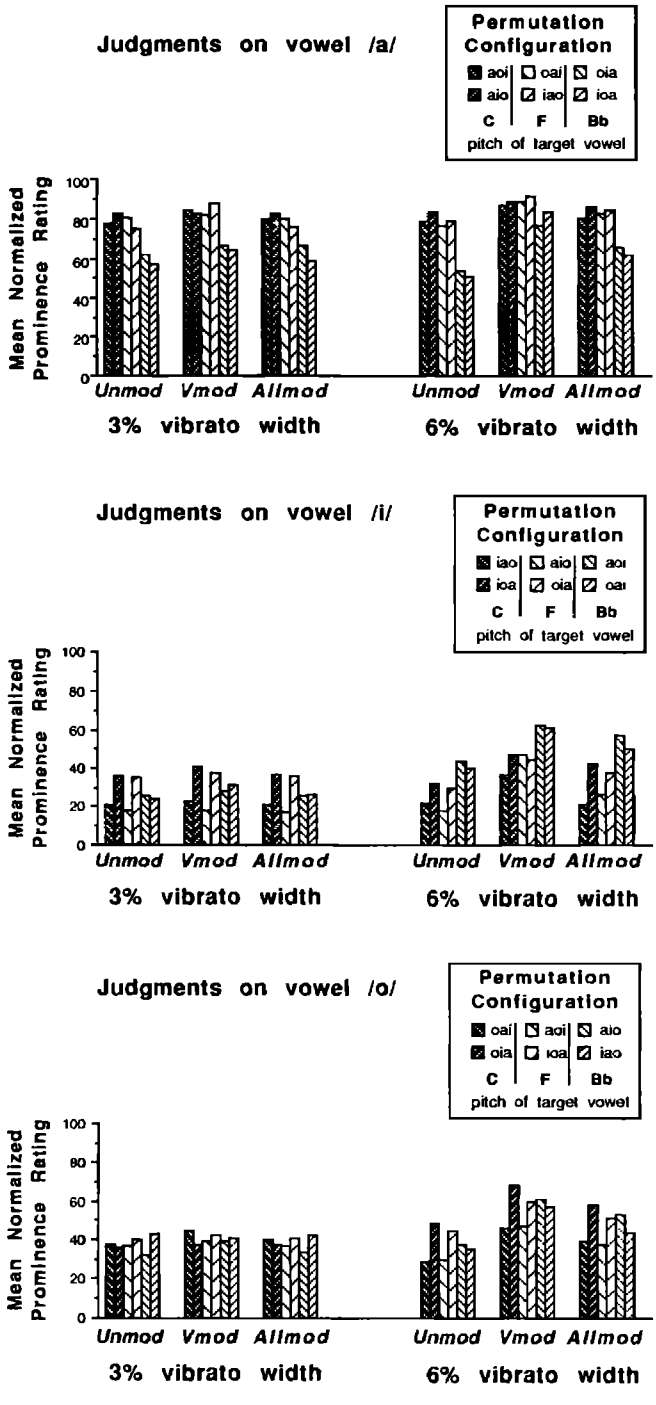

FIG. 5. Mean normalized prominence ratings across ten subjects within each vibrato width condition and across tracing conditions for judgments on vowels /a/, /i/, and /o/. The six original modulation conditions have been collapsed according to the modulation states of the component vowels (see text). Permutation configurations are grouped according to modulation condition and are arranged from left to right in order of increasing F_0 of the target vowel.

frequency modulation for the perceived prominence of a vowel in the presence of other concurrently presented vowels, the lack of effect of frequency modulation coherence among concurrent vowels, and the dependence of prominence ratings on the relative pitch positions of the vowels, though the details on this latter point differ in the two studies. In addition, several new findings are presented here concerning (1) the lack of influence of spectral envelope tracing on perceived prominence and, by inference, on vowel source separation, (2) the importance of the width of frequency modulation in concurrent sound segregation, and (3) a tendency for a modulated vowel's prominence to be reduced by the presence of other modulating vowels, whether this modulation be coherent or not with the vowel being judged.

A. Spectral envelope tracing

We wanted to test one possible aspect of the coherent behavior of resonant structures that had been interpreted by McAdams and Rodet (1988) as contributing to vowel discrimination and identification. We found, however, that this cue had either no effect or a negative effect on ratings of perceived prominence. In the case of a 3% vibrato width, there was no significant difference between prominence ratings for target vowels in the *Trace* condition and those in the *NoTrace* condition. For a 6% width, the *NoTrace* stimuli were judged to be more prominent than the *Trace* stimuli, though the effect is very small compared to the effects of modulation state and pitch permutation. This result completely refutes the experimental hypothesis according to which the grouping of a vowel's harmonics and the identification of its formant frequencies were partially based on a tracing of the spectral envelope.

One might ask whether the amount of amplitude modulation induced by vibrato under the vowel spectral envelopes was large enough to be useful to listeners. We computed the AM depth inducted by a 6% vibrato on harmonics within each formant for all vowels at the three F_0's used. Only harmonics that were within 10 dB of the formant peak were analyzed. For all vowels at all pitches, at least one and as many as three harmonics satisfied the criterion in each formant. The only exception was /o/ at the highest pitch where the first two harmonics straddled the F_1 peak and had levels of -17 and -15 dB *re*: peak, respectively. For all other cases, the induced AM varied between 1.3 and 13.3 dB, values generally above AM detection threshold. The average induced AM present on harmonics forming each formant (across vowels and pitches) was 2.7 dB for F_1 (range 1.5–4.6), 4.9 dB for F_2 (range 1.7–8.4), 6.7 dB for F_3 (range 3.2–10.3), 8.3 dB for F_4 (range 3.2–13.3), and 7.1 dB for F_5 (range 1.3–13.0). The negative result concerning the contribution of spectral envelope tracing to sound source separation cannot thus be discounted on the basis of insufficient AM on the harmonics.

It remains possible, following the earlier-mentioned suggestions of Huggins (1952, 1953), that isolating a single, analytic aspect of resonance behavior such as spectral envelope tracing, does not capture a behavior to which the auditory system is sensitive. It does not seem farfetched to suppose that this system has been attuned by the process of evolution to the complex behavior of physical resonators that are so ubiquitous in the sound environment. When envelope tracing is not accompanied by, for example, an appropriate phase modulation in the region of the formant center frequency, the auditory system, if it were biased toward the analysis of resonant structures, might not recognize such behavior as "meaningful." Such speculation remains to be verified experimentally.

B. Frequency modulation

When a vowel was presented simultaneously with two others, its prominence was generally judged to be greater when it was modulated than when it was not. This effect depended, however, on the frequency-modulation width and the modulation states of the other vowels. Vowels that were modulated with a peak-to-peak vibrato width of 6% were judged to be more prominent than those modulated with a 3% width. Many differences due to modulation state conditions were significant for all three target vowels at the 6% width, but were either very small (for /a/ judgments) or inexistent (for /i/ and/o/ judgments) at the 3% width (see Table III, Fig. 5).

Several effects due to the modulation states of concurrent vowels were revealed at the 6% vibrato width. Modulation of a nontarget vowel had no impact on prominence ratings of unmodulated target vowels either at 3% or 6% modulation. When all three vowels were modulated, there was no effect on prominence ratings of their being modulated coherently or independently. This same tendency was found by Chalikia and Bregman (1989) who presented subjects pairs of vowels for which the F_0's were steady, gliding in parallel or gliding such that the contours crossed one another midway through the stimulus duration. In addition, the maximum F_0 separation between vowels varied from 0 to 12 semitones. They found that, while there was a tendency for parallel glides (equivalent to our coherent modulation) to give lower vowel identification scores than crossing glides (similar to our independent modulation condition), the differences were not significant at F_0 separations of $\frac{1}{2}$, three, and six semitones. The latter value is roughly equivalent to our five semitone F_0 separation. However, there was a significant decrease in identification scores for parallel glides when the F_0 separation was one octave; i.e., a perfectly harmonic relation existed between the components of both vowels. No such decrease was found for crossing glides. This would lead one to suppose that coherent modulation in and of itself is not sufficient to make two harmonic series fuse together. Coherent modulation may increase the fusion of harmonically related components, but not inharmonically related ones. Thus harmonicity may be considered a constraining factor on the grouping power of coherent modulation. Harmonicity has little or no effect on perceptual fusion due to coherent amplitude modulation (Bregman *et al.*, 1990). In the stimuli in our study, the components of individual vowels can be fused since they are harmonically related, but the groups of components across coherently modulated vowels cannot be fused since they are not harmonically related.

A result in the present study not found in McAdams (1989) was an increase in ratings for a vowel modulated alone compared to when it was modulated in the presence of other modulating vowels (*Vmod* vs *Allmod* in Fig. 5). This implies a reduction in prominence due to a mutual interference of multiple modulating sources that perhaps perturbs, without completely obscuring, the information (most likely of a temporal nature) needed by the auditory system to separate and identify the individual sources. While the frequency components were relatively noncoincident, due to the separation of the three fundamental frequencies by a musical fourth, there was significant overlap of activity patterns on the basilar membrane in the higher harmonics of all three vowels.

C. Differences observed with McAdams (1989)

The main discrepancies with the 1989 study may be summarized as follows. (1) Effects of frequency modulation at a 3% modulation width in that study were not found at that width in the present study, but were found at a 6% width. (2) In the 1989 study, no effect of the modulation state of nontarget vowels was found when the target vowel was modulated, whereas in these data *Allmod* stimuli were judged to be less prominent than *Vmod* stimuli.

There were two differences in the synthesis of the stimuli and one methodological difference between the two experiments. The 1989 study used a time-domain formant-wavefunction (FOF) synthesis algorithm whose behavior is closer to (though not identical to) that of a true resonator than the additive synthesis algorithm used here. This difference may implicate a sensitivity in the auditory system to the phase structure of resonators. The other synthesis difference was the use of jitter (1.6% rms modulation width) combined with a sinusoidal vibrato (3% peak-to-peak width) in the earlier study. Only the vibrato was used in the present study. While the presence of jitter may have increased the effective modulation width to some extent, it seems unlikely, given the size of the effects reported in both studies, that such a difference could be entirely responsible for the discrepancies in results.

The methodological difference lay in the amount of time each subject had to listen to each stimulus configuration. In the 1989 study, subjects listened continuously to the repeating 2-s stimulus while making the prominence ratings for the three vowels in succession. Some subjects may have listened to a given stimulus for as long as 30 to 60 s (10–20 presentations). In the present study, each stimulus was presented twice for each vowel judgment, or a total of six times per repetition. The kind of listening strategy developed may have been quite different in the two cases, and a prolonged presentation of the stimulus could allow the listener more time to focus on cues related to frequency modulation.

IV. CONCLUSIONS

The present study found that spectral envelope tracing did not increase the perceived prominence of vowel sounds embedded among other vowels compared to sounds in which the component amplitudes remained constant. The presence of coherent frequency modulation on the components of a single vowel increased its perceived prominence compared to when the vowel was not modulated. The prominence of a modulated vowel was reduced if the other vowels were also modulated, and this reduction was independent of whether the modulations among the vowels were coherent or not, though these latter conditions still resulted in higher vowel prominence than was obtained for unmodulated vowels. All of the modulation effects depended on the frequency modulation width: they were small or nonexistent at a 3% peak-to-peak width, and were much stronger at a 6% width. We conclude that coherent, subaudio frequency modulation on

a harmonic sound source contributes to its segregation from other concurrent sounds if the modulation width is large enough, but that coherent modulation among harmonic sources does not cause them to fuse together if the resulting combined spectrum is not harmonic. Modulation incoherence on harmonic sources at different F_0's does not always increase the separation due to the F_0 difference. Thus harmonicity is probably a stronger grouping cue than frequency modulation coherence. The latter cue can, nonetheless, reinforce the segregation effect of harmonicity in situations of multiple complex harmonic sources.

ACKNOWLEDGMENTS

We would like to thank Bennett Smith for programming and technical support, Xavier Rodet for help with sound synthesis, Laurent Demany and Catherine Semal for methodological suggestions at an early stage of this work, and Marie-Claire Botte for letting us use the computer in her office at odd times to do several of the statistical analyses. Helpful comments were made by Jean-Sylvain Liénard, John Grose, and two anonymous reviewers. The preparation of this article was made possible in part by a graduate research fellowship to C. M. H. Marin from the D. C. N. of the French Ministry of Defense and by a grant to S. McAdams from the Fyssen Foundation, Paris, France.

APPENDIX: DISCRIMINABILITY OF MULTIPLE VIBRATO RATES

In order to facilitate the interpretation of the effects of frequency modulation coherence on vowel prominence ratings, it was necessary to establish that the presence of single or multiple vibrato rates could be distinguished. This is particularly important for the comparison between *CohMod* and *SepMod* conditions. If these could not be discriminated, one would not expect prominence judgments to be different between them.

1. Stimuli

Stimuli were constructed from pairs of 16-component, flat-spectrum harmonic series based on the three F_0's used in the main experiment (130.8, 174.6, 233.1 Hz). This yielded three possible F_0 pairs. Each harmonic series was modulated with a 3% peak-to-peak sinusoidal frequency modulation. Tones were 2 s in duration with 200-ms raised cosine attack and decay ramps. The stimuli were synthesized at a sampling rate of 16 kHz. The sound presentation system was identical to that in the main study. Each trial consisted of two tones separated by a 500-ms silence. In the standard tone, the vibrato rates on both sets of harmonics were identical. In the comparison tone, a different vibrato rate was present on either the higher or the lower harmonic series. The 5.1-Hz rate was compared with both 5.7- and 6.3-Hz rates (rate differences of 0.6 and 1.2 Hz, respectively). The order of standard and comparison tones was randomized. Subjects were asked to decide which of two tones in a trial had a single vibrato rate present on both harmonic series. In essence, they were to detect the tone that had coherent frequency modulation. Performance was measured as the percentage of correct responses. Each configuration of vibrato comparison and pitch pair (24 total) was repeated five times in block randomized order (120 trials per session). Eight subjects (including the authors) each completed two sessions. Only data from the second session were analyzed.

TABLE AI. Results of the concurrent vibrato rate discrimination experiment (mean percent correct responses).

F_0 pair (Hz)	Vibrato rate pair (Hz) 5.1/6.3	5.1/5.7
130.8/174.6	98	94
174.6/233.1	97	96
130.8/233.1	51	58

2. Results

The results are summarized in Table AI. The data were submitted to an analysis of variance with Pitch pair and Vibrato rate pair as factors (3×2). There was no effect of vibrato rate nor was the interaction term significant. Thus the 0.6- and 1.2-Hz differences in vibrato rates are equally distinguishable within each pitch pair. There was, however, a significant effect of pitch pair [$F(2,186) = 156.8$, $p < 0.0001$]: performance is close to perfect when the pitch separation is five semitones and falls almost to chance when the separation is ten semitones. In essence, the latter condition is not relevant to our study since adjacent vowels were always separated by five semitones. This auxiliary study thus rules out the possibility that the lack of difference in prominence judgments between *CohMod* and *SepMod* conditions was due to listeners' inability to discriminate between them.

[1]Two subjects made two judgments for each condition; three subjects made six judgments and one subject made seven judgments. There were no significant differences among these three groups.

[2]Figures 2–4 show that, across conditions, /a/ was judged as more prominent than /o/, which was judged more prominent than /i/. The same qualitative result was found in McAdams (1989) and was attributed in that paper to masking effects among vowels (Sec. II C 6, p. 2155).

[3]Significant interactions of secondary interest to this study that are not discussed in the text include the following: modulation × tracing for /a/ [$F(5,90) = 6.60$, $p < 0.001$]; modulation × vibrato width × tracing for /a/ [$F(5,90) = 3.16$, $p < 0.05$]; permutation × modulation × tracing for/a/ [$F(25,450) = 2.07$, $p < 0.01$]; permutation × tracing for/i/ [$F(5,90) = 2.70$, $p < 0.05$].

[4]Seven of the 18 contrasts performed on permutations within target vowel pitch positions were statistically significant (e.g., between *aoi* and *aio* for /a/). The prominence relations of target vowels at different pitches qualitatively reflect the same structure found in the 1989 study for vowels /a/ and /i/. Due to the differences *within* pitch position, however, we cannot conclude that the principal effect of permutation is due to pitch position of the target vowel, as was the case in that earlier study.

Bregman, A. S., Levitan, R., and Liao, C. (1990). "Fusion of auditory components: Effects of the frequency of amplitude modulation," Percept. Psychophys. 47, 68–73.

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

Chalikia, M., and Bregman, A. S. (1989). "The perceptual segregation of simultaneous auditory signals: Pulse train segregation and vowel segregation," Percept. Psychophys. 46, 487–496.

Chowning, J. (1980). "Computer synthesis of the singing voice," in *Sound*

Generation in Winds, Strings, Computers (Royal Swedish Academy of Music, Stockholm), Pub. No. 29.

Huggins, W. H. (**1952**). "A phase principle for complex frequency analysis and its implications in auditory theory," J. Acoust. Soc. Am. **24**, 582–589.

Huggins, W. H. (**1953**). "A theory of hearing," in *Communication Theory*, edited by W. Jackson (Butterworths, London), pp. 303–379.

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

Marin, C. M. H. (**1987**). "Rôle de l'enveloppe spectrale dans la perception des sources sonores," DEA thesis, Universite Paris III, Paris.

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 R. Crozier and A. Chapman (North-Holland, Amsterdam), pp. 298–324.

McAdams, S. (**1984b**). "Spectral fusion, spectral parsing and the formation of auditory images," unpublished Ph.D. dissertation, Stanford University, Stanford, CA.

McAdams, S. (**1989**). "Concurrent sound segregation. I: Effects of frequency modulation coherence," J. Acoust. Soc. Am. **86**, 2148–2159.

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.

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.