Extracting spectral envelopes: Formant frequency matching between sounds on different and modulated fundamental frequencies

Pascal Dissard and C. J. Darwin[a)]
Experimental Psychology, University of Sussex, Brighton BN1 9QG, United Kingdom

(Received 28 December 1998; revised 17 May 1999; accepted 18 October 1999)

The four experiments reported here measure listeners' accuracy and consistency in adjusting a formant frequency of one- or two-formant complex sounds to match the timbre of a target sound. By presenting the target and the adjustable sound on different fundamental frequencies, listeners are prevented from performing the task by comparing the absolute or relative levels of resolved spectral components. Experiment 1 uses two-formant vowellike sounds. When the two sounds have the same $F0$, the variability of matches (within-subject standard deviation) for either the first or the second formant is around 1%–3%, which is comparable to existing data on formant frequency discrimination thresholds. With a difference in $F0$, variability increases to around 8% for first-formant matches, but to only about 4% for second-formant matches. Experiment 2 uses sounds with a single formant at 1100 or 1200 Hz with both sounds on either low or high fundamental frequencies. The increase in variability produced by a difference in $F0$ is greater for high $F0$'s (where the harmonics close to the formant peak are resolved) than it is for low $F0$'s (where they are unresolved). Listeners also showed systematic errors in their mean matches to sounds with different high $F0$'s. The direction of the systematic errors was towards the most intense harmonic. Experiments 3 and 4 showed that introduction of a vibratolike frequency modulation (FM) on $F0$ reduces the variability of matches, but does not reduce the systematic error. The experiments demonstrate, for the specific frequencies and FM used, that there is a perceptual cost to interpolating a spectral envelope across resolved harmonics. *© 2000 Acoustical Society of America.* [S0001-4966(00)00202-2]

PACS numbers: 43.66.Jh, 43.71.Es [RVS]

INTRODUCTION

This paper addresses the general question of what representations of sound mediate between a peripheral, spectral representation and the abstract categories of speech such as vowels. Specifically, it measures how accurately and consistently listeners can match a formant frequency in complex sounds under conditions which force them to use abstract representations of a sound's spectrum.

The idea behind the experiments is that if listeners have ready access to a particular representation such as a spectral envelope or a formant frequency, they should be able to make perceptual matches on the basis of that representation despite variation in other dimensions of the stimulus.

Although a number of experiments have been carried out on the ability of listeners to discriminate changes in formant frequency for single-formant sounds (see, for example, Lyzenga and Horst, 1995, 1997) or for multi-formant synthetic vowels (see for example, Kewley-Port and Watson, 1994; Kewley-Port, 1995; Kewley-Port *et al.*, 1996), these experiments can all be performed either by a simple identity match of (part of) the excitation pattern (Moore and Glasberg, 1983) or, where signal levels are roved, by profile analysis (Green, 1988)—comparing *relative* levels in the excitation pattern. Listeners need not extract either an interpolated spectral envelope or a formant frequency in order to perform the task.

By presenting sounds on different fundamental frequencies ($F0$), we can reduce the value to listeners of being able to compare absolute or relative levels at corresponding places in the excitation pattern. If they are to obtain a reliable, veridical match, they must then use a more abstract representation which is closer to the spectral envelope.

Whether the simple strategy of comparing corresponding places on the excitation pattern is useful depends on the relationship between the formant frequency and the fundamental frequency. When a formant is excited by high-numbered, unresolved harmonics, the spectral envelope is represented explicitly in the excitation pattern. Sounds on different fundamentals could then be matched by comparing absolute (or relative) levels of excitation patterns. However, when the formant frequency is a smaller multiple of $F0$, so that harmonics close to the formant peak are resolved, their local peaks prevent the spectral envelope being explicitly represented in the excitation pattern. If the spectral envelope is to be used by listeners, it must be derived from the excitation pattern. This process may have a perceptual cost: listeners may have relatively greater difficulty in making different-$F0$ matches for sounds presented on high $F0$'s than for sounds presented on low $F0$'s. The main object of this paper is to test whether there is such a perceptual cost.

How the spectral envelope might be extracted from an excitation pattern of resolved harmonics has not been dis-

[a)]Author to whom correspondence should be addressed. Electronic mail: cjd@biols.sussex.ac.uk cussed in the literature, although the analogous problem of vernier acuity in spatial vision has received some attention. Observers can locate the spatial position of a spatially periodic pattern with a precision as high as 5–10-s arc, even though the pattern is coarsely sampled at an interval over ten times that amount (Morgan and Watt, 1982). They probably achieve this by interpolating luminance profiles on the basis of a few samples (Kontsevich and Tyler, 1998).

A number of papers have addressed the more complex problem of how the first formant frequency ($F1$) might be extracted. An early suggestion that listeners simply equate $F1$ with the frequency of the dominant harmonic (Mushnikov and Chistovich, 1972) has been discounted in favor of methods which either form a weighted sum of two (Carlson *et al.*, 1975; Assmann and Nearey, 1987) or more (Darwin and Gardner, 1985) harmonic frequencies close to the formant. Klatt (1986) has pointed out that such weighting models, together with alternative approaches such as spectral smoothing and linear predictive coding (LPC) analysis produce formant frequency estimates which are biased in the direction of the dominant harmonic. Such bias can produce formant estimation errors as large as 16%. In contrast, vowel identification experiments have found no explicit evidence for such shifts (Florén, 1979; Klatt, 1985). Experiment 2 of this paper provides evidence of such a bias.

The third issue addressed in this paper is whether adding frequency modulation (FM) to $F0$ changes the accuracy and veridicality of formant-frequency matches. In principle, a changing $F0$ can provide additional information about the value of the spectral envelope over a range of values around each harmonic, and, through amplitude modulation of the individual harmonics, information about the slope of the spectral envelope. However, previous work on the effects of vibratolike FM of $F0$ on the perception of formants or of vowels has given mixed results.

For example, on the one hand, McAdams and Rodet (1988) showed that differences in the slope of spectral envelopes could be discriminated and identified in the presence of a small amount of vibrato. They interpreted their data in terms of listeners using the spectral tracing produced by vibrato to discriminate and identify spectral envelopes with different formant frequencies (although whether the dynamic aspects of the stimulus were necessary was not established). On the other hand, Marin and McAdams (1991) could find no evidence for such spectral tracing increasing the prominence that vibrato gives a vowel against a background of other steady vowels. Similarly, rather little effect of vibrato on vowel identification thresholds was found by Demany and Semal (1990) for vowels ($F0=100\,\mathrm{Hz}$) masked either by noise or by a different-$F0$ pulse-train.

Beneficial effects of vibrato on vowel identification are, in principle, more likely to be found at high than at low $F0$'s since the sparser sampling of the spectral envelope can make vowel identification worse at high $F0$'s (Ryalls and Lieberman, 1982) although not invariably (Hillenbrand and Nearey, 1999). It is surprising then that Sundberg (1975, 1977) was unable to find such a beneficial effect. He examined the influence of ±0.5 semitone (±3%) vibrato on the identification of 12 synthetic Swedish vowels with $F0$'s between 300 and 1000 Hz. Overall, the effects were small, and in the majority of cases where the vibrato affected the response, the vowel identification became somewhat harder when the stimulus was presented with than without vibrato. Carlson *et al.* (1975) also comment that a modulated $F0$ slightly decreased the reliability of vowel labelling judgements. In experiments 3 and 4 we ask whether FM improves listeners' ability to match single-formant sounds on different $F0$'s.

In summary, the experiments described in this paper measure listeners' ability to match a formant frequency in complex sounds under stimulus conditions which force them to use a representation of a sound's spectrum that is more abstract than a simple excitation pattern. Subjects adjust the frequency of a single formant of either a one- or a two-formant sound so that it matches a similar sound, which can be played on the same or a different $F0$. We use the within-subject standard deviation of matches as a measure of the difficulty of the task.

The first experiment, which establishes the viability of the technique, uses two-formant vowels, with subjects adjusting either the first or the second formant. The sounds that we have used last 500 ms. We use relatively long vowels in order to improve our chances of finding effects due to the dynamic changes in $F0$ that are introduced in experiments 3 and 4. Experiment 2 compares matching reliability for single-formant sounds when the target and the adjustable sounds have either the same $F0$ or different $F0$'s and when the target formant frequency is in a region of resolved harmonics (high $F0$) or of unresolved harmonics (low $F0$). Experiments 3 and 4 ask whether a sinusoidally modulated $F0$ increases the reliability of matching.

I. EXPERIMENT 1

A. Stimuli and procedure

On each trial subjects heard two 500-ms sounds: a target sound followed after 500 ms by an adjustable sound. They adjusted the frequency of one formant of a periodically excited two-formant complex to match the timbre of the similar target sound, by moving a roller-ball up or down. Instructions given to subjects were to match the timbre of the sounds, in other words to try to get the same quality of sound between target and adjustable sound. The adjustable sound could have either the same or a different fundamental frequency from the target ($F0$ values: 90, 120, 150 Hz). The pair of sounds could be repeated as often as necessary on each trial by pressing the roller-ball's button.

In experiment 1a, the second formant ($F2$) of both sounds was kept constant at 2100 Hz (bandwidth 200 Hz) and the target's first formant ($F1$) could be either 400, 550, or 700 Hz (bandwidth 100 Hz). In experiment 1b, $F1$ was fixed at 550 Hz and the target's $F2$ could be either 1500, 2100, or 2600 Hz. Within a block of trials the target could have either of the three formant values, but the $F0$ of the adjustable sound, and whether the target had the same $F0$ as the adjustable sound or a different $F0$ was fixed. When the $F0$'s were different, the $F0$ of the target sound also varied within a block of trials. Consequently, there were three different targets in blocks where the $F0$ was the same, and six

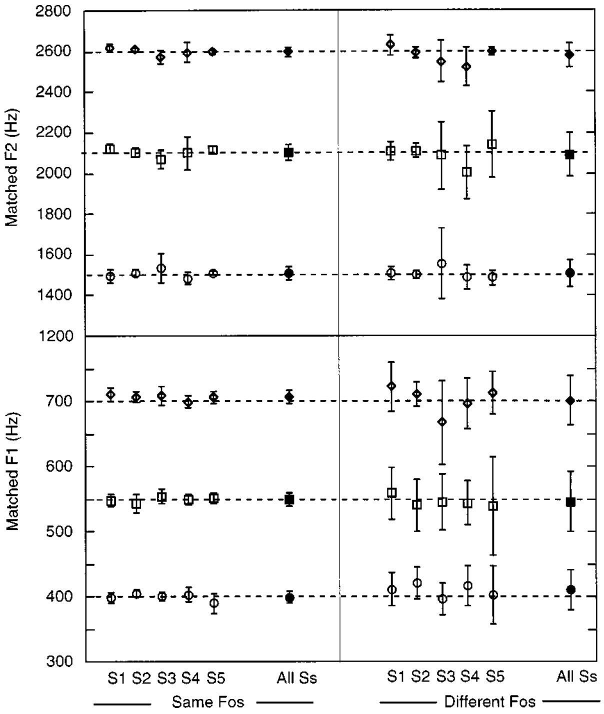

FIG. 1. Open symbols represent mean matches with standard deviations across five replications for individual subjects in experiment 1a ($F1$ matching), lower panel, and in experiment 1b ($F2$ matching), upper panel. Closed symbols represent average matches across subjects together with the average within-subject standard deviation.

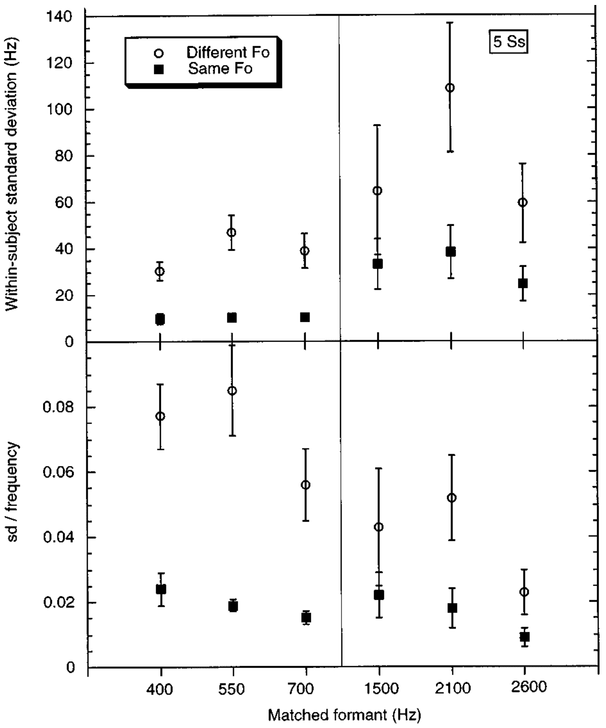

FIG. 2. The upper panel shows the mean within-subject standard deviations (and their across-subject standard errors) for formant matches in experiment 1. The lower panel plots the same data as a fraction of the center frequency of the matched formant.

different targets in blocks that had different $F0$'s. Each target sound was matched five times in a quasi-random order. The order of the 12 experimental blocks were randomized across subjects so that each block appeared equally in each serial position across subjects.

Sounds were synthesized in real time at 22.05 kHz using the parallel branch of SenSyn PPC™ (Sensimetrics, Cambridge, MA) incorporated into custom software. Formant amplitudes were set equal in the SenSyn tables and the overall output level was around 60 dB SPL. Voice source parameters were set to their default values, which are the same as described in Klatt (1980). Sounds were output through a Digidesign Protools board and presented through Sennheiser HD414 headphones in an IAC booth. An Apple Power Macintosh 7100 computer controlled the experiment.

At the beginning of each trial the formant frequency of the adjustable sound was chosen at random from the permitted range (150 to 850 Hz in experiment 1a, 1000 to 3100 Hz in experiment 1b). As subjects moved the roller-ball to adjust the formant frequency of the comparison sound a screen cursor also moved. The cursor was recentered after each sound pair so that subjects could not base their adjustment on the cursor's position. In experiment 1a, moving the cursor by half a screen led to a change of about 33 Hz (fine adjustment) or 100 Hz (coarse adjustment). In experiment 1b, the adjustment was either 33 Hz (fine) or 310 Hz (coarse). Subjects could toggle between the coarse and fine adjustments. If the formant frequency was adjusted outside the permitted range, it was reset to a random value within the range and a warning sound played.

Six subjects (including the two authors) participated in the experiment. Subjects were university students or staff and were paid for their services. All had pure-tone thresholds within the normal range at octave frequencies between 250 Hz and 4 kHz. Subjects were introduced to the task by undertaking between 10 and 40 trials with the experimenter providing feedback to ensure that they understood the task. One subject's data were removed because of grossly inconsistent matches for sounds that had the same $F0$.

B. Results

Average matches across five replications for individual subjects and their within-subject standard deviation are shown by the open symbols in Fig. 1. Within-subject standard deviations were calculated for each target/adjustable sound condition across the five replications. These standard deviations were then averaged across conditions to give the values plotted in Fig. 1. The filled symbols show the mean across all five subjects together with the average within-subject standard deviation.

Averaging across subjects, mean matches correspond closely to their target values, with no systematic differences whether or not the sounds differ in $F0$. However, as is apparent from the within-subject standard deviation error bars in Fig. 1, matches are more variable when the sounds have different $F0$'s than when they have the same $F0$ [$F(1,4) = 16.5$, $p < 0.02$]. These standard deviations are themselves plotted in Fig. 2 both as raw (Hz) values (upper panel) and as a proportion of the target formant value (lower panel). The increase in variability when the sounds have different $F0$'s rather than the same $F0$ is significant both for the raw

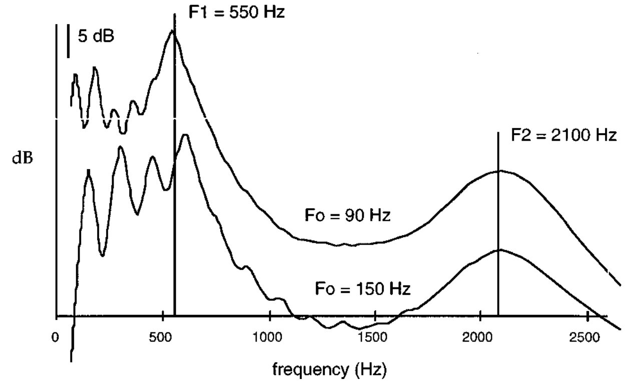

FIG. 3. Excitation patterns for two of the target sounds in experiment 1. Both sounds $F1$ at 550 Hz and $F2$ at 2100 Hz (marked by vertical lines). The upper curve has a fundamental frequency of 90 Hz, the bottom of 150 Hz. The curves are vertically displaced for clarity.

$[F(1,4)=16.5,\ p<0.02]$ and proportional scores $[F(1,4)=31.8,\ p<0.005]$.

Figure 2 also shows that a difference in $F0$ increases the proportional within-subject variability more for $F1$ matches than for $F2$ $[F(1,4)=20.4,\ p=0.01]$. The size of the $F0$ difference (i.e., 30 or 60 Hz) had no significant effect on match variability.

C. Discussion

The first experiment has shown that subjects can make formant matches across two-formant vowellike sounds on different fundamentals. However, the within-subject reliability of these matches is less than those made across sounds that have the same $F0$. Moreover, the difference in $F0$ impairs performance more for the $F1$ matches than for the $F2$ matches.

This pattern of results can be explained by considering the excitation pattern produced by the sounds we have used. When matches are being made between sounds that have the *same* $F0$, listeners do not need to extract the spectral envelope, but can simply match the absolute or relative levels of the excitation pattern at corresponding frequencies around the formant frequency. The resulting matches are accurate and show low within-subject variability. The variability of matches for sounds on the same $F0$ is comparable with discrimination threshold data from Lyzenga and Horst (1998). They found jnd's (equivalent to 63% correct and using a roving level) of 1%–3% for discrimination of changes to $F1$ in the presence of a static $F2$ (at around 2 kHz) on a fundamental of 100 Hz.

Placing sounds on *different* fundamentals complicates the listener's task in two ways. First, the sounds can no longer be adjusted to be identical: there is always a difference in pitch which can distract listeners from their timbral judgements. This complication applies to both the $F1$ matches and the $F2$ matches.

Second, for the $F1$ matches, where the harmonics close to the matched formant frequency are resolved by the cochlea, the excitation pattern shows marked variations with $F0$ around the formant frequency (corresponding to individual harmonic peaks). Figure 3 shows excitation patterns (Moore and Glasberg, 1983) for sounds with $F1$ at 550 Hz and $F2$ at 2100 Hz on $F0$ of either 90 or 150 Hz. Since the excitation pattern in the region of $F1$ is very different for sounds on different $F0$'s subjects cannot make veridical formant matches by simply matching relative levels of the excitation pattern.[1] However, a more abstract representation, such as the spectral envelope, would allow veridical matches to be made. On the other hand, for the $F2$ matches, where the harmonics close to the matched formant are *not* resolved, the excitation pattern varies little with $F0$ around the formant peak and so can be used as an explicit basis for the perceptual match.

This analysis provides a natural explanation for the changes that we have observed in within-subject variability. For both the $F1$ and $F2$ matches, variability increases when the matches are made between sounds with different $F0$'s primarily because of the distracting effect of a difference in $F0$. For the $F1$ matches, variability increases further because of the need for listeners to use a more abstract representation than explicit properties of the excitation pattern to match the formant frequency.

The above analysis is weakened by the fact that the comparison of resolved and unresolved harmonics is confounded with the different spectral regions of formants $F1$ and $F2$. It is also the case that part of the increase in variability produced by a difference in $F0$ could be due to vowel quality changes that cannot be compensated for by changing the single formant over which the subject has control. The next experiment uses simpler, single-formant stimuli which allow a more direct comparison of the effect of a difference in $F0$ for resolved and unresolved harmonics.

II. EXPERIMENT 2

Experiment 2 uses single-formant sounds and a slightly different experimental design to investigate whether the increased difficulty of making matches on different $F0$'s is greater when there are resolved harmonics in the region of the formant (for a formant on a high $F0$) than it is with unresolved harmonics in the region of the formant (with a low $F0$). A single formant at around 1100 Hz is used since at this frequency $F0$'s within a normal vocal range generate harmonics around the formant peak that are either clearly resolved ($F0$ around 250 Hz) or unresolved ($F0$ around 80 Hz). Figure 4 shows excitation patterns (Moore and Glasberg, 1983) for a single formant at 1100 Hz on two different fundamentals: 80 and 250 Hz. The harmonic ripple is not evident around the formant frequency for the low fundamental, but is clearly present for the high fundamental.

A. Stimuli and procedure

Listeners had to adjust the formant frequency of a periodically excited single-formant complex sound to match the timbre of a similar sound with a formant frequency of either 1100 or 1200 Hz (bandwidth 100 Hz). Three factors were varied orthogonally across 8 blocks of 20 trials: whether both sounds in a trial were from the low (80 or 90.4 Hz) or the high (221.2 or 250 Hz) $F0$ range, whether they had the same or a different $F0$ and whether the target sound had an $F0$ that was the higher or lower value in its range. On each trial

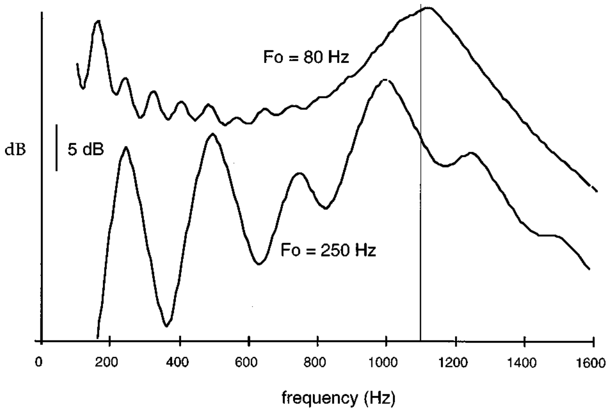

FIG. 4. Excitation patterns for two of the target sounds in experiment 2. Both sounds have a single formant at 1100 Hz (marked by the vertical line). The upper curve has a fundamental frequency of 80 Hz, the bottom of 250 Hz. The curves are vertically displaced for clarity.

the two sounds' fundamental frequencies $(F0)$ were in the ratio 1:1.13. The specific values of $F0$ for each block of trials are shown in Fig. 5. Within blocks the target formant frequency was randomly either 1100 or 1200 Hz. Each target sound was matched ten times in a quasi-random order; the number of matches per condition was increased from that used in experiment 1 in order to increase the reliability of estimates of within-subject variability. Each block took about 30 min to complete and the order of experimental blocks was randomized across subjects.

Stimulus synthesis and presentation were similar to experiment 1a except that the permitted adjustment range was 800–1500 Hz. Eleven subjects (including the two authors), who were university students or staff, participated in the experiment. All had pure-tone thresholds within the normal range at octave frequencies between 250 Hz and 4 kHz. For each subject, matches that deviated from the mean by more than two standard deviations were classified as errors and were ignored. Such matches amounted to less than 3% of the total in this and subsequent experiments.

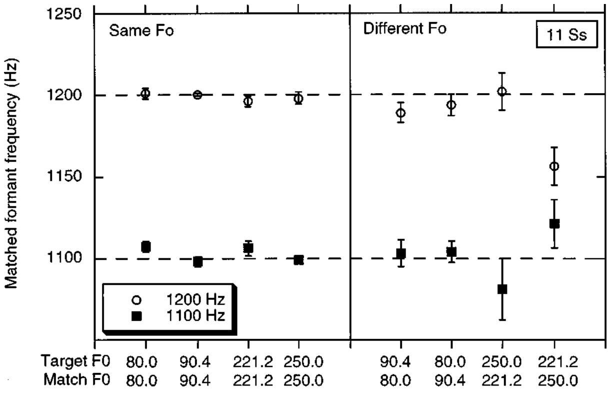

FIG. 5. Mean matched target formant frequencies in experiment 2 for target formant frequencies of 1100 and 1200 Hz. Error bars are standard errors of the mean over 11 subjects. The left panel shows matches made when the target and match had the same $F0$, the right-hand panel shows matches when they had different $F0$'s.

TABLE I. Dominant harmonic and estimated formant frequencies.

	Dominant harmonic		Burg-estimated formant	
Formant	$F0=221.2$	$F0=250$	$F0=221.2$	$F0=250$
1100	**1106**	*1000*	1092	1050
1200	*1106*	**1250**	1166	1210

B. Results and discussion

1. Mean matches

The mean matched formant frequencies for each condition across the 11 subjects are shown in Fig. 5 together with their standard error. Matches made on the same $F0$ are not systematically different from the target formant frequency and are very consistent across subjects.

The matches to sounds on different $F0$'s are also close to the target values when the $F0$'s are low (with unresolved harmonics near the formant peak), but there are significant discrepancies between the target and matched formant frequencies when the $F0$ is high, as seen in the four right-most data points in Fig. 5. The interaction is significant both as a four-way interaction involving all the data in the figure $[F(1,10)=9.9,\ p=0.01]$ and as a simple interaction involving only the four right-most data points $[F(1,10)=9.1,\ p<0.02]$. This interaction between target formant frequency and the relative $F0$ of the target and the match, which is confined to those conditions where the target and the match are on different $F0$s and the individual harmonics are resolved, can be interpreted by examining how individual harmonics align with the formant frequency. For our target sounds, the harmonics close to the formant peak had an asymmetric level distribution, which varied depending on the particular $F0$'s and formant frequencies used. For some sounds the most intense harmonic was lower in frequency than the formant frequency, for others it was higher.

Table I shows the dominant (highest amplitude) harmonic for each of the four combinations of $F0$ and formant frequency in question. Harmonic frequencies in bold are above the formant frequency, those in *italics* are below the formant frequency. For example, for a 1200-Hz formant, the dominant harmonic from a 221.2-Hz $F0$ is at 1106, considerably below the true formant frequency. The interaction in the data could be due to listeners tending to hear the formant frequency as displaced in the direction of the dominant harmonic.

Klatt (1986) has pointed out that various methods of estimating formant frequencies such as spectral smoothing and LPC analysis are also prone to a similar displacement. That this is true for our stimuli is shown in Table I using LPC analysis as an example. This table gives formant frequency estimates of our stimuli using an implementation (Press, 1993) of the Burg method (identical results were obtained using an LPC covariance method). With these methods of estimating formant frequency an increase in $F0$ from 221.2 to 250 Hz decreases the estimated frequency for 1100 Hz by about 40 Hz but increases it by about the same amount for 1200 Hz. These shifts are similar to the perceptual data. Although this explanation is couched in terms of listeners

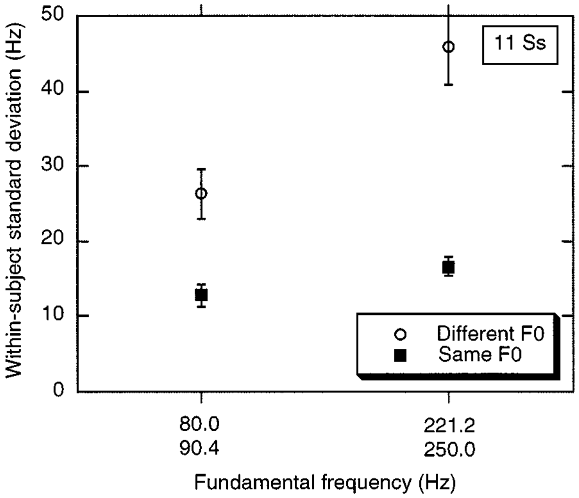

FIG. 6. Average within-subject standard deviations of matches in experiment 2 with their standard errors over 11 subjects.

making explicit formant frequency comparisons, it is important to reiterate that the demands of this experiment do not force listeners to use an explicit formant frequency, rather than the spectral envelope. A similar explanation could be couched in terms of the shape of the spectral envelope derived by spectral smoothing or by LPC analysis.

2. Match variability

Figure 6 shows the mean within-subject standard deviation of matches within each block of trials, together with the standard error of these means across the 11 subjects. Within-subject variability was not significantly different between the two matched formant values or between different values of $F0$ within an $F0$ range,[2] so Fig. 6 pools data across these dimensions. As in Experiment 1, while within-subject variability of matching was low for sounds on the same $F0$, it increased when the adjustable sound's $F0$ was different from the target's. This increase was significantly larger for sounds on high $F0$'s (with resolved harmonics) than for those on low $F0$'s (with unresolved harmonics) [$F(1,10)=7.1$, $p<0.025$].

These results have confirmed the main findings from experiment 1 using single-formant sounds. When the target and match sounds are on the same $F0$, formant frequency matches are on average veridical with low within-subject variability whether both sounds are presented on a low or a high $F0$. When the target and match sounds are on different $F0$'s, the variability of matching varies with the range of $F0$. On low $F0$'s variability increases significantly from the case where $F0$'s are the same. On high $F0$'s however, there is a significantly larger increase in within-subject variability. This interaction can be attributed to the need for listeners to interpolate a spectral envelope when the individual harmonics near the formant peak are resolved.

In addition, this experiment has shown that when individual harmonics near the formant peak are not resolved, the mean matches to sounds on different $F0$'s deviate significantly from veridicality. This result is explicable if listeners tend to perceive the formant peak as shifted towards the dominant harmonic—a shift which is also seen in formants estimated from our sounds by the Burg or LPC covariance methods.

3. Testing an excitation pattern match model of matching

We have interpreted the results of this experiment and experiment 1 on the assumption that listeners cannot perform the matching task for resolved harmonics on different $F0$'s on the basis of explicit properties of the excitation pattern. The following section tests this assumption.

It is possible that listeners might solve the formant matching task even when sounds are on different fundamentals by minimizing the discrepancy between the excitation pattern of the target vowel and that of the adjustable vowel as suggested for same-$F0$ formant discrimination by Sommers *et al.* (1996). We tested whether this strategy is feasible by measuring the rms error (in units of log power) between excitation patterns for various combinations of single-formant sounds. Excitation patterns were calculated using the formula proposed by Moore and Glasberg (1983), using 44 channels spaced at 3% frequency increments from 500 to 1836 Hz. The upper panel of Fig. 7 shows the rms dB error between a fixed formant at either 1100 or 1200 Hz and $F0$ of 250 Hz, and a variable formant with an $F0$ of 221.2 Hz whose frequency is given on the abscissa. The middle panel is similar except the $F0$'s of the fixed and variable formants are reversed. In the lower panel the $F0$'s of the two formants are the same at 250 Hz. The figure also shows the mean matched position for each condition from the experimental data, together with error bars that show the average within-subject standard deviation.

As one would expect, when there is no difference in fundamental (in the bottom panel), there is a clear minimum in the rms difference score, which corresponds with the matched value. However, when there *is* a difference in fundamental (top two panels), the minima are much shallower and the actual minima correspond neither to the true (synthesized) formant values nor to the experimental matched ones.

Let us assume that in the control condition (for which the $F0$'s are the same) listeners are adopting a strategy of finding the minimum error in the excitation pattern. Let us also assume that the variability of listeners' performance on the matching task is limited by the accuracy with which they estimate the average rms difference. If we take an accuracy of 1 dB as an arbitrary value, we can then covert this into a measure of the accuracy with which the minimum is defined. In the bottom panel, each curve remains within 1 dB of the minimum value over an average range of about 45 Hz. This value corresponds to about 1.4 times ±1 standard deviation of the matching scores. If listeners are employing a similar strategy in matching the conditions with different $F0$'s, we would expect this ratio to remain constant. In fact, the corresponding figure for the upper panel is 200 Hz corresponding to 2.6 times ±1 standard deviation of the appropriate matching scores, and for the middle panel 280 Hz, again corresponding to 2.6 times ±1 standard deviation of the appropriate matching scores.

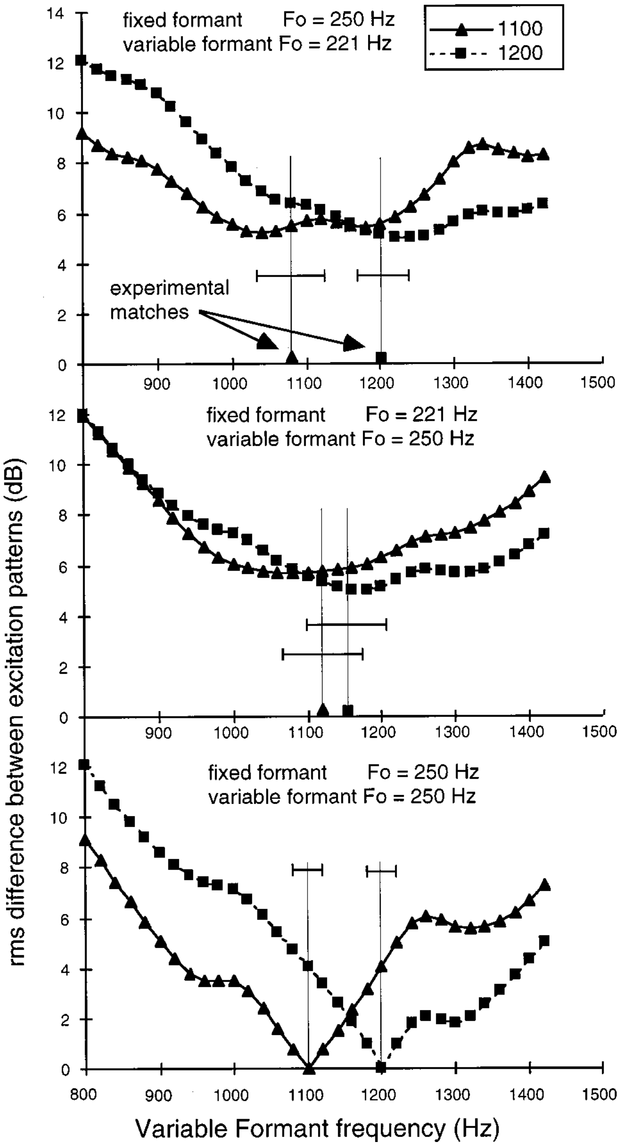

FIG. 7. The rms difference in dB between excitation patterns of single-formant sounds. In the upper panel, a fixed formant at either 1100 or 1200 Hz and $F0$ of 250 Hz is compared with a variable formant with an $F0$ of 221.2 Hz whose frequency is given on the abscissa. The middle panel is similar except the $F0$'s of the fixed and variable formants are reversed. In the lower panel the $F0$'s of the two formants are the same at 250 Hz. The vertical lines show the mean matched position (±1 within-subject s.d.) for each condition from the experimental data.

Therefore, with different $F0$'s, subjects do better by a factor of about 2 than predicted by this excitation pattern model from their performance on the same $F0$. Since it is also likely that some aspect of their absolutely inferior performance with different $F0$'s arises because of the distracting effect of a difference in pitch, a direct comparison of excitation patterns does not provide an adequate model for explaining listeners' performance on this task.

III. EXPERIMENT 3

The aim of experiment 3 is to examine whether adding frequency modulation (FM) to $F0$ changes the accuracy and veridicality of formant-frequency matches.

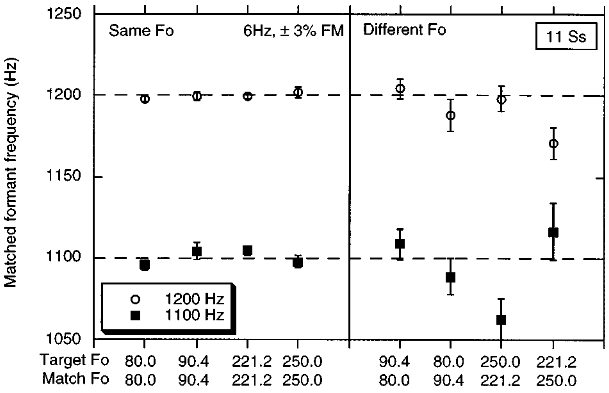

FIG. 8. Mean matched formant frequencies with standard errors over 11 subjects in experiment 3 for stimuli with frequency-modulated $F0$'s.

In our experimental paradigm we look for two possible effects of vibrato on formant matching. First, vibrato might generally make the task harder since it replaces a steady sound with one that is changing in an irrelevant dimension ($F0$). If this factor is at work, then matches ought to be more variable with FM than with a steady $F0$ in all conditions, but in particular, matches on the *same* $F0$ might show larger within-subject variation with vibrato than without. Second, if the spectral-envelope tracing produced by vibrato is useful to listeners, then different-$F0$ matches on high $F0$'s would be both more veridical and show less within-subject variability with vibrato than without.

A. Stimuli and procedure

The stimuli and procedure were similar to experiment 2 except that the steady $F0$'s of experiment 2 were replaced with $F0$'s that were frequency-modulated at 6 Hz and ±3% depth. The same 11 subjects participated in this experiment after they had taken experiment 2.

B. Results

1. Mean matches

The mean matched formant frequencies are show in Fig. 8. The overall pattern of the data is very similar to that from the previous experiment without FM. Matches on the same $F0$ are veridical and consistent across listeners. Matches on different $F0$'s are also close to the target values when the $F0$'s are low (with harmonics near the formant peak unresolved), but there are again significant discrepancies between the target and matched formant frequencies when the $F0$ is high. The pattern of this interaction is identical to that in the previous experiment and is again significant both as a four-way interaction involving all the data in Fig. 8 [$F(1,10) = 37.4$, $p = 0.001$] and as a simple interaction involving only the four right-most data points [$F(1,10) = 14.8$, $p < 0.005$].

2. Match variability

The average within-subject standard deviations of matches are shown in Fig. 9. Matches to sounds on the same $F0$ show similar variability in this experiment to those in the previous one where there was no FM. It is thus unlikely that FM is distracting listeners from the matching task. However,

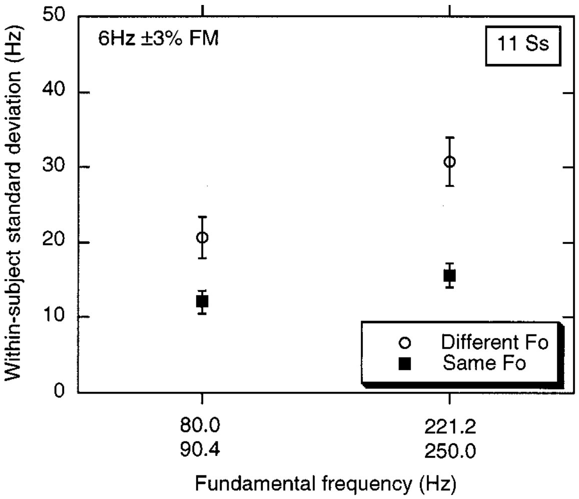

FIG. 9. Average within-subject standard deviations of matches in experiment 3 with standard errors over 11 subjects.

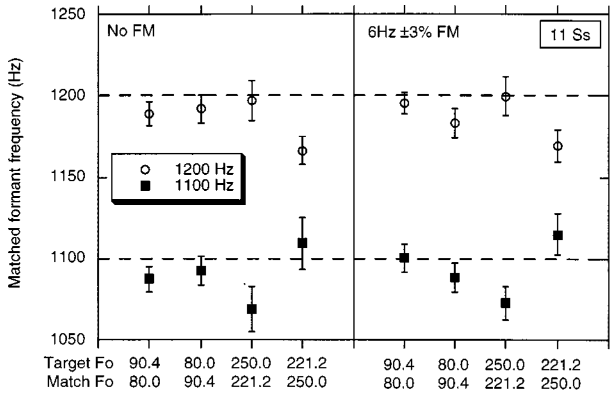

FIG. 10. Mean matched formant frequencies with standard errors over 11 subjects in experiment 4 for stimuli with or without frequency-modulated $F0$'s.

the variability of matches with different $F0$'s is lower in this experiment, giving a marginally significant interaction between the two experiments and same/different $F0$ [$F(1,10) = 4.7$, $p = 0.06$]. Although this reduction in variability of the more difficult conditions between the two experiments could be due to the introduction of FM, it could also be due to increasing practice of listeners on the task, since this experiment was conducted subsequent to experiment 2 and on the same listeners. The next experiment removes this ambiguity.

IV. EXPERIMENT 4

The previous experiment showed a marginally significant decrease of within-subject variability for matches to sounds on different, high $F0$'s when the $F0$ is modulated. To show that this increased consistency with FM is not due to subjects being more practised, experiment 4 presents the different-$F0$ conditions used in experiments 2 (no FM) and 3 (FM) in counter-balanced order.

A. Stimuli and procedure

The different-$F0$ conditions from experiments 2 and 3 were used, with the order of blocks with FM or no FM counterbalanced across subjects. The 11 subjects from experiments 2 and 3 participated after they had taken experiment 3.

B. Results and discussion

All the matches in this experiment were done on sounds that had different $F0$'s. The mean matched formant frequencies for all conditions and their standard errors across 11 subjects are shown in Fig. 10. The pattern of mean matches is very similar whether the sounds had frequency modulation or not and replicates the pattern seen in the different-$F0$ conditions of experiments 2 and 3. For matches made on a low $F0$, giving unresolved harmonics in the region of the formant, matches are generally veridical and are little affected by whether the target or the match has the higher $F0$. On the other hand, matches made on a high $F0$, are less veridical and again interact with the relative $F0$ of the target and mask. The three-way interaction between $F0$ range, target formant frequency, and specific $F0$ values is highly significant [$F(1,10) = 35.1$, $p < 0.001$], but does not interact with whether the $F0$ has FM or not [$F(1,10) = 0.01$].

In contrast to the absence of any effect of FM on the mean matched formant frequencies, FM reduces the within-subject variability of matches made on high $F0$'s (resolved harmonics), compared with those made on low $F0$'s (unresolved harmonics). This significant interaction [$F(1,10) = 5.8$, $p < 0.05$] is illustrated in Fig. 11 and shows that the corresponding difference between experiments 2 and 3 was not simply due to the fact that subjects took experiment 3 after experiment 2 and were therefore more practised at the task. The mere presence of FM does not appear to have a general distracting effect, since even in the low-$F0$ condition, FM matches are not significantly more variable than are the no-FM matches.

In summary, this experiment provides direct evidence that frequency-modulating $F0$ improves listeners' reliability at matching a formant frequency that is cued by resolved harmonics. However, although matches were less variable with FM, they were no more veridical: the tendency for matches with unresolved harmonics to be influenced by the

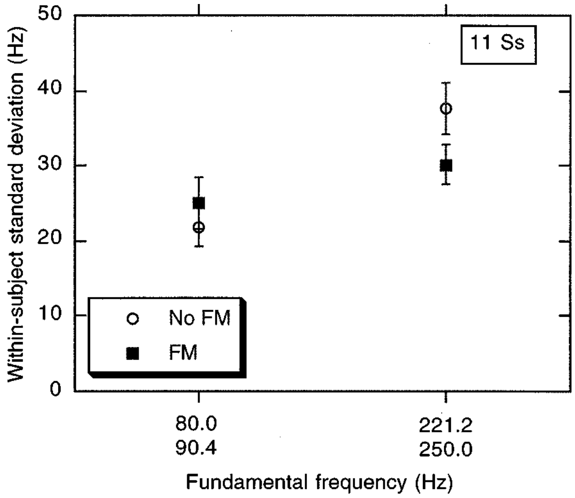

FIG. 11. Average within-subject standard deviations of matches in experiment 4 with standard errors over 11 subjects for stimuli matched on different $F0$'s, with or without FM.

frequency of the dominant harmonic is just as strong with FM as without it.

V. SUMMARY AND GENERAL DISCUSSION

The four experiments described here have explored how a difference in $F0$ affects listeners' ability to match sounds according to formant frequency. When the individual harmonics close to a formant peak are resolved by the cochlea, and so produce individual peaks in the excitation pattern, a difference in $F0$ prevents listeners from performing the matching task by simply comparing across corresponding frequencies the absolute or relative levels of the excitation patterns produced by the target and adjustable sounds—a strategy which accounts for much of the data on formant-frequency discrimination.

The main result is that a difference in $F0$ increases the variability of matches for sounds with resolved harmonics more than it increases the variability of those with unresolved harmonics. This increase in variability is probably due to the need to interpolate a spectral envelope in order to perform the task with resolved harmonics but not with unresolved. It is not necessary for listeners to explicitly extract the formant frequency in order to perform the task. We plan to report in a subsequent paper experiments which require listeners to match sounds which also differ in bandwidth. The bandwidth difference prevents listeners from performing the matching task on the basis of the (interpolated) spectral envelope, thereby forcing them to use a more abstract representation such as the envelope peak (formant frequency).

An unexpected finding was the way that small differences in $F0$ affected the mean values of the matched formant frequencies with resolved harmonics. It is probable from experiments on vowel identification that have manipulated the amplitude of individual resolved harmonics close to the $F1$ frequency (Carlson *et al.*, 1975; Darwin and Gardner, 1985) that the perceived formant frequency is influenced by the amplitude of the small number of resolved harmonics close to the formant peak—although it has been claimed that it is simply determined by the frequency of the harmonic closest to the formant frequency (Mushnikov and Chistovich, 1972). The effect that we found could be explained by listeners' formant frequency estimates (or spectral envelope interpolations) being unduly influenced by the dominant harmonic in the spectrum. A similar effect was also shown by automatic formant frequency estimation using the Burg method.

The experiments have also shown that giving the fundamental frequency vibratolike FM reduces the variability of matches. The matches carried out with FM, however, still demonstrate the same effect of the dominant harmonic as those without FM. This failure of FM to produce more veridical matches (as defined by the synthesis procedure) argues against listeners being sensitive to the spectral envelope tracing produced by FM, but further experiments using deeper modulation and more systematic variation of the alignment of harmonic frequencies and formant peaks are needed to test this speculation. Given the lack of any change in the veridicality of matches produced under FM, the reason for the reduced variability of FM matches may be more to do with the greater perceptual integrity of the timbral percept from the $F0$-modulated vowels than from the steady-state ones. FM has been shown to increase perceptual integration for harmonically related sounds (McAdams, 1989; Darwin *et al.*, 1994) even though differences in FM cannot be used to segregate different sound sources (Gardner and Darwin, 1986; Gardner *et al.*, 1989; Carlyon, 1991; Culling and Summerfield, 1995).

In summary, the formant-matching task introduced here has been shown to be sensitive to the increased perceptual difficulty of extracting a spectral envelope from resolved rather than unresolved harmonics. In addition, it has demonstrated a beneficial effect of FM in reducing the variability of subjects' formant-frequency matches. It therefore provides a tool for further investigation of the intermediate representations that listeners employ for complex sounds. The experiments have also demonstrated a specific effect of the alignment of harmonics with a formant peak on subjects' perception of the formant. Our data are compatible with shifts of the perceived formant frequency towards the strongest resolved harmonic—an effect also shown in automatic methods of formant extraction.

ACKNOWLEDGMENTS

P. Disscard was supported by Grant No. GR/L03422 from the UK EPSRC. Paul Russell wrote the software for formant matching based on SenSyn for the PowerPC. We are grateful to Rob Hukin for technical assistance, to Dr. George Mather for discussions on vernier acuity, to Peter Assmann and another anonymous reviewer for numerous helpful suggestions, and to Oded Ghitza for drawing our attention to Dennis Klatt's 1986 paper.

[1] This assertion is tested in the context of experiment 2.

[2] The largest within-subject standard deviations for matches on the same $F0$ in Fig. 6 occur for the two conditions where a harmonic frequency is closest to the formant frequency (1100/221.2 and 1200/250). In discrimination experiments, formant frequency thresholds are also larger when a harmonic frequency is close to a formant frequency (Kewley-Port and Watson, 1994; Lyzenga and Horst, 1995).

Assmann, P. F., and Nearey, T. M. (**1987**). ''Perception of front vowels: the role of harmonics in the first formant region,'' J. Acoust. Soc. Am. **81**, 520–534.

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

Carlyon, R. P. (**1991**). ''Discriminating between coherent and incoherent frequency modulation of complex tones,'' J. Acoust. Soc. Am. **89**, 329–340.

Culling, J. F., and Summerfield, Q. (**1995**). ''The role of frequency modulation in the perceptual segregation of concurrent vowels,'' J. Acoust. Soc. Am. **98**, 837–846.

Darwin, C. J., and Gardner, R. B. (**1985**). ''Which harmonics contribute to the estimation of the first formant?'' Speech Commun. **4**, 231–235.

Darwin, C. J., Ciocca, V., and Sandell, G. R. (**1994**). ''Effects of frequency and amplitude modulation on the pitch of a complex tone with a mistuned harmonic,'' J. Acoust. Soc. Am. **95**, 2631–2636.

Demany, L., and Semal, C. (**1990**). ''The effect of vibrato on the recognition of masked vowels,'' Percept. Psychophys. **48**, 436–444.

Florén, Å. (**1979**). ''Why does [a] change to [ɔ] when $F0$ is increased?: Interplay between harmonic structure and formant frequency in the perception of vowel quality,'' PERILUS (Institute of Linguistics, University of Stockholm) **1**, 13–23.

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

Gardner, R. B., Gaskill, S. A., and Darwin, C. J. (**1989**). ‘‘Perceptual grouping of formants with static and dynamic differences in fundamental frequency,’’ J. Acoust. Soc. Am. **85**, 1329–1337.

Green, D. M. **(1988)**. *Profile Analysis* (Oxford U.P., New York).

Hillenbrand, J., and Nearey, T. M. (**1999**). ‘‘Identification of resynthesised /hVd/ utterances: Effects of formant contour,’’ J. Acoust. Soc. Am. **105**, 3509–3523.

Kewley-Port, D. (**1995**). ‘‘Thresholds for formant-frequency discrimination of vowels in consonantal context,’’ J. Acoust. Soc. Am. **97**, 3139–3146.

Kewley-Port, D., and Watson, C. S. (**1994**). ‘‘Formant-frequency discrimination for isolated english vowels,’’ J. Acoust. Soc. Am. **95**, 485–496.

Kewley-Port, D., Li, X. F., Zheng, Y. J., and Neel, A. T. (**1996**). ‘‘Fundamental-frequency effects on thresholds for vowel formant discrimination,’’ J. Acoust. Soc. Am. **100**, 2462–2470.

Klatt, D. H. (**1980**). ‘‘Software for a cascade/parallel formant synthesizer,’’ J. Acoust. Soc. Am. **67**, 971–995.

Klatt, D. H. (**1985**). ‘‘The perceptual reality of a formant frequency,’’ J. Acoust. Soc. Am. **78**, S81–S82.

Klatt, D. H. **(1986)**. ‘‘Representation of the first formant in speech recognition and in models of the auditory periphery,’’ in Proceedings of the Montreal symposium on speech recognition; McGill University, Montreal, pp. 5–7.

Kontsevich, L. L., and Tyler, C. W. (**1998**). ‘‘How much of the visual object is used in estimating its position?’’ Vision Res. **38**, 3025–3029.

Lyzenga, J., and Horst, J. W. (**1995**). ‘‘Frequency discrimination of band-limited harmonic complexes related to vowel formants,’’ J. Acoust. Soc. Am. **98**, 1943–1955.

Lyzenga, J., and Horst, J. W. (**1997**). ‘‘Frequency discrimination of stylized synthetic vowels with a single formant,’’ J. Acoust. Soc. Am. **102**, 1755–1767.

Lyzenga, J., and Horst, J. W. (**1998**). ‘‘Frequency discrimination of stylized synthetic vowels with two formants,’’ J. Acoust. Soc. Am. **104**, 2956–2966.

Marin, C. M. H., and McAdams, S. (**1991**). ‘‘Segregation of concurrent sounds. II. Effects of spectral envelope tracing, frequency-modulation coherence, and frequency-modulation width,’’ J. Acoust. Soc. Am. **89**, 341–351.

McAdams, S. (**1989**). ‘‘Segregation of concurrent sounds. 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. Jorst, and H. P. Wit (Academic, London)

Moore, B. C. J., and Glasberg, B. R. (**1983**). ‘‘Suggested formulae for calculating auditory-filter bandwidths and excitation patterns,’’ J. Acoust. Soc. Am. **74**, 750–753.

Morgan, M. J., and Watt, R. J. (**1982**). ‘‘Mechanisms of interpolation in human spatial vision,’’ Nature (London) **299**, 553–555.

Mushnikov, V. N., and Chistovich, L. A. (**1972**). ‘‘Method for the experimental investigation of the role of component loudness in the recognition of a vowel,’’ Sov. Phys. Acoust. **17**, 339–344.

Press, W. H., Teukolsky, S. A., Vettering, W. T., and Flannery, B. R. **(1993)**. *Numerical recipes in C* (Cambridge C.U.P.).

Ryalls, J. H., and Lieberman, P. (**1982**). ‘‘Fundamental frequency and vowel perception,’’ J. Acoust. Soc. Am. **72**, 1631–1634.

Sommers, M. S., and Kewley-Port, D. (**1996**). ‘‘Modeling formant frequency discrimination of female vowels,’’ J. Acoust. Soc. Am. **99**, 3770–3781.

Sundberg, J. (**1975**). ‘‘Vibrato and vowel identification,’’Speech Transmission Laboratory, Quartery Progress and Status Report (KTH, Stockholm) STL-QPSR **2-3**, 49–60.

Sundberg, J. (**1977**). ‘‘Vibrato and vowel identification,’’ Arch. Acoust. **2**, 257–266.