Perceptual and computational separation of simultaneous vowels: Cues arising from low-frequency beating

John F. Culling[a] and C. J. Darwin
Laboratory of Experimental Psychology, University of Sussex, Falmer, Brighton, E. Sussex BN1 9QG, England

(Received 12 January 1993; revised 26 July 1993; accepted 24 September 1993)

Identification of simultaneous speech sounds, such as pairs of steady-state vowels (double vowels), is more accurate when there is a difference in fundamental frequency (F_0). Accuracy of identification for double vowels increases with increasing F_0 difference (ΔF_0) asymptoting above 1 semitone. The experiment described here attempts to distinguish two mechanisms underlying this effect: first, perceptual separation by grouping together harmonic components of a common F_0; and, second, exploitation of the fluctuations in the spectral envelope of the composite stimulus that result from beating between unresolved components. The beating is mainly caused by interactions between corresponding harmonics of the two vowels with a small ΔF_0. Identification accuracy for normal, harmonically excited double vowels was compared with that for double vowels composed from the same components, but whose constituent vowels were excited by a mixture of the two harmonic series. These double vowels were designed to produce similar beating patterns to the normal double vowels. Both harmonically and inharmonically excited constituents improved identification with increasing ΔF_0, but the increase was larger for harmonically excited vowels. A computational model based upon psychophysical measurements of auditory frequency and temporal resolution correctly predicted an increase in accuracy of identification with increasing ΔF_0 which was attributable to beating. The results are interpreted in terms of a spectral change cue in the identification of double vowels with ΔF_0's which complements grouping by F_0, and which plays a dominant role for ΔF_0's smaller than 1 semitone.

PACS numbers: 43.71.Es, 43.66.Hg

INTRODUCTION

Speech is often heard against a background of other sounds, particularly competing speech sounds. Cherry (1953), proposed a number of cues to explain our ability to distinguish speech against a background of competing speech. Recent experiments have predominantly addressed the role of a difference in fundamental frequency (ΔF_0) between two simultaneous streams of speech. Direct evidence that ΔF_0 improves intelligibility of attended speech against a competing speech background comes from experiments by Brokx and Nooteboom (1982). In order to control the size of ΔF_0 available as a cue, they used continuous speech which was monotonically resynthesized from an LPC analysis. Subjects attended to one of two competing voices and reported the words of that voice more accurately the greater the ΔF_0 up to 2 semitones. For 12 semitones (1 octave) ΔF_0, performance was not better than for zero ΔF_0.

More tightly controlled, but less naturalistic experiments have shown that accuracy of identification for simultaneous synthetic vowels improves with increasing ΔF_0 between the two vowels (Scheffers, 1983; Zwicker, 1984; Chalikia and Bregman, 1989; Assmann and Summerfield, 1990). For double vowels of 200 ms or more, accuracy of identification (for both vowels correct) is well above chance when there is no ΔF_0, but rises sharply with only a $\frac{1}{4}$ or $\frac{1}{2}$ semitone ΔF_0, before asymptoting above 1 semitone. The results of previous experiments (Culling and Darwin, 1993) show that ΔF_0's in the first formant ($F1$) region are mainly responsible for this improvement in double-vowel identification with increasing ΔF_0. Since harmonics in the $F1$ region of a single vowel are usually well resolved, it might be thought that this effect of a ΔF_0 is mediated by a mechanism which selects resolved harmonics.

An harmonic sieve (Duifhuis *et al.*, 1982) could provide an appropriate mechanism for segregating the harmonics of two ΔF_0's (Scheffers, 1983). Scheffers' used a harmonic sieve to estimate the F_0's of two vowels and then sample the cochlear excitation pattern at harmonic intervals. There is perceptual evidence that mistuned harmonics can be excluded from vowel categorization (Darwin and Gardner, 1986; Roberts and Moore, 1990, 1991). However, both Scheffers (1983) and Assmann and Summerfield (1990, "place" model) found that a computational model of a harmonic sieve performed poorly in separating vowels and gave little improvement in identification accuracy as ΔF_0 increased from 0 to 4 semitones. On the other hand, models which exploit the periodicity within each frequency channel using autocorrelation (Assmann and Summerfield, 1990, "place-time" model; Meddis and Hewitt, 1992) can account for improvements in double-vowel identification with a ΔF_0 of only $\frac{1}{4}$ semitone. It is difficult to account for the failure of Scheffers' harmonic sieve model with ΔF_0's of 2 and 4 semitones, but the failure of Scheffers' model to

[a] New address: MRC Institute of Hearing Research, University of Nottingham, NG7 2RD, UK.

explain the improvement in double-vowel identification with a small ΔF_0 is probably due to the bandwidths of peripheral auditory filters. A semitone difference in frequency ($\cong$6%) at, for instance, 500 Hz is much smaller ($\cong$30 Hz) than the auditory filter bandwidth at that frequency ($\simeq$75 Hz). Consequently, the auditory system will inadequately resolve corresponding components from competing vowels that are separated by only 1 semitone. A harmonic sieve might provide some separation of the first formant region for vowel pairs with a ΔF_0 of a semitone if the amplitudes of the corresponding harmonics were sufficiently different to allow the F_0 of at least the dominant one to be estimated accurately; if each vowel dominates the other at the frequency of its first formant peak then the formants may be separable.

Perceptual evidence that two separate harmonic series cannot be extracted by the auditory system from vowel pairs that differ by a semitone or less comes from Summerfield (personal communication). He found that having successfully identified two simultaneous vowels subjects were poor at ranking their pitches; taking only those trials on which both vowels were correctly identified, subjects performed at chance (50%), when ranking the pitches of vowels which differed by $\frac{1}{4}$ semitone and correctly ranked less than 65% of pairs with $\frac{1}{2}$ and 1 semitone ΔF_0's. If vowels are separated by selecting components at multiples of each F_0, it is difficult to see why the listeners do not have more conscious access to the F_0 information that their auditory systems are using.

The present experiment investigated the possibility that the auditory system is employing a completely different cue at small ΔF_0's. Given two harmonic series with F_0's of 100 and 103 Hz, the 1st harmonics will beat together at 3 Hz, the 2nd harmonics at 6 Hz the 3rd harmonics at 9 Hz and so on. Since these modulation frequencies are integer multiples of the ΔF_0, the result is a cyclic pattern of "spectral modulation" with a fundamental frequency of modulation equal to the ΔF_0. The cyclic pattern of modulation will not systematically assist the perception of either vowel; the components of each vowel will beat at different rates and with independent phases, distorting the spectrum in various ways (see simulation in Fig. 4), but this distortion may, at some point in the cycle, make important features more prominent. Assmann and Summerfield (1990), attributed a *dip* in identification accuracy at $\frac{1}{4}$ semitone ΔF_0 for 50-ms double vowels to such waveform interaction, but at longer duration waveform interactions may be beneficial. There are two ways in which vowel pairs with small ΔF_0's may be identified more easily due to the resulting spectral modulation.

(1) The stimulus as a whole may sound more like one or other of the constituent vowels at different times: First like one vowel then like the other.

(2) Subjects can often identify both constituents of a double vowel from their combined spectrum when they are on the same F_0; when a small ΔF_0 causes this combined spectrum to undergo change, there may come a moment when the simultaneous identification task becomes particularly easy.

TABLE I. Formant frequencies and 3-dB bandwidths used to synthesize the five individual vowels.

	Vowel					
Formant	EE	AR	OO	ER	OR	Bandwidth
F1	250	650	250	450	350	90
F2	2250	950	850	1250	750	110
F3	3050	2950	1950	2650	2850	170
F4	3300	3300	3300	3300	3300	250
F5	3850	3850	3850	3850	3850	300

Although it is difficult to imagine a listener employing both strategies simultaneously, both strategies may be available to listeners. These possibilities are borne out to some extent by subjective impressions of the experimental materials, because they clearly have an unstable timbre. The following experiment tests whether such changes are responsible for the improvement in vowel identification.

I. EXPERIMENT

The experiment set out to test whether the amplitude modulation produced by the beating of corresponding harmonics is responsible for the increased identification at small ΔF_0's. Stimuli were devised whose frequency composition gave rise to similar patterns of spectral modulation cues to those produced by normal double vowels, but whose frequency composition was not harmonic for either vowel (each vowel was excited by a mixture of the two harmonic series) and so would mislead an harmonic selection mechanism, whether it is based purely on the excitation pattern, or also exploits temporal information. If subjects' identification accuracy increases as the ΔF_0 of these vowels is increased, the improvement cannot be attributed to a harmonic selection mechanism, but it could be due to exploitation of the spectral modulation.

A. Stimuli

The five British-English tense vowels (/ɑ/, /i/, /ɜ/, /u/, and /ɔ/; notated here as AR, EE, ER, OO, and OR) were synthesized using the same formant specifications (Table I) and the same six F_0's (100, 101.455, 102.930, 105.946, 112.246, and 125.992 Hz) as Assmann and Summerfield (1990). The vowels were 1000 ms in duration, including 10-ms raised cosine onset and offset ramps. They were synthesized with 12-bit quantization and with a sampling rate of 10 kHz. Six F_0's×5 vowels gave 30 vowels in a "set." As in Culling and Darwin (1993), the intensity of "AR" was reduced by 6 dB, in order to reduce the variation in level across vowels.

Three sets of single vowels were prepared, differing in the pattern of frequency components which sampled their spectral envelopes. One set, the normal vowels, used the same F_0 for all the components of a particular vowel (Fig. 1, rows 1 or 2). The spectrum of vowels in the second, inharmonic set consisted of the odd harmonics of 100 Hz F_0 and the even harmonics of the F_0 (Fig. 1, row 4). The spectra of the third, also inharmonic set consisted of the even harmonics of 100 Hz F_0 and the odd harmonics of the other F_0 (Fig. 1, row 5). Each component was added with

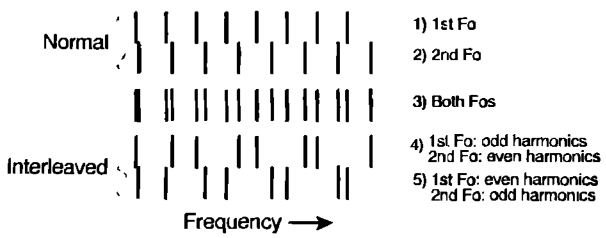
Normal
Interleaved
Frequency
1) 1st Fo
2) 2nd Fo
3) Both Fos
4) 1st Fo: odd harmonics 2nd Fo: even harmonics
5) 1st Fo: even harmonics 2nd Fo: odd harmonics

FIG. 1. The component structure of Interleaved ΔF_0's. Rows 1 and 2 show two different harmonics series; row 3 shows the two series in combination; row 4 shows the odd harmonics of the first F_0 and the even harmonics of the second F_0; row 5 shows the even harmonics of the first F_0 and the odd harmonics of the second F_0.

the amplitude and phase appropriate to a component of that frequency produced by a cascade formant synthesizer [Klatt, 1980, Eq. (6)].

Two types of double vowels, Normal and Interleaved, were created from the three sets of single vowels. Normal vowels were combined with other normal vowels to give a set of normal double vowels (Fig. 1, row 1+row 2). Vowels from the second and third sets, which had complementary component structures, were combined to give a set of interleaved double vowels (Fig. 1, row 4+row 5). The interleaved double vowels contained the same frequency components as the corresponding normal double vowels, but in an interleaved double vowel a particular harmonic series contained even harmonics which sampled one vowel's spectral envelope and odd harmonics which sampled the other vowel's envelope. An harmonic selection process operating on an interleaved double vowel would thus be unable to recover the envelope of either vowel. By contrast, the pattern of beating generated by the normal and by the interleaved double vowels at each frequency would be very similar. The frequency of modulation would be identical, because the same frequency components were present in both normal and interleaved double vowels. The depth and phase of modulation would differ because, for each vowel in an interleaved stimulus, half of the components are now harmonics of a different F_0 and are therefore given different amplitudes and phases by the synthesizer [Klatt, 1980, Eq. (6)]. Since the amplitude and phase spectra generated by the synthesizer change smoothly with frequency, these differences tend to become progressively larger as the corresponding harmonics of the two F_0's diverge in frequency. They are therefore larger at higher frequencies and for larger ΔF_0's. In contrast, the interest of this investigation lies at low frequencies, where Culling and Darwin (1993) showed that most of the separation effect produced by ΔF_0's is mediated, and small ΔF_0's whose effects are most difficult to explain using harmonic selection mechanisms.

In common with Scheffers (1983) and Culling and Darwin (1993), only vowels which differed in their phonemic identities were paired. With five vowels there are ten such exclusive pairs. Two versions of each vowel pair were prepared at each ΔF_0 and in each condition. In each version, the two F_0's or harmonic structures were allocated to different vowels, making 20 vowel combinations altogether. This procedure was followed even when the F_0's were both 100 Hz. With six ΔF_0's$\times$20 vowel combinations, this made 120 stimuli of each type and 240 stimuli altogether.

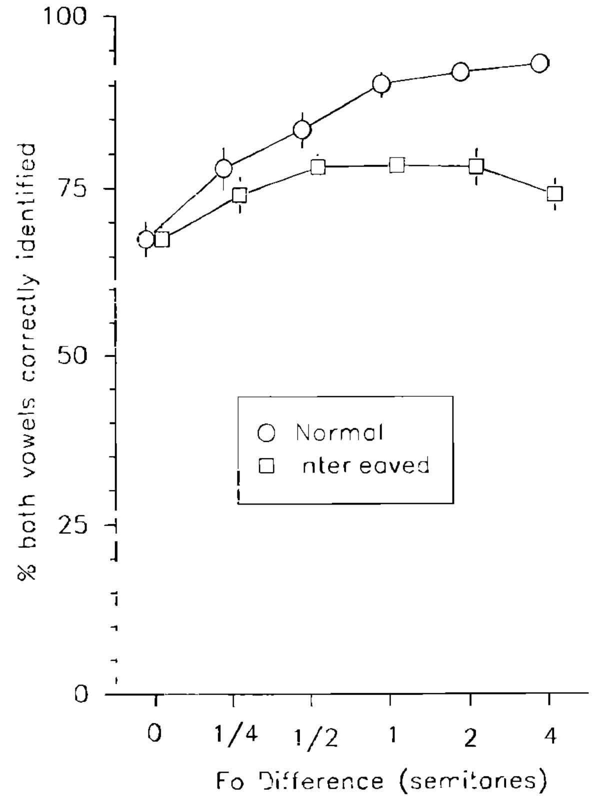

FIG. 2. Subjects' means and standard errors for percent both vowels correctly identified with (1) normal stimuli (solid); (2) interleaved stimuli (dashed).

The stimuli were played at a 10 kHz sampling rate via a 12-bit digital to analogue converter and passed through an antialiasing filter (4.5-kHz low pass) before being presented to subjects over Sennheisser HD414 headphones in a sound attenuating booth. The resulting presentation levels for individual vowels lay in the range 77–85 dB(A).

B. Procedure

Nine subjects, all of whom were experienced in double-vowel experiments, attended two hour-long sessions. Before each session subjects completed a practice containing all 90 single normal and inharmonic vowels in a random order.

In each experimental session the subjects received each double vowel twice in a randomized order. Three randomizations were used, with each subject receiving a different order in each of the two sessions.

C. Results

The results (Fig. 2) replicate the previously found improvement for identification of normal double vowels with ΔF_0, asymptoting at one semitone. For small ΔF_0's, the interleaved stimuli give similar improvements in identification to that of the normal stimuli, but a higher ΔF_0's ($\geqslant 1$ semitone) the normal stimuli gave significantly greater improvements. An analysis of variance covering harmonic allocation (normal/interleaved), ΔF_0 (0, $\frac{1}{4}$, $\frac{1}{2}$, 1, 2, 4 semi-

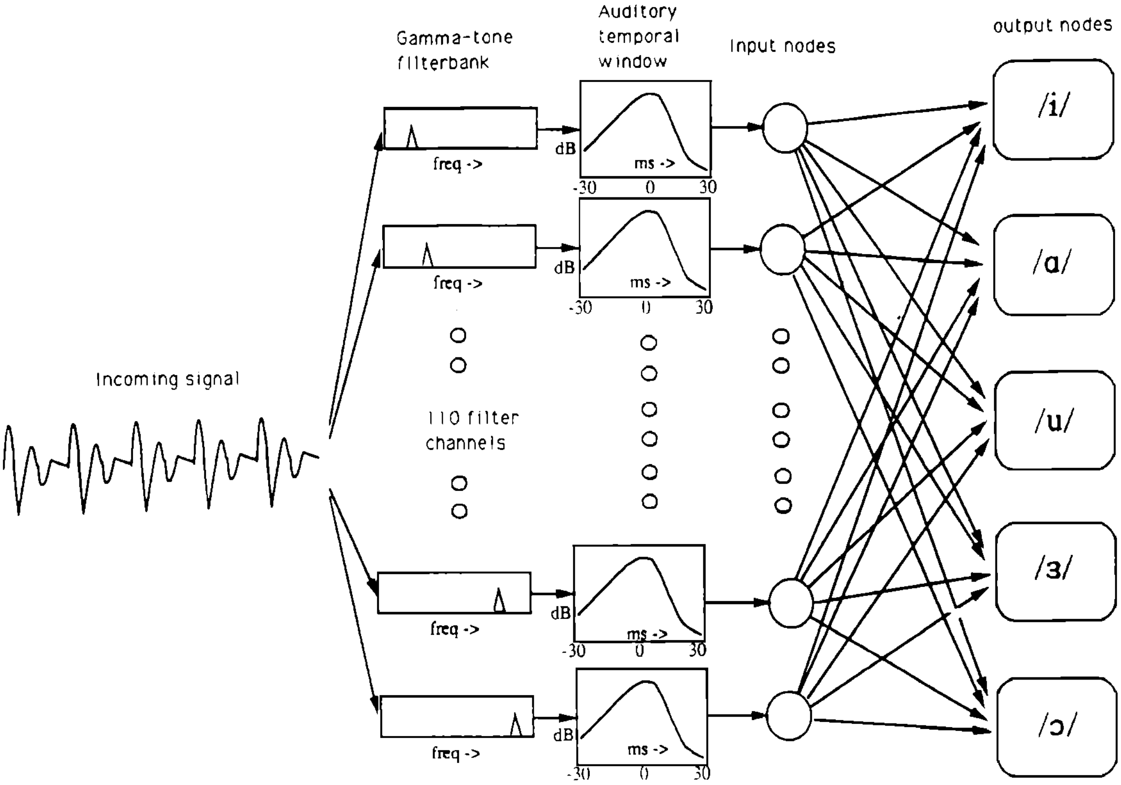

FIG. 3. Overview of the computational model architecture.

tones) and two replications (from different sessions), revealed a significant main effect of ΔF_0 [$F(5,40)=14.89$, $p<0.0001$] reflecting the general increase in identification with ΔF_0 and a significant main effect of harmonic allocation [$F(1,8)=86.22$, $p<0.0001$] due to better identification of the normal than of the interleaved stimuli. There was also an interaction between ΔF_0 and harmonic allocation [$F(5,40)=15.44$, $p<0.0001$], reflecting a lower asymptote for interleaved than for normal stimuli and possibly decline at 4 semitones ΔF_0. No effects involving the replication dimension were significant.

The interaction between the harmonic allocation and ΔF_0 was analyzed further using Tukey's pairwise comparisons. This analysis showed that the normal and interleaved conditions differed significantly only for ΔF_0's of 1, 2, and 4 semitones ($q=7.06$, 8.23, and 11.30, respectively, $p<0.01$). It also showed that correct identification in both the normal and interleaved conditions increased significantly above the $\Delta F_0=0$ baseline. For the normal condition, ΔF_0's of $\frac{1}{2}$, 1, 2, and 4 semitones gave identification significantly higher than 0 semitones ($q=6.48$, 9.20, 9.88, and 10.33, respectively, $p<0.01$). The 1, 2, and 4 semitone conditions also gave significantly more accurate identification than the $\frac{1}{4}$ semitone condition ($q=4.98$, 5.66, and 6.12, respectively). For the interleaved condition, ΔF_0's of $\frac{1}{2}$, 1, and 2 semitones gave identification significantly higher than 0 semitones ($q=4.27$, 4.38, and 4.27, respectively, $p<0.05$). No other effects were significant.

D. Discussion and conclusions

If harmonic selection were the only mechanism responsible for the improved identification accuracy of double vowels with increasing ΔF_0, identification would decrease with ΔF_0 in the interleaved condition, because the mechanism would group harmonics which have same F_0, but which sample different spectral envelopes, resulting in incomprehensible spectra. However, the results show an increase with ΔF_0 from 0 to $\frac{1}{2}$ semitone. This increase cannot be due to harmonic selection, but may be due to the exploitation of spectral amplitude modulation. Harmonic selection may, however, contribute to the continuing improvement in identification of the normal double vowels from $\frac{1}{2}$ to 1 semitone ΔF_0 and to maintaining this level out to 4 semitones ΔF_0. Identification accuracy in the interleaved condition remains steady from $\frac{1}{2}$ to 2 semitones and shows signs of decline at 4 semitones.

II. COMPUTATIONAL MODEL

A. Model design

The experiment above has shown that the improvement in double-vowel identification at small ΔF_0's may be due to low-frequency beating. In order to demonstrate the plausibility of this cue, a computational model was designed (Fig. 3), which exploits the beating between corresponding harmonics in order to classify pairs of simultaneous vowels. Beating gives rise to periodic changes in the power spectrum of the combined vowel pair over time, which are heard as an unstable, fluctuating timbre. The model, therefore, incorporates a representation of changing timbre, which incorporates temporal and frequency resolution, based on psychophysical measurements.

From its representation of dynamic timbre, the model derives recognition scores for the five possible constituent vowels at different points in time. It then exploits the changes in these values produced by beating in order to identify constituents vowels. The predictions of the model were tested using each of the two vowel selection strategies described in the introduction; the vowels were either selected individually at different points in the duration of the stimulus or in combination at the same point.

To test the model, the following experimental findings, which have been attributed to beating effects, were simulated.

(1) The increase in accuracy of identification of vowels in interleaved stimuli with the introduction of small ΔF_0's (predicted identification accuracy for normal stimuli should also match listeners' identification accuracy for interleaved stimuli rather than their accuracy for normal stimuli, since for the normal stimuli the subjects may also use harmonic selection while the model will not).

(2) The dip in the profile of identification scores at $\frac{1}{4}$ semitone ΔF_0 for stimuli of 50 ms (Assmann and Summerfield, 1990; Culling and Darwin, 1993).

A more severe test of the model would be to establish whether it reproduced the pattern of results across different vowel pairs which was found in the experiment, rather than just the overall accuracy of identification. In order to make this comparison the scores for individual vowel pairs produced by the model must be probabilistic predictions of the identification accuracy.

1. A dynamic spectrum

Frequency resolution in the model was given by Patterson *et al.*'s (1987, 1988) gamma-tone filterbank. The filterbank produces a set of filter output waveforms for any number of frequencies evenly distributed on an ERB scale (Moore and Glasberg, 1983). In the present model the filterbank had 110 channels with center frequencies equally spaced on a scale of ERB rate between 20 Hz and 4.8 kHz.

The time-varying rms power in each of the array of waveforms produced by the filterbank was calculated using Moore *et al.*'s (1988) "auditory temporal window." Attenuation, W was defined as a "rounded exponential" or "roex" function (Patterson and Moore, 1986) of temporal displacement t from the center of the window:

$$\begin{aligned} W(t) &= (1-w)\left[1+\frac{2t}{T_p}\right]\exp\left[\frac{-2t}{T_p}\right] \\ &\quad + w\left[1+\frac{2t}{T_s}\right]\exp\left[\frac{-2t}{T_s}\right] \\ &= \text{roex}(T_p,w,T_s). \end{aligned}$$

The formula is used with separate coefficients (T_p, w, and T_s) for the two skirts of the window, in order to reflect its asymmetry. The parameter T_p determines the decay rate of the exponential and the width of the temporal window, w determines the relative weighting of a second rounded exponential, and T_s determines the decay rate of the second rounded exponential, and hence the shape of the filter's tail.

The coefficients were estimated using data from Plack and Moore (1990) taken at probe tone frequencies of 300, 900, 2700, and 8100 Hz and at three different (frequency dependent) levels. For each frequency the data from the highest level was selected because these levels were closest to that used in the present experiment and by Culling and Darwin (1993). Plack and Moore found that the equivalent rectangular duration (ERD) of the window was around 13 ms at 300 Hz, shortening to around 8 ms at 8.1 kHz, so temporal detail with a shorter duration than this will be averaged out by the window. Their roex(T_p,w,T_s) coefficients were linearly interpolated to obtain coefficients for each of the model's 110 filter channels. Below 300 Hz the coefficients were held constant.

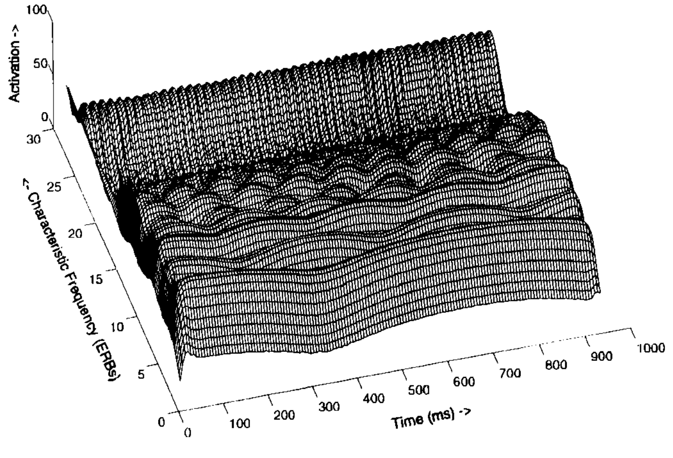

FIG. 4. The "dynamic spectrum" of the 1000-ms vowel pair AR+OR with $\frac{1}{4}$ semitone ΔF_0. 110 gammatone filter outputs were sampled 150 times by the temporal window with coefficients interpolated between those given by Plack and Moore (1990).

Figure 4 illustrates the output of the model for the AR+OR vowel pair with $\frac{1}{4}$ semitone ΔF_0 sampled at 6.67-ms intervals. At each sampling point the rms of the waveform was calculated, weighted by the appropriate window shape in each of the 110 frequency channels. The resulting values were log compressed and stored as 110 8-bit unsigned integers. These values were used as input activation values in the perceptron below. The temporal window was calculated over the range ±33 ms about the sampling point (increasing the range of the window did not alter the output once quantized to 8 bits). The 110 values which formed a spectral sample were collected at 150 sampling points spaced evenly throughout the 1000 ms duration (i.e., at 6.67-ms intervals). All the stimuli from the experiment were analyzed in this way. In order to simulate identification of the 50- and 200-ms stimuli from Assmann and Summerfield (1990) and Culling and Darwin (1993) the model simply considered the appropriate subset of these samples at the identification stage (the first eight for 50 ms; the first 30 for 200 ms). Unfortunately, although the stimuli were gated with 10-ms, raised cosine onset and offset ramps, the model's vowel selections were erratic when the temporal window was placed close to the onset or offset of the stimulus. Consequently, we were obliged to ignore the first two and last two spectrum samples; only the remaining 146 samples made a contribution to the results for 1000-ms stimuli and only six and 28 samples, respectively, made a contribution to the model's 50- and 200-ms results.

2. *Vowel identification using a two-layer perceptron*

A linear, two-layer associative network, or perceptron (Rosenblatt, 1959), was used for the purpose of vowel identification. The perceptron was trained to respond to each of the five vowels with activation on one of its five output nodes.

Unlike the method of vowel identification used by Scheffers (1983) and by Assmann and Summerfield (1990), a perceptron learns to use the whole stimulus spectrum and not just the formant peaks. A perceptron was used for simplicity and transparency, but may be too simple; Assmann and Summerfield (1989) found that Scheffers' identification algorithm, based on formant peak frequencies, was a better predictor of the pattern of vowel identification scores in the double-vowel paradigm that one which utilized the whole spectrum.

3. *Model structure and training*

The network had 110 input units, each receiving input from each of the 110 frequency channels and connected to five output units, which each coded one of the five vowels. Input activation, a_i, was propagated through connections of weight w_{ij} to activation of the five output units o_j by linear activation rule,

$$o_j = \sum_{i=1}^{i=110} a_i w_{ij}.$$

The weights of the perceptron were initialized to zero. It was then presented with a randomly selected stored spectral sample, from each of the five individual vowels in turn. The sequence of five vowels was repeated for each of the six F_0's in turn. This sequence of 30 stimuli was presented 150 times. The order of presentation was immaterial in the long term, since a perceptron always approaches the same solution to a given problem (Kohonen, 1984). In each training trial the perceptron was presented with one of the 150 spectral samples from that vowel, chosen at random. The target output, T, for each presentation to the model was 1.0 for the target vowel and 0.0 for each of the other vowels. Weight modifications, δw_{ij}, were derived using the "delta" rule with a learning rate, η, of 4×10^{-7}:

$$\delta w_{ij} = \eta \ a_i(T_{j-oj})$$

4. *Training results*

The model easily learned to identify the individual vowels; target vowel activation error typically declined to around 5% for all stimuli after 900 trials on each vowel (five vowels × six F_0's) × 150 presentations = 4500 trials altogether). The model did not simply learn each of the individual stimuli, as spectra from vowels synthesized at other F_0's (e.g., 150 Hz) were also correctly identified. An advantage of using a simple two-layer network compared to a more complex three-layer network, is that the weight vectors associating each input node with each output can easily be viewed. Figure 5 shows the weight vectors which resulted from this training. The model mainly learned the differences between the characteristic spectral envelopes of the training vowels. For instance, the ER vector shows clear peaks at frequencies of around 350 and 1250 Hz corresponding to its formant frequencies (Table I).

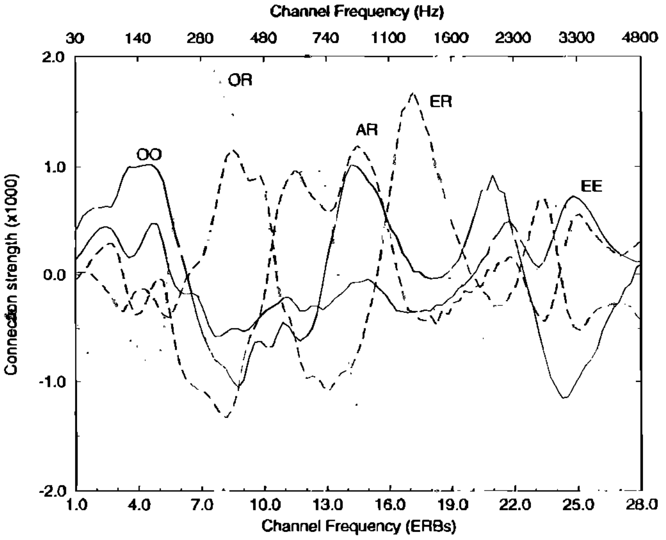

FIG. 5. Perceptron weight vectors leading to each output node after 900 training trials for each vowel.

B. Modeling the identification of double vowels

The identification of double vowels was modeled in two ways, reflecting two different strategies which listeners might employ. First, in the both-at-once strategy, the perceptron's output activations for each of the five individual vowels were combined to give response probabilities for each of the ten vowel-pair categories at each sampling point; the values of these probabilities at one sampling point, which produced the clearest, winning vowel pair were then taken as the overall response probabilities. Second, in the one-at-a-time strategy, the highest output activations produced by the perceptron for each of the five individual vowels across the different sampling points were recorded and then combined to give response probabilities for each of the ten vowel-pair categories.

1. *Generating response probabilities*

The output activations from the perceptron were not used to produce discrete vowel selections as in some previous models (e.g., Scheffers, 1983; Meddis and Hewitt, 1992). In common with Assmann and Summerfield (1989, 1990), individual vowel scores were converted into "response probabilities" which could vary continuously between 0 and 1. In making such a conversion, two basic principles were observed; the rank order of activations was preserved and the sum of calculated probabilities for each stimulus was 1.0. In order to fulfill these conditions each output activation was divided by the sum of all of the output activations *under consideration* (see below). However, the perceptron could produce negative output activations (which cannot be used in this conversion procedure).

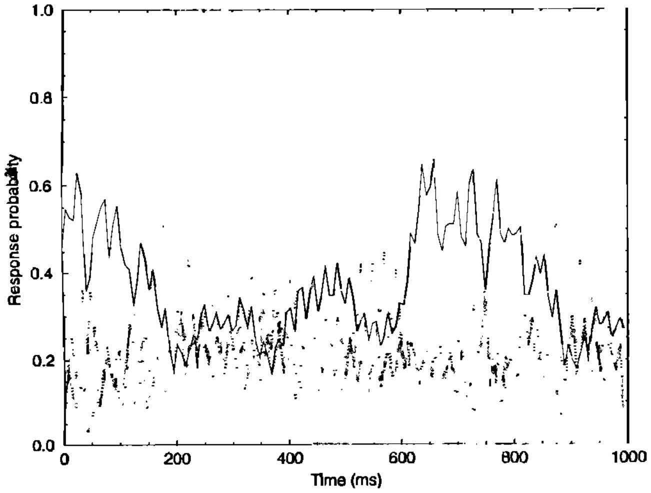

FIG. 6. Response probability contours for each of the 10 possible vowel pairs at 150 points during 1 s of the vowel pair OO+ER, with a $\frac{1}{4}$ semitone ΔF_0. The OO+ER contour has a solid line, while each of the other contours has a dotted line.

Attempts to use normalization of the activations in order to overcome this problem proved inappropriate, since the presence of strongly negative output activations could make zero activations yield quite high response probabilities. Elman (personal communication) recommended that response probability, p_j, follow an expansive nonlinear relation:

$$p_j \propto e^{(ko_j)}.$$

This relation contains a constant, k, which was used as a free parameter to scale the model's overall performance of that of subjects.

Since, in the experiment, the subjects knew that the two vowels were always different and always made two different responses, the model assumed that the first vowel selection was ***eliminated from consideration*** in calculating the probability of each remaining vowel as a second selection. Aside from this modification, the method of converting raw scores (activations) into probabilities was similar to that of Assmann and Summerfield [1989, Eq. (6)].

The probability of selecting two different vowels (1 and 2), *in that order*, given output activations $o_{1\to 5}$ is given by the following equation:

$$p(1 \text{ and } 2)=\frac{e^{ko_1}}{\sum_{i=1}^{5} e^{ko_i}} \times \frac{e^{ko_2}}{\sum_{i=2}^{5} e^{ko_i}}.$$

2. Response probability contours

In the simple one-at-a-time strategy this equation is used only once and applied to the highest of the output activations produced for each of the five vowels. In the case of the both-at-once strategy the equation is used at each sampling point to give a response probability for each of the ten vowel-pair categories at that point; the response probability for a particular vowel pair thus forms a contour across time. Figure 6 illustrates a set of such contours for the vowel-pair OO+ER with a ΔF_0 of $\frac{1}{4}$ semitone. In most cases, as in the figure, the dominant contour is that for the correct vowel pair, and some contours remain at around zero probability for the duration of the stimulus. Without a ΔF_0 the contours are quite steady, aside from a rapid and regular variation which coincides with the glottal pulsing in the stimuli. Once a ΔF_0 has been introduced, however, long-term changes occur in the contours and different contours can, as illustrated in the figure, be dominant at different times.

In the both-at-once strategy, the model assumes that subjects exploit their ability to derive two vowels from a spectral contour at the moment when the changing contour sounds most clearly like a particular vowel pair. The model takes the highest point reached by any response probability contour during the stimulus and takes the response probabilities of each vowel pair at that moment as those for the complete stimulus.

The effect of a spectral change induced by a ΔF_0 on the response probability contours and upon the predicted responses is complex. There are three likely scenarios:

(1) When the correct vowel pair has the dominant contour, the effect of the changing spectrum produced by a ΔF_0 is likely to make this contour sometimes more dominant sometimes less. Since the model will select the moment when it is most dominant, spectral changes and hence ΔF_0's favor the dominant contour. The model may predict better performance with the ΔF_0 than without.

(2) Occasionally, however, spectral change induced by a ΔF_0 will cause an incorrect and normally subordinate contour to briefly become dominant and reach a higher peak than is attained by the correct and normally dominant contour. At such a moment, the response probability from the contour for the correct pair will tend to be relatively low and the model will do worse with a changing spectrum than with a stable spectrum.

(3) When the correct vowel pair does not have the dominant contour, another vowel pair must be dominant. Since predicted responses are likely to be based upon a high point in the dominant contour, all the subordinate contours, including that for the correct vowel pair will tend to be relatively low. Hence the model may again predict worse performance with the ΔF_0 than without.

C. Model performance for normal and interleaved stimuli

1. Overall performance

Figure 7 shows the performance of the model using each vowel selection strategy at each ΔF_0, and for both normal and interleaved stimuli with various values of the free parameter, k, juxtaposed against the data from the experiment. For higher ΔF_0's ($\geqslant 1$ semitone) the subjects perform much better than the model in identifying normal double vowels, but, with interleaved double vowels, subjects' identification is comparable to that of the model at each ΔF_0. At the higher ΔF_0's the subjects may be performing better than the model because they alone are able to employ harmonic selection using these larger ΔF_0's.

For the both-at-once strategy [Fig. 7(a)] the model and subject performance is closest when $k=6$ or 7. For the one-at-a-time strategy [Fig. 7(b)] the model and subject

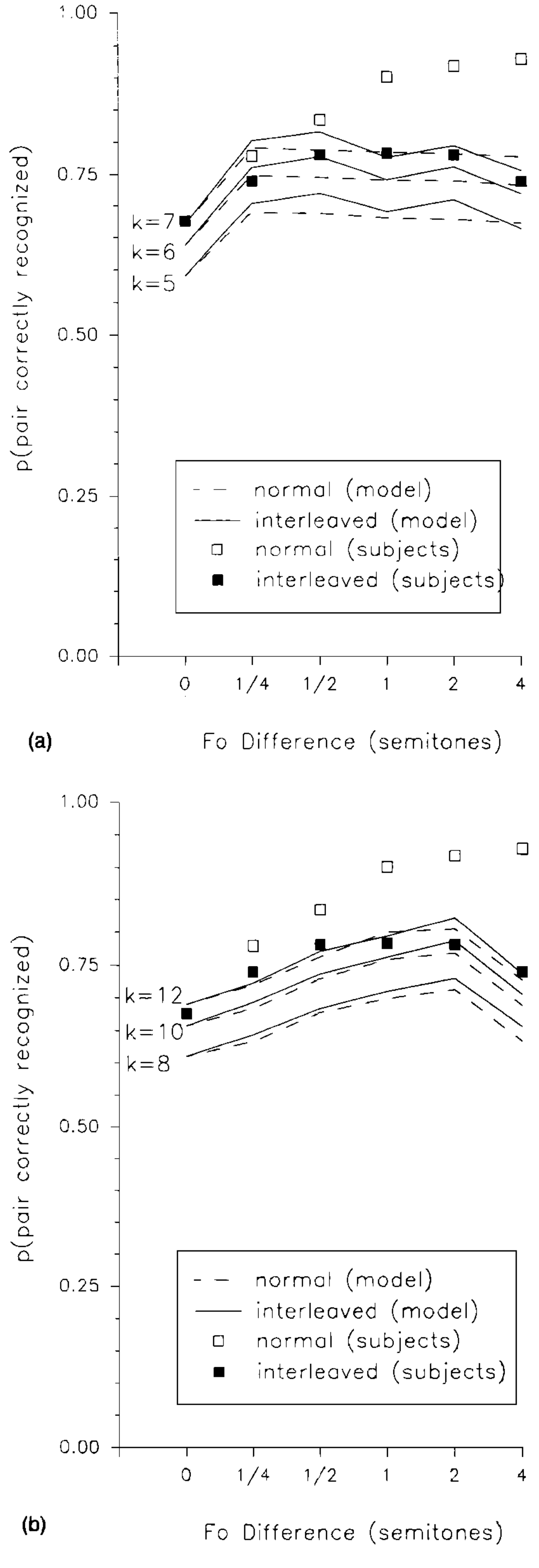

FIG. 7. Comparison of model's predicted and listener's actual identification accuracy for (a) both-at-once model and (b) one-at-a-time model. Each panel shows recognition performance for subjects with (1) interleaved stimuli (triangles), (2) normal stimuli (hexagons) and for the model with (3) interleaved stimuli (dashes), and (4) normal stimuli (solid) for three different values of k.

performance is closest when $k=12$. Further evaluation of the model used data obtained with k set to 6 and 12, respectively. For the both-at-once strategy, the model's correct identification probability increases as ΔF_0 increases to $\frac{1}{4}$ semitone is roughly level as ΔF_0 increases further. For the one-at-a-time strategy, the model's correct identification probability increases progressively as ΔF_0 increases up to 2 semitones, and then drops. The model's correct identification probabilities are similar for both normal and interleaved double vowels and correspond fairly well with the human accuracy of identification for interleaved stimuli, which the model is designed to emulate.

2. Performances with individual vowel pairs

Simply modeling overall performance, as done previously by Scheffers (1983) and Meddis and Hewitt (1992), is a weak test of any model. Assmann and Summerfield (1990) calculated the Pearson's r correlations between the listeners accuracy of identification and that predicted by the different versions of their model for each vowel pair, but a stronger test of the action of a model is to see whether the model predicts the observed *changes* in identification of *individual* vowel pairs across different ΔF_0's. The degree of improvement/decline in performance was used as a measure in order to factor out any differences between the model and the subjects in the overall identifiability of each vowel pair. Table II shows the Pearson's r correlations between the model's and the subjects' changes in performance with the introduction of each ΔF_0. The correlations across vowel pairs are all positive and often significantly so. The correlations produced using the both-at-once strategy are stronger than those produced by the one-at-a-time strategy.

One trivial explanation of the correlations in Table II. If, when $\Delta F_0=0$, the different vowel pairs are similarly identifiable for both model and subjects, similarity in the improvement may then be the result of shared ceiling effects among those pairs which are well identified by both. However, the correlations between model and subject scores for zero ΔF_0 are modest (0.2218 for both-at-once; 0.2655 for one-at-a-time).

3. Discussion

The interleaved stimuli of the experiment were designed to prevent subjects from using harmonic selection mechanisms to improve identification scores, while allowing the use of spectral change cues. The computational model was designed to exploit only spectral change cues to improve its identification score, whether presented with normal or interleaved stimuli. The similarity between the subjects' percent correct identification for interleaved stimuli and the model's correct identification probabilities for both the interleaved and the normal double vowels is therefore very satisfactory. With normal stimuli the subjects performed increasingly better than with the interleaved stimuli (and than the model) for ΔF_0's$>\frac{1}{2}$ semitone, reflecting the intervention of genuine harmonic selection mechanisms.

The invariably positive and often significant correlations between the model's and the subjects' performance improvement for individual vowel pairs reinforce the conclusion that the model is exploiting the same cues as the

TABLE II. Correlations between model and subject correct identification probability improvements with the introduction of each ΔF_0 for normal and interleaved stimuli using the both-at-once and one-at-a-time vowel selection strategies.

Stimulus type	ΔF_0 (semitones)	Both-at-once correlation	Both-at-once one-tail prob.	One-at-a-time correlation	One-at-a-time one-tail prob.
Normal	$\frac{1}{4}$	0.7808	0.0077**	0.5788	0.0796
	$\frac{1}{2}$	0.7580	0.0111*	0.4745	0.1658
	1	0.7989	0.0056**	0.6708	0.0337*
	2	0.6442	0.0444*	0.4862	0.1542
	4	0.6339	0.0491*	0.6189	0.0564
Interleaved	$\frac{1}{4}$	0.4163	0.2315	0.3518	0.3188
	$\frac{1}{2}$	0.4921	0.1485	0.4409	0.2021
	1	0.7556	0.0115*	0.7612	0.0105*
	2	0.4154	0.2325	0.3992	0.2530
	4	0.7647	0.0010**	0.5589	0.0931

subjects. It illustrates how subjects may be improving their performance for stimuli with ΔF_0's without selecting out two harmonic series.

D. Model performance for different stimulus durations

1. Overall performance

Figure 8 shows the performance of the model using only the first eight (50 ms) or 30 (200 ms) samples from the normal stimuli, juxtaposed against the 50- and 200-ms data for the eight higher scoring subjects from Culling and Darwin (1993, expt. 1). The subjects' overall scores are relatively low, because these data were collected from inexperienced subjects, but the pattern of their data is comparable to that of the model. Using the both-at-once strategy the model reproduces the $\frac{1}{4}$ semitone performance dip for eight sample (50 ms) stimulus segments, whilst showing a similar performance profile for 30 sample (200 ms) segments to that for 150 samples (1000 ms), shown in Fig. 7. Using the one-at-a-time strategy, the results are less satisfactory; the dip at $\frac{1}{4}$ semitone ΔF_0 occurs for both the eight and 30 sample analyses. In addition, Culling and Darwin (1993)'s listeners showed a strong improvement between $\frac{1}{4}$ and $\frac{1}{2}$ semitone ΔF_0 for 200-ms stimuli, which the model was unable to reproduce, suggesting that harmonic selection may have played a role at only $\frac{1}{2}$ semitone ΔF_0.

2. Performance with individual vowel-pairs

A Pearsons *r* correlation across vowel pairs compared the decline in performance described above which resulted from the introduction of a $\frac{1}{4}$ semitone ΔF_0 for the subjects and the model. This correlation was slightly negative ($r = -0.0348$).

The model therefore offers only limited support (in the form of a dip in the overall predicted identification accuracy) to Assmann and Summerfield's suggestion that waveform interaction impairs subjects performance on 50-ms stimuli with $\frac{1}{4}$ semitone ΔF_0 .

E. Discussion

Though simple, the model provides an illustration of how spectral modulation might be exploited by a listener in order to improve identification of double vowels at very small ΔF_0's. The model serves to illustrate the potential of spectral changes rather than to provide a formal theory. Other, more sophisticated theories, using some kind of temporal integration could be invoked. Further empirical work must investigate such possibilities. For instance, Assmann and Summerfield (1994) has investigated why the 50-ms stimuli used in Assmann and Summerfield (1990) did not give the improvement with increasing ΔF_0 which was found for 200-ms stimuli. They compared the effects of temporal integration with spectral change by presenting repeated or successive 50-ms portions of the 200-ms stimuli. If temporal integration is important then repeating the same segment should elicit an improvement at the larger ΔF_0's, but if spectral changes are important then improvement will come only from the spectral differences between successive portions of the stimuli. They found that the repetition of the same segment does not improve recognition with a ΔF_0, while a succession of 50-ms segments does. This finding corroborates the suggestion that a stimulus with a ΔF_0 changes in identifiability over time. In addition, Assmann and Summerfield used their technique to show that scores for the 200-ms segments were no better than for the best of the four 50-ms segments, corroborating the suggestion that listeners decisions may be based on particular parts of the stimulus in which the vowels are easier to identify.

III. GENERAL DISCUSSION

The experiment and model reported here have tried to explain how accuracy of identification of double vowels of at least 200-ms duration can improve markedly with the introduction of only $\frac{1}{4}$ semitone ΔF_0 (Scheffers, 1983; Culling and Darwin, 1993). For such a small ΔF_0, a harmonic selection/rejection mechanism would require a spectral representation of extremely high resolution—considerably higher than that produced by the peripheral auditory sys-

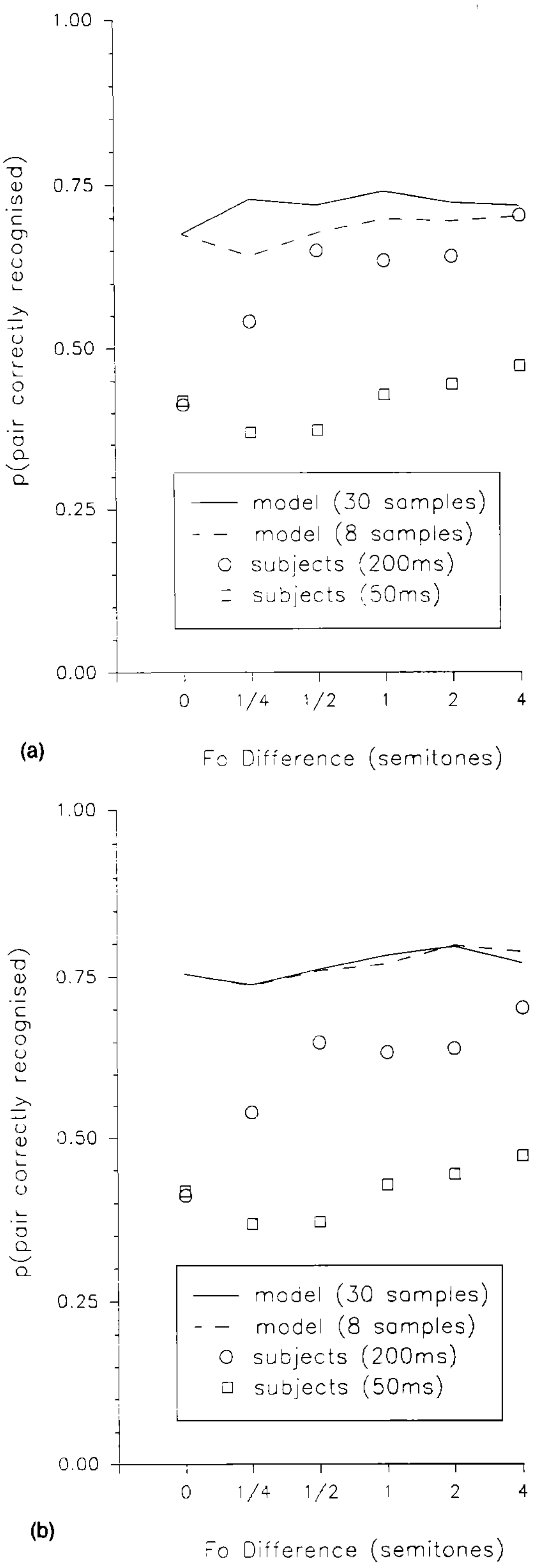

FIG. 8. Comparison of model's predicted and listener's actual identification accuracy for (a) both-at-once model and (b) one-at-a-time model. Each panel shows the model's predicted probabilities of correct identification for (1) 30 samples (200 ms—solid); (2) eight samples (50 ms—medium dash) and subjects' actual probability of correct identification for (3) 200-ms stimuli (long dash); (4) 50-ms stimuli (short dash).

tem. Assmann and Summerfield (1990) and Meddis and Hewitt (1992) have suggested that phase-locking information from the auditory nerve may supplement peripheral filtering to provide the necessary resolution through a mechanism similar to autocorrelation (Licklider, 1951). The above experiment and model show, however, that it is possible to explain the data for small ΔF_0's without recourse to an analysis of temporally encoded periodicity.

The experiment reported here used double vowels whose component structure was designed to confuse mechanisms which use harmonic selection or rejection; alternate harmonics of each F_0 sampled the two vowels' different spectral envelopes. The accuracy of identification for these stimuli improved when the corresponding components from competing vowels were slightly mistuned, as though a ΔF_0 between the vowels had been introduced. The improvement in identification accuracy was attributed to the spectral modulation which occurs when the corresponding components are mistuned, and their waveforms interact. These results suggest that the effect of small ΔF_0's on double-vowel identification may be mediated by listeners' exploitation of the spectral modulation rather than of the harmonic structure of the component vowels. The experiment shows that the spectral modulation is a sufficient cue, but cannot prove that this cue is exploited by subjects when listening to normal stimuli.

A computational model was developed to illustrate how listeners might be exploiting the spectral modulation. The model employed a psychophysically realistic spectral representation of changing timbre, using a gamma-tone filterbank (Patterson *et al.*, 1987, 1988) followed by an implementation of the auditory temporal window (Moore *et al.*, 1988; Plack and Moore, 1990). Based on this representation a perceptron rated the similarity of the spectrum to that of each candidate vowel from moment to moment. The model made the most accurate predictions when the overall response probabilities were based on one moment when the perceptron identified two different candidates most clearly. So configured, the model represents only one, rather simple strategy which subjects may use in order to exploit dynamic spectral cues, but nonetheless the model reproduced subjects' ability to exploit spectral modulation.

Both the experiment and the computational model have shown that part of the improvement in identification with ΔF_0's can be accounted for by the exploitation of changing timbre cues. Mechanisms based on harmonic selection may make little or no contribution to the improvement in double vowel recognition for ΔF_0's of less than 1 semitone. If so, one would expect models based purely on harmonic selection to show improvements in identification only at ΔF_0's of 1 semitone or more. In contrast, published models of double-vowel separation have either shown no reliable improvement in vowel identification with increasing ΔF_0 (Scheffers, 1983; Assmann and Summerfield, 1990, "place" model) or have displayed improvement which begins at $\frac{1}{4}$ semitone ΔF_0 (Assmann and Summerfield, 1990, "place-time;" Meddis and Hewitt, 1992).

Assmann, P. F., and Summerfield, Q. (**1989**). "Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency," J. Acoust. Soc. Am. **85**, 327–338.

Assmann, P. F., and Summerfield, Q. (**1990**). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680–697.

Assmann, P. F., and Summerfield, Q. (1994). "Some effects of duration on the perception of concurrent vowels," J. Acoust. Soc. Am. 95, 471–484.

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

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

Cherry, E. C. (1953). "Some experiments on the recognition of speech with one and two ears," J. Acoust. Soc. Am 25, 975–979.

Culling, J. F., and Darwin, C. J. (1993). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F_0," J. Acoust. Soc. Am 93, 3454–3467.

Darwin, C. J., and Gardner, R. B. (1986). "Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality," J. Acoust. Soc. Am. 79, 838–844.

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

Klatt, D. J. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. 67, 838–844.

Kohonen, T. (1984). *Self-organization and Associative Memory* (Springer-Verlag, Berlin).

Licklider, J. C. R. (1951). "A duplex theory of pitch perception," Experientia 7, 128–133.

Meddis, R., and Hewitt, M. J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. 91, 233–45.

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

Moore, B. C. J., Glasberg, B. R., Plack, C. J., and Biswas, A. K. (1988). "The shape of the ear's temporal window," J. Acoust. Soc. Am. 83, 1103–1116.

Roberts, B., and Moore, B. C. J. (1990). "The influence of extraneous sounds on the perceptual estimation of first formant frequency in vowels," J. Acoust. Soc. Am. 88, 2571–2583.

Roberts, B., and Moore, B. C. J. (1991). "Modeling the effects of extraneous sounds on the perceptual estimation of first formant frequency in vowels," J. Acoust. Soc. Am. 89, 2933–2951.

Patterson, R. D., and Moore, B. C. J. (1986). "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing* (Academic, London).

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1987). "An efficient auditory filter-bank based on the gammatone function," paper presented to the IOC speech group on auditory modeling at RSRE, Dec. 14–15.

Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (1988). "Spiral vos final report, Part A: The auditory filter bank," Cambridge Electronic Design, Contract Report (APU 2341).

Plack, C. J., and Moore, B. C. J. (1990). "Temporal window shape as a function of frequency and level," J. Acoust. Soc. Am. 88, 2178–2187

Rosenblatt, F. (1959). "Two theorems of statistical separability in the perceptron," in *Mechanization of Thought Processes*, Proceedings of a Symposium Held at the National Physical Laboratory, November 1959 (H. M. Stationery Office, London), Vol. I.

Scheffers, M. T. M. (1983). "Sifting vowels: Auditory pitch analysis and sound segregation," Ph.D. thesis, University of Gronigen.

Summerfield, Q., and Assmann, P. F. (1991). "Perceptual separation of concurrent vowels: Effects of pitch pulse asynchrony and harmonic misalignment," J. Acoust. Soc. Am. 89, 1364–1377.

Zwicker, U. T. (1984). "Auditory recognition of diotic and dichotic vowel pairs," Speech Commun. 3, 265–277.