Modeling the identification of concurrent vowels with different fundamental frequencies

Ray Meddis and Michael J. Hewitt
Department of Human Sciences, University of Technology, Loughborough LE11 3TU, United Kingdom

(Received 14 June 1990; accepted for publication 5 August 1991)

Human listeners are better able to identify two simultaneous vowels if the fundamental frequencies of the vowels are different. A computational model is presented which, for the first time, is able to simulate this phenomenon at least qualitatively. The first stage of the model is based upon a bank of bandpass filters and inner hair-cell simulators that simulate approximately the most relevant characteristics of the human auditory periphery. The output of each filter/hair-cell channel is then autocorrelated to extract pitch and timbre information. The pooled autocorrelation function (ACF) based on all channels is used to derive a pitch estimate for one of the component vowels from a signal composed of two vowels. Individual channel ACFs showing a pitch peak at this value are combined and used to identify the first vowel using a template matching procedure. The ACFs in the remaining channels are then combined and used to identify the second vowel. Model recognition performance shows a rapid improvement in correct vowel identification as the difference between the fundamental frequencies of two simultaneous vowels increases from zero to one semitone in a manner closely resembling human performance. As this difference increases up to four semitones, performance improves further only slowly, if at all.

PACS numbers: 43.66.Ba, 43.66.Hg, 43.71.Cq, 43.71.Es

INTRODUCTION

The human ability to attend selectively to one speech signal in a mixture of speech sounds has received considerable attention (e.g., Broadbent, 1952; Brokx and Nooteboom, 1982; Cherry, 1953; Darwin, 1981, 1984; Egan *et al.*, 1954; Gardner *et al.*, 1989; Halikia and Bregman, 1984; Hartmann, 1988; Parsons, 1976; Stubbs and Summerfield, 1988, 1990; Triesman, 1960; Weintraub, 1985, 1987). Various factors have been shown to influence this ability. We consider here only the role of voice pitch. A consistent finding is that listeners are able to separately identify two simultaneously presented synthesized vowels significantly better than chance even when they have approximately the same amplitude, when they start and stop at the same time, are both presented to the same ear and both have the same fundamental frequency (f_0) (Assmann and Summerfield, 1989, 1990; Chalikia and Bregman, 1989; Scheffers, 1983a; Zwicker, 1984). Also, all investigators find that performance improves substantially if a difference in f_0 is introduced between the two vowels. Figure 1 shows that correct identification of both vowels shows an improvement of 18% for f_0 differences up to four semitones. Most of the improvement, however, is restricted to the first semitone f_0 separation.

The problem for the modeler is to devise conceptual schemes that can give an account of this process. Both Scheffers (1983a) and Assmann and Summerfield (1990) have developed sophisticated models capable of identifying simultaneously presented vowels at approximately the same level of success as human listeners. Despite this achievement, both models have experienced difficulty in reproducing the gradual improvement in performance with increasing f_0 separation as shown in Fig. 1. In this article, we present an example of a model which does show this important property.

The vowel identification systems of both Scheffers (1983a) and Assmann and Summerfield (1990) extract pitch and use this to assist the generation of two separate templates for matching purposes. However, their difficulty

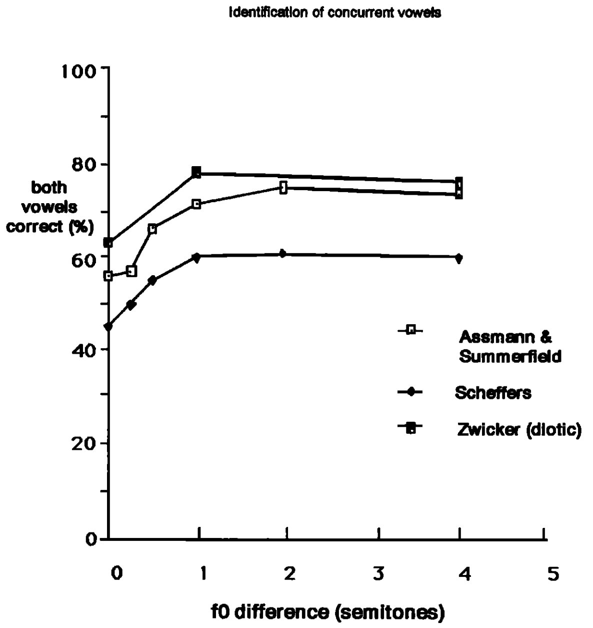

FIG. 1. Listeners' ability to correctly identify both vowels of a pair of simultaneous presented vowels as a function of the separation of the f_0 of the two vowels. Results taken from Scheffers (1983a), Zwicker (1984), and Assmann and Summerfield (1990).

in showing an improvement in performance with increasing f_0 separation cannot be explained in terms of any inadequacy in the pitch extraction algorithms because the difficulty persists even when the models are given explicit information about the true pitch values. Nor can the problem be assigned to any failure of the vowel identification algorithms because these work adequately when there is no f_0 separation between the vowels. The problem must lie elsewhere.

Scheffers used a variant of the harmonic sieve technique based on work by Goldstein (1973) and Gerson and Goldstein (1978) and explicitly modeled by Duifhuis *et al.* (1982) and Scheffers (1983b). In this method, the input signal is passed through a bank of bandpass filters configured to simulate many of the known mechanical filtering properties of the auditory periphery. The power output of the filters is passed to an algorithm which attempts to identify the fundamental or fundamentals which might give rise to the observed pattern of peaks and valleys in the spectral profile. Each low-frequency peak is assumed to be a resolved harmonic of one or both of the two f_0's and is assigned to one or both groups on this basis or, alternatively, it might be rejected altogether. Two new spectral profiles are reconstructed on the basis of these two sets of peaks using a process of interpolation. Another algorithm, then estimates the formant frequencies for each reconstructed spectral profile. These frequencies are used in a template-matching algorithm which indicates the optimum classification for the two vowels.

Assmann and Summerfield (1990) followed Scheffer's basic plan but introduced a number of variations which, in effect, constituted four different models. Two "place" models estimated pitch using an analysis of the distribution of power output across the channels of the filter bank as did Scheffers' model. Two "place-time" models achieved the same goals using a periodicity analysis of the waveforms in each channel. To do this, they computed the autocorrelation function (ACF) for each channel separately before pooling the functions by summing across channels. Major peaks in the pooled ACF were, with certain restrictions, identified with component pitches.

Each place and place-time method was studied in two versions. The "linear" version operated directly on the waveform emerging from the filters as did Scheffer's model. The "nonlinear" version applied a compressive nonlinearity to the output of the filters to simulate one of the properties of the mechanical-to-neural transduction process at the hair cell in the cochlea. Their results showed that the nonlinear, place-time model produced the most satisfactory performance in terms of (a) accuracy of predicting pitches, (b) overall mean accuracy in identifying both component vowels, and (c) the ability to specify the pattern of correct responses and confusions for individual stimuli. While these two innovations (nonlinearity and time domain periodicity analysis) represent important advances in sophistication in this context, none of their models was able to show a gradual monotonic improvement in performance with increasing separation of f_0.

The work to be described below seeks to build on the modeling work of Assmann and Summerfield and uses, for comparison purposes, the human performance data they collected. Like their system, our model simulates a number of aspects of peripheral auditory processing (middle/outer ear frequency effects, cochlea filtering and mechanical-to-neural conversion at the inner hair cell). We also use a digital-bandpass simulation of the basilar membrane mechanical frequency selectivity. Another common feature was the inner hair cell model developed in our laboratory (Meddis, 1986, 1988; Meddis *et al.*, 1990). In the next stage, our model also extracts the pitches of the two sounds using a periodicity analysis method similar in many respects to Licklider's (1951, 1959) early suggestion (see, also, Gardner, 1989; Lazzaro and Mead, 1989; Moore, 1982).

The important difference occurs later in the system. One difference involves using the decision concerning pitch values to segregate frequency-selective channels into two mutually exclusive sets of channels, one for each vowel. Another difference involves the use of periodicity information (combined across channels belonging to a subgroup of channels) to produce a pooled ACF for each vowel. Identification of the component vowels is then based on these two separate periodicity profiles. Both innovations represent departures from current models. In the first case, component vowels are characterized using information from *only* one of two mutually exclusive subsets of channels identified on the basis of pitch. The segregation of channels into two sets only becomes possible as the fundamental frequencies of the two vowels diverge. Moreover, the segregation becomes more secure as the difference in f_0 increases. This is what gives rise to the gradual improvement in performance.

In the second major departure, the identification is based entirely on the pooled periodicity profiles which are summed across channels. At this stage, place information is entirely abandoned. We do not maintain that only periodicity information is relevant to hearing. On the contrary, we recognize that there are strict limitations to the ability of the nervous system to extract and preserve periodicity information concerning individual frequencies above 4–5 kHz. However, in the case of speech, most of the relevant information is carried by lower frequencies. For the synthesized stimuli used, the model performs very adequately using only periodicity information.

I. THE MODEL

The early stages of the model are exactly the same as those used in our exploration of pitch phenomena (Meddis and Hewitt, 1991). It has already been shown to give good estimates of pitch which are consistent with a wide range of psychophysical studies of human pitch perception. Two new modules have been added to deal with vowel identification. These are (i) a procedure for using pitch information to segregate channels into two sets corresponding to the two sound sources and (ii) a template matching procedure for the purpose of identifying the vowels.

The total system now consists of a concatenation of eight modules which: (1) simulate middle- and outer-ear, low- and high-frequency frequency attenuation effects, (2) simulate the mechanical frequency-selectivity of the basilar membrane, (3) simulate mechanical to neural transduction at the inner hair cell, (4) calculate running ACFs in individual channels, (5) perform cross-channel summation of the ACFs to form a pooled ACF, (6) perform pitch identification using peaks in the pooled ACF, (7) segregate channels into two mutually exclusive subsets, (8) perform vowel identification using a template-matching procedure applied to the pooled ACF of each subset of channels.

Stages (1)–(6) are summarized in Fig. 2 and a detailed specification of the peripheral processing aspects of the model is given in Meddis and Hewitt (1991). However, in the interests of clarity, the following account illustrates the first six stages of the model's response to one of the five single synthesized vowels as used by Assmann and Summerfield.

The stimulus shown in Fig. 3(a) is a 30-ms segment of the vowel "ah" ($f_0 = 100$ Hz) at amplitude 50 dB1[1] just before the end of the 200-ms duration of the stimulus. In Fig. 3(b), the stimulus has been passed through a 100 bandpass digital-filter system. The equivalent rectangular bandwidth (ERB) of each filter is based on measures of the psychophysical critical bandwidth in human subjects which, in turn, correspond reasonably closely (at least, at moderate amplitudes) with the tuning curves of individual auditory-nerve fibers (Moore, 1986). The center frequencies of the overlapping filters in the filterbank are equally spaced on an ERB scale 0.24 ERBs apart between 80 Hz and 4 kHz (Moore and Glasberg, 1987). For clarity in reproducing the figures, only one in four of the channels are shown. The vertical graph to the right shows the power output from each channel and represents the excitation function. The amplitude responses of the filters, as well as their ERBs, are based on human psychophysical studies. However, at the presentation levels used here, these functions are a reasonable approximation to the frequency tuning curves when measured electrophysiologically in other mammals.

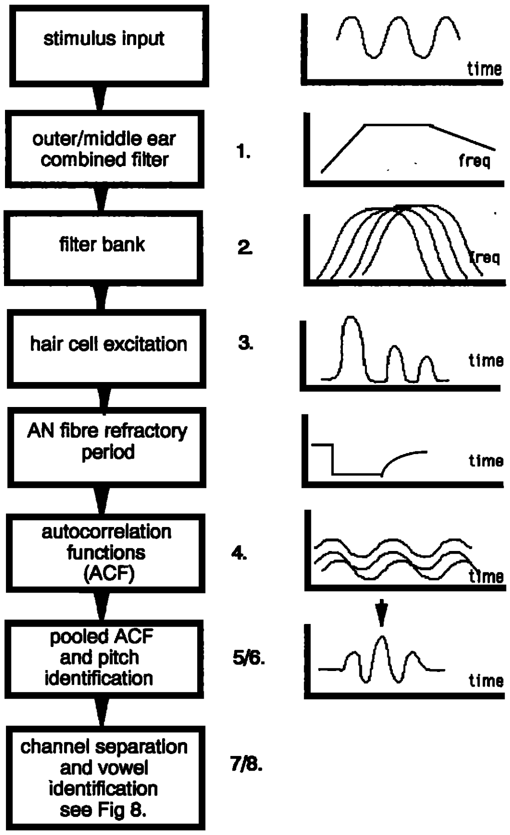
stimulus input
outer/middle ear combined filter
filter bank
hair cell excitation
AN fibre refractory period
autocorrelation functions (ACF)
pooled ACF and pitch identification
channel separation and vowel identification see Fig 8.
1.
2.
3.
4.
5/6.
7/8.
time
freq
freq
time
time
time
time

FIG. 2. Processing sequence for the computational model. See text for explanation of numbered steps. Channel separation and vowel identification algorithms are represented in Fig. 8.

Figure 3(c) shows the response of the inner hair cells within each channel in terms of the *probability* of an action potential in the corresponding auditory-nerve fiber. We assume that a large number of inner hair cells are active within a single channel and that the aggregate within channel firing-rate will be similar to the probability function for a single fiber. The model calculates the amount of transmitter in the hair-cell/nerve-fiber synapse and assumes that the probability of firing is a linear function of that amount. At low intensities (less than 20 dB1), the firing-probability function follows the filtered input function fairly closely. At higher intensities, the output is increasingly half-wave rectified in character. This model of hair-cell functioning is only one of many and the relative merits are thoroughly discussed in Hewitt and Meddis (1991).

The vertical graph at the right-hand side of Fig. 3(c) shows the average event rate for each channel (calculated over the 30-ms interval shown) and represents the "rate-place" profile for this vowel.

A running ACF was generated separately for each channel. Licklider (1951) suggested that the summation over time should be limited by a time constant, Ω, of approximately 2.5 ms:

$$h(t,\partial t) = \sum_{i=1}^{\infty} p(t-T)p(t-T-\partial t)e^{-T/\Omega}$$

$$(T = i\,dt). \qquad (1)$$

Here ∂t is the autocorrelation lag, dt is the sample period, and t is the time at which the ACF is sampled. Licklider does not explain why the time constant should be 2.5 ms but more recent work by Viemeister (1979) on the temporal modulation transfer function suggests a similar value (3 ms). Recent work on the "temporal window" by Plack and Moore (1990) suggests a window width of 8–10 ms but a somewhat different function from the exponential decay implied above. In the context of double vowel separation, we found that a time constant between 10 and 25 ms was more satisfactory.

For $T > 3\Omega$ expression (1) returns only very small values. Accordingly, the ACF was, in practice, computed only over a period equal to three times the time constant.

Figure 3(d) shows the running autocorrelation function for each channel, immediately before the end of the 200-ms presentation. The function is computed with a time constant of 10 ms. As a consequence, the more distant the event in time, the less influence it has on the running autocorrelation function. The function is computed over time lags from 0.1 to 12.5 ms in steps of 0.1 ms. For pure tone inputs this represents a range from 10 kHz to 80 Hz. The ACFs reveal the periodicities present in each channel.

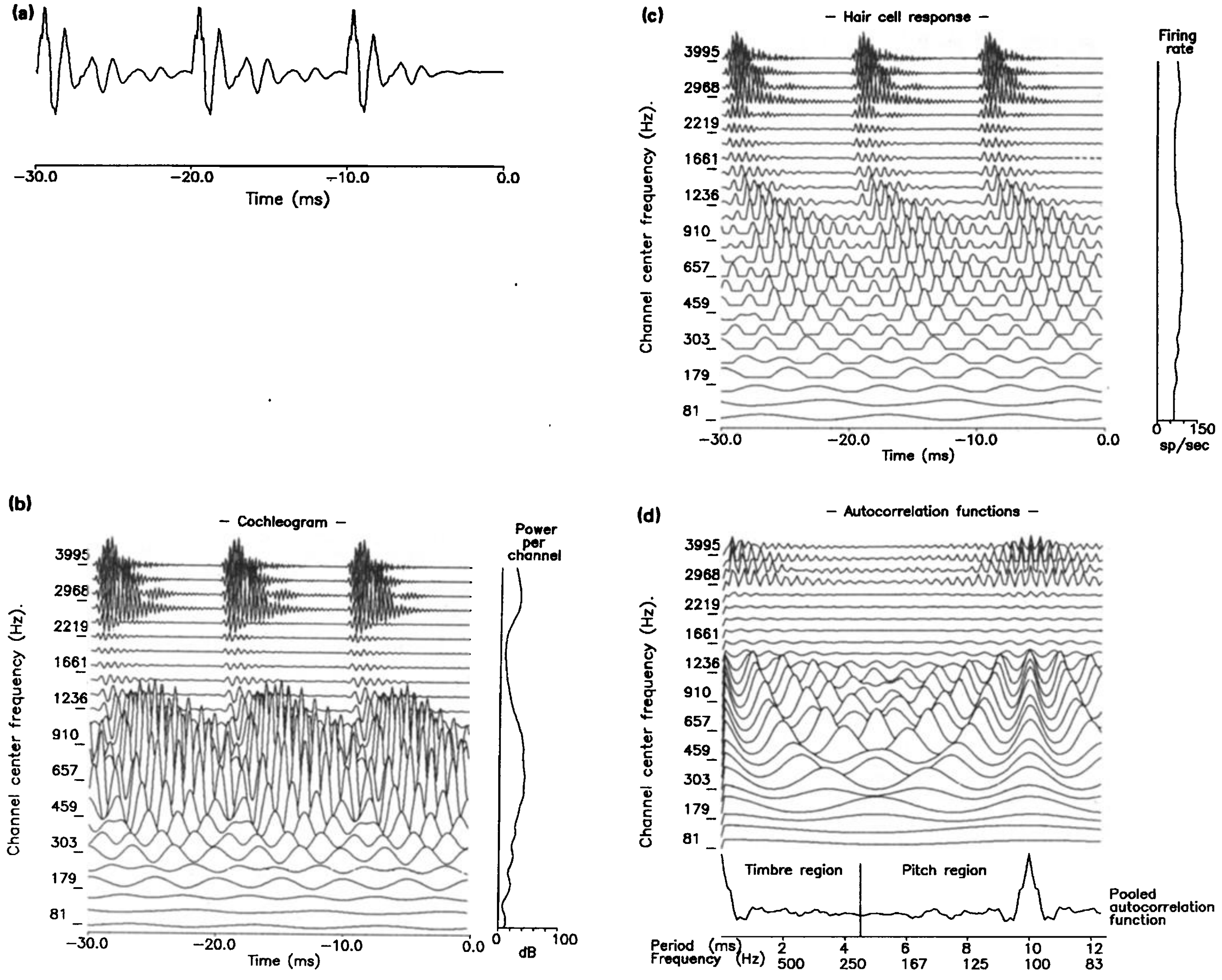

FIG. 3. Output of the model at intermediate stages. The figures show the last 30 ms of the model's operation in response to a 200 ms signal. (a) Stimulus input consisting of the single synthesised vowel "ah." (b) Cochleogram; response of individual channel bandpass filters. The vertical function to the right is the power output across channels. (c) Response of the simulated hair cells. Each function represents the probability of firing in a group of auditory-nerve fibers responding to a single location on the basilar membrane. The vertical function to the right is the firing rate in each channel over the period shown. (d) Channel running autocorrelation functions (ACF) and the pooled ACF formed by vertical summation across channels. The "timbre region" of the pooled ACF lies between periods 0.0001 and 0.0045 s. The "pitch region" lies between periods 0.0045 and 0.0125 s. Note that the very first point in each ACF is used to mark the baseline for that channel.

At the foot of Fig. 3(d), the pooled ACF is shown. This is computed by summing vertically all of the functions in the figure. Its main purpose is to highlight common features in the individual channels. A strong peak at 10 ms (100 Hz) shows that a common periodicity corresponding to the pitch of the stimulus is present in many channels. We define the region between 0.0125 and 0.0045 s (80 to 222 Hz) of the pooled ACF as the "pitch region" and use peaks in this region to identify possible pitches in the stimulus.

To the left of this region, between 0.0045 and 0.0001 s (222 Hz to 10 kHz), the pooled ACF gives information about the higher-frequency components or "timbre" of the stimulus. This "timbre region" is used by the model to identify the stimulus. The pitch region is excluded because it shows variation between different utterances of the same vowel. This algorithm compares the timbre region of the pooled ACF of the stimulus with a set of five templates, one for each of the five single vowels used in the study.

We accept that the values which define the pitch and timbre regions are somewhat arbitrary and will need to be given more attention in future studies. Our present concern here was to establish the *general principle* of using pitch estimates to segregate channels when separating sound sources. When dealing with the voices of children and some women, it will clearly be necessary to allow these to regions to overlap. For the moment, we have set this important issue to one side. We have also temporarily ignored the problem of "sub-octave" pitch estimates. When using the autocorrelation method of estimating pitch, a prominent peak in the ACF is always accompanied by similar peaks at half, third, etc. of

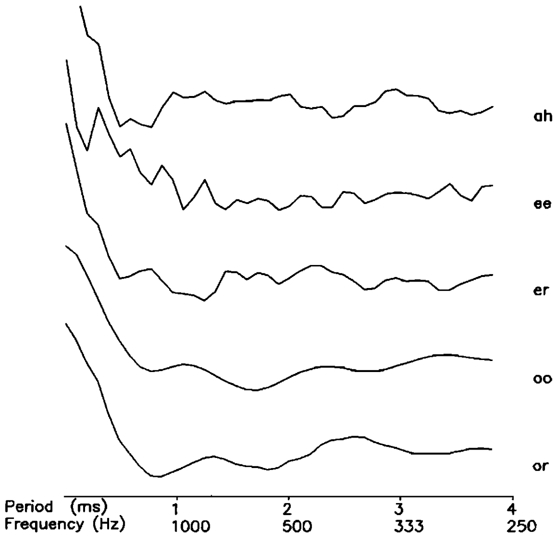

FIG. 4. Templates for single vowels. Templates are based on the "timbre region" of the pooled autocorrelation function (periods between 0.0001 and 0.0045 s). Each template is the average of six pooled ACFs for that vowel synthesised at pitches of 100, 101.5, 103, 106, 112, and 126 Hz.

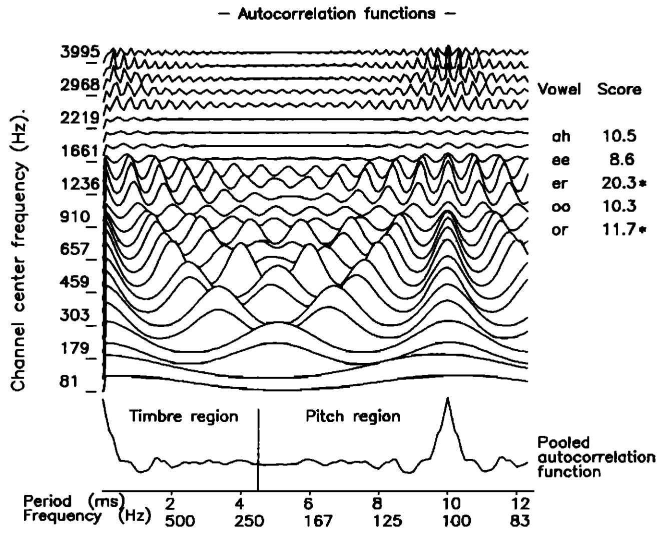

FIG. 5. ACFs and pooled ACF for a stimulus consisting of two vowels ("or" and "er") where both vowels are synthesised with the same f_0 (100 Hz). Eighty nine channels showed a pitch peak at the dominant pitch of 100 Hz. The inset values show the results of the matching algorithm. The starred values show the two best matching vowels.

the frequency of the first peak. In this implementation, we have simply taken the highest frequency peak in such a series. We achieve this automatically by restricting the ACF to values greater than 80 Hz while using pitches close to 100 Hz.

The template for for each vowel is created by averaging the pooled ACFs for six presentations of the vowel. Each vowel was presented in isolation with f_0's 100, 101.5, 103, 106, 112, and 126 Hz. Only the timbre region of the pooled ACF was used in the template. These pooled ACFs were then standardized so that

$$\sum t_i = 0 \quad \text{and} \quad \frac{\Sigma t_i^2}{N} = 1, \qquad (2)$$

where N is 40 (the number of points in the timbre region of the running autocorrelation function), and t_i are the points in the template corresponding to the timbre region of the pooled ACFs.

The five templates used are shown in Fig. 4. In the 100-Hz condition for a single vowel, the target vowel was always correctly recognized using these templates which is slightly better than the subjects of Assman and Summerfield who averaged 96% correct. The templates contain only periodicity information and do not refer in any direct way to the individual places (channels) where those periodicities were generated. They do not, therefore, necessarily show any pronounced peaks at periods corresponding to formants. This representation is not a simple transform of a spectral analysis of the stimuli.

The template matching was carried out using an inverse Euclidian distance measure

$$m = \left(\sum (t_i - s_i)^2\right)^{-1}, \qquad (3)$$

where t_i is the ith element of the standardized template and s_i is the ith element of the timbre region of the standardized pooled ACF for the stimulus. A larger value of m represents a better match to the template. We use m_1 to represent the best match and m_2 to represent the second best match.

A. Single f_0 double vowel example

Figure 5 shows the response of the model to a double vowel ("or" and "er") where both vowels have the same fundamental frequency of 100 Hz. A dominant pitch peak can be seen at the 10-ms period in the pooled ACF. A peak is regarded as dominant when it is the highest peak in the pitch region of the ACF.

The next stage in the recognition algorithm requires that all channels which do not show a peak at this pitch be excluded by setting the channel ACF to zero along its length. However, 89 channels do have such a pitch peak and only 11 channels have had to be excluded. This is taken as evidence that only one pitch is present by using the following decision rule: Rule 1: *one pitch is judged to be present in the stimulus when more than 80% of channels show a peak in their ACF at the period of the dominant pitch (at the highest pitch peak) in the pooled ACF.*

The parameter 80% was chosen because it results in error-free performance in discriminating single-f_0 from double-f_0 stimuli. This was a clear-cut discrimination and a range of criteria between 79% and 84% would have all served equally well.

Next it must be decided if there are one or two vowels present in this utterance. Figure 5 shows the result of matching the timbre region of the pooled ACF with all five templates. Vowels "er" and "or" have the best matches and would be chosen by the model as its best estimate of the two vowels. We accept that there are two vowels rather than one using decision rule 2: Rule 2: *In the single pitch condition (see rule 1), only one vowel is judged to be present if the match*

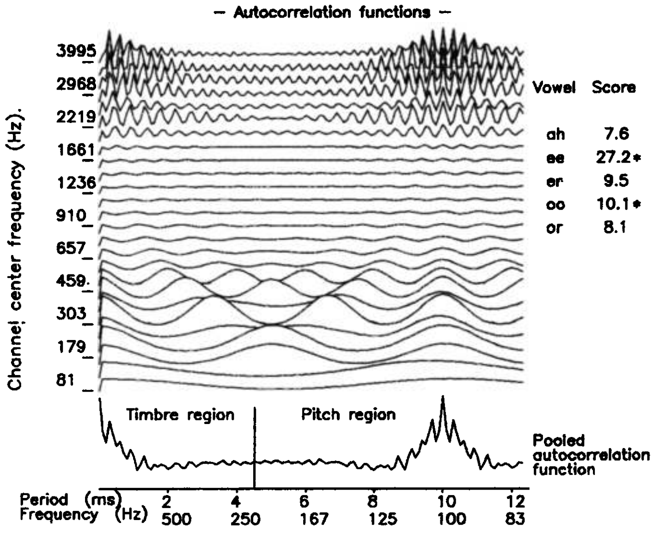

FIG. 6. ACFs and pooled ACF for a stimulus consisting of a single vowel ("ee") synthesized with a pitch of (100 Hz). Eighty eight channels showed a pitch peak at 100 Hz. The inset values show the results of the matching algorithm. The starred values show the two best matching vowels. Because the highest matching statistic ("ee"—27.2) is more than twice as great as the next highest ("oo"—10.1), the algorithm judges only one vowel to be present.

statistic for the second best match was less than half the matching statistic for the best match.

In our example, the match statistic for the second best match ("or"; $m_2 = 11.7$) is greater than half of the best match ("er"; $m_1 = 20.3$) and we, therefore accept that two vowels are present.

The choice of the ratio 2:1 was a somewhat arbitrary value and some fine tuning of the model might be attempted by changing it. However, rule 2 only applies to the case where a single pitch was found (20% of all stimuli). The balance between "hits" and "false positives" was such that modest changes in the ratio had little effect on the final results.

B. Single f_0, single-vowel example

Figure 6 shows the response of the model to a single vowel, "ee," presented at a f_0 of 100 Hz. The highest peak in the pooled ACF is at 0.01 s (100 Hz). Here, 88% of channels have peaks in their ACFs at this period. Rule 1 dictates that only one pitch be judged to be present. Figure 6 also indicates the match statistics of the five candidate vowels. Vowel "ee" is the best match because the value of its matching statistic is the highest of the five. Rule 2 dictates that only one vowel sound be judged to be present because the matching statistic for "ee" ($m_1 = 27.2$) is greater than twice that for "er" ($m_2 = 10.1$).

C. Two f_0 s, two-vowel example

Figure 7 shows a case where the two vowels ("er" and "ah") are presented with different fundamentals (100 and 112 Hz, respectively). To find the first vowel, we take the dominant pitch peak in the pooled ACF and note all channels which have a peak at the same period in their ACFs. The highest pitch peak in Fig. 7(a) is at 112 Hz and it can be seen that the ACF in some channels shows a peak at this value. In fact, 49% of channels show such peaks and we decide, using rule 1, that a second pitch must be present.

We retain these channels but set all other channel ACFs to zero and we obtain the representation shown in Fig. 7(b) which contains only channels characterized by a peak at 112 Hz. We presume that these channels are maximally excited by only one vowel. The pooled ACF at the bottom of the figure is based on these channels only and is used to identify the first of the component vowels. Figure 7(b) shows that "ah" is the best match.

The next step, therefore, is to return to the original set of ACFs and *remove* those channels associated with the highest pitch peak so as to investigate the identity of the second vowel. We set to zero all channel ACFs which have a pitch peak at the same value as the highest peak in the pooled ACF [Fig. 7(c)]. The remaining channels are the complement of the set used for the identification of the first vowel. The pooled ACF derived from these channels is presumed to relate to the second vowel and is used as the basis for the template matching procedure. The best match statistics in the figure show that vowel "er" must be chosen as the best candidate for the second vowel.

The steps involved in the separation and matching algorithm are summarized in the flow diagram given in Fig. 8. Clearly, the above account describes only successful examples for the purpose of illustration; performance was typically in the region of 45%–75% correct identification of both vowels in a pair.

II. MODEL EVALUATION

The model was evaluated by simulating an experiment involving human identification of double vowels (Assmann and Summerfield, 1990). They presented two vowels simultaneously to three subjects who were required to identify both vowels. The vowels were synthesised using Klatt's (1980) algorithm for cascade formant synthesis at a sampling rate of 10 kHz and lasted 200 ms (see footnote 2). Five different monophthongal (British) English vowels were used referred to here as "ah," "ee," "er," "oo," and "or." The ASCII approximations to IPA notation are /ɑ/, /i/, /ɜ/, /u/, /ɔ/.

A version of each vowel was prepared at each of six fundamental frequencies of 100, 101.45, 102.93, 105.95, 112.25, and 125.99 Hz (representing differences from 100 Hz of 0, 0.25, 0.5, 1, 2, and 4 semitones). Each stimulus consisted of a pair of these vowels. All possible pairs were used except that one vowel of the pair always had a f_0 of 100 Hz. The stimuli included double versions of the same vowel. For equal f_0's these are referred to as "single vowels" because they were physically indistinguishable from a single vowel (except for amplitude). There were 150 stimuli. Vowels began and ended simultaneously. Assmann and Summerfield supplied their full range of double-vowel stimuli to us in digitized form, exactly as used in their experiment.

Figure 9 shows the results obtained by Assmann and

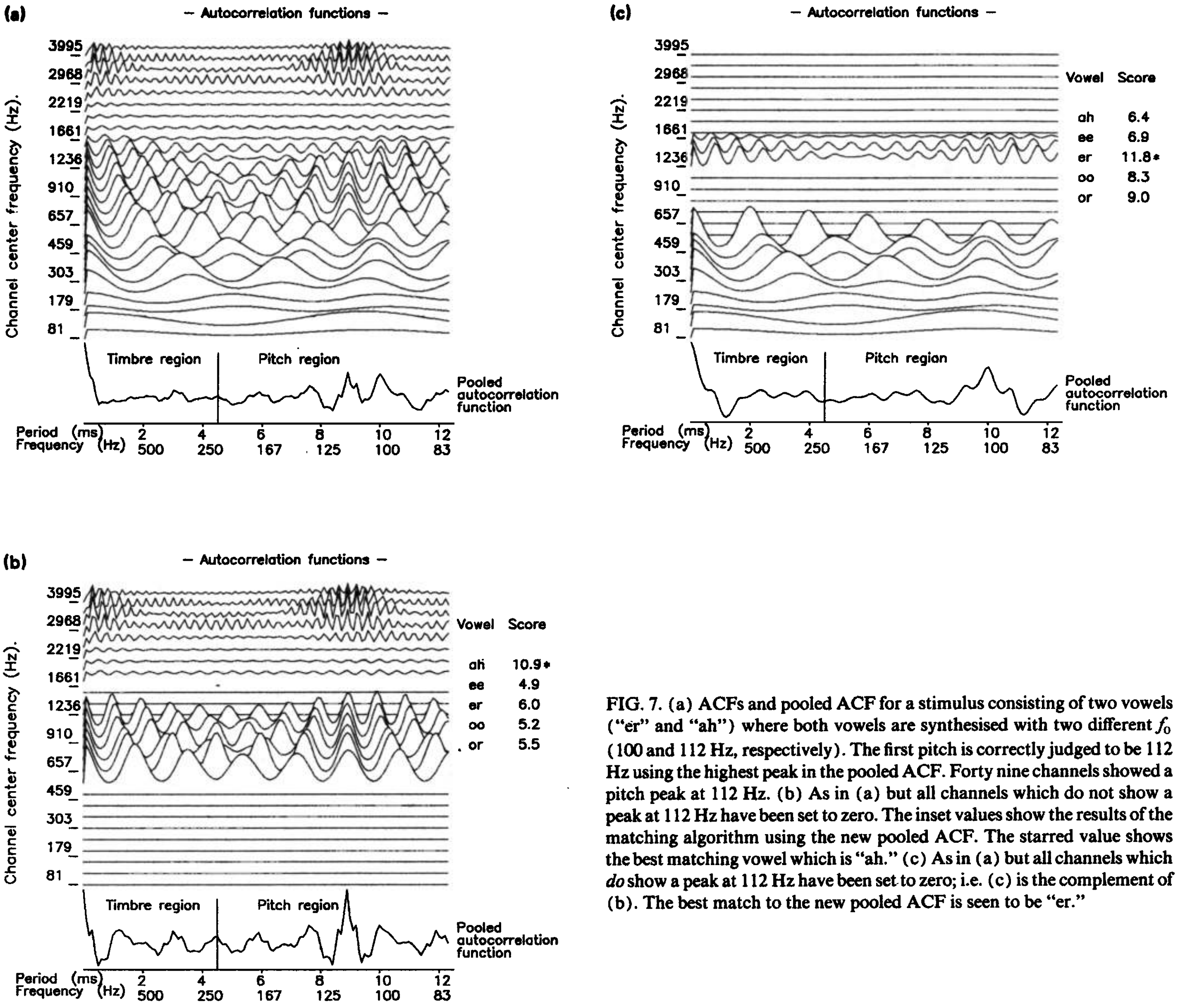

FIG. 7. (a) ACFs and pooled ACF for a stimulus consisting of two vowels ("er" and "ah") where both vowels are synthesised with two different f_0 (100 and 112 Hz, respectively). The first pitch is correctly judged to be 112 Hz using the highest peak in the pooled ACF. Forty nine channels showed a pitch peak at 112 Hz. (b) As in (a) but all channels which do not show a peak at 112 Hz have been set to zero. The inset values show the results of the matching algorithm using the new pooled ACF. The starred value shows the best matching vowel which is "ah." (c) As in (a) but all channels which *do* show a peak at 112 Hz have been set to zero; i.e. (c) is the complement of (b). The best match to the new pooled ACF is seen to be "er."

Summerfield (1990). Subjects were able to identify correctly both vowels of a pair on approximately half of the presentations even when both vowels had the same f_0. On a chance basis, only 7% double-correct responses would be expected. As the f_0 difference between the vowels was increased, the number of correct double identifications also increased. At a separation of four semitones, performance was 18% higher than at no f_0 separation. This improvement was restricted to the condition where vowels were 200 ms long; a second condition where the stimuli were 51.2 ms long showed no improvement. It is the longer condition which is simulated in this study.

The five templates were prepared in the manner described above and stored. Each of the 150 stimuli were processed by the model which generated estimates of the identity of the two component vowels of each pair. The model's response was judged to be correct only if both of the vowels of a pair were correctly identified, otherwise it was deemed to be in error.

The performance of a 173-channel version of the model is shown in Fig. 9(a) along with the data for Assmann and Summerfield's human listeners for comparison purposes. The overall level of correct responding is broadly comparable for the model and the human listeners. This correspondence should not be overemphasized, however, because of the many opportunities which the modeler has to optimize performance on a small data set. The important feature of the results, however, is the gradual rise in performance as the f_0 difference is increased.

A small dip at two semitones spoils the overall appearance of the model results but this is not a reliable feature. Small changes in the parameters can produce minor fluctuations because some decisions are very nicely balanced and a small change in the number of correct decisions has a large effect on the appearance of the graph. However, the overall picture is one of improvement as f_0 difference increases. In Fig. 9(b), we show the effect of varying the number of channels and all three show an overall trend of an increasing

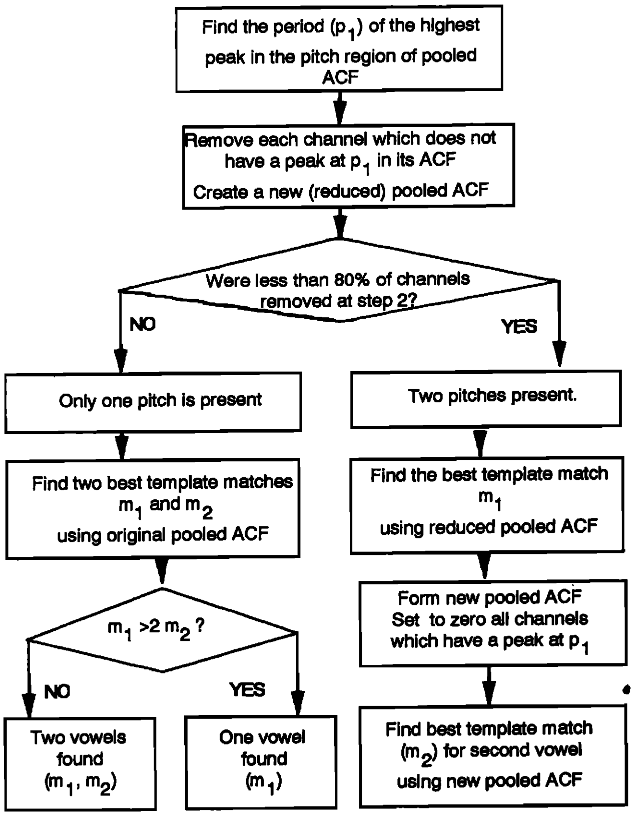
Find the period (p₁) of the highest peak in the pitch region of pooled ACF
Remove each channel which does not have a peak at p₁ in its ACF
Create a new (reduced) pooled ACF
Were less than 80% of channels removed at step 2?
NO
YES
Only one pitch is present
Two pitches present.
Find two best template matches m₁ and m₂ using original pooled ACF
Find the best template match m₁ using reduced pooled ACF
m₁ >2 m₂ ?
Form new pooled ACF Set to zero all channels which have a peak at p₁
NO
YES
Two vowels found (m₁, m₂)
One vowel found (m₁)
Find best template match (m₂) for second vowel using new pooled ACF

FIG. 8. Flow diagram for the steps and decision processes associated with the identification of the pairs of vowels. Rule 1 decides whether we have one pitch or two. When only one pitch is found, rule 2 (see text) decides if we have one vowel or two.

number of correct decisions as the f_0 difference increases.

It is interesting to note that the model does almost as well with 28 channels as it does with 173 channels. The pattern of improvement is, in fact, clearer with this reduced number of channels. We shall consider the number of channels again below when we look at the effect of introducing a random element into the firing of the simulated auditory-nerve fibers.

We considered the possibility that the results would have shown the improvement with f_0 difference even if the model had not used f_0 information; i.e., a simple matching stratagem based on the initial pooled ACF could have proved equally satisfactory. To check this, we forced the 173-channel version [see Fig. 9(a)] of the model to assume that only one pitch could be found; i.e., we forced it to take the left-hand path in Fig. 8. Rule 2, used for deciding whether there was one or two vowels, was left in place. These results are given in Fig. 10. These do not show a consistent rise in success as the f_0 of the second vowel rises. It is clear that the gradual improvement with f_0 separation in the model results was due to the channel separation procedure.

The algorithm almost always estimated the f_0 of the first vowel correctly (97%)—i.e., the first ACF value chosen was the closest possible value (given a bin width of 0.1 ms) to the true f_0 of one of the vowels—which is a result similar to that of Scheffers (1983a). The recognition algorithm does not require that the second pitch be estimated which is fortunate because it was typically much poorer (< 50%). The separation of channels into two subsets requires only one pitch estimate. Channels are segregated into those which do

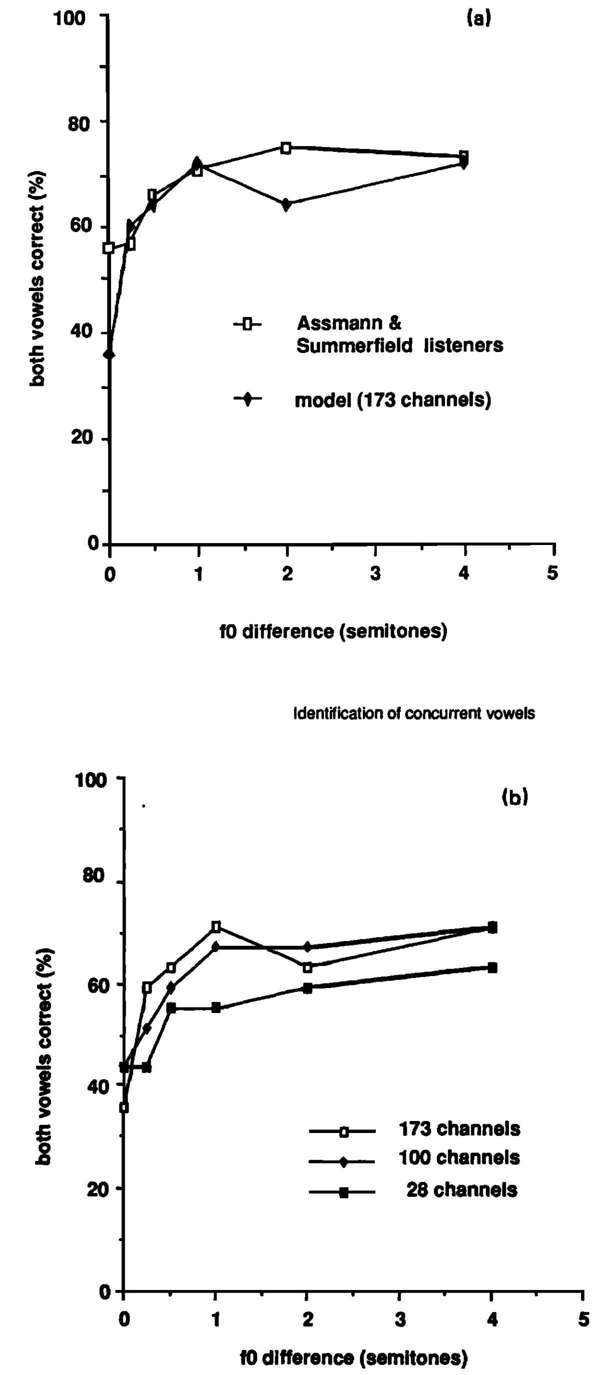

FIG. 9. (a) Response of a 173-channel version of the model to 150 paired vowels at various degrees of f_0 separation. Results for Assmann and Summerfield's (1990) listeners are given for comparison. The measure used is the percentage of vowel pairs where both vowels were correctly identified. (b) Model performance compared using 173, 100, and 28 channels.

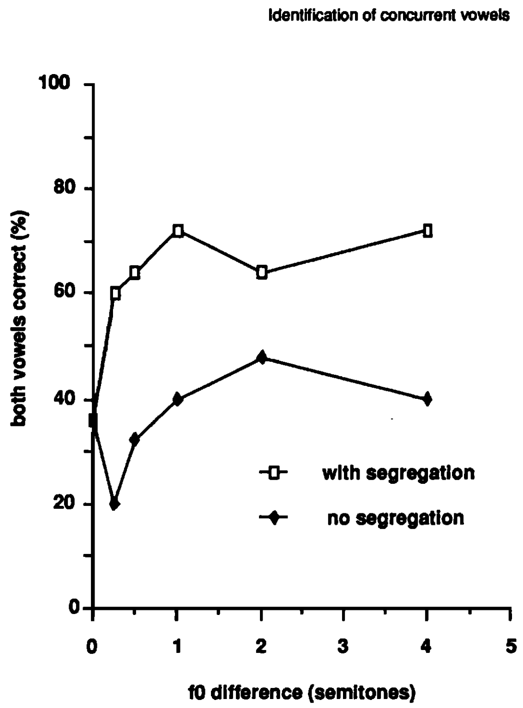

FIG. 10. Response of the model to the same stimuli as Fig. 9(a) except that the model was forced to respond without using any pitch information (i.e., rule 1 was set to judge only one pitch present for all stimuli). Results from Fig. 9(a) are included for comparison purposes.

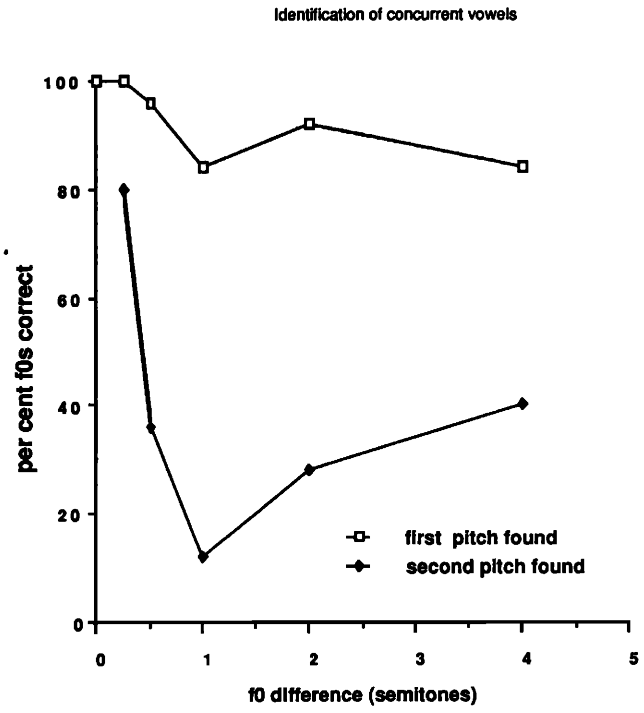

FIG. 11. Percentage correct estimation of the vowel fundamental frequency. The value for the second vowel is based only on those cases where the first vowel fundamental was estimated correctly.

and those which do not show an activity peak at this periodicity. Figure 11 shows the pitch estimation success as a function of f_0 separation. It does not improve as the two f_0's are more widely separated. We may infer that the improvement in recognition performance is not related to success in pitch estimation. It is more reasonable to assume that the separation of channels into two subsets becomes more clear cut as the f_0 separation increases because, at low separations, it is more likely that a given channel will be misclassified.

The model segregates the channels into two groups on the basis of its estimate of f_0 for the first vowel only. It does not attempt to estimate the f_0 of the second vowel. For interest we explored the model's accuracy in estimating the pitch of the second vowel. Figure 11 shows the likelihood of estimating the fundamental frequency of the second vowel correctly assuming that the model had correctly estimated the fundamental of the first vowel. Performance is relatively poor and it would appear that the estimate of the f_0 of the second vowel would form an inadequate basis for segregating channels and the model does not attempt to do so. Unfortunately, we have no human data concerning the ability of human listeners to estimate the pitch of a second simultaneous harmonic sound and the model's performance in this respect cannot be fully evaluated. Although, Beerends and Houtsma (1989) have reported that listeners can often identify the pitches of simultaneous two-tone complexes correctly for f_0 differences of two semitones or more.

The results shown above were calculated using an ACF time constant of 10 ms. Licklider (1951) had originally suggested a time constant of 2.5 ms in connection with pitch extraction. Figure 12 compares the time constants of 2.5, 10, and 25 ms using a 173-channel version of the model. The model is more accurate for equal f_0 vowel pairs when the

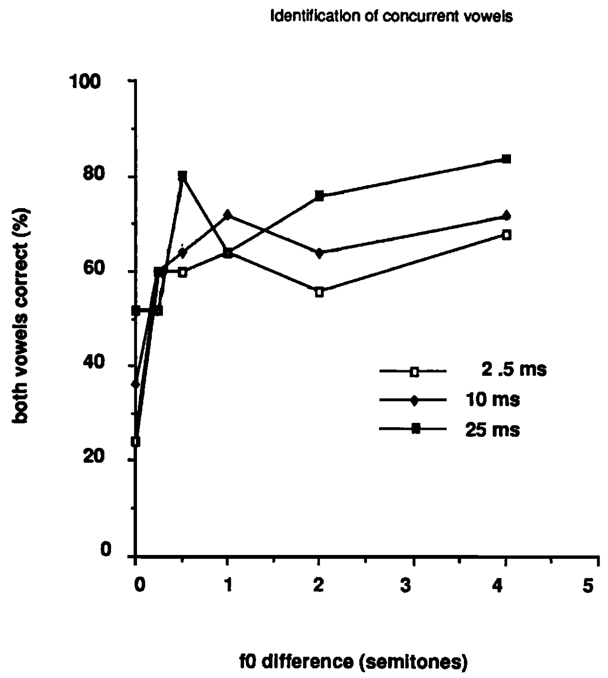

FIG. 12. A comparison of model performance for three running ACF time constants.

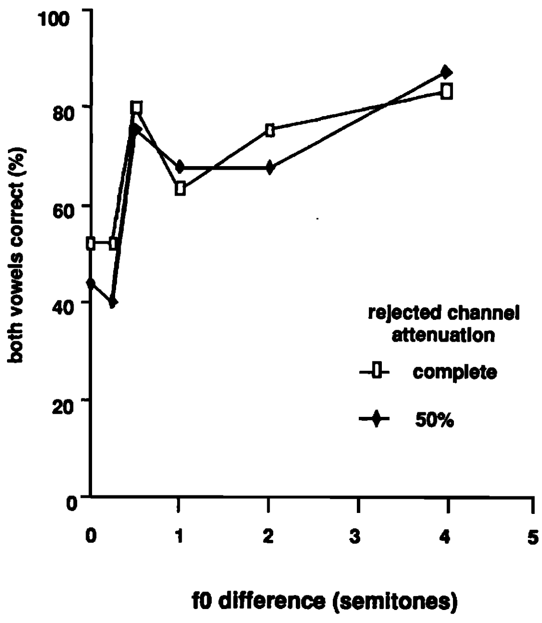

FIG. 13. A comparison of model performance for two methods of segregating channels. The standard method of attenuating a channel completely when it does not have a pitch peak in the appropriate location is compared with an attenuation of 0.5 in the height of the ACF values for that channel.

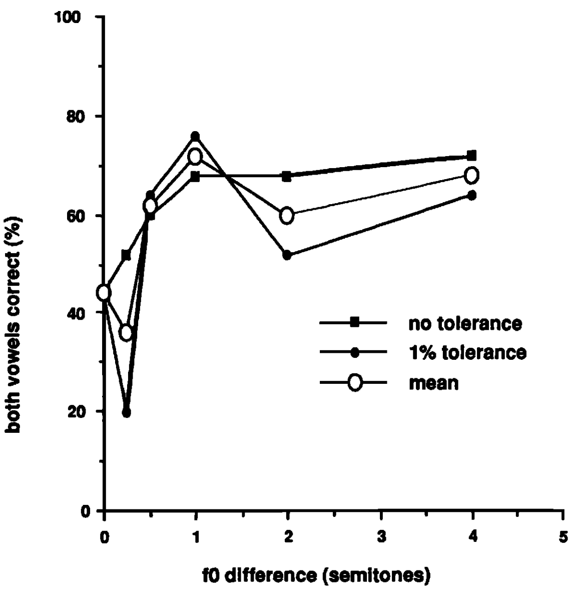

FIG. 14. A comparison of the performance of a 100-channel version of the model for two levels of tolerance for including individual channel autocorrelation functions within the first of two complementary sets of channels used for recognizing component vowels (see text).

time constant is lengthened. At greater f_0 separations this improvement is less consistent.

In an effort to keep the model simple, we had adopted the drastic expedient of dividing the channels into two quite separate groups on the basis of the first pitch found. In the event, this proved satisfactory but we considered the less drastic option of attenuating channels to 50% of their original strength rather than eliminating them altogether. We ran the computer program twice using a time constant of 25 ms; the first run was the same as that reported in Fig. 12; the second run was the same but involved an attenuation of only 0.5 of "rejected" channels. Figure 13 shows that this had very little effect on the results. It may be that the less extreme model will have advantages in some situations but this can only be tested when we have more data for human listeners to be used as a basis for comparison.

When assigning channels to one of the two sets, we adopted the rule that a channel would only be included in the first set if it had a peak in its autocorrelation function in the same bin as the highest peak in the pooled autocorrelation function. Otherwise, it was assigned to the complementary set. To study the effect of varying this rule, we relaxed the inclusion criterion to allow peaks within 1% (± 1 bin width) of the main pitch peak to qualify the channel for inclusion in the first set. The performance of a 100-channel version of the model is given in Fig. 14 along with the corresponding results for the zero tolerance condition [see Fig. 9(b)]. Performance drops markedly at 0.25 semitones before recovering to a raised performance at higher f_0 differences. Clearly, the increased tolerance does not improve the model's performance.

Unfortunately, the 10-kHz sampling of the model does not permit us to make intermediate tolerance levels. The average of the two functions (zero tolerance and 1% tolerance) is given for interest as a possible indication of what might have resulted if an intermediate tolerance had been possible. It is interesting that the average function gives similar values for 0 and 0.25 semitone f_0 differences which is a feature of the listeners' data [see Fig. 9(a)] suggesting that an intermediate tolerance would have given us a better fit between the model performance and the empirical results.

All the above demonstrations of the model's performance have been deterministic in nature in the sense that repeated runs of the program will give exactly the same result as long as no parameter has been changed. However, we do know that the firing of the auditory nerve is essentially stochastic in nature. To study this, we assigned a number of auditory-nerve fibers to each channel and, using the probability of firing function in combination with a random number generator, we were able to characterize the activity of each fiber in terms of action potentials. The output from each channel was then characterized as the number of action potentials which actually occurred within each time step. This was closely linked to the firing probability but had a random element associated with it which meant that each run was unique.

Human beings have approximately 30 000 afferent nerve fibers and we estimated that about 17 000 would probably be involved between center frequencies of 80 Hz and 4 kHz. Assuming that we had 100 channels, this would allow us 170 fibers per channel. Preliminary trials with this number of fibers showed that performance was abysmal both

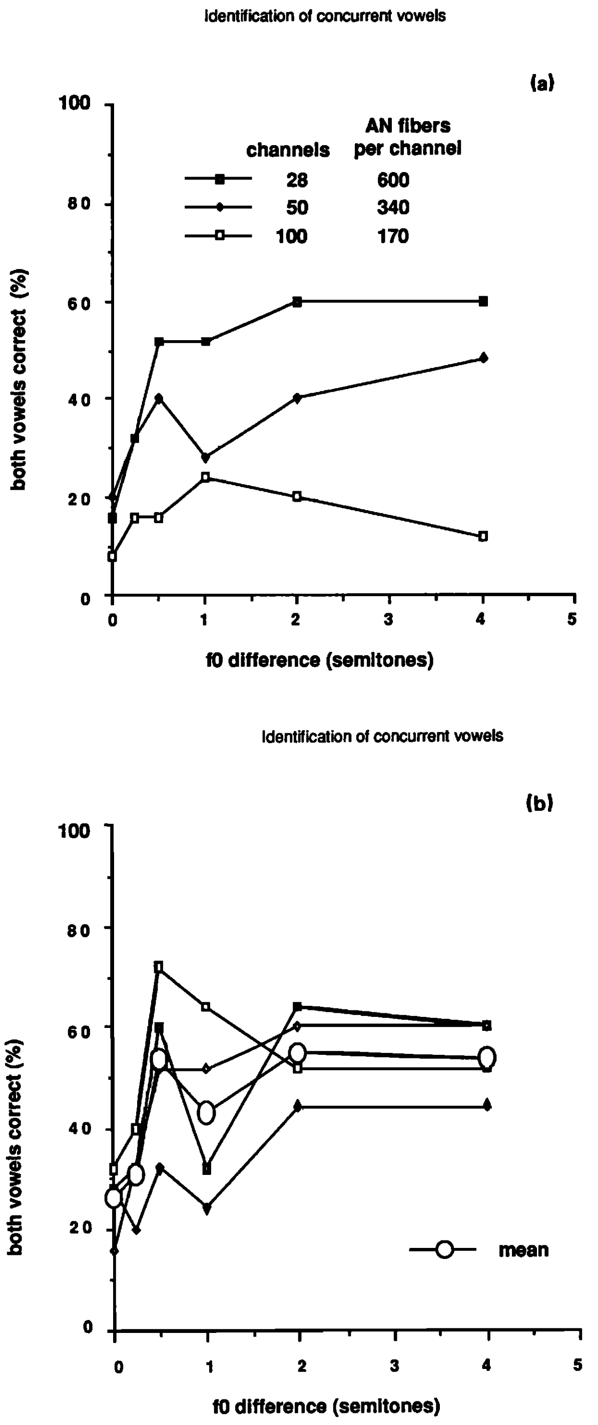

FIG. 15. (a) Performance of a stochastic version of the model (see text) compared for 100 channels with 170 fibers per channel, 50 channels with 340 fibers, and 28 channels with 600 fibers based on a single run for each. (b) Results for four runs of a 28 channel model with 600 fibers.

in terms of the recognition success which was only a little better than chance and the overall trend which was opposite to that expected [Fig. 15(a)]. This occurred because the individual channel autocorrelation functions were so noisy that the pitch peaks were rarely aligned appropriately.

Increasing the number of fibers per channel was capable of solving the problem but this was contrary to the known physiological limits. However, with only 28 channels and 600 fibers per channel the results were much better while remaining physiologically plausible by using 16 800 fibers. Using 50 channels with 340 fibers produced results which are intermediate and these are also shown in Fig. 15(a).

Figure 15(b) shows the results of four different runs using 28 channels and 600 fibers per channel. Each run is different and deviates somewhat from the expected pattern but the overall picture shows a sharp improvement over the first-half semitone with little improvement thereafter.

III. DISCUSSION

The model successfully simulates an important aspect of the perceptual data of Assmann and Summerfield (1990), i.e., the tendency to improved performance with increased fundamental frequency (f_0) difference between the two vowel components of the stimuli. The ability to capitalize on the presence of two different pitches is a structural feature of the model. Neither Scheffers (1983a) nor Assmann and Summerfield (1990) were able to model this crucial aspect of the data.

It is important to stress that our model is merely intended to illustrate one possible approach to the problem of identifying two simultaneous vowels. It is not a general purpose vowel recognizer. The template matching procedure was introduced merely to allow the model to demonstrate that it had segregated the information appertaining to the two vowels. Moreover, the two rules used for deciding whether the number of vowels present was one or two are manifestly *ad hoc*. It is also true that the rate of successful identifications obtained by the model is subject to fine tuning by the authors. The purpose of the article was to show that a principle has been uncovered whereby two simultaneously presented vowels can be segregated using pitch information such that the success rate increases as a function of the f_0 difference between the two vowels. To our knowledge, this is the only model of this process which has been successfully simulated numerically.

The speculative intuition guiding our investigation was that individual frequency-selective physiological channels have the capacity to inhibit one another but do not do so if they are responding to the same sound source. Evidence for a common sound source would arise if stimulus onset, offset, or other kind of amplitude modulation were correlated in two or more channels. Pitch is a particular example of an amplitude modulation which may be simultaneously present in a number of channels. Yost *et al.* (1989) have recently suggested a similar idea in the context of infrapitch (1–100 Hz) sinusoidal amplitude-modulation of signals.

In the case of a double vowel with two different steady pitches, we might expect two mutually inhibitory sets of channels to emerge. When one set dominates the other, we expect a single vowel to be more easily identifiable against the background of a competing vowel. A key question arises as to how the second vowel can be identified if one subset dominates and inhibits the other. This may be made possible by switching from one set to the other by applying a positive bias to the inhibited set before the stimulus has ceased. Assmann and Summerfield (1990) note that recognition is not assisted by separating the fundamentals of the two vowels if the stimulus is very short (51.2 ms) even though the recognition rate for double vowels with the same f_0 is just as good. In this case, we might suppose that the switch could not take place quickly enough. The challenge for the future is to establish this idea as a functioning neural network. The current study has mainly served to show that, in principle, it may be possible. This concept of mutual inhibition between channels is not essential to the model. It is mentioned here simply to give the reader some insight into the model's origin. In future work, we intend to explore the detailed implications of mutually inhibitory channels in a model which is more explicitly physiological.

Although most of the early testing of the model used 173 channels, we were surprised to note that a reduction to 28 channels did not seriously affect recognition accuracy and, if anything, produced a more faithful reproduction of the listener's improvement with increasing f_0 difference. When we later introduced stochastic effects into the auditory-nerve fiber activity, we again found that the best results were obtained with small numbers of channels. Large numbers of fibers per channel were required to give useful results and the restriction to approximately 17 000 fibers meant that these were better deployed in large groups over a small number of channels than in small groups over a large number of channels. Whether this is a pointer to the actual number of channels in the system, it is, of course, too early to say.

An interesting feature of the model is its sole reliance on periodicity information to achieve acceptable levels of vowel identification in the testing context of simultaneous vowels. We are not proposing that the nervous system uses only periodicity information in hearing, but we believe that this study shows that a great deal can be achieved when using only such information particularly when the sounds are harmonically structured. The tonotopic organization of the auditory system is exploited by the model to help separate information from different sound sources but the final identification is achieved on the basis of cross-channel aggregation of periodicity information. However, this identification takes place without *direct* reference to any place information whatsoever.

The model relies on two kinds of periodicity information; one related to pitch or amplitude information and the other related to timbre or short-term effects closely linked with the frequency components of the signal. Both have been extracted from our autocorrelation functions and pooled autocorrelation functions but we suspect that they are handled quite separately in the nervous system. Pitch information may be extracted by neurons in the cochlea nucleus (Frisina, 1983; Kim and Leonard, 1988). These cells can respond to and follow amplitude modulations between about 50 Hz and 500 Hz. In the inferior colliculus, units respond similarly but to a more restricted range of frequencies (Rees and Palmer, 1989; Rees, 1988). It is not known how higher-frequency periodicity information might be extracted although short-duration interval extraction represents less of a challenge to the modeling of nervous processing than the slower amplitude modulations involved in pitch.

Assmann and Summerfield (1990) have already shown that place-time models are superior to place models in the context of vowel identification. We have built upon their work by incorporating these principles. The innovation of the present model is to use a periodicity representation of the sounds as part of the identification process. This has allowed us to segregate the individual filtered channels into two groups in a way which would create difficulties if we were using a place representation of the sounds; to remove channels would produce gaps in a place representation. None of this establishes that place methods are inappropriate or not viable. Ingenuity alone may be required to create a suitably successful alternative place theory. That would be a useful development and spur to the development of crucial tests of these two opposing approaches.

IV. CONCLUSIONS

The model has shown that the segregation of simultaneous vowels can be assisted by a system of tonotopic channel segregation with vowel recognition taking place separately within the two complementary sets of channels. The explorations of the model have also shown that the number of channels can be as low as 28 and still show the basic phenomenon of improved recognition with an increase in f_0 difference. The basic result can be reproduced using either a deterministic model or a stochastic variant which respects the total number of fibers available in the human auditory nerve.

ACKNOWLEDGMENTS

We are particularly indebted to Peter Assman and Quentin Summerfield for their help in supplying stimuli, data, and explanations essential to this study. We would also like to thank Quentin Summerfield, Peter Assman, Brian Moore, Chris Darwin, John Culling, Bob Carlyon, Trevor Shackleton, and the two anonymous reviewers who read and commented on the first draft of this paper. This work was supported by a SERC research Image Interpretation Initiative grant.

[1] In the model, the signal is purely a number sequence and has no physical dimensions but we use the convention that a signal rms of 1 is treated as 0 dB1 (decibel *re*: rms = 1). Since the scale is arbitrary, we have chosen values which show a rough parallel with SPL ratings in psychological studies.

[2] Assmann and Summerfield (1990) have adapted this algorithm to accept noninteger values for the stimulus fundamental frequency.

Assmann, P. F., and Summerfield, Q (**1989**). "Modelling 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**). "Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680–697.

Beerends, J. G., and Houtsma, A. J. M. (**1989**). "Pitch identification of simultaneous diotic and dichotic two-tone complexes," J. Acoust. Soc. Am. **85**, 813–819.

Broadbent, D. E. (**1952**). "Failures in selective listening," J. Exp. Psychol. **44**, 428–433.

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phon. **10**, 23–26.

Chalikia, M. H., and Bregman, A. S. (**1989**). "The perceptual segregation 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.

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

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

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.

Egan, J. P., Carterette, E. C., and Thwing, E. J. (**1954**). "Some factors affecting multichannel listening," J. Acoust. Soc. Am. **26**, 774–782.

Frisina, R. D. (**1983**). "Enhancement of responses to amplitude modulation in the gerbil cochlear nucleus: single-unit recordings using an improved surgical approach," Special Report ISR-S-23, Ph.D thesis, Institute for Sensory Research, Syracuse University, Syracuse, NY.

Gardner, R. B. (**1989**). "An algorithm for separating simultaneous vowels," Br. J. Audiol. **23**, 170–171 (abs).

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.

Gerson, A., and Goldstein, J. L. (**1978**). "Evidence for a general template in central optimal processing for the pitch of complex tones," J. Acoust. Soc. Am. **63**, 498–510.

Goldstein, J. L. (**1973**). "An optimal processor theory for the central formation of pitch of complex tones," J. Acoust. Soc. Am. **54**, 1496–1516.

Halikia, H. M., and Bregman, A. S. (**1984**). "Perceptual segregation of simultaneous vowels presented as steady states and as crossing glides," J. Acoust. Soc. Am. Suppl. 1 **75**, S83.

Hartmann, W. M. (**1988**). "Pitch perception and the segregation and integration of auditory entities," in *Auditory Function: Neurobiological Bases of Hearing*, edited by G. M. Edelman, W. E. Gall, and W. M. Cowan (Wiley, New York), pp. 623–646.

Hewitt, M. J., and Meddis, R. (**1991**). "An evaluation of eight computer models of mammalian inner hair-cell function," J. Acoust. Soc. Am. **90**, 904–917.

Kim, D. O., and Leonard, G. (**1988**). "Pitch-period following response of cat cochlear nucleus neurons to speech sounds," in *Basic Issues in Hearing*, edited by H. Duifhuis, J. W. Horst, and H. P. Wit (Academic, London), pp. 252–260.

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

Lazzaro, J., and Mead, C. (**1989**). "Silicon modeling of pitch perception," Proc. Natl. Acad. Sci. USA. **86**, 9597–9601.

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

Licklider, J. C. R. (**1959**). "Three auditory theories," in *Psychology: A Study of a Science*, edited by S. Koch (McGraw-Hill, New York).

Meddis, R. (**1986**). "Simulation of mechanical to neural transduction in the auditory receptor," J. Acoust. Soc. Am. **79**, 702–711.

Meddis, R. (**1988**). "Simulation of auditory-neural transduction: Further studies," J. Acoust. Soc. Am. **83**, 1056–1063.

Meddis, R., and Hewitt, M. J. (**1991**). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery: I. Pitch identification," J. Acoust. Soc. Am. **89**, 2866–2882.

Meddis, R., Hewitt, M. J., and Shackleton, T. M. (**1990**). "Implementation details of a computational model of the inner hair-cell/auditory-nerve synapse," J. Acoust. Soc. Am. **87**, 1813–1818.

Moore, B. C. J. (**1982**). *An Introduction to the Psychology of Hearing* (Academic, London).

Moore, B. C. J. (**1986**). "Parallels between frequency selectivity measured psychophysically and in cochlear mechanics," Scand. Audiol. Suppl. **25**, 139–152.

Moore, B. C. J., and Glasberg, B. R. (**1987**). "Formulae describing frequency selectivity as a function of frequency and level, and their use in calculating excitation patterns," Hear. Res. **28**, 209–225.

Nedzelnitsky, V. (**1980**). "Sound pressures in the basal turn of the cat cochlea," J. Acoust. Soc. Am. **68**, 1676–1689.

Parsons, T. W. (**1976**). "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am. **60**, 911–918.

Plack, C. J., and Moore, B. C. J. (**1990**). "Temporal window shape as a function of frequency and level," J. Acoust. Soc. Am. **87**, 2178–2187.

Rees, A. (**1988**). "The influence of noise on neuronal responses to pure and amplitude-modulated tones in the guinea pig inferior colliculus," in *Basic Issues in Hearing*, edited by H. Duifhuis, J. W. Horst, and H. P. Wit (Academic, London), pp. 261–269.

Rees, A., and Palmer, A. R. (**1989**). "Neuronal responses to amplitude-modulated and pure tone stimuli in the guinea-pig inferior colliculus, and their modification by broadband noise," J. Acoust. Soc. Am. **85**, 1978–1994.

Scheffers, M. T. M. (**1983a**). "Sifting vowels: Auditory pitch analysis and sound segregation," Ph.D. thesis, University of Groningen, The Netherlands.

Scheffers, M. T. M. (**1983b**). "Simulation of auditory analysis of pitch: An elaboration of the DWS pitch meter," J. Acoust. Soc. Am. **74**, 1716–1725.

Stubbs, R. J., and Summerfield, Q. (**1988**). "Evaluation of two-voice separation algorithms using normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **84**, 1236–1249.

Stubbs, R. J., and Summerfield, Q. (**1990**). "Algorithms for separating the speech of interfering talkers: evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **87**, 359–372.

Triesman, A. M. (**1960**). "Contextual cues in selective listening," Q. J. Exp. Psychol. **12**, 242–248.

Viemeister, N. F. (**1979**). "Temporal modulation transfer functions based on modulation thresholds," J. Acoust. Soc. Am. **66**, 1364–1380.

Weintraub, M. (**1985**). "A theory and computational model of monaural auditory sound separation," unpublished Ph.D. thesis, Stanford University, Stanford, CA.

Weintraub, M. (**1987**). "Sound separation and auditory perceptual organisation," in *The Psychophysics of Speech Perception*, edited by M. E. H. Schouten (Martinus Nijhoff, Dordrecht, The Netherlands).

Yost, W. A., Scheft, S., and Opie, J. (**1989**). "Modulation interference in detection and discrimination of amplitude modulation," J. Acoust. Soc. Am. **86**, 2138–2147.

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