Concurrent vowel identification. III. A neural model of harmonic interference cancellation

Alain de Cheveigne
Centre National de la Recherche Scientifique/Université Paris 7, 2 place Jussieu, case 7003, F-75251 Paris
Cédex 05, France and ATR Human Information Processing Research Laboratories, 2-2 Hikaridai,
Seika-cho Soraku-gun, Kyoto 619-02, Japan

(Received 19 December 1995; revised 12 August 1996; accepted 22 November 1996)

This paper presents a “neural cancellation filter” capable of segregating weak targets from competing harmonic backgrounds, and a model of concurrent vowel segregation based upon it. The elementary cancellation filter comprises a delay line and an inhibitory synapse. Filters within each peripheral channel are tuned to the period of the competing sound to suppress its correlates within the neural discharge pattern. In combination with a pattern matching model based on autocorrelation functions summed over channels, the cancellation filter forms a model of concurrent vowel identification. The model predicts the number of vowels reported for each stimulus (when subjects are allowed to report one or two) and identification rates. It belongs to the class of “harmonic cancellation” models that are supported by experimental evidence that vowel identification is better when competing sounds are harmonic than inharmonic. Two alternative schemes using the same filter are also considered. One derives a “place” representation from the magnitude of the filter output. The other uses the ratio of filter input/output to select channels. © 1997 Acoustical Society of America. [S0001-4966(97)04104-0]

PACS numbers: 43.71.An, 43.71.Cq, 43.64.Bt, 43.66.Hg [WS]

INTRODUCTION

In “double-vowel” experiments, subjects identify concurrent steady-state synthetic vowels better when they differ in fundamental frequency ($F_0$) than when $F_0$’s are the same. Identification is also improved when vowels are modulated in frequency in a fashion that introduces instantaneous $F_0$ differences ($\Delta F_0$) (Summerfield, 1992; Summerfield and Culling, 1992a; Culling and Summerfield, 1995b), or when concurrent synthetic or natural voices have different intonation patterns (Brook and Nooteboom, 1982). A number of “$F_0$-guided” segregation models have been proposed to explain this effect [see de Cheveigne (1993a) for a review], that can be classified into three categories: spectral, spectro-temporal, and temporal. Spectral models, derived from Parsons’ (1976) harmonic selection method, require a spectral representation that can resolve individual harmonics. Assmann and Summerfield (1990) implemented such a model based on a computer model of peripheral filtering, and found that it lacked the necessary resolution. Spectro-temporal models (Assmann and Summerfield, 1990; Lea, 1992; Meddis and Hewitt, 1992; Weintraub, 1985) are less demanding in terms of resolution: Whereas a spectral model needs to resolve individual partials, a spectro-temporal model needs only to ensure that some channels are dominated by one vowel, and others by the other vowel.

Among spectro-temporal models, the model of Meddis and Hewitt (1992) is perhaps the most effective in terms of predictive power and physiological plausibility (Palmer, 1992). It works by partitioning auditory channels between those that respond with the same period as the “dominant” vowel, and those that do not. The dominant period is determined from the largest peak in a summary autocorrelation function (Licklider, 1959, 1962; Meddis and Hewitt, 1991).

Each group of channels is used to identify a vowel, and the model works as long as all channels are not dominated by the same vowel. If all channels are dominated by the same vowel, then there is no partition, and thus no advantage when vowels have different $F_0$’s. The model was applied to the stimulus set of de Cheveigne et al. (1997a) that included conditions for which one vowel was weaker than the other by 10 or 20 dB. For certain vowel pairs all channels were dominated by the stronger vowel, and the model therefore predicted no improvement with $\Delta F_0$. Nevertheless, strong $\Delta F_0$ effects were observed for these vowel pairs.

$F_0$-guided segregation models of the third category (temporal) do not explicitly require peripheral filtering to split information between vowels. De Cheveigne (1993a) proposed a time-domain filtering model that partitioned neural information within channels, rather than between channels, using a “neural cancellation filter.” That model was tested with physiological data recorded from the guinea pig auditory nerve, but it was not tested in detail with stimuli of double-vowel identification experiments. The purpose of this paper is to do so. The “neural cancellation filter” is first described in a simplified form that operates on discharge probability. The behavior of the filter is then illustrated with stimuli employed in double-vowel identification experiments, and three possible segregation schemes are outlined. Finally, one of these schemes is developed into a quantitative model of concurrent vowel identification, and its predictions are compared to the response counts and identification rates that were measured experimentally using the same stimuli.

I. THE NEURAL CANCELLATION FILTER

The time-domain “neural comb filter” involves delay and inhibition, and operates within auditory channels.
(auditory-nerve fibers or groups of similar fibers). It removes those spikes that form, with a preceding spike, an interval equal to the delay. The result is that the spike train contains fewer such intervals. Supposing that the filter is tuned to remove intervals equal to the period of some harmonic interference, correlates of that interference are cancelled. In this respect the filter is compatible with recent data showing that vowel identification is better when interference is harmonic than inharmonic (Summerfield and Culling, 1992b; Lea, 1992; de Cheveigné et al., 1995, 1997a, b). As in Meddis and Hewitt’s (1992) model, the neural filter partitions neural data into correlates of the dominant vowel, that are removed, and a remainder that is used to identify the weaker vowel. In contrast to their model, however, the partition occurs within individual channels rather than between channels.

A “neural cancellation filter” is implemented by a neuron with two synapses [Fig. 1(a)]. One is excitatory and fed by the input spike train, the other is inhibitory and fed by a delayed version of the input. Synapse and neuron properties are such that the neuron fires each time an input spike arrives, unless a spike arrives simultaneously along the delayed path. The result is that some spikes are weeded out, reducing the number of intervals equal to the delay [Fig. 1(b)]. de Chevigné (1993a) tested the filter with auditory-nerve fiber data recorded in the guinea pig in response to double-vowel stimuli (Palmer, 1990), and found that it partitioned neural data into patterns that clearly reflected the individual constituent vowels.

The behavior of a physiological mechanism following this principle would depend on many factors (number of afferents, pattern of innervation, spike generation process, spike rate, integration time constants, etc.). These are ignored in a simpler formulation, appropriate to a large ensemble of neurons, that captures the “vowel segregation” properties of the filter:

\[
o(t) = \text{MAX}(0, i(t) - \Phi[i(t-T)]),
\]

where \(o(t)\) is the output neuron’s firing probability, \(i(t)\) is input spike probability, and \(T\) is a lag parameter. The “MAX” operation reflects the fact that firing probability cannot be negative. \(\Phi[\cdot]\) is an operator designed to represent temporal smearing due to temporal integration, etc. In the following we further simplify the filter by supposing \(\Phi=\text{identity}:\)

\[
o(t) = \text{MAX}(0, i(t) - i(t-T)).
\]

The behavior of the filter is illustrated in Fig. 2.

Figure 2 represents discharge probability within an auditory-nerve fiber (or group of fibers) in response to a 100-Hz pure-tone pulse. The representation is idealized—ringing and adaptation effects are ignored, and transduction is represented as a simple half-wave rectification. The first peak of the cancellation filter input (a) finds its way to the output (b), but the following peaks are suppressed. In Fig. 2(c), the filter input is a half-wave rectified sum of sines, as might occur for example within an auditory-nerve channel in response to concurrent vowels with different \(F_0\)’s. One of the sines, the “target,” is the same 100-Hz sine as in (a), the other competing sine is lower in frequency (80 Hz) and higher in level by 10 dB. The filter is tuned to cancel the stronger competing sine. The output (d) consists of a peak followed by a modulated series of small peaks with the same spacing as the weaker target sine. This filter can be exploited to segregate two vowels on the basis of \(\Delta F_0\) according to at least three schemes that all suppose that the neural cancellation filter is tuned to cancel the period of the dominant vowel. The segregation schemes are illustrated in Sec. II with double-vowel stimuli used in experiments described in companion papers (de Chevigné et al., 1997a,b). In Sec. III...

FIG. 1. (a) Schema of neural filter. (b) Input and output spike trains. The delay is equal to delay parameter \(T\) of the filter. Input spikes that coincided with delayed spikes are no longer present in the output.

FIG. 2. (a) 100-Hz half-wave rectified sine wave input and (b) output of a neural cancellation filter. (c) Half-wave rectified sum of two sines of frequencies 80 and 100 Hz, differing in amplitude by 10 dB; (d) output of neural cancellation filter. The neural filter was tuned to cancel the stronger 80-Hz component.
one scheme is further developed into a model of concurrent vowel segregation, and compared with the experimental results.

The stimuli were based on Japanese vowels /a/, /i/, /u/, /e/, and /o/, synthesized at $F_0$’s of 125 and 132.5 Hz at a sampling rate of 16 kHz. Each stimulus consisted of either a single vowel or the sum of two vowels at a relative amplitude of 0 dB. The filter was tuned to cancel each of the constituent vowels in turn. Symbols indicate the frequencies of the first formants of each vowel. (b) Same, for vowel pair /a+/o/ (125 Hz, 8 ms) at a relative amplitude of 0 dB. The filter is tuned to the same two periods as in (a), one of which is the common period of both vowels.

that vowel might be identified. This is illustrated in Fig. 3(a) for the vowel pair /a+/o/ at $\Delta F_0=6\%$. The thick, continuous line corresponds to the profile of unfiltered activity over channels of the gammatone/hair-cell model. The dotted lines correspond to the profile of cancellation filter outputs when the filter delay parameter is set to each of the vowel periods in turn. When the filter is tuned to the period of /o/ (7.55 ms) the profile has peaks at the first two formants of /a/ (indicated by symbols). When it is tuned to the period of /u/ (8 ms) the profile may be interpreted as reflecting the first two formants of /o/. Filtered profiles might thus serve as a substrate for vowel identification. When the vowel $F_0$’s are the same, almost all activity is suppressed if the filter is tuned to the common period, 8 ms [Fig. 3(b), dot-dash line near axis], whereas if it is tuned to a different period (7.55 ms) the profile is not specific to either vowel (dotted line). The fact that filtered profiles reflect constituent vowels better at $\Delta F_0=6\%$ than at $\Delta F_0=0\%$ could explain why identification is better in the former case than in the latter.

The filter is still effective when /u/ is at 0, −10, or even −20 dB (lower thin lines in Fig. 4) relative to the competing vowel /o/. At 10 dB the profile is not so clear, and at 20 dB it is the same as that obtained for /u/ in isolation (thick continuous line). Both the filtered and the unfiltered profiles for /u/ in isolation (uppermost in Fig. 4) lack clear peaks at formants $F_1$ and $F_2$ of /u/, due to the compressive saturation properties of the hair-cell model. The driving level was cho-

<table>
<thead>
<tr>
<th>$F_0$</th>
<th>125</th>
<th>3.7</th>
<th>8.00</th>
<th>132.5</th>
<th>3.8</th>
<th>7.55</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>312</td>
<td>7.5</td>
<td>3.21</td>
<td>468</td>
<td>9.9</td>
<td>2.14</td>
</tr>
<tr>
<td>$F_2$</td>
<td>1219</td>
<td>16.9</td>
<td>0.820</td>
<td>781</td>
<td>13.4</td>
<td>1.28</td>
</tr>
<tr>
<td>$F_3$</td>
<td>2469</td>
<td>22.7</td>
<td>0.405</td>
<td>2656</td>
<td>23.3</td>
<td>0.377</td>
</tr>
<tr>
<td>$F_4$</td>
<td>3406</td>
<td>25.3</td>
<td>0.294</td>
<td>3281</td>
<td>25.0</td>
<td>0.304</td>
</tr>
<tr>
<td>$F_5$</td>
<td>4200</td>
<td>27.0</td>
<td>0.238</td>
<td>4200</td>
<td>27.0</td>
<td>0.238</td>
</tr>
</tbody>
</table>

TABLE I. Formant frequencies in units of Hz and ERB (Moore and Glasberg, 1983), and equivalent formant periods in ms of the two vowels used to illustrate the segregation schemes.

II. THREE SEGREGATION SCHEMES

A. Place scheme

In channels that respond only to the dominant vowel, the output of the cancellation filter should show an initial peak followed by zero output [Fig. 2(a)]. If the channel is influenced by the other vowel, even only weakly, the output following the initial pulse should not be zero [Fig. 2(b)]. Supposing that the amplitude of this residue is proportional to the relative amplitude of the weaker vowel at the frequency of the channel, the profile of cancellation filter outputs over channels should produce a place representation from which...
When both vowels have the same filtered SACFs differ from each other, and from the unfiltered response is overall dominated by that vowel. When filtering is not effective to read, we plot the square root of the SACF. Figure 5(a) (continuous line) shows the square-root SACF for the vowel pair /u/+/o/ at 0 dB relative amplitude and $\Delta F_0=6\%$. There is a peak at the period of /o/ (7.55 ms), suggesting that the unfiltered response is overall dominated by that vowel. When the filter is tuned to the period of /o/ (7.55 ms), there is a peak at the period of /u/ (8 ms). When it is tuned to the period of /u/, the peak is instead at the period of /o/. The two filtered SACFs differ from each other, and from the unfiltered SACF. When both vowels have the same $F_0$ (125 Hz), filtering is not effective [Fig. 5(b)]. If the filter is tuned to the common period of the vowels (8 ms), it wipes out all activity. If it is tuned to a different period (7.55 ms), it produces a pattern similar to the unfiltered pattern. This could explain why subjects identify vowels less well at $\Delta F_0=0\%$ than when $F_0$’s are different.

The filter tuned to cancel /o/ remains effective even when /u/ is weak. Figure 6 displays square-root SACFs restricted to lags shorter than 4.5 ms (the “timbre region”) that Meddis and Hewitt’s model used to identify vowels. The unfiltered square-root SACF for the isolated vowel /u/ is represented by the topmost line. Other lines represent SACFs obtained after filtering with a cancellation filter tuned to suppress 8 ms (period of /o/). Each line represents a different proportion of /u/ in the input stimulus. The filtered response for /u/ alone is represented by the thick continuous line, and thin lines represent responses for the vowel pair /u/+/o/. All filtered patterns are similar to the filtered pattern for /u/ alone, which itself is similar (with the exception of the dip near 0.5 ms) to the unfiltered pattern for /u/ alone. None of these patterns reflects the presence of /o/, even at $-20$ dB where the /u/+/o/ stimulus contains mostly /o/. A concurrent vowel identification model based on this scheme is presented in Sec. III.
The SACFs of Fig. 6 are not easy to interpret in terms of spectral features, apart from the fact that peaks at 0.8 ms and multiples seem to reflect the F2 of /u/ (1219 Hz). Figure 7 shows Fourier transforms derived from the first 8-ms portion of the SACFs, together with the spectral envelope used to synthesize /u/ (topmost line). For the vowel pair /u/+/o/ at 0 and −20 dB, the filtered SACF Fourier transform shows peaks at F2 and F3 of /u/ (lowest thin lines). Both patterns resemble the filtered and unfiltered patterns for /u/ alone (thick lines). All show similarities with the spectral envelope used to synthesize the vowel /u/ (uppermost line), apart from a complete lack of evidence of F1. The Fourier transform reinforces the conclusion that the filtered SACFs carry evidence of /u/ (F2 and F3), and no evidence of /o/. It is, however, debatable whether the auditory nervous system could carry out a Fourier transform to exploit this evidence.

C. Channel selection scheme

Instead of calculating ACFs for all cancellation-filtered channels and pooling them, we may use the filter output-to-input ratio as a criterion to select channels, as in Meddis and Hewitt’s (1992) model. The main difference with their scheme is that the ratio threshold may be tuned to be more or less stringent in the selection of channels. Channel selection is also analogous to the scheme of Sec. II B, in that channels are given nonuniform weights before summation. The weighting is graded in one case, all-or-none in the other. A more important difference is that channels were filtered by the cancellation filter in the previous scheme, whereas here they are not. Filtering causes distortion that might interfere with template matching (compare filtered and unfiltered SACFs for /u/ in Fig. 6), so the lack of filtering in the present channel selection scheme might be an advantage.

Channel selection was implemented by filtering all channels with the cancellation filter tuned to the period of the competing vowel, and calculating the ratio between rms filter output and input (integrated over a 100-ms square window starting 50 ms from stimulus onset). Channels for which the output/input ratio fell below a threshold (arbitrarily set to 0.25) were considered dominated by the competing vowel, and discarded. ACFs for remaining unfiltered channels were calculated and summed to form a SACF representative of the target vowel. Figure 8 illustrates the result of this process for target /u/ mixed with competing vowel /o/. The uppermost line is the SACF for /u/ alone, calculated from all channels. The lowermost line is for /o/ alone, for which no channels were selected. The remaining lines are SACFs in response to the vowel pair /u/+/o/, calculated by summing channels that survived the selection process. When /u/ is at −20 dB, a few channels are selected in the vicinity of 17 ERB, close to the F2 of /u/ (see Fig. 4), as suggested by the period of the ripple in the SACF. However, the peak near 7.5 ms in the SACF shows that those channels were predominantly responding with the period of /o/. The same is true when /u/ is at −10 dB, and to a certain degree, 0 dB. At all three amplitudes the SACFs are rather different in shape from the SACF obtained in response to /u/ alone. In the previous scheme of Sec. II B, a cancellation filter processed each channel and eradicated all evidence of the competing vowel. In the present scheme, unfiltered selected channels are more subject to “talk through” from the competing vowel. The channel selection scheme could form the basis of a concurrent vowel identification model similar to that of Meddis and Hewitt (1992), but we do not attempt to do so in this paper.

III. A MODEL OF CONCURRENT VOWEL IDENTIFICATION

This section takes the segregation scheme of Sec. II B and develops it into a model of concurrent vowel identification. For each stimulus, the model predicts (a) the number of vowels reported and (b) the name(s) of the vowel(s) reported. The aim is to prove that such a model can be built on the basis of the neural cancellation filter, and to work out enough detail to make a comparison with experimental data. The important question is whether the neural cancellation filter is consistent with major qualitative features of experimental results (dependency on $\Delta F_0$, amplitude, harmonicity, etc.). The quantitative match and specific details of the identification model are considered secondary.
A. Implementation details

After peripheral filtering and hair-cell transduction, the running autocorrelation function (ACF) of the auditory-nerve discharge probability within each channel was calculated with a time constant of integration of 12.5 ms. ACFs were sampled at 150 ms from stimulus onset, summed, and an estimate of the “dominant period” was obtained from the largest peak within the SACF in the range 5–10 ms. That range was chosen to include both periods of the stimulus vowels, and exclude their subperiods (harmonics) and super-periods. Next, the auditory-nerve discharge probability within each channel was fed through a cancellation filter tuned to the dominant period, and ACFs were calculated and summed to obtain a second, filtered SACF pattern. The average rms output-to-input ratio of the cancellation filter was also calculated. Following Meddis and Hewitt (1992), vowel identification was performed by matching SACFs to templates obtained for single vowels. SACFs were restricted to the portion between 0 and 2.5 ms to reduce the influence of $F_0$, and normalized to have a mean of zero and a variance of one. The uniform Euclidean distance was calculated between each token and all templates, and the closest template was chosen as the match.

Based on these elements (output-to-input ratio, filtered and unfiltered SACF patterns, templates) and two threshold parameters T1 and T2, the model predicts the number of vowels reported for each stimulus and their names. The algorithm is as follows. The output-to-input ratio is compared to threshold T1. If it is larger, the model decides that two vowels are present. The first is determined from the best match to the unfiltered SACF. The second is determined from the best match to the filtered SACF, unless this produces the same match as the first vowel. In that case the second best match to the unfiltered SACF is taken instead (this rule enforces the restriction that both vowels must be different). If the output-to-input ratio is smaller than T1, the model discards the filtered SACF and attempts two matches of the unfiltered SACF. If the distance ratio between the two matches is above threshold T2, the model predicts two responses: the two matches. Otherwise it predicts one response: the best match. The model thus predicts a single response only if the cancellation filter residual is small and the nonfiltered SACF can unambiguously be matched by a single template.

B. Predicted effects of $\Delta F_0$ and amplitude

The symbols in Fig. 9 represent experimental results obtained by de Cheveigne et al. (1997a), and the lines are predictions of the model for the same stimuli. Consider first $\Delta F_0=0$% (filled symbols and thick lines). Threshold T2 was adjusted to obtain a reasonable match for the number of vowels reported [Fig. 9(a)]. The value chosen (0.2) also produced identification rates that are close to experimental values [Fig. 9(b)].

Consider now results at $\Delta F_0=6$% (open symbols and thin lines). Threshold T2 was maintained at its previous value and threshold T1 was set to 0.01. The number of vowels reported is well predicted [Fig. 9(a)], but the prediction for identification rate is somewhat less good [Fig. 9(b)]. The discrepancy at 0 and 10 dB may be attributed to the inaccurate period estimation in the first stage of the model: If estimates are corrected, the predictions follow the dotted line in Fig. 9(b). The discrepancy at $-20\,\text{dB}$ may be attributed to the fact that the cancellation filter is implemented linearly and works well over a large dynamic range (Fig. 6), whereas a physiological implementation would certainly have a more limited dynamic range. The model successfully explains the one aspect of the experimental data that Meddis and Hewitt’s (1992) model had difficulty accounting for the improvement in identification with $\Delta F_0$ when the target is very weak. It also nicely predicts the number of vowels reported for each stimulus.

One should not give excessive weight to the quantitative match between predictions and experimental data: Each condition was tested using only 20 vowel pairs, so predicted rates have a granularity of 5%. Small changes in parameters may cause large jumps in predicted rate. Assmann and Summerfield (1989) and Assmann (1995) used techniques to transform distances (or neural net activations) to better-behaved “probability” estimates. We did not follow that course, because it introduces extra assumptions and parame-
eters, and because our simpler method was sufficient for the
points we wished to make.

C. Predicted effects of harmonicity

The model was applied to the stimuli of de Cheveigne et al. (1997b), using the same parameters as before. The symbols in Fig. 10 are experimental results, and the lines are predictions of the model. The model correctly predicts that target harmonicity has no effect on identification rate, and that identification is better when the competing vowel is harmonic rather than inharmonic [Fig. 10(a)]. Overall, predicted identification rates are too high, as was observed in Fig. 9(b) at $\Delta F_0=6\%$ and $-20$ dB. The number of responses predicted and observed is plotted in Fig. 10(b). The predictions are again a bit too high, and neither the predictions nor the experimental data vary much between conditions.

IV. DISCUSSION

Section I presented the neural cancellation filter. Section II showed that it could be exploited in several ways to segregate concurrent vowels that differ in fundamental frequency. Finally Sec. III elaborated one of those schemes into a model of concurrent vowel identification.

A. Neural cancellation

Removing certain spikes from the spike train (Fig. 1) suppresses evidence of a competing vowel, but we would be wrong in concluding that that evidence was “contained” in those particular spikes. Rather, the filter adjusts the interval statistics of the spike train so that they no longer reflect the competing vowel. Equation (1) is a crude account of the statistics of the spike selection process, but it produces patterns similar to simulations with recordings from the auditory nerve (de Cheveigne, 1993a). A linear version of Eq. (1) (time-domain comb filter) was used for speech interference rejection by de Cheveigne (1993b). Subtraction in Eq. (1) is analogous to the cancellation step in the equalization and cancellation model of binaural interaction (Durlach, 1963; Culling and Summerfield, 1995a), which could be implemented by a spike selection process similar to the one illustrated in Fig. 1.

So far there is little physiological evidence for delay lines of the length required to cancel vowels with fundamental periods of the order of 10 ms (100 Hz), although similar delays would be needed to extract their pitch with an autocorrelation model (Licklider, 1959, 1962; Meddis and Hewitt, 1992; Cariani and Delgutte, 1996). Shorter delay lines have been shown to exist for binaural processing of interaural time delay cues (Yin and Chan, 1990).

B. Peripheral selectivity

Whereas most segregation models rely critically on peripheral filtering (including the schemes of Sec. II A, and C), the scheme of Sec. II B developed in Sec. III does not give it a major role. Cancellation is performed within channels, to be sure, but channel bandwidth is not a critical parameter, and channels separated by the cochlea are merged after the cancellation filter. Peripheral filtering can, however, be assigned two roles as follows. (a) It isolates channels in which the target-to-background ratio is more favorable than within the stimulus itself, which is useful if the linearity or dynamic range are limited. Thus peripheral selectivity might play a secondary role in the principle of the model but be essential for its implementation. (b) A signal that is split into band-limited channels can be reconstructed despite within-channel nonlinearities (Slaney et al., 1994). Peripheral filtering may thus protect acoustic information from being degraded by the cascaded nonlinearities of peripheral transduction and cancellation filter.2 In other words, the cochlea splits the acoustic signal redundantly into channels that differ by their phase and spectral content so that information that is degraded within each channel may survive in the ensemble. The neural cancellation filter should be seen as a segregation mechanism that operates together with, rather than instead of, peripheral filtering.

C. The place alternative

The scheme of Sec. II B was chosen as a basis for the identification model of Sec. III on criteria of ease of description and similarity, in its pattern-matching stage, with the model of Meddis and Hewitt (1992). The alternative schemes are worth considering. The place profile (Sec. II A) has the advantage of displaying information in a familiar spectrum-like form. A difficulty is that the profile for isolated vowels shows little vowel-specific evidence (Fig. 4), as observed by Sachs and Young (1979) in auditory-nerve fiber populations. Profiles might be better preserved in a population of low spontaneous rate fibers. One of our examples showed that the place profile can carry information, such as the formant $F_1$ of /a/ (Fig. 4) that is apparently not coded in the SACF (Fig. 7). The auditory system does not share the need for simplicity that constrains our models, so it might exploit both the SACF and the place profile, or even the entire ACF array
from which both derive. This is reminiscent of a many-to-
many neural network proposed by Licklider (1959) for pitch
perception.

D. The channel selection alternative

The channel selection scheme of Sec. II C eliminated
channels that left little residual after cancellation, whereas
Meddis and Hewitt (1992) eliminated channels that had a
peak in the ACF at the competing vowel’s period. The
scheme of Sec. II C has the advantage that it depends upon a
threshold that may be tuned to avoid rejecting all channels
when a target vowel is weak. A version of Meddis and
Hewitt’s model based on this scheme might thus account for
the improvement of identification with $\Delta F_0$ at low target-to-
background levels (de Cheveigné et al., 1997a). That idea is
not without problems, however, as a less stringent selection
criterion retains channels that are more strongly affected by
the competing vowel.

E. Recent experimental results

McKeown and Patterson (1996) found that one vowel of
acoustic pairs was always identifiable whatever the $\Delta F_0$,
even if the stimulus was very short (one fundamental pe-
riod). Identification of the other vowel improved with $\Delta F_0$,
but only if the stimulus was at least three of four periods in
duration. The model of Sec. III is consistent with this behav-
or; the first vowel is determined independently from any
segregation process, and the second after application of a
cancellation filter tuned to an estimate of the period of the
first. An $F_0$ estimate of sufficient precision for effective can-
cellation probably requires some time to develop. Indeed,
Robinson and Patterson (1995) found that perception of the
pitch of a vowel required several fundamental periods, whereas perception of the timbre needed only a single pe-
riod.

Culling and Darwin (1993) synthesized concurrent vow-
els with $\Delta F_0$’s interchanged between vowels in the region of
the second formant (the $F_2$ of one vowel belonged to the
same harmonic series as the $F_1$ of the other). A harmonic
cancellation model tuned to either period should produce
categorical patterns comprising the $F_1$ of one vowel and the
$F_2$ of the other. Nevertheless, Culling and Darwin found that
identification improved with $\Delta F_0$, although less than for nor-
mal concurrent vowels. Several tentative explanations can be
offered: (a) The chimerical segregated patterns may support
identification better than the unsegregated pattern available
at $\Delta F_0=0\%$ (for example, the pattern-matcher might make
do with one formant while ignoring the other); (b) the $\Delta F_0$-in-
harmonicity of the stimulus may make it easier to interpret as containing two vowels, and thus indi-
rectly improve identification; (c) harmonic cancellation may
use different delays in different frequency regions, in a fash-
on analogous to Culling and Summerfield’s (1995a) modi-
ification of the equalization and cancellation model of binaural
interaction, and (d) Segregation at small $\Delta F_0$’s may be due
to beats in the $F_1$ region rather than $F_0$-guided segregation.
This last explanation was favored by Culling and Darwin (1993).

Culling and Darwin (1994) found further that identifica-
tion improves with $\Delta F_0$ (up to one semitone) for concurrent
vowels represented by alternating partials of two harmonic
series (successive partials of each vowel belonged to differ-
ent series). This result is also difficult to explain on the basis
of harmonic cancellation only. Darwin (1996) presented sub-
jects with pairs of concurrent sentences synthesized each
with a fixed $F_0$. Identification increased greatly with $\Delta F_0$,
but only beyond one semitone. At one semitone, identifica-
tion was no better than at $\Delta F_0=0\%$. This result casts some
doubt on the relevance of effects demonstrated in double-
vowel experiments to the segregation of real speech.

V. SUMMARY AND CONCLUSION

The first two articles in this series confirmed earlier data
indicating that the auditory system may segregate sounds by
exploiting the harmonic structure of interference to suppress
it. Alternative hypotheses such as enhancement of harmonic
targets, beats, or pitch-period-asynchrony are unlikely to ac-
count for segregation of weak target vowels ($-10$ to $-20$
$\text{dB}$) at relatively large $\Delta F_0$’s ($6\%$), although we cannot ex-
clude that they might play a role at higher target levels or
smaller $\Delta F_0$’s. Together, the results support a rather wide
class of models capable of harmonic cancellation. Among
those models, that of Meddis and Hewitt (1992) stands out
because of its plausibility and predictive power. However, it
could not explain one aspect of our data: the improvement
with $\Delta F_0$ of the identification of weak targets ($-20\text{ dB}$). For
some vowel pairs, the model predicted no effect of $\Delta F_0$ for
the weaker vowel, yet for the same pairs we observed a clear
effect. The problem appeared to lie in the model’s all-or-
nothing selection principle that assigns each peripheral chan-
nal as a whole to one or the other vowel.

The “neural cancellation filter” described in the present
paper processes auditory-nerve fiber discharge patterns
within channels to isolate correlates of concurrent vowels.
Certain spikes are removed from the spike train, so that its
statistics no longer reflect the competing vowel. The compet-
ing vowel must be harmonic for the filter to work, which is
consistent with the principle of harmonic cancellation sup-
ported by our data. The filter allows individual channels to
be searched for residual information reflecting a vowel too
weak to dominate any one channel. It can in principle be
exploited in a variety of ways, either to implement the chan-
ell selection stage of a revised version of Meddis and
Hewitt’s model, or else to produce segregated “temporal” or
“place” representations of each vowel. Of the two physi-
ological elements required by the filter, one (inhibitory gat-
ing) is plausible but the other (delay lines of up to 10 or 20
$\text{ms}$) has yet to be found in the auditory system.

The concurrent vowel identification model of Sec. III
added to the neural cancellation filter a set of assumptions
that allowed it to make quantitative predictions. The details
of those assumptions are secondary; their purpose was to
produce an incarnation of the model for which reasonable
predictions could be formulated. The focus of this paper is
the neural cancellation filter that underlies the model, and
beyond it, the concept of harmonic cancellation that has
emerged from various experimental results as an important principle underlying $F_0$-guided segregation.

ACKNOWLEDGMENTS

This paper is based on experimental data obtained at ATR Human Information Processing Research Laboratories under a research agreement between ATR and the Centre National de la Recherche Scientifique. John Culling of the MRC Institute of Hearing Research provided part of the software used to build the models. Thanks to him, Ray Meddis, Alan Palmer, and one anonymous reviewer for detailed comments on previous versions of the draft.

1The illustration is valid for a population of high-spontaneous rate fibers. If low-spontaneous rate fibers are considered, place features may be preserved at high driving levels.

2On average, 75% of the samples of the signal within a channel are set to 0; half-wave rectification sets all negative samples to zero, and of the remainder, all those such as $s(t)<s(t-T)$ are set to 0 in Eq. (1). Only 25% of the samples retain their original value.


Darwin, C. J. (1996). (Personal communication.)


