# Waveform interactions and the segregation of concurrent vowels

Alain de Cheveigné
*Laboratoire de Linguistique Formelle, CNRS/Université Paris 7, 2 place Jussieu, case 7003, 75251, Paris, France and ATR Human Information Processing Research Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan*

(Received 2 April 1999; revised 12 July 1999; accepted 16 July 1999)

Two experiments investigated the effects of small values of fundamental frequency difference ($\Delta F_0$) on the identification of concurrent vowels. As $\Delta F_0$'s get smaller, mechanisms that exploit them must necessarily fail, and the pattern of breakdown may tell which mechanisms are used by the auditory system. Small $\Delta F_0$'s also present a methodological difficulty. If the stimulus is shorter than the beat period, its spectrum depends on which part of the beat pattern is sampled. A different starting phase might produce a different experimental outcome, and the experiment may lack generality. The first experiment explored the effects of $\Delta F_0$'s as small as 0.4%. The smallest $\Delta F_0$ conditions were synthesized with several starting phases obtained by gating successive segments of the beat pattern. An improvement in identification was demonstrated for $\Delta F_0$'s as small as 0.4% for all segments. Differences between segments (or starting phase) were also observed, but when averaged over vowel pairs they were of small magnitude compared to $\Delta F_0$ effects. The nature of $\Delta F_0$-induced waveform interactions and the factors that affect them are discussed in detail in a tutorial section, and the hypothesis that the improvement in identification is the result of such interactions (beat hypothesis) is examined. It is unlikely that this hypothesis can account for the effects observed. The reduced benefit of $\Delta F_0$ for identification at smaller $\Delta F_0$'s more likely reflects the breakdown of the same $F_0$-guided segregation mechanism that operates at larger $\Delta F_0$'s. © *1999 Acoustical Society of America.* [S0001-4966(99)01811-1]

PACS numbers: 43.71.Es, 43.71.Pc, 43.66.Hg, 43.66.Ba [JH]

## INTRODUCTION

A number of cues are useful when one tries to hear speech in a noisy environment (Cherry, 1953; Brokx and Nooteboom, 1982; Darwin and Carlyon, 1995). When both target and competitor are harmonic (for example, both voiced), a difference in fundamental frequency ($F_0$) is beneficial. This effect has been studied by many authors using pairs of synthetic vowels (see de Cheveigné *et al.*, 1997a for a review). When a $\Delta F_0$ is introduced between vowels, identification generally improves up to about one semitone (6%), after which it remains constant before deteriorating again at the octave. The largest jump in identification rate usually occurs between $\Delta F_0 = 0\%$ and the smallest nonzero value used in the study (typically 6%, 3%, or 1.5%). However, the region below 1.5% where most of the improvement occurs has not been explored in detail.

Small $\Delta F_0$'s present a methodological difficulty. The shape of the compound stimulus fluctuates at a rate equal to $\Delta F_0$. If the stimulus is shorter than the beat period $1/\Delta F_0$, both its long-term spectrum, and the set of short-term spectra that can be sampled within it, depend on which part of the beat pattern it spans, which in turn depends on the starting phase spectra. A different starting phase might produce a different experimental outcome, and so the generality of the experiment may be in question. What appears like a $\Delta F_0$ effect might be the chance result of some particularly unfavorable phase spectrum at $\Delta F_0 = 0$, and/or a particularly favorable segment of the beat pattern at $\Delta F_0 \neq 0$.

A first aim of this study was to verify the generality of improvement of identification with $\Delta F_0$ by assessing the effects of starting phase. It is impossible to test all possible phase spectra, but useful conclusions can be drawn by sampling several phase conditions. Differences among them tell us about the approximate size of phase effects, and comparisons of $\Delta F_0$ effects across phase conditions tell us of their generality. In experiment 1 of this paper, at the smallest nonzero $\Delta F_0$, the stimulus set included 4 consecutive portions of a double-vowel waveform, each shorter than the beat pattern. In experiment 2 it included two particular phase conditions: same-phase and antiphase. A second aim of this study was to test the hypothesis that beats contribute to the segregation of concurrent sounds. Beat-induced fluctuations might be sampled by the auditory system to enhance identification of a vowel pair. This so-called ''beat hypothesis'' has been proposed to explain the effects of small $\Delta F_0$'s (Culling and Darwin, 1993, 1994; Assmann and Summerfield, 1994). The experiments allow the effects of such fluctuations to be measured, so one can decide whether or not they are capable of explaining the $\Delta F_0$ effects.

A major obstacle in dealing with waveform interactions on the basilar membrane is their complexity. The simplest beats (those between two partials) have dimensions of rate, phase, depth, and carrier frequency, which vary among channels of the peripheral filter. In response to the sum of two vowels, the shape of the waveform in each channel depends on channel characteristics (selectivity, phase distortion) as well as stimulus characteristics (amplitude and phase spectra of both vowels). It is also affected by demodulation or trans-

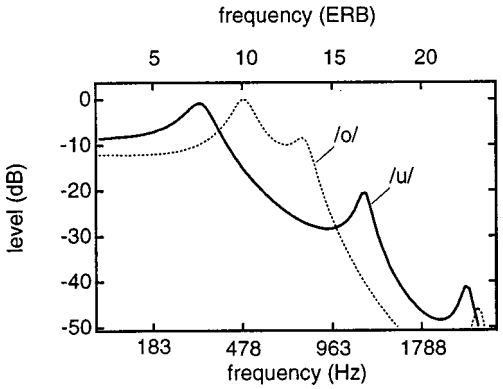


FIG. 1. Spectral envelopes of Japanese vowels /o/ and /u/. The abscissa represents frequency, warped to an ERB scale [uniform density in terms of equivalent rectangular bandwidth, Moore and Glasberg (1983)].

duction characteristics: nonlinearity, temporal integration, etc. To facilitate understanding of these phenomena, the next section offers a tutorial on $\Delta F_0$-induced waveform interactions on the basilar membrane. It is followed by a section on models of how the auditory system might exploit beats to enhance identification. Next comes the description of two experiments that probe the effects of small $\Delta F_0$'s while controling for phase-dependent effects of waveform interactions. Finally, in Sec. V we weigh evidence for and against several models of segregation.

## I. A TUTORIAL ON $\Delta F_0$-INDUCED WAVEFORM INTERACTIONS

Beats are a temporal phenomenon, but the conditions that they depend upon are best described in the frequency domain. Actually, two frequency axes must be considered: the frequency axis of the *spectrum* of the stimulus (or of the vibration waveform at some point in the cochlea), and the *tonotopic axis* of the basilar membrane. If peripheral filters were infinitely narrow, each would select a single partial and these two axes could be merged. Unfortunately, waveform interactions occur precisely because filters allow several partials to pass through. In the following graphs the nature of the axis can be determined by checking whether it is labeled *frequency* or *filter CF*.

Figure 1 shows the spectral envelopes of the Japanese vowels /o/ and /u/. The abscissa here is frequency. For uniformity with following figures, it is warped so that frequencies are uniformly distributed on a scale of equivalent rectangular bandwidth (ERB) (Moore and Glasberg, 1983). The spectral envelope is not the spectrum of a waveform, but rather a function that determines the amplitude of each partial according to its frequency. It is a complex function that defines both level and phase, but the figure shows only the level. The phase spectrum usually has little effect on the sound of the vowel. When the vowel is produced, the spectral envelope is sampled at multiples of $F_0$ to obtain the actual spectrum of the vowel: densely if $F_0$ is low, or sparsely if it is high.

Figure 2 shows the magnitude of the short-term spectrum of the vowel /u/ synthesized at $F_0 = 132$ Hz. The frequency axis is again warped to an ERB scale. The spectrum

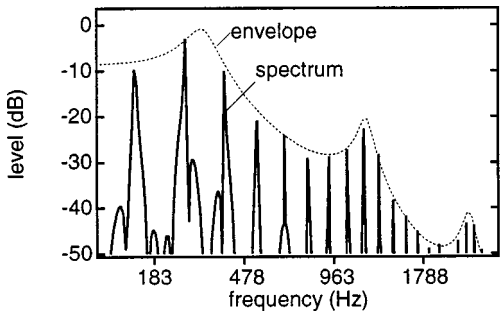


FIG. 2. Short-term spectrum of vowel /u/, calculated over a 250-ms stimulus. The abscissa represents frequency, warped to an ERB scale.

consists of a series of peaks. Their shape and width depend on the shape and width of the analysis window, itself constrained by the duration of the stimulus (in this case 250 ms including 20 ms raised cosine onset and offset). The vowel spectrum can be seen as the product of the line spectrum of a wideband periodic pulse train by the previous spectral envelope.

The spectrum of the waveform at any point of the basilar membrane also consists of peaks at harmonics of $F_0$. Each filter responds to several partials, but most of them are of low amplitude. Figure 3 shows their rms levels as a function of the filter's characteristic frequency (CF). Each line is for a different partial, the first few of which are labeled. The thin dotted line represents the total rms output in response to all partials together (excitation pattern). The figure uses the same warped scale as the preceding plots, but here it represents filter CF (or position along the basilar membrane) instead of the frequency axis of a spectrum. Basilar membrane filtering was simulated using the Auditory Toolbox software of Slaney (1993).

The number of partials in the filter output differs according to its CF. In the low-CF region, filters tuned near an individual partial respond mainly to that partial and exclude all others. Filters tuned between two partials respond to both, but weakly. In the high-CF region, all filters respond to several partials. The distance between the full line belonging to a given partial, and the dotted line, represents the proportion (in dB) of other partials in the total response. In the time domain, the waveform at the output of a low-CF filter tuned to a partial is approximately a sine-wave. That of other filters is a composite waveform that beats at the fundamental period. Such fast fluctuations at the pitch period are *not* what is meant by beats in the context of this paper.

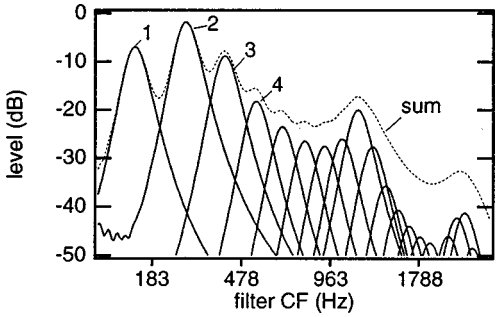


FIG. 3. Full lines: rms output of a cochlear filter bank as a function of characteristic frequency (CF), for each of the partials of vowel /u/. Dotted line: rms output in response to the entire vowel.

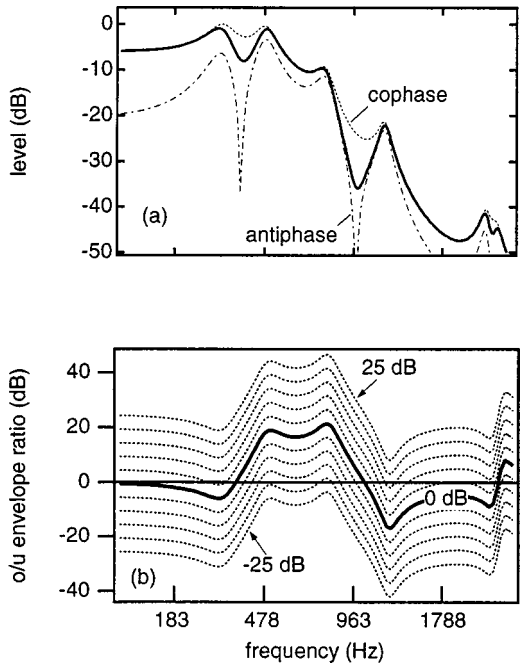


FIG. 4. (a) Vector sum of spectral envelopes of /o/ and /u/. Dotted line: both vowels have same phase spectra. Dot–dash: opposite phase spectra. Full line: both vowels are in Klatt phase [produced by the synthesizer of Klatt (1980) that approximates the phase spectrum of naturally produced vowels]. (b) Relative level in dB between magnitude spectral envelopes of /o/ and /u/ for values of the overall o/u rms relative level ranging from −25 to 25 dB in 5-dB steps. Note that for o/u ratio=20 and 25 dB the spectrum is entirely dominated by /o/, and at o/u ratio=−25 dB it is dominated by /u/.

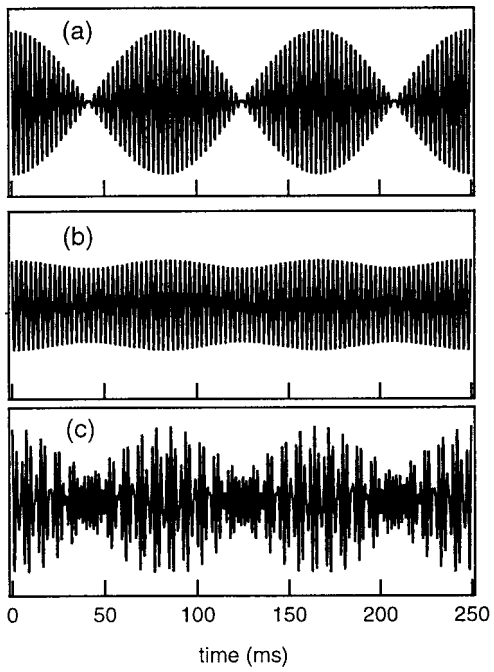


FIG. 5. (a) Beats between equal-amplitude partials of frequencies 390 and 402 Hz. The beat rate is 12 Hz. (b) Same, but the partials differ in level by 20 dB. (c) Same as (a) with the addition of a third partial of frequency 520 Hz.

When two vowels are added, the previous analysis can still be applied as long as they have the same $F_0$. The compound waveform consists of partials of that common $F_0$, with levels determined by a spectral envelope that can be calculated by *vector summation* of the complex envelopes of both vowels. The sum depends not only on the levels of both, but also on their relative *phases*. This is illustrated in Fig. 4(a), for vowels /o/ and /u/ with equal rms levels. The dotted line represents the sum supposing the vowels' phase spectra are identical (an unlikely occurrence), and the dot–dash line represents the difference if they are opposite (equally unlikely). For arbitrary phase spectra the envelope is somewhere between the two. The full line represents the sum if both vowels are in ''Klatt phase'' [the phase spectrum produced by the Klatt synthesizer (Klatt, 1980) that approximates the phase of natural vowels].

The experiments reported in this paper used pairs of vowels of unequal amplitude. Figure 4(b) shows the *relative level*, at each frequency, between envelopes of vowels /o/ and /u/, scaled with an overall rms relative level that was varied between −25 and +25 dB in 5-dB steps. The effect of phase on summation is largest where the vowel envelopes are of similar amplitudes, that is, where the plot crosses or approaches the 0-dB line. These plots are interesting also in that they show which parts of the spectrum are ''dominated'' by either vowel at a given overall relative level. At extreme levels the spectrum is entirely dominated by one vowel or the other (/o/ at +20 or +25 dB, /u/ at −25 dB). At intermediate levels it is ''partitioned'' between the two vowels (the plots cut the 0-dB line). We will see presently how this dominance pattern might be modified locally by cochlear filtering.

So far, the $F_0$'s of both vowels were the same. If they are different but close, the previous analysis can be extended by interpreting partials of equal rank as having the same frequency but a progressively increasing (or decreasing) phase shift. The level of the sum varies between the limits described above [Fig. 4(a)], at a rate equal to $n\Delta F_0$ (where $n$ is the rank of the partial). Partials of all ranks beat in this way, but with differences in *rate* (proportional to rank), *depth* (depending on the relative level between partials), and *phase* (depending on the difference between their starting phases).

Figure 5(a) gives an example of a beat between two partials of equal amplitude. Figure 5(b) illustrates the shallower beats that occur when their levels differ by 20 dB. The waveform of the stimulus (double vowel) is the superposition of various such beat patterns. In the context of the ''beat model,'' we are not directly interested in fluctuations of the acoustic waveform. Nor are we interested in the beats of individual partials, as they might occur if the partials were somehow isolated from the rest. Rather we are interested in the waveform fluctuations that actually occur on the *basilar membrane*, as a result of filtering the stimulus waveform. It is those fluctuations that would be exploited by a mechanism sensitive to beats.

The effects of filtering must thus be taken into account. First, filtering alters the relative amplitudes of partials of a pair, some channels favoring one partial and others the other. The depth of their beats may thus vary somewhat among channels. Second, the dispersive properties of the basilar membrane affect their relative phase. This effect is channel dependent, so beats may occur with different phases in different channels. Third and most importantly, filtering reduces

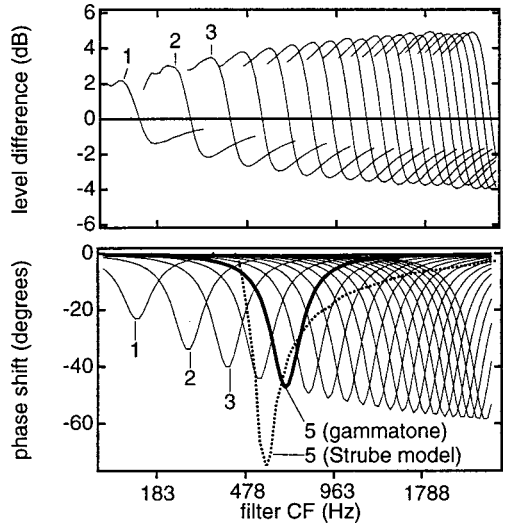


FIG. 6. Top: relative level between partials of same rank belonging to two equal-amplitude harmonic series of frequencies 128 and 132 Hz ($\Delta F_0 = 3\%$) filtered by a basilar membrane model, as a function of CF. Each curve is for a different rank, and is limited to CFs for which both partials are attenuated by less than 40 dB. Bottom: phase shift between partials of same rank belonging to two harmonic series of frequencies 128 and 132 Hz as a function of CF. Each curve is for a different rank.

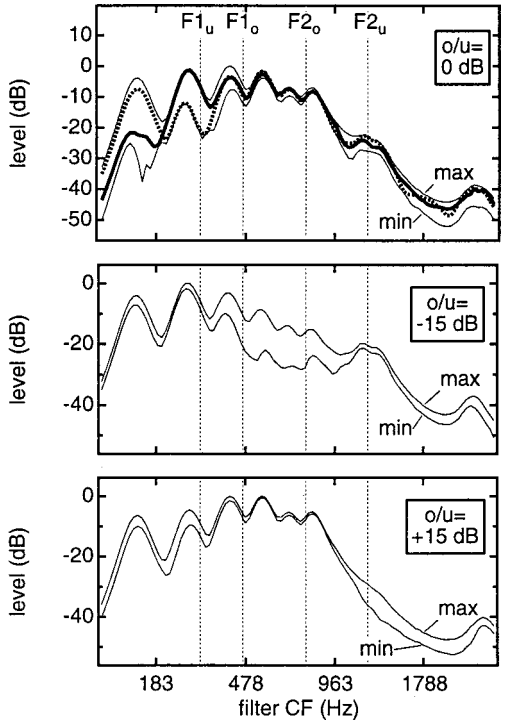


FIG. 7. Top, thin lines: limits of beat-induced variations of the rms output of a gammatone filter bank in response to the sum of vowels /o/ and /u/ for o/u relative level=0 dB, as a function of CF. Thick lines: profile of output at two instants (chosen for their dissimilarity). Middle and bottom: same, for o/u ratio= $-15$ and $+15$ dB. The simulation was performed with very slow beats, to avoid smoothing by temporal integration in the rms calculation stage. Vertical dotted lines indicate the positions of formants $F1$ and $F2$ of both vowels.

the number of partials that interact, with the result that the fluctuations of a filter output are simpler (and usually deeper) than those of the acoustic waveform. Some channels isolate individual partial pairs. The beat pattern at their output is similar to that shown in Fig. 5 (a or b). Other channels allow several partial pairs to interact together. The waveform at their output, which is more complex, can be understood as the superposition of two or more beat patterns.

The first two factors may be quantified. Figure 6 (top) shows the relative level, as a function of position along the basilar membrane, between partials of same rank from two harmonic series. The series had equal amplitudes, and $F_0$'s $=128$ and 132 Hz ($\Delta F_0 = 3\%$). Each line is for a different rank, and each is limited to CFs for which both partials are attenuated by less than 40 dB by the filter. The shift (in dB) is positive for channels with CFs below the partial's frequency, and negative above. The result of this shift is to modify, within each channel, the pattern of dominance of Fig. 4(b). Within each channel, the ratio of partials of same rank differs from that specified in Fig. 4(b) by the amount specified in Fig. 6 (top). This plot is for $\Delta F_0 = 3\%$, for other values the magnitude of the shift would be in proportion. The simulation used a software implementation of the gammatone filterbank (Slaney, 1993; Patterson *et al.*, 1992).

A similar analysis can be made for phase. Figure 6(b) shows the *phase shift* introduced between two partials of same rank ($\Delta F_0 = 3\%$) as the result of the dispersive properties of the basilar membrane. The shift is proportional to the *slope* of the phase characteristic, which for the gammatone is steepest for the channel tuned to the frequency of the partial. A word of warning: this simulation depends critically on the choice of the *gammatone filter* to model peripheral selectivity. If cochlear filters have different dispersive properties, the magnitude and pattern of phase shifts must be different. For example, the dotted line shows similar data obtained with a software implementation of the model of Strube (1985), which according to Kohlrausch (1995) better matches the phase characteristics of the basilar membrane than the gammatone. The phase shift is greatest within channels tuned slightly below the partials.

The third factor (isolation of individual partial pairs) is crucial for the existence of deep beats in the waveform at the output of a filter. The reason is easy to grasp. Beat patterns of partial pairs have rates and phases that vary according to their rank. The minimum of one pattern is unlikely to coincide with a minimum of the others, with the result that the peak-to-valley ratio is reduced when the patterns are superposed. Figure 5(c) shows the same two partials as Fig. 5(a), together with a neighboring partial. The depth of the valley is reduced. Based on this reasoning, we expect that deep beats are most likely to occur in channels that are dominated by a single beating pair. Going back to Fig. 3, we see that this may be the case in low-CF channels tuned to a partial. Low-CF channels intermediate between two partials respond to the superposition of two beat patterns, and high-CF channels respond to even greater numbers of beating partials, and this is likely to produce limited peak-to-valley ratios.

Figure 7 (thin lines) shows the minimum and maximum levels in each channel in response to the mixture of /o/ and /u/ at relative levels of 0, $-15$, and $+15$ dB as a function of filter CF. Also shown (at 0 dB, top) are two samples of the instantaneous excitation pattern chosen for their dissimilarity (thick lines). At 0 dB, beats are deepest within channels with

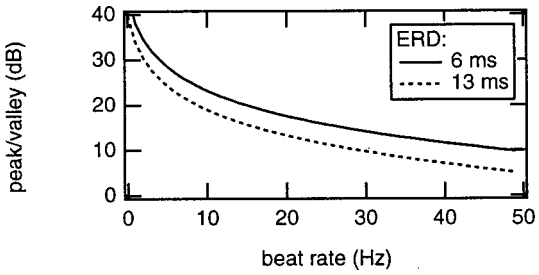


FIG. 8. The effect of temporal smoothing on beat amplitude. Peak/valley ratio of a maximally deep beat [e.g., Fig. 5(a)] after smoothing by a window of equivalent rectangular duration (ERD) 6 or 13 ms, as a function of beat rate. The peak/valley ratio of faster beats is reduced.

CFs below the first formants ($F1$) of /u/ and /o/. At an o/u ratio of $-15$ dB, beats are deepest in the vicinity of $F1$ and $F2$ of /o/ (middle plot). At $+15$ dB they are relatively shallow over all channels (bottom).

An important issue is the rate of beats. The overall beat pattern repeats itself at a rate equal to $\Delta F_0$. In the experiment of Sec. III, at the smallest $\Delta F_0$ (0.4%) the beat period was 2 s, or eight times the stimulus duration. At the largest $\Delta F_0$ (6%) it was 125 ms, or one-half the stimulus duration. Beats between partial pairs of rank $n$ (supposing they can be isolated) occur at a faster rate: $n\Delta F_0$. Restricted parts of the pattern of channel outputs may thus appear to pulsate faster than the overall pattern. In the vowel set used in the experiments, partials closest to $F1$ had ranks ranging from 2 to 6, and those closest to $F2$ ranks from 6 to 17. The fastest beats in the $F1-F2$ range thus occurred at 8.5 Hz for $\Delta F_0 = 0.4\%$, and 136 Hz for $\Delta F_0 = 8\%$ (at this spacing it makes little sense to distinguish beats between partials of same rank from those between partials of different rank).

A final consideration is that beats might be smoothed by temporal integration in the auditory system. Figure 8 shows the peak/valley ratio of a maximally deep beat pattern [such as that of Fig. 5(a)] after smoothing by a temporal window with an equivalent rectangular duration of 6 or 13 ms. These values are estimates obtained by Plack and Moore (1990). The former (6 ms) was obtained at frequencies of 900, 2700, and 8100 Hz, the latter (13 ms) was obtained at 300 Hz. Beat amplitude is reduced progressively with increasing beat rate. Note, however, that the detection of AM may not be limited by such low-pass filtering (Dau *et al.*, 1997). This question is discussed again later on.

## II. BEATS AND SEGREGATION

It has been suggested that beats that arise in response to concurrent vowels with different $F_0$'s might promote their identification, particularly at $\Delta F_0$'s too small to support other $F_0$-guided mechanisms (Culling and Darwin, 1993, 1994; Assmann and Summerfield, 1990, 1994). This section examines the various forms that can be given to this hypothesis, and discusses them critically in the light of the previous section.

### A. The ''glimpsing'' model of Culling and Darwin

Culling and Darwin (1994) suggested that beats might cause the excitation pattern to momentarily assume a shape favorable for identification. Selected samples or ''glimpses'' of this pattern would benefit identification. Their model involved a perceptron pattern-recognizer, of which there were two variants. The first (''one-at-a-time'' strategy) used two sampling points, one for each vowel, chosen to give the highest and second-highest activation scores of the perceptron. The second (''both-at-once'' strategy) sampled the excitation pattern at a single point in time, chosen to give the highest value of a pairwise compound measure derived from the perceptron outputs. The process that produces the excitation pattern (filtering, transduction, and smoothing) involves integration over time, so each sample is actually derived from a windowed portion of the stimulus.

The both-at-once strategy was the more successful. The single sample was more often classified as the correct vowel pair than the constant pattern evoked at $\Delta F_0 = 0$. Another way to put it is that beat-induced fluctuations produced a cloud of points in feature space (instead of the single point at $\Delta F_0 = 0$), with an extent that was fortunately greatest in a direction that led to correct classification. The model exploited this favorable aspect of beat-induced variability. However, it is not clear that it would be as successful in the presence of intraclass variability of natural speech sounds, or variability induced by noise. Temporal sampling is antithetical to smoothing schemes used to deal with noise.

### B. Serial differences between excitation patterns

An alternative hypothesis is that identification might benefit from dynamic cues, such as the *difference* between successive excitation patterns ($\Delta$EP). It is well known that dynamic cues are important for vowel identification (e.g., Strange *et al.*, 1998). Kuwabara *et al.* (1983) found that a vowellike stimulus (''*X*'') with a spectrum intermediate between two vowels (''*A*'' and ''*B*'') was identified as vowel ''*A*'' when it appeared as the central portion of dynamic vowellike spectrum of shape ''*BXB*,'' and as vowel ''*B*'' when it appeared in an ''*AXA*'' pattern. Summerfield *et al.* (1984) found that subjects could perceive a vowel from a flat-spectrum complex if it were preceded by the ''complement'' of a vowel spectrum in which formant peaks were replaced by valleys. Summerfield *et al.* (1987) found further that a uniform-spectrum precursor enhanced the identification of vowellike stimuli with shallow envelopes. In both cases the auditory system seemed to exploit the difference (latter minus former) between two spectral shapes. The result was extended by Summerfield and Assmann (1987) to the case where the precursor was shaped like a first vowel, to which was later added a second vowel. Steps in spectral shape as small as 2 dB were effective. One might imagine that $\Delta$EPs produced by beats are exploited in a similar fashion.

There are several difficulties with this proposition. First, Summerfield *et al.* (1984) found that transitions *toward* the target vowel's spectrum alone were effective. With beats, the auditory system would need to select transitions in the right direction and ignore the (possibly confusing) opposite transitions. Second, they also found that a transition had to be preceded by a precursor of at least 125 ms. Beat-induced valleys of modulation are often shorter than that, particularly

the sharp dips that occur within deep beats [Fig. 5(a)]. Finally, beats occur with different rates (and phases) in different channels, implying a rather disorderly succession of ΔEPs.

### C. Modulation profile

A third proposition is that the profile of beat-induced pulsations across the basilar membrane supports identification, according to a mechanism similar to that which produces the sensation of roughness. Contrary to the previous proposals, the relative phases of beats between channels, or the sign of transitions, are indifferent. The auditory system must, however, be able to detect the pulsations and keep track of their distribution across the channels.

The detection threshold of sinusoidal modulation of an isolated high-frequency pure-tone carrier corresponds to a modulation ratio (ratio of peak excursion to average) $m = -30$ dB or a peak/valley ratio of 0.55 dB (Dau *et al.*, 1997). However, detection sensitivity for harmonics in vowels might be reduced by at least two factors. A first factor is the reduction of sensitivity to modulations at high rate. This is classically described as following a low-pass characteristic, but Dau *et al.* (1997) argue that modulation detection is best understood as involving a bank of *bandpass* filters tuned to modulation rates extending from 0 to at least 200 Hz. These filters are wide ($Q=2$), and therefore presumably liable to leakage from modulation at $F_0$, implying a low-pass characteristic similar to low-pass filtering in the modulation domain. In any case, sensitivity is likely to be reduced for the faster modulations that occur between partials of high rank or at large $\Delta F_0$'s. A second factor is modulation detection interference (MDI) (Yost and Sheft, 1989), by which modulation in one part of the spectrum interferes with the detection of modulation in other parts. Detection of a cue to identification might be hindered by beats that occur in other parts of the mixed-vowel spectrum.

Supposing that modulation is detected, it must be ''localized,'' that is, assigned to the right part of the spectrum. Marin and McAdams (1996) measured modulation thresholds for the detection and correct assignment of the center frequency (375, 750, or 1250 Hz) of a ''formant'' defined by the amplitude modulation of two or three consecutive partials. The modulation waveform was complex and comprised 13 harmonics of 5 Hz with amplitudes following a $1/f$ law. The threshold rms modulation index $m_{\mathrm{rms}}$ was in the range $-19$ to $-13$ dB (for sinusoidal modulation this would have corresponded to a modulation ratio $m$ in the range $-16$ to $-10$ dB, or a peak/valley ratio of 2.8 to 5.7 dB). Overall, Marin and McAdams estimated that thresholds for correct ''localization'' of modulation were about three times higher than for its mere detection. Figure 9 shows the peak/valley ratio of beats as a function of CF for a mixture of vowels /o/ and /u/, for different o/u relative levels. The largest beat ratios are observed in low-frequency channels at relative levels near 0 dB. They correspond to beats with envelopes shaped as rectified cosines [as in Fig. 5(a)].[1] At other relative levels modulations are strongest in other channels. For example, at $-15$ dB there are beats near $F2$ of /o/, and at 0 dB near $F1$ and $F2$ of /u/. Assuming they can be detected and properly localized, $\Delta F_0$-induced modulations might assist identification.

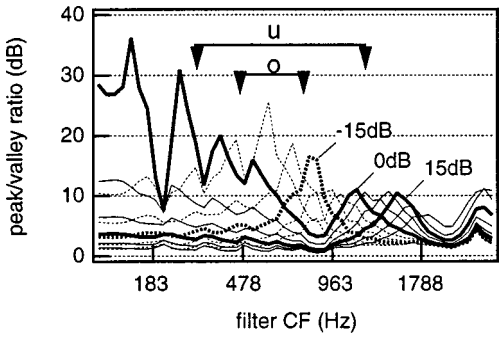


FIG. 9. Modulation profile: peak/valley ratio as a function of filter CF, for o/u relative levels ranging from $-25$ to $+25$ dB in 5-dB steps. Arrows indicate formant frequencies $F1$ and $F2$ of both vowels.

There are several difficulties with this proposition. First, the position of maximum beats is not the same at different relative levels, and does not always correspond to a formant of either vowel. This might confuse a vowel identification mechanism. Second, strong $\Delta F_0$ effects are found in conditions for which beat amplitudes are small in all channels. Third, beats depend on the difference between partial frequencies, and should thus affect both vowels symmetrically. Yet a previous study found that if one vowel was harmonic and the other inharmonic, the inharmonic vowel was identified better (de Cheveigné *et al.*, 1995, 1997b). The same argument can be used against the other two schemes (glimpsing and ΔEP). A final argument against the modulation profile hypothesis comes from an experiment of Moore and Alcántara (1995). Subjects could identify ''vowels'' that were defined by amplitude modulation, in the region of their formants, of an otherwise flat-spectrum complex, but only if components started in cosine phase. For random starting phase they could not identify the vowels.

### D. The generality of $\Delta F_0$ experiments

The choice of starting phase spectra in concurrent vowel experiments is largely arbitrary. It is common to use the phase spectra produced by the synthesizer of Klatt (1980) that approximates natural vowels, and to give both vowels the particular starting phase produced by default by the software. This choice is not necessarily typical of natural situations, as different sources need not start out in synchrony, different path lengths from sources to ear(s) add to the phase shift of one source with respect to the other, and room acoustics may further scramble the phase spectra of both voices.

Manipulation of the phase spectrum is known to have little effect on the quality of isolated vowels. For concurrent vowels it could affect identification in three ways. Starting phase determines (a) the *set* of excitation patterns that may occur during a beat period, (b) the *order* in which they appear, and (c) the *subset* of these patterns that are available within a stimulus shorter than the beat period. A previous study (de Cheveigné *et al.*, 1997b) found that (a) and (b) had negligible effects: identification of concurrent vowels with durations twice the beat period was the same for all the

TABLE I. Formant frequencies and bandwidths of vowels.

| | /a/ | /i/ | /u/ | /e/ | /o/ | BW |
|---|---|---|---|---|---|---|
| *F*1 | 750 | 281 | 312 | 469 | 468 | 90 |
| *F*2 | 1187 | 2281 | 1219 | 2031 | 781 | 110 |
| *F*3 | 2595 | 3187 | 2469 | 2687 | 2656 | 170 |
| *F*4 | 3781 | 3781 | 3406 | 3375 | 3281 | 250 |
| *F*5 | 4200 | 4200 | 4200 | 4200 | 4200 | 300 |

starting-phase spectra investigated (both sine, both ''random'' with the same pattern, both random with different patterns, one sine and the other random). On the other hand (c) is likely to have a strong effect at small $\Delta F_0$'s for stimuli shorter than the beat period. As a related concern, starting phase determines the spectrum of the stimulus in the $\Delta F_0 = 0$ condition against which $\Delta F_0 \neq 0$ conditions are compared. Improvements observed with nonzero $\Delta F_0$ might be specific to a particularly unfavorable starting phase spectrum at $\Delta F_0 = 0$.

The experiments reported in the next two sections explore the parameter region of small $\Delta F_0$'s, using various starting phase conditions to test the generality of the effects observed. They also challenge segregation models based on waveform interactions, insofar as those models lead us to expect phase and $\Delta F_0$ effects to be of similar magnitude and to covary in an orderly fashion.

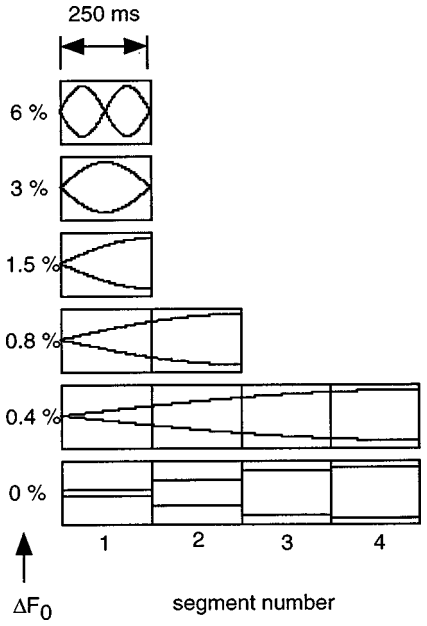


FIG. 10. Schematized beat patterns for each $\Delta F_0$ and segment. Stimuli contain two beat periods at $\Delta F_0 = 6\%$ (top), one at 3%, and one half at 1.5%. The 0.8% condition is realized with two consecutive segments, and the 0.4% condition with four segments. The $\Delta F_0 = 0\%$ condition is realized in four versions, with phases equal to the ongoing phase at the centers of the 0.4% segments. The shape of real beat patterns is, of course, more complex than schematized here.

## III. EXPERIMENT 1: SMALL $\Delta F_0$'s

### A. Methods

Methods were similar to those of de Cheveigné *et al.* (1997a, b). Stimuli were constructed from synthetic tokens of Japanese vowels /a/, /e/, /i/, /o/, /u/ (formant frequencies and bandwidths are listed in Table I). Stimuli were 270 ms in duration, with 20-ms raised-cosine ramps at onset and offset (250-ms ''effective duration'' between −6-dB points). Vowels were synthesized at a 20-kHz sampling rate using a frequency-domain additive synthesizer (Culling, 1996) that emulates Klatt's cascade synthesizer (Klatt, 1980), and scaled to a standard rms value. To obtain double vowels, single-vowel tokens were paired, one vowel was scaled by an amplitude factor, both were summed, and their sum was scaled to a standard rms value. The result was output diotically to earphones from the NeXT computer. The sound pressure levels ranged from 63 to 70 dB (A) depending on the vowel pair, as measured by a Bruel and Kjaer artificial ear.

Fundamental frequencies ($F_0$) were chosen by pairs centered on 132 Hz, to obtain $\Delta F_0$'s of approximately 6% (128, 136 Hz), 3% (130, 134 Hz), 1.5% (131, 133 Hz), 0.75% (131.5, 132.5 Hz), and 0.375% (131.75, 132.25 Hz). For convenience, the latter $\Delta F_0$ values are rounded to 0.8% and 0.4%, respectively, in the rest of the text. Both vowels were given a ''random'' phase spectrum that was the same for both vowels and all conditions. Partials of same rank thus had the same starting phase, allowing beat patterns to be more easily predicted. ''Random'' phase was preferred over alternatives such as sine or cosine, because it produces less peaky waveforms. ''Klatt'' phase (produced by default by the Klatt synthesizer) was not used because it differs between vowels.

The $\Delta F_0$ results in a progressive phase shift between partials of same rank, with magnitude proportional to rank, $\Delta F_0$, and time. Collectively, the beats produce a pattern that repeats with a period of $1/\Delta F_0$, as represented schematically in Fig. 10. To control for phase effects in stimuli shorter than the beat period, stimuli for the smallest $\Delta F_0$ conditions were synthesized by windowing successive parts of a longer waveform. For the $\Delta F_0 = 0.4\%$ condition, a 1020-ms stimulus was synthesized from which four segments were windowed, each of 250-ms ''effective duration.'' The stimulus set also contained four stimuli at $\Delta F_0 = 0\%$, each with a starting phase spectrum equal to the phase spectrum of a $\Delta F_0 = 0.4\%$ segment, sampled at its center. These were obtained by synthesizing single vowel waveforms at $F_0 = 132$ Hz and time shifting them by $(0.5 \pm n)T_0/16$, $n = 0,1,2,3$, where $T_0$ is the fundamental period, before adding. Corresponding double-vowel segments at 0% and 0.4% therefore had similar long-term spectra, and differed from each other by $\Delta F_0$ only. Segments at 0% and 0.4% were numbered 1, 2, 3, 4. In a similar fashion, two segments were prepared at $\Delta F_0 = 0.8\%$ by windowing consecutive 250-ms portions of a beat pattern.

To summarize, there was *one* segment each for $\Delta F_0 = 6\%$, 3%, and 1.5%, *two* segments for $\Delta F_0 = 0.8\%$, and *four* segments each for $\Delta F_0 = 0.4\%$ and 0%, a total of 13 $\Delta F_0$-segment conditions. The stimuli cover two beat periods at 6%, one at 3%, and one-half of a beat period at 1.5%, 0.8%, and 0.4%. When designing the experiment, it was incorrectly assumed that beat patterns were symmetric in time,

and that covering half a period was equivalent to covering it all. Our sample is thus less complete than intended, but nevertheless sufficient to reveal any strong phase effects. When paired, vowels had either the same level (0 dB), or one vowel was weaker or stronger than the other by 15 dB. Strong $\Delta F_0$ effects are expected for weak targets ($-15$ dB), but beats might be stronger at 0 dB and this condition was included to allow comparisons with other studies. Vowels within a pair were always different. There were a total of 780 double-vowel stimuli (vowel pair=/ae/,/ai/,/ao/,/au/,/ei/,/eo/,/eu/,/io/,/iu/,/ou/)$\times$(relative level$=-15,0,15$ dB)$\times(2F_0$orders)$\times(13\Delta F_0$-segment conditions). Ideally, the stimulus set should also have contained single vowels to make it consistent with the description made to the subjects (see below). However, the set was already very large, and so single-vowel conditions were not included.

Subjects were 15 Japanese students (seven male and eight female, aged 18 to 22 years) recruited for a series of ten experiments on concurrent vowel identification and paid for their services. Experiments 1 and 2 described in this paper were respectively the fourth and eighth of that series. Each stimulus was presented once. The subjects were told that it could be either a single vowel or two simultaneous different vowels. They were instructed to choose either one or two vowels as a response according to what they heard. If the response was inappropriate (more than two vowels, two identical vowels, a nonvowel, etc.), a message reminded them of the options and requested a new answer. They could pause at will, in which case the last stimulus presented before the pause was repeated after the pause (subjects paused on average five times per session). There was no feedback.

The response for each double-vowel stimulus was scored twice: each vowel in turn was nominated the ''target,'' the other being a ''competitor.'' A target was deemed identified if its name was among the one or two vowels reported by the subject. The *proportion of targets correctly identified* (constituent-correct or target-correct identification rate) was calculated for each target/competitor condition. The *average number of vowels reported* per stimulus was also recorded.

### B. Results

Two overlapping subsets of the conditions are considered separately. The first subset consists of ($\Delta F_0=0\%,0.4\%$)$\times$(segment$=1,2,3,4$). The second subset consists of conditions ($\Delta F_0=0\%,0.4\%,0.8\%,1.5\%,3\%,6\%$), the first two of which are the ($\Delta F_0=0\%,0.4\%$) conditions of the first subset averaged over phase conditions. Segment conditions at $\Delta F_0=0.8\%$ showed no interesting differences and are not discussed in detail. Results are considered at target-to-competitor ratios of $-15$ and 0 dB. Identification rates at $+15$ dB were essentially perfect and are not discussed.

#### *1. Segment effects at 0% and 0.4%*

Target-correct identification rates for the subset ($\Delta F_0=0\%,0.4\%$)$\times$(segment$=1,2,3,4$) were submitted to a repeated-measures ANOVA with factors $\Delta F_0$ and segment. This analysis was performed for target/competitor ratio $=-15$ and 0 dB.

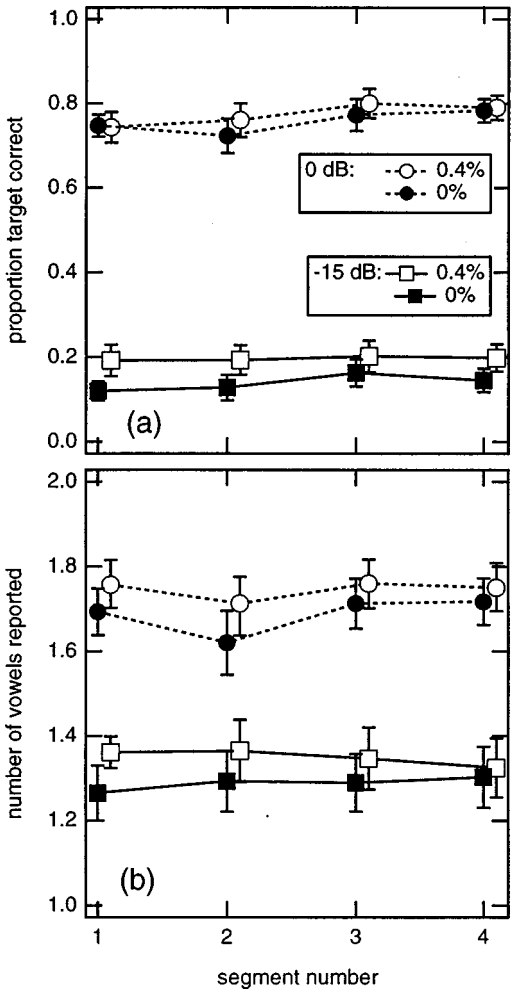


FIG. 11. Experiment 1. Target-correct identification rate (a) and number of vowels reported per stimulus (b) as a function of segment number, for $\Delta F_0=0\%$ (filled symbols) and 0.4% (open symbols), at 0 dB (circles) and $-15$ dB (squares). Error bars represent $\pm$ one standard error of the mean.

At $-15$ dB, the main effect of $\Delta F_0$ was significant [$F(1,14)=17.42$, $p=0.0009$], as was that of segment [$F(3,42)=3.59$, $p=0.044$, GG$=0.63$] [Probabilities reflect, where necessary, a correction factor applied to the degrees of freedom to compensate for the correlation of repeated measures (Geisser and Greenhouse, 1958)]. Their interaction was not significant. Identification rates are plotted in Fig. 11(a) (square symbols). Identification was better at $\Delta F_0=0.4\%$ than at 0%. The average number of vowels reported per stimulus was also submitted to an ANOVA with factors $\Delta F_0$ and segment. The main effect of $\Delta F_0$ was again significant [$F(1,14)=14.65$, $p=0.0018$], that of segment was not, and their interaction was significant [$F(3,42)=4.78$, $p=0.0079$, GG$=0.90$]. The number of vowels reported per stimulus is plotted in Fig. 11(b) (square symbols).

At 0 dB, neither the main effect of $\Delta F_0$ nor that of segment, nor their interaction were significant. The lack of segment effect or interaction was somewhat unexpected, as the similarity in amplitudes of partials of both vowels was expected to result in strong waveform interactions. In the

study of Assmann and Summerfield (1994), in which segment effects were found, vowels in a pair were excited by the same source (equal ''vocal effort'') and their rms levels were almost the same. The number of vowels reported was significantly affected by both $\Delta F_0$ [$F(1,14)=9.43$, $p=0.008$] and the segment factor [$F(3,42)=3.78$, $p=0.04$, GG$=0.62$]. Their interaction was not significant. The identification rate and number of vowels reported are plotted in Fig. 11(a) and (b) (round symbols). An explanation for the rather small segment effects at 0 dB may be found in the pairwise response data described further on.

### 2. $\Delta F_0$ effects

At target/competitor relative level$=-15$ dB, differences between segments were significant but not large, so one can reasonably average scores over segments for the lowest $\Delta F_0$'s. Target-correct identification rates for conditions ($\Delta F_0=0\%,0.4\%,0.8\%,1.5\%,3\%,6\%$) were submitted to a repeated-measures ANOVA with the single factor $\Delta F_0$. The effect of $\Delta F_0$ was significant [$F(5,14)=79.97$, $p<0.0001$, GG$=0.36$]. Identification rate is plotted with squares in Fig. 12(a). The average number of vowels reported per stimulus is plotted with squares in Fig. 12(b). There is a gradual increase of both measures with increasing $\Delta F_0$. The question of the significance of the step between $\Delta F_0=0\%$ and 0.4% is discussed in Sec. III B. At target/competitor ratio$=0$ dB, effects were smaller than at $-15$ dB, as observed in previous experiments (de Cheveigné *et al.*, 1997a, b). Identification rates and number of vowels reported are plotted as circles in Fig. 12(a) and (b).

### 3. Segment effects for individual vowel pairs

Waveform interactions are expected to favor the identification of certain segments over others, but there is no reason why the pattern across segments should be the same for all vowel pairs. Data were reanalyzed with vowel pair as a factor (20 levels). Four separate analyses were performed, one on each of the subsets: $(0,-15\,\text{dB})\times(0\%,0.4\%)$. For each, an ANOVA was performed with factors (segment$=1,2,3,4)\times$(vowel pair).

At 0 dB and $\Delta F_0=0\%$, the main factors of segment and vowel pair were significant [$F(3,42)=5.57$, $p=0.0084$, GG $=0.69$, and $F(19,266)=9.21$, $p<0.0001$, GG$=0.27$ respectively]. The segment$\times$pair interaction was also significant [$F(57,798)=2.71$, $p=0.005$, GG$=0.17$]. Identification rates are plotted in Fig. 13 as a function of segment number for all pairs, six of which are labeled. For u/a and u/e, identification dropped between the first and second segments, and increased thereafter. For o/u and e/i, it instead increased and then leveled off or dropped, while for o/a the greatest change was between the second and third segment. Patterns are indeed different for different vowel pairs.

At 0 dB and $\Delta F_0=0.4\%$, the pair effect was significant [$F(19,266)=6.46$, $p<0.0001$, GG$=0.27$]. The segment effect was also significant [$F(3,42)=3.2$, $p=0.044$, GG $=0.825$] but small and the interaction with pair was not significant. Thus, the large pair-specific segment effects observed at $\Delta F_0=0\%$ were not found at $\Delta F_0=0.4\%$. At $-15$

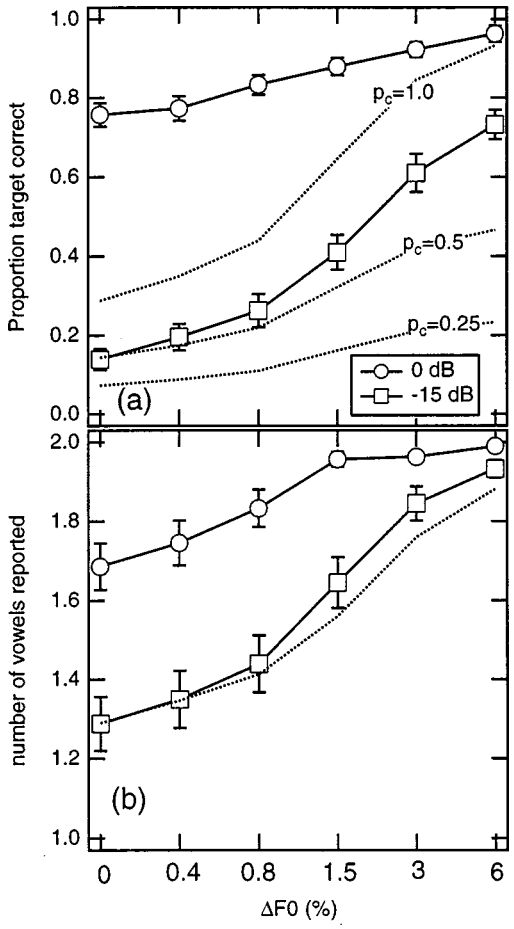


FIG. 12. Experiment 1. Target-correct identification rate (a) and number of vowels reported per stimulus (b) as a function of $\Delta F_0$, at 0 dB (circles) and $-15$ dB (squares). Data points at the lowest three $\Delta F_0$'s are averaged over segment conditions. The dotted lines are predictions of one measure based on the other. The dotted lines in (a) show the identification rate expected supposing that the probability of the second vowel being correct is constant and equal to 1.0 (top), 0.5 (middle), or 0.25 (chance, bottom), and that responses are determined entirely by the subjects' tendency to report two vowels. The dotted line in (b) supposes instead that the number of vowels reported is determined by the identifiability of the second vowel as measured by the identification rate (see text).

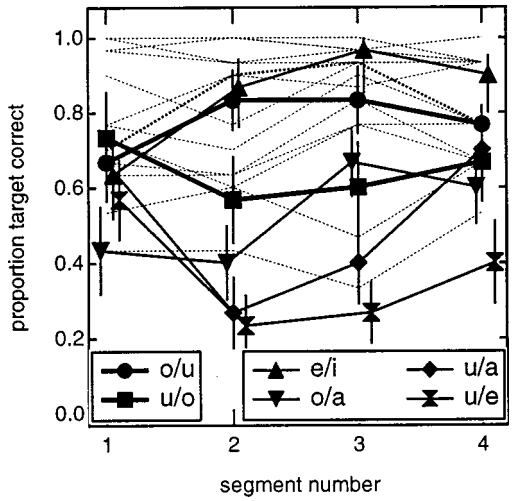


FIG. 13. Experiment 1. Identification rate as a function of segment number for individual vowel pairs, for $\Delta F_0=0\%$ and target/competitor ratio$=0$ dB.

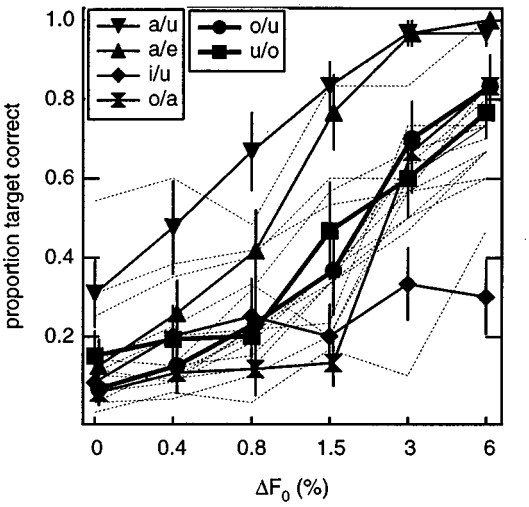


FIG. 14. Experiment 1. Identification rate as a function of $\Delta F_0$ for individual vowel pairs, at target/competitor ratio $=-15$ dB (averaged where appropriate over segment).

dB, at both $\Delta F_0$'s, the main factors of segment and vowel pair were both significant but their interaction was not. The pair-specific segment effects observed at 0 dB and $\Delta F_0 = 0\%$ did not generalize to the case where the target vowel was weaker than its competitor. To summarize, strong pair-specific segment effects were found, but only at 0 dB and only for $\Delta F_0 = 0$.

#### 4. $\Delta F_0$ effects for individual vowel pairs

It is likewise interesting to know whether $\Delta F_0$ effects were different between vowel pairs. The data for $-15$-dB targets were submitted to a repeated-measures ANOVA with factors $(\Delta F_0 = 0\%, 0.4\%, 0.8\%, 1.5\%, 3\%, 6\%) \times$ (vowel pair). The main effect of $\Delta F_0$ was highly significant as found before, as was the main effect of pair. Their interaction was also significant [$F(95,1330) = 2.84$, $p = 0.0018$, GG $= 0.12$], indicating differences between pairs in the pattern of dependency of identification on $\Delta F_0$. Data are plotted in Fig. 14 for all 20 pairs, 6 of which are labeled. Pairs o/u and u/o are typical of most. The $\Delta F_0$ effect was largest for pair a/e and smallest for pair i/u. For o/a, identification hardly improved with $\Delta F_0$ until 1.5%, after which it jumped sharply. The increment between 0% and 0.4% was positive for 18 out of 20 pairs. In summary, the average rates plotted in Fig. 12 are representative of most of the pair-specific trends.

## IV. EXPERIMENT 2: VERY SMALL $\Delta F_0$'s

Experiment 1 found significant effects at the lowest $\Delta F_0$ (0.4%), suggesting that measurable effects might be found at still lower values. Experiment 2 introduced $\Delta F_0$'s of 0.2% and 0.1%. Experiment 1 found segment effects that were either small or inconsistent across vowel pairs, but the segment conditions represented only a small sample of possible starting phases. Experiment 2 introduced two phase conditions likely to produce more radical waveform interaction effects: same and opposite phase. In Experiment 1, all $F_0$'s were clustered around 132 Hz, and it is conceivable that many repetitions might have allowed the auditory system to ''tune in'' to this frequency, and perform segregation with an unnatural degree of accuracy. In experiment 2, $F_0$'s were roved between three regions (124, 128, and 132 Hz) to discourage any such hypothetical fine tuning. The 0-dB inter-vowel relative level of experiment 1 was replaced by 5 dB, in the hope that identification rates at $+5$ dB would not be so high as to be at a ceiling, and thus possibly informative.

### A. Methods

Methods were as in experiment 1. Single vowels were synthesized at $F_0$'s of 124, 128, and 132 Hz, and at $F_0$'s higher by 0.125, 0.25, 0.5, and 1 Hz. Single vowels were paired and added with a relative level of 5 or 15 dB, to form double vowels with $\Delta F_0$'s of approximately 0%, 0.1%, 0.2%, 0.4%, and 0.8% (the precise percentage depends on the baseline $F_0$). The 0.4% and 0.8% conditions at 15 dB were identical to conditions of experiment 1, apart from their starting phases and $F_0$'s. Vowels were added in phase, denoted ''0,'' or else the polarity of one vowel was reversed before summation, denoted ''$\pi$.'' There were a total of 800 double vowel stimuli (vowel pair=/ae/,/ai/,/ao/,/au/,/ei/,/eo/, /eu/,/io/,/iu/,/ou/)$\times$(relative level$=-15,-5,5,15$ dB)$\times$(phase $=$''0,''''$\pi$'')$\times(5\Delta F_0$'s)$\times(2F0$orders). Absolute $F_0$ was assigned randomly from trial to trial. Again, the stimulus set lacked single vowels.

Note that the in-phase condition of experiment 2 is not quite the same as the segment 1 condition of experiment 1. The first segment at $\Delta F_0 = 0\%$ in experiment 1 was the sum of two vowels out of phase by $0.5 \pm T_0/16$, rather than perfectly in phase as in experiment 2. At $\Delta F_0 = 0.4\%$ and 0.8% the in-phase conditions of experiment 2 are the same as the segment 1 conditions of experiment 1.

### B. Results

At target/competitor ratio $=-15$ dB, identification scores were submitted to a repeated-measures ANOVA with factors $\Delta F_0$ and phase. The main effect of $\Delta F_0$ was significant [$F(4,56) = 10.73$, $p = 0.0002$, GG $= 0.56$], as were those of phase [$F(1,14) = 18.44$, $p = 0.0007$] and their interaction [$F(4,56) = 7.46$, $p = 0.0015$, GG $= 0.57$]. Identification rates are plotted as full symbols in Fig. 15 as a function of $\Delta F_0$ for the same-phase condition (downward pointing triangles) and the antiphase condition (upward pointing triangles). Also plotted are rates obtained in experiment 1 averaged over segment (dotted line).

Because of the difference in identification rate between phases at $\Delta F_0 = 0\%$, patterns of variation with $\Delta F_0$ are not the same for both phase conditions. They cannot meaningfully be averaged, and one cannot speak of a ''$\Delta F_0$ effect'' on the basis of these data. Nevertheless, given that phase effects were small for $\Delta F_0 = 0.4\%$ and 0.8%, one can compare corresponding data points of experiments 1 and 2 and conclude that roving the $F_0$ did not affect identification for $\Delta F_0$'s that size. At $\Delta F_0 = 0\%$, identification was better for same- than for antiphase, possibly because formants of the weaker vowel tended to produce ''bumps'' in the compound spectrum in the first case, and ''dips'' in the second. Spectral peaks are known to be perceptually more prominent than

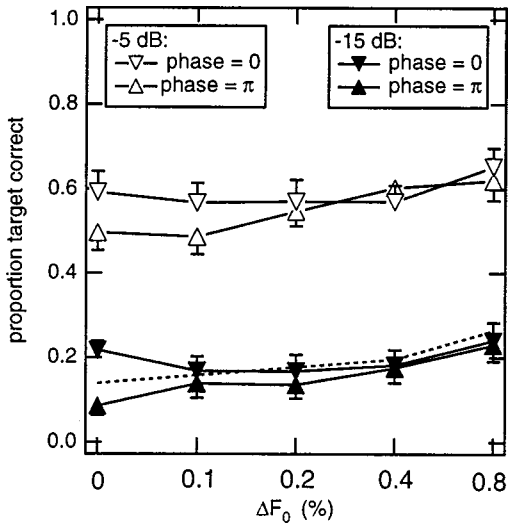


FIG. 15. Experiment 2. Identification rate a function of $\Delta F_0$, at target/competitor ratios of $-5$ dB (open symbols) and $-15$ dB, for the same phase (downward pointing triangles) and opposite phase (upward pointing triangle) conditions. To avoid overlap, only half error bars are plotted. The dotted line represents scores from experiment 1 averaged over segment conditions.

valleys. Experiment 1 gave identification rates intermediate between these two, possibly because the phases used in experiment 1 produced spectra that were less radically favorable or unfavorable. For nonzero $\Delta F_0$'s, the differences between phase conditions were smaller, presumably because ongoing phase shifts quickly disrupted the favorable or unfavorable spectral summation that occurred at $\Delta F_0 = 0$.

For $-5$ dB targets, identification rates for the weaker vowel were submitted to a repeated-measures ANOVA with factors $\Delta F_0$ and phase. The main effect of $\Delta F_0$ was significant [$F(4,56) = 13.55$, $p < 0.0001$, GG$= 0.71$], as were those of phase [$F(1,14) = 9.01$, $p = 0.0095$] and their interaction [$F(4,56) = 4.92$, $p = 0.0035$, GG$= 0.84$]. Identification rates are plotted as open symbols in Fig. 15 as a function of $\Delta F_0$ for the same-phase condition (downward pointing triangles) and the opposite-phase condition (upward pointing triangles). As observed at $-15$ dB, at 0 dB and $\Delta F_0 = 0\%$ the opposite-phase condition was unfavorable for identification compared to the same-phase condition. For $+5$ dB targets, contrary to what was hoped, identification rates were at a ceiling and no interesting effects were observed.

## V. DISCUSSION

### A. $\Delta F_0$ effects

The $\Delta F_0$ effects generally agree with other studies. At 6% they reproduce effects found previously with the same task and similar stimuli (de Cheveigné *et al.*, 1997a, b). At 1.5% and 3% they agree (allowing for differences in experimental procedure) with the observations of Assmann and Summerfield (1990, 1994) and Culling and Darwin (1993, 1994), and show that those observations were not particular to the starting phases they used. Effects found at 0.8% and 0.4% extend those studies and show that identification of concurrent vowels can benefit from $F_0$ differences that are very small indeed. For weak targets ($-15$ dB), effects did not differ radically within the set of starting phases we sampled, nor among different vowel pairs. The step in identification rate between $\Delta F_0 = 0$ and 0.4% was positive for 18 pairs out of 20, and 6 phase conditions out of 6.

On casual listening, stimuli at small $\Delta F_0$'s did not seem to have two pitches, and did not even sound strikingly inharmonic. In natural speech $F_0$'s are unlikely to remain in such close proximity for any period of time. If they did, the imperfect periodicity of real speech would probably prevent such small $\Delta F_0$'s from being exploited. The ecological value of $F_0$-guided segregation mechanisms is no doubt limited to larger $\Delta F_0$'s. Nevertheless, the limit of small $\Delta F_0$'s is of theoretical interest as it reveals the pattern of breakdown of segregation mechanisms as $\Delta F_0$ cues vanish. The gradual decrease of effect size suggests the gradual degradation of a single mechanism rather than a transition to a different mechanism.

### B. Segment and phase effects

Segment effects in experiment 1 were either nonsignificant or small when averaged over vowel pairs. Individual pairs showed relatively strong effects, but only at $\Delta F_0 = 0$ and target/competitor ratio$= 0$ dB. However, these pair-specific effects were not significant at $\Delta F_0 = 0.4\%$, nor at any $\Delta F_0$ at $-15$ dB. Overall, segment effects were small in comparison to $\Delta F_0$ effects.

The segment conditions of experiment 1 represent one particular sample of starting phase spectra (various degrees of uniform delay applied to all partials of both vowels). Experiment 2 used another sample (in phase and antiphase) and found larger effects at both 0 and $-15$ dB. In-phase was favorable and antiphase unfavorable, but this effect was strong only at $\Delta F_0 = 0\%$ and decreased rapidly with $\Delta F_0$'s as small as 0.1% or 0.2%. At its largest, it was of the same order as the $\Delta F_0$ effect observed between 0% and 0.4%, that is, small.

### C. Evidence for a beat model

Given the small size of segment effects in some conditions, or inconsistency across vowel pairs in others, it is unlikely that the waveform interactions that produced them also produced the large and consistent $\Delta F_0$ effects. An interesting comparison can be made between identification rates at $\Delta F_0 = 0.4\%$ and 1.5%. Beat patterns at different $\Delta F_0$'s differ by a mere stretching factor. For example, the patterns found within a 62.5-ms portion of a stimulus with $\Delta F_0 = 1.5\%$ are also found within a 250-ms portion of a stimulus with $\Delta F_0 = 0.375\%$ (nominally: 0.4%). Neglecting onset effects, the four $\Delta F_0 = 0.4\%$ stimuli thus collectively contain the same set of patterns as the single stimulus used at $\Delta F_0 = 1.5\%$. Supposing that a ''best'' pattern exists somewhere within the latter stimulus, it also appears within one of the former. Based on this reasoning, the best rate over segments at 0.4% should be similar to that obtained at 1.5%. This prediction was tested by means of a repeated-measures ANOVA with factors ($\Delta F_0 = 0.4\%, 1.5\%$) $\times$ (pair), using for each pair in the $\Delta F_0 = 0.4\%$ condition the segment that gave

the best scores for that pair (on average over subjects). Identification at $\Delta F_0 = 1.5\%$ was significantly better (0.41) than the best score over 0.4% segments (0.26) [$F(1,14) = 25.7$, $p = 0.0002$]. This was despite the fact that the segment selection process biased the 0.4% scores positively, and that the slower fluctuations at that $\Delta F_0$ should have made spectral samples easier to exploit.

Another difficulty for the ''glimpsing'' hypothesis is that strong $\Delta F_0$ effects occur in conditions for which beat-induced variations of the excitation pattern are small. An example is the sum of /o/ and /u/ at an o/u relative level of 15 dB (Fig. 7, bottom). Beat amplitudes are small, and yet $\Delta F_0$'s produce a strong increase in identification rate for the weaker vowel (/u/) of this pair (Fig. 14). A previous study (de Cheveigné *et al.*, 1997a) found no tendency for $\Delta F_0$ effects to be largest at relative levels that favor strong beats near a target vowel's formants.

It remains to account for the evidence that Assmann and Summerfield (1994) found in favor of the ''glimpsing'' hypothesis. They presented subjects with four consecutive 50-ms segments gated from a 200-ms double vowel. $\Delta F_0$'s ranged from $\frac{1}{4}$ to 4 semitones (approximately 1.5, 3, 6, 12.5, and 26 Hz given the baseline $F_0$ of 100 Hz), and therefore the ''effective duration'' of the windowed stimulus (38.6 ms between $-6$-dB points) was in most cases shorter than the beat period (667, 333, 167, 80, and 38.5 ms, respectively). Identification varied between segments, except at $\Delta F_0 = 0$, where the four segments were identical, and at four semitones where the beat period was equal to the effective stimulus duration. The identification rate of the 200-ms stimulus could be predicted from that of the best 50-ms segment that it contained. This prediction failed to generalize to the present study, where segment effects were observed in some conditions, but were inconsistent across vowel pairs and vanished when averaged. However, one cannot exclude that such effects might have survived averaging with a different vowel set, for example, that used by Assmann and Summerfield. A factor that may have enhanced segment effects in their study was that their stimuli were short (38.6 ms between $-6$-dB points) in relation to the pitch period (about 10 ms). They were also synthesized with Klatt phase, which produces relatively peaky waveforms, so that differences in alignment of the small number of peaks (three to four) relative to segment boundaries might have led to stronger effects than observable with the longer stimuli (250 ms), shorter periods (7.6 ms), and random phase spectra of the present study.

### D. Evidence for a ''multiplicity cue''

The task allowed subjects to report one or two vowels at will. One can speculate whether their choice was the result of the second vowel becoming more identifiable, or the result of a mere ''multiplicity'' cue that signaled that more than one vowel was present.

Suppose that subjects reported a second vowel entirely on the basis of a ''multiplicity cue,'' and that the probability of the second vowel being correct was unaffected by $\Delta F_0$. The identification rate should follow a linear function of the proportion of two-vowel responses. The lowest dotted line in Fig. 12(a) shows the rate expected if identification of the second vowel were at chance ($p_c = 0.25$). The middle dotted line shows the rate for $p_c = 0.5$, and the uppermost for $p_c = 1.0$ (identification limited only by the reluctance of subjects to report two vowels). The match is not particularly good, indicating that subjects' responses are unlikely to result entirely from a ''multiplicity'' cue. $\Delta F_0$ evidently also improved the identifiability of the second vowel.

At the other extreme is the hypothesis that subjects reported a second vowel when it was correctly identifiable, possibly in addition to a fixed number of incorrect answers. The dotted line in Fig. 12(b) predicts the response count as an affine function of the identification rate with slope one, supposing 15% of incorrect responses. Given the similarity to the measured data this hypothesis is tenable, but it is of course not the only possible account for the data. It is likely that $\Delta F_0$ enhances *both* identifiability *and* a general perception of multiple sources, and that both of these aspects determined responses.

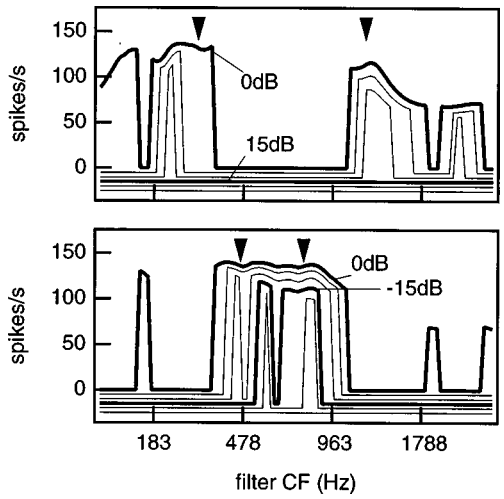


FIG. 16. Channel selection model of Meddis and Hewitt (1992). Top: output of channels *not* dominated by the periodicity of /o/ for o/u ratios of 0 to 25 dB in 5 dB steps. Nonselected channels are set to zero, and curves are offset vertically for clarity. Note that for o/u ratios$\geqslant$15 dB *all* channels are dominated by /o/, and the output is everywhere 0. Bottom: same for channels not dominated by /u/, for o/u ratios of $-25$ to 0 dB in 5-dB steps. In this case the model succeeds in partitioning channels at all ratios except at o/u$=-25$ dB. Arrows indicate formants $F1$ and $F2$ of /u/ (top) and /o/ (bottom).

### E. Evidence for a channel-selection model

The concurrent vowel identification model of Meddis and Hewitt (1992) partitions the set of peripheral channels into two sets, dominated by one vowel or the other, on the basis of their periodicity as measured from the position of the ''period peak'' of the autocorrelation function (ACF). The partition is illustrated in Fig. 16. The top panel shows the output of channels left over after eliminating channels dominated by the periodicity of /o/ (channels dominated by /o/ are set to zero). Only data for positive o/u relative levels are shown. The lower panel similarly shows channels selected because they were not dominated by /u/, for negative o/u relative levels. Such partitions do not depend on the size of the $\Delta F_0$, so *in principle* the model is effective for arbi-

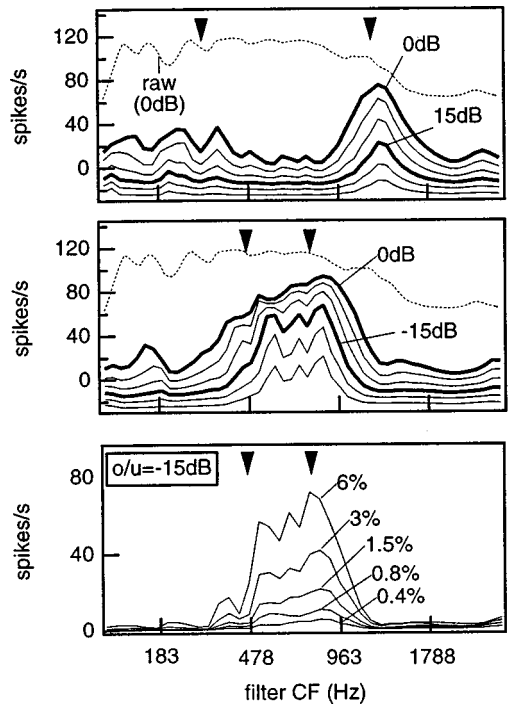


FIG. 17. Within-channel cancellation model of de Cheveigné (1997b). Top: output (full lines) of the harmonic cancellation filter as a function of CF, for o/u ratios of 0 to 25 dB in 5-dB steps, at $\Delta F_0 = 6\%$. The thin dotted line represents the input at 0 dB. The filter was tuned to cancel the period of /o/ (stronger vowel). Curves are offset vertically for clarity. Middle: same for o/u ratios of $-25$ to 0 dB. In this case the filter was tuned to cancel the period of /u/. Bottom: output of the cancellation filter tuned to cancel /u/, for o/u ratio $= -15$ dB and various values of $\Delta F_0$. Arrows indicate formants $F1$ and $F2$ of /u/ (top) and /o/ (middle and bottom).

trarily small $\Delta F_0$'s. In practice, of course, smaller $\Delta F_0$'s strain the temporal resolution of the model. For example, discrimination between channels when $F_0$'s differ by 0.4% (at 132 Hz) requires that ACF peak positions be determined with a time resolution of about 30 $\mu$s.

The pattern of selected channels characterizes the weaker vowel in a straightforward way, and it should be easy to exploit (Meddis and Hewitt's model used the low-lag portion of the summary autocorrelation function, but other schemes are possible). The main weakness of the scheme is that it breaks down if there is a level mismatch and all channels are dominated by the same vowel. Such is the case for o/u ratios of $-25$, 15, 20, and 25 dB (Fig. 16). Meddis and Hewitt's model cannot explain the $\Delta F_0$ effects for u/o plotted in Fig. 14.

### F. Evidence for a within-channel cancellation model

The author recently proposed a concurrent vowel identification model based on within-channel filtering to suppress responses at the period of the stronger vowel (de Cheveigné, 1997b). The top panel of Fig. 17 shows the output across channels of an array of cancellation filters tuned to the period of the vowel /o/ (data are restricted to positive o/u relative levels). The thin dotted line shows the profile before filtering (for o/u ratio $= 0$ dB). The middle panel shows the same for a filter tuned to suppress instead the period of /u/ (for negative o/u relative levels). As in Meddis and Hewitt's model, the weaker vowel is represented in a straightforward fashion. The advantage, over their model, is that the representation does not break down at extreme relative levels. It can thus explain $\Delta F_0$ effects observed for very weak targets (de Cheveigné, 1997a, 1999).

Like Meddis and Hewitt's model, this model can work *in principle* for $\Delta F_0$'s however small. The lower panel of Fig. 17 shows the output of the filter array for various values of $\Delta F_0$. The magnitude of the output decreases approximately linearly with $\Delta F_0$, but in terms of ratio the pattern remains as prominent. In practice, of course, the implementation requires fine temporal resolution (as Meddis and Hewitt's model), and the neural representation must be sufficiently precise to allow the output of the cancellation filter to emerge from noise.

Like Meddis and Hewitt's model, this model assumes that the competitor is harmonic. It could possibly be extended to explain the results of Culling and Darwin (1993) ($F_0$'s swapped between first and second formant regions) by assuming that cancellation filters are tuned differently in different channels. However, it is difficult to explain the results of Culling and Darwin (1994) with interleaved harmonics (every other partial swapped between vowels). Those results remain a mystery as far as this model is concerned.

## VI. CONCLUSIONS

The $F_0$ differences between concurrent vowels result in complex patterns of beats and waveform interactions on the basilar membrane. These can affect identification in a phase-dependent way, but it is unlikely that they account entirely for improvements in identification with $\Delta F_0$, even at small $\Delta F_0$'s. Effects of $\Delta F_0$ decrease as this parameter tends to zero, but values as small as 0.4% ($\frac{1}{16}$ of a semitone) still produce a measurable benefit in terms of identification rate. To summarize, we have the following.

(1) For weak targets (target/competitor ratio $= -15$ dB), identification was improved by the presence of $\Delta F_0$'s as small as 0.4%. The effect was observed for six different phase conditions, and is unlikely to be specific to a particular choice of starting phase.
(2) Identification was in some cases affected by the choice of starting phase, confirming earlier observations by Assmann and Summerfield (1994). Effects were, however, small. In one condition (target/competitor ratio $= 0$ dB, $\Delta F_0 = 0$), strong segment effects were observed for certain pairs. However, these pair-specific effects did not generalize to $\Delta F_0 \neq 0$, or to target/competitor ratio $= 0$ dB.
(3) Given the small size of starting phase effects, it is unlikely that improvements in identification with $\Delta F_0$ are the result of a mechanism exploiting waveform interactions or beats, at least for weak targets ($-15$ dB).
(4) Effects of small $\Delta F_0$'s can, in principle, be explained by the channel-selection model of Meddis and Hewitt (1992), except at extreme relative levels for which the model breaks down. The model requires high temporal resolution (on the order of 30 $\mu$s).

(5) They can also, in principle, be explained by the within-channel cancellation model of de Cheveigné (1997b). That model works, in principle, at all relative levels. In addition to high temporal resolution, it requires good linearity and a wide dynamic range.

## ACKNOWLEDGMENTS

The experiments were carried out at ATR Human Information Processing Research Laboratories, within a research agreement between ATR and the Centre National de la Recherche Scientifique and University of Paris 7. The author thanks ATR for its kind hospitality, and the CNRS for leave of absence. Hideki Kawahara participated in the preparation, and Rieko Kubo supervised the experiments. Thanks to John Culling for providing the software for stimulus synthesis, and to him, Quentin Summerfield, Ray Meddis, and Sébastien Dupuis for comments. Thanks to Brian Moore, an anonymous reviewer, and the editor for their criticism. A previous version of this paper benefited from useful criticism of Winifred Strange, Chris Darwin, and one anonymous reviewer.

[1]According to Hartmann (1997, p. 462), beats between two partials are as detectable as if their envelope were reduced to its dc and fundamental component. For the deepest beats, the effective modulation percentage is thus $m=\frac{2}{3}$, implying a peak/valley ratio of 12 dB rather than infinite. For beats that are less deep, the envelope shape is closer to sinusoidal, and therefore two-partial beats are approximately equivalent to sinusoidal beats of similar peak/valley ratio.

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

Assmann, P. F., and Summerfield, Q. (**1994**). ''The contribution of waveform interactions to the perception of concurrent vowels,'' J. Acoust. Soc. Am. **95**, 471–484.

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

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

Culling, J. F. (**1996**). ''Signal processing software for teaching and research in psycholinguistics under UNIX and X-windows,'' Behav. Res. Methods Instrum. Comput. **28**, 376–382.

Culling, J. F., and Darwin, C. J. (**1993**). ''Perceptual separation of simultaneous vowels: Within and across-formant grouping by *F*0,'' J. Acoust. Soc. Am. **93**, 3454–3467.

Culling, J. F., and Darwin, C. J. (**1994**). ''Perceptual and computational separation of simultaneous vowels: Cues arising from low frequency beating,'' J. Acoust. Soc. Am. **95**, 1559–1569.

Darwin, C. J., and Carlyon, R. P. (**1995**). ''Auditory grouping,'' in *Handbook of Perception and Cognition: Hearing*, edited by B. C. J. Moore (Academic, New York), pp. 387–424.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**). ''Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers,'' J. Acoust. Soc. Am. **102**, 2892–2905.

de Cheveigné, A. (**1997a**). ''Ten experiments in concurrent vowel segregation,'' ATR Human Information Processing Research Labs technical report, TR-H-217.

de Cheveigné, A. (**1997b**). ''Concurrent vowel segregation III: A neural model of harmonic interference cancellation,'' J. Acoust. Soc. Am. **101**, 2857–2865.

de Cheveigné, A. (**1999**). ''Vowel-specific effects in concurrent vowel segregation,'' J. Acoust. Soc. Am. **106**, 327–340.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (**1997a**). ''Concurrent vowel identification. I: Effects of relative level and *F*0 difference,'' J. Acoust. Soc. Am. **101**, 2839–2847.

de Cheveigné, A., McAdams, S., and Marin, C. (**1997b**). ''Concurrent vowel identification. II: Effects of phase, harmonicity and task,'' J. Acoust. Soc. Am. **101**, 2848–2856.

de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (**1995**). ''Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement,'' J. Acoust. Soc. Am. **97**, 3736–3748.

Geisser, S., and Greenhouse, S. W. (**1958**). ''An extension of Box's results on the use of the F distribution in multivariate analysis,'' Ann. Math. Stat. **29**, 885–889.

Hartmann, W. H. (**1997**). *Signals, Sound, and Sensation* (AIP, Woodbury, NY).

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

Kohlrausch, A., and Sander, A. (**1995**). ''Phase effects in masking related to dispersion in the inner ear. II: Masking period patterns of short targets,'' J. Acoust. Soc. Am. **97**, 1817–1829.

Kuwabara, H. (**1983**). ''Vowel identification and dichotic fusion of time-varying synthetic speech sounds,'' Acustica **53**, 143–151.

Marin, C., and McAdams, S. (**1996**). ''The role of auditory beats induced by frequency modulation and polyperiodicity in the perception of spectrally embedded target sounds,'' J. Acoust. Soc. Am. **100**, 1736–1753.

Meddis, R., and Hewitt, M. J. (**1992**). ''Modeling the identification of concurrent vowels with different fundamental frequencies,'' J. Acoust. Soc. Am. **91**, 233–245.

Moore, B. C. J., and Alcántara, J. I. (**1995**). ''Identification of flat-spectrum vowels on the basis of amplitude modulation,'' J. Acoust. Soc. Am. **97**, 3274.

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

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (**1992**). ''Complex sounds and auditory images,'' in *Auditory Physiology and Perception*, edited by Y. Cazals, K. Horner, and L. Demany (Pergamon, Oxford), pp. 429–446.

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.

Slaney, M. (**1993**). ''An efficient implementation of the Patterson-Holdsworth auditory filter bank,'' Apple Computer technical report, 35.

Strange, W., and Bohn, O.-S. (**1998**). ''Dynamic specification of coarticulated German vowels: Perceptual and acoustical studies,'' J. Acoust. Soc. Am. **104**, 488–504.

Strube, H. W. (**1985**). ''A computationally efficient basilar-membrane model,'' Acustica **58**, 207–214.

Summerfield, Q., and Assmann, P. A. (**1987**). ''Auditory enhancement and the perception of concurrent vowels,'' Percept. Psychophys. **45**, 529–536.

Summerfield, Q., Sidwell, A., and Nelson, T. (**1987**). ''Auditory enhancement of changes in spectral amplitude,'' J. Acoust. Soc. Am. **81**, 700–708.

Summerfield, Q., Haggard, M., Foster, J., and Gray, S. (**1984**). ''Perceiving vowels from uniform spectra: phonetic exploration of an auditory aftereffect,'' Percept. Psychophys. **35**, 203–213.

Yost, W. A., and Sheft, S. (**1989**). ''Across-critical-band processing of amplitude-modulated tones,'' J. Acoust. Soc. Am. **85**, 848–857.