Concurrent vowel identification. I. Effects of relative amplitude and F_0 difference[a)]

Alain de Cheveigné
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

Hideki Kawahara, Minoru Tsuzaki, and Kiyoaki Aikawa
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)

Subjects identified concurrent synthetic vowel pairs that differed in relative amplitude and fundamental frequency (F_0). Subjects were allowed to report one or two vowels for each stimulus, rather than forced to report two vowels as was the case in previously reported experiments of the same type. At all relative amplitudes, identification was better at a fundamental frequency difference (ΔF_0) of 6% than at 0%, but the effect was larger when the target vowel amplitude was below that of the competing vowel (-10 or -20 dB). The existence of a ΔF_0 effect when the target is weak relative to the competing vowel is interpreted as evidence that segregation occurs according to a mechanism of cancellation based on the harmonic structure of the competing vowel. Enhancement of the target based on its own harmonic structure is unlikely, given the difficulty of estimating the fundamental frequency of a weak target. Details of the pattern of identification as a function of amplitude and vowel pair were found to be incompatible with a current model of vowel segregation.
 [S0001-4966(97)03904-0]

PACS numbers: 43.71.An, 43.71.Es, 43.66.Ba, 43.66.Lj [WS]

INTRODUCTION

Various experiments have shown that identification of two synthetic vowels that are mixed together improves when the vowels differ in fundamental frequency (F_0) (Scheffers, 1983; Summerfield and Assmann, 1991; Culling and Darwin, 1993, 1994; Assmann and Summerfield, 1994). The results of several of these studies are shown in Fig. 1. Despite differences in task, vowel set, subjects, etc., a common trend is that identification improves as the F_0 difference (ΔF_0) increases between vowels. However, it has been noted that identification is far above chance at $\Delta F_0 = 0\%$, and yet remains less than perfect when F_0's are different.

The limited effect of F_0 differences may be due partly to ceiling effects. If identification is perfect at $\Delta F_0 = 0\%$ for certain subjects or conditions, there is no room left for improvement with ΔF_0. De Cheveigné *et al.* (1995a) reasoned that such might be the case for one vowel within a pair if it dominated its companion. They tried to determine corrective amplitude factors to balance relative dominance and mutual masking. However, that reasoning was flawed: There is nothing to guarantee that, once the balance is attained, identification of *both* vowels will not be at ceiling. In the present study, we followed the opposite course and introduced a systematic amplitude imbalance to reduce ceiling effects for the weaker vowel. The first aim of the study was to test that idea, and determine good amplitude imbalance factors for future experiments. For that purpose, we measured identification rate as a function of both ΔF_0 and relative amplitude between vowels. This experiment allows a link to be made between the classic paradigm in which identification rates are measured with constant stimuli (Fig. 1), and more recent paradigms in which thresholds of identification are determined adaptively (Demany and Semal, 1990; Assmann and Summerfield, 1990; Summerfield, 1992; Summerfield and Culling, 1992; Culling *et al.*, 1994; Culling and Summerfield, 1995). Finally, the dependency of identification on relative amplitude and ΔF_0 may be used to test theories of vowel perception or segregation.

McKeown (1992) performed a similar experiment using double-vowel stimuli with ΔF_0's of 0%, 25%, and 100% (relative to the lower F_0). Relative amplitude between vowels was varied between -14 and 14 dB in 2-dB steps, by reducing the amplitude of either vowel from the equal-amplitude condition (thereby reducing the overall stimulus level). The experiment we report here used a larger range (-20–20 dB), a larger step (10 dB), and overall rms amplitude was held constant for all stimuli. McKeown required subjects to identify a ''dominant vowel'' and a second vowel, and analyzed results separately for each. We also scored the constituents of vowel pairs separately, but according to stimulus properties rather than a subjective judgement.

In the classic double vowel identification paradigm, subjects are presented with two vowels and requested to identify both of them, whether they are both audible or not. This has several drawbacks: (a) The task is uncomfortable when one vowel is inaudible and the subject must guess; (b) the subject may use one particular vowel as a default response, and thus

[a)]Portions of this work were presented at a meeting of the Acoustical Society of Japan and as an ATR Human Information Processing Research Laboratories technical report (de Cheveigné, 1995; de Cheveigné *et al.*, 1995b).

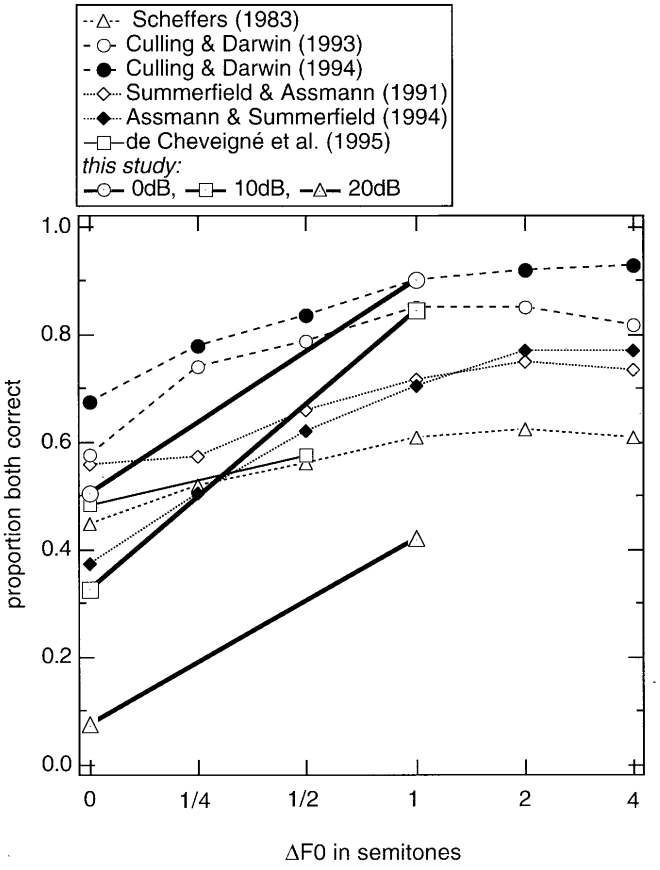

FIG. 1. Proportion of ''both correct'' identification as a function of ΔF_0 for previous studies (thin lines), and for this report (thick lines).

unwittingly score perfect identification on that vowel; (c) segregation cues that signal the multiplicity of sources are ignored; and (d) subjects are under pressure to improve their identification. This might enhance training effects that may contribute to reduce the size of ΔF_0 effects (Assmann and Summerfield, 1994). Instead of requiring two vowel responses, we told our subjects that the stimuli contained either single or double vowels, and we requested them to give either one or two answers. This task is typical of natural situations where segregation occurs, and the number of vowels reported is an interesting measure.

I. METHODS

Steady-state Japanese vowels /a/, /i/, /u/, /e/, and /o/ were synthesized at two fundamental frequencies: 125 and 132.5 Hz, with equal rms signal amplitude. Details of the vowels and synthesis technique are given in the Appendix. Double vowels were created by scaling one vowel by a factor (-20, -10, 0, 10, or 20 dB), adding the two vowels, and setting the rms amplitude of the sum to a standard value. Stimuli were 200 ms in duration, with 20-ms raised-cosine onset and offset ramps. They were presented to subjects via headphones (Stax SR-Λ), at a sound-pressure level between 63 and 70 dBA. The sound system was calibrated using a Bruel&Kjaer artificial ear (sound level meter type 2231, half-inch microphone type 4134).

For a given vowel pair, all four combinations of the two F_0's were used to produce ΔF_0's of 0% and 6%. Vowels within a pair were always different. There were three repetitions of each condition, for a total of 600 double-vowel stimuli (2 ΔF_0's)$\times$(5 amplitude differences)$\times$(10 unordered vowel pairs)$\times$(2 F_0 orders)$\times$(3 repetitions). To these were added 240 single vowel stimuli (2 F_0's)$\times$(5 vowels)$\times$(24 repetitions). A relatively large number of single vowels were included to ensure that the stimulus set was consistent with the description given to the subjects. It also allowed us to verify that vowel quality was acceptable and unaffected by synthesis parameters such as F_0.

Subjects were six native speakers of Japanese, aged 18 to 26 years. Two were male (K, S) and four female (N, M, T, U). Two were members of the ATR staff (N, M), and the other four were students paid for their services. Each subject performed five sessions on different days. Subjects were seated in a sound-treated booth or room, in front of a computer terminal that was used to give prompts and gather results. Each stimulus was presented once, and the subjects were requested to give one or two responses, according to the number of vowels they perceived to be present. The subjects were informed that the stimulus contained one or two vowels belonging to the set /a/, /i/, /u/, /e/, /o/, and that, in the case of double vowels, vowels within a pair were different. They were not allowed to answer twice the same vowel, but they had the option to answer ''*x*'' instead of a vowel that they could not identify.[1] They could pause at will, in which case the last stimulus before the pause was presented again after the pause. A session typically lasted between one and two hours. There was no feedback.

The answers were scored to obtain three measures: (a) Average *number of vowels reported* per stimulus. This is the proportion of stimuli that elicited two responses, regardless of whether they were correct or not, and regardless of whether the stimulus contained one vowel or two (b) *Combination-correct identification rate*. This is the proportion of double-vowel stimuli for which both constituents were correctly identified. It is the measure most commonly reported for double-vowel experiments (Fig. 1) (c) *Constituent-correct identification rate* (Lea, 1992; de Cheveigné *et al.*, 1995a). In the case of double vowels, this rate was obtained by scoring each stimulus twice, once for each vowel. A constituent was scored as correctly identified if it was matched by the response vowel or, if two responses were given, by either of the two response vowels. In the case of single vowels, each stimulus was scored as correct if the stimulus vowel was matched by the response vowel (or either of the response vowels if two responses were given).

For the first two measures, the order of vowels within a pair is ignored. In counting conditions we consider three levels of intervowel amplitude difference (0, 10, 20 dB), and ten conditions of unordered vowel pair, in addition to two levels of ΔF_0 (0%, 6%). The third measure distinguishes the order of vowels within a pair (target/background). There are 20 levels of ordered vowel pair, and 5-levels of target-to-background rms level (-20, -10, 0, 10, 20 dB) according to whether the vowel being considered (the ''target'') is the weaker or the stronger within the pair. In addition, there are

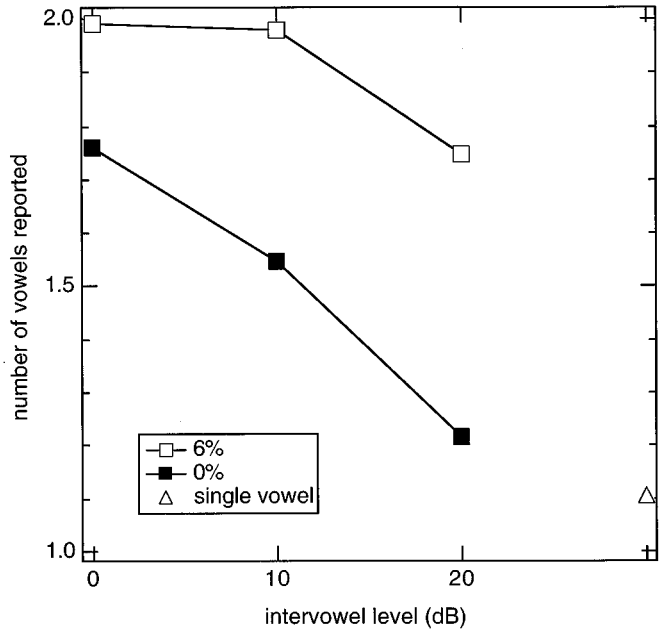

FIG. 2. Number of vowels reported as a function of rms intervowel amplitude difference, for ΔF_0=0% (filled symbols) and ΔF_0=6% (open symbols). The triangle represents single vowels.

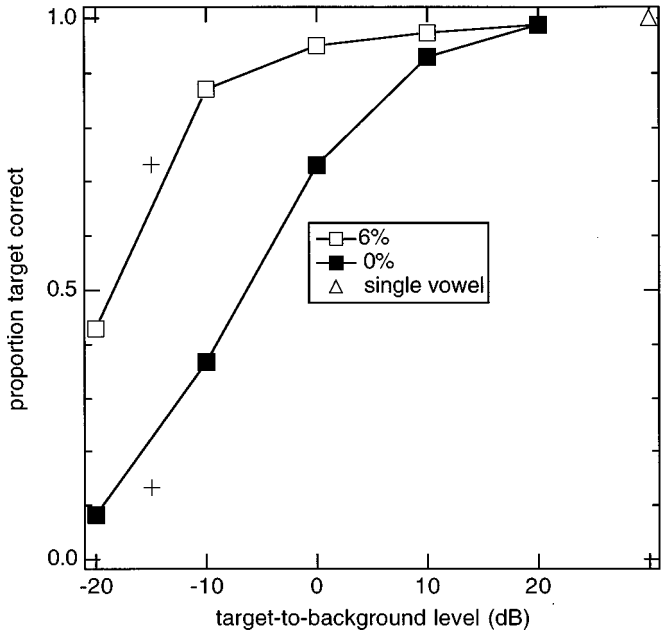

FIG. 3. Target-correct identification rate as a function of rms target-to-background level, for ΔF_0=0% (filled symbols) and ΔF_0=6% (open symbols). Crosses represent rates obtained with the same subjects at −15 dB in a similar experiment (de Cheveigné *et al.*, 1997). The triangle represents single vowels.

two levels of ΔF_0 (0%,6%), and also two levels of absolute F_0 (low/low versus high/high, and low/high versus high/low).

II. RESULTS

A. ΔF_0 and relative amplitude

The number of vowels reported per stimulus is plotted in Fig. 2 as a function of amplitude difference for both values of ΔF_0. The triangle is for single vowels. For double vowels at ΔF_0=0%, subjects tended to report two vowels when both stimulus vowels had the same rms amplitude (0 dB). When either vowel was stronger (10 and 20 dB), they more often reported a single vowel, but they still reported two vowels for a certain proportion of trials, even when the stimulus contained only one vowel (triangle). At ΔF_0=6% they almost always reported two vowels. Thus ΔF_0 seems to function as a cue indicating the *multiplicity* of sources within the stimulus.

Combination-correct scores are plotted in Fig. 1 as a function of ΔF_0, for all three values of amplitude difference. At 0 dB the ΔF_0 effect is similar, if somewhat larger, to that reported in previous studies. At 10 and 20 dB the overall rates are lower but the ΔF_0 effect remains large.

Constituent-correct identification rates for double vowels were analyzed in more detail. Rates averaged over session (5) and repetition (3) were subjected to a repeated-measures analysis of variance (ANOVA). Probabilities reflect, where necessary, an adjustment of the degrees of freedom by a factor that corrects for the inherent correlation of repeated measurements (Geisser and Greenhouse, 1958). Only the three lowest (−20, −10, 0 dB) of the five levels of target-to-background AMPLITUDE were retained, as responses were essentially perfect at 10 and 20 dB, with relatively little variability. The other factors were ΔF_0(2), ordered vowel PAIR (20) and absolute F_0(2). Significant effects were noted for factors ΔF_0 [$F(1,5)=51.80$, $p=0.0008$], AMPLITUDE [$F(2,10)=200.31$, $p<0.0001$, GG=0.64] and PAIR [$F(19,95)=5.91$, $p=0.0045$, GG =0.18], as well as for interactions between ΔF_0 and AMPLITUDE [$F(2,10)=11.45$, $p=0.008$, GG=0.7], AMPLITUDE and PAIR [$F(38,190)=4.00$, $p=0.013$, GG=0.11], and ΔF_0 and PAIR [$F(19,95)=4.65$, $p=0.02$, GG=0.15]. The main effect of F_0 order was not significant, nor were any of the interactions involving it. In other words, at ΔF_0=0% it made no difference whether both vowels were at 125 or 132.5 Hz, and, at ΔF_0=6%, it made no difference whether the target was the lower or the higher of the two frequencies, whatever the vowel pair or amplitude difference. Given the various ways in which F_0 can interact with formant structure, this result is perhaps surprising.

Constituent-correct identification rates are plotted in Fig. 3 as a function of amplitude difference for both values of ΔF_0. As one might expect, identification of a vowel was better when its relative amplitude was greater. It was also better at ΔF_0=6% than at ΔF_0=0%, particularly when the target amplitude was low (−20 or −10 dB).

B. Vowel pairs

Results for two particular pairs are described in detail. The first pair, /o/+/u/, serves later on in Sec. III E to test a model of double-vowel segregation (Meddis and Hewitt, 1992). The second pair, /e/+/o/, was chosen to illustrate the variety of response patterns for different pairs. Figure 4 shows the number of vowels reported for both pairs. The pattern for /o/+/u/ is typical of the average (Fig. 2), but that

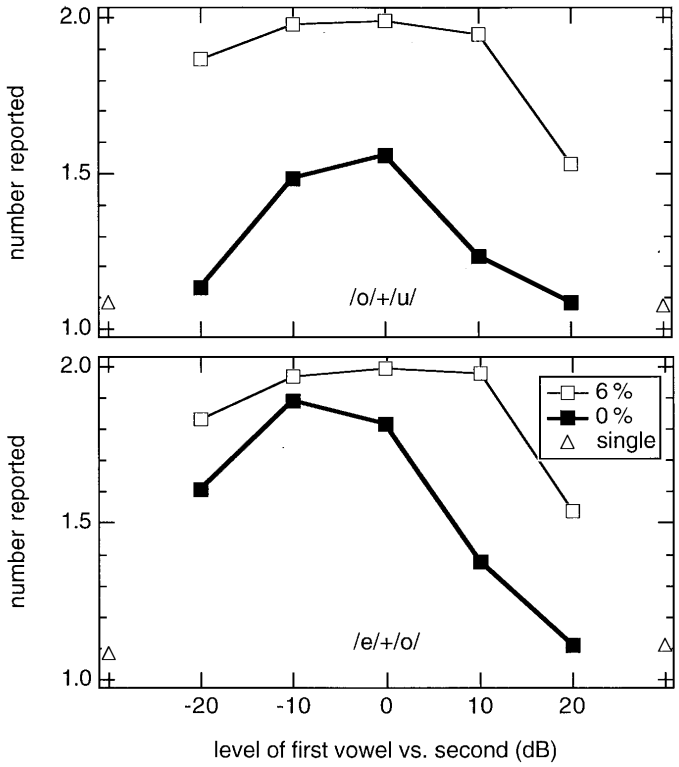

FIG. 4. Number of vowels reported for vowel pairs /o/+/u/ (top) and /e/+/o/ (bottom) as a function of the amplitude of the first vowel relative to the second, at ΔF_0=0% (filled symbols) and ΔF_0=6% (open symbols). Triangles represent single vowels.

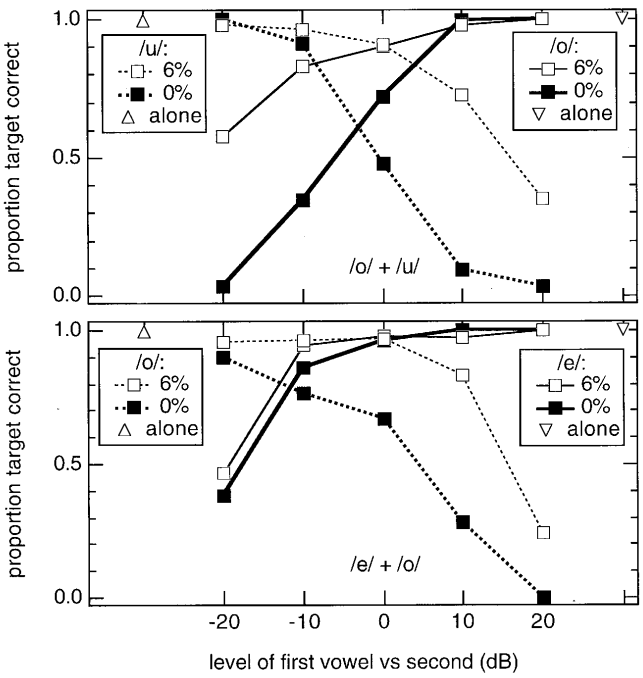

FIG. 5. Target-correct identification rate for vowel pairs /o/+/u/ (top) and /e/+/o/ (bottom) as a function of the amplitude of the first vowel relative to the second, at ΔF_0=0% (filled symbols) and ΔF_0=6% (open symbols). Ascending lines (continuous) are for the first vowel, descending lines (dotted) are for the second vowel in the pair. Triangles represent single vowels.

for /e/+/o/ is asymmetric: ΔF_0 has less effect on the response count when /o/ is dominant and /e/ is weak than vice versa. Figure 5 shows the identification rate for the same pairs. The pattern of identification for components of /o/+/u/ is typical of the average (Fig. 3), and the same is true for the /o/ vowel of /e/+/o/ (Fig. 5, lower panel, descending, dotted lines). However for /e/ the effect of ΔF_0 is very small, mainly because the identification rate is high at ΔF_0=0%. The patterns for the vowels within this pair are quite asymmetric. Vowel-pair specificities are discussed in terms of possible models of vowel segregation in Secs. III D and E.

C. Subject differences

Results of three subjects (K, T, and M) are presented in Figs. 6–8 to illustrate differences between subjects. The effect of ΔF_0 on the number of vowels reported was overall larger for T and M than for K (Fig. 6). The same was true for the identification rate (Fig. 7). This was mainly due to the fact that K had higher identification rates at ΔF_0=0% than T and M, and also a lower rate at −20 dB and ΔF_0=6%. The identification rate *conditional* on a two-vowel response (Fig. 8) reveals marked intersubject differences, mainly at ΔF_0=0%. Conditional rates for subject M are high, perhaps because she was more conservative than other subjects in reporting two vowels, and thus more often correct when she did so. However, they are also high for subject K, despite the fact that he was the least conservative in reporting two vowels. The one-or-two response task probably exaggerates in-

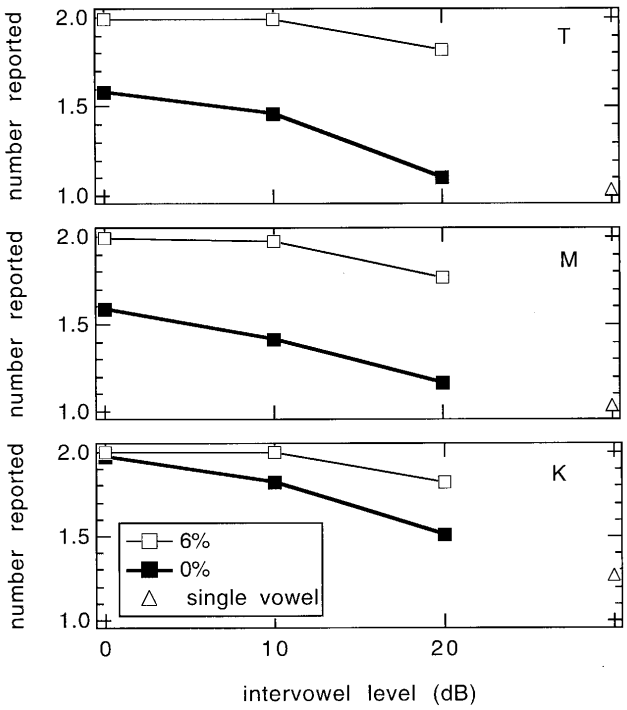

FIG. 6. Number of vowels reported for three subjects as a function of rms intervowel amplitude difference, at ΔF_0=0% (filled symbols) and at ΔF_0=6% (open symbols). Triangles represent single vowels.

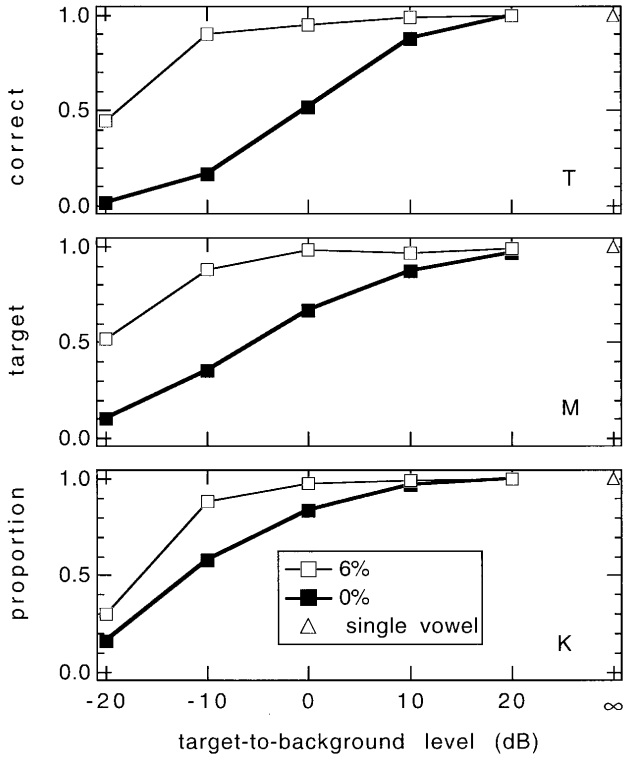

FIG. 7. Identification rate as a function of rms target-to-background level for three subjects, at ΔF_0=0% (filled symbols) and at ΔF_0=6% (open symbols). Triangles represent single vowels.

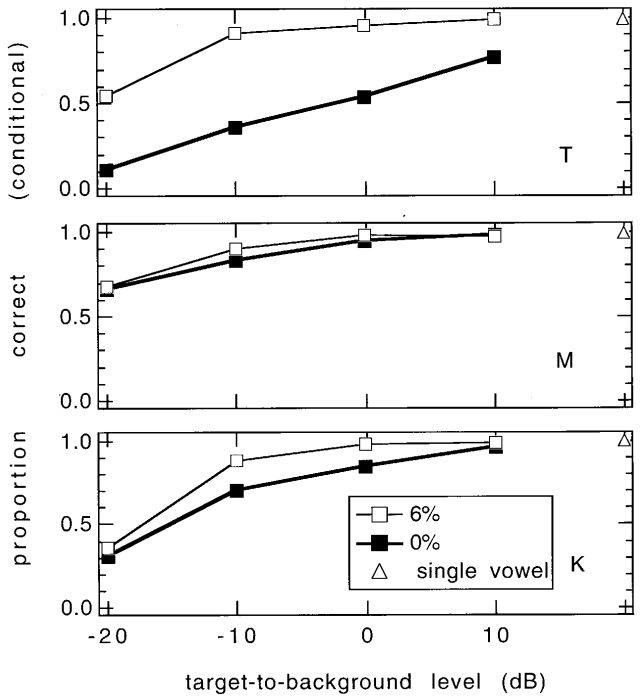

FIG. 8. Identification rate conditional on a two-vowel response (proportion of correct responses for stimuli that evoked two responses) for three subjects as a function of rms target-to-background level at ΔF_0=0% (filled symbols) and at ΔF_0=6% (open symbols). Triangles represent single vowels.

tersubject differences, as subjects may use their freedom differently when permitted to report one or two vowels.

D. Single vowels

The stimulus set contained 240 single vowels in addition to 600 double vowels. For single vowels, the overall identification rate was 99.75%. The lowest rate for a subject (subject N) was 99.3%, and the lowest rate for a vowel (/i/) was 99.2%. Evidently subjects had no difficulty identifying any of the vowels. About 10% of all single vowels evoked two-vowel responses, with considerable differences between subjects (27% for subject K, 2% for subject U), but only small differences between vowels. No effect of F_0 was evident.

III. DISCUSSION

A. ΔF_0 effect

To allow comparison with previous results, combination-correct scores were plotted in Fig. 1. The ΔF_0 effect at 0 dB appears to be larger than in previous studies. Our one-or-two response task probably enhanced effect size, as discussed by Cheveigné *et al.* (1997). Assmann and Summerfield (1994) also found relatively large effects, possibly because their subjects were allowed to make ''both-same'' responses. These may have played a role similar to ''single-vowel'' responses in our task, with the result of relatively low identification rates at ΔF_0=0%, as in our experiment.

B. Interaction with relative amplitude

The effects of $\Delta F0$ at low target amplitudes are relevant to voice segregation in everyday life, since naturally competing voices rarely stay at the same amplitude. When a target voice is weaker it can benefit from F_0-guided segregation, whereas when it is stronger, segregation is less necessary. Drawing a horizontal line in Fig. 3 at a performance level of 70%, the ΔF_0 effect appears to be equivalent to an amplitude difference of about 13 dB. This may be compared with the 17-dB shift in masked threshold (70% correct) found by Culling *et al.* (1994) using an adaptive technique. There are at least two possible explanations for the increase in effect size as targets get weaker. One is that lower identification rates reduce ceiling effects, as we aimed for when we designed the experiment. The other is that estimation of the competing vowel's F_0 (required by segregation models that invoke harmonic cancellation) is more accurate when targets are weak. If the first explanation were correct, then other manipulations that make the task more difficult should also produce large effects. Instead, Assmann and Summerfield (1990) found that reducing stimulus duration from 200 to 52 ms practically abolished ΔF_0 effects. De Cheveigné *et al.* (1995a) used a stimulus set with high intravowel variability to reduce overall performance, yet they obtained relatively small ΔF_0 effects. Shackleton *et al.* (1994) manipulated formant frequencies to increase the chance of confusions and reduce ceiling effects, but the ΔF_0 effect sizes they obtained were not particularly large. Within their data, identification rate and effect size both varied considerably across subjects and binaural conditions, but there was no systematic relationship between the two quantities.

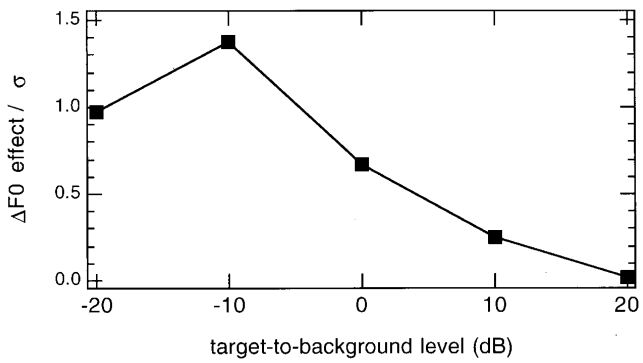

FIG. 9. Ratio between the ΔF_0 effect (difference between identification rate at ΔF_0=6% and 0%) and its standard deviation (calculated over levels of SUBJECT, SESSION, and PAIR), as a function of rms target-to-background level.

Whatever the explanation, an amplitude mismatch increases ΔF_0 effect size for the weaker vowel. Large effects are of practical interest because they are relatively easy to demonstrate with statistical confidence. However, that benefit would be lost if large effects were accompanied by an equally large variability. Figure 9 shows that this is not the case: the ratio of effect size to standard deviation (calculated over levels of SUBJECT, PAIR, and SESSION) is greater at -20 and -10 dB than at 0, 10, and 20 dB. In general, if an experiment measures small effects relative to a baseline, the identification rate for that baseline should not be too high (to avoid ceiling effects) nor too low (to avoid floor effects and subject frustration). Taking 70% as a compromise, and the ΔF_0=6% condition of our experiment as a baseline, a target amplitude of about -15 dB should be appropriate to avoid ceiling effects. This value is used by de Cheveigné *et al.* (1997) to design a sensitive test of phase and harmonicity effects. If the baseline were ΔF_0=0%, a target amplitude of 0 or -5 dB might be better. The most effective amplitude bias depends on the experiment.

C. Enhancement versus cancellation

The ΔF_0 effect was strong when the amplitude of the target vowel was -10 or -20 dB relative to the competing vowel. In contrast, the ΔF_0 effect measured by McKeown (1992) vanished beyond a 12-dB amplitude mismatch, but that may have been the result of floor effects (identification rates were overall lower). A ΔF_0 effect at low amplitudes can hardly be explained by a segregation strategy using the target's F_0 (harmonic enhancement), because that parameter is difficult to estimate when the target-to-background ratio is small. Harmonic enhancement might work when the target is strong, but it is difficult to demonstrate segregation in that case because of ceiling effects.

On the other hand the result is compatible with the hypothesis of harmonic cancellation, already supported by other experimental data (Lea, 1992; Summerfield and Culling, 1992; de Cheveigné, 1994; de Cheveigné *et al.*, 1995a, 1997). Cancellation requires estimation of the competing vowel's F_0, and this is relatively easy when the target's amplitude is low.

D. Segregation based on beats

It has been suggested that, for small ΔF_0's, segregation mechanisms might exploit beat patterns that occur when there is a difference in F_0 (Assmann and Summerfield, 1994; Culling and Darwin, 1994; Culling and Summerfield, 1995). For example, beats near the formants of a weaker vowel might reveal the presence of those formants (de Cheveigné *et al.*, 1997, Fig. 1). Beats at a vowel formant will be strong if the spectral envelopes of both vowels have the same amplitude near that formant.[2] If formant $F1$ (respectively, $F2$) determines a target vowel's identity, ΔF_0 effects should be large at a relative amplitude for which spectral envelopes of both vowels coincide in the $F1$ (respectively, $F2$) region of the target. In other words, for each vowel pair we should be able to predict the size of the ΔF_0 effect given the difference in envelope amplitude of the two vowels in the region of the formants of the target.

To test whether beats might have played a role in the current experiment, we chose two parameters representing the absolute amplitude difference between vowel envelopes at formant $F1$ (respectively, $F2$) of the target vowel. (This corresponds to the vertical distance between curves in Fig. A1 at each of the first two formants of the target.) A linear regression model was then formed, based on these two parameters, to predict the difference between identification rate at ΔF_0=0% and at ΔF_0=6%. This model was compared to a second regression model predicting simply that the ΔF_0 effect is uniformly greatest at a target amplitude of -10 dB. The two-parameter model fits *less* well than the one-parameter model (r^2=0.20 vs 0.26), despite its larger number of parameters. This model is clearly inadequate, possibly because it is too crude, possibly because beats did not determine segregation in this experiment. The question of beats is examined in further detail in a companion paper (de Cheveigné *et al.*, 1997).

E. Meddis and Hewitt's channel selection model

Meddis and Hewitt (1992) proposed a model for the identification of concurrent vowels with same or different F_0's. The model is based on an array of autocorrelation functions (ACF) calculated within individual channels output by a model of peripheral filtering and hair-cell transduction. The pattern is summed across channels to obtain a summary autocorrelation function (SACF). The largest peak in the summary autocorrelation serves to estimate the period of the ''dominant'' vowel. In the same authors' pitch perception model (Meddis and Hewitt, 1991), a similar peak served as a cue to the pitch. To account for better identification of double vowels when there is a difference in F_0, the model supposes that individual ACFs that show a peak at the period of the dominant vowel are selected and attributed to that vowel. The remaining channels are attributed to the other vowel. ACFs within each group are summed, and the sums are matched to templates to identify both vowels. When there is an amplitude mismatch between vowels, the partition between channels is likely to be determined on the basis of the periodicity of the stronger vowel, and thus the weaker vowel can be said to have been segregated according to the

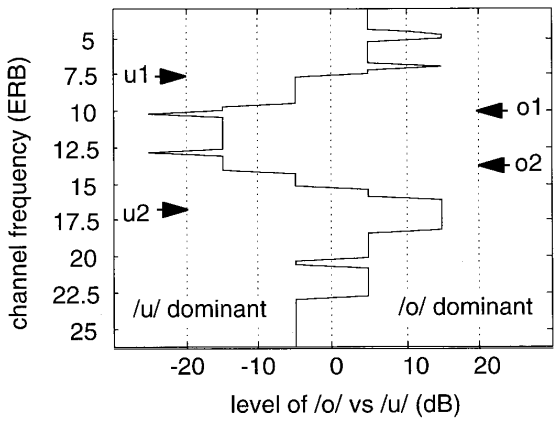

FIG. 10. Dominance of auditory-model channels by periods corresponding to the fundamental of /u/ (125 Hz), left of crooked line, and the fundamental of /o/ (132.5 Hz), right of crooked line, as a function of channel frequency in ERB (vertical scale) and amplitude of /o/ relative to /u/ (horizontal scale). Arrows indicate formant frequencies of each vowel.

principle of harmonic cancellation. In this respect the model is compatible with our findings, as discussed in Sec. III C.

We implemented Meddis and Hewitt's model using a software model of peripheral filtering and hair-cell transduction (Holdsworth *et al.*, 1988; Meddis, 1988; Culling, 1996) that produced a nerve fiber discharge firing probability within channels spaced along an ERB scale (equivalent rectangular bandwidth; Moore and Glasberg, 1983) at a density of four channels per ERB, between 80 Hz and 4 kHz. ACFs were calculated and summed to obtain a SACF from which the dominant period was determined. Each channel was assigned to one or the other vowel according to whether or not its ACF attained its highest value within 3% of the period of the dominant vowel.[3] The partition between channels at different amplitudes is illustrated in Fig. 10 for the vowel pair /o/+/u/ at ΔF_0=6%. To the left of the crooked line channels are dominated by /u/, to the right they are dominated by /o/. As either vowel is made stronger, the partition changes to its advantage, as more channels respond with that vowel's periodicity. When /o/ is 20 dB below /u/ (left-hand side) the partition isolates channels near $F1$ and $F2$ of /o/, from which that vowel might be identified. However when /o/ is 20 dB above /u/, its periodicity dominates *all* channels and leaves no channels to represent /u/. As ΔF_0 does not produce a partition, the model cannot predict the improvement with ΔF_0 that we observed for the vowel /u/ when it was 20 dB weaker than /o/ (Fig. 5, top panel, dotted lines, abscissa=20 dB). This problem occurs for 5 pairs out of 20.

Our implementation of Meddis and Hewitt's model failed to predict ΔF_0 effects for weak targets. However, it would be unwise to reject the model on that account, for at least two reasons. The first is that it might perform better with filters narrower than those used in the gammatone simulation. The second is that the auditory system might use a more sensitive criterion to detect that a channel is not completely dominated by the period of the stronger vowel. A scheme involving such a criterion is investigated in a companion paper (de Cheveigné, 1997), together with a model that allows for segregation *within* channels instead of between channels. To summarize, Meddis and Hewitt's model can explain the main aspects of our data, but not the ΔF_0 effects observed at low target-to-background levels for certain vowel pairs.

IV. CONCLUSIONS

Identification of concurrent synthetic vowels was measured as a function of amplitude difference and ΔF_0, using a task in which subjects could report one or two vowels. Main results were as follows.

(1) The number of vowels reported and the identification rate were greater at ΔF_0=6% than at ΔF_0=0%. For equal-amplitude vowel pairs, the effect of ΔF_0 on identification was larger than previously reported, presumably because effects were enhanced by the one-or-two response task.

(2) The ΔF_0 effect size was greater when the target vowel was weaker by 10 to 20 dB relative to the competing vowel. This may be explained as due either to reduced ceiling effects, or to more effective harmonic cancellation, since the competing vowel's F_0 is easier to estimate when the target is weak.

(3) The existence of a ΔF_0 effect when the target is weak relative to the competing vowel is difficult to reconcile with the hypothesis of harmonic enhancement of the target.

(4) Patterns of response for specific vowel pairs were not compatible with a simple model that assumed that identification should be better when large beats occur near formant frequencies of the target vowel.

(5) ΔF_0 effects observed at low target-to-background levels (-20 dB) for certain vowel pairs were not compatible with Meddis and Hewitt's (1992) channel selection model.

ACKNOWLEDGMENTS

The experiments were carried out at ATR Human Information Processing Laboratories, within a research collaboration agreement with the Centre National de la Recherche Scientifique. The first author thanks ATR for its kind hospitality and CNRS for leave of absence. Cécile Marin, Jean Laroche, and Steve McAdams participated in the preparation of these experiments. Hiroaki Kato and Ikuyo Masuda contributed useful ideas and advice and Rieko Kubo supervised the experiments. Thanks to John Culling of the MRC Institute of Hearing Research for providing the software for stimulus synthesis, and to him and Peter Assmann for detailed comments on previous versions of the manuscript.

APPENDIX: VOWEL SYNTHESIS

Vowels were /a/, /i/, /u/, /e/, and /o/ of Japanese. The first four formants had frequencies suggested by Hirahara and Kato (1992), the fifth was set to 4200 Hz for all vowels. The same bandwidths were used for all vowels, as in Culling and Summerfield (1995). Formant frequencies and bandwidths are shown in Table AI. Vowels were synthesized using an implementation of Klatt's synthesizer (Klatt, 1980; Culling, 1996). Table AI indicates the rms levels produced by the synthesizer (so-called ''equal effort'' levels). Waveforms were then scaled so the rms amplitudes of all vowels were the same. Table AI shows the resulting sound-pressure

TABLE AI. Frequencies and bandwidths of vowels, together with the rms amplitude produced by the synthesizer for each vowel, and the sound-pressure level at the earphone measured with an artificial ear (after rms amplitudes had been made equal).

	/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
dB rms after synthesis	46.9	40.6	40.4	41.8	44.5	
dB (A) SPL	70.0	63.0	63.6	67.4	66.2	

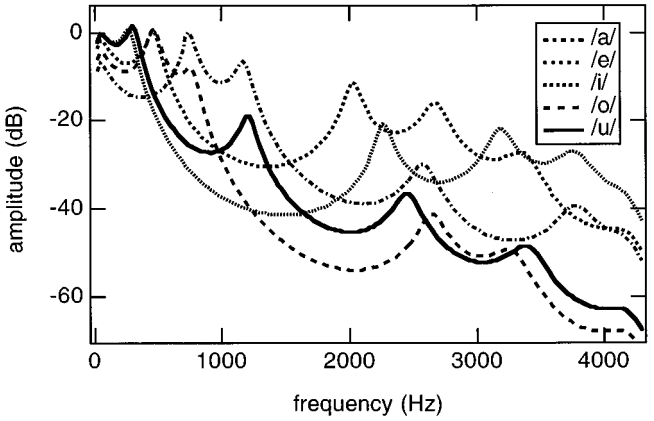

FIG. A1. Spectral envelopes of all five vowels.

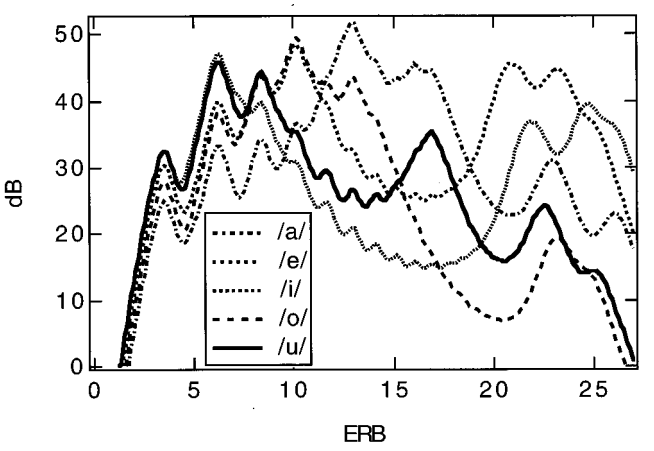

FIG. A2. Estimates of basilar-membrane excitation as a function of channel frequency (on an ERB scale) for all five vowels (fundamental is 125 Hz).

levels delivered by the earphones, as measured by the artificial ear. Spectral envelopes (scaled by the same amount as the waveforms) are plotted in Fig. A1. Estimates of basilar membrane excitation patterns for each vowel are plotted in Fig. A2. Excitation patterns were calculated by taking the FFT of a 16-ms Hanning-shaped window of a 125-Hz vowel (two periods), and applying spectral smoothing according to formulas of Moore and Glasberg (1983).

[1]The ''x'' answer was counted as if the subject had reported an incorrect vowel. The option was rarely used by the subjects, and in subsequent experiments it was removed.

[2]In the low-frequency region where partials are resolved, the amplitude of beats of adjacent partials depends upon the levels of their excitation patterns within each channel. Excitation may locally be equal even if spectral envelopes do not coincide. However, given the small ΔF_0, excitation patterns of the lowest partials tend to overlap, and the range of intervowel levels over which they may coincide is therefore narrow and near the intervowel level for which spectral envelopes coincide. At higher frequencies where partials are not resolved, excitation patterns are wide and best beats also occur when the spectral envelopes coincide.

[3]This selection criterion is a departure from Meddis and Hewitt's model, in the interest of clarity. In his model, the dominant period was derived from the largest peak in the summary ACF (within a certain range), and individual channels were classified according to whether their ACF had a ''peak'' at the dominant period. A peak was defined as any point higher than its left and right neighbors. This definition can lead to somewhat erratic results when ACF's are ''noisy.'' Our modification produced patterns that are better suited for illustration purposes. It does not betray the spirit of the model or reduce its chance of success.

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.

Culling, J. (**1996**). ''Signal processing software for teaching and research in psycholingustics 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.

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

Culling, J. F., Summerfield, Q., and Marshall, D.-H. (**1994**). ''Effects of simulated reverberation on the use of binaural cues and fundamental frequency differences for separating concurrent vowels,'' Speech Commun. **14**, 71–95.

de Cheveigné, A. (**1994**). ''Strategies for voice separation based on harmonicity,'' Proceedings of the International Conference on Speech and Language Processing, Yokohama (Acoustical Society of Japan), pp. 1071–1074.

de Cheveigné, A. (**1995**). ''Experiments in vowel segregation,'' ATR Human Information Processing Research Labs Tech. Report No. TR-H-154.

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

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

de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (**1995b**). ''Sensitive experimental techniques for the study of sound segregation,'' ASJ autumn meeting (Acoustical Society of Japan), pp. 373–374.

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

Demany, L., and Semal, C. (**1990**). ''The effect of vibrato on the recognition of masked vowels,'' Percept. Psychophys. **48,** 436–444.
Hirahara, T., and Kato, H. (**1992**). ''The effect of *F*0 on vowel identification,'' in *Speech Perception, Production and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Ohmsha, Tokyo), pp. 89–112.
Holdsworth, J., Nimmo-Smith, I., Patterson, R. D., and Rice, P. (**1988**). ''Implementing a GammaTone filter bank (SVOS final report, annex C),'' MRC Applied Psychology Unit Tech. Report.
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.
Klatt, D. H. (**1980**). ''Software for a cascade/parallel formant synthesizer,'' J. Acoust. Soc. Am. **67,** 838–844.
Lea, A. (**1992**). ''Auditory models of vowel perception,'' Ph.D. dissertation, Nottingham (unpublished).
McKeown, J. D. (**1992**). ''Perception of concurrent vowels: The effect of varying their relative level,'' Speech Commun. **11**, 1–13.
Meddis, R. (**1988**). ''Simulation of auditory-neural transuction: further studies,'' J. Acoust. Soc. Am. **83,** 1056–1063.
Meddis, R., and Hewitt, M. J. (**1991**). ''Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification,'' J. Acoust. Soc. Am. **89,** 2866–2882.
Meddis, R., 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 Glasberg, B. R. (**1983**). ''Suggested formulae for calculating auditory-filter bandwidths and excitation patterns,'' J. Acoust. Soc. Am. **74,** 750–753.
Scheffers, M. T. M. (**1983**). ''Sifting vowels,'' Ph.D. thesis, Gröningen (unpublished).
Shackleton, T. M., Meddis, R., and Hewitt, M. J. (**1994**). ''The role of binaural and fundamental frequency difference cues in the identification of concurrently presented vowels.'' Q. J. Exp. Psychol. A **47**, 545–563.
Summerfield, Q. (**1992**). ''Roles of harmonicity and coherent frequency modulation in auditory grouping,'' in *The auditory processing of speech: from sounds to words*, edited by M. E. H. Schouten (Mounton de Gruyter, Berlin), pp. 157–166.
Summerfield, Q., and Assmann, P. F. (**1991**). ''Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony,'' J. Acoust. Soc. Am. **89,** 1364–1377.
Summerfield, Q., and Culling, J. F. (**1992**). ''Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency,'' J. Acoust. Soc. Am. **92,** 2317(A).