Ten experiments on vowel segregation

Alain de Cheveigne

last revised: 16 May 97

ABSTRACT

This report describes ten experiments on concurrent vowel segregation and identification. Experiments are numbered from 1 to 10.

Experiment 1 was designed to be sensitive to a variety of hypothetical mechanisms by which frequency modulation (FM) might affect identification. The results were mostly negative, in the sense that no effect was found that could not be attributed to other factors. The only "FM effect" observed was that identification was better for incoherent than for coherent modulation. However this effect was small, and one cannot rule out that it was caused by unavoidable differences in the pattern of instantaneous ΔF_0 between FM conditions.

Experiment 2 explored the identification of 3 concurrent vowels. As for in the case of 2 concurrent vowels, a difference in F_0 between vowels aided identification.

Experiment 3 explored the effects of ΔF_0 and amplitude differences between vowels over a relatively wide range. Presence of a ΔF_0 helped identification when the target/competitor amplitude ratio was low (down to -25 dB). The effect disappeared at -35 dB. In general identification was better at 3 % than at 0 %, but there was little difference between ΔF_0 = 3 %, 6 % or 12 %. One might have expected larger ΔF_0s to be more effective at low target amplitudes. Such was not the case.

Experiment 4 explored the region of very small ΔF_0s, while controlling for phase effects and beats. As it turned out, the smallest ΔF_0 used, 0.375 %, was sufficient to cause segregation. This did not seem to be the consequence of beat patterns caused by the ΔF_0.

Experiment 5 explored ΔF_0 effects at short durations (125 and 62.5 ms), while again controlling for phase effects. ΔF_0 effects were somewhat weaker at 62.5 and at 125 ms than at 250 ms, but they were still quite large and significant.

Experiment 6 attempted to find evidence for harmonic enhancement. Double-vowel stimuli were divided into two short pulses separated by a silence. The F_0s of the target and competitor shared the same value in the first pulse, and one, the other or both could differ from this value by 6 % in the second pulse. It was expected that a jump in target F_0 might impair harmonic enhancement and reduce the identification rate. No such effect was found.

Experiment 7 reproduced the 3-vowel experiment with a 3-vowel forced response task, instead of the 1,2 or 3 response task of Exp. 2 . The 3-vowel forced response task is less affected by "multiplicity" cues. A comparison between Exp. 2 and Exp. 7 allows other cues to be factored out, so the role of multiplicity cues can be assessed.

Experiment 8 was an extension of Exp. 4 to even smaller ΔF_0s (0.1 and 0.2 %). An additional intervowel phase relation (antiphase) was also included. As in Exp. 4, effects

of ΔF_0 were observed at 0.8 and 0.4 % (equivalent to the 0.75 and 0.375 % conditions of Exp. 4) but not at 0.2 or 0.1 %. Phase (same phase vs antiphase) had little effect at 0.2, 0.4 or 0.8 %. It had some effect at 0 and 0.1 %.

Experiment 9 investigated the effect of formant bandwidth on segregation. Formant bandwidth is known to have surprisingly little effect on vowel identification, but it affects the "peakiness" of the spectrum and so is likely to affect the way a vowel's features emerge from the spectrum of a concurrent vowel pair. Such was indeed the case: in general a vowel was much better identified if its formant bandwidths were narrower than normal (by a factor of 2), rather than wider than normal (by a factor of 2). Somewhat unexpectedly, identification was better if the interfering vowel had wide bandwiths rather than narrow. Narrowing the formant bandwidths of a vowel has effects similar to raising its RMS amplitude.

Experiment 10 attempted to find evidence of harmonic enhancement (improved identification based on the harmonic structure of a target) by measuring identification of static or frequency modulated diphthongs (sequential vowel pairs) that were partially masked by a noise (harmonic or inharmonic) with a vowel-like spectrum. Enhancement was expected to cause better identification of targets with a static F_0. No such effect was observed.

1 Introduction

A previous series of experiments carried out at ATR (de Cheveigné 1995, 1996, 1997, de Cheveigné et al. 1995, 1997a,b) used the concurrent vowel identification paradigm to study effects of ΔF_0, amplitude differences between vowels, and phase. The classic paradigm was modified in several ways: a) a systematic intervowel level mismatch was introduced to avoid ceiling effects, b) the task allowed one or two responses per stimulus (rather than two-vowel in the classic task), thus making it sensitive to cues that signal the multiplicity of sources within a stimulus, c) vowels within the stimulus were scored separately rather than together, resulting in *constituent correct* scores sensitive to factors that affect constituent vowels asymetrically.

A motivation in designing these experiments was the hope that the modified paradigm might bring new insight to issues for which experimentation has so far yielded puzzling or unconclusive results. Such is the case of FM (frequency modulation). We also wished to investigate new issues (such as the perception of more than two concurrent sources), and the effects of certain parameters (amplitude ratio, ΔF_0) in ranges hitherto unexplored.

Some of the ten experiments are related to each other, some are not. Some are relatively straightforward. Others involve complex issues, a complex design, or a complex pattern of results. They are not necessarily described in the clearest possible fashion in this report (hopefully they will be in future papers). The reader is warned of the risk of "information overload".

After an initial section on methods common to all experiments, each experiment is described in more or less self-contained fashion. The details of the statistical analyses are all grouped in an appendix at the end of the report. No F- or p-levels are reported in the body of the report.

2 General Methods

2.1 Subjects

There were four subsets of subjects:

1. 5 Japanese subjects (ATR staff) each performed a session of Experiment 1.
2. 15 Japanese subjects (students, 7 male and 8 female, aged 18 to 22 years, paid for their services) each performed a session of Experiments 1, 2, 3, 4, 5, 7, 8, and 10. They also performed preliminary versions of Experiments 6 and 9.
3. 15 Japanese subjects (students, 8 male and 7 female, aged 18 to 22 years, paid for their services) each performed a session of Experiments 2, 6, 7 and 9.
4. 12 French subjects each performed a session of Experiment 1 (using stimuli based on French vowels, but otherwise equivalent to those used for Japanese subjects).

In summary, Experiment 1 (on FM) was run with a total of 32 subjects (20 Japanese, 12 French). Experiments 2 and 7 (on triple vowels) were run with 30 subjects. The other experiments were run with 15 subjects.

2.2 Stimuli

Stimuli were either single, double or in some cases triple vowels. Vowels were synthetic tokens of the five vowels /a/, /i/, /u/, /e/, /o/ of Japanese (or, for the 12 french subjects in Exp. 1, of French). Formant frequencies and bandwidths for Japanese are listed in de Cheveigné et al. (1997a, Table I), and for both languages in de Cheveigné and Marin (1996). Vowel tokens were obtained by additive synthesis using a software implementation of Klatt's synthesizer (Klatt, 1980; Culling, 1996) at 20 kHz with double floating point precision. They were 270 ms in duration, with onsets and offsets shaped by 20 ms raised-cosine ramps, leading to an "effective" duration of 250 ms between -6dB points. Starting phases were set to a "random" pattern that was the same for all conditions and experiments (pattern "R" of de Cheveigne et al. 1997b). Fundamental frequencies (F_0) ranged between 124 and 140 Hz. F_0s were static for all experiments except Experiments. 1, 6 and 10, where they could be either static or modulated. Vowels were always harmonic, except in Exp. 1 where they were either harmonic or inharmonic.

After synthesis, all single vowels were scaled to a standard rms value and stored on disk in single precision floating point format. During the experiments, double and triple vowels were created "on the fly" by adding single vowels (eventually with a level mismatch), and setting the sum to a standard rms value. Stimuli (single, double or triple vowels) were converted to 16 bit integer format and output diotically to earphones from the NeXT. The gain was adjusted so that the sound pressure level was in the range 63-70 dB(A), as measured by a Bruel& Kjaer artificial ear (sound level meter type 2231, half-inch microphone type 4134, flat coupler plate).

2.3 Experiment design and task

In most cases, the experiments used the conventional concurrent vowel identification paradigm (Scheffers, 1983; Assmann and Summerfield, 1990; Culling and Darwin 1993), with the following three modifications (de Cheveigné et al. 1997a,b):

- Each stimulus was scored as many times as it contained vowels. Single vowels were scored once, double vowels twice, and triple vowels three times. When scoring a stimulus, each vowel in turn was nominated the "target". The target was deemed identified if its name was among the vowels reported by the subject for that stimulus. This outcome was recorded according to the target's nature, and the nature of the other vowel or vowels that were mixed with it ("competitors"). Roles of target and competitor(s) were then exchanged, leading to as many scores per stimulus as there were vowels. This procedure yielded "constituent-correct" rates, rather than the more commonly used "combination-correct" rates that count trials for which all vowels are correctly identified. Constituent-correct rates are possibly a more sensitive measure. For example, an effect might affect each vowel of a pair in a different direction, leading to a null effect in terms of combination-correct score. The constituent-correct score remains sensitive in this case. Effects of asymmetric configurations (for example a modulated vowel with an unmodulated competitor, etc.) may also be investigated in detail.

- For all experiments except 7 and 10, subjects were allowed to report a variable number of vowels on each trial. In general the stimulus set included stimuli made up of a variable number of vowels, and subjects were informed of that fact. This is typical of natural situations where the number of sources to attend to is not known a priori. The number of vowels reported is a measure sensitive to "multiplicity" cues.

  The stimulus set of Exp. 2 contained single, double and triple vowels, and the subjects could answer 1, 2 or 3 vowels for each stimulus. The stimulus set of Experiment 7 contained only triple vowels, and subjects had to report three vowels. Stimuli of experiment 10 were partially masked diphthongs, and the subject had to report both vowels in each diphthong.

- In general, an amplitude mismatch was introduced between vowels to enhance the sensitivity of identification of the weaker vowel to conditions of interest. This is typical of natural situations in which competing voices rarely share the same level. Identification scores for the stronger vowel were usually perfect and were ignored.

Subjects were seated in a sound treated booth or room, in front of a computer screen that gave prompts and instructions, and they responded by means of a keyboard.

2.4 Data analysis

Data were analyzed using repeated-measures ANOVA. This analysis compares the average effect (main effects or interactions) with the variability of the effect between subjects. It is thus insensitive to effects that are specific only to certain subjects.

Each experiment was designed as a combination of simpler experiments sharing certain conditions. This was to avoid repetition of common conditions in the interest of economy. Analysis of variance was performed on subsets of the data. There was usually some overlap between subsets.

Each experiment involved a rather large number of tests, and with 10 experiments in all, the probability of a false positive is rather large. We treat the F- and p-levels produced by the ANOVAs and contrasts as descriptive quantities, and we do not attempt to apply corrective factors. We leave it to the reader to adjust his or her confidence in the significance of these results. Caveat lector!

3 Experiment 1: FM

3.1 Motivation

Frequency modulation (FM) has been cited as a prime example of "common fate" in Auditory Scene Analysis (McAdams, 1984; Bregman, 1990). Partials that are coherently modulated "move together" accross the spectrum, and should stand out if the background is static, or is modulated incoherently with the target. In a striking experiment cited by McAdams (1984), vibrato made a sound "pop out" from a harmonic background of same periodicity.

Careful experiments have shown that the effect of FM can usually be explained by the instantaneous ΔF_0 that is induced by the modulation (Summerfield, 1992; Carlyon, 1991; Culling, Summerfield and Marshall, 1994; Culling and Summerfield, 1995; Marin and McAdams 1991). However it is difficult to accept there are no effects specific to modulation per se, because:

- The "common fate" model is appealing.
- FM (vibrato) is commonly used by musicians, and it is logical to guess that its role to enhance segregation of the part played by the musician from the musical background.
- It is conceivable that previous experiments failed to reveal the effects of FM because of lack of sensitivity, because other effects (such as harmonicity) were overwhelming, or because the experimental context somehow forced subjects to ignore FM-based cues that would nevertheless be used in everyday life.
- Certain genuine FM effects have been demonstrated (see de Cheveigné and Marin 1996 for a discussion).

Experiment 1 was designed to test for a wide range of imaginable FM-based mechanisms, using a relatively sensitive paradigm. In this way we hoped to reveal effects that had previously escaped detection. Failing that, and supposing that our efforts were convincing, our lack of success would be evidence that the hypothesized effects do not exist. We would then be relatively confident that the hypotheses that predicted them were false. According to this logic, we give their chance to some hypotheses that are a priori unlikely, given current knowledge, to make sure that "no stones were left unturned". The reader is warned that the set of hypotheses is rather heteroclite.

As it turned out, most hypotheses were not supported by the data. In order for this negative result to have some value, we must explain in detail why each mechanism *might* have been effective, and how the experiment was designed to be sensitive to it.

Hypotheses were:

- *Modulation of the target affects its identification*. For example, suppose that segregation of targets occurs according to a mechanism that sensitive to their periodicity[1]. If modulation *aids* F_0 estimation, then identification should be improved. If it *hinders* F_0 estimation, then identification should instead be impaired. In both cases we expect an effect specific to target modulation.
- *Modulation of the competing vowel affects identification of the target*. There is strong evidence that segregation occurs according to a mechanism of harmonic

[1] Harmonic segregation hypothesis - So far we have no evidence that this hypothesis is true.

cancellation that suppresses the competing vowel. Granted this, and supposing that modulation of the competing vowel aids estimation of its F_0, identification should be enhanced by modulation. If instead F_0 estimation is hindered by the modulation, identification should be impaired. In both cases we expect identification of a target vowel to be affected by modulation of the vowel that competes with it.

- *Identification depends on the shape of modulation ("n" vs "u").* Demany and Clément (1995) demonstrated that pitch discrimination of peaks in modulation is better than that of dips (at least for wide modulation amplitudes). It is conceivable that a similar asymmetry might affect the F_0 estimation step that is required for segregation. Identification might thus depend on the shape of target modulation (in the hypothesis of harmonic enhancement), or the shape of ground modulation (in the hypothesis of harmonic cancellation).

- *Modulation might affect the number of vowels reported* and thus indirectly identification. For example modulation might make the stimulus more inharmonic (because FM produces sidebands) and increase the number of sources perceived. Or it might instead enhance the "cohesion" of the stimulus, and decrease the number of sources perceived. This latter effect, supposing it exists, might be stronger for inharmonic vowels that "lack cohesion", and therefore tend to evoke the perception of multiple sounds.

- Auditory Scene Analysis theory (Bregman 1990) leads us to expect that *incoherent* modulation of target and competitor might provoke the perception of more sources, and thus indirectly lead to better identification.

- Results obtained by Summerfield (1992) for pairs of inharmonic vowels lead us to expect that identification might be better for a *modulated* target on an *static* ground, rather than, either a static target (whatever the ground), or a modulated target on a modulated ground (whether the ground is modulated coherently or incoherently with the target).

The stimulus set was designed to test all of these hypotheses, plus any additional unforseen mechanisms that they might trigger. We were thus relatively confident at the outstart in our chances of finding real FM effects.

3.2 Stimuli

The stimulus set contained both single and double vowels. Single vowels were synthesized at average F_0s of 124, 128 and 132 Hz. Vowels could thus be paired with ΔF_0s of 0 and approximately 3 and 6 %. F_0 was either constant, or else modulated with a single cycle of a cosine-shaped modulator. Peak modulation was 3 % (approximately 4 Hz), and the modulation rate was 4 Hz. Modulator phase could be either cosine, denoted as "u", or its opposite, denoted as "n" (unmodulated vowels are denoted as "_"). Vowels were either harmonic or inharmonic. Inharmonic vowels were obtained by randomly shifting each partial frequency of a harmonic vowel by -3, 0 or 3 % (approximately 4 Hz). This "random" pattern of partial frequencies was the same for all repetitions of all inharmonic vowels, whatever the F_0.

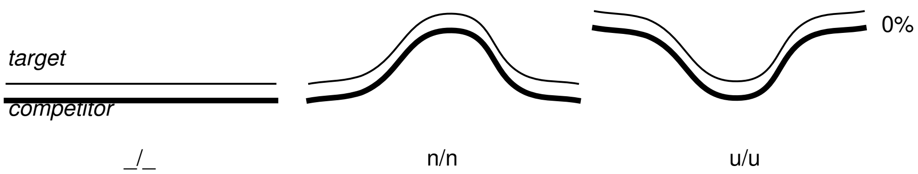
This choice of F_0s, inharmonicity pattern, and FM rate ensured that all components (including those induced by the FM) were multiples of 4 Hz, inverse of the duration of the stimulus (250 ms between -6 dB points). The stimulus was thus actually periodic with a period equal to the stimulus duration. This allows a better control of the effect of starting phase on the long-term spectrum magnitude[2].

Single vowel conditions were all 18 combinations of (F_0 = 124, 128, 132 Hz) x (harmonicity = I, H) x (modulation shape = _,n,u). Each was repeated twice for each of the 5 vowels, leading to a total of 180 single vowels within the stimulus set.

Double vowels were formed by adding single vowels with a level mismatch of 15 dB. One vowel (the "target") was thus weaker than the other by 15 dB. Identification of the stronger vowel ("the background") tended to be perfect and therefore uninteresting. We measured only identification of the weaker vowel.

Ignoring the order of F_0 (low vs hi), there are a priori 108 possible double vowel conditions: (3 ΔF_0s) x (2 target harmonicities) x (2 ground harmonicities) x (3 target modulation shapes) x (3 background modulation shapes). This set is too large to be practical. We therefore selected a subset of 16 conditions that are sufficient to test our hypotheses. These conditions were (notation X/Y indicates that X is the state of the target or weaker vowel, and Y that the stronger or competing vowel):

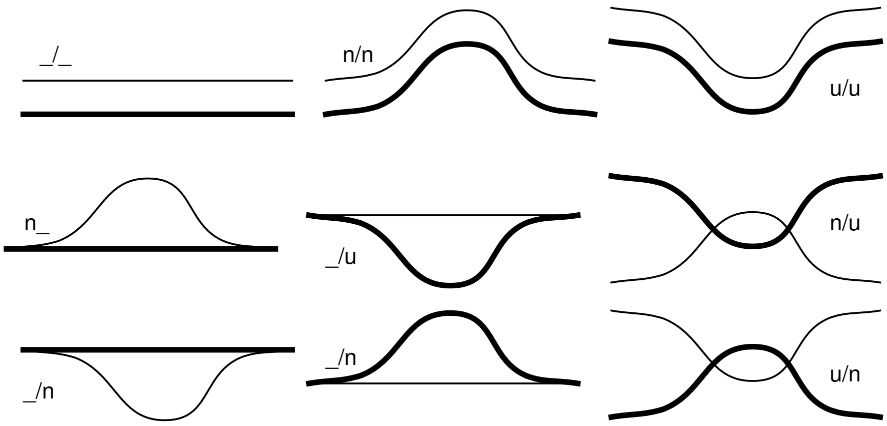
- (harmonicity = H/H) x (ΔF_0 = 0) x (modulation = _/_, n/n, u/u). The instantaneous ΔF_0 is everywhere 0.



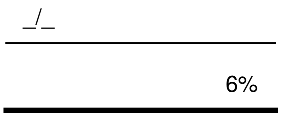
- (harmonicity = H/H) x (ΔF_0 = 3%) x (modulation = _/_, n/n, u/u, n/u, u/n, n/_, u/_, _/n, _/u). All combinations of target and competitor modulation are included. For conditions _/_, n/n and u/u, the instantaneous ΔF_0 is constant and equal to 3%. For conditions (_/n, _/u, n/_, u/_) it is variable but equal to 3% on average (it starts at zero, peaks at 6%, and ends at zero). For conditions (n/u, u/n), the F_0

[2]The long-term spectrum magnitude was independant of starting phase of all partials that did not coincide. It depended on the phase of partials that did coincide. At ΔF_0 = 0 with coherent modulation, all partials coincided. Phases being the same for all vowels, partials added up in phase, and so the spectrum was independent of starting phase. At $\Delta F_0 \neq 0$, for static vowels, no partials coincided, so the spectrum was again independent of starting phase. However, in the case of modulated vowels, each partial was "split" into a series of partials spaced by 4 Hz, and the compound spectrum thus depended on starting phases in a way that is unfortunately difficult to predict or control.

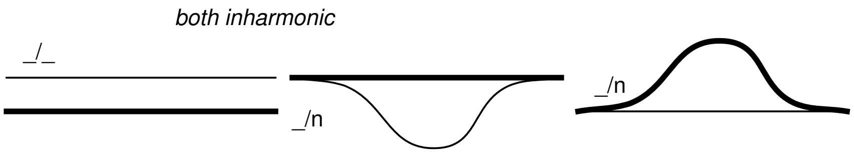
tracks cross and the average ΔF_0 is somewhat larger than 3 %.



- (harmonicity = H/H) x (ΔF_0 = 6%) x (modulation = _/_).



- (harmonicity = I/I) x (ΔF_0 = 3%) x (modulation = _/_, _/n, n/_).



These 16 conditions were crossed with 20 ordered vowel pairs (vowels within a pair distinct). They were also crossed with two values of absolute F_0 or F_0 order. From previous experiments we expected no effect of this factor, but we wished to avoid the possibility that a subject might associate a particular condition with a particular F_0. An exception to this rule was made for conditions H3_/n and H3n/_: the *unmodulated* vowel (_) always had the *lower* F_0, so the F_0s of both vowels started and stopped at the same value. Likewise, in conditions H3u/_ and H3_/u the unmodulated vowel had the higher F_0, again so that the F_0 tracks of both vowels started and stopped at the same value[3]

The stimulus set thus comprised (16 interesting conditions) x (2 F_0s) x (20 vowel pairs) for a total of 640 double vowels. These were mixed randomly with the 180 single vowels for a total of 820 stimuli. These were presented in a single session that typically took 40 to 90 minutes to complete.

[3]This convention was intended to make it easier to interpret possible differences between "u" and "n"-shaped modulation. A "n"-shaped modulator has one maximum, but also two "half" minima, at onset and offset. F_0 estimation at these points might might also depend on modulation shape, in an opposite way from the central part. However, given our choice of modulation patterns, the ΔF_0 at onset and offset is zero, so accuracy of F_0 estimation at these points cannot affect identification. Shape effects, if they exist, are thus limited to the central part of the modulator waveform.

3.3 Results

Results reported here include data for subject subsets 1,2, and 4. The design and stimulus set was identical in all cases, except that French subjects heared stimuli based on French vowels. Graphs display results for both the entire subject set, and for subset 2 to allow comparison with Experiments 2-10.

Scores were averaged over F_0s (because we expect no significant effect), and over vowels or vowel pairs (because we are not interested in this effect). Details of ANOVA and contrasts are given in Appendix A.

3.4 Effect of ΔF_0

As expected from previous experiments, ΔF_0 had a strong effect on both the *number of vowels reported*, and the *identification rate* of the weaker vowel (-15 dB). Fig. 1 shows the number of vowels reported (a) and the identification rate (b) as a function of ΔF_0 for static F_0 conditions. Data for subset 2 are plotted as dotted lines. Data of a previous experiment, also at -15 dB but with different subjects, are plotted as dot-dash lines (de Cheveigné et al. 1977b). The difference between 3 % and 6 % is significant for both measures.

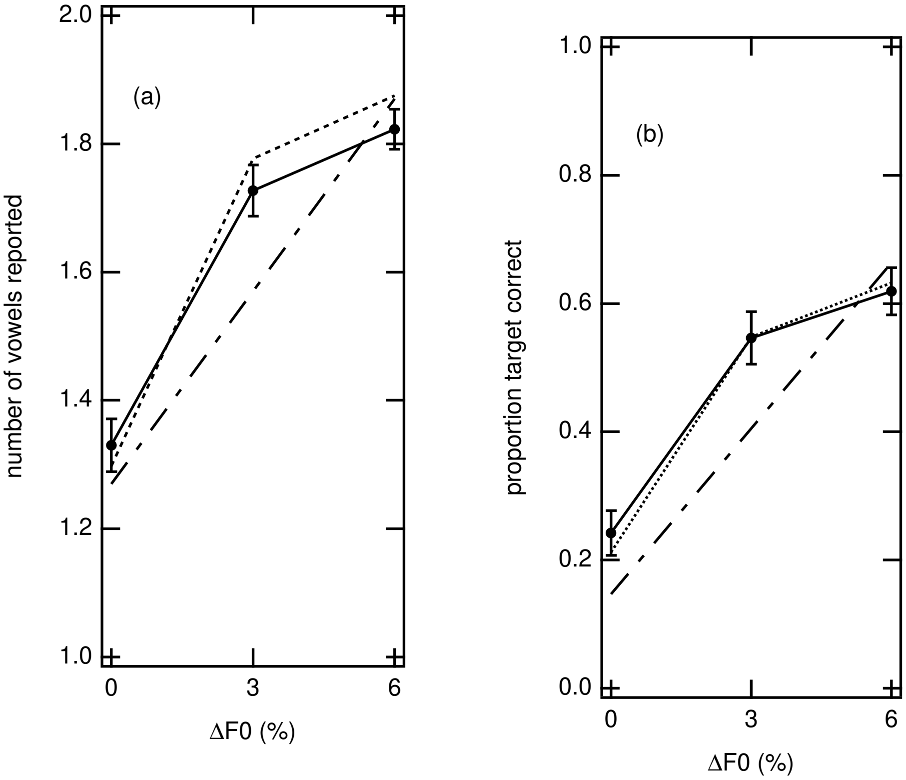


Fig. 1 *(a) Average number of vowels reported per stimulus as a function of ΔF_0, averaged over all 32 subjects (subsets 1, 2, 4). Dotted line represents data for subset 2. Dot-dash lines are data obtained in a previous experiment with the same stimuli but different subjects (de Cheveigné et al. 1997b). Error bars represent one standard error. (b) Target-correct identification rate as a function of ΔF_0.*

3.4.1 Effects of harmonicity

Inharmonic stimuli were included in the stimulus set mainly to test the hypothesis that FM might counteract their lack of "coherence", leading to an effect of modulation on the number of‘ vowels reported. Harmonicity per se was not our major concern, but we nevertheless report the effects observed.

Harmonicity did not affect identification of *single* vowels, whatever their modulation state. It did however increase the number of vowels reported: 1.42 for inharmonic vowels vs 1.11 for harmonic vowels (Sect. A.1).

For double vowels (nominal ΔF_0 of 3 %) harmonicity had the opposite effect. Subjects reported on average 1.60 vowels per pair of inharmonic vowels, vs 1.76 per pair of harmonic vowels (at the same nominal ΔF_0). Harmonicity also affected identification: 57.7 % for harmonic and 33.5 % for inharmonic vowels. From past results, we can attribute this effect to the fact that harmonic cancellation is less effective if the competitor is inharmonic.

3.5 FM effects

3.5.1 coherent FM

The presence of coherent FM of both vowels had no effect, either on the number of vowels reported, or on the identification rate. There was no difference between conditions _/_, n/n, u/u. This was true both at ΔF_0 = 0 and ΔF_0 = 3 %.

3.5.2 Shape of FM ("n" vs "u")

The *shape* of modulation had no effect. It made no difference whether modulation had the shape of a peak ("n") or a valley ("u"). This was true whether the target was modulated (n/_ vs u/_), the ground (_/n vs _/u) or both (n/n vs u/u), (n/u vs u/n).

3.5.3 Conditions (n/u, u/n) versus others

The F_0 tracks for conditions (n/u, u/n) cross, implying that the average instantaneous ΔF_0 is larger than for other conditions. Indeed, they evoked more answers, and produced greater identification rates than the other conditions with a nominal ΔF_0 of 3%. This difference is not particularly informative, so we won’t mention these conditions again in the following.

3.5.4 Target FM

We evoked the hypothesis that FM of the target vowel might determine segregation. Interaction between target and ground modulation was highly significant, and it is best to ignore the significant main effect of target FM on the number of vowels reported and identification rate and concentrate on simple effects.

When the ground was static, target FM had a highly significant effect (Fig. 2, full symbols). The same was true when the ground was modulated (Fig. 2, open symbols). However these effects were opposite in sign. The data does not reflect a mechanism sensitive to target modulation per se.

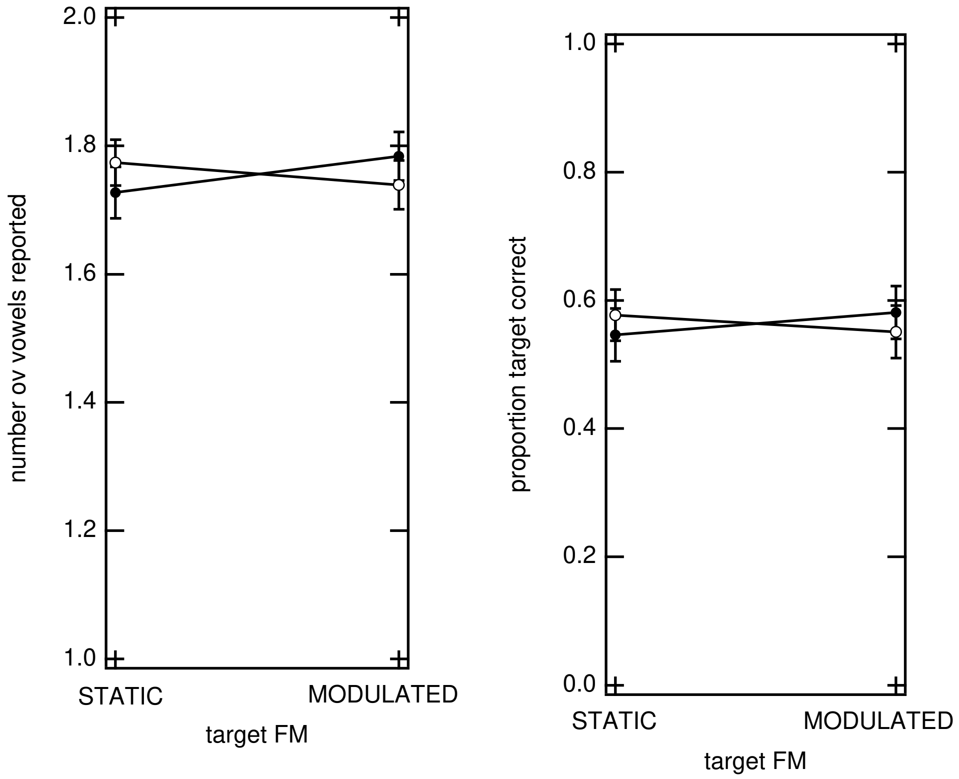


Fig. 2 *Left: average number of vowels reported per stimulus as a function of target modulation, averaged over all 32 subjects. Full symbols: static ground, open symbols: modulated ground. Error bars represent one standard error. Right: target-correct identification rate.*

3.5.5 Ground FM

We also evoked the hypothesis that identification of the target might be affected by modulation of the competing vowel. The main effect of ground was not significant. When the target was static (Fig. 2, left), ground FM had a significant effect, and the same was true when the target was modulated (Fig. 2, right). However these effects were opposite in sign . The data cannot be described as reflecting a ground modulation-specific mechanism.

3.5.6 Modulated target on static ground

Another hypothesis was that modulated targets with static competitors might be easier to identify than either static targets, or targets - modulation indifferent - with modulated competitors. Summerfield (1992) found evidence of this effect for inharmonic but not harmonic vowel pairs. We found no evidence of the effect for either harmonic or inharmonic vowels. The effect would have shown up as an asymmetry between n/_ and _/n, which we did not observe.

3.6 Coherence of FM

A final hypothesis was that *coherence* of FM might determine segregation. Conditions for which target and ground were modulated incoherently (more precisely: one modulated, the other static) evoked significantly more responses (1.78 vs 1.74), and gave a significantly higher identification score (58.9 % vs 54.8 %) than conditions for which both vowels were modulated coherently, or both were static.

The data can thus be described as reflecting an effect of FM coherence. However this effect is small, especially when compared to ΔF_0 or harmonicity effects. It is concievable that it is a consequence of the greater peak instantaneous ΔF_0 (6 %) observed in incoherently modulated conditions.

That interpretation would have been eliminated if identification rate or number of vowels were greater than for a static ΔF_0 of 6 %. Such was not the case: scores for incoherently modulated pairs fell between those for static vowel pairs at ΔF_0 = 3 % (1.73, 54.6 %) and ΔF_0 = 6 % (1.82 and 61.9 %). We thus cannot guarantee that this "FM" effect is really the result of modulation per se.

3.7 Conclusion

Despite our efforts to tap a wide variety of FM mechanisms, and despite the sensitivity of the experiment, we found no evidence that FM per se affects segregation, apart from an effect of FM coherence. This effect was small in comparison to effects of ΔF_0 and harmonicity, and we cannot exclude that it was caused by the difference in maximum instantaneous ΔF_0 between coherently and incoherently modulated conditions.

4 Experiment 2: Segregation of 3 concurrent vowels

4.1 Introduction

[Warning: this experiment is complex and difficult to understand. It is best to read about Exp. 7 first, and then come back to Exp. 2]

Concurrent vowel segregation experiments usually involve two vowels. This is a reasonable restriction, to keep the experiments simple, but it leaves open the question of how the auditory system deals with the very common situation of more than two sources present in the environment.

One thesis is that the auditory system prepares an internal representation, usually some form of spectrotemporal map, within which correlates of each source are separated. What the subject actually hears depends on the result of a sort of "perceptual shopping" within this representation. This is implicit in computational auditory scene analysis (CASA) models such as those of Mellinger (1991), Cooke (1991), Brown (1992) or Ellis (1996). A priori,the representation might accomodate an arbitrary number of sources. One can thus imagine perceiving three or more vowels at a time, given sufficient segregation cues such as F0 differences.

Another thesis is that there are never more than two entities involved: a "target", and whatever else is in the auditory environment (the "background", or "competing sounds"). When a sound is the object of attention, the auditory system singles out its correlates, and those of all competing sounds are lumped together to form a "background". It is conceivable that the auditory system might maintain a number of "target-ground" parses at a low level. At any moment the auditory system would choose among these dichotomies. The main difference with the previous thesis is that the auditory system would never manipulate more than two entities at a time. At each instant one would hear one vowel among the three, together with an undifferentiated background.

A third thesis is that the auditory system proceeds by suppressing each sound in turn. The difference with the first thesis is mainly one of emphasis: the first thesis concentrated on the target(s), this one concentrates on each sound considered as an interferer. The success of the previous schemes presumably depended on target characteristics (such as periodicity). The success of this scheme depends on characteristics of the interfering sound(s), according to whether or not they are easy to suppress. Again, one can conceive that the auditory system might maintain a number of parses at a low level, each the result of cancelling the correlates of one or more sources.

The "suppression" thesis is partially supported by experimental results that suggest that F_0-guided segregation is the result of *harmonic cancellation* of competing sounds. Targets are easier to hear if the interference is periodic than if it is not. This is why identification of two concurrent vowels is better when each has a different F_0. However when there are *three* concurrent vowels, each vowel is in competition with *two* other vowels. One can consider these two vowels as a single undifferentiated competitor (thesis 2). In that case it should be hard to suppress if the two competing vowels have different F_0s, because their sum is not periodic. However one can also consider the vowels as periodic competitors that can be removed one after the other (thesis 3). In that case identification of the target vowel should be relatively easy.

A related question pertains to the *number of sources heard*. Where subjects are free to report one vowel or two for stimuli containing one or two vowels, they tend to report two if a) the stimulus is inharmonic (vowels have different F_0s), or b) the stimulus contains two different vowels with similar amplitudes. They report only one vowel if a) the stimulus is harmonic (a single vowel or two vowels with same F_0), or b) its spectral envelope is close to that of a single vowel (single vowel, or vowels mixed with a large amplitude difference). What happens if the subjects are free to report one, two or *three* vowels, and the stimulus contains one, two or three vowels, with one, two or three different F_0s? Presumably, when there is one vowel and one F_0, they will tend to report a single vowel. With multiple vowels and multiple F_0s, they will report several vowels more often, but what is the actual pattern? Do they report more vowels with three vowels than two? More with three F_0s than two?

Part of the answer comes from an experiment by Kashino and Hirahara (1995). Subjects listening to the concurrent speech of several speakers accurately reported their number when this number was one or two. As the number of speakers increased to 11, the number of voices reported increased much more slowly, with an asymptote at 4. This asymptote might indicate the maximum number of sources that can be resolved. However it might also be related to the imperfect periodicity of speech, or to a cognitive limit on the number of voices that can be kept track of at higher levels.

The aim of Experiment 2 was to explore this question using stimuli containing a variable number of synthetic vowels (1, 2 or 3) and a task in which the subject was free to report 1, 2 or 3 vowels. The parameters of this experiment were the number of vowels present and the relationship between their F_0s (all same, two same one different, all different).

4.2 Methods

4.2.1 Subjects

The experiment was first performed with 15 Japanese subjects (subset 2). Some effects of interest were only marginally significant, so it was decided to extend the experiment to 15 more subjects (subset 3), for a total of 30 subjects.

4.2.2 Stimuli

The stimulus set comprised single, double, and triple vowels. For ease in stimulus specification and generation, every stimulus was the sum of *three* vowels, different or not. For triple vowels, the three components were different. For double vowels, two component vowels were identical and the third different. For single vowels, all three components were the same. Ignoring order, there are 35 different combinations of 3 component vowels:

- [aaa, eee, iii, ooo, uuu]. These produced single vowels.
- [aee, aii, aoo, auu, eaa, eii, eoo, euu, iaa, iee, uoo, iuu, oaa, oee, oii, ouu, uaa, uee, uii, uoo]. These produced double vowels. One vowel appears twice in the sum, so its amplitude is 6 dB greater than that of the other.
- [aei, aio, aou, aeo, aiu, aeu, eio, eou, eiu, iou]. These produced genuine triple vowels.

To balance the number of single, double and triple vowels within the stimulus set, single vowels were repeated 4 times and triple vowels twice, leading to a total of 60 "triplets".

Component vowels were synthesized at 3 F_0s: 124, 132 and 140 Hz. This allowed three different patterns of F_0: all same, all different, two same-one different. The set of F_0 patterns was crossed with that of vowel patterns. This is easy to conceive in the case of triple vowels: the three vowels had either all the same F_0, or all three had different

F_0s, or two vowels shared an F_0 different from the third vowel. It is also easy to conceive in the case of single vowels: the single vowel was either *harmonic* (all F_0s the same), or *inharmonic* (made up of two or three harmonic series).

The case of double vowels is a bit more subtle. Both vowels could have the same F_0, or the weaker vowel could have one F_0 and the stronger vowel another F_0. However the stronger vowel could also be *inharmonic* (made up of two harmonic series). This case can be split into two: either the F_0s of the interference were both different from that of the target, or one of them was the same as that of the target.

Number of vowels and number of F_0s (independent harmonic series) were thus crossed almost orthogonally. A few precautions were necessary to balance the stimuli with respect to F_0 (to avoid that some conditions would have only high F_0s while others only low F_0s, etc.) [The reader might want to skip this]:

- All different. The middle frequency (132 Hz) has a special status: it is equidistant from the other two. Each vowel must have an equal chance to play the "odd man out", so the all-different condition must be realized in 3 ways.
- Two same, one different. Two F_0s are involved. We restrict ourselves to the case where they are contiguous and exclude the case where they are extreme (that would imply a larger ΔF_0). Two vowels have the same F_0 and one a different F_0. Each vowel must have an equal chance to play "odd man out", so we must realize this condition in three ways.
- All same F_0. For uniformity with the other two cases, this one is repeated 3 times.

To summarize, each of the three conditions of F_0 pattern was realized in three ways (F_0 orders). This led to (3 F_0 patterns) x (3 F_0 orders) x (60 triplets) = 540 conditions.

Absolute F_0 was not expected to have an effect (de Cheveigné 1997a), but it seemed wise to balance the probability of occurence of each F_0 or F_0 order, to avoid the possibility that a subject might learn to associate a given condition with a given F_0. F_0s were assigned at random, and this assignment was renewed at each session.

4.2.3 Task

Subjects were informed that each stimulus was a single, double or triple vowel, and they were requested to report one, two or three vowels for each stimulus.

4.2.4 Scoring

Let us distinguish the cases of single, double and triple vowels.

For single vowels, we measured the number of vowels reported and the identification rate (probably perfect) as a function of the patterns of F_0 (xxx, xxy, xyz).

For double vowels, we considered each vowel in turn. One was weak (-6 dB), the other strong (+6 dB). For the weaker vowel, there were 4 F_0 patterns to consider (notation target/ground): x/xx, x/yy, x/xy and x/yz. In the first, both vowels were harmonic with the same F_0. In the second, both were harmonic with different F_0s. In the third, the ground was the superposition of two harmonic series, one of which was the same as the target. In the fourth, both harmonic series of the ground were different from that of the target. In all four cases, the *target* was harmonic.

For the stronger vowel there were also 4 F_0 patterns to consider: xx/x, xx/y, xy/y, and xy/z. In the first, both vowels were harmonic with the same F_0. In the second, they were both harmonic with different F_0s. In the third, the target was the superposition of two harmonic series, one of which was the same as that of the ground. In the fourth, the target was the superposition of two harmonic series that both differed from that of the ground. In all four cases, the *ground* was harmonic.

Finally, for triple vowels we considered each of the three vowels in turn. Each was "weak" in the sense that it was in competition with two other vowels. There were four F_0 patterns: x/xx, x/yy, x/xy, x/yz. In the first, the target had the same F_0 as its two competitors. In the second, the competitors had the same F_0, different from the target. In the third, the competitors had different F_0s, one of which was the same as the target. In the fourth, the F_0 of both competing vowels were different from that of the target.

We measured the target identification rate and number of vowels reported in all of these cases.

4.3 Results

The details of ANOVAs and contrasts are given in Appendix B.

4.3.1 Number of vowels reported

Figure 3 shows the average number of vowels reported as a function of the number of different F_0s present in the stimulus, for single vowels (squares), double vowels (triangles) and triple vowels (circles). For double vowels made up of two F_0s, one can distinguish two cases. In the first (downward pointing triangles), each vowel was harmonic and had its own F_0. In the second (upward pointing triangles), one vowel was harmonic, but the other was the sum of two different harmonic series (one of which had the same F_0 as the other vowel). The number of vowels reported increased monotonically with the number of different vowels within the stimulus, and also with the number of different

F_0s within the stimulus.

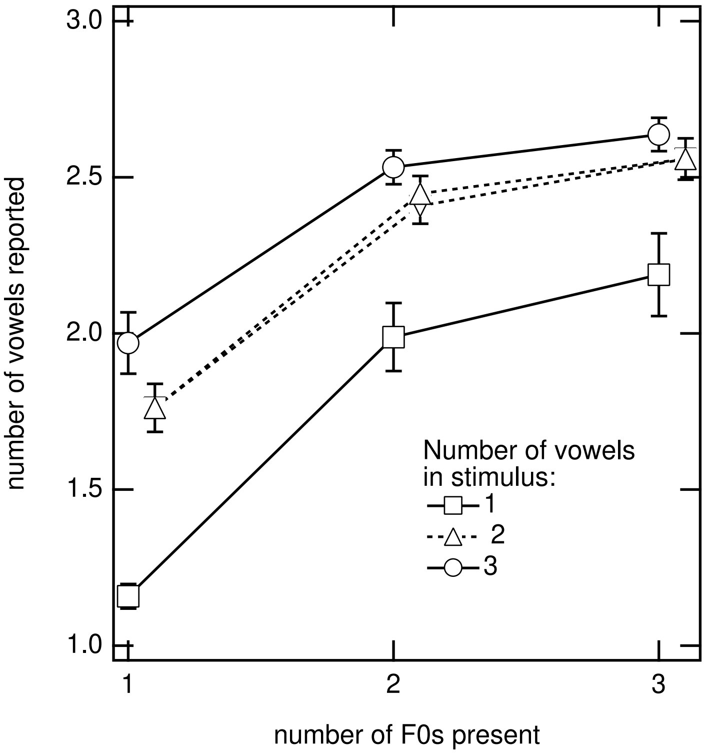


Fig. 3 *Number of vowels reported as a function of the F_0 pattern, for single (squares), double (triangles) and triple vowels (circles). See text for the difference between upward and downward pointing triangles. Error bars represent one standard error.*

Data for the two different two-vowel two-F_0 conditions were pooled together, and the data set was subjected to a repeated-measures ANOVA with factors number of vowels (1, 2, 3) and number of F_0s (1, 2, 3). Both main factors were highly significant. Their interaction was significant but rather small. The number of vowels reported was greater for two vowels than one, and greater for three than 2 (this last difference was only marginally significant). The number of vowels reported also increased with the number of F_0s present. It was greater for two than for one F_0, but the difference between 2 and 3 was not significant.

When the stimulus was a *single* vowel with *one* F_0, subjects rarely reported more than 1 vowel (average: 1.16). They reported more vowels if the stimulus was either inharmonic (2 or 3 F_0s) or was the mixture of 2 or 3 vowels. However even with 3 vowels and 3 F_0s, the number of vowels reported was less than 3 (average: 2.64). Segregation was certainly not sufficient for the subjects to realize that there were 3 vowels on all trials. It is also interesting to note that a *double* vowel with *two* F_0s evoked a relatively high average number of responses (2.44). Segregation was apparently not sufficiently good to convince the subjects that the stimulus contained only 2 vowels.

In addition to the "average number of vowels reported" score, the proportion of trials for which subjects reported 1, 2 and 3 vowels might be of interest. We did not attempt to analyze the data in that way.

4.3.2 Identification

Single vowels were identified almost perfectly, whatever the number of F_0s involved (Sect. B.4).

For double vowels, we distinguished weak (-6dB) and strong (+6dB) targets.

For weak targets (Fig. 4(a)), identification was poor when the competitor was harmonic and had the same F_0 as the target (x/xx). It was best when the competitor was harmonic and had a different F_0 from the target (x/yy). When the competitor was inharmonic and contained F_0s that were both different from the target (x/yz), identification was slightly impaired (the contrast with x/yy is marginally significant). It was more severely impaired if one of the two competing F_0s was the same as the target (x/xy).

These results are overall consistent with the hypothesis that a) the auditory system removes interference by harmonic cancellation of the competitor (s), b) several harmonic series may be cancelled, but c) two harmonic series are somewhat harder to remove than one, and d) the task is harder still if one of the competing F_0s is the same as that of the target (removing it would remove the target).

The task of segregating three vowels thus does not seem much much harder than segregating two. A word of caution however: much of the difference between x/xx and other conditions may be due to *multiplicity* cues that affect the tendency to report several vowels, and thus indirectly the identification rate (this is controlled for in Exp. 7).

The conditions x/xx and x/yy may be compared to conditions used in a previous double-vowel experiment with a ΔF_0 of 6 % (vs 6.45 % here) and a level mismatch of 10 dB (vs 6 dB here) (de Cheveigné 1997a). Those data are plotted as the dotted lines in Fig. 4(a, b).

For stronger targets (Fig. 4(b)), identification was overall better than for weaker targets. The difference between x/xx and the other three conditions was significant, but the differences among the latter were not. In all those three conditions the weaker competing vowel was harmonic. Previous experiments (de Cheveigné 1997b) found that identification was not degraded when the target was inharmonic rather than harmonic. It is thus not surprising that identification was not degraded when the target contained two F_0s and was thus inharmonic (xy/z), even if one of these two F_0s was the same was that of the competitor (xy/y). However identification rates were overall very high, so the lack of difference between xx/y, xy/y and xy/z may also be explained by a ceiling effect.

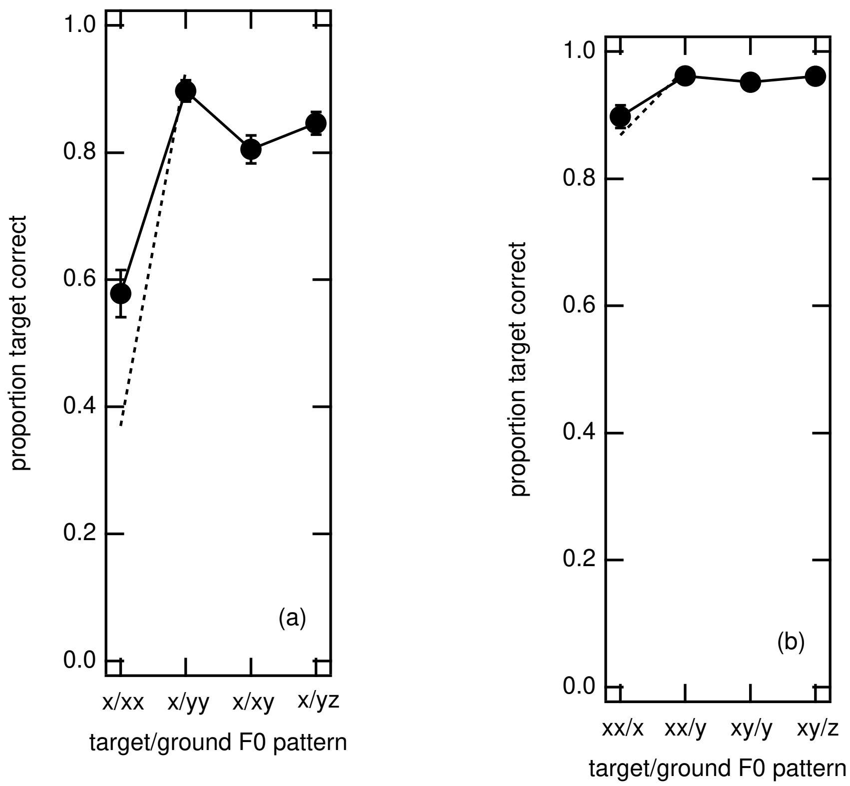


Fig. 4 *(a) Identification rate of the weaker (-6 dB) vowel as a function of the target/ground F_0 pattern. Error bars represent one standard error. The dotted line represents rates measured in a similar experiment at -10 dB (de Cheveigné et al. 1997a). (b) Same for the stronger (+6 dB) vowel. The dotted line represents rates measured in a similar experiment at +10 dB.*

For triple vowels (Fig. 5) identification was poor when both competitors had the same F_0 as the target (x/xx). It was best when both competitors had the same F_0, different from the target (x/yy). Identification was somewhat impaired if the two competitors had different F_0s (x/yz) (the contrast with x/yy was marginally significant). It was even more impaired if one of the two competing F_0s was the same as that of the target (x/xy).

These results are consistent with the hypotheses mentioned above (double vowels).

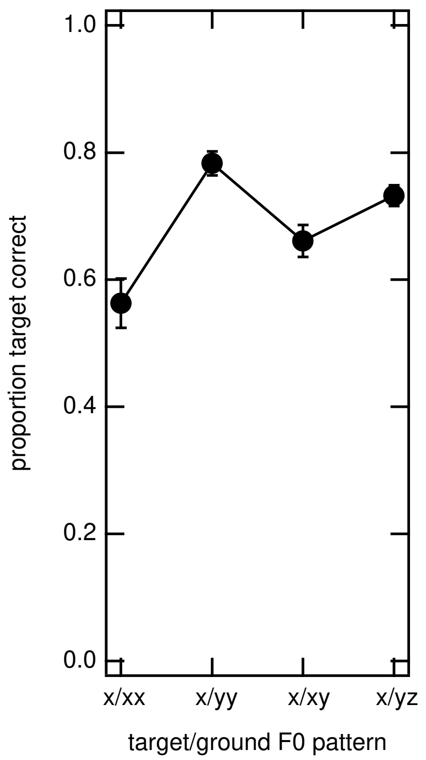


Fig. 5 *Identification rate of vowels within a triple vowel, as a function of the target/ground F_0 pattern.*

Figure 6 compares the identification of a vowel mixed with two different competing vowels (triple vowel) to that of a vowel mixed with twice the same competing vowel (double vowel, weak target). For the different-F_0 conditions, identification was better when the competitor consisted of twice the same vowel, rather than two different vowels. The reduced identification in the latter case can be attributed to more effective masking (masker energy is distributed rather than located at a few formants), or a cognitive confusion effect, due to the greater number of vowels. However neither explanation accounts for the fact that there was *no effect* when F_0s were the same.

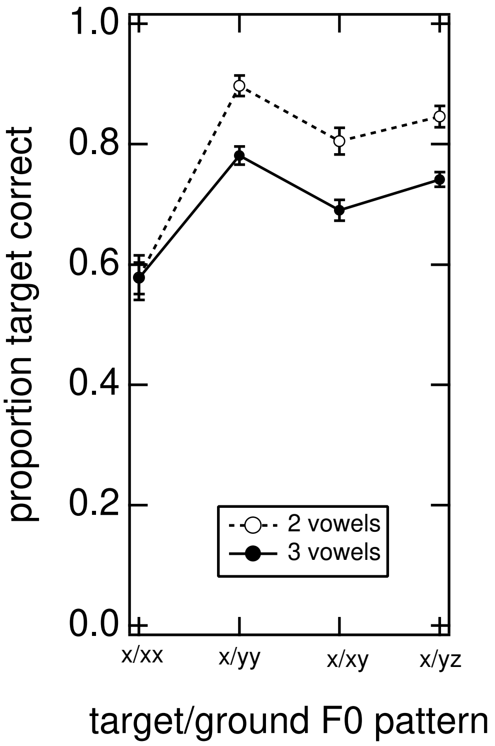


Fig. 6 *Target identification rate as a function of the target/ground F_0 pattern. Open symbols: the competitors are different vowels (3 vowel). Filled symbols: they are twice the same vowel (2 vowel).*

4.4 Discussion

Subjects reported more vowels when the stimulus contained two vowels than one, but the smaller difference between three and two was only marginally significant. They also reported more vowels when the stimulus contained two F_0s rather than one, but the smaller difference between three and two was not significant.

This is consistent with the results of Kashino and Hirahara (1995) that found that subjects underestimated the number of concurrent voices when this number was greater than 2.

Identification of members of triple vowels improved when the target had a different F_0 from both of its competitors. This improvement was slightly reduced when the competitors had different F_0s among themselves, and therefore constituted a form of *inharmonic* interference. This result is to be compared with previous data that showed that identification was impaired when interference was inharmonic (de Cheveigné 1997b).

Identification in the x/yz condition, although reduced relative to the x/yy condition, was significantly better than in the x/xx condition. This might be interpreted as meaning that the auditory system can perform simultaneous cancellation of several interfering harmonic interferers.

However, much of the effects are due to low identification rates in the x/xx state, where all F_0s are the same and the stimulus is harmonic. Subjects tend to report fewer vowels in that case, and this of course affects the identification rate. Experiment 7 replicates the three-vowel experiment with a task in which subjects had to report 3 vowels for all stimuli.

5 Experiment 3: The effect of ΔF_0 over a wide range of amplitudes

5.1 Introduction

Previous experiments explored the interaction between level and ΔF_0 (McKeown, 1992; de Cheveigné et al. 1997a), and found that ΔF_0 effects were relatively large for weak targets, and smaller for strong targets. However the range of levels and ΔF_0s explored was limited. Here we use a wider range. We wish to determine how weak a target may become before ΔF_0 effects vanish, and whether this limit depends on the size of the ΔF_0.

5.2 Methods

The task was similar to that used in Experiment1, and previous experiments (de Cheveigné et al. 1997a,b). The stimulus set consisted of double vowels made by adding single vowels. Single vowels were synthesized with F_0s of 124, 128, 130, 132, 134, 136, 140 Hz, with "random" phase. F_0s were paired to obtain ΔF_0s of 0 % (132, 132 Hz), 3 % (130, 134 Hz), 6 % (128, 136 Hz) and 12 % (124, 140 Hz). Vowels were added with amplitude differences of 5, 15, 25 and 35 dB (leading to target/competitor ratios ranging from -35 to 35 dB in 10 dB steps). There were (4 F_0s) x (4 amplitude differences) x (20 ordered pairs) x (2 F_0 orders) = 640 pairs. The stimulus set contained no single vowels (we reasoned that double vowels with an amplitude mismatch of 25 or 35 dB are very close to being single vowels, and that inclusion of single vowels was not necessary for the stimulus set to be consistent with the description made to the subjects).

5.3 Results

Details of ANOVA and contrasts are shown in Appendix. C.

5.3.1 Number of vowels reported

Figure 7 shows the average number of vowels reported as a function of amplitude mismatch between vowels. For each ΔF_0 the number of vowels reported decreased with amplitude mismatch. This is understandable, as a large amplitude mismatch makes the stimulus similar to a single vowel. At all amplitudes except 35 dB, the number of vowels reported was significantly greater when $\Delta F_0 \neq 0$ than when $\Delta F_0 = 0$ (at 35 dB there was no significant effect). The difference between 3 % and (6, 12 %) was marginally significant at 25 dB, but not at other amplitudes. Most of the ΔF_0 effect occured between

0 and 3%, and very little beyond.

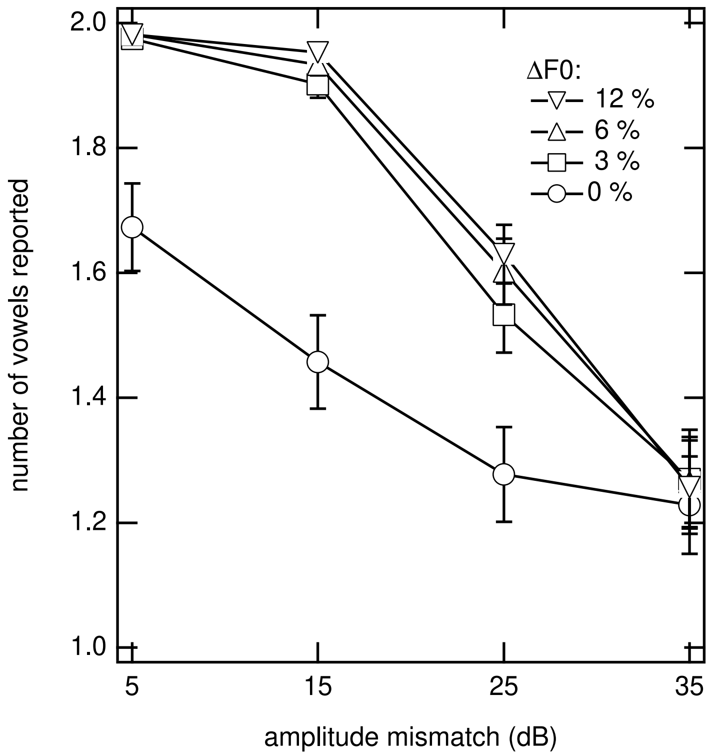


Fig. 7 *Number of vowels reported as a function of intervowel amplitude difference, for each* ΔF_0.

5.3.2 Identification rate

Figure 8 shows the target-correct identification rate as a function of the target/competitor ratio, for each ΔF_0. Identification was better when the target was strong relative to the competitor. It was also better when $\Delta F_0 \neq 0$ than when ΔF_0 = 0 at target/competitor ratios of -25, -15, -5 and 5 dB (there was no significant effect at -35 dB, or at 15, 25 or 35 dB). The difference between 3 % and (6, 12 %) was significant at -15 dB but not at other levels. Again, the ΔF_0 effect is mostly limited to the difference between 0 and 3

%.

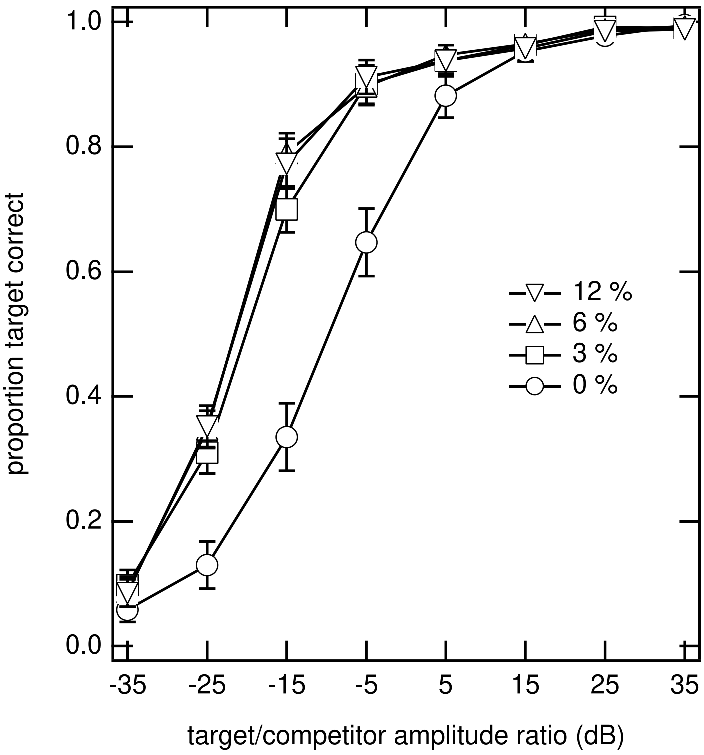


Fig. 8 *Target-correct identification rate as a function of intervowel amplitude difference, for each* ΔF_0.

5.4 The effect of absolute F_0

In previous experiments, we found no evidence that identification might be affected by the *absolute* F_0 of the target or competitor (de Cheveigné et al. 1997a,b). However the range of F_0s was smaller in that experiment. Here the range is larger, since we included a ΔF_0 = 12 % condition.

Response data for ΔF_0 = 3, 6, and 12 % were reanalyzed to measure the target identification rate as a function of target F_0 (low vs high). The data were analyzed separately for each amplitude ratio, by an ANOVA with factors (ΔF_0 = 3, 6, 12 %) x (F_0 = low, high).

At -25 dB, there was a significant main effect of F_0 and no interaction with ΔF_0. Targets were identified less well at the higher F_0 (0.29 %) than at the lower F_0 (0.38 %). At all other amplitude ratios, F_0 and its interaction with ΔF_0 were not significant.

The lack of effect at -15 dB and higher is consistent with our previous observation of a lack of effect at -20 and -10 dB target/competitor ratio. Why the effect is evident at -25 dB is a mystery. It is conceivable that the ΔF_0 effect observed at -25 dB is due entirely to one or two pairs, that happen to show a sensitivity to F_0. We have not investigated this question.

5.5 Discussion

Results were similar to those reported by de Cheveigné et al. (1997b). The new experiment adds the following information:

- ΔF_0 affects both identification and the number of vowels reported for targets as weak as -25 dB relative to the competing vowel. Our previous data showed improvement down to -20 dB (de Cheveigné et al. 1997a). At -35 dB the effect was too small to be measurable. We have thus an idea of the lower limit of measurable effects.

- Most of the ΔF_0 effect occured between 0 and 3 %, beyond which scores were hardly affected by ΔF_0. This pattern was observed by previous authors for equal-amplitude targets. Our results extend it to weak targets. A priori, the plateau of identification rate at larger ΔF_0s could have been attributed to a ceiling effect. If so, weak targets would have benefitted more from a larger ΔF_0s. Such was not the case.

The detailed pattern as a function of vowel pair should be used to test models of vowel perception. We did not attempt to do so here. Experiment 3 confirmed that most ΔF_0 effects occur in the region of small ΔF_0s. This emphasizes the need to explore this region where most of the changes in segregation occur. That is the purpose of Experiment 4.

6 Experiment 4: ΔF_0 effects at small ΔF_0s

6.1 Introduction

In double vowel experiments, most of the improvement in identification occurs within the range of smallest ΔF_0s, between 0 % and the lowest non-zero value of the parameter set (typically 6, 3 or 1.5 %). This region of the parameter space has not been investigated in detail (most efforts have concentrated on the range of large ΔF_0s, where differences in identification are small).

A difficulty with small ΔF_0s is that they give rise to beat patterns with long periods. When the beat period is longer than the stimulus, the long-term spectrum depends on starting phase, and is not unequivocally determined by ΔF_0. This problem is often not recognized or controlled for, and it is possible that effects reported for small ΔF_0s were specific to the particular portion of the beat period that happened to be included in the stimulus (or more generally, to the particular starting phase relationship between partials of both vowels).

The aim of this experiment was to explore the small-ΔF_0 region, while controlling for the effects of phase-dependent interaction.

6.2 Methods

The task was the same as for Experiments 1 and 3. Subjects were told that a stimulus could consist of one or two vowels, and they were free to report one vowel or two. However, as in Experiment 3, the stimulus set contained only double vowels. [4]his is a departure from our doctrine of ensuring that the stimulus set is consistent with the description made to the subjects. The stimulus set contained a large number of conditions with an amplitude mismatch of 15 dB, and a ΔF_0 at or near zero. We reasoned that describing the stimulus set as containing "both single and double vowels" would seem plausible to the subjects, and that single vowels were not necessary. Including a significant proportion of single vowels would have made the stimulus set too big.

The double vowels were made by adding pairs of single vowels with F_0s centered on 132 Hz: ΔF_0 = 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). F_0s were placed symmetrically about 132 Hz to ensure that subjects would not be influenced by differences in mean F_0.

The stimulus duration of 250 ms (between -6 dB points) is equal to the beat period at ΔF_0 = 3 % (4 Hz). When the ΔF_0 is smaller than 3 %, the stimulus contains only a fraction of a beat period. For example at 0.375 % the stimulus represents one eighth of a beat period. The overall spectrum of the stimulus thus depends on the particular portion of the beat period that was selected, and it is conceivable that this might affect the way it is identified. To control for this possibility, conditions were repeated with as many successive segments as necessary to cover a half beat period [5]e reasoned (incorrectly) that the beat pattern would be symmetrical in time, and that sampling one half of its period was sufficient. Our stimulus set is therefore incomplete. Nevertheless, with 2 successive segments at 0.75 % and 4 segments at 0.375 %, the sample is complete enough to reveal eventual phase effects. In any case, the stimulus set was already very large and could not have accomodated more stimuli. While this is but a small sample of possible phase relationships, it is sufficiently wide to alert us to a possible phase-specificity of ΔF_0 effects at small ΔF_0s.

[4]T
[5]W

The ΔF_0 = 0.375 % condition was thus synthesized in 4 different versions. Each was a successive segment of a beat pattern, that is the two component vowels were summed with a increasingly large delay (1/16, 3/16, 5/16, 7/16 of a period). We also synthesized four versions of the ΔF_0 = 0 % with the same intervowel delay. In this way, each of the four ΔF_0 = 0.375 % segments could be compared to a ΔF_0 = 0 % segment of similar global spectrum. A previous study showed that, at small ΔF_0s, successive segments excised from a double-vowel may produce different identification scores (Assmann and Summerfield 1994). It has been proposed such beat patterns might enhance identification of vowels within pairs, and might thus account for "ΔF_0ëffects at small ΔF_0s. If such were the case, we would observe a) differences between different segments (intervowel phases), at both ΔF_0 = 0.375 % and ΔF_0 = 0 %, and b) no difference between corresponding segments as a function of ΔF_0.

There were thus: one phase pattern at ΔF_0 = 6, 3, 1.5 %, two at ΔF_0 = 0.75 %, and four at ΔF_0 = 0, 0.375 %. Each condition was realized with two F_0 orders (low/high and high/low), and with an amplitude mismatch of 0 and 15 dB. Conditions were doubled at 15 dB, in order that each condition be realized at least once with a weak (-15 dB) target. There were thus (13 F_0 and phase conditions) x (2 F_0 orders) x (3 amplitudes) x (10 unordered vowel pairs) = 780 stimuli within a stimulus set.

From previous results, we expect effects to be clearest for a 15 dB amplitude mismatch. A 0 dB amplitude mismatch was nevertheless included to allow comparison with previous reports of beat effects. The amplitude of interactions such as beats is likely to be largest when both vowels have the same amplitude.

6.3 Results

Details of ANOVA and contrasts are given in Appendix D. Results

6.3.1 Phase effects at small ΔF_0s

The number of vowels reported is plotted in Fig. 9 for vowel amplitude ratios of 15 dB (a) and 0 dB (b). At both amplitudes the main effect of ΔF_0 was significant: subjects reported two vowels more often when there was a ΔF_0. The main effect of phase was barely significant at 0 dB, and not significant at 15 dB. The interaction was significant at 15 dB, but not at 0 dB.

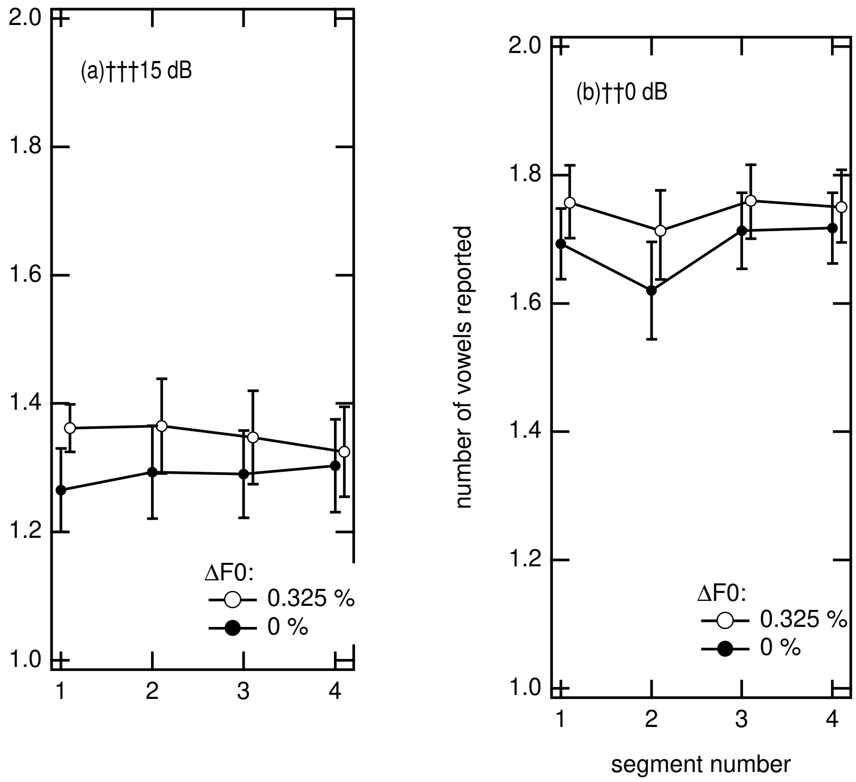


Fig. 9 *(a) Number of vowels reported as a function of the segment number (phase or delay between vowels) for ΔF_0 = 0 % (filled symbols) and 0.375 % (open symbols), for an amplitude mismatch of 15 dB. (b) Same, for an amplitude mismatch of 0 dB.*

The target-correct identification rate is plotted in Fig. 10 for target/competitor ratios of -15 dB (a) and 0 dB (b). The main effect of ΔF_0 was not significant at 0 dB, but it was at -15 dB: subjects identified targets more accurately when ΔF_0 was not zero. Phase and interaction were not significant.

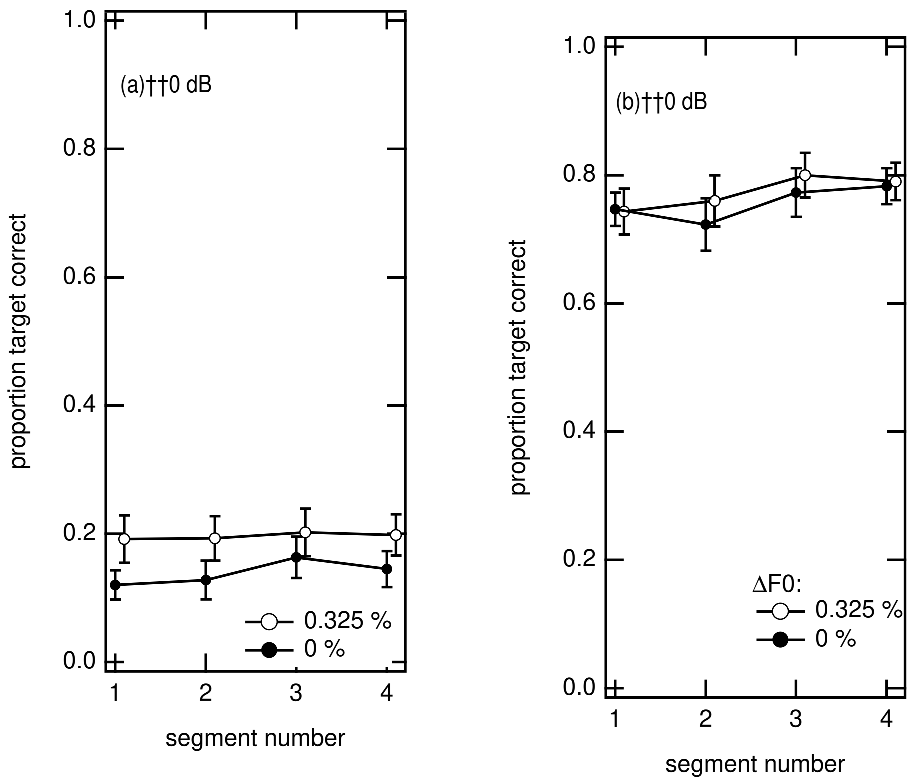


Fig. 10 *(a) Target-correct identification rate as a function of the segment number (phase) for ΔF_0 = 0 % (filled symbols) and 0.375 % (open symbols), for an amplitude mismatch of 15 dB. (b) Same, for an amplitude mismatch of 0 dB.*

6.3.2 ΔF_0 effect

Scores at ΔF_0 = 0, 0.375 and 0.75 % were averaged over segments (phase). Figure 11 shows the number of vowels reported as a function of ΔF_0 for amplitude differences of

15 dB (squares) or 0 dB (circles).

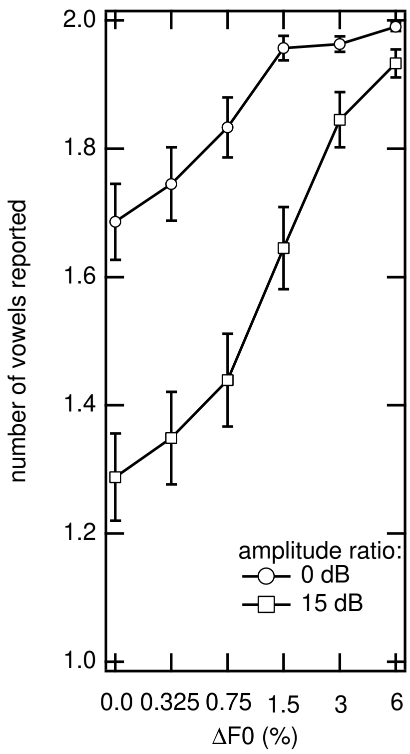


Fig. 11 *Number of vowels reported as a function of ΔF_0 for an amplitude difference of 0 dB (circles) or 15 dB (squares).*

Figure 12 shows the target-correct identificaton rate as a function of ΔF_0 for a target/competitor ratio of 15 dB (squares) or 0 dB (circles).

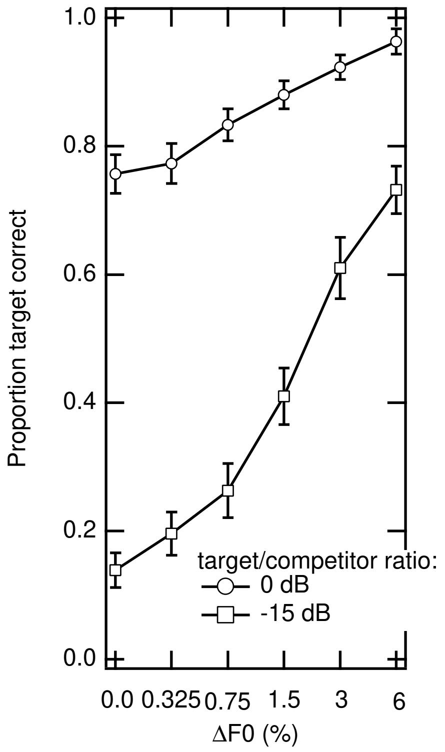


Fig. 12 *Target-correct identification rate as a function of ΔF_0 for a target/competitor ratio of 0 dB (circles) or 15 dB (squares).*

6.4 Discussion

Both the number of vowels reported and the identification rate increased with ΔF_0, even for a ΔF_0 as small as 0.375 % (1/16th of a semitone, or 0.5 Hz). Contrary to a theory that has been put forward to explain segregation at small ΔF_0s (Assmann and Summerfield, 1994; Culling and Darwin, 1994), this ΔF_0 effect does not seem to be the result of beats. In agreement with that theory, we found evidence that scores were dependent on the particular segment of the beat pattern that was chosen for the stimulus. However, each successive segment at ΔF_0 = 0.375 % was better identified than a corresponding segment at ΔF_0 with the same average phase relationship.

It is interesting that identification rate and number of vowels reported are enhanced by a ΔF_0 so small that the stimulus contains only 1/8th of a beat period.

7 Experiment 5: Duration

7.1 Introduction

Assmann and Summerfield found that ΔF_0 effects observed at 200 ms duration vanished at 52.5 ms duration. Assmann and Summerfield (1994) found that successive 50 ms segments excised from a 200 ms double-vowel stimulus were not equally identifiable.

Short stimulus durations pose the same problem as small ΔF_0s: if the stimulus is shorter than the beat period, its global spectrum depends on the part of the beat period that it occupies. In other words the global spectrum is phase-sensitive, and without specification of this phase the description of the stimulus is not complete.

The aim of Experiment 5 was to explore effects of the duration parameter in the same fashion as Experiment 4 explored the ΔF_0 parameter. We controlled for phase effects by repeating the smallest duration stimuli twice, each time with a different portion of the beat period.

7.2 Methods

Double vowels were synthesized with ΔF_0s of 0 % (132, 132 Hz), 3 % (130, 134 Hz) and 6 % (128, 136 Hz). Stimulus durations were 145 ms and 82.5 ms, with 20 ms raised cosine onsets and offsets. The durations between -6 dB points were 125 and 62.5 ms (as compared to 250 ms for Exp. 4).

The stimuli were synthesized with two different intervowel starting phases, corresponding to the ongoing phases at the center points of two successive quarters of a beat period. [6]

Levels were 0dB and $\pm$15 dB. There were a total of (3 ΔF_0s) x (2 phases) x (2 durations) x (2 F_0 orders) x (10 unordered pairs) = 720 stimuli.

7.3 Results

Figure 13 shows the number of vowels reported as a function of ΔF_0, for each of the durations and phase conditions, at an amplitude mismatch of 15 dB (a) and 0 dB (b). Dotted lines (circles) represent data obtained with 250 ms stimuli in Exp. 4 [7]. Whatever the duration, the number of vowels reported increased with increasing ΔF_0. When there was a level mismatch of 15 dB, the increase with ΔF_0 was smaller at short than at long durations. Duration affected more the step between 0 and 3 % than that between 3 and

[6] Stimuli at 125 ms should have been synthesized with only one phase, by mistake they were synthesized with both. Responses for these two phases were pooled.

[7] The stimulus sets of Exp. 4 and 5 were not the same, so this comparison is not perfectly fair

6 %.

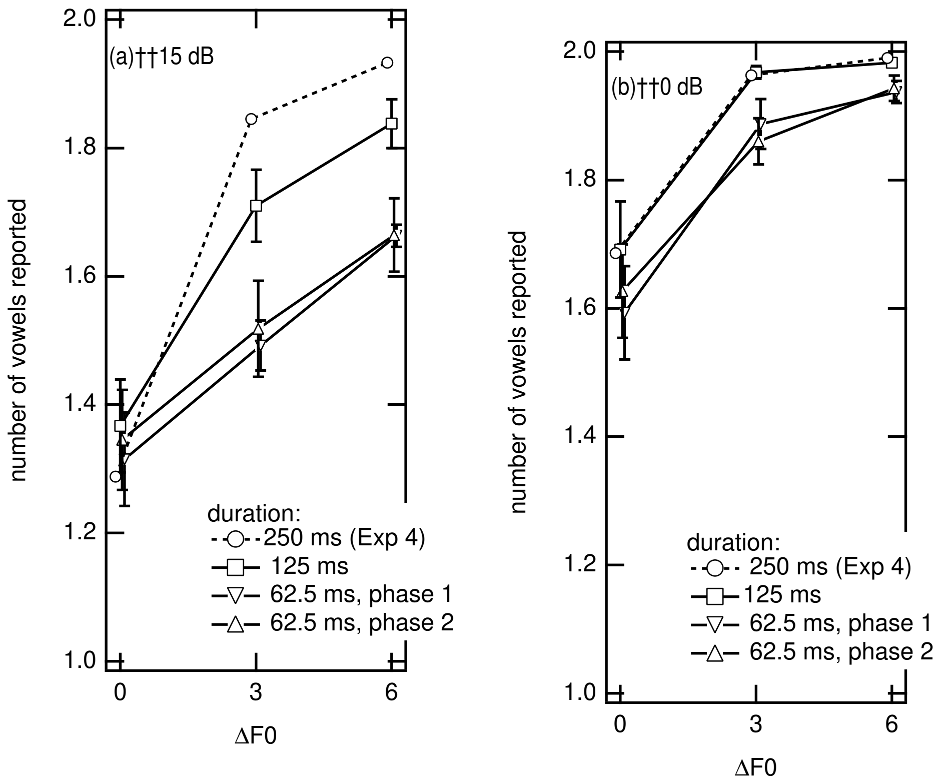


Fig. 13 *Number of vowels reported as a function of ΔF_0 for each duration, at 15 dB (a) and 0 dB (b) amplitude difference. Dotted lines and circles represent data obtained at 250 ms in Experiment 4.*

Figure 14 shows the target-correct identification rate as a function of ΔF_0, for each of the durations and phase conditions at -15 dB (left) and 0 dB (right). Dotted lines (circles) represent data obtained with 250 ms stimuli in Exp. 4 Whatever the duration, the identification rate increased with ΔF_0. When the target/competitor ratio was -15 dB, the increase with ΔF_0 was smaller at short than at long durations. Duration affected more the step between 0 and 3 % than that between 3 and 6 %.

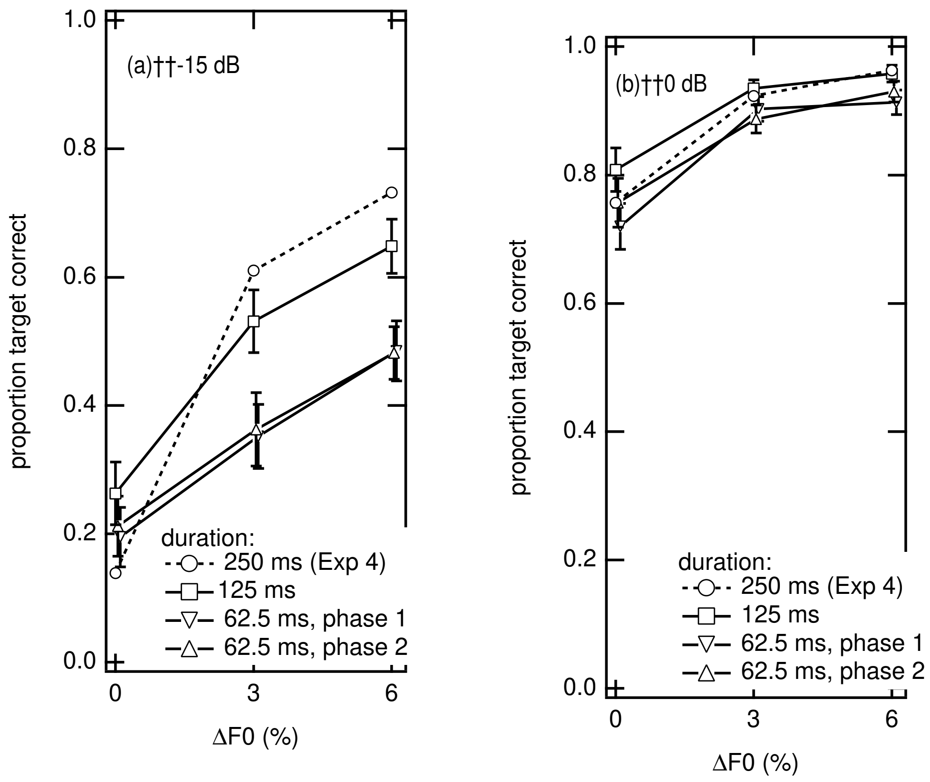


Fig. 14 *Target-correct identification rate as a function of ΔF_0 for each duration, at -15 dB (a) and 0 dB (b) target/competitor ratio. Dotted lines and circles represent data obtained at 250 ms in Experiment 4.*

7.4 Discussion

In contrast to the results of Assmann and Summerfield (1990) for 52.5 ms stimuli, we observed clear ΔF_0 effects for stimuli of 62.5 ms. This difference might be explained by the difference in task. Our subjects were allowed to report one or two vowels, whereas those of Assmann and Summerfield had to report two vowels for each stimulus. Our subjects may have been influenced by multiplicity cues, whereas theirs had to ignore them and use only "unmasking" cues. It is conceivable that those cues vanish for short stimuli, while multiplicity cues remain. One should note also that our effects were relatively small for equal amplitude vowel pairs.

In contrast to the results of Assmann and Summerfield (1994) for successive 50 ms portions excised from a 200 ms double-vowel stimulus, we found no significant effect of the phase factor (equivalent to successive segments of a beat pattern). This may be simply because our sampling of the phase factor was different.

8 Experiment 6: In search of harmonic enhancement

8.1 Introduction

Models of harmonic segregation come in two flavors: harmonic enhancement and harmonic cancellation (de Cheveigné, 1993; de Cheveigné et al., 1995). According to the former, the harmonic structure of a target sound is used to "pull it out" of interference. According to the latter, the harmonic structure of interference may be used to suppress it. So far, the balance of evidence is in favor of harmonic cancellation, and against harmonic enhancement (Lea 1992; Summerfield and Culling, 1992; de Cheveigné et al. 1995, 1997a,b). Nevertheless, the principle of harmonic enhancement underlies many models, in particular "auditory scene analysis" models, as well as many algorithms for "speech enhancement". It is an appealing principle, and it would be surprising if the auditory system made no use of the harmonic structure of targets whatsoever.

The aim of this experiment, as well as of Experiment 8, is to attempt to find a role for target harmonicity. So far we found no evidence that it serves for simultaneous grouping or segregation; our hypothesis here is that it serves for "sequential grouping".

This experiment is based on the hypotheses that segregation is relatively ineffective for short durations (Assmann and Summerfield 1990; see also Experiment 5), and that the F_0 estimation mechanism is somewhat "sluggish", and may be fooled by certain patterns of transitions.

8.2 Methods

Stimuli consisted of double vowels, formed by pairing single vowels (same or different) and adding them with an amplitude difference of 15 dB. Each single vowel itself consisted of two consecutive 82.5 ms pulses, shaped with 20 ms raised cosine ramps (62.5 ms between -6 dB points), separated by 105 ms of silence. The total duration of each stimulus was 270 ms (250 ms between -6 dB points) as in most of our other experiments.

The F_0 of each pulse was chosen among 124, 128, 132 and 136 Hz. The F_0s of the two consecutive pulses could be the same, or they could differ by 3 or 6 % (4 or 8 Hz). A single vowel can thus be seen as modulated in amplitude (with a "hole" in the middle) and in some cases in frequency (a transition accross the "hole").

Single vowels were paired to form double vowels with the following constraint: a) the ΔF_0 of the first pulse was zero, b) within the second pulse, the F_0 of each vowel was either the same as in the first pulse, or different by 6 % (8 Hz). The accepted transitions are illustrated in Figure 15. The target is shown as a dotted. line.

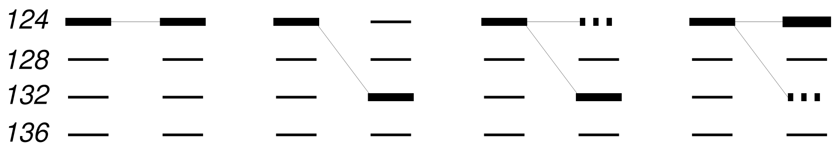


Fig. 15 *F_0 patterns used in Experiment 6. Only patterns "anchored" on 124 Hz are shown. The stimulus set also included patterns "anchored" on 128 and 124 Hz that are not shown, for a total of 16 F_0 patterns. Dotted lines represent the target (weaker vowel).*

There were (16 patterns) x (25 ordered pairs) x (2 repetitions) = 800 stimuli. Pairs in which both vowels were the same constituted single vowels (harmonic or inharmonic, according to the F_0 pattern). There were 160 such "single vowels".

The task was as in other experiments, to report one or two vowels for each stimulus.

8.3 results

Details of ANOVA and contrasts are given in Appendix F. Results reported are for 14 subjects (data for one subject was still missing at the time of writing).

Results were averaged over all four starting F_0s. There are four different F_0 patterns:

1. F_0 = 0, no jump in F_0,
2. F_0 = 0, a 6 % jump in F_0,
3. F_0 = 6 % (on second pulse), no jump in target, competitor jumps by 6 %,
4. F_0 = 6 % (on second pulse), no jump in competitor, target jumps by 6 %.

The number of vowels reported was greater when there was a difference in F_0 during the second pulse (1.81) than when the F_0s were the same (1.55). The identification rate was also greater (0.51 vs 0.35). However in the ΔF_0 = 0 % condition it made no difference whether or not there was a 6 % jump in F_0 of both vowels, and in the ΔF_0 = 6 % condition it made no difference whether the target or the competitor jumped by 6 %.

8.4 Discussion

Both measures were affected by the ΔF_0 of the second pulse. Neither measure was significantly affected by the pattern of F_0 transitions of either the target or the competing vowel. In particular in the ΔF_0 = 6 % condition, there was no evidence that a jump in target F_0 impaired identification. Either our hypothesis of segregation based on target F_0 continuity was false, or our efforts to impair the target F_0 tracking ability of the auditory system were insufficient.

9 Experiment 7: Triple vowels with a 3-vowel forced response task

9.1 Introduction

Experiment 2 investigated the segregation and identification of mixtures of three vowels. Subjects were allowed to answer one, two or three vowels for each stimulus, according to what they heard. Identification of a hard-to-hear vowel was mainly determined by whether or not the subject decided to report three vowels. That is, identification rate was strongly dependent on "multiplicity" cues.

This experiment investigates the same situation, but when multiplicity cues are disabled (because the subject must report three vowels).

9.2 Methods

Stimuli were a subset of the stimuli of Experiment 2. Only genuine triple vowels were included. There were (10 triplets) x (3 F_0 patterns) x (3 F_0 orders) x (8 repetitions) = 720 stimuli. The task was similar to that of Experiment 2, except that the stimulus set only contained triple vowel stimuli, and subjects were required to report 3 vowels for each stimulus.

9.3 Results

Details of ANOVA and contrasts are given in Appendix G.

Figure 16 shows the target-correct identification rate as a function of F_0 pattern, together with a similar score measured for triple vowels in Experiment 2. Identification is worst when all vowels have the same F_0 (x/xx) and best when both competitors have the same F_0, different from the target (x/yy). Identification is less good when competitors have different F_0s, both distinct from the target (x/yz). However it is better than when all vowels have the same F_0 (x/xx). This suggests that segregation is still effective, despite the inharmonicity of the "masker" (sum of competing vowels). In other words, harmonic cancellation of two competing harmonic series is possible, although it is much less effective than one. This is confirmed by the fact that identification is impaired if one of the competitors has the same F_0 as the target (cancellation of that competitor would cancel the target). Identification in the (x/xy) condition is not significantly different from that in the (x/xx) condition.

Identification was overall better that that measured in Experiment 2. This is easy to understand as a consequence of the 3-vowel response forced task (although one cannot exclude a contribution of training). Comparing patterns across experiments, the pattern across conditions (x/yy), (x/xy) and (x/yz) is similar between the two experiments, apart from the overall difference in rate. The (x/xx) condition is different. The difference between that condition and the others was much greater in Exp. 2 than in Exp. 7, essentially because subjects tended to (and were allowed to) report fewer vowels. One last thing to note is that standard errors are smaller in Exp. 7 than in Exp. 2.

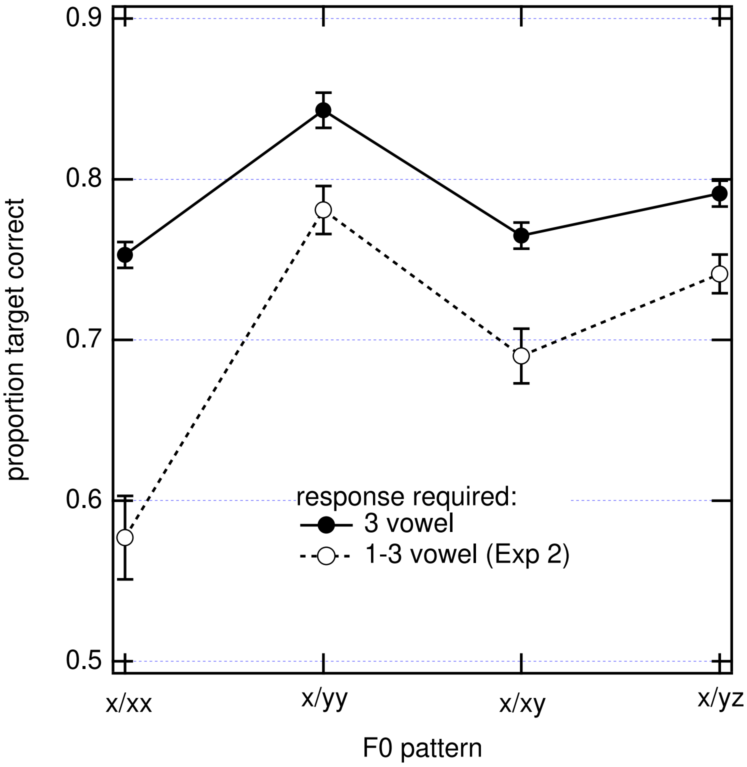


Fig. 16 *Target-correct identification rates as a function of the target/competitor F_0 pattern obtained in Exp. 7 (filled symbols) and Exp. 2 (open symbols).*

10 Experiment 8: VERY small ΔF_0

10.1 Introduction

Experiment 4 explored ΔF_0 effects for small ΔF_0s. Unexpectedly, effects were found with ΔF_0s as low as the lowest value in the range (0.375 %). The first aim of this experiment was to extend the range to ΔF_0s that were smaller still.

In Experiment 4 (as in many double-vowel experiements) the F_0s all clustered around a single baseline value (132 Hz). The auditory system might conceivably use this fact to enhance filtering properties near that frequency. If so, performance might be degraded if the baseline F_0 were roving. A second aim of this experiment was to test for that possibility. Delta F_0 conditions were constructed with F_0s near 124, 128 and 132 Hz.

An outcome of Experiment 4 was that performance depended slightly on phase at ΔF_0 = 0, 0.375 %. The space of possible phase relations is vast, and the conditions tested in Experiment 4 (adjacent portions of a beat pattern) were but a sample. The present experiment takes one more sample: comparison is made between vowel pairs with same starting phase, and opposite (pi) starting phase.

10.2 Methods

The task was the same as in Exp. 4. Single vowels were synthesized at F_0s of 124, 128 and 132 Hz, and also at F_0s that were higher by 0.125, 0.25, 0.5 and 1 Hz. They were paired and added with an amplitude mismatch of 5 or 15 dB, to form double vowels with ΔF_0s of 0.125, 0.25, 0.5 and 1 Hz, or approximately 0.1, 0.2, 0.4, 0.8 % (the 0.4 and 0.8 % conditions are equivalent to the 0.375 % and 0.75 % conditions of Exp. 4). Both vowels were added in phase, or one vowel was inverted (π phase) before addition.

There were: (5 delta f0s) x (2 phases) x (2 amplitudes) x (2 F_0 orders) x (20 ordered pairs) = 800 stimuli. Absolute F_0 was chosen randomly from trial to trial.

10.3 Results

Details of ANOVA and contrasts are given in Appendix H.

10.3.1 Effects of phase

Number of vowels reported At a 15 dB amplitude difference, a repeated-measures ANOVA with factors ΔF_0 and starting phase (same phase vs opposite phase) indicated a significant effect of ΔF_0 and a marginally significant effect of phase.

Subjects reported slightly fewer vowels when vowels were added in antiphase than in phase. The interaction with ΔF_0 was not significant. Figure 17 (a) shows the number of vowels reported as a function of the starting phase relationship beween vowels of a pair, at each ΔF_0.

At 5 dB the ΔF_0 effect was again highly significant, but neither the phase effect nor the interaction were significant (Fig. 17 (b)).

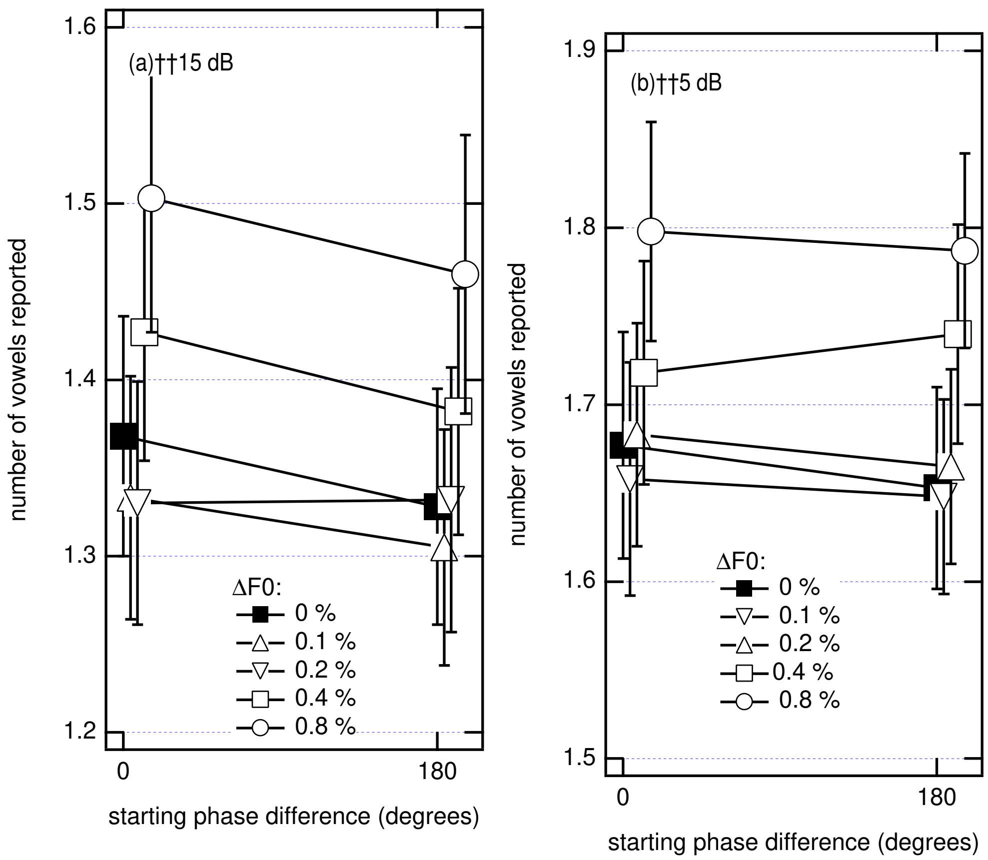


Fig. 17 *Number of vowels reported as a function of starting phase relationship (in phase vs opposite phase), for several values of* ΔF_0*. Left: 15 dB amplitude difference, Right: 5 dB amplitude difference.*

Identification rate When the target/competitor ratio was -15 dB, the main effects of ΔF_0 and phase were significant, as was their interaction. Figure 18 (left) shows the identification rate as a function of phase, for different values of ΔF_0. When there was a non-zero ΔF_0, the effect of phase was not significant. When ΔF_0 = 0 it was highly significant. Adding vowels at unison in opposite phase greatly impaired identification of the weak target.

When the target/competitor ratio was -5 dB, the main effects of ΔF_0 and phase and their interaction were again significant (Fig. 18 (right)).

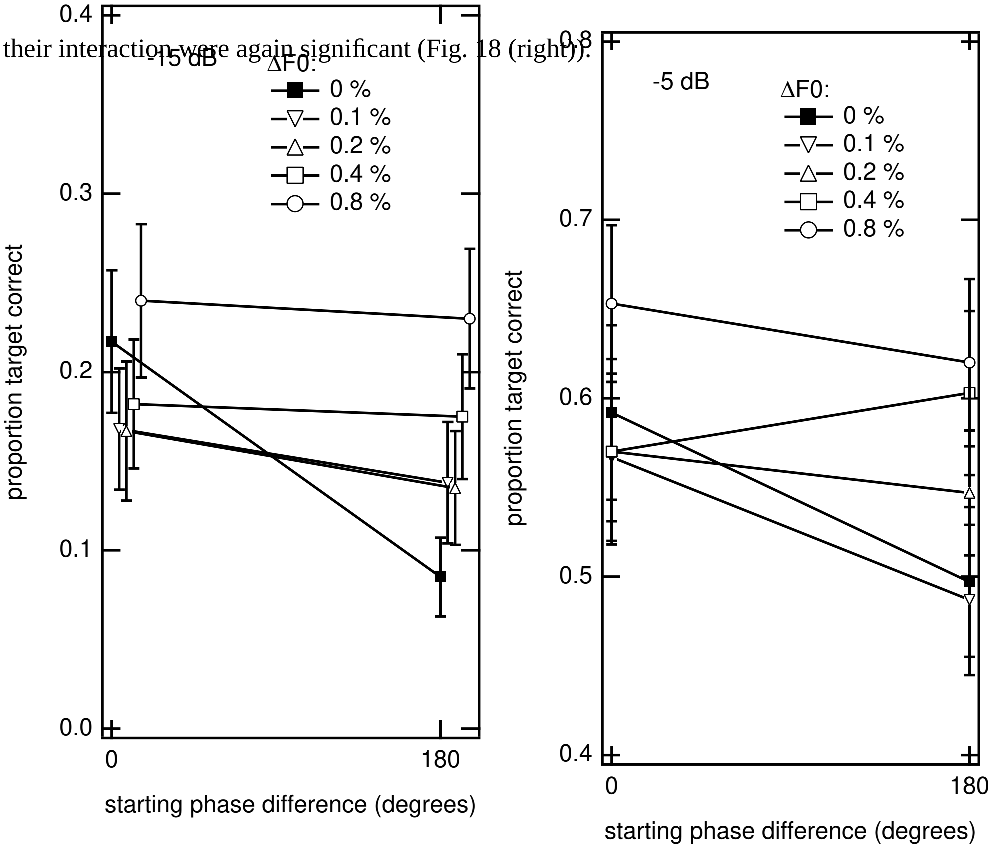


Fig. 18 *Identification rate as a function of starting phase relationship (in phase vs opposite phase), for several values of ΔF_0. Left: 15 dB amplitude difference, Right: 5 dB amplitude difference.*

10.3.2 Effects of ΔF_0

Number of vowels reported Figure 19 shows the number of vowels reported as a function of ΔF_0 (averaged over phase), for an amplitude mismatch of 15 dB (left) or 5 dB (right). Dotted line in Fig. 19 (left) represents data obtained in Experiment 4 for similar stimuli.

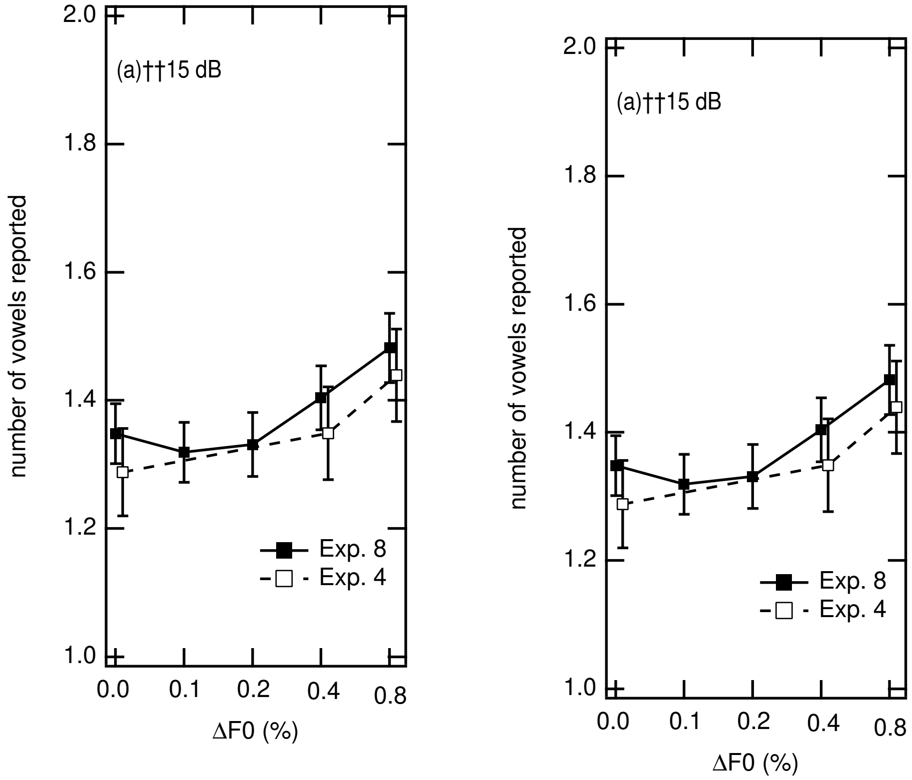


Fig. 19 *Number of vowels reported as a function of* ΔF_0*. Left: 15 dB amplitude difference. The dotted line represents data obtained in Experiment 4. Right: 5 dB amplitude difference.*

Figure 20 shows the identificaton rate as a function of ΔF_0 (averaged over phase), for a target/competitor ratio of -15 dB (left) or -5 dB (right). Dotted line in Fig. 20 (left) represents data obtained in Experiment 4 for similar stimuli.

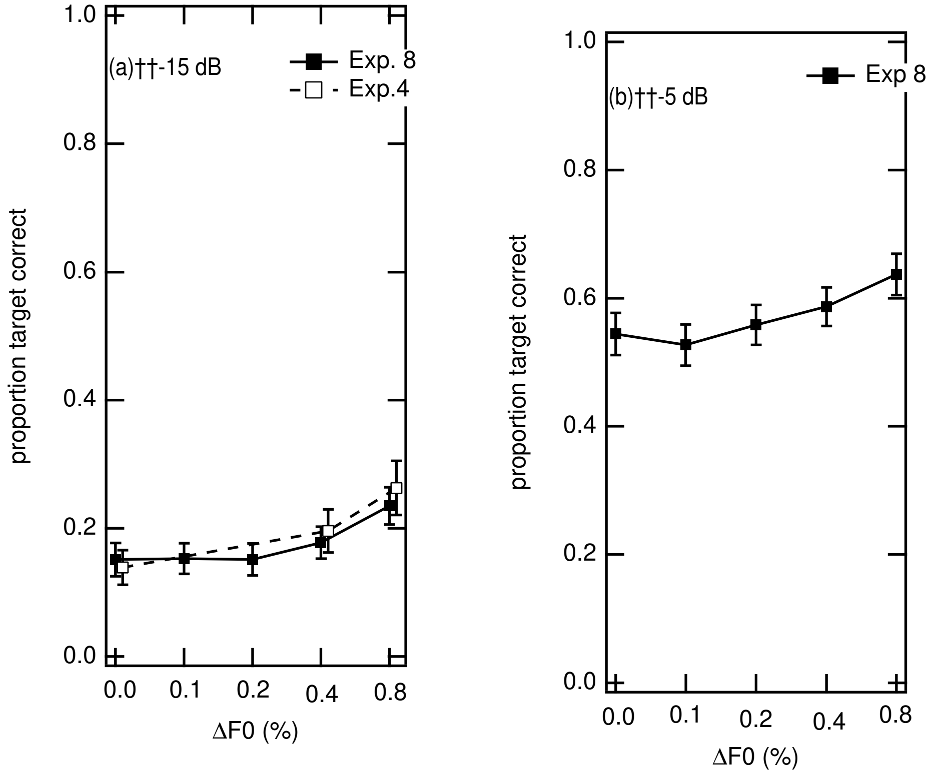


Fig. 20 *Target identification rate as a function of* ΔF_0*. Left: -15 dB target/competitor ratio. Right: -5 dB target/competitor ratio. Dotted line in represents data obtained in Experiment 4.*

10.4 Discussion

The data are overall consistent with those of Experiment 4. There are some phase effects, the most striking being the effect of a phase reversal at $\Delta F_0 = 0$ % when the target/competitor ratio = -15 dB. Both vowels have the same starting phase, so a phase reversal actually creates *dips* in the compound spectrum at the formants of the weaker vowel. This explains why identification is poor at a 180 degree phase shift. It is interesting to note that a ΔF_0 of 0.1 % is sufficient to abolish this penalty.

A ΔF_0 as small as 0.4 % is useful for segregation but effects of ΔF_0s smaller than that are indistiguishable from experimental noise and spurious phase effects. We have thus an estimate of the lower limit of ΔF_0s that support F_0-guided segregation.

11 Experiment 9: Effects of formant bandwidth

11.1 Introduction

Changes in formant bandwidth have a surprising small effect on vowel identification (Rosner and Pickering, 1994). They affect neither phoneme boundaries, nor the ability to classify spectra reliably as vowels, even if when bandwidths are doubled or quadrupled relative to their "normal" values. Identification is impaired only if the bandwidths are so wide that the spectrum is severely smeared.

On the other hand, in our concurrent vowel identification experiments we have found that subjects can detect the presence of vowels that are as much as 20 dB (de Cheveigné et al. 1997a) or 25 dB (exp. 3) weaker than a competing vowel, even without the aid of a ΔF_0. Having detected its presence, they can also identify it on a good proportion of trials[8]. In the absence of ΔF_0-based segregation cues, the auditory system presumably looks for evidence of the weaker vowel in the spectral region between the formants of its competitor.

Narrow formants should be beneficial to segregation in two ways: a) spectral features of the target should "emerge" more easily because they are sharp, and b) masking by the competitor should be less severe, because the valleys between its formants are deeper. The present experiment investigates the effects of making the bandwidth of either vowel (or both) either wider than "normal" by a factor of 2, or narrower than normal by a factor of two. Target-correct scoring of narrow/wide vs wide/narrow pairs allows us to decide between a) and b) above.

11.2 Methods

Single vowels were synthesized with formant bandwidths that were either one half ("narrow") or twice ("wide") their normal values, at F_0s of 124 and 132 Hz. Spectral envelopes for narrow and wide vowels are plotted in Fig. 21. The range is limited to 2500 Hz for clarity. Double vowels were created by adding single vowels with an amplitude ratio of 5, 15 or 25 dB. F_0s were the same (ΔF_0 = 0 %) or different(6 %). The bandwidths of both vowels could be either wide or narrow (n/n, n/w, w/n, w/w).

[8] We have not checked formally whether identification is above chance at the lowest amplitudes.

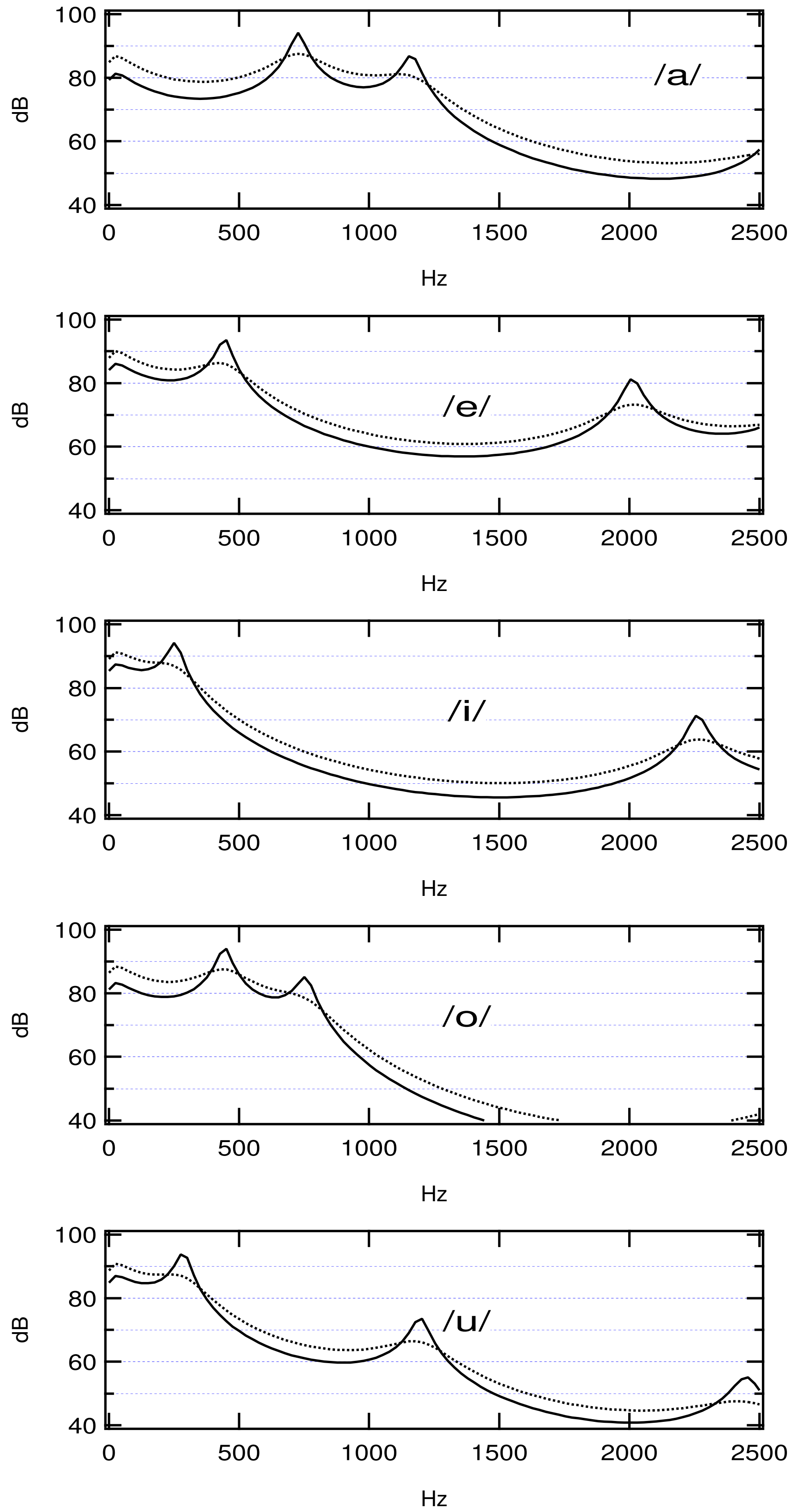

/a/
/e/
/i/
/o/
/u/
dB
Hz
100
80
60
40
0
500
1000
1500
2000
2500

Fig. 21 *Spectral envelopes of vowels. Dotted lines: wide formants (twice normal bandwidth). Continuous lines: narrow formants (half normal bandwidth). The abscissa is limited to the region of the lower two formants.*

There were (20 ordered pairs) x (3 amplitude ratios) x (2 ΔF_0s) x (2 F_0s) x (4 bandwidth combinations) = 960 double vowel conditions. Given the large size number of double vowels, single vowels were not included (pairs with an amplitude ratio of 25 dB are sufficiently similar to single vowels for the stimulus set to match the description made to the subjects: a set containing "both double and single vowels").

11.3 Results

Details of ANOVA are given in Appendix I. Anovas were performed separately at each amplitude ratio.

11.3.1 Number of vowels reported

15 dB amplitude ratio The pattern of results is most orderly at 15 dB amplitude ratio. Formant bandwidth may be treated as two orthogonal factors: bandwidth of the weaker vowel (wbw), and bandwidth of the stronger vowel (sbw). The ANOVA shows that the main factors of ΔF_0, wbw and sbw are highly significant, but their interaction is either not significant ($\Delta F_0 \times$ sbw, $\Delta F_0 \times$ wbw, $\Delta F_0 \times$ wbw $\times$ sbw) or marginally significant but small (wbw $\times$ sbw). Each factor has its own effect, independent from other factors, and the effects of all add up linearly.

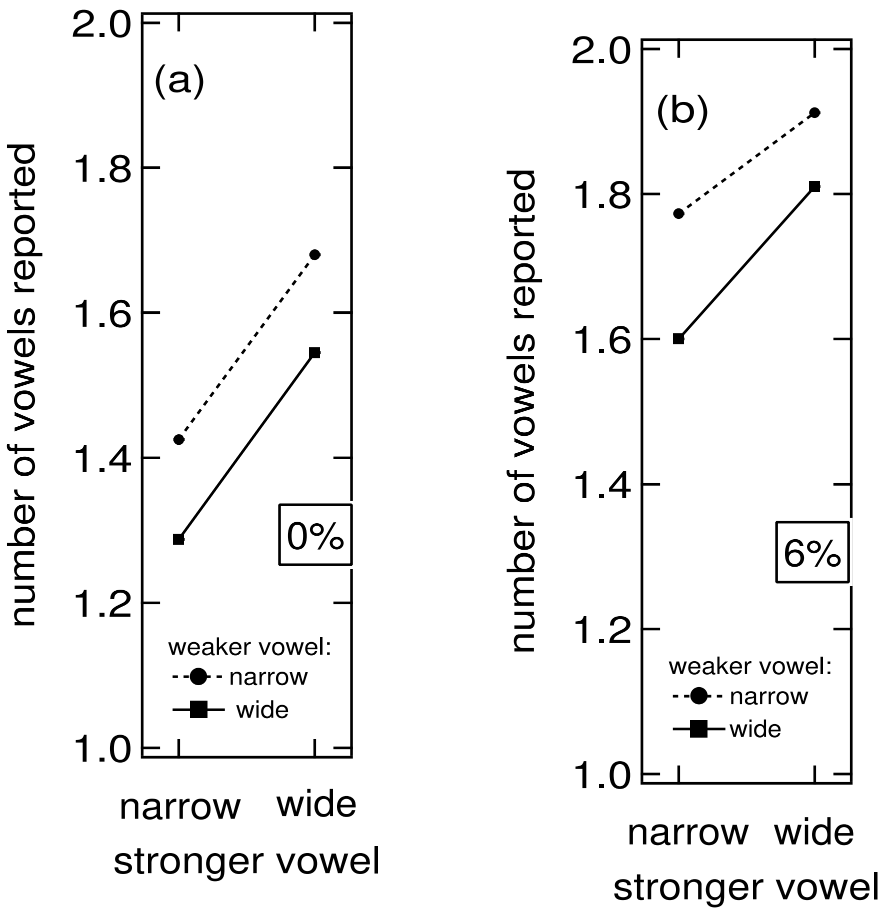


Fig. 22 *Number of vowels reported as a function of formant bandwidth of the stronger vowel (abscissa), and weaker vowel (dotted: narrow, continuous: wide). (a) ΔF_0 = 0 %, (b) ΔF_0 = 6 %.*

Results for ΔF_0 = 0 % are shown in Fig. 22 (a). The number of vowels reported is greater when the bandwidth of the *stronger vowel is wide rather than narrow*. It is also greater when the bandwidth of the *weaker vowel is narrow rather than wide*. The phase patterns of both vowels are identical, so their spectral envelopes add up linearly, and the spectral envelope of the sum is relatively easy to predict in terms of the formant positions, amplitudes, and shapes of the constituent vowels. If formants of the weaker vowel are narrow, they will emerge within the valleys between formants of the stronger vowel. It they are wide they will be less conspicuous. This explains the effect of formant bandwidth of the weaker vowel.

When formants of the stronger vowel are narrow, they emerge clearly from the compound spectrum and the stimulus sounds like a single vowel. When they are wide, the spectrum is less like that of a single vowel and a two-vowel response is more likely. This explains the effect of formant bandwidth of the stronger vowel.

At ΔF_0 = 6 % (Fig. 22 (b), the number of vowels reported is greater than at ΔF_0 = 0 %, as observed previously (de Cheveigné et al. 1997a,b). The effect of formant bandwidth is similar to that observed at ΔF_0 = 0 %. It is interesting to compare the magnitude of the effects of target bandwidth (0.14) and competitor bandwidth (0.21) with that of ΔF_0 (0.29). The effect of formant bandwidth is surprisingly large, although not quite as large as that of ΔF_0.

Formant bandwidth effects can also be compared with that of a 10 dB change in amplitude ratio, from 15 to 25 dB (0.16). Formant peak amplitude of a narrow-formant vowel is about 7 dB greater than that of a wide-formant vowel, and the valley between formants is about 5 dB deeper. If bandwidth effects were due only to the local amplitude differences they induce, one would have expected their effects to be smaller than observed.

Comparison between 5, 15 and 25 dB amplitude ratios The number of vowels reported for pairs with the same F_0 is plotted in Fig. 23 as a function of the bandwidths of the weaker and stronger vowel of each pair (notation: weak-vowel-bandwidth/strong-vowel-bandwidth), for amplitude ratios of 5, 15 and 25 dB. Effects of bandwidth are similar at 15 and 25 dB, but smaller and less orderly at 5 dB.

At ΔF_0 = 6 % the pattern is similar (Fig. 24. Effects at 5 dB are smaller still, possibly because of a ceiling effect.

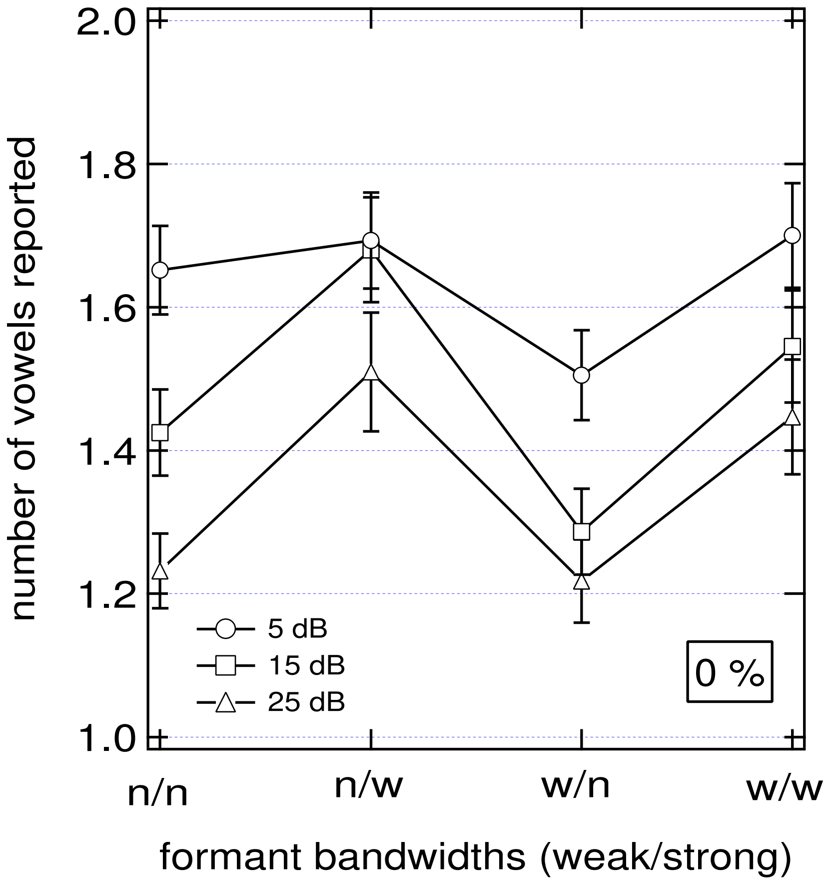


Fig. 23 *Number of vowels reported as a function of formant bandwidth of the stronger and weaker vowels (notation: weak-vowel-bandwidth/strong-vowel-bandwidth).* ΔF_0

= 0 %.

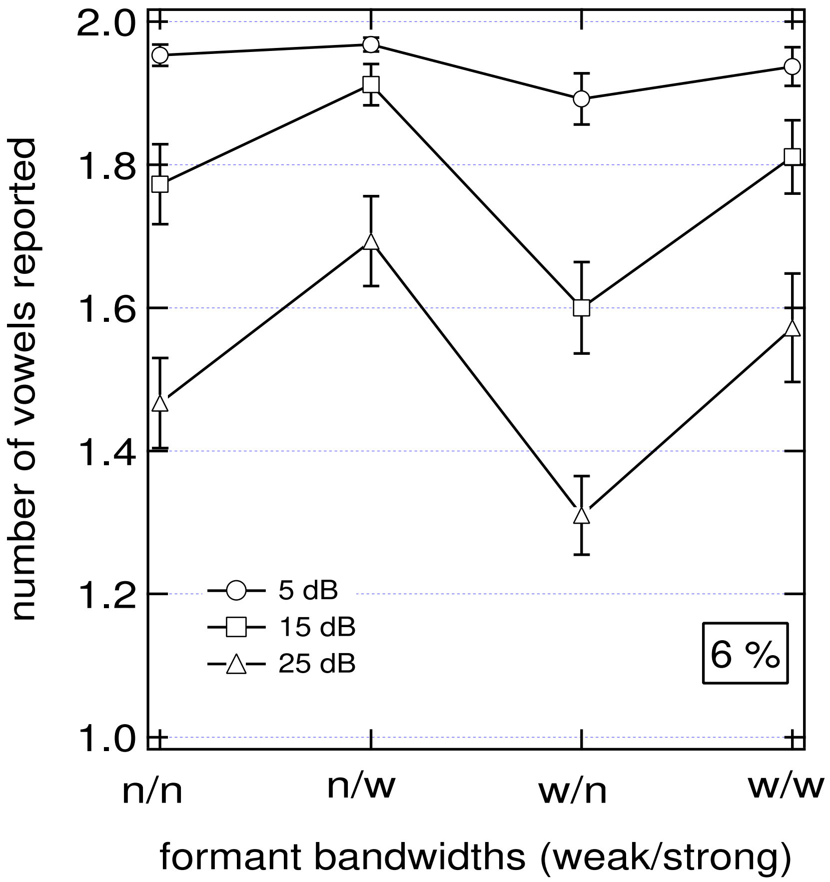


Fig. 24 *Number of vowels reported as a function of formant bandwidth of the stronger and weaker vowels (notation: weak-vowel-bandwidth/strong-vowel-bandwidth).* ΔF_0 *= 6 %.*

11.4 Identification

Identification is measured separately for each vowel in a pair, so we reason in terms of *target/competitor amplitude ratio* (rather than just amplitude ratio between vowels), *formant bandwidth of the target* (same as the weaker vowel if the target/competitor amplitude ratio is negative), and bandwidth of the competing vowel (same as the stronger vowel if the target/competitor amplitude ratio is negative).

-15 dB target/competitor ratio At -15 dB the pattern of identification appears quite orderly if bandwidth is treated as two orthogonal factors. The main effects of target bandwidth, competitor bandwidth and ΔF_0 are highly significant, and their interactions are marginally or non- significant, and in any case small.

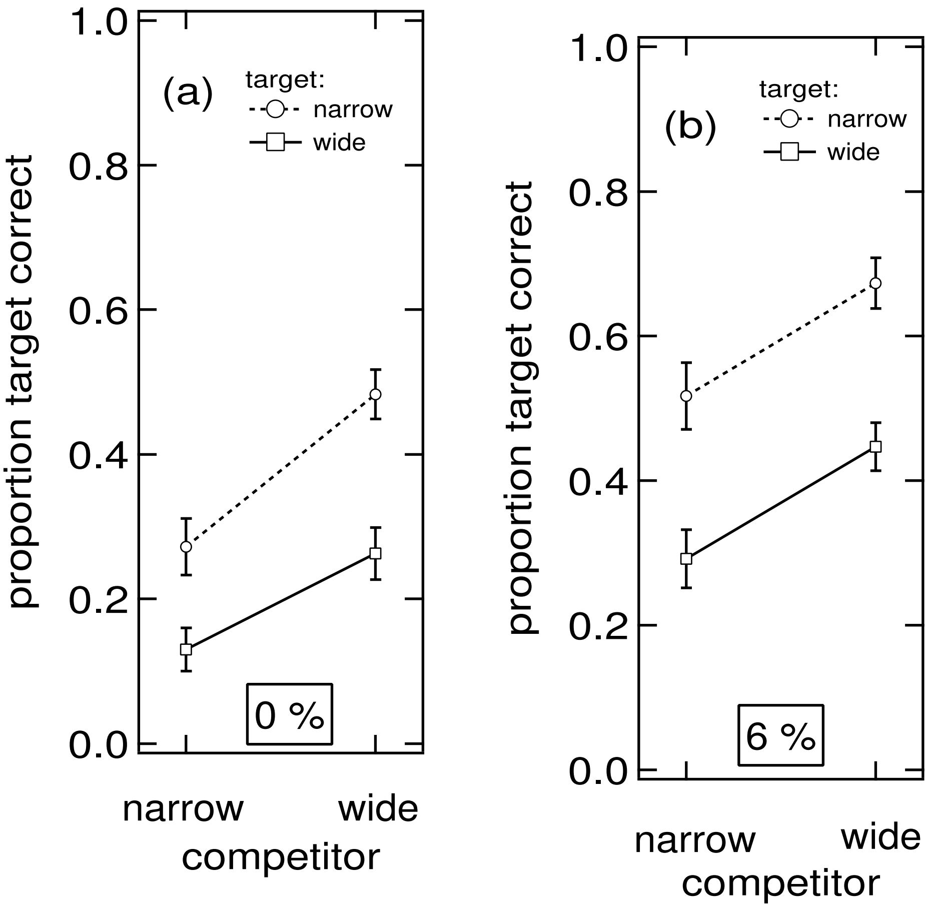


Fig. 25 *Identification rate as a function of formant bandwidth of the competing vowel (abscissa), and the target (dotted: narrow, continuous: wide). (a) ΔF_0 = 0 %, (b) ΔF_0 = 6 %.*

Consider first the pattern at $\Delta F_0 = 0$ % (Fig. 25 (a)). Identification is better if the target formants are narrow than wide. This is understandable, as narrow formants should allow the peaks of the weaker target vowel to better emerge within a spectrum dominated by the stronger competitor. Identification is also better if the formants of the competing vowel are wide rather than narrow. This is somewhat surprising, as valleys between formants of the competitor are *less deep* when bandwidths are wide. Masking should therefore be greater for target formants falling in these valleys. The result can be explained by assuming that competitor formants are "less competitive" when they are wide rather than narrow, and don’t disrupt identification of the target formants so much. It is also conceivable that identification is determined mainly by the tendency to report two vowels (when only one vowel is reported, the other is necessarily not identified). Replication of the experiment with a two-vowel forced response task might resolve this issue.

Identification is overall better at $\Delta F_0 = 6$ % (Fig. 25 (b)). The effects of bandwidth are similar to those observed at 0 %. At this amplitude ratio, the effects of target bandwidth, competitor bandwidth and ΔF_0 are independent and affect identification additively.

Other amplitude ratios For $\Delta F_0 = 0$ %, the identification rate is plotted in Fig. 26 as a function of target and competitor bandwidths, for each target/competitor amplitude ratio. The bandwidth effect is overall the same at all amplitudes, with evidence of a floor effect at low amplitudes (-25 dB) and a strong ceiling effect at high amplitudes (15 and 25 dB). The pattern is similar at $\Delta F_0 = 6$ % (Fig. 27), but identification rates are overall higher.

Restricting ourselves to the -15 and -5 dB levels, where effects of ΔF_0 and bandwidth are similar and regular, it is interesting to compare the magnitude of the effects of target bandwidth (0.20), competitor bandwidth (0.16), ΔF_0 (0.19) and a 10 dB step in amplitude ratio (0.28). One could argue that the approximately 7 dB greater formant peak amplitude of a narrow-formant vowel can account for the effect of target bandwidth. The approximately 5 dB lower valley amplitude of narrow formant competitors would lead us to expect an effect of the magitude similar to that which was observed (0.16), but of opposite sign.

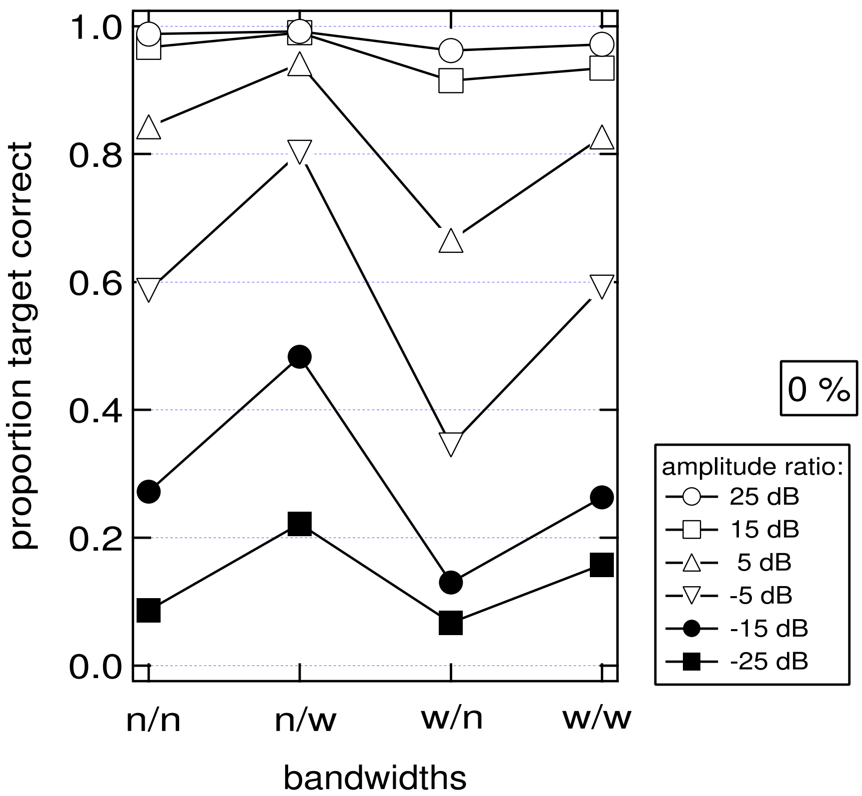


Fig. 26 *Target-correct identification rate as a function of formant bandwidth of the target and competitor (notation: target-bandwidth/competitor-bandwidth) for several target/competitor amplitude ratios.* $\Delta F_0 = 0$ *%.*

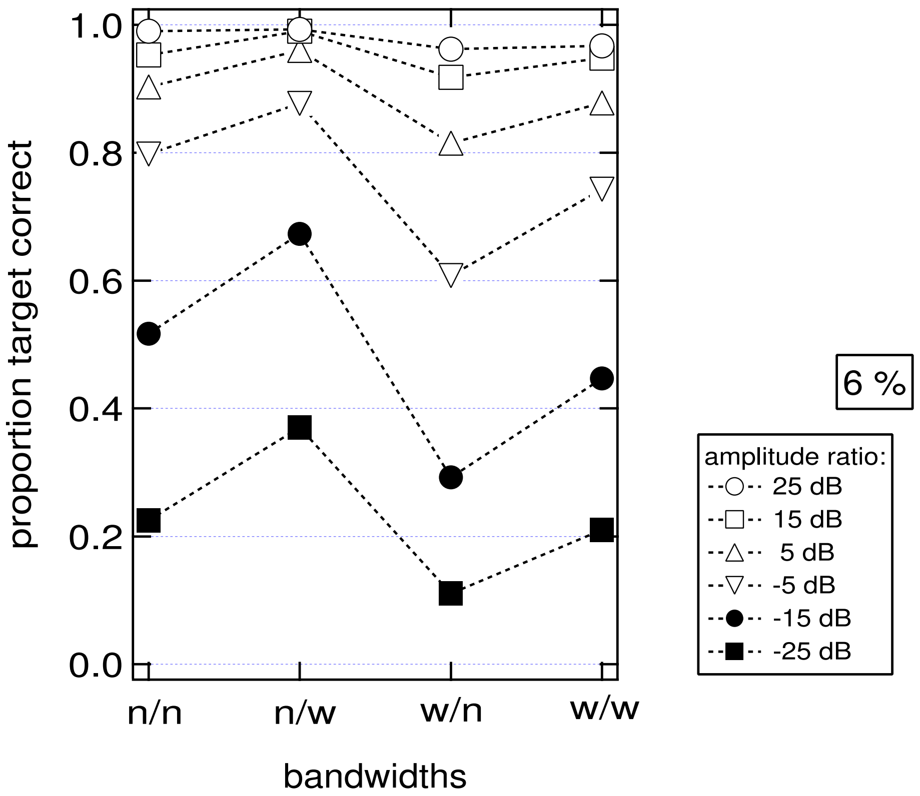


Fig. 27 *Target-correct identification rate as a function of formant bandwidth of the target and competitor (notation: target-bandwidth/competitor-bandwidth) for several target/competitor amplitude ratios.* ΔF_0 *= 6 %.*

11.5 Discussion

Formant bandwith has a strong effect on the number of vowels reported and the identification rate. The effects of target and competitor bandwidth are more or less independent, and of the same order of magnitude as the effect of 6 % ΔF_0, with which they combine more or less additively.

When both vowels have the same bandwidth, identification is usually better if this common bandwidth is narrow. However the effect is small. It is surprising that narrowing the competitor's formants does not enhance target identification, as the narrowing also deepens the valleys between the interfering formants and should help the target's formants to emerge (Fig. 28).

The strong effect of formant bandwidth on segregation is in stark contrast to its negligeable role in identification of isolated vowels. A voice with narrow formant bandwidths has a competitive advantage relative to voices with narrower bandwidths. It would be interesting to know whether individual voices that "stand out" have this characteristic, and whether formant bandwidth is a correlate of stressed or emotional speech. The detailed pattern as a function of vowel pair could be used to test models of vowel perception. We did not attempt to do so here.

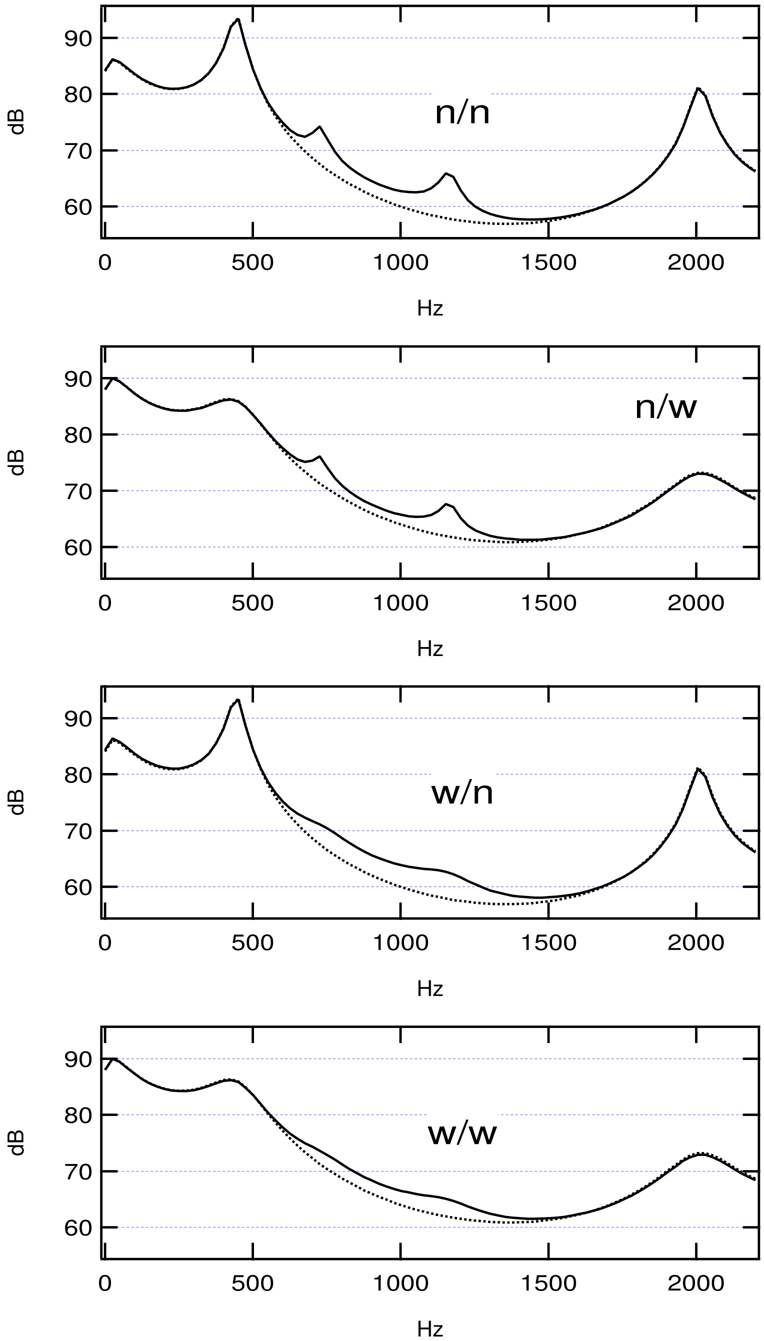


Fig. 28 *Spectral envelopes of an /i/ masker (dotted lines), and created by the addition to the masker of a weak /a/ (continuous line). Target/competitor amplitude ratio is 25 dB. Each plot is for a different combination of target/competitor formant bandwidth.* ΔF_0 *= 6 %.*

12 Experiment 10: In search of enhancement, part II.

12.1 Introduction

As pointed out in the Introduction to Experiment5, the hypothesis of harmonic enhancement (improved identification of harmonic targets) is attractive, and many models of concurrent harmonic sound separation (eg speech separation) models and methods are based on it. However there is little experimental evidence in favor of it: the harmonic state of targets has little effect on their identification.

One possibility is that the *continuity* of the target F_0 is a useful cue for segregation in noise. For example the target's F_0 might be estimated while the interference is weak, and that knowledge applied when the interference is strong. F_0 estimation is difficult when the interference is strong, so such the extrapolation of F_0 information might be useful. The mechanism would not be triggered with static stimuli, and this could explain why our previous attempts to find enhancement were unsuccessful.

In this experiment, we use dynamic stimuli that approximate sequential vowel pairs. We shall improperly use the term "diphthong" to describe such pairs (properly speaking, a diphthong is a vowel-like phoneme with a changing spectrum). The spectrum starts out resembling one vowel and terminates resembling another, after a smooth transition. The F_0 may stay the same during the transition, or it may change together with the spectral envelope. Each stimulus is partially masked by an interfering noise that starts at the beginning of the transition ramp. The subject's task is to report both vowels. The first is easy to hear, but the second vowel is severely masked, so identification should be strongly dependent on eventual segregation phenomena. For example a mechanism dependent on F_0 continuity might be more effective if the F_0 remains the same, than if it changes between the first and the second parts of the diphthong.

By using dynamic stimuli, this experiment takes one (modest) step in the direction of "real" speech.

12.2 Methods

Stimuli consisted of targets partially masked by noise. Targets were synthetic sequential vowel pairs, based on the five Japanese vowels. All pairs were allowed, including pairs with the same vowel. Stimulus duration was 270 ms with 20 ms raised cosine onsets and offsets. The 250 ms portion between -6 dB points was divided into four equal 62.5 ms segments. The spectral envelope was constant during the first and last segment, with a linear transition of envelope parameters (formant frequencies) during the middle two segments. The F_0 could be either constant throughout the stimulus (124, 128, 132 or 136 Hz), or else constant during the first and last segments, with a linear transition of about 6 % (124–>132 Hz, 136–>128 Hz, etc.) during the middle two segments.

The masker was 270 ms in duration and shaped with a single raised-cosine window. The masker was delayed by 62.5 ms relative to the target. The masker onset thus preceded the beginning of the target transition ramp by 10 ms, and the maximum of the masker envelope coincided with the end of the transition ramp. The masking level thus increased during the ramp. It decreased somewhat during the final steady-state portion of the target, but remained at a level sufficient to mask that portion effectively.

The masker was either periodic or noise-like. A periodic masker was made by summing all five vowels with equal RMS amplitude. Its F_0 could be 124, 128, 132 or 138 Hz. This F_0 was either the same as the final F_0 of the target (ΔF_0 = 0 %), or else different (ΔF_0 = 6 %). Noise-like maskers were made by shaping gaussian noise with the same spectral envelope. There were four different noise-like maskers, based on four different noise tokens.

There were thus three different masker states: noise-like, periodic (same F_0 during masked part), periodic (different F_0 during masked part). The masker RMS amplitude was 12 dB greater than that of the target for periodic maskers, and 2 dB for noise-like maskers. RMS amplitudes were calculated over the 270 ms duration of the target, and the 270 ms duration of the (delayed) masker.

There were (25 vowel pairs) x (4 starting F_0s) x (2 target F_0 patterns) x (3 masker states) = 600 stimuli.

If enhancement were effective, we would expect identification to be better for targets with static than modulated F_0s, at least when the masker is noise-like or periodic with ΔF_0 = 6 %. When ΔF_0 = 0 %, the static F_0 should be of no benefit.

12.3 Results

The results are illustrated in Fig. 29. When the masker was noise-like, or periodic with an F_0 that differed from that of the latter part of the target, identification rates were *not* better for a static than for a modulated target (they were, non-significantly, slightly worse). Our prediction based on the enhancement hypothesis was *not* confirmed. When the masker was periodic and an F_0 that was the same as that of the last segment of the target, identification was slightly but significantly better for unmodulated targets relative to modulated targets. This effect is not easy to explain. In the modulated case, the F_0 of the target differed from that of the masker during the ramp, and the auditory system might have gleaned information from that part to improve information. Such was not the case.

The significant difference between "S" and "D" conditions corresponds to the classic ΔF_0 effect. Comparisons between either of these and the "N" condition are meaningless, as target amplitudes are different. (The fact that identification was more or less the same for a +12 dB harmonic masker and a +2 dB noise-like masker indicates that the noise-like masker was considerably more disruptive.)

Once again, we have failed to find support for the harmonic enhancement hypothesis.

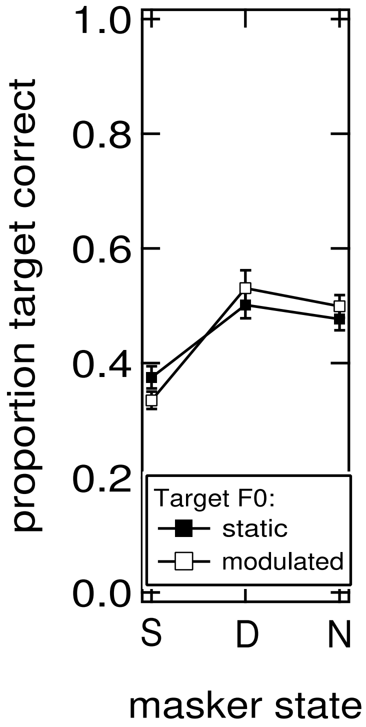


Fig. 29 *Identification rate as a function of masker state (S = harmonic, same F_0, D = harmonic, different F_0, N = noise like), for targets with a constant F_0 (filled symbols) or a ramped F_0 (open symbols).*

Acknowledgements

Analysis of this data was performed at the Laboratoire de Linguistique Formelle, CNRS / Universite Paris 7. Cecile Marin participated in the preparation of Experiment 1, and gathered data for French subjects, and she and Steve McAdams participated in discussions that led to these experiments. Hideki Kawahara contributed useful advice. Particular thanks go to Rieko Kubo who ran the experiments. For stimulus generation we relied heavily on John Culling's software (Culling 1996).

References

[1] 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.

[2] 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.

[3] Bregman, A. S. (1990). "Auditory scene analysis," Cambridge, Mass., MIT Press.

[4] Brown, G. J. (1992), "Computational auditory scene analysis: a representational approach," Sheffield, Department of Computer Science unpublished doctoral dissertation.

[5] Carlyon, R. P. (1991). "Discriminating between coherent and incoherent frequency modulation of complex tones," J. Acoust. Soc. Am. 89, 329-340.

[6] Carlyon, R., Demany, L., and Semal, C. (1992). "Detection of across-frequency differences in fundamental frequency," J. Acoust. Soc. Am. 91, 279-292.

[7] Carlyon, R. (1994). "Further evidence against an accross-frequency mechanism specific to the detection of frequency modulation (FM) incoherence between resolved frequency components," J. Acoust. Soc. Am. 95, 949-961.

[8] Cooke, M. P. (1991), "Modeling auditory processing and organisation," Sheffield, Department of Computer Science unpublished doctoral dissertation.

[9] Culling, J. F., and Darwin, C. J. (1993). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0," J. Acoust. Soc. Am. 93, 3454-3467.

[10] 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.

[11] 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 Comm. 14, 71-95.

[12] Culling, J. F., and Summerfield, Q. (1995). "The role of frequency modulation in the perceptual segregation of concurrent vowels," J. Acoust. Soc. Am. 98, 837-846.

[13] Culling, J. F. (1996). "Signal processing software for teaching and research in psycholinguistics under UNIX and X-windows," Behavior Research Methods, Instruments, and Computers 28, 376-382.

[14] de Cheveigné, A. (1993). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. 93, 3271-3290.

[15] de Cheveigne, A. (1995), "Experiments in vowel segregation," ATR-HIP technical report, TR-H-154.

[16] de Cheveigne, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (1995). "Sensitive experimental techniques for the study of sound segregation.", Proc. ASJ automn meeting, 373-374.

[17] de Cheveigné, A. (1996). "Ségrégation de voyelles simultanées: effets du niveau relatif et de la différence de F0.", Proc. Journées d'Étude de la Parole, 55-58.

[18] de Cheveigné, A., and Marin, C. (1996). "The segregation of frequency-modulated concurrent harmonic sounds," J. Acoust. Soc. Am. 100, 2718.

[19] de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (1997a). "Concurrent vowel segregation I: Effects of relative level and F0 difference," J. Acoust. Soc. Am. 101, 2839-2847.

[20] de Cheveigné, A., McAdams, S., and Marin, C. (1997b). "Concurrent vowel segregation II: Effects of phase, harmonicity and task," J. Acoust. Soc. Am. 101, 2848-2856.

[21] de Cheveigné, A. (1997c). "Concurrent vowel segregation III: A neural model of harmonic interference cancellation," J. Acoust. Soc. Am. 101, 2857-2865.

[22] Demany, L., and McAnally, K. I. (1993). "The perception of frequency peaks and troughs in wide frequency modulations," submitted to J. Acoust. Soc. Am.

[23] Ellis, D. (1996), "Prediction-driven computational auditory scene analysis," MIT unpublished doctoral dissertation.

[24] 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.

[25] Kashino, M., and Hirahara, T. (1995). "How many concurrent talkers can we hear out?", Proc. ASJ Automn meeting (in Japanese), 467-468.

[26] Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. 67, 838-844.

[27] Lea, A. (1992), "Auditory models of vowel perception," Nottingham unpublished doctoral dissertation.

[28] Marin, C., and A., d. C. (1997). "Rôle de la modulation de fréquence dans la séparation de voyelles.", Proc. SFA, .

[29] Marin, C., and McAdams, S. (1991). "Segregation of concurrent sounds. II: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width," J. Acoust. Soc. Am. 89, 341-351.

[30] Mellinger, D. K. (1991), "Event formation and separation in musical sound," Stanford Center for computer research in music and acoustics unpublished doctoral dissertation.

[31] McAdams, S. (1984), "Spectral fusion, spectral parsing, and the formation of auditory images," Stanford unpublished doctoral dissertation.

[32] McAdams, S. (1989). "Segregation of concurrent sounds. I: Effects of frequency modulation coherence," J. Acoust. Soc. Am. 86, 2148-2159.

[33] McKeown, J. D. (1992). "Perception of concurrent vowels: The effect of varying their relative level," Speech Communication 11, 1-13.

[34] Rosner, B. S., and Pickering, J. B. (1994). "Vowel perception and production," Oxford, Oxford University Press.

[35] Scheffers, M. T. M. (1983a), "Sifting vowels," Groningen unpublished doctoral dissertation.

[36] 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, Berlin, Mouton de Gruyter, 157-166.

[37] Summerfield, Q., and Culling, J. F. (1992). "Auditory segregation of competing voices: absence of effects of FM or AM coherence," Phil. Trans. R. Soc. Lond. B 336, 357-366.

[38] Summerfield, Q., and Culling, J. F. (1992). "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency.", Proc. 124th meeting of the ASA, 2317(A).