Acoustical Society of America and Acoustical Society of Japan
Third Joint Meeting
Honolulu, Hawaii, 2-6 December 1996.

THE SEGREGATION OF FREQUENCY-MODULATED CONCURRENT HARMONIC SOUNDS.

Alain de Cheveigné (CNRS/Université Paris 7, and ATR Human Information Processing Research Laboratories, Japan), alain@linguist.jussieu.fr
Cécile Marin (IRCAM, Paris), marin@ircam.fr

Abstract

Subjects were presented with pairs of concurrent vowels, and were requested to report one or two vowels for each stimulus. Vowels within a pair had either the same average fundamental frequency (F_0), or F_0s different by 3 or 6%. The F_0 of each vowel was either constant, or else frequency-modulated (FM) at a rate of 4Hz with a peak excursion of 3%. Vowels were either both harmonic or both inharmonic (partials randomly displaced by -3%, 0% or 3%). One vowel within each pair was always stronger than the other by 15 dB, to reduce ceiling effects for the weaker vowel ("target"). The proportion of trials for which the weaker vowel was identified was recorded, together with the proportion of two-vowel responses. The experiment was expected to be sensitive to several mechanisms by which the identification of a target vowel might be affected by frequency modulation of that vowel, or of the competing vowel. The two-vowel response and identification rates were greater when the average F_0s differed by 3% or 6% than 0%. Neither measure was affected by the presence of FM within the stimulus, or the modulation of the target or competing vowel. Identification was slightly less good when both vowels were modulated coherently (or static) than when modulation was incoherent. However incoherent modulation also implied greater maximum ΔF_0, and one cannot rule out the possibility that this caused better identification. Overall, results agree with previous studies that suggest that FM has little or no effects beyond those attributed to FM-induced differences in fundamental frequency.

A. Introduction

Fourier analysis allows harmonic sounds, such as voiced speech, to be represented as a number of discrete components equally spaced in frequency. This has led to two expectations concerning the way such sounds are perceived in the presence of competing sounds. 1) The regular pattern of components, similar to the visible texture of a surface, should contribute to the perception of a harmonic sound that is partially occluded by interference. Components that share the regularity belong to the target, and any remaining components may be attributed to interference. 2) When the fundamental frequency of the sound changes, all its components shift in parallel on a logarithmic scale. This common movement should further aid the auditory system to group together components belonging to the target, and exclude components of other sounds that don't follow the same movement. Differently modulated sounds should be easy to segregate and perceive as separate entities. This idea, originally put forward by Helmoltz (1885), has become especially appealing within the framework of Auditory Scene Analysis (Bregman, 1990), as a prime example of grouping by "common fate".

Experimentation with mixtures of harmonic and inharmonic vowels has failed to confirm the first expectation. According to that expectation, a target vowel mixed with a competing vowel should be less easy to hear if it is inharmonic rather than harmonic. This is not the case. The target is however easier to identify if the *competing* vowel is harmonic, rather than inharmonic (Summerfield and Culling 1992, Lea 1992, de Cheveigné et al. 1995, 1997a,b). To take a visual analogue, it is as if objects of any texture were visible against a striped background, but striped objects themselves were not particularly conspicuous, whatever the texture of the background. Indeed, if a target is harmonic and has the same F_0 as a competing sound, the benefit of harmonicity of the competitor is lost. This explains the low identification rates observed at ΔF_0=0% in double-vowel experiments (Scheffers 1983).

Likewise, the second expectation has failed to be confirmed experimentally. Identification of a harmonic target may indeed be enhanced when it is modulated incoherently with a harmonic masker, however this occurs only if modulation introduces F_0 differences (ΔF_0) that would not have occurred with coherent modulation. If for example FM is superimposed on a static F_0 difference, so that the average ΔF_0 remains the same, there is no difference between coherent modulation and incoherent modulation (Summerfield 1992). Subjects do not seem to take advantage of the common movement of target partials, or its incoherence with the background. Carlyon (1991) showed that subjects cannot detect coherent modulation of two partials, unless their frequencies form a harmonic relation that incoherent modulation would disrupt. He argued that if the cue cannot be detected, it cannot be used to group partials belonging to a target. Effects reported by Wilson et al. (1990) and Cohen and Chen (1992) seemed to contradict this conclusion, but Carlyon (1994) showed that they should be attributed to differences in harmonicity, combination tones, or beats that co-occur with incoherent FM, rather than to a mechanism that compares modulation patterns of partials accross frequency. Nevertheless, Furukawa and Moore (1996) found that detection of the *presence* of FM on pairs of pure tones (as opposed to detection of the coherence of modulation between them) benefits from accross-frequency integration of information, and depends on whether the modulation is coherent or not. This implies that FM coherence should be detectable between components just above their modulation threshold.

Summerfield (1992) reasoned that FM effects might be observable when harmonicity cues due to FM-induced ΔF_0 are reduced. He measured identification thresholds for synthetic vowels in the presence of vowel-like maskers. For static harmonic targets and maskers, thresholds were lower for different F_0s than for same F_0s, but modulating either the masker or the target gave no further benefit, as noted earlier. For inharmonic targets and maskers, however, a modulated target with a static competitor did have a lower threshold than a) a *static* target, whatever the competitor, b) a *modulated* target with a *modulated* competitor. In the latter case it made no difference if the modulation of the competitor was coherent with that of the target or not. This appears to be a genuine FM effect: harmonicity cues were absent or weak, as evident in the fact that static targets with modulated competitors showed no lowering of thresholds. However it is not what one would expect on the basis of grouping by common fate of target partials: a) A static target mixed with a modulated competitor reaps no benefit, nor does a modulated target with a modulated competitor, despite incoherence of modulation. b) The target components do not need to be modulated coherently for the effect to occur (Culling and Summerfield 1995).

Summerfield (1992) suggested that the target modulation effect could be explained by FM-induced amplitude modulation (AM) arising on the skirts of peripheral auditory filters near the formant peaks of the target. AM would signal the presence of the formants among relatively stronger static correlates of the masker. Such is not the case for a modulated masker, because the strong AM pattern that occurs all over the spectrum is hardly affected when the weaker target is present. However, Marin and McAdams investigated in detail the nature and effectiveness of FM-induced AM cues. Subjects matched the formant frequency of a harmonic one-formant sound embedded in a harmonic background. AM cues were restricted to those produced by partials moving along the skirts of peripheral filters, by using "flat" formants comprised of two or three adjacent harmonics, and maskers for which the corresponding harmonics were absent. Using a computer simulation of peripheral filtering, they estimated the amount of FM-induced AM produced within each channel at threshold. They compared this to the amount of modulation produced at threshold in a similar task in which formants were amplitude-modulated rather than frequency modulated. The amount of modulation produced in the FM task was far below that required in the AM task, from which they concluded that FM-induced AM was *not* the cue used by their subjects. They suggested that the cue might involve beat patterns caused by the interaction of adjacent components *mistuned from a harmonic ratio*, rather than simply close in frequency. Such interactions do not modulate the envelope of filter outputs.

Thus it appears that the increased intelligibility observed when competing speakers have different intonation patterns (Cherry 1953, Brokx and Nooteboom 1982) results from two mechanisms:

1) Instantaneous F_0 differences allow harmonic portions of an interfering voice to be suppressed (Summerfield and Culling 1992, de Cheveigné et al. 1995, 1997a,b). Without them, targets would be suppressed together with the masker. FM of the masker is actually detrimental in reverberant conditions, because periodicity is smeared (reflected sounds arriving at the ear are produced at various phases of the FM) (Culling et al., 1994). Modulated *targets* are not adversely affected by reverberation, which reinforces the conclusion that F_0-guided segregation does not operate by enhancement of harmonic targets, but rather by suppression of harmonic interference.

2) Frequency modulation of a target may induce amplitude modulation or beats in spectral regions important for the identification of a target. This is of benefit only if the masker is static (otherwise modulation would occur in all regions) and inharmonic (otherwise the masker could be suppressed by more effective harmonic cancellation). FM-induced AM arises whether partials of the target are modulated coherently or not, which explains why the benefit is the same in both cases (Culling and Summerfield 1995).

In addition to these two mechanisms that affect identification directly, FM affects perception in other ways. McAdams (1989) and Marin and McAdams (1991) presented subjects with "triads" of vowels synthesized with F_0s in a ratio of 1.33, and asked them to rate the "prominence" of each vowel. Vowels were judged to be more prominent when they were modulated, particularly if the other two vowels were *not* modulated. When the two competing vowels were modulated, it made no difference whether the modulation was coherent between all three vowels, or incoherent. This pattern resembles that observed by Summerfield (1992): the mistuned *pair* of competing vowels in Marin and McAdams' experiment is similar to the inharmonic masker in Summerfield's experiment, and prominence ratings and identification rates depend on target and masker modulation in a similar fashion, as noted by Culling and Summerfield (1995). One difference is that, whereas masker modulation eliminated the effect of target modulation in Summerfield's experiment, it only reduced it in Marin and McAdams's.

There is also evidence that FM may affect the number of sources perceived (Gardner et al. 1989, McAdams 1984). Paradigms such as concurrent-vowel identification tasks (Scheffers, 1983) are not sensitive to multiplicity cues, because subjects are required to report two vowels whatever the stimulus. A modified task that allows one or two responses for each stimulus allows the task to be sensitive to such cues (de Cheveigné et al. 1997a,b). This provided the motivation for the present experiment which was designed to be sensitive to a variety of hypothetical mechanisms by which FM might affect the perception of speech.

B. Experiment

In this section we first outline some differences between our paradigm and the conventional concurrent vowel identification paradigm. We then describe our stimuli and synthesis methods, and the set of conditions selected for the experiment. Finally, we list a series of hypothetical mechanisms by which FM might affect perception of concurrent vowels. Each is associated with specific "predictions" concerning the outcome of the experiment.

1. Design

The experiment was based on the conventional concurrent vowel identification paradigm (Scheffers, 1983; Assmann and Summerfield, 1990; Culling and Darwin 1993), with three modifications designed to make it more sensitive and versatile (de Cheveigné et al. 1997a,b):

Stimuli were scored twice, once for each vowel within the pair (Lea 1992). Each vowel in turn was nominated the "target", and the outcome of its identification was recorded according to the nature of that vowel and that of the other vowel (the "competitor"). Roles of target and competitor were then reversed, leading to two scores per stimulus. This procedure yields "constituent-correct" rates instead of the conventional "combination-correct" rates that count trials for which *both* vowels were correctly identified. It is conceivable that some cue might affect each vowels in a different direction, and lead to a null effect on combination-correct scores. Or else one vowel might be systematically mis-identified, masking effects affecting the other. Constituent-correct scores remain sensitive in both cases. They also allow effects specific to asymmetric configurations (for example a modulated vowel with an unmodulated competitor, etc.) to be investigated.

The stimulus set contained both single and double vowels, and subjects were requested to report one or two vowels on each trial (rather than the two-vowel response typically required in such experiments). Both the identification rates and the two-vowel response rates are thus sensitive to cues that signal the presence of *multiple sources* within the stimulus. The task is typical of natural situations where the number of sources to attend to is not known a priori.

An amplitude mismatch of 15 dB was introduced between vowels, to enhance the sensitivity of identification of the weaker vowel to conditions of interest. This is also typical of natural situations, in which competing voices rarely share the same level. Identification scores for the stronger vowel are usually perfect and therefore discarded.

2. Subjects

There were 10 japanese and 12 french subjects.

3. Stimuli

Synthetic vowels represented the five vowels /a/, /i/, /e/, /o/, /u/ of French (french subjects) or of Japanese (japanese subjects). Formant frequencies and bandwidths are listed in Table 1. Vowel tokens were 270ms in duration with onset and offsets shaped by 20 ms raised-cosine ramps (250 ms between -6dB points). Fundamental frequencies were either constant at 124, 128 or 132Hz, or sinusoidally modulated around these values with a peak excursion of 3 %. Double-vowels could thus be synthesized with ΔF_0s of approximately 0%, 3% and 6%. The modulation rate was 4 Hz, and the period of modulation (250 ms) was therefore equal to the effective duration of the stimulus. The modulation had either a "u" shape (cosine phase), or a "n" shape (opposite phase). The unmodulated (static) state is noted "_". Tokens were also synthesized with partials following an inharmonic series obtained by randomly shifting the frequency of each partial of a harmonic series by -3%, 0% or 3%. These conventions ensure, at least for static stimuli, that all beats occur at a multiple of 4Hz, and thus complete an integer number of periods within the 250ms stimulus. The long-term spectrum thus does not depend on starting phase (de Cheveigné et al. 1997b). Additional components introduced by FM also have 4Hz spacing, so this property is preserved. All stimuli started out with the same "random" phase (pattern "R" used by de Cheveigné et al. 1997b).

The "F_0" of an inharmonic vowel is defined as that of the harmonic vowel before shifting. The same inharmonic pattern was used for all vowels and all F_0s. Vowels were synthesized at a sampling rate of 16 kHz in double floating-point precision, using an implementation of Klatt's synthesizer (Klatt 1980, Culling 1996). After synthesis, all single vowel tokens were scaled to the same RMS value and stored on disk in floating point format. Double vowels were created "on the fly" during the experiment by taking two single vowel tokens, scaling one by a factor of 15 dB, adding them, and scaling the sum to a standard RMS value. The result was converted to short integer and output diotically to earphones at a level between 63-70 dB as measured by a Bruel&Kjaer artificial ear (sound level meter type 2231, half-inch microphone type 4134).

Table 1. Formant frequencies and bandwidths used for synthesizing Japanese (Hirahara and Kato, 1992) and French (Boe et al. 1993) vowels.

	/a/		/e/		/i/		/o/		/u/		BW
	jp	fr	jp	fr	jp	fr	jp	fr	jp	fr	
F1	750	742	469	395	281	252	468	399	312	276	90
F2	1187	1266	2031	2027	2281	2202	781	829	1219	733	110
F3	2595	2330	2687	2552	3187	3242	2656	2143	2469	2171	170
F4	3781	3457	3375	3438	3781	3937	3281	3445	3406	3506	250
F5	4200	4200	4200	4200	4200	4200	4200	4200	4200	4200	300

4. Conditions

The stimulus set contained both single vowels and double vowels. The single vowels conditions were those described in the previous section. For double vowels, a subset of 16 conditions of ΔF_0 and modulation was selected from a total of 108 possible conditions ((3 ΔF_0s) x 2 (target harmonicies) x (3 target modulation patterns) x (2 ground harmonicities) x (3 ground modulation patterns)). The subset was designed to allow all the hypotheses to be tested while keeping the number of stimuli within reasonable limits. In all cases, both vowels within a pair have the same harmonicity state, either harmonic (H) or inharmonic (I). The notation x/y indicates that the target (weaker vowel) is modulated in state x while the competitor (stronger vowel) is in state y. For example "_/m" signifies a static target with a modulated competitor. When necessary, the shape of modulation ("n" or "u") may be specified, as well as the ΔF_0 and harmonic state. For example "H3_/n" specifies that the competitor's modulation waveform was shaped as an "n", both vowels were harmonic, and the ΔF_0 was 3%. Conditions are:

- Both harmonic, ΔF_0=0% (H0). Modulation: _/_, n/n and u/u.
- Both harmonic, ΔF_0=3% (H3). Modulation: all combinations of _ , n and u (_/_, n/n, u/u, n/u, u/n, n/_, u/_, _/n, _/u).
- Both harmonic, ΔF_0=6% (H6). Modulation: _/_.
- Both inharmonic, ΔF_0 = 3% (I3). Modulation: _/_, _/n, n/_.

The reasons for chosing this particular subset will become apparent in Sec. 5, that describes hypothetical mechanisms of sensitivity to FM. These 16 ΔF_0 and modulation conditions were crossed with all 20 ordered pairs of distinct vowels. They were also crossed with 2 values of absolute F_0 (at ΔF_0=0) or absolute F_0 order (at $\Delta F_0 \neq 0$). In other words, each condition was repeated with the assignment of F_0s to vowels reversed. From previous experiments we expected no effect of absolute F_0, but we wished to avoid the possibility that a subject might associate a particular condition with a particular absolute F_0. An exception to this rule was made for conditions 3_/n and 3n/_: the unmodulated vowel (_) always had the lower F_0, so the two vowels started and stopped at the same F_0. For 3_/u and 3u/_ the unmodulated vowel was the higher one, so both vowels again started and stopped together. The justification of this exception is explained in Section XXXX (hypotheses 7 and 8).

At a nominal ΔF_0 of 3%, the instantaneous ΔF_0 for conditions (_/_, n/n, u/u) is constant and equal to 3%. For conditions (n/_, u/_, _/n, _/u), the instantaneous ΔF_0 varies between 0 and 6%, but its average remains equal to 3%. In conditions (n/u, u/n), the F_0 tracks cross and the average instantaneous ΔF_0 is slightly greater.

The stimulus set thus comprised 640 double vowel conditions. To these were added (3 F_0s) x (3 modulation patterns) x (2 harmonicity states) x (10 repetitions) = 180 single vowels. The single vowels allow us to control for effects specific to the vowels themselves rather than their combinations. Their relatively large number ensures that the stimulus set is consistent with the description made to the subjects. The total stimulus set comprised 820 stimuli, presented in a single session.

Subjects were seated in a sound treated booth or room (at ATR for japanese subjects, at IRCAM for french subjects). They were seated in front of a computer screen that gave instructions and prompts, and they responded by means of the keyboard. They were informed that each stimulus could contain one or two vowels, and they were instructed to press one or two keys (a, e, i, o, u) according to what they heard. Pressing the return key recorded this response and triggered a new presentation. Subjects were allowed to interrupt the session at will by pressing "Q" instead of an answer. In this case, the stimulus preceding the pause was presented again after the pause. Sessions took between 40 and 90 minutes to complete.

5. Hypotheses

The stimulus set allows a range of hypotheses to be tested. Some are mutually exclusive. Others contradict current knowledge, but were nevertheless included according to the principle of "suspension of disbelief", in the hope that the experiment may reveal unexpected effects. For each hypothesis we present the experimental outcomes that it predicts.

1) *The presence of FM within a stimulus increases the impression of multiple sources.* This assumption is reasonable since FM can be interpreted as a form of inharmonicity. Double-vowel responses should be evoked more often for stimuli that contain modulated constituents than for those that don't. If we denote "N(condition)" as the proportion of double-vowel responses for that condition:

N(H_) < N(Hn, Hu), N(I_) < N(In, Iu) (single vowels)
N(H0_/_) < N(H0n/n), N(H0u/u)
N(H3_/_) < N(H3n/n, H3u/u, H3n/u, H3u/n, H3n/_, H3u/_, H3_/n, H3_/u)
N(I3_/_) < N(I3n/_, I3_/n)

2) *The presence of FM within a stimulus decreases the impression of multiple sources*. This assumption is reasonable if one assumes that FM reinforces the cohesion of coherently modulated partials. This might be the case particularly for inharmonic vowels that lack cohesion. Predictions are opposite those of the previous hypothesis. If we add the restriction that modulation must be coherent among *all* partials (not just within each constituent vowel), then the predictions are reduced to:

N(H_) > N(Hn, Hu), N(I_) > N(In, Iu) (single vowels)
N(H0_/_) > N(H0n/n), N(H0u/u)
N(H3_/_) > N(H3n/n), N(H3u/u)

3) *Identification is better if the competitor is static rather than modulated* For example this might be the case if a) segregation occurs by harmonic cancellation of the competing vowel, and b) the F_0 estimation process is somewhat sluggish (Moore and Sek, 1996):

I(H3n/_) > I(H3n/u, H3n/n)
I(H3u/_) > I(H3u/n, H3u/u)
I(H3_/_) > I(H3_/n, H3_/u)

4) *Identification is better if the competitor is modulated rather than static*. For example this might be the case if estimation of the competing vowel's F_0 were easier when it is modulated rather than static. Predictions are the opposite of the previous hypothesis.

5) *Identification is better if the target is static rather than modulated.* For example this might be the case if segregation occured by harmonic enhancement of the target, and if estimation of its F_0 were easier when it is static rather than modulated:

I(H3_/n) > I(H3n/n, H3u/n)
I(H3_/u) > I(H3u/u, H3n/u)
I(H3_/_) > I(H3n/_, H3u/_)

6) *Identification is better if the target is modulated rather than static.* Predictions of this hypothesis are opposite those of the previous one.

7) *Identification is better if the competing vowel's modulation has an "n" shape rather than a "u" shape*. For example this might be the case if segregation occured by harmonic cancellation, and if F_0 estimation were easier or more precise for a peak of modulation than for a dip. Pitch discrimination has been shown to be better at peaks than at dips of wide modulation patterns (Demany and Clement, 1995). Modulation might also be less detectable for one direction than another (Carlyon and Stubbs, 1989). Actually, an "n"-shaped modulation has one peak, but also two half-dips, one at each extremity. Better F_0 estimation at the peak might be compensated by worse estimation at each extremity. To avoid this, the stimuli were designed so that, in the _/n and _/u conditions, the static and modulated vowels started and stopped with the same F_0. For example, in the _/n condition, the static vowel (_) always had the lower F_0, and the modulated vowel (n) the higher F_0. In the _/u condition the F_0s were

reversed, and vowels again started and stopped at unison. In the absence of a ΔF_0, differences in shape at the extremities should have no effect.

9) *Identification is better if the target modulation has an "n" shape rather than an "u" shape*. This hypothesis is analogous to the previous one, replacing the assumption of harmonic cancellation by that of harmonic enhancement.

In addition to hypothetical effects related to these hypotheses, there are others that we expect from previous results.

10) From results of Summerfield (1992) we expect modulated targets with a static competitor to be better identified than static targets with any competitor, or modulated targets with a modulated competitor, for example:

I(H3n/_) > I(H3n/n, H3n/u, H3_/_, H3_/n, H3_/u)

11) From results of conventional double-vowel experiments, we expect identification to be better when there is a ΔF_0 between static or coherently modulated vowels:

I(H0_/_) < I(H3_/_, H6_/_)

I(H0n/n, H0u/u) < I(H3n/n, H3u/u)

The 0% and 6% static harmonic conditions were used in previous experiments (deCheveigné et al. 1997a,b) and may serve as a comparison basis.

12) From previous experiments on harmonicity (de Cheveigné et al. 1995), we expect inharmonicity at 3% nominal ΔF_0 to degrade indentification relative to equivalent harmonic conditions:

I(I3_/_) < I(H3_/_)

I(I3n/_, I3_/n) < I(H3n/_, H3_/n).

C. Results

1. Analysis

Identification rates for double-vowels were submitted to a repeated-measures analysis of variance with a single 16-level factor. Probabilities reflect, where necessary, an adjustment of the degrees of freedom by a factor that corrects for the inherent correlation of repeated measurements (Geisser and Greenhouse, 1958). The effect was significant [$F(21,441)=47.32$, $p<.0001$, GG=0.108, and will be described in detail in in the following paragraphs. Identification of single vowels was essentially perfect.

Two-vowel response rates for both single and double-vowel conditions were likewise submitted to a repeated-measures analysis of variance with a single 22-level factor (single vowel conditions were included, as they evoked an appreciable number of two-vowel responses). The single effect was significant [$F(15,315)=50.04$, $p<.0001$, GG=0.233], and will also be described in the following paragraphs.

2. ΔF_0

The two-vowel response rate for static harmonic conditions is plotted in Fig. 1 (a) as a function of ΔF_0, together with data from equivalent conditions of a previous experiment (de Cheveigné et al. 1997b). The number of responses increases between 0 and 3% [$F(1,441) = 44.13$, $p=.0002$] and then asymptotes (the difference between 3% and 6% just misses significance [$F(1,441)=6.46$, $p=.059$]). The target-correct identification rate is plotted in Fig. 1 (b). It increases between 0% and 3% [$F(1,315)=90.14$, $p<0.0001$] and from there to 6% [$F(1,315)=6.52$, $p=0.044$]. Effects for both measures are smaller than in the previous experiment.

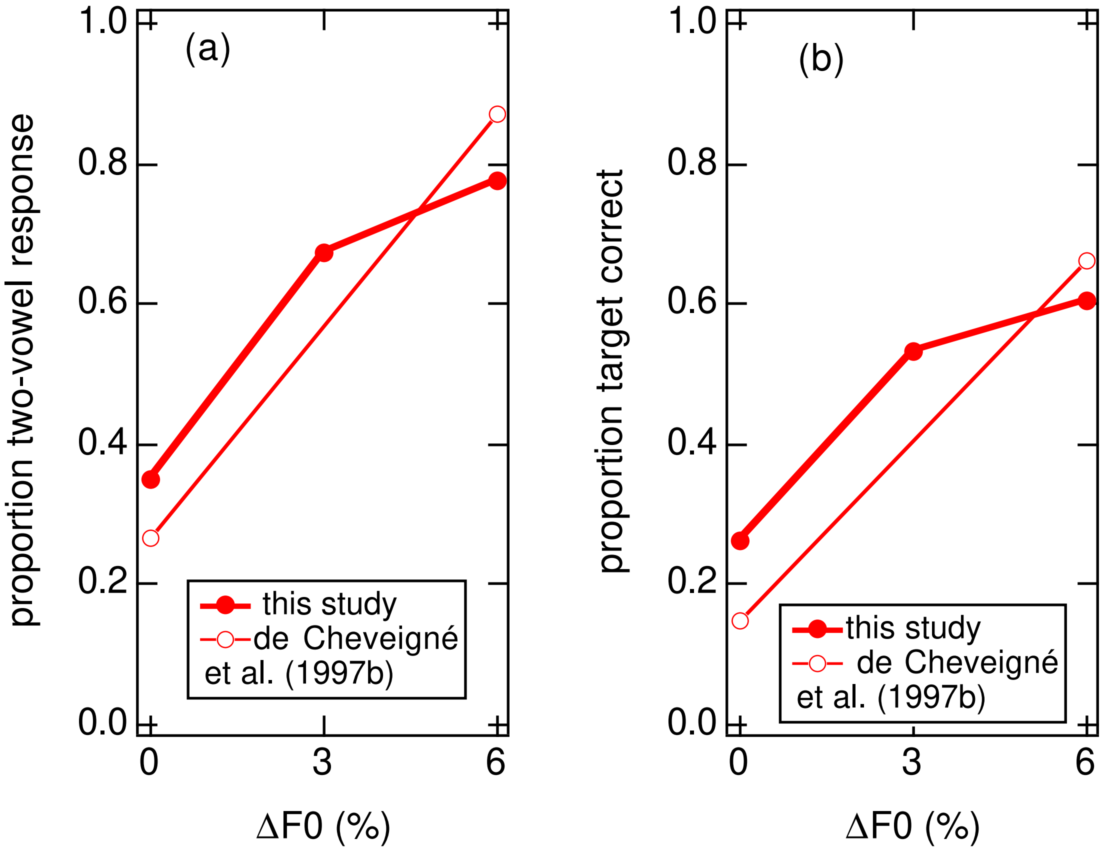

Fig. 1 (a) Proportion of two-vowel responses and (b) proportion of correct responses for the target (weaker) vowel, as a function of ΔF_0 for static harmonic vowel pairs in this study (filled symbols) and a previous study with equivalent conditions (open symbols) (de Cheveigné et al. 1997b)

3. Harmonicity

Comparisons can be made between conditions that exist with both harmonic and inharmonic vowels. For single vowels, the proportion of two-vowel responses is plotted in Fig. 2 (a) (triangles). It is greater for vowels that are harmonic rather than inharmonic [$F(1,441)=48.41$, $p=0.0016$], as observed in a previous experiment (de Cheveigné et al. 1997b). In contrast, for double vowels (conditions _/_, n/_, _/n at $\Delta F_0=3\%$), the two-vowel response rate is *smaller* when both vowels are inharmonic rather than harmonic (Fig. 2 (a), triangles) [$F(1,441)=48.28$, $p=.0016$]. This result is understandable if one considers that targets are less easy to segregate from an inharmonic than a harmonic background (de Cheveigné et al. 1995, 1997b).

Identification of single vowels is essentially perfect and won't be discussed. For of double vowels, the target identification rate is better when both vowels are harmonic rather than inharmonic (Fig. 2(b)) [$F(1,315)=228.96$, $p<.0001$]. From previous results (de Cheveigné 1997b) we can attribute this to the fact that the competing vowel is more difficult to cancel when it is inharmonic rather than harmonic.

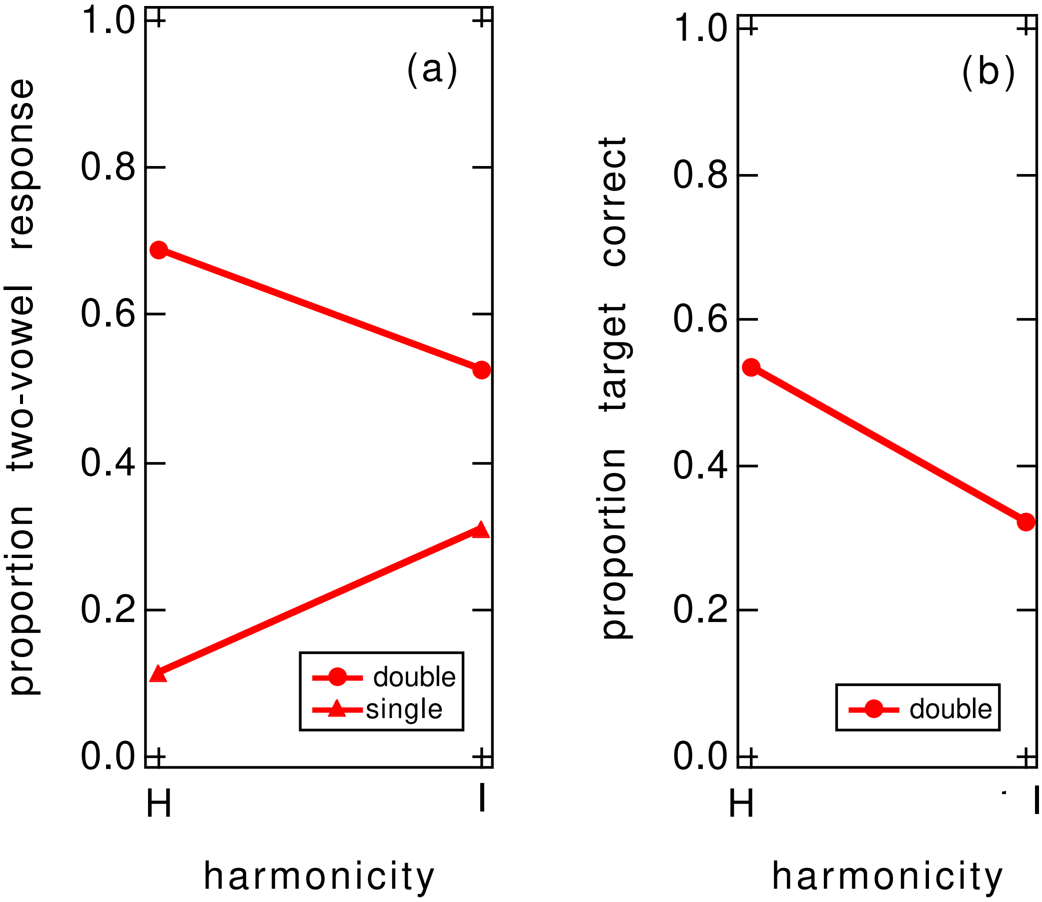

Fig. 2 (a) Proportion of two-vowel responses, and (b) proportion of target-correct responses, as a function of the harmonic state of the single vowel (triangles), or of both vowels in a pair (circles).

4. Presence of FM

We expected that the presence of FM within a stimulus might affect the number of vowels reported, either augmenting it or diminishing it, and that it might thus also affect identification. Fig. 3 compares two-vowel response rates (a) and target-correct identification rates (b) for comparable conditions that differ by the presence or absence of FM. The differences between the two are not significant.

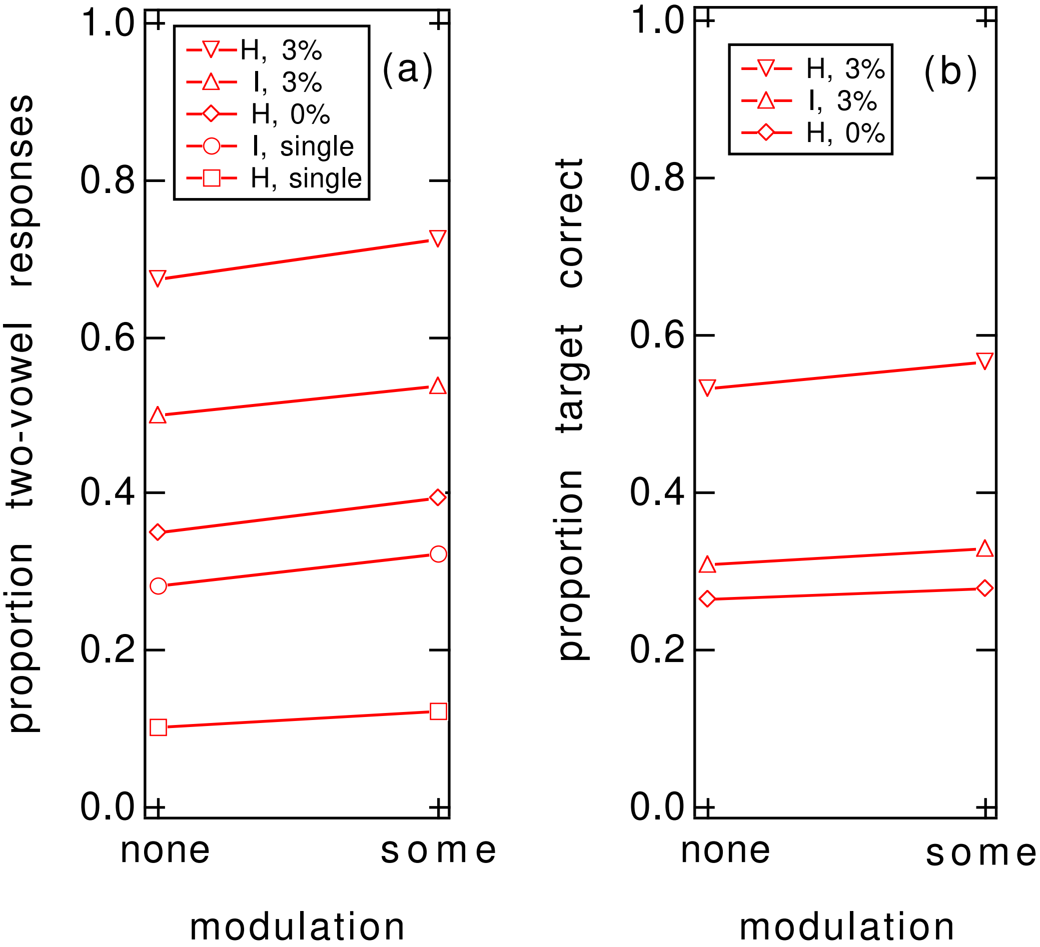

Fig. 3 (a) Proportion of two-vowel responses, and (b) target-correct identification rate according to whether FM is absent ("none") or present ("some") within the stimulus. Inverted triangles: (H3_/_) vs (H3n/n, H3u/u, H3n/_, H3u/_, H3_/n, H3_/u); triangles: (I3_/_) vs (I3n/_, I3_/n); losanges: (H0_/_) vs (H0n/n, H0u/u); circles (I_) vs (In, Iu); squares: (H_) vs (Hn, Hu).

5. FM coherence

A comparison can be made between conditions in which the modulation is coherent (H3_/_, H3n/n, H3u/u) or incoherent (H3n/_, H3u/_, H3_/n, H3n/_). The difference in two-vowel response rate (Fig 4(a)) misses significance at the 0.05 level [$F(1,441)=3.67$, $p=0.067$], but the difference in identification rates is marginally significant [$F(1,315)=8.84$, $p=0.028$]: targets are better identified (59.1% vs 53.5%) when both vowels have incoherent modulation patterns (or one is modulated and the other not) than when they have coherent modulation patterns (or are both static).

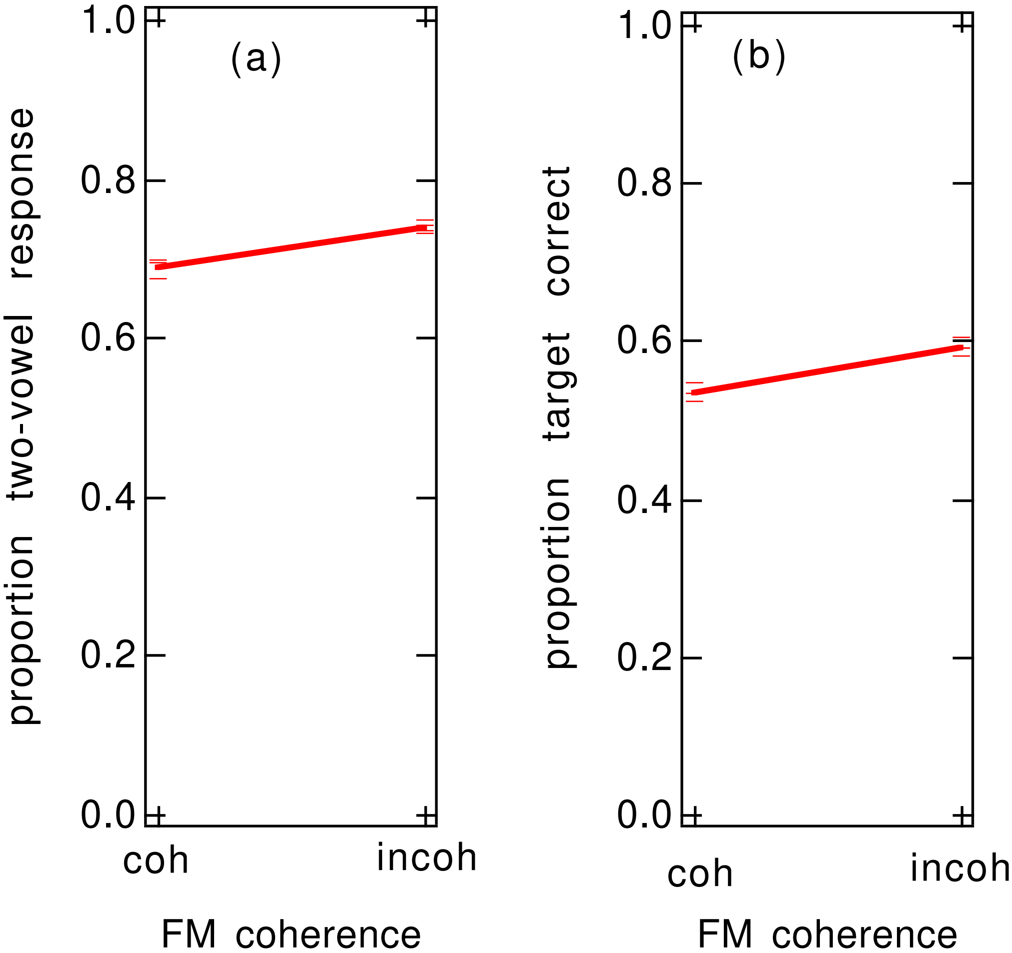

Fig. 4. (a): Proportion of two-vowel responses as a function of FM coherence, for harmonic pairs at ΔF_0=3%. Markers on the left are for (_/_,n/n,u/u), and those on the right are for (n/_, u/_, _/n, _/u). The line joins the means of both groups. (b): Same, proportion of target-correct responses.

6. FM of target

Comparisons between conditions that differ in modulation state of the target, for example (n/_, u/_) vs (_/_), or (n/n, u/u) vs (_/u, _/n) yeild no significant differences.

7. FM of competitor

Comparisons between conditions that differ in modulation state of the competing vowel, for example (_/n, _/u) vs (_/_), or (n/n, u/u) vs (u/_, n/_) yeild no significant differences. From Summerfield's (1992) results, one would have expected identification to be better for modulated targets with a static competitor (u/_, n/_) might be better than either unmodulated targets (_/_, _/n, _/u) or modulated targets with a modulated competitor (n/n, u/u, n/u, u/n). No such effect was found, either for harmonic vowels, or inharmonic vowels (I3n/_ vs I3_/n).

8. FM shape

If the effectiveness of FM-induced F_0 differences were dependant on the modulation shape, as we argued we should expect a difference between _/u and _/n, or else between u/_ and n/_. No difference was found between these pairs, or any other that differ by the shape of their modulation.

D. Discussion

Effects of ΔF_0 are similar to those observed previously, but somewhat smaller than we found in a similar experiment with equivalent conditions and different subjects. Inharmonicity increases the number of perceived sources in the case of single vowels. For double vowels at ΔF_0=3%, the tendency is countered by the increased difficulty in suppressing a non-harmonic competing vowel. A previous experiment with inharmonic stimuli (de Cheveigné et al. 1997b) lacked the I/I condition, and thus had not revealed this effect. Inharmonicity also reduces identification relative to that obtained at ΔF_0=3% with harmonic stimuli. In fact it is not significantly better than at ΔF_0=0% with harmonic stimuli.

Effects involving FM were either weak (coherence effect) or inexistent (presence of FM, target FM, competitor FM, shape of FM). Incoherent FM results in larger peak values of ΔF_0 (6%) than coherent FM (3%), and this might account for the coherence effect. Indeed, two-vowel response rates and identification rates were similar for incoherently modulated pairs at ΔF_0=3% and static pairs at 6%. Such confusion is difficult to avoid in practice.

Using inharmonic target and masking vowels, Summerfield (1992) found that a modulated target with a static masker was better identified than a static target with any masker, or any target with a modulated masker. Based on this result we would expect I(I3n/_) > I(I3_/n). We found no such effect. A possible explanation may lie in the different definition of "inharmonic" vowels. Partials of our inharmonic vowels were displaced by -3%, 0% or 3%, and thus formed three sparse harmonic series with rather close fundamentals. The same pattern of displacements was used for all vowels at all F_0s, and so partials of an inharmonic double vowel at ΔF_0=3% fell on four harmonic series. In contrast, partial frequencies of Summerfield's inharmonic vowels were randomly distributed over a much wider range ($-F_0/2$, $+F_0/2$). His stimuli were also longer (400 ms 250 ms here), and static F0s were larger (2 or 4 semitones, against 1/2 semitone here).

From the prominence results of McAdams (1989) and Marin and McAdams (1991) we expected FM might affect the number of sources heard, perhaps in the absence of identification effects. This failed to be the case.

Overall, our experiment provided no conclusive evidence that FM can affect segregation, other than effects mediated by the instantaneous differences of F_0 that FM may induce.

E. Conclusion

A double-vowel identification experiment was designed to be sensitive to several hypothetical mechanisms by which frequency modulation can affect the perception of concurrent speech sounds. Coherently modulated vowels were slightly less well identified than incoherently modulated vowels, but the effect was small and compounded with differences in maximum instaneous ΔF_0. None of the other hypothesized FM effect was found. Overall, the data are consistent with the hypothesis that segregation effects sometimes observed between frequency modulated sounds are entirely the consequence of FM-induced instantaneous F_0 differences.

F. Ackowledgements

The experiments were partly carried out at ATR Human Information Processing Laboratories, within a research collaboration agreement with the Centre National de la Recherche Scientifique. The first author thanks ATR for its kind hospitality and the CNRS for leave of absence. Hideki Kawahara, Minory Tsuzaki, and Ikuyo Masuda contributed ideas and advice and Rieko Kubo supervised the experiments in Japan. Thanks to John Culling of the MRC Institute of Hearing Research for providing the software for stimulus synthesis.

References

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.

Boe, L.-J., J.-L. Schwartz, et al. (1993). La prédiction des structures sonores, Programme Pluriannuel Rhone Alpes en Sciences Humaines (unpublished report).

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

Brokx, J. P. L. and S. G. Nooteboom (1982). "Intonation and the perceptual separation of simultaneous voices." Journal of Phonetics 10: 23-36.

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

Carlyon, R. P. (1994). "Further evidence against an accross-frequency mechanism specific to the detection of frequency modulation (FM) incoherence between resolved frequency components." JASA 95: 949-961.

Carlyon, R. P. and R. J. Stubbs (1989). "Detecting single-cycle frequency modulation imposed on sinusoidal, harmonic and inharmonic carriers." JASA 85: 2563-2574.

Cherry, E. C. (1953). "Some experiments on the recognition of speech with one, and with two ears." JASA 25: 975-979.

Cohen, M. F. and X. Chen (1992). "Dynamic frequency change among stimulus components: effects of coherence on detectability." JASA 92: 766-772.

Culling, J. (1996). "Signal processing software for teaching and research in psychoacoustics," Behavior Research Methods, Instruments, and Computers, in press.

Culling, J. F. and C. J. Darwin (1993). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F_0." JASA 93: 3454-3467.

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 Communication, 14, 71-95.

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.

de Cheveigné, A. (1997). "Concurrent vowel segregation III: A neural time-domain model of harmonic interference cancellation," J. Acoust. Soc. Am., in press.

de Cheveigné, A., Kawahara, H., Tsuzaki, M. and Aikawa, K. (1997a). "Concurrent vowel segregation I: Effects of relative amplitude and F_0 difference," J. Acoust. Soc. Am, in press.

de Cheveigné, A, McAdams, S. and Marin, C. (1997b). "Concurrent vowel segregation II: Effects of phase, harmonicity and task," J. Acoust. Soc. Am., THIS ISSUE.

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

Demany, L. and S. Clément (1995). "The perception of frequency peaks and troughs in wide frequency modulations. III. Complex carriers." JASA 98: 2515-2523.

Furukawa, S. and B. C. J. Moore (1996). “Across-channel processes in frequency modulation detection.” JASA 100: 2299-2311.
Gardner, R. B., S. A. Gaskill, et al. (1989). “Perceptual grouping of formants with static and dynamic differences in fundamental frequency.” JASA 85: 1329-1337.
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.
Helmoltz, H.L.F. (1885), "On the sensations of tone as a physiological basis for the theory of music" (english translation, 1954, Dover, New York).
Hirahara, T. and Kato, H. (1992). "The effect of F_0 on vowel identification," in *Speech perception, production and linguistic structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (Ohmsha, Tokyo), 89-112.
Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am., 67, 838-844.
Lea, A. (1992). "Auditory models of vowel perception," unpublished doctoral thesis, Nottingham.
Marin, C. and S. McAdams (1991). “Segregation of concurrent sounds. II: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width.” JASA 89: 341-351.
Marin, C. and S. McAdams (1996). “The role of auditory beats induced by frequency modulation and polyperiodicity in the perception of spectrally embedded complex target sounds.” JASA 100: 1736-1753.
McAdams, S. (1984). Spectral fusion, spectral parsing and the formation of auditory images. Stanford.
Scheffers, M. T. M. (1983). "Sifting vowels," unpublished doctoral thesis, Gröningen.
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 (Mouton-de Gruyter, Berlin), 157-166.
Summerfield, Q. and Culling, J. F. (1992). "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency". 124th meeting of the ASA [J. Acoust. Soc. Am. 92, 2317 (A)].
Wilson, A. S., J. W. Hall, et al. (1990). “Detection of frequency modulation (FM) in the presence of a second FM tone.” JASA 88: 1333-1338.