EXPLORING THE CONDITIONS FOR THE PERCEPTUAL SEPARATION OF CONCURRENT VOICES USING F_0 DIFFERENCES

J F Culling,

Experimental Psychology, University of Sussex, Brighton BN1 9QG

1. INTRODUCTION

Scheffers [1] found that two concurrent synthesized vowels (from a selection of 8) are recognised more easily when there is an F_0 difference (ΔF_0) between them than when they are on the same F_0. There have been a number of attempts to model Scheffers' data computationally without further investigations of the Psychological processes involved.

This paper describes the models which have been advanced to account for the effect of ΔF_0s by using harmonic selection. Two experiments are described which demonstrate an effect of ΔF_0s for combinations of synthesized vowels which possess potentially misleading departures from correct harmonic structure. A further model, designed to account for some of these new Psychophysical data without using harmonic selection, is then advanced.

2. MODELS USING HARMONIC SELECTION

2.1 The Harmonic Sieve Model

Scheffers modelled the separation process using a simulated cochlea filterbank. Harmonic sieves were applied to a 'cochlea power spectum' from this filterbank. The sieves admitted spectral energy within 4% of the first 12 harmonic frequencies of a specified F_0, and all energy at higher frequencies (where 4% sieve slots begin to overlap). The F_0s of the sieves could be selected to match those of the two constituent vowels, and vowels recognised from the two sieved spectra.

Scheffers found some improvement in the model's performance with increasing ΔF_0s, centred on 151Hz, but the increase in performance was erratic and did not asymptote like the Human data at 2-4 semitones. Increases in performance were also highly dependent on the frequencies of the F_0s used; F_0s centred on 156Hz produced a decline in the model's performance with increasing ΔF_0. Listeners, on the other hand, are little affected by the F_0s used. Zwicker [2], Assmann & Summerfield [3] and Chalikia & Bregman [4] have reproduced asymptotic performance profiles using various different F_0 values.

2.2 Autocorrelation Models

More recent attempts to model the effect of ΔF_0s have used autocorrelation methods based on Licklider's [5] model of pitch perception. While the harmonic sieve uses only the cochlea power spectrum, autocorrelation models make use of the waveform which emerges from each cochlea filter channel.

Typically, the waveform in each channel is autocorrelated using a range of delays to produce an autocorrelation function (ACF). The strength of autocorrelation at a delay equal to a constituent vowel's fundamental period is taken as a reflection of the contribution to that channel from that vowel. Two output spectra for vowel template matching are derived from the contributions from each constituent across different frequency channels. Assmann & Summerfield [3] have demonstrated that, using the same model of cochlear filtering and transduction, an autocorrelation model gives superior performance to a harmonic sieve model.

3. ACROSS FORMANT INCONSISTENCIES IN F_0

3.1 Introduction

There are two ways in which harmonic selection might improve the recognition of concurrent vowels with ΔF_0s. First, the selection of two harmonic series from two overlapping formants may allow better formant frequency estimation for each of the formants. Second, formants in different frequency regions which share the same F_0 may be allocated to the same vowel, while those with different F_0s are allocated to separate vowels.

In support of the latter hypothesis, Gardner et al. [6] have shown that when F2 in the synthesized syllable /ru/ is on a different F_0 from F1, F3 & F4, it may begin to be heard separately as an isolated buzzing sound, while the remaining formants are heard as a different syllable, /li/. Perceptual exclusion of the second formant, as reflected in the frequency of /li/ percepts, increased steadily across a wide range of ΔF_0s. The present experiment was designed to test the role of grouping of formants by common F_0 in Scheffers' concurrent vowel paradigm. The constituent vowels displayed across-formant inconsistencies in F_0, which should mislead any formant grouping mechanism.

3.2 Method

The 5 English tense vowels (i, ɑ, u, ɜ & ɔ) were synthesized using an additive Klatt software synthesizer according to the parameters used by Assmann & Summerfield [3]. Vowels were either synthesized using the same F_0 for all components, or with an abrupt change in F_0 at the spectral minimum between F1 and F2.

The vowels were combined into 3 types of vowel pair; 'normal', 'F_0-swapped' and 'same-F_0-for-F2/3'. Normal vowel pairs used each F_0 throughout the spectrum of each constituent vowel. F_0-swapped pairs used each F_0 for the F1 region of one vowel and the F2/3 region of the other vowel; formant grouping mechanisms should group the F1 of one vowel with the F2 of the other and vice versa. Same-F_0-for-F2/3 pairs used one F_0 thoughout the spectrum of one vowel and in the F2/3 region of the competing vowel, while the second F_0 was used only in the F1 region of the second vowel.

3.3 Results

Figure 1 shows subjects' recognition performance with the 3 types of vowel pair at each ΔF_0. There is little difference between each of the 3 conditions, and all show significant improvements with increasing ΔF_0. There are signs of a decline in the F_0-swapped and same-F_0-for-f2/3 conditions at 2-4 semitones.

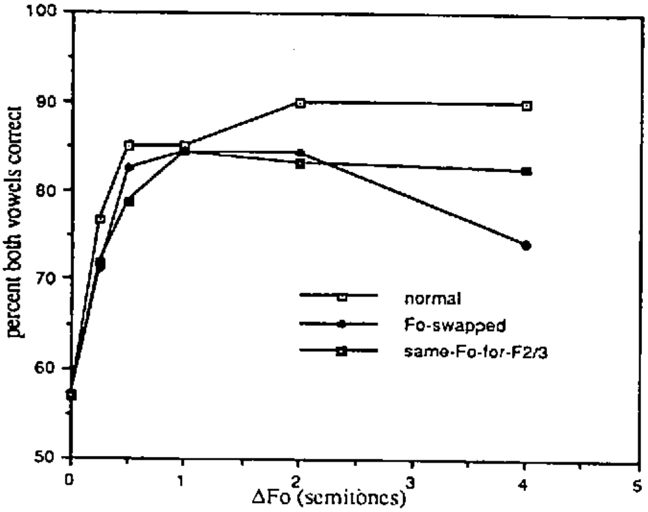

Figure 1: performance for normal, Fo-swapped and same-Fo-for-F2/3 stimuli at each ΔFo

3.4 Conclusions

The F_0-swapped condition appears to have had very little influence upon the vowel separation effect, suggesting that across formant grouping can only play a minor role for ΔF_0s < 4 semitones. This appears to conflict with the results of Gardner et al. However, in their experiment, relatively large ΔF_0s (more than 2 semitones) were required to perceptually exclude F2 on a majority of occasions.

The same-F_0-for-F2/3 condition provides some insight into the F_0-swapped results. With ΔF_0s in only the F1 region, results were almost identical to those of the other conditions, suggesting that the F1 region is largely responsible the effect of ΔF_0s < 4 semitones.

4. THE ROLE OF BEATING IN THE F1 REGION

4.1 Introduction

The dominance of the F1 region in the ΔF_0 effect accords with the harmonic sieve model, which can only separate the first 12 components of each vowel. At the small ΔF_0s used in Scheffers' paradigm, however, harmonics of the same number from different vowels are only slightly mistuned. In the F1 region the mistuning is so small that a high resolution FFT can resolve little detail of the two harmonic series; a vowel pair with a ΔF_0 is almost indistinguishable from from one without.

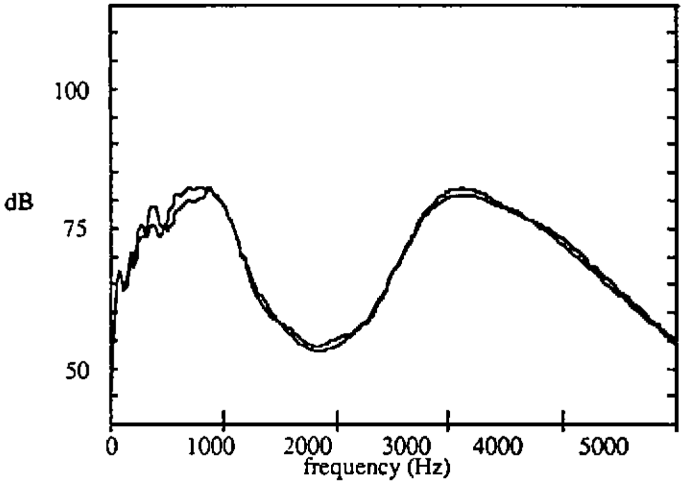

Figure 2: cochlea power spectra of /ɑ/ + /ɔ/ from 30ms frames at 30ms and 90ms

Paradoxically, spectra of lower resolution, comparable to that of the ear, show large differences in the F1 region when a ΔF_0 is introduced. The slightly mistuned harmonics in the F1 region beat together to procduce spectral change during the stimulus. Figure 2 shows the spectrum of a vowel pair (/ɑ/+/ɔ/) at two instants. The spectrum of a pair without a ΔF_0 would be almost static.

Since we have already seen that the F1 region is responsible for most of the ΔF_0 effect, we should examine the possibility that these spectral changes are involved. Perhaps the spectrum resembles one vowel at one point in time and the other at another point. Perhaps there is a point in time when the spectrum is more amenable to spectral analysis into its constituent vowels, while elsewhere it is dominated by one of them. In order to test these possibilities stimuli were devised which would mislead a harmonic selection mechanism completely, while still providing a similar pattern of beating.

4.2 Method

Two conditions were prepared, termed 'normal' and 'interleaved'. The normal vowel pairs were identical to those in the first experiment. For the interleaved pairs the odd harmonic frequencies of one F_0 and the even harmonic frequencies of the other F_0 were assigned to one vowel, while the remaining harmonics of each F_0 were assigned to the second vowel (see figure 3). The spectral envelopes the two vowels were thus excited by a mixture of the two F_0s; any harmonic selection mechanism should select out two series which each sample the two spectral envelopes alternately.

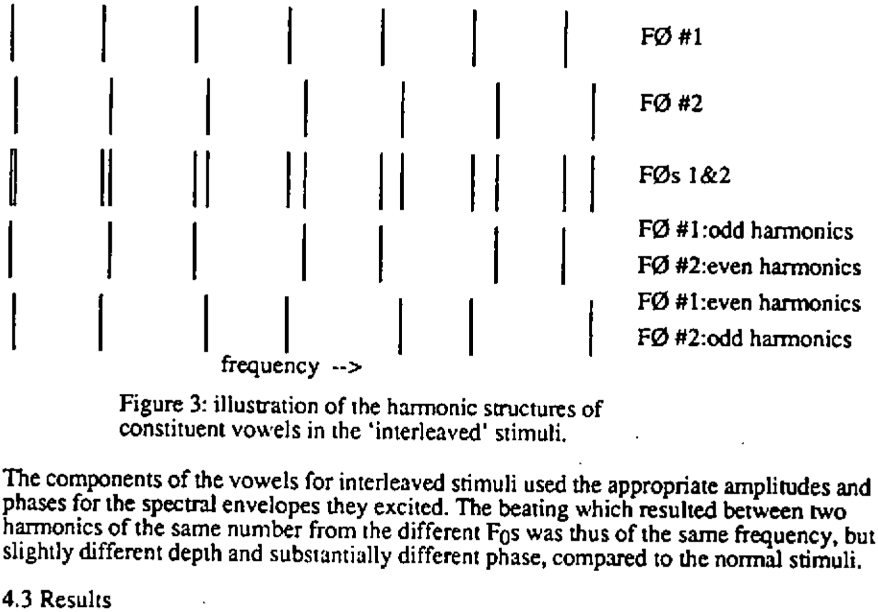

Figure 3: illustration of the harmonic structures of constituent vowels in the 'interleaved' stimuli.

The components of the vowels for interleaved stimuli used the appropriate amplitudes and phases for the spectral envelopes they excited. The beating which resulted between two harmonics of the same number from the different F_0s was thus of the same frequency, but of slightly different depth and substantially different phase, compared to the normal stimuli.

4.3 Results

The interleaved stimuli produced a significant improvement with the introduction of ΔF_0s (figure 4). This improvement began to fall off at 4 semitones. The normal condition, however, shows higher recognition rates, particularly with $\Delta F_0s \geq 1$ semitone.

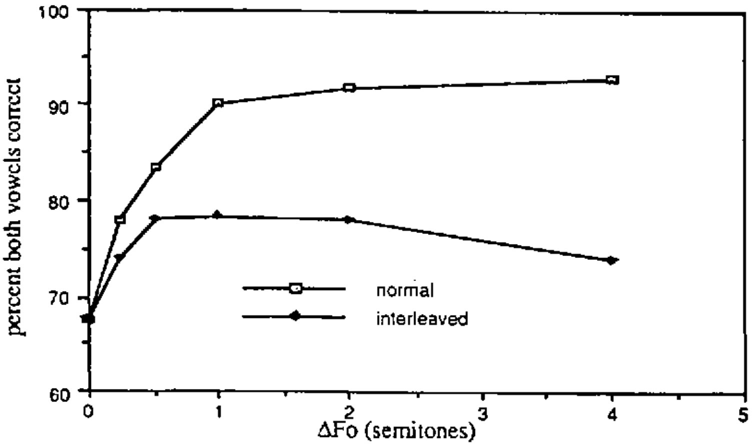

Figure 5: model performance for normal and interleaved stimuli at each ΔFo

4.4 Conclusions

The improvement for interleaved stimuli with ΔF_0s cannot be attributed to any harmonic selection mechanism. A mechanism based upon beating may therefore be responsible. The improvement for interleaved stimuli falls short of that for normal stimuli at all ΔF_0s, but is particularly marked for ΔF_0s ≥1semitone. There may be two effects at work here. First, a harmonic selection mechanism may be increasingly effective with increasing ΔF_0s, reflected in the progressive divergence of the two performance profiles. Second the altered phases in the beating pattern may have reduced the effectiveness of beating cues in the interleaved stimuli, in which case harmonic selection mechanisms may simply be inactive at small ΔF_0s.

5. COMPUTATIONAL MODEL

5.1 Design

A computational model was designed to exploit timbral changes in the stimuli of the second experiment. A gamma-tone filterbank output [7]was sampled 33 times at 30ms intervals by an auditory temporal window [8], to provide a suitable rate-place representaion with no fine timing information of the kind used by autocorrelation models.

A 2-layer PDP network was trained to recognise the individual 'normal' vowels using each F_0. Output activations produced in response to the paired stimuli were then converted into "response probabilities" for selecting both vowels correctly. Negative activations were zeroed and the probability for each vowel response was then taken to be proportional to activation. These probabilities were combined using the following formula:

$$p(a\&b) = p(a).p(b|a) + p(b).p(a|b)$$

The probability of a correct response was taken to be the highest "response probability" for the corect combination across the 33 samples.

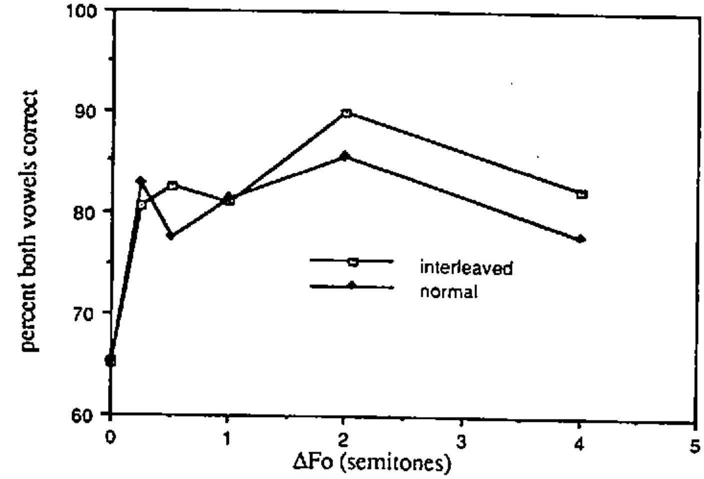

Figure 5: model performance for normal and interleaved stimuli at each ΔFo

5.2 Results
Figure 5 shows the results for 'normal' and 'interleaved' stimuli at each ΔF_0. Performance for both normal and interleaved stimuli shows a marked increase with the introduction of ΔF_0s. There is some sign of a fall at 4 semitones.

5.3 Conclusions
The model has produced performance profiles for each type of stimulus which are simlilar in form both to each other and to Human performace with interleaved stimuli. Since the interleaved stimuli were designed to provide only timbral cues and the model was designed to exploit only timbral cues, these results support the notion that timbral change can be, and is, exploited by listeners in experiments using concurrent vowels. Harmonic selection mechanisms may therefore make a correspondingly smaller contribution to the ΔF_0 effect.

The normal stimuli produce slightly higher scores at 1/4 semitone ΔF_0, providing some support for the notion that Human performance with interleaved stimuli may have been depressed by the altered beat phases. At higher ΔF_0s, however, the interleaved stimuli facilitate higher scores then the normal stimuli.

REFERENCES

[1] M T Scheffers, 'Sifting vowels: Auditory pitch analysis and sound segregation', Ph.D., Groningen University, The Netherlands, 1983

[2] U T Zwicker, 'Auditory recognition of diotic and dichotic vowel pairs', Speech Comm, 3, 265-78 (1984)

[3] P F Assmann & A Q Summerfield, 'Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies.', J. Acoust. Soc. Amer., 88, 680-697 (1990)

[4] M H Chalikia & A S Bregman, 'The perceptual segregation of simultaneous auditory signals: Pulse train segregation and vowel separation', Percep. Psychophys., 46, 487-496 (1989)

[5] J C R Licklider, 'A duplex theory of pitch perception', Experientia, 7, 128-136 (1951)

[6] R B Gardner, S A Gaskill & C J Darwin, 'Perceptual grouping of formants with static and dynamic differences in fundamental frequency.', J Acoust Soc Amer, 85, 1329-37 (1989)

[7] R Patterson, I Nimmo-Smith J Holdsworth & P Rice, 'Spiral VOS final report, Part A: The auditory filter bank' Cambridge Electronic Design, Contract Report (APU 2341) (1988)

[8] C J Plack & B C J Moore, 'The auditory temporal window shape as a function of frequency and level', J. Acoust. Soc. Amer., 88, 680-697 (1990)