PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Auditory Segregation of Competing Voices: Absence of Effects of FM or AM Coherence [and Discussion]

Quentin Summerfield, John F. Culling and A. J. Fourcin

Phil. Trans. R. Soc. Lond. B 1992 **336**, 357-366
doi: 10.1098/rstb.1992.0069

Auditory segregation of competing voices: absence of effects of FM or AM coherence

QUENTIN SUMMERFIELD AND JOHN F. CULLING

MRC Institute of Hearing Research, University Park, Nottingham NG7 2RD, U.K.

SUMMARY

Four experiments sought evidence that listeners can use coherent changes in the frequency or amplitude of harmonics to segregate concurrent vowels. Segregation was not helped by giving the harmonics of competing vowels different patterns of frequency or amplitude modulation. However, modulating the frequencies of the components of one vowel was beneficial when the other vowel was not modulated, provided that both vowels were composed of components placed randomly in frequency. In addition, staggering the onsets of the two vowels, so that the amplitude of one vowel increased abruptly while the amplitude of the other was stationary, was also beneficial. Thus, the results demonstrate that listeners can group changing harmonics and can segregate them from stationary harmonics, but cannot use coherence of change to separate two sets of changing harmonics.

1. INTRODUCTION

There are at least two reasons for studying the auditory and perceptual processes which listeners use to attend selectively to one voice in a mixture of voices. First, speech is generally heard against a background of other sounds, including other voices. Thus an account of speech perception should include descriptions of the ways in which the elements of a voice are identified, grouped together, and separated from other sounds. Second, many hearing-impaired listeners have difficulty understanding speech in noise, particularly when the noise consists of other voices. Understanding the processes by which speech is normally extracted from interfering sounds, and the ways in which those processes break down in pathology, could lead to improved algorithms for speech enhancement in hearing aids.

The basic problem in segregating voices was set out by Broadbent & Ladefoged (1957): when two talkers speak concurrently, the spectrum of the sound reaching listeners' ears contains evidence of the formants of both voices. What cues enable listeners to assign each formant to the appropriate source? When both talkers produce voiced speech, the problem is that of correctly assigning each of the harmonics that define the formant peaks. Much work has sought to identify the mechanisms of spectral and temporal analysis that exploit the 'harmonicity' of the harmonics of a voice; i.e. the fact that the harmonics are found at frequencies that are integer multiples of their common fundamental frequency (F_0) (e.g. Broadbent & Ladefoged 1957; Darwin 1981; Scheffers 1983; Zwicker 1984; Gardner *et al.* 1989; Darwin & Culling 1990; Assmann & Summerfield 1990; Summerfield & Assmann 1991; Meddis & Hewitt 1992*b*). In this paper, we are concerned with an additional issue: the role of time-varying cues. The experiments ask whether listeners can use correlated changes in the frequencies or amplitudes of harmonics, in addition to their harmonicity, to segregate competing voices.

2. EFFECTS OF COHERENT FREQUENCY MODULATION

When the fundamental frequency of a voiced vowel changes, the frequencies of its harmonics change coherently: they all rise or fall by the same percentage of their starting frequency. We shall refer to this example of common fate as 'coherent frequency modulation' (CFM) and ask: Can CFM help to group the harmonics of one voice and segregate them from the harmonics of a competing voice that are undergoing a different pattern of CFM? (For economy in writing, we shall describe voices that have different patterns of CFM as being 'incoherently modulated' and voices that have the same pattern of CFM as being 'coherently modulated'.) A demonstration by McAdams suggested that CFM might be a powerful grouping principle. He summed the waveforms of three synthetic vowels sung on different pitches. Applying CFM to the harmonics of one member of the triad caused it to stand out perceptually from the other two. Subsequent experiments (McAdams 1989; Marin & McAdams 1990) confirmed that CFM increased the perceptual prominence of one vowel in a mixture. However, its prominence was not affected by the status of the other two vowels. Prominence did not decrease when the other vowels were modulated coherently with the first, nor did it increase when the other vowels were modulated incoherently with the first. Thus McAdams concluded that CFM does not aid segregation. His

Phil. Trans. R. Soc. Lond. B (1992) **336,** 357–366
Printed in Great Britain

conclusion has been reinforced by the results of studies that have used accuracy of identification to measure the ability of listeners to segregate competing vowels or syllables (Chalikia & Bregman 1989; Gardner *et al.* 1989; Darwin & Culling 1990; Demany & Semal 1990). All show that a difference in F_0 is a potent cue for segregation, but that CFM does not make an independent contribution.

Carlyon (1991; see also this symposium) explained this outcome by arguing that listeners may not be able to use CFM. He demonstrated that listeners cannot distinguish coherent FM from incoherent FM carried on a small set of inharmonic tones. For example, in one experiment listeners were presented with tones at 400 Hz and 700 Hz, each modulated at a rate of 5 Hz with a modulation depth (zero-peak) of 5%. Listeners could not distinguish the case where the tones were modulated in phase from the case where their modulating waveforms were 180° out of phase. Carlyon argued that because listeners cannot detect whether components are modulated coherently or incoherently, they cannot be expected to use coherence of FM as a basis for grouping.

A different explanation is implicit in the writings of Chalikia & Bregman (1992). They argued that harmonicity is such a powerful cue for grouping that it permits all the segregation that can be achieved, leaving nothing for CFM to contribute. Chalikia & Bregman proposed that the way to demonstrate effects of CFM is to prevent harmonicity from playing a role, by synthesizing competing sounds whose components are placed randomly in frequency rather than harmonically. They carried out such an experiment. Inharmonic ('random') vowels were created from harmonic vowels by (i) randomly displacing harmonics in frequency within a circumscribed range and (ii) adjusting the amplitudes of the displaced harmonics to reinstate the original spectral envelope. In this way it was possible to convert a harmonic stimulus with a F_0 of, say, 100 Hz, into an inharmonic stimulus with a 'nominal F_0' of 100 Hz. A second inharmonic sound, whose nominal F_0 differed from the first by, say, 2 semitones, could then be created by changing the frequency of each component by 2 semitones.

Chalikia & Bregman used these procedures to generate pairs of inharmonic vowels whose nominal F_0s either rose or fell by 6 semitones over a duration of 2 s. In one set of pairs, the nominal F_0 values maintained a constant difference of 6 semitones. In another set, the initial and final differences were 6 semitones but the contours crossed. The members of the crossing pairs were identified slightly, but significantly, more accurately (91% correct compared to 87%) than the members of the parallel pairs, suggesting a small role for CFM.

Chalikia & Bregman's experiment was ingenious, but its implementation suffered from the problem that different stimuli, involving different ranges of nominal F_0, were presented in the different conditions. Thus, results could have been confounded by differences in the phonetic distinctiveness of the vowels depending on the precision with which components defined the locations of formant peaks. Experiment 1 sought to distinguish Chalikia & Bregman's account of the role of CFM from Carlyon's account using a more rigorous psychophysical procedure.

3. MEASURING SEGREGATION THROUGH MASKING

We measure the effectiveness of cues for segregating voices using a masking procedure (Summerfield 1992) derived from procedures used by Demany & Semal (1990) and Summerfield & Assmann (1991). The procedure determines the minimal signal-to-noise ratio at which listeners can identify 'target' vowels in the presence of 'masking' vowels. An increased ability to segregate targets from maskers is revealed as a fall in the listener's masked threshold and can be quantified in dB. We use a two-interval, five-alternative forced-choice task. Maskers are presented in both intervals at a mean level of 60 dB (A). A target vowel is presented in one interval, chosen randomly. To score a correct response, listeners must indicate which interval contained the target and what its identity was. An adaptive staircase controls the target-to-masker ratio (TMR) and estimates the TMR giving 71% correct responses. The targets are exemplars of the five British–English vowels /a/, /i/, /ɜ/, /u/, and /ɔ/, with unchanging formant frequencies synthesized with a version of the cascade synthesizer described by Klatt (1980). The maskers are also five-formant sounds, but differ from the targets. To prevent listeners using unintended cues, two parameters are varied randomly between the intervals: overall level, so that an increase in loudness cannot be used to locate the interval containing the target; and the spectrum of the masker, so that the spectrum of the target cannot be recovered by computing the difference between the spectra of the sounds presented in the two intervals. Conditions are distinguished by changing the maskers not the targets. Thus, differences between the phonetic distinctiveness of the targets cannot confound the results. The maskers are drawn randomly from a set of ten. Thus, it is unlikely that listeners perform the task by learning the sound of each target combined with each masker.

Three experienced listeners with normal hearing took part in each experiment. Each listener provided two thresholds in each condition. Results are reported averaged over listeners. The test–retest reliability is such that differences between conditions of 3 dB are significant.

4. EXPERIMENT 1: EFFECTS OF CFM WITH HARMONIC AND INHARMONIC VOWELS

The experiment involved seven conditions which were distinguished by different relationships between the F_0 contours of maskers and targets. These relationships are shown schematically in the panels at the top of figure 1. Maskers and targets were 400 ms in duration with onsets and offsets shaped by 20 ms raised-cosine functions. The components of the targets were always modulated. Mean F_0s were chosen randomly from the set 100.0, 112.2, 126.0, and 141.4 Hz, whose members

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

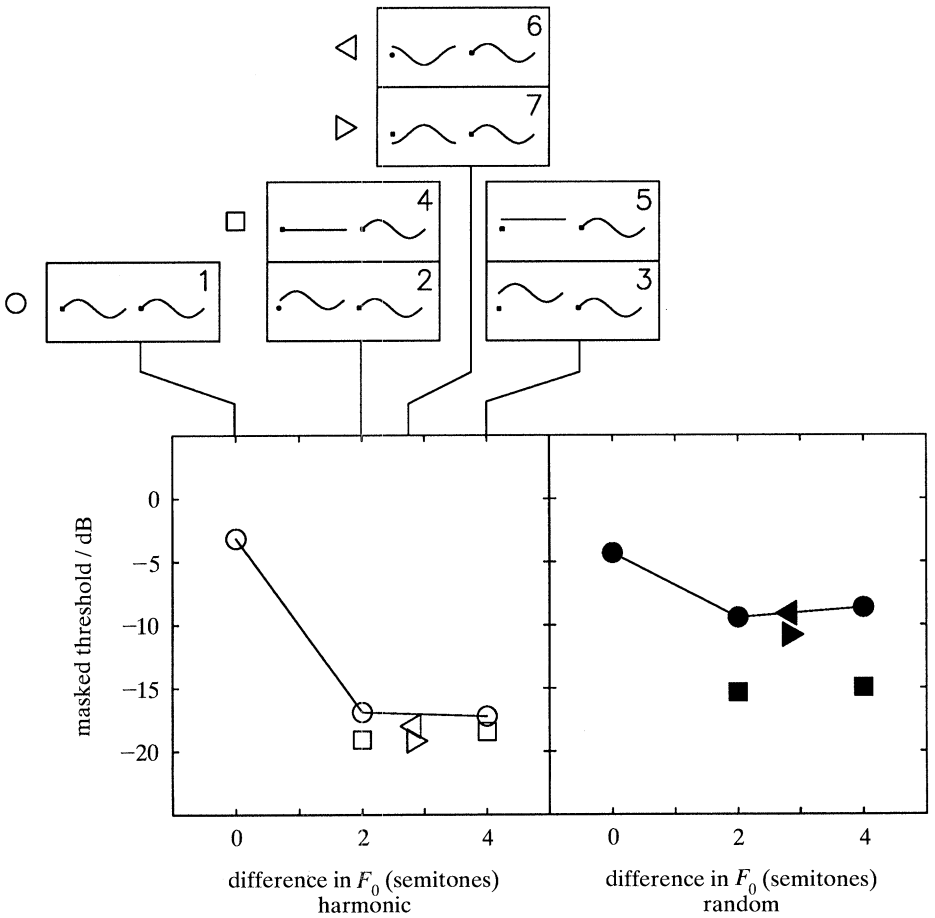

Figure 1. Results of experiment 1 for harmonic stimuli (open symbols, left graph) and inharmonic stimuli (filled symbols, right graph). Panels at the top of the plot illustrate the relationships between the F_0 contours of maskers (left trace in each panel) and targets (right trace). The small squares mark a constant point in time-frequency for reference. The numbers 1–7 identify the different conditions. Plotting symbols are shown beside each panel. Results in conditions with coherent modulation are shown by circles, in conditions with incoherent modulation by triangles, and in conditions where maskers were not modulated by squares.

are 2 semitones from their nearest neighbours. The modulation rate was 2.5 Hz and the modulation depth (zero-peak) was two semitones (12.2%). When maskers were modulated, their rate and depth were also 2.5 Hz and two semitones.

(*a*) *Role of CFM with harmonic stimuli*

Results obtained with harmonic stimuli are plotted in the left-hand panel of figure 1. Three effects can be seen. First, the circles show that thresholds fell by 14 dB when a difference of 2 semitones was introduced between maskers and targets (compare conditions [1] and [2]), but fell no further when the difference was increased to 4 semitones [3]. The result is compatible with earlier results showing that the benefits of F_0 differences reach a plateau at a difference of about 2 semitones (e.g. Summerfield & Assmann 1991). Second, thresholds were not significantly lower when targets were modulated against static maskers giving maximum differences in F_0 of 2 or 4 semitones (squares: conditions [4] and [5], respectively) than in the corresponding conditions where maskers and targets were both modulated with constant differences of 2 or 4 semitones (circles: conditions [2] and [3]). Third, the triangles show that there was no significant advantage from modulating maskers and targets incoherently by either advancing [6] or retarding [7] the phase of the masker modulation by 90°. In these conditions, the maximum instantaneous difference in F_0 was 2.8 semitones. Thresholds were not significantly lower in conditions [6] and [7] (triangles) than in conditions [2] and [3] (circles) where maskers and targets were modulated coherently with constant differences of either 2 or 4 semitones. In other words, there was no advantage from giving maskers and targets different patterns of CFM over and above the advantage that would be expected from the maximum instantaneous difference in F_0 occurring during the modulation cycle. The results are compatible with the reports noted above which suggested that CFM plays no independent role in segregating harmonic sounds.

(*b*) *Effects of onset asynchronies with harmonic stimuli*

According to Chalikia and Bregman's explanation, no effect of CFM was shown in conditions [6] and [7] because the instantaneous differences in F_0 had already allowed thresholds to fall as far as they could

in conditions [2] and [3]. We tested this aspect of their explanation by checking whether the introduction of an additional cue for segregation would cause thresholds to fall further. The further cue was a difference in onset time. Maskers started 200 ms before targets; they then continued for 400 ms and ended together. Staggering the onset times of competing sounds generally facilitates their segregation. For example, a harmonic that starts before the remaining harmonics in a complex makes a reduced contribution to the pitch of the complex (Darwin and Ciocca, 1991) and to the phonetic quality of a vowel (Darwin, 1984). Similarly, identification of the second vowel in a pair is more accurate if the second vowel starts after the first (Summerfield and Assmann, 1989). In the present experiments, we should expect thresholds to fall when maskers start before targets, unless differences in F_0 have already allowed thresholds to fall as far as they can.

The lengths of the open bars in Figure 2 show the amounts by which thresholds fell when maskers started 200 ms before targets, compared to the results plotted in Figure 1 where they started together. In the condition where there was no difference in F_0 between maskers and targets [1], thresholds fell significantly, showing that onset asynchrony can aid segregation. However, thresholds did not fall significantly when maskers and targets were modulated coherently with a constant difference of 2 semitones [2], nor in any of the other conditions. Thus, the outcome is compatible with the idea that no effect of CFM is found with harmonic stimuli because differences in F_0 allow thresholds to fall as far as they can. However, it is also possible that listeners simply cannot use CFM for segregation. The results obtained with the inharmonic stimuli, described in the next section, distinguish these alternatives.

(*c*) *Role of CFM with inharmonic stimuli*

The filled symbols in the right-hand panel of figure 1 show the results obtained with inharmonic stimuli. The circles show that thresholds fell by 5 dB when a difference of 2 semitones [2] or 4 semitones [3] was introduced between corresponding components in maskers and targets. The fall probably occurred for the following reasons. When maskers and targets had the same nominal F_0, they were composed of components with the same frequencies but different amplitudes and phases. Summation of the two waveforms distorted the spectral envelopes of both signals. Introducing a difference of 2 or 4 semitones between corresponding components reduced the interference, improving the definition of formant peaks in the targets. This outcome suggests that only 9 dB of the 14 dB fall obtained with harmonic stimuli in the analogous conditions should be attributed to effects of harmonicity.

The squares show that it is relatively easy to identify a modulated target against a static masker. In conditions [4] and [5] (squares), targets were modulated while maskers were static. Thresholds fell significantly compared to conditions [2] and [3] (circles) where maskers and targets were both modulated. The result shows that a sound defined by changing components can 'stand out' against a background of static components.

However, the triangles show that a sound defined by one set of coherently modulated components does not stand out against a sound defined by components which are given a different pattern of CFM. Thresholds were not significantly lower when maskers and targets were modulated incoherently [6] and [7] (triangles) rather than coherently [2] and [3] (circles). Thus, CFM made no contribution to segregation beyond the contribution expected from the maximum instantaneous difference in (nominal) F_0 occurring during the modulation cycle.

(*d*) *Effects of onset asynchronies with inharmonic stimuli*

The solid bars in figure 2 show the effects of introducing a 200 ms onset asynchrony between inharmonic maskers and targets. In general, thresholds fell significantly when they were high in the synchronous case in figure 1 and failed to fall significantly when they started low, as in condition [4]. The crucial comparisons involve conditions [2], [6] and [7]. In conditions [2] and [6], thresholds fell significantly when the onset asynchrony was introduced. This result indicates that, if listeners had been sensitive to CFM, thresholds should also have fallen when FM incoherence was introduced. (This conclusion must be tempered slightly by the failure of condition [7] to show a significant fall.)

(*e*) *Summary*

Experiment 1 has demonstrated that CFM does not help listeners to segregate concurrent vowels. No advantage might be expected in the case of harmonic vowels because harmonicity allows all the segregation that can be achieved. However, CFM also failed to aid the segregation of inharmonic vowels. This outcome is compatible with Carlyon's (1991) conclusion that listeners cannot use CFM for segregation. It runs counter to Chalikia & Bregman's predictions.

Gardner *et al.* (1989) speculated that listeners have not included CFM in their armoury of grouping weapons because the uneven frequency responses of natural reverberant communication channels distort evidence of FM. More generally, Summerfield (1992) suggested that it is likely that CFM is not used because its exploitation would be computationally demanding, and is unnecessary on ecological grounds. To exploit CFM, listeners would have to track individual harmonics and compare their frequency contours. The problem might be soluble when only one source is present but could be intractable in the presence of a competing voice where each set of harmonics would have to be tracked across the changing background of the competing set. Instead, given the low incidence of natural sound sources generating discrete components at inharmonic frequencies, auditory analysis exploits harmonicity for grouping. Harmonicity can be exploited by

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

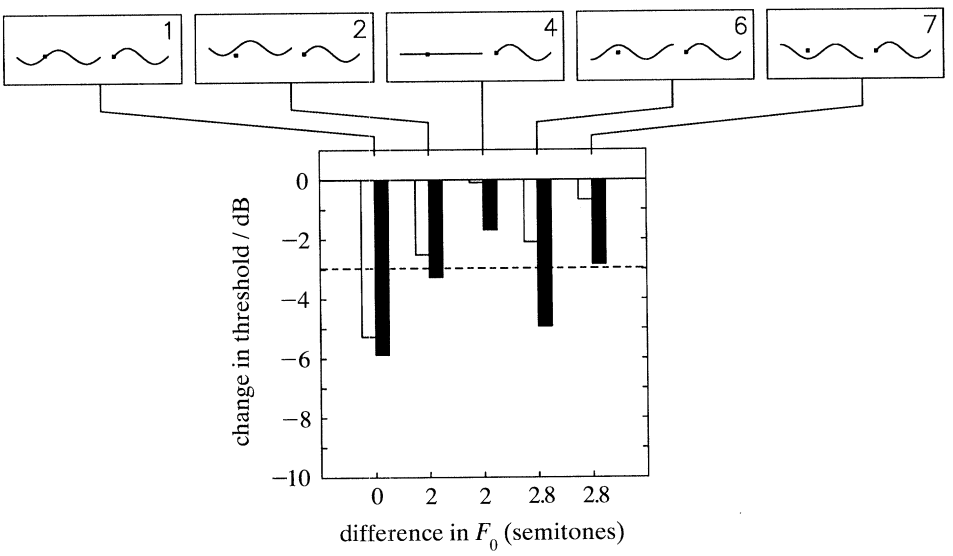

Figure 2. Effects of onset asynchrony in experiment 1 for harmonic stimuli (open bars) and random stimuli (filled bars). The dashed line shows the 3 dB difference between conditions required for significance. Panels at the top of the plot illustrate the relationships between the F_0 contours of maskers (left trace in each panel) and targets (right trace).

across-channel processes without the need to track individual harmonics (Assmann & Summerfield 1990; Meddis & Hewitt 1991)†. Moreover, as shown by the open symbols in figure 1 and the open bars in figure 2, sensitivity to CFM is unnecessary because harmonicity allows all the segregation that can be achieved.

The inability of listeners to use CFM for grouping could reflect a specific limitation in anlaysing the frequencies of individual harmonics, or a general difficulty in separating incoherently modulated sounds. Accordingly, the following experiments describe initial explorations of the ability of listeners to use coherent changes in amplitude to separate concurrent vowels.

5. EFFECTS OF COHERENT AMPLITUDE MODULATION

The experiments study effects of amplitude modulation in the sub-audio-frequency range which extends up to about 50 Hz. Such modulations are produced in speech by the control of air flow and acoustic radiation through the mouth by movements of the jaw, lips, and tongue. In the output of a bank of auditory filters, the modulations are found in a correlated form across a wide range of frequency channels. Their patterning can cue some phonetic distinctions (Rosen, this symposium) and their preservation in communication channels is important for intelligibility. Listeners use the modulations to group energy in different audio-frequency regions. For example, the release from masking demonstrated in co-modulation masking release (CMR) (Hall *et al.* 1984; Hall & Grose, this symposium) can be interpreted as a consequence of grouping: the on-frequency band of noise and its flanking companion band are grouped together by virtue of their correlated patterning in amplitude, thereby allowing the unmodulated signal tone to be heard out from the on-frequency band. CMR has been viewed as a manifestation of mechanisms that segregate co-modulated speech formants from background noises.

A further experiment (Hall & Grose 1990) demonstrated that listeners can segregate concurrent signals carrying different patterns of AM. Hall & Grose measured CMR in conditions where the on-frequency band was centred on 1 kHz and six co-modulated flanking bands were centred on six multiples of 200 Hz around 1 kHz. Two 'co-deviant' bands centred on 900 Hz and 1100 Hz were introduced and modulated together but with a different envelope from the co-modulated bands. Their presence reduced the amount of release from masking compared with the release expected from the co-modulated bands. However, introducing six more co-deviant bands centred on other odd harmonics of 100 Hz around 1 kHz reinstated some of the lost CMR. Hall and Grose argued that increasing the number of co-deviant bands caused them to be grouped separately from the co-modulated bands and thereby prevented them from interfering with the unmasking effect. Although it is not clear whether it was important that the bands in each group were harmonically related, the result demonstrates that concurrent signals with different patterns of AM can be separated. A similar conclusion has been drawn from studies of 'modulation masking' (Bacon & Grantham 1989) which have demonstrated that AM at one rate (e.g. 8 Hz) can mask the detection of AM at the same rate more effectively than at other rates (e.g. 4 or 16 Hz; Houtgast 1989). It might be expected therefore that it should be easier to identify a target vowel in our paradigm if it is given a different

† In some other circumstances listeners can track individual harmonics, because a harmonic that starts before the others in a complex makes a reduced contribution to the pitch of the complex (Darwin & Ciocca 1991) and its vowel colour (Darwin 1984).

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

pattern of amplitude modulation from the masking vowels.

However, other results make the outcome less certain. There is considerable evidence that it is difficult to make judgements about modulations in one frequency region if there are concurrent modulations occurring in different frequency regions. For example, compared with conditions where energy in different frequency regions is not modulated, when it is modulated it is harder to detect amplitude modulation itself (Yost & Sheft 1989) or to detect changes in the phase of modulation (Yost & Sheft 1989), the rate of modulation (Yost *et al.* 1989), or the depth of modulation (Moore *et al.* 1991; Moore & Shailer, this symposium). Moreover, the tuning of these effects of 'modulation detection interference' (MDI) to modulation rate is quite broad. As a result, Moore (1992) has suggested that across-channel masking may hinder the segregation of competing voices, despite the presence of short-term differences in modulation rate between the voices that might be expected to promote segregation.

From these results, it is difficult to predict whether it should be easier or more difficult to separate concurrent vowels if they are given different rates of AM rather than the same rate. Experiment 2 examined this issue.

6. EXPERIMENT 2: EFFECTS OF AMPLITUDE-MODULATION RATE?

Figure 3 shows the relationship between the intensity envelopes of maskers (thicker lines) and targets (thinner lines) in three of the five conditions. The modulation amplitude (zero-peak) of maskers and targets was 5 dB. The modulation rate of the targets was 8 Hz; that of the maskers ranged from 3.4 Hz to 19.0 Hz. A peak in the intensity envelope of the target coincided with a valley in the envelope of the masker half-way through each stimulus. Thus, if listeners can do no more than take advantage of the maximum instantaneous difference in level between maskers and targets, thresholds should be constant across the five conditions. Alternatively, if listeners can use a difference in modulation rate to separate targets from maskers, thresholds should decline as the difference between the modulation rates increases. To obtain some generality, conditions were run in which maskers and targets had the same F_0 of 100 Hz and, separately, in which the F_0 of the maskers (105.9 Hz) was 1 semitone above the F_0 of the targets (100 Hz).

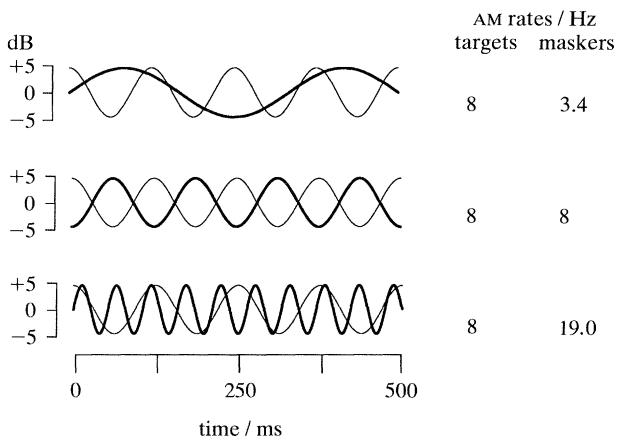

Figure 3. Amplitude envelopes of a subset of the maskers (thicker lines) and targets (thinner lines) used in experiment 2.

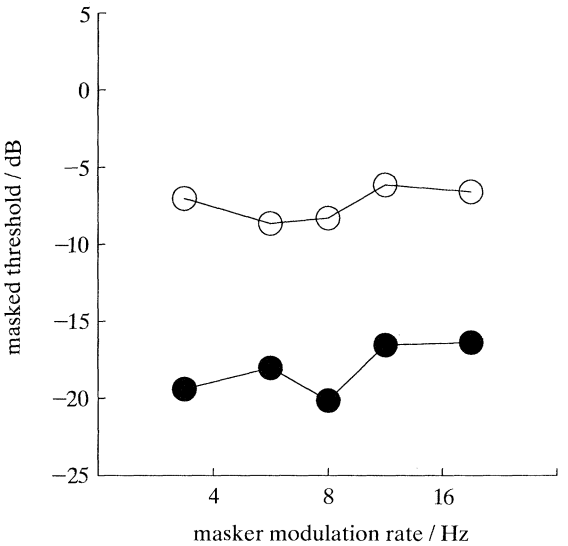

Figure 4. Results of experiment 2 for conditions in which the difference in F_0 between maskers and targets was 0 semitones (open circles) or 1 semitone (filled circles).

Mean thresholds from three listeners are shown in figure 4. As in experiment 1, there is a large effect of harmonicity; thresholds were 10–15 dB lower when maskers and targets possessed different F_0s. However, the results provide no evidence that listeners can use a difference in AM rate between two concurrently modulated 500 ms vowels to separate them perceptually. In fact, thresholds increased by 3–4 dB when maskers were given faster modulation rates than targets and had a different F_0. Thus, introducing a difference in AM rate slightly disrupted segregation.

7. EXPERIMENTS 3 AND 4: EFFECTS OF AMPLITUDE-MODULATION PHASE?

The modulating waveforms of maskers and targets in experiment 2 had different phases, even when they had the same 8 Hz rate. Thus, targets and maskers were always modulated incoherently. The incoherence itself might have permitted a material amount of segregation since the formants of one vowel were rising in amplitude at times when the formants of the other were falling. Accordingly, Experiment 3 asked whether a difference in modulator phase aids segregation when maskers and targets are modulated at the same rate.

Maskers and targets had a duration of 400 ms and were modulated at a rate of 2.5 Hz either coherently (in-phase) or incoherently (by advancing the phase of the masker modulation by 180°). Modulation depth (zero-peak) was varied from 1 dB to 5 dB. The first line of figure 5, labelled 'condition 1', shows the relationship between the amplitude envelopes of

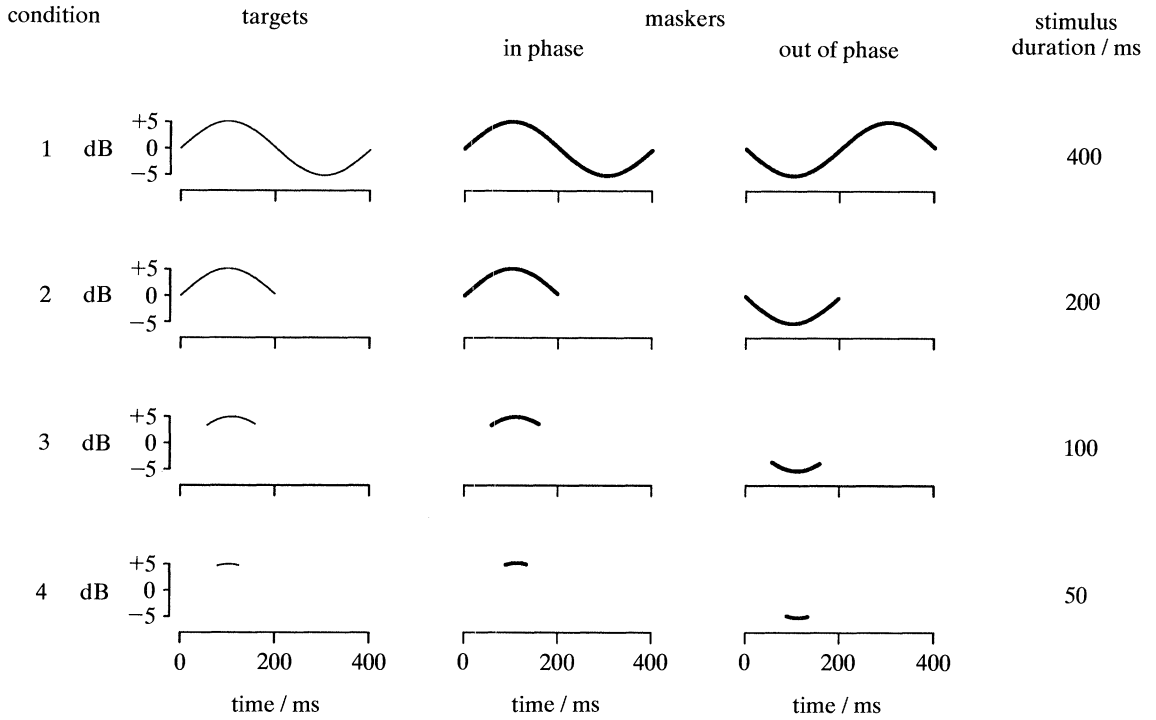

Figure 5. Amplitude envelopes of maskers (thicker lines) and targets (thinner lines). Experiment 3 involved condition 1 only. Experiment 4 involved all four conditions.

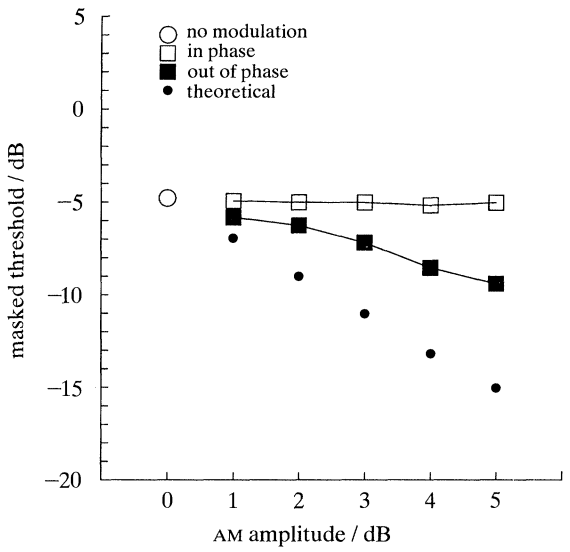

Figure 6. Results of experiment 3.

maskers and targets in the in-phase and out-of-phase conditions for the case where the modulation depth was 5 dB.

The experiment was intended to establish whether AM coherence plays an independent role in segregating voices. In other words, do the benefits of incoherent modulation exceed the advantages expected from the maximum instantaneous difference in level between maskers and targets that occurs during the combined stimulus?

Figure 6 shows thresholds obtained with in-phase stimuli (open squares) and out-of-phase stimuli (filled squares) as modulation depth increased from 1 dB to 5 dB. Thresholds were constant in the in-phase condition, but fell in the out-of-phase condition. The small filled circles have been plotted below the thresholds obtained in the in-phase condition (open squares) by an amount equal to the peak-to-peak modulation depth. Thus, the small circles plot the thresholds that would have occurred in the out-of-phase condition if listeners could take advantage of the maximum instantaneous difference in level between maskers and targets that occurred during the combined stimulus. In fact, the thresholds measured in the out-of-phase condition (filled squares) are higher than these theoretical points. Thus, not only could listeners not take advantage of AM incoherence, they could not even take advantage of the maximum instantaneous difference in level between maskers and targets.

A possible explanation for this result is based on the idea that a shorter duration of the targets was detectable in the out-of-phase condition than in the in-phase condition. Consider condition 1 in figure 5 again. In the in-phase condition, the local TMR is constant throughout the 400 ms duration of the combined stimulus. In the out-of-phase condition, in comparison, the local TMR varies over the duration of the stimulus. It is maximal, momentarily, at the point 100 ms after the start of the stimulus and is minimal at the 300 ms point. Imagine that a target is added to a masker at the same overall TMR in both the in-phase and out-of-phase conditions. The local TMR in the out-of-phase condition would be higher at the 100 ms point than it would be at any point in the in-phase condition. Hence, as was observed in figure 6, thresholds would be expected to be lower in the out-of-phase condition (filled squares) compared to the in-phase

condition (open squares). Now consider the situation that would arise if the overall TMR was reduced until the local TMR at the 100 ms point in the out-of-phase condition equalled the constant TMR observed at threshold in the in-phase condition. In this situation listeners would be able to detect only a brief segment of the target in the out-of-phase condition close to the 100 ms point in the stimuli. This segment would obviously be shorter than the 400 ms segment that could be detected in the in-phase condition. It is well established that performance in tasks requiring detection or discrimination deteriorates as the duration of the stimulus is reduced (e.g. Viemeister & Wakefield 1991). Hence, thresholds observed in the out-of-phase condition (filled squares in figure 6) would be higher than those predicted (small filled circles) from performance in the in-phase condition. In essence, listeners would not achieve predicted performance because they would be basing their judgements on a shorter effective duration of the targets in the out-of-phase condition than in the in-phase condition. We shall refer to this explanation as the 'time-intensity trading' account of the results of experiment 3.

Experiment 4 sought to verify the 'time-intensity trading' account and to establish whether another factor, described below, might also have played a role. The experiment compared performance in all four of the conditions illustrated in figure 5. Moving from condition 1 to condition 4, the stimuli were progressively restricted to a 50 ms segment centred on the point where the local TMR is maximal in the out-of-phase condition. The rationale is as follows. Suppose that the time-intensity trading account holds. In which case, in the out-of-phase conditions of experiment 3 listeners would have based their judgements on a brief segment of the target close to the 100 ms point in the stimuli. Let that segment have a duration of D ms. Thus, in experiment 4, thresholds should remain constant in the out-of-phase condition until the duration of the stimuli is reduced to a value less than D. In the in-phase conditions of experiment 3, in comparison, listeners could accumulate evidence of the targets over the full duration of the stimulus. Thus, here in the in-phase conditions of experiment 4, performance should suffer as stimulus duration is reduced and thresholds should rise.

The results‡ of experiment 4, shown in figure 7, are compatible with these predictions. As the duration of the stimuli was reduced, thresholds rose in the in-phase conditions (open squares) but stayed constant in the out-of-phase conditions (filled squares). The constancy of the thresholds in the out-of-phase conditions suggests that the duration D could be as short as 50 ms.

An additional factor which might have affected the results of experiment 3 is that listeners may not have known 'when to listen'. Because of MDI, or for some other reason, they might not have been able to select the moment in the stimuli when the TMR was maximal. Instead, they might have averaged evidence of the target over the part of the stimulus giving a reasonably good TMR. If so, performance in the out-of-phase conditions of experiment 4 might have been expected to improve as the duration of the stimuli was reduced, because listeners would be able to focus on the moment giving the maximal TMR. However, performance did not improve as stimulus duration was reduced (filled squares in figure 7). Thus, there is no evidence that MDI, or any other process mediated specifically by AM, significantly limited performance in experiment 3. Rather, the results of that experiment are explained by the time-intensity trading account.

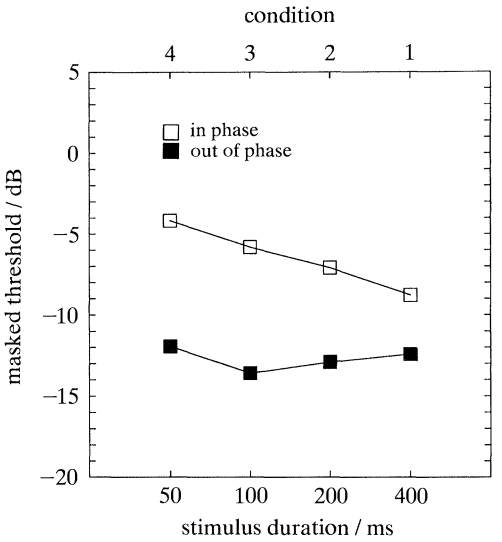

Figure 7. Results of experiment 4.

In summary, experiments 3 and 4 have shown that segregation of concurrent vowels is not facilitated by incoherent amplitude modulation. Establishing the generality of these conclusions, however, requires further experiments using faster modulation rates than the relatively slow (2.5 Hz) rate used here.

8. CONCLUSIONS

We have found no evidence that listeners can use coherent changes in the frequencies or amplitudes of the harmonics of a vowel to separate that vowel from a competing vowel whose harmonics are undergoing a different pattern of modulation. Neither form of 'common fate' was useful when both vowels were modulated. However, there were benefits when the components of one vowel were modulated in frequency while the components of the other were stationary. In addition, staggering the onsets of the vowels, so that one underwent an abrupt increase in amplitude while the other was static, was also beneficial. Thus, there is evidence that certain types of change in frequency or amplitude can help segregate the changing vowel from a static one. However, even these benefits were small in relation to the benefits from an absolute difference in F_0 between the vowels. The potency of harmonicity in relation to other cues for voice segregation has also been noted by Shackle-

‡ The 4 dB difference in overall performance level between experiments 3 and 4 can be attributed to the participation of different subjects in the two experiments.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

ton *et al.* (1991) who compared its benefits with those of binaural cues. Together, these results reinforce the idea (e.g. Stubbs & Summerfield 1991) that signal-processing approaches to voice segregation should exploit harmonicity as one of the primary cues.

We thank Mark Haggard, Chris Darwin and Bob Carlyon for constructive comments on drafts of this paper.

REFERENCES

Assmann, P.F. & Summerfield, Q. 1990 Modeling the perception of concurrent vowels: vowels with different fundamental frequencies. *J. acoust. Soc. Am.* **88**, 680–697.

Bacon, S.P. & Grantham, D.W. 1989 Modulation masking: effects of modulation frequency, depth, and phase. *J. acoust. Soc. Am.* **85**, 2575–2580.

Broadbent, D.E. & Ladefoged, P. 1957 On the fusion of sounds reaching different sense organs. *J. acoust. Soc. Am.* **29**, 708–710.

Carlyon, R.P. 1991 Discriminating between coherent and incoherent frequency modulation of complex tones. *J. acoust. Soc. Am.* **89**, 329–340.

Chalikia, M.H. & Bregman, A.S. 1989 The perceptual segregation of simultaneous auditory signals: Pulse train segregation and vowel segregation. *Percept. Psychophys.* **46**, 487–496.

Chalikia, M.H. & Bregman, A.S. 1992 The perceptual segregation of simultaneous vowels with harmonic, shifted, and random components. (In preparation.)

Darwin, C.J. 1981 Perceptual grouping of speech components differing in fundamental frequency and onset time. *Q. Jl exp. Psychol.* **33**A, 185–207.

Darwin, C.J. 1984 Perceiving vowels in the presence of another sound: constraints on formant perception. *J. acoust. Soc. Am.* **76**, 1636–1647.

Darwin, C.J. & Ciocca, V. 1992 Grouping in pitch perception: effects of onset asynchrony and ear of presentation. *J. acoust. Soc. Am.* (In the press.)

Darwin, C.J. & Culling, J.F. 1990 Speech perception seen through the ear. *Speech Communic.* **9**, 469–475.

Demany, L. & Semal, C. 1990 The effect of vibrato on the recognition of masked vowels. *Percept. Psychophys.* **48**, 436–444.

Gardner, R.B., Gaskill, S.A. & Darwin, C.J. 1989 Perceptual grouping of formants with static and dynamic differences in fundamental frequency. *J. acoust. Soc. Am.* **85**, 1329–1337.

Hall, J.W., Haggard, M.P. & Fernandes, M. 1984 Detection in noise by spectro-temporal pattern analysis. *J. acoust. Soc. Am.* **76**, 50–57.

Hall, J.W. & Grose, J.H. 1990 Comodulation masking release and auditory grouping. *J. acoust. Soc. Am.* **88**, 119–125.

Houtgast, T. 1989 Frequency selectivity in amplitude-modulation detection. *J. acoust. Soc. Am.* **85**, 1676–1680.

Klatt, D.H. 1980 Software for a cascade/parallel formant synthesizer. *J. acoust. Soc. Am.* **67**, 971–995.

Marin, C.M.H. & 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.

McAdams, S. 1989 Segregation of concurrent sounds. I: Effects of frequency modulation coherence. *J. acoust. Soc. Am.* **86**, 2148–2159.

Meddis, R. & Hewitt, M.J. 1991 Virtual pitch and phase sensitivity studied using a computer model of the auditory periphery: Pitch identification. *J. acoust. Soc. Am.* **89**, 2866–2882.

Meddis, R. & Hewitt, M.J. 1992 Modelling the identification of concurrent vowels with different fundamental frequencies. *J. acoust. Soc. Am.* **90**, 233–245.

Moore, B.C.J. 1992 Across-channel masking and co-modulation masking release. In *Audition, speech, and language* (ed. M. E. H. Schouten). Berlin: Mouton. (In the press.)

Moore, B.C.J., Glasberg, B.R., Gaunt, T. & Child, T. 1991 Across-channel masking of changes in modulation depth for amplitude- and frequency-modulated signals. *Q. Jl exp. Psychol.* **43**A, 327–348.

Scheffers, M.T.M. 1983 Sifting vowels: auditory pitch analysis and sound segregation. Ph.D. thesis, University of Groningen, The Netherlands.

Shackleton, T.M., Meddis, R. & Hewitt, M.J. 1992 The role of binaural cues in the identification of simultaneously presented vowels. *Q. Jl exp. Psychol.* (Submitted.)

Summerfield, Q. 1992 Roles of harmonicity and coherent frequency modulation in auditory grouping. In *Audition, speech, and language* (ed. M. E. H. Schouten). Berlin: Mouton. (In the press.)

Summerfield, Q. & Assmann, P.F. 1989 Auditory enhancement and the perception of concurrent vowels. *Percept. Psychophys.* **45**, 529–536.

Summerfield, Q. & Assmann, P.F. 1991 Perception of concurrent vowels: effects of harmonic misalignment and pitch-period asynchrony. *J. acoust. Soc. Am.* **89**, 1364–1377.

Summerfield, Q., Sidwell, A. & Nelson, T. 1987 Auditory enhancement of changes in spectral amplitude. *J. acoust. Soc. Am.* **81**, 700–708.

Stubbs, R.J. & Summerfield, Q. 1991 Effects of signal-to-noise ratio, signal periodicity, and degree of hearing impairment on the performance of voice-separation algorithms. *J. acoust. Soc. am.* **89**, 1383–1393.

Viemeister, N.F. & Wakefield, G.H. 1991 Temporal integration and multiple looks. *J. acoust. Soc. Am.* **90**, 858–865.

Yost, W.A. & Sheft, S. 1989 Across-critical-band processing of amplitude-modulated tones. *J. acoust. Soc. Am.* **85**, 848–857.

Yost, W.A., Sheft, S. & Opie, J. 1989 Modulation interference in detection and discrimination of amplitude modulation. *J. acoust. Soc. Am.* **86**, 2138–2147.

Zwicker, U.T. 1984 Auditory recognition of diotic and dichotic vowel pairs. *Speech Communic.* **3**, 265–277.

Discussion

A. J. Fourcin (*Department of Phonetics and Linguistics, University College London, U.K.*). One of the findings reported was that the listener makes no use of fundamental frequency contour – intonation related information – in segregating the outputs of two competing 'speakers'. This result may perhaps be modified if speech-like patterns are used. Although the normal listener has an excellent intrinsic knowledge of the intonation patterning of normal spoken language and can use this as a basis for speaker identification, phrase-level sinusoidal fundamental frequency or intonation contours will be foreign to his or her experience. A useful extension to the present experiments might come from the use of contours which are based on real utterances. These can be manipulated, so that they are of experimentally convenient centre fre-

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

quency and range, but conserved in respect of their idiosyncratic shapes (a technique which has been shown to preserve speaker identity information (Abberton & Fourcin 1978)).

A similar possibility for exploring the utility of using more natural, listener-experienced, stimuli comes from the employment of amplitude contour information which is speech derived. This also has been shown to be a source of speaker identification information (Atal 1968).

In both of these cases, the prediction is that the enhanced use of cognitive constraints coming from prior speech knowledge will enable listeners better to disentangle competing speech stimuli. More generally, it may always prove advantageous in exploring speech perceptual processing to pay close attention to the structure of speech itself.

References

Abberton, E. & Fourcin, A. 1978 Intonation and speaker identification. *Lang. Speech* **21**, 305–318.

Atal, B.S. 1968 Automatic speaker recognition based on pitch contours. Ph.D. thesis Polytechnic Institution of Brooklyn.

Q. Summerfield. It is important to distinguish low-level processes, which are used to group the harmonics of a single voice, from higher-level processes, which ensure that the groups of harmonics created by the low-level analyses are linked appropriately over time. A related distinction has been drawn by Bregman (1991) between primitive and schema-based grouping principles. Our experiments concern primitive processes reflecting physical constraints. Professor Fourcin's comment concerns schema-driven processes reflecting linguistic constraints.

The distinction can be illustrated by considering the computational problem faced by a system which attempts to separate sentences spoken concurrently by two talkers. A first step could be to locate harmonics. The next step would be to group the harmonics into two sets, one for each talker. The cues that might be used to do this include the following primitive grouping principles: (i) harmonicity: components whose frequencies are multiples of a common fundamental should be grouped together; (ii) coherent frequency modulation: components whose frequencies change in the same direction by the same percentage of their starting frequency should be grouped together; (iii) onset–offset synchrony: components that start and stop at the same time should be grouped together; (iv) coherent amplitude modulation: components whose amplitudes rise or fall coherently should be grouped together; and (v) concurrent change: components whose frequencies are changing should be grouped separately from components whose frequencies are static. The experiments described in our paper show that factors (i), (iii), and (v) can be used by listeners. We did not find evidence that listeners could use factors (ii) or (iv).

At this stage, the system has formed two groups of components at each of a succession of moments in time. The next problem is to string together the appropriate members of each group. For example, suppose that at time t_1 the system has established that two groups of harmonics are present, with F_0s of (i) 150 and (ii) 190 Hz, while at time t_2, there are two groups with F_0s of (iii) 170 and (iv) 160 Hz. The task now is to establish whether group (iii) is the continuation of group (i) or group (ii). It is at this stage that the linguistic schema-based principles of the type mentioned by Professor Fourcin are likely to play a role. For example, continuity of pitch would help to solve the problem of grouping (iii) with (i) or (ii), particularly if the continuity accorded with linguistic rules.

We believe that the best way to study the primitive principles is to use heavily constrained stimuli. Clearly, however, more natural stimuli should be used to study the schema-based linguistic principles, as Professor Fourcin suggests. By using the two approaches, Brokx & Nooteboom (1982) demonstrated that both primitive and schema-based principles play a role when listeners are required to identify words in sentences spoken by competing talkers.

References

Bregman, A.S. 1991 *Auditory scene analysis: the perceptual organisation of sound.* Cambridge, Massachusetts: MIT Press.

Brokx, J.P.L. & Nooteboom, S.G. 1982 Intonation and the perceptual separation of simultaneous voices. *J. Phonet.* **10**, 23–36.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES