The Auditory Processing of Speech

From Sounds to Words

Edited by

M. E. H. Schouten

pp 133-147.

Mouton de Gruyter
Berlin · New York 1992

ent vowels:
)-697.

), in: *The*
Granstrom

-state tones

Psych., 27,

D. Thesis,

tribution, *J.*

of sounds,

mants with
Am., 85, 3,

owels with

f individual
1853-1860.
ity patterns
sing, (W.A.

on, in: *The*
f.

Listening to two Things at Once

C.J. Darwin
Experimental Psychology
University of Sussex
Brighton BN1 9QG
U.K.

1. Introduction

Our auditory system allows us to listen to a particular sound source, largely unperturbed by the presence of other sounds. Understanding how it does this is a major challenge. At present, we know rather little of the mechanisms responsible for this ability, although there have been a number of demonstrations of factors that are important. This paper will outline different types of mechanism that might contribute and then describe some new experimental data aimed at illuminating the relationship between mechanisms of pitch perception and those used in auditory grouping.

2. Mechanisms for Source Separation

Albert Bregman's recent book "Auditory Scene Analysis" (Bregman, 1990, p. 408) distinguishes two types of mechanism that can be used to organise sound from more than one source. On the one hand the listener might use primitive grouping mechanisms that partition the input on the basis of simple stimulus properties. On the other hand, the listener might use schema-governed mechanisms that select an array of data that meets certain criteria. The primitive grouping mechanisms use simple general properties of a single sound source (such as common onset time or frequency continuity) to which the listener may be innately attuned and so do not depend on specific experience. The schemata that govern the second mechanism on the other hand are learned and so depend on the listener's specific experience. The first mechanism could be caricatured as bottom-up, the second as top-down.

Bregman's distinction has a parallel in computational approaches to speech processing. Primitive grouping mechanisms based on pitch have been used with some success to segregate the voices of different speakers (Parsons, 1976; Stubbs et al., 1988; Weintraub, 1987), but they are limited to the voiced portions of speech and have difficulty maintaining source continuity across silent or voiceless intervals. This failure highlights the basic problem with the purely bottom-up grouping approach for speech segregation: there are no simple properties that distinguish the components of one speaker's voice from those of another.

The alternative, top-down, schemata-based approach has been used in speech processing to recognise two simultaneous sounds by Roger Moore and his colleagues at the RSRE in Malvern (Moore et al., 1991). They have taken the interesting step of extending the Hidden Markov Model (HMM) approach in speech recognition to the case where there are two (known) sounds present at the same time. An HMM contains stochastic templates: the

program must first be trained on a representative selection of sounds that might occur, to allow it to develop a statistical model of each possible sound category. In recognition, there are a variety of algorithms that find which sound's model is most likely to have generated the presented sound. The RSRE group have extended the normal 'Viterbi' decoding algorithm to provide the correct 'maximum likelihood' recognition and optimal decomposition of the signal into two contributing sound sources. The algorithm is capable of performing speech recognition on two speakers simultaneously and shows substantial improvements in recognition accuracy for a single speaker against a background sound (such as a machine gun!) for which the program has a model.

The two approaches complement each other. The top-down approach produces good separation of speech from known sounds but fails on novel sounds that do not come close to any of the stored templates. It would gain rather little benefit from the presence of primitive grouping cues. By contrast, the bottom-up approach is agnostic as to the identity of the sounds that it is separating. It will work well if there are primitive grouping cues, and will fail if they are absent.

It is reasonably clear that both types of mechanism operate in human speech perception. Primitive grouping cues substantially improve recognition of one voice against a background of another (Brokx et al., 1982; Scheffers, 1983) and they also limit the possible interpretations that the listener makes of multiple sound-source data (Darwin, 1981; Darwin, 1991; Gardner et al., 1989). On the other hand, there is experimental evidence for a schema-driven mechanism that we use for selecting components that form possible speech sounds. Listeners can overcome inappropriate grouping cues to gain a clear speech percept (Cutting, 1976; Remez, 1987; Remez et al., 1981) and can identify remarkably well simultaneous pairs of known speech sounds which cannot be separated by primitive grouping cues (Scheffers, 1983). It is perhaps worth noting in passing that this schema-driven mechanism might not be learned (Mattingly et al., 1991).

3. Pitch Perception and Source Segregation

Difference in fundamental frequency is the best-documented primitive cue for the separation of different speech sound sources. The auditory system can exploit the fact that a periodic sound consists of frequency components that are integer multiples of the fundamental. Simultaneous speech sounds on different fundamentals are more easily identified than those on the same fundamental (Assmann et al., 1989; Assmann et al., 1990; Brokx et al., 1982; Scheffers, 1983); a single formant may be perceptually removed from a syllable if it is on a different fundamental from the rest of the syllable (Darwin, 1981; Darwin, 1991; Gardner et al., 1989) and a single harmonic may be perceptually removed from a steady vowel if it is not sufficiently close in frequency to a harmonic of the fundamental (Darwin et al., 1986).

Although the effect of fundamental frequency on speech sound source separation is indisputable, the mechanism or mechanisms by which fundamental frequency exercises its influence is less clear. Do similar mechanisms operate in the perception of the pitch of two simultaneous tones and in the perceptual segregation of two simultaneous voiced speech sounds?

I will first summarise previous experimental evidence on this question and then presen new data that we have recently obtained at Sussex, both on speech source separation and o pitch perception.

We have previously argued that the harmonic sieve (Duifhuis et al., 1982) might operat in sound source separation as well as in pitch peception (Moore et al., 1985) on the ground that similar amounts of mistuning are necessary for a harmonic to make a reduce contribution to both pitch and vowel quality (Darwin et al., 1986). The sieve (partly) block the (sufficiently) mistuned harmonic from both the subsequent pitch and the subsequent vowe classifiers.

The harmonic sieve, by definition, can only operate on resolved harmonics, and so i ineffective at separating unresolved harmonics. Harmonics may be unresolved because the are too close in frequency either to the next harmonic in the same series, or to a harmoni from a different series. High-numbered harmonics (above say the 8th) are excluded on th first criterion, so any source segregation that takes place for them must be occurring by som mechanism other than the harmonic sieve. Two periodic sounds whose fundamentals are ver close together will also have (low-numbered) harmonics that are close together in frequenc and so are unresolved from each other. For this reason the harmonic sieve cannot explain th improvement that occurs in identification scores for pairs of vowels when fundament frequency differences of a semitone or less are imposed on them.

Although the harmonic sieve cannot segregate unresolved harmonics, it is clear th sounds which have no harmonic structure (Burns et al., 1981) or consist only of unresolve harmonics (Houtsma et al., 1990) can give a sensation of pitch. The pitch sensation produce by these sounds is less clear and pitch discrimination is much poorer for these sounds tha for sounds that contain resolved components (Houtsma, 1984; Houtsma et al., 1990; Patterso et al., 1977). This weaker sensation of pitch is likely to be carried by the temporal pattern c amplitude modulation of the auditorially filtered waveform. Pitch discrimination becomes eve poorer for sounds that consist only of unresolved components when the phase of thes components is altered to reduce the depth of amplitude modulation of the auditorially filtere waveform (Houtsma et al., 1990).

At Sussex, we have recently carried out an experiment to see whether the distinctio between resolved and unresolved harmonics affects formant segregation in a similar way t pitch discrimination. Is it harder to segregate a formant when it only contains unresolve harmonics than when it contains resolved harmonics?

To address this question we used the "ru-li" experimental paradigm (Darwin, 198 Gardner et al., 1989). The basic sound used in these experiments is represented schematicall in Fig. 1. A three-formant "base" syllable /li/ has added to it an extra (second) formant whic is the F_2 of /ru/ to produce a composite four-formant syllable. When all four formants a excited by the same fundamental listeners report hearing /ru/, but if the second formant excited by a sufficiently different fundamental then listeners can hear the syllable /li/ with th second formant emerging as a separate sound source. With smaller differences in fundament frequency between the second and the other formants subjects report a "duplex" percept c a separate second formant but still /ru/. There are thus two stages in the segregation of formant. With small F_0 differences the formant stands out as a separate sound source but still phonetically integrated into the syllable to maintain the /ru/ percept; this stag

corresponds to "duplex" perception. With larger F_0 differences the formant phonetically segregates and the syllable is heard as /li/.

In our earlier experiments (Darwin, 1981; Gardner et al., 1989), the second formant generally had some resolved harmonics. The base syllable - formants 1,3&4 - was excited at 110 Hz and the second formant was excited over a range of values from 110 to 174 Hz. The second formant frequency varied between 780 and 1400 Hz. 780 Hz corresponds roughly to the 7th harmonic of 110 Hz and the 4th harmonic of 174 Hz, whereas 1400 Hz corresponds roughly to the 13th harmonic of 110 Hz and the 8th harmonic of 174 Hz.

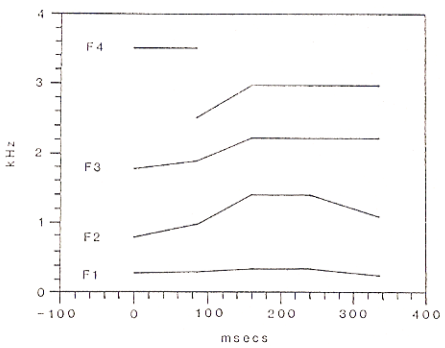

Fig. 1 Schematic spectrogram of formant tracks for a composite /ru/-/li/ syllable. Formants 1, 2 & 3 make /ru/ and formants 1, 3 & 4 make /li/. The percept of the whole syllable changes from /ru/ to /li/ when the fundamental frequency difference between the second formant and the rest is increased.

In this new experiment we used three different fundamentals for the base syllable: 80, 120 and 200 Hz, and then made the F_0 of F_2 either 0, 2.5, 5, 10, 20, 40 or 80% higher than the F_0 of the base. The experiment was run as a student project by Charles Edwardes on a variety of listeners - naive undergraduates and more experienced lab members. About half of the 12 listeners (all naive undergraduates) heard almost no /li/ sounds in the experiment and their data has been excluded. The remaining 5 more experienced listeners (though not more experienced in this task) gave very clear data which is shown in Fig. 2. The subject was given the choice of four responses to each stimulus. In each response category (except /li/ alone which we would expect to be almost never heard) there was a substantial effect of the F_0 of the base syllable. The pattern of results follows from the hypothesis that segregation is harder for unresolved than for resolved harmonics.

Fig. 2 shows the percentage of trials on which a particular response was given as a function of the percent F_0 difference between F_2 and the base syllable, with the base syllable's F_0 as parameter. In the top left panel the decline of "/ru/" responses with increasing F_0 difference is more rapid for the two higher base F_0s (200 and 120 Hz) than for the lowest F_0 (80 Hz). The decline in "/ru/" responses is complemented by an increase in both "/ru/+buzz" and in "/li/+buzz" responses. Here "buzz" refers to the sound of F_2.

With a base F_0 of 200 Hz, an increase of only 2.5% for the F_0 of F_2 is sufficient to give clear phonetic segregation with almost 100% "/li/+buzz" responses. Here the second formant frequency range corresponds to harmonics 4 through 7, all of which are clearly resolved.

With a base F_0 of 120 Hz, a more complex picture emerges. For small (2.5 and 5%) F_0 differences, the dominant response is a "duplex" one: subjects hear "/ru/+buzz". A difference of around 10% is needed before the phonetic segregation occurs and "/li/+buzz" becomes the dominant response. At an F_0 of 132 Hz (120+10%) the second formant frequency range

corresponds to harmonics 6 through 11, so there will be a mixture of resolved and unresolv harmonics.

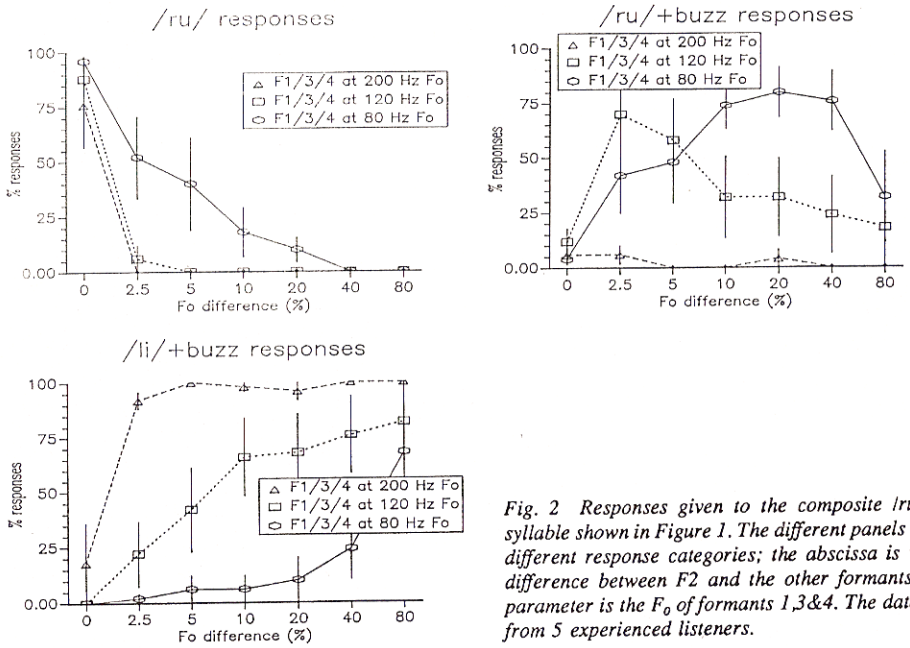

Fig. 2 Responses given to the composite /ru/-syllable shown in Figure 1. The different panels sh different response categories; the abscissa is % difference between F2 and the other formants; parameter is the F_0 of formants 1,3&4. The data from 5 experienced listeners.

With a base F_0 of 80 Hz the pattern of results is similar to that at 120 Hz, but larger differences are needed to give both types of segregation. Now the "/ru/+buzz" response do not dominate until there is a 10% F_0 difference, and the "/li/+buzz" response requires a f 80%. At 88 Hz (80+10%) the second formant frequency range corresponds to harmonics through 16 and so none would be resolved. However, at 144 Hz (80+80%) the second forma frequency range corresponds to harmonics 6 through 10, so again there will be a mixture resolved and unresolved harmonics.

In summary, phonetic segregation of a formant can be achieved with a much smaller difference when the segregated formant contains resolved harmonics than when it or contains unresolved harmonics. In fact there is no evidence from our experiment that phone segregation of a formant *can* be achieved at all when only unresolved harmonics specify t formant. Identification of a formant as a separate sound (i.e. "/ru/+buzz" responses) *can* achieved with only unresolved harmonics (as we have shown previously for F_4 segregatio The experiment thus extends to sound segregation the auditory system's greater sensitivity F_0 differences for resolved harmonics found by Houtsma & Smurzynski.

A consequence of this conclusion is that when two speech sounds have only a small difference, segregation should be much clearer in the first formant region than in the hig

formants. Such a difference has in fact already been found by John Culling. Culling started from the observation (Scheffers, 1983) that identification of pairs of vowels improves when there is a small (0.5 or 1 semitone) F_0 difference between the two vowels. In a series of experiments (Culling, 1990; Culling, 1991; Darwin et al., 1990) Culling demonstrated that this improvement in identification over the first semitone is due entirely to the first formant region. Higher formants only contribute to vowel separation when the F_0 difference is around 4 semitones.

In summary, we have shown that speech segregation shares with pitch perception sensitivity to the difference between resolved and unresolved harmonics. This common influence could arise because both pitch perception and speech segregation are subject to the same limitations of auditory coding, or it could arise because the process of sound segregation uses mechanisms common to pitch processing or even the results of pitch processing.

Carrying out experiments on the effect of fundamental frequency differences on speech segregation brought home to me how little work had been done on the perception of the pitch of simultaneous complex sounds. In particular we know almost nothing of how auditory primitive grouping cues influence pitch perception when there are multiple sound sources. Apart from the pioneering work of Beerends and Houtsma (Beerends, 1989; Beerends et al., 1986; Beerends et al., 1988; Beerends et al., 1989; Houtsma et al., 1984) and McAdams (McAdams, 1980; McAdams, 1984) some preliminary experiments by Elisabeth Cohen (Cohen, 1980) and work on the harmonic sieve (Duifhuis et al., 1982; Moore, 1987; Moore et al., 1985) neither pitch theorists nor experimenters have addressed the issue of how we perceive multiple pitches in complex sounds. The field is potentially a very rich one since it has important implications in both speech and music perception.

4. Auditory Grouping and Pitch Perception

Is pitch perception influenced by auditory grouping? How does the pitch perception mechanism determine which of the frequency components that are present at a particular time (or indeed across different times (Hall et al., 1981)) should contribute to a particular pitch percept? Over the last year or so Valter Ciocca and I (together with Roel Smits, Antonio Buffa, and Deirdre Williams) have investigated these questions exploiting a paradigm introduced by Brian Moore and his colleagues (Moore et al., 1985). We have asked the following questions:

1. Can a difference in onset time prevent a harmonic from contributing to the pitch of a complex sound?
2. Can a difference in amplitude modulation reduce a harmonic's contribution to the pitch of a complex sound?
3. Are sounds which arrive at the same ear (or from the same direction in space) more likely to contribute to a common pitch than those that do not?
4. Can a particular harmonic contribute only to the pitch of one of a number of simultaneous sounds? Alternatively: Does the Principle of Disjoint Allocation hold for pitch perception?

The paradigm allows the experimenter to measure the contribution that a particular harmonic makes to the pitch of a complex sound. The subject's task is to adjust the pitch of the second of two consecutive sounds until it matches the pitch of the first. The second sound is always strictly harmonic. The first sound typically is based on a harmonic series, but may have one of its components mistuned. The mistuned component may alter the pitch of the complex, and this alteration will be apparent in the fundamental frequency of the matched harmonic sound. The original work with this paradigm (Moore et al., 1985) demonstrated that a component mistuned by up to about 3% of its harmonic frequency made a full contribution to the pitch of the complex, but that above that its contribution declined until by about 8% mistuning it no longer made any contribution. T pitch shift produced by a mistuned component is thus a measure of the extent to which i integrated into the pitch percept.

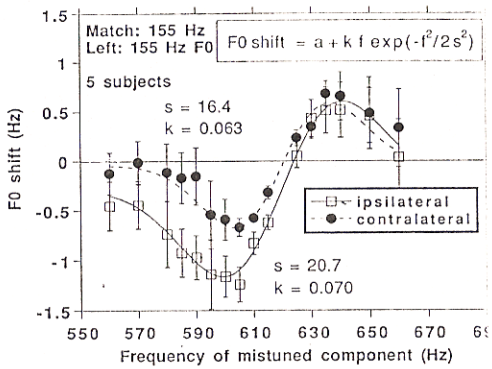

Fig. 3 Pitch shift produced by mistuning the fou harmonic of a 5-harmonic complex presented either to same or opposite ear as the rest of the complex.

We have replicated Moore's results with more values of mistuning than used by Mo (Darwin et al., 1991). Fig. 3 shows a subset of that data, using the first five harmonics o 155-Hz fundamental, with the fourth harmonic mistuned and presented either ipsi- contralaterally to the other harmonics. Here we have modelled the pitch shift by assuming t the contribution that a harmonic makes is reduced according to a Gaussian curve as frequency deviates from harmonic. The measured pitch shift from the in-tune value (ΔF_0 thus:

$$\Delta F_0 = a + k\ \Delta f\ \exp(-\Delta f^2/2s^2)$$

where Δf is the deviation of the mistuned component from harmonic, s is the stand deviation of the Gaussian (proportional to the width of the sieve's hole), k is a measure the maximum contribution the harmonic makes, and a allows for inaccuracy in measuring zero baseline for ΔF_0. This equation gives a good fit to our data. It would be difficult distinguish the fit of this equation from one that was linear for small mistunings.

4.1 Grouping by Onset Asynchrony

To illustrate the use of the paradigm let us look first at an experiment that Valter Ciocca and I recently carried out at Sussex. We were interested in whether a component that started before the others in a complex would make less of a contribution to the pitch of the complex than if it started simultaneously. We already know that onset time differences are a powerful cue for segregating an individual harmonic from a vowel percept (Darwin, 1984; Darwin et al., 1984) and for hearing out an individual component in a complex (Rasch, 1981; Rasch, 1984), but it is not clear whether they preclude a component from the pitch calculation.

In this experiment we used two different durations (80 ms and 410 ms) of a basic 12-harmonic complex on a fundamental frequency of 155 Hz. We used two different durations since Moore (Moore, 1987) found that a mistuned harmonic was more likely to be incorporated into the pitch of a short (50 ms) than a long (410 ms) complex (200 Hz F_0) tone. For each of the two durations we allowed the fourth harmonic to vary both in onset time and in frequency. The nominally 620-Hz component was mistuned by ± 0, 10, 20, 50 Hz, and could start 0, 30 or 300 ms before the remaining 12 harmonics. 6 subjects performed 5 matches to each of these stimuli, making their adjustments with a Rollerball attached to a MacII. The sounds were generated from the MacII in real time, using specially developed 56001 assembler code (Russell et al., 1991) running on a Digidesign Audiomedia™ card.

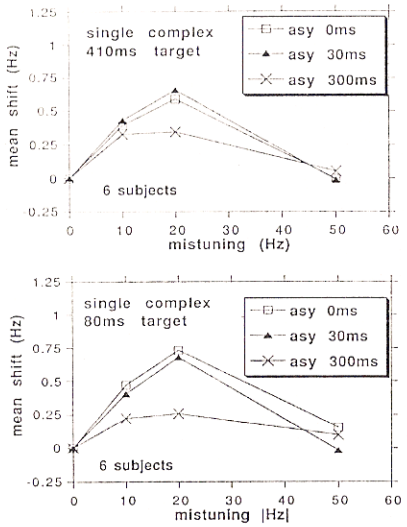

Fig. 4 Mean pitch shift produced by mistuning the 4th harmonic of a 12-harmonic 155 Hz complex tone. The parameter is the onset asynchrony between the 4th harmonic and the rest of the complex. The two panels show results at two different durations for 6 subjects.

The results of the experiment are shown in Fig. 4 as the mean matched pitch shift for each absolute amount of mistuning. This mean shift is calculated as half the difference between the matched pitches to the upward and the downward mistuned conditions. The results partly replicate Moore's findings and show a clear effect of a 300 ms onset asynchrony.

For both the 80 and 410 ms durations there is a large and statistically reliable pitch shift for the 0 and 30 ms asynchrony conditions which is greatest at 20 Hz (3%) mistuning, and has returned to zero at 50 Hz (8%). The 0 and 30 ms conditions do not differ at all from each other, and rather little across the two durations. There is very little evidence for the mistuned component being more integrated into the shorter than the longer sound as found by Moore (Moore, 1987).

300 ms onset asynchrony shows a clear effect. Pitch shifts are less with 300 ms than wi 0 or 30 ms asynchrony at both durations, but the reduction in pitch shift is greater for 80 r than for 410 ms duration.

In summary, a 300 ms onset asynchrony reduces the contribution that a mistun harmonic makes to the pitch of a complex tone. The reduction is greater for an 80 ms th for a 410 ms tone. The pitch of a complex tone, like the quality of a vowel, is not simp determined by the frequency components that are present at a particular time. It also depen on the time at which the sounds started.

Our results are compatible with experiments done using a different paradigm by V (1991). He found no effect of a 20 ms onset asynchrony on the perceptual separation of t simultaneous complex tones. Since natural musical instruments show quite substant differences in onset time of different harmonics (Risset et al., 1982) it is reassuring that t system is tolerant of some asynchrony between components.

We cannot say whether or not primitive auditory grouping mechanisms are operati before the pitch perception mechanism on the basis of these results, since they are al compatible with an explanation in terms of the rapid adaptation that occurs in the auditc response to a tone over its first few tens of milliseconds (Summerfield et al., 1987).

4.2 Grouping by Amplitude Modulation

There has been considerable and justified interest in the possible role that common amplitude modulation might play in auditory grouping, particularly in the context of comodulation masking release (Hall et al., 1990) and other across-channel masking phenomena (Moore et al., 1990; Yost et al., 1989). We have recently investigated whether a common pattern of amplitude modulation might provide a primitive grouping principle for components that belong to the same complex tone.

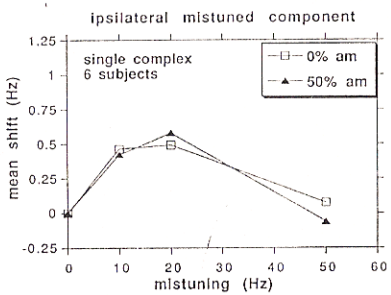

Fig. 5 Mean pitch shift produced by mistuning the harmonic of a 12-harmonic 155 Hz complex tone. T mistuned component could be 50% amplitude modula at 5 Hz; the other components were not modulated.

Our experiment used the same pitch matching paradigm with which we had found ons asynchrony effects. One component of a 410 ms, 155 Hz 12-harmonic complex was mistun and given either no amplitude modulation or 50% 5 Hz AM. The remaining 11 compone were not amplitude modulated.

The results for 6 subjects are shown in Fig. 5. Amplitude modulating the mistun component does not reduce its contribution to the pitch of a complex whose other compone are not amplitude modulated.

4.3 Grouping by Ear

Although the classical selective attention literature (Broadbent, 1958) showed that listeners can selectively attend more easily to spatially separated voices, spatial separation has not proved to be an overwhelming primitive grouping principle. Specifically, for identification of the pitch of two simultaneous 2-tone complexes, Beerends and Houtsma concluded that "Identification performance is only weakly dependent on the manner of distributing partials between the ears" (Beerends et al., 1989).

We have investigated the importance of ear of presentation in the mistuning paradigm. The simplest case is where all the in-tune components of a complex go to one ear, and the mistuned either to the same ear (ipsilateral) or the opposite ear (contralateral). We can then measure the extent of the pitch shift produced by the mistuned component as a function of the degree of mistuning and the distribution of the components across the ears. Fig. 3 shows the results of such an experiment conducted by Valter Ciocca and myself for a 5-component 155-Hz, 410 ms complex. There is a clear effect of ear: the mistuned component makes less of a contribution to the pitch of the complex when it goes to the opposite ear than when it goes to the same ear as the other, harmonic components. Although contralateral presentation clearly reduces the contribution of the mistuned component, it does still make a significant contribution to the perceived pitch.

4.4 Disjoint Allocation

In all the pitch matching experiments that I have described so far, there has been but a single basic harmonic complex. There has been no other complex sound competing for the mistuned component. We now turn to the more complex question of how we decide whether a particular component belongs to one or the other or even both of two possible complex tones. The basic design of the following experiments was conceived with Roel Smits (Darwin et al., 1990) and has since been extended with Valter Ciocca.

In these experiments we have set up the following situation: the listener is presented with two simultaneous complex tones, with fundamentals at, say, 155 Hz and 200 Hz. These two fundamentals have been chosen so that the third harmonic of 200 Hz (600Hz) and the fourth harmonic of 155 Hz (620 Hz) are separated by 3%. These two components, however, are replaced with a single component whose frequency can be varied. We then ask the listeners to match the pitch of either the 155 Hz or the 200 Hz complex as a function of the frequency of the single mistuned component.

The single mistuned component could potentially contribute to the pitch of either or both of the complex tones. To help clarify the issues here let us oppose two extreme hypotheses.

First, decisions about the pitches of the two complexes might be taken independently; all the available data is used for both decisions. This hypothesis predicts that the pattern of pitch matches to either complex tone is the same as we would expect if the other complex were absent. That is, the pitch shift should reach a maximum at 3% and decline to around zero at around 8%.

Second, decisions might be taken dependently; for example, if a particular component is exactly in tune with one harmonic series, it might be "captured" by that series and prevented

from contributing to the pitch of the other series. This hypothesis follows the Principle of Disjoint Allocation (Bregman et al., 1975). The particular frequencies that we have used were chosen with this hypothesis in mind. When the mistuned component is exactly in tune with the 200 Hz complex (at 600 Hz), it is potentially able to produce the largest pitch shift in the other complex. But if it is captured by the 200 Hz complex then it cannot contribute to the 155 Hz.

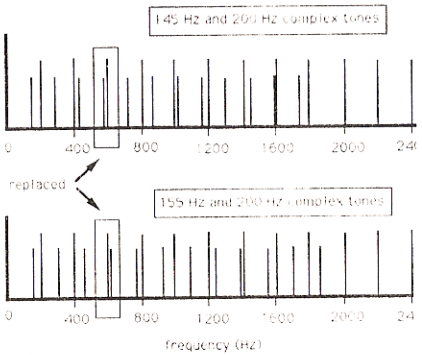

Fig. 6 Frequencies of harmonics of sounds used to investigate disjoint allocation in pitch perception. The two components enclosed in each box were replaced by a single variable-frequency component. All components had the same amplitude - the height difference here is used to visually separate the two harmonic series.

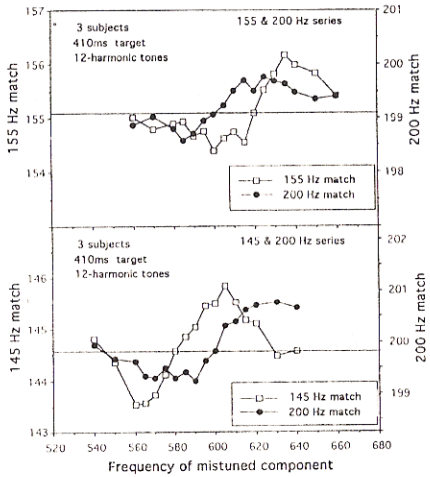

Fig. 7 Mean pitch matches averaged across three subjects to the double harmonic series shown in Figure 6. In the upper panel, subjects made a pitch match to either the 155 Hz or the 200 Hz series, in the lower to either the 145 Hz or 200 Hz series. Their matches are shown as deviation from their pitch match to the appropriate in-tune condition. The data shows that a frequency component that is exactly in tune with one harmonic series still contributes to the pitch of another harmonic series even though it is mistuned by 3% from this second series.

Our experiment in fact used two different pairs of fundamentals, 155 & 200 Hz as described above, and also 145 & 200 Hz (Fig. 6). The same design principles apply to this second pair of fundamentals. Our results clearly support the first hypothesis: decisions about the pitches of the two tones appear to be taken independently. In Fig. 7 we show the mean pitch matches for three subjects. The results are displayed differently from those in the preceding figures in that we are not averaging the upward and downward pitch shifts. The graphs are all aligned to the pitch match for the appropriate in tune condition. It is clear from

the data that a component that is precisely in-tune with one harmonic series can still make a full (mistuned) contribution to the pitch of another simultaneous complex. So for instance in the 145 & 200 Hz results, a 600 Hz mistuned component, though perfectly in tune with the 200 Hz complex, is producing a maximal pitch shift in the 145 Hz complex.

5. CONCLUSIONS

In this paper we have explored the relationship between pitch perception and speech separation. We have presented new experimental evidence that grouping for speech categorisation can take place on the basis of common harmonicity, and have extended previous findings by showing that this grouping is considerably less sensitive to F_0 differences when it is based on unresolved rather than on resolved harmonics. We have also presented new experimental evidence on auditory grouping in pitch perception itself. We have shown that both onset time and ear of presentation can be used as grouping cues for pitch. A mistuned component that starts at a different time from the rest of a sound makes less of a contribution to the pitch of that sound than if it had started at the same time. Similarly, a mistuned component that arrives at a different ear makes less of a contribution to the pitch of a complex than does one that arrives at the same ear. Amplitude modulation, by contrast, does not appear to act as a grouping cue for pitch. We have also demonstrated that the principle of Disjoint Allocation does not apply to pitch perception.

Acknowledgments

This research was supported by the Image Interpretation and Computational Science Initiatives of the S.E.R.C. through grants GR/F 34060 and GR/F 45356.

References

Assmann, P.F. and Summerfield, A.Q. (1989). Modelling the perception of concurrent vowels: Vowels with the same fundamental frequency, *J. Acoust. Soc. Am.*, 85, 327-338.

Assmann, P.F. and Summerfield, A.Q. (1990). Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies, *J. Acoust. Soc. Am.*, 88, 680-697.

Beerends, J.G. (1989). The influence of duration on the perception of pitch in single and simultaneous complex tones, *J. Acoust. Soc. Am.*, 86, 1835-1844.

Beerends, J.G. and Houtsma, A.J.M. (1986). Pitch identification of simultaneous dichotic two-tone complexes, *J. Acoust. Soc. Am.*, 80, 1048-1055.

Beerends, J.G. and Houtsma, A.J.M. (1988). The influence of duration on the perception of single and simultaneous two-tone complexes, in *Basic Issues in Hearing*, edited by H. Duifhuis, J.W. Horst, and H.P. Wit, Academic, London.

Beerends, J.G. and Houtsma, A.J.M. (1989). Pitch identification of simultaneous diotic and dichotic two-tone complexes, *J. Acoust. Soc. Am.*, 85, 813-819.

Bregman, A.S. (1990). *Auditory Scene Analysis: the perceptual organisation of sound*, Bradford Books, MIT Press, Cambridge, Mass.

Bregman, A.S. and Rudnicky, A. (1975). Auditory segregation: stream or streams?, *J. Exp Psychol: Hum. Perc. & Perf.*, 1, 263-267.

Broadbent, D.E. (1958). *Perception and Communication*, Pergamon, London.

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

Burns, E.M. and Viemeister, N.F. (1981). Played again SAM: Further observations on the pitch of amplitude-modulated noise, *J. Acoust. Soc. Am.*, 70, 1655-1660.

Cohen, E.A. (1980). *The influence of nonharmonic partials on tone perception*, Stanford University. Ph.D.

Culling, J.F. (1990). Exploring the conditions for perceptual separation of concurrent voices using F_0 differences, *Proc. I.O.A.*, 12, 559-566.

Culling, J.F. (1991). *The perceptual separation of concurrent vowels*, University of Sussex DPhil thesis.

Cutting, J.E. (1976). Auditory and linguistic processes in speech perception: inferences from six fusions in dichotic listening, *Psych Rev*, 83, 114-140.

Darwin, C.J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time, *Q Jl exp Psychol*, 33A, 185-208.

Darwin, C.J. (1984). Perceiving vowels in the presence of another sound: constraints on formant perception, *J. Acoust. Soc. Am.*, 76, 1636-47.

Darwin, C.J. (1991). The relationship between speech perception and the perception of other sounds, in *Modularity and the motor theory of speech perception*, edited by I.G Mattingly and M.G. Studdert-Kennedy, Erlbaum, Hillsvale, N.J., pp. 239-259.

Darwin, C.J., Buffa, A., and Smits, R.L.H.M. (1990). Pitch of simultaneous complex tone with a single mistuned frequency component, *Proc. I.O.A.*, 12, 499-506.

Darwin, C.J., Buffa, A., Williams, D., and Ciocca, V. (1991). Pitch of dichotic complex tone with a mistuned frequency component, in *Auditory physiology and perception*, edited by Y. Cazals, L. Demany, and K. Horner, Pergamon, Oxford.

Darwin, C.J. and Culling, J.F. (1990). Speech perception seen through the ear, *Speech Comm* 9, 469-475.

Darwin, C.J. and Gardner, R.B. (1986). Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality, *J. Acoust. Soc. Am.*, 79, 838-45.

Darwin, C.J. and Sutherland, N.S. (1984). Grouping frequency components of vowels: when is a harmonic not a harmonic?, *Q Jl Exper Psychol*, 36A, 193-208.

Duifhuis, H., Willems, L.F., and Sluyter, R.J. (1982). Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception, *J. Acoust. Soc. Am.*, 71 1568-1580.

Gardner, R.B., Gaskill, S.A., and 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. and Grose, J.H. (1990). Comodulation masking release and auditory grouping, *J Acoust. Soc. Am.*, 88, 119-125.

Hall, J.W. and Peters, R.W. (1981). Pitch from nonsimultaneous successive harmonics in quiet and noise, *J. Acoust. Soc. Am.*, 69, 509-513.

Houtsma, A.J.M. (1984). Pitch salience of various complex sounds, *Music Percepn.*, 1, 296-307.

Houtsma, A.J.M., Canning, J.M., and Beerends, J. (1984). A preliminary study of identification of harmonic intervals made by simultaneous complex tones, in: *Computational models of hearing and vision*, edited by Estonian Academy of Sciences, Tallinn, pp. 19-23.

Houtsma, A.J.M. and Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics, *J. Acoust. Soc. Am.*, 87, 304-310.

Mattingly, I.G. and Studdert-Kennedy, M.G. (Ed) (1991). *Modularity and the motor theory of speech perception*, Erlbaum, Hillsvale, N.J.

McAdams, S. (1980). The effects of spectral fusion on the perception of pitch for complex tones, *J. Acoust. Soc. Am.*, 68 S1, 109.

McAdams, S. (1984). *Spectral fusion, spectral parsing and the formation of auditory images*, Stanford University. unpublished Ph.D. dissertation.

Moore, B.C.J. (1987). The perception of inharmonic complex tones, in *Auditory processing of complex sounds*, edited by W.A. Yost and C.S. Watson (Erlbaum, Hillsdale, N.J.), pp. 180-189.

Moore, B.C.J., Glasberg, B.R., Gaunt, T., and Child, T. (1990). Across-channel masking of changes in modulation depth for amplitude- and frequency-modulated signals, *Q. Jl exp. Psychol.*

Moore, B.C.J., Glasberg, B.R., and Peters, R.W. (1985). Relative dominance of individual partials in determining the pitch of complex tones, *J. Acoust. Soc. Am.*, 77, 1853-60.

Moore, R.K., Varga, A.P., and Kadirkamanatha, M. (1991). *Automatic separation of speech and other complex sounds using hidden Markov model decomposition*, Institute of Acoustics Speech Group Meeting; February 1991, Brighton, Sussex.

Parsons, T.W. (1976). Separation of speech from interfering speech by means of harmonic selection, *J. Acoust. Soc. Am.*, 60, 656-60.

Patterson, R.D. and Johnson-Davies, D. (1977). Detection of a change in the pitch of AM noise, in *Psychophysics and physiology of hearing*, edited by E.F. Evans and J.P. Wilson, Academic Press, London.

Rasch, R. (1981). *Aspects of the perception and performance of polyphonic music*, Drukkerij Elinkwijk BV, Utrecht.

Rasch, R.A. (1984). The perception of simultaneous notes such as in polyphonic music, *Acustica*, 40, 22-33.

Remez, R.E. (1987). Units of organization and analysis in the perception of speech, in *The psychophysics of speech perception*, edited by M.E.H. Schouten, Martinus Nijhoff, Dordrecht, pp. 419-432.

Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.D. (1981). Speech perception without traditional speech cues, *Science*, 212, 947-950.

Risset, J.-C. and Wessel, D.L. (1982). Exploration of timbre by analysis and synthesis, in *The Psychology of Music*, edited by D. Deutsch, Academic, New York, pp. 25-58.

Russell, P. and Darwin, C.J. (1991). Real-time synthesis of complex sounds on a Mac II with 56001 DSP chip, *Brit. J. Audiol.*, 25, 59-60.

Scheffers, M.T. (1983). *Sifting vowels: Auditory pitch analysis and sound segregation*, Groningen University, The Netherlands. Ph.D.

Stubbs, R.J. and Summerfield, A.Q. (1988). Evaluation of two voice-separation algorithms using normally-hearing and hearing-impaired listeners, *J. Acoust. Soc. Am.*, 84, 1236-1249.

Summerfield, A.Q. and Assmann, P. (1987). Auditory enhancement in speech perception, in *The psychophysics of speech perception*, edited by M.E.H. Schouten, Martinus Nijhoff, Dordrecht, pp. 140-150.

Vos, J. (1991). *Perceptual separation in musical intervals with simultaneous complex tones: the effect of slightly asynchronous onsets*, Abstract submitted to Institute of Acoustics Speech Group Meeting; February 1991, Brighton, Sussex.

Weintraub, M. (1987). Sound separation and auditory perceptual organisation, in *The psychophysics of speech perception*, edited by M.E.H. Schouten, Martinus Nijhoff, Dordrecht, pp. 125-134.

Yost, W.A. and Sheft, S. (1989). Across-critical-band processing of amplitude-modulated tones, *J. Acoust. Soc. Am.*, 85, 848-857.