The role of temporal fine structure in pitch and speech perception

This dissertation is submitted for the degree of Doctor of Philosophy

Helen M. Jackson
St Catharine's College

September 2011

Auditory Perception Group
Department of Experimental Psychology
University of Cambridge

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

No part of this dissertation has already been, or is currently being submitted by the author for any other degree or diploma or other qualification.

The word count does not exceed 60000 excluding bibliography, figures, tables and appendices.

Typeset in Times New Roman with LaTeX.

Acknowledgements

A number of people have helped me with the completion of the research presented in this thesis. In particular, I wish to thank Brian Moore for his invaluable guidance throughout my research, for his patient responses to never-ending questions and for his incredibly quick provision of detailed feedback.

The Auditory Perception Group as a whole has contributed much both to my success and to my enjoyment of the last four years. Brian Glasberg has given up a great deal of time to help me, particularly with statistics; I am grateful to him for his friendship, kindness and good humour, and for understanding the critical importance of having good pens. Michael Stone and Kathryn Hopkins showed kindness and patience, gently prodding me in the right direction when I needed it. Simon Goldman, Sarah Creeke, Marina Salorio-Corbetto, Sarah Madsen, Hugh Greenish and Jo Robinson were the perfect labmates, offering listening ears, advice regarding grammar and cups of tea as necessary. All members of the extended group have had a role in my time in the lab, and, without giving a long list of names, I would like to thank them all.

St Catharine's College has supported me throughout both my undergraduate and postgraduate degrees. I am grateful not only for the academic and financial assistance I have received, but also, probably more importantly, for the thriving intellectual and social environment it has offered.

I would like to thank Deafness Research UK for funding the main part of this research. I would also like to thank the Cambridge Philosophical Society for awarding me funding for the final stages. St Catharine's College Research Fund kindly covered the cost of my use of Amazon Elastic Computing Cloud.

Simon Jackson, Kristian Glass, Ian Davies and Meredydd Luff have provided much practical help with the technical aspects of this project, for which I am extremely grateful.

Lastly, I thank my parents, grandparents and friends — particularly all those named above — for their love, advice and patience, and for putting up with the standard difficulties of having a close relationship with a graduate student.

The role of temporal fine structure in pitch and speech perception

Helen M. Jackson

This dissertation is submitted for the degree of Doctor of Philosophy

September 2011

A narrow band complex signal can be considered as having two parts: the envelope, which is the slow variation in amplitude with time, and the temporal fine structure (TFS), which is the fast-moving fluctuations of the signal within the overall envelope shape. Changes in the envelope of a signal are coded via changes in auditory nerve neurone firing rate, whereas changes in the TFS are coded via the precise timing of nerve spikes, which necessarily relies on phase locking. These temporal features, as well as spectral features of a signal, play a role in pitch perception and speech perception, as well as in the localisation of signals.

Some of the perceptual difficulties encountered by listeners with sensorineural hearing loss may be accounted for by a loss of sensitivity to the TFS of signals. These problems can include difficulty hearing speech in noise and ambiguous pitch. Not all hearing-impaired listeners lose sensitivity to TFS information to the same degree. It has been suggested that the magnitude of the deficit in TFS-processing ability for a given hearing-impaired listener could affect the best choice of hearing aid. Therefore it is important to understand fully the role of TFS in auditory perception to inform future technologies to help the hearing impaired.

Complex tone pitch perception is thought to depend on at least two mechanisms: pattern-recognition mechanisms, using the frequencies of individual low-

ranking components in the stimuli, and temporal mechanisms, using the overall periodicity of the signal derived from the interaction between high-ranking components. To use the information from individual components, they must be resolved on the basilar membrane: there must be an individual point of maximum vibration evoked by a given component, and this must be spatially distinct from those evoked by neighbouring components. The peripheral resolvability of components in a complex tone thus depends on the "sharpness" of auditory filters, and the degree of sharpness is a matter of current debate. An overview of the debate is given in chapter 2 of this thesis.

The role of TFS in the pitch perception of complex tones is also a matter of current debate, which begins with the observation that frequency shifting the components in a complex tone with harmonics below the 14^{th} produces a change in pitch with no accompanying change in envelope repetition rate: the pitch-shift effect. Since only the lowest 7 or 8 harmonics are thought to be resolved, this supports the idea that the use of TFS underlies good pitch perception for tones with harmonics in the range 8-14. TFS-processing ability is assumed to have an upper limit itself; above this, the envelope of the signal is the only remaining cue for pitch. However, some authors have proposed that the limit of peripheral resolution is higher, and hence that good pitch perception for intermediate harmonic ranks relies on some residual resolvability beyond the 8^{th} harmonic. A review of current issues in this area is also given in chapter 2 of this thesis.

A recently developed test was aimed at assessing the extent to which listeners can discriminate tones on the basis of their TFS alone, all other cues, such as excitation pattern and envelope cues, being unusable. The method was based on the pitch-shift effect described above: the same value in Hertz was added to the frequency of each component in a harmonic complex tone (H), resulting in an inharmonic tone (I). One interval in the two-alternative forced-choice task contained four H tones - "HHHH" - and the other contained alternating H and I tones - "HIHI". Subjects indicated which interval contained HIHI. The stimuli were bandpass filtered to reduce excitation pattern cues, and components were added in random phase to reduce envelope cues. This thesis provides evidence corroborating the claim that sensitivity to TFS can be measured in this way, presenting computer models in chapter 3 and data from human listeners in chapter

4. Chapter 5 uses another paradigm to show how listeners use TFS to determine the pitch of complex tones.

Pitch is not the only aspect of auditory processing where there is a role for TFS. Recently, it has been proposed that the ability to process TFS is useful for understanding speech, particularly against a fluctuating noise background. If this is the case, the intelligibility of a target speaker presented against a competing background speaker should be improved by an increase in F0 separation between the speakers when TFS information is present in the signal, but should not be when it is removed, via use of a vocoder. An experiment confirming this, as well as background information in this area, is presented in chapter 6.

To summarise, the aim of this thesis is to improve understanding of the role of TFS in the perception of complex signals such as tones and speech. It is shown that the sensitivity of an individual to TFS information in a signal can be measured directly. Measurements such as this also provide indirect evidence that the limit of peripheral resolution in the auditory system is no higher than the 7th or 8th harmonic. Furthermore, it is shown that TFS information enhances the intelligibility of target speech against a competing speaker. Conclusions on the ability of normal hearing listeners to process the TFS of signals are drawn.

Contents

List of Figures

Chapter 1

Introduction

A narrow band complex signal can be considered as having two parts: the envelope, which is the slow-moving general amplitude variation with time, and the temporal fine structure (TFS), which is the fast-moving fluctuations of the signal within the overall envelope shape. These are shown in Figure 1.1. Changes in the envelope of a signal are coded via changes in auditory nerve neurone firing rate, whereas changes in the TFS are coded via the precise timing of nerve spikes, which necessarily relies on phase-locking. These temporal features, as well as spectral features of a signal, play a role in pitch perception and speech perception, as well as in the localisation of signals.

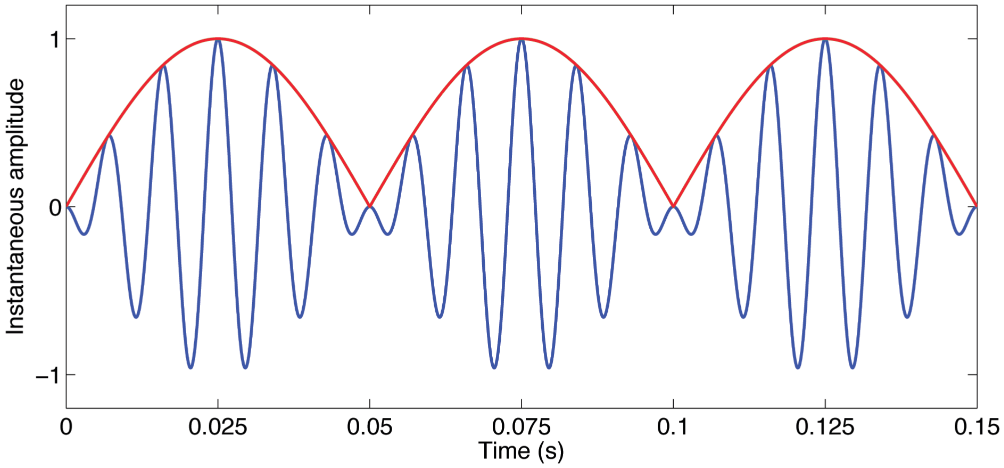


Figure 1.1: The TFS (blue line) and envelope (red line) of a complex signal.

There are a number of ways in which the envelope and TFS of a signal can be calculated mathematically. The envelope drawn in Figure 1.1 was calculated using a Hilbert transform. Another possible method for decomposition is half-wave rectification of the signal followed by lowpass filtering. It is important to note that the TFS and envelope of a signal are not entirely independent from one another: the envelope is correlated with the TFS to some extent. This point has important implications for psychophysical experiments, and will be returned to in chapter 6.

It has been proposed (Moore, 2008b) that some of the perceptual difficulties encountered by listeners with sensorineural hearing loss can be accounted for by a loss of sensitivity to the TFS of signals. These difficulties can include difficulty hearing speech in noise and ambiguous pitch. Not all hearing-impaired listeners lose sensitivity to TFS information to the same degree. It has been suggested that the magnitude of the deficit in TFS-processing ability of a given hearing-impaired listener could affect the best choice of hearing aid (Moore, 2008a; Moore and Sek, 2009a). Therefore it is important to understand fully the role of TFS in auditory perception in order to inform future development of technologies to help the hearing-impaired. In particular, it is important to identify a reliable method for determining the sensitivity of a given listener to TFS information quickly.

The pitch of a sinewave, or pure tone, may be extracted using both “place” mechanisms, utilising the tonotopicity of the basilar membrane to analyse spectral content, and “temporal” mechanisms, relying on the phase-locking of auditory nerve neurones to its TFS (a sinewave has a constant envelope corresponding to its amplitude). Increasing the frequency of a sinewave results in an increase in pitch, predicted by both these mechanisms. The upper frequency limit of human phase-locking to a pure tone is about 4–5 kHz (Kiang *et al.*, 1965; Palmer and Russell, 1986), though there is some debate as to the exact upper limit for complex tones in humans (Moore and Sek, 2009b; Moore and Ernst, 2012). Above this limit, pitch salience is low; the ability of subjects to perform octave matches decreases (Ward, 1954), as does the ability of subjects with absolute pitch to name intervals accurately (Ohgushi and Hatoh, 1991). The reason for this deterioration in phase-locking is that any inaccuracy in the initiation of the action potential in a neurone becomes a progressively larger proportion of the period as the frequency increases, so inter-spike intervals become a progressively less accurate representation of the period of the stimulus. More evidence for the role of temporal mechanisms comes from the observation that changing the level of a pure tone with a frequency of 1 kHz results in a shift in the peak in the pattern of excitation on the basilar membrane towards the base (high-frequency end) of the cochlea, whereas the corresponding perceptual shift varies in direction and magnitude across subjects.

For complex tones — those made up of more than one sinewave — the relative contributions of such place and temporal mechanisms have been a matter of debate spanning a number of decades. The pitch of a complex tone usually corresponds to its fundamental frequency (F0). Pattern-recognition models, such as those proposed by Goldstein (1973) and Terhardt (1974) assume that perceiving the pitch of a complex tone relies on being able to extract the individual frequencies of components in it. Here, frequencies of the individual components would be extracted using either the place or temporal mechanisms described for a pure tone. Then a central analyser would calculate a fundamental frequency either based on the observed harmonic series (Goldstein, 1973) or by generating sub-harmonics (Terhardt, 1974) and then looking for coincidences. In order to use the information from individual harmonics, they must be resolved on the basilar membrane: there must be an individual point of maximum vibration evoked by a given component, and this must be spatially distinct from those evoked by neighbouring components. The peripheral resolvability of components in a complex tone thus depends on the frequency selectivity of the basilar membrane, which depends on auditory filter shape. The "sharpness" of auditory filters — defined as the half-power bandwidth — and hence the limit of peripheral resolution of components in a complex tone, is a matter of current debate. An overview of this debate is given in chapter 2 of this thesis.

In contrast to these pattern-recognition theories, others (Schouten, 1940b, 1970; de Boer, 1956a) have proposed temporal mechanisms for pitch perception whereby the pitch is extracted from the envelope, or TFS resulting from the envelope of the interaction between higher-ranking components. It has been shown that auditory nerve neurones can phase-lock to the envelope or TFS of a signal, providing an accurate internal representation of its period (Cariani and Delgutte, 1996). The phenomenon of the "missing fundamental" can thus be explained by either pattern-recognition or temporal models of pitch perception. Complex tones containing low-ranking components have a more salient pitch (Plomp, 1967; Ritsma, 1967; Moore *et al.*, 1985; Houtsma and Smurzynski, 1990). Furthermore, experiments where components of a complex tone were presented to different ears (Houtsma and Goldstein, 1972) or successively in time (Hall and Peters, 1981) show that interference on the basilar membrane is not necessary for pitch perception. However, the fact that consecutive complex tones of high harmonic rank can evoke the clear perception of musical intervals (Houtsma and Smurzynski, 1990) or of a melody (Moore and Rosen, 1979) shows a clear role for temporal mechanisms in pitch perception.

The question of how the physical TFS of a signal is represented by a neural code for the pitch perception of complex tones is a matter of more debate, which begins with the observation that frequency shifting the components in a complex tone with harmonics of intermediate rank (between about 7 and 14) produces a change in pitch with no accompanying change in envelope repetition rate: the pitch-shift effect (Schouten *et al.*, 1962; Moore and Moore, 2003). Some authors (Hopkins and Moore, 2007; Moore *et al.*, 2009b) believe the limit of peripheral resolution to be about the 7$^{\text{th}}$ or 8$^{\text{th}}$ harmonic (for intermediate F0s), and support the idea that the use of TFS underlies good pitch perception above this rank, as documented in many studies (Houtsma and Smurzynski, 1990; Moore *et al.*, 2006; Moore and Sek, 2009a). TFS processing ability is assumed to have an upper limit itself; above this the envelope of the signal is the only remaining cue for pitch, and "classic" temporal mechanisms apply. Other authors prefer a more parsimonious account, avoiding a role for TFS other than for extracting the pitch of individual low-ranking components (Oxenham *et al.*, 2009; Shackleton and Carlyon, 1994). These explanations assume that the limit of peripheral resolution is higher, and hence that good pitch perception for intermediate harmonic ranks relies on some residual resolvability beyond the 8$^{\text{th}}$ harmonic. It is clear, then, that the debate surrounding resolvability and TFS is linked. A review of current issues in this area will also be given in chapter 2 of this thesis.

Recently, Moore and Sek (2009a) developed a test to determine a given listener's sensitivity to TFS. The aim of this test was to show that listeners could discriminate tones on the basis of their TFS alone, all other cues such as excitation-pattern and envelope cues being absent from the stimuli. The method they used was based on the pitch-shift effect (Moore and Moore, 2003; Hopkins and Moore, 2007) described above: the same value in Hertz was added to the frequency of each component in a harmonic complex tone (H) with a given F0 and harmonic rank, resulting in an inharmonic tone (I). One interval in the two-alternative forced-choice task contained four H tones — "HHHH" — and the other contained alternating H and I tones — "HIHI". Subjects were asked to indicate the interval containing the HIHI stimulus. The stimuli were bandpass filtered to reduce excitation-pattern cues, and components were added in random phase to reduce envelope cues. Moore and Sek proposed that this test could be used in a clinical setting to diagnose the early stages of sensorineural hearing loss.

This thesis will provide evidence to corroborate the claim that sensitivity to TFS can be measured in the way described by Moore and Sek, presenting computer models in chapter 3 and data from human listeners in chapter 4. It is shown that it is highly likely that listeners use TFS cues to discriminate the tones in Moore and Sek's task, and experiments investigating the sensitivity of normal-hearing listeners to TFS are presented. Chapter 5 uses another paradigm to show how normal-hearing listeners use TFS to extract the pitch of complex tones.

Pitch is not the only aspect of auditory processing where there is a role for TFS. Recently, it has been proposed that the ability to process TFS is useful for understanding speech, particularly against a fluctuating noise background (Hopkins *et al.*, 2008). Hopkins *et al.* proposed that the TFS of a signal can be glimpsed "in the dips" of the masker, facilitating auditory grouping. If this holds, intelligibility of a target speaker presented against a competing background speaker would be improved by an increase in F0 separation between the speakers when TFS information is present in the signal, but not to the same extent when it is not. An experiment confirming this, as well as background information in this area, is presented in chapter 6.

TFS also plays a role in localisation. The binaural localisation of pure tones requires detecting inter-aural phase differences (IPDs) between the two ears. To use IPDs, neurones must phase-lock accurately to the signal entering each cochlea, so the timing difference between the two ears is preserved and can be analysed at higher levels in the auditory pathway. If this were not possible, localisation of low-frequency pure tones would be difficult, as the contribution of interaural level differences (ILDs) to localisation of these frequencies is negligible due to diffraction around the head (Moore, 2003a). The ability of listeners to use IPDs could, as for Moore and Sek's task described above, be used in a clinical setting to diagnose the early stages of sensorineural hearing loss. Such a test has been developed recently by Hopkins and Moore (2010a).

To summarise, the aim of this thesis is to improve understanding of the role of TFS in the perception of complex signals such as tones and speech. It will be shown that the sensitivity of an individual to TFS information in a signal can be measured directly. Measurements such as this also provide indirect evidence that the limit of peripheral resolution in the auditory system is no higher than the 7th or 8th harmonic. Furthermore, it will be shown that TFS information enhances the intelligibility of target speech against a competing speaker. Conclusions on the ability of normal hearing listeners to process the TFS of signals will be drawn.

Chapter 2

A review of current issues regarding peripheral resolvability

2.1 Introduction

Discrimination of the fundamental frequency (F0) of complex tones is usually good when the tones contain low harmonics, but worsens when the number of the lowest harmonic, N, increases above about 7, reaching a plateau when N is about 14 (Hoekstra and Ritsma, 1977; Houtsma and Smurzynski, 1990; Bernstein and Oxenham, 2003; Moore *et al.*, 2006). The worsening as N is increased from 7 to about 14 has been interpreted by some as resulting from a progressive reduction of the ability to resolve the components in the complex tone (Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994): the resolvability hypothesis. However, others have interpreted the worsening as resulting from a progressive loss of the ability to use TFS information (Moore *et al.*, 2006; Hopkins and Moore, 2007; Ives and Patterson, 2008; Moore and Sek, 2009a; Moore *et al.*, 2009b): the TFS hypothesis. The decision as to which interpretation is more nearly correct depends on the extent to which harmonics with numbers in the range 7 to 14 are resolved, and judgments about this depend on the definition of resolvability and how it is measured.

This chapter presents some definitions of resolvability, and shows that any discussion of the TFS hypothesis depends on the bandwidth and sharpness of auditory filters.

2.2 Definitions of resolvability

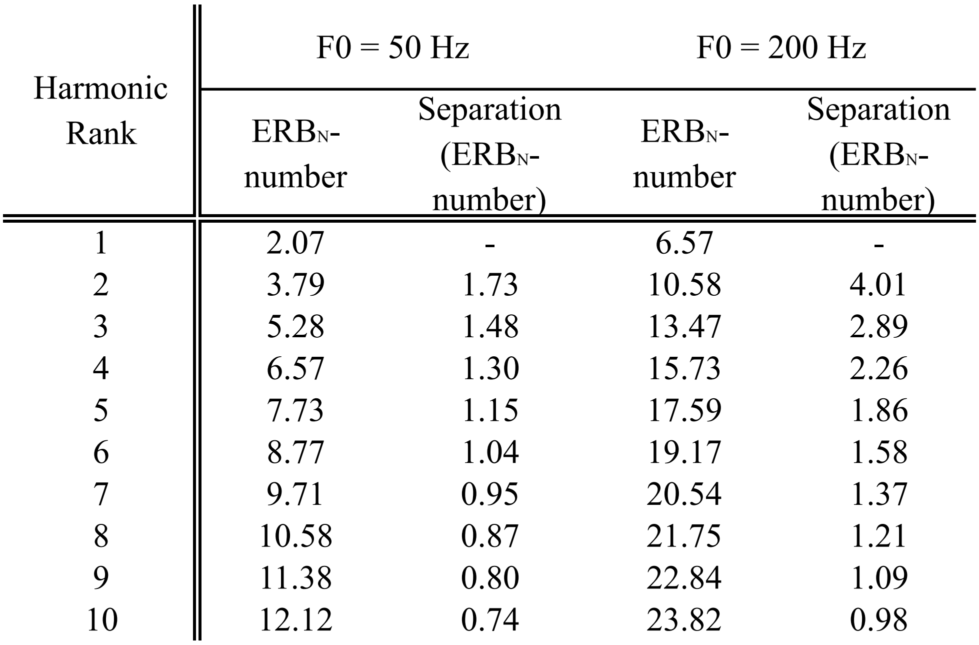
2.2.1 "Hearing out" of individual components

One definition of resolvability is connected with whether the sinusoidal components in a complex tone can be perceived individually or "heard out". The limits of this ability have been measured by asking listeners to compare the frequency of an isolated pure tone and a component in a harmonic (Plomp, 1964), or inharmonic (Moore and Ohgushi, 1993) complex tone. These studies showed that performance was good for harmonics with numbers up to 5–8, or, alternatively, that partials could be heard out when their spacing from their neighbours exceeded $1.25 * \mathrm{ERB_N}$, where $\mathrm{ERB_N}$ is the average value of the equivalent rectangular bandwidth of the auditory filter for normal-hearing listeners at moderate sound levels (Glasberg and Moore, 1990). Expressed another way, harmonics were heard out when their spacing from their neighbours exceeded 1.25 on the $\mathrm{ERB_N}$-number scale. Table 2.1 gives the separation in terms of $\mathrm{ERB_N}$-number between successive components of complex tones with fundamental frequencies of 50 Hz and 200 Hz, using the formula:

$$\mathrm{ERB_N}\text{-number} = 21.4log10(4.37f + 1) \tag{2.1}$$

where f is centre frequency expressed in kHz (Glasberg and Moore, 1990).

For the F0 of 200 Hz, it can be seen that the first six harmonics are separated from the adjacent harmonics by more than $1.25{*}\mathrm{ERB_N}$-number. The 7$^{\text{th}}$ harmonic is separated from the 8$^{\text{th}}$ harmonic by slightly less than $1.25 * \mathrm{ERB_N}$-number, and above this rank, the separation is much less than $1.25 * \mathrm{ERB_N}$-number. For the F0 of 50 Hz, only the first three harmonics are separated from the adjacent harmonics by more than $1.25 * \mathrm{ERB_N}$-number. This outcome appears to favour the TFS hypothesis: multiple studies have shown a progressive increase in F0DLs for tones where N is between 7 and 14 (Hoekstra and Ritsma, 1977; Houtsma and



Harmonic Rank	F0 = 50 Hz		F0 = 200 Hz	
	ERB$_\mathrm{N}$-number	Separation (ERB$_\mathrm{N}$-number)	ERB$_\mathrm{N}$-number	Separation (ERB$_\mathrm{N}$-number)
1	2.07	-	6.57	-
2	3.79	1.73	10.58	4.01
3	5.28	1.48	13.47	2.89
4	6.57	1.30	15.73	2.26
5	7.73	1.15	17.59	1.86
6	8.77	1.04	19.17	1.58
7	9.71	0.95	20.54	1.37
8	10.58	0.87	21.75	1.21
9	11.38	0.80	22.84	1.09
10	12.12	0.74	23.82	0.98

Table 2.1: ERB$_\mathrm{N}$-number separation between successive components of complex tones with fundamental frequencies of 50 Hz (third column) and 200 Hz (final column).

Smurzynski, 1990; Bernstein and Oxenham, 2003; Moore *et al.*, 2006), and if there is no peripheral resolution higher than the 7th component then this worsening cannot be due to a progressive reduction in resolvability. There has been much less work published for low F0s, but Moore, Hopkins, and Cuthbertson (2009b) showed a worsening in performance between $N = 7$ and $N = 13$ for an F0 of 50 Hz. This clearly favours the TFS hypothesis. Chapter 5 of this thesis will present some experiments using low F0s in order to provide more evidence.

A modification of the above method was introduced by Bernstein and Oxenham (2003). They pulsed the target harmonic in the complex tone on and off to resolve perceptual confusion effects and to make it more clear to the listener which component in the complex tone was to be compared with the isolated pure tone. Using this method, they found that harmonics up to about the 10th could be heard out, which is more in favour of the resolvability hypothesis; it is assumed that harmonics can only be heard out if discrete information from each one can be extracted from the basilar membrane. Even more recently, the method used by Bernstein and Oxenham has been criticised on the grounds that the results may have been influenced by adaptation in the auditory nerve (Moore *et al.*, 2009a, 2012). Overall, it seems likely that only harmonics with numbers up to about 7 or 8 can be heard out from complex tones with equal-amplitude, contiguous harmonics, but the exact upper limit remains in some doubt.

2.2.2 Ripples in the excitation pattern on the basilar membrane

A second approach to defining resolvability is based on calculation of the excitation pattern of a complex tone (Glasberg and Moore, 1990; Moore *et al.*, 1997). For complex tones with equal-amplitude harmonics, ripples in the excitation pattern occur, with peaks corresponding to the frequencies of the individual harmonics. The peak-to-valley ratio of the ripples decreases with increasing harmonic number, and only the lowest harmonics evoke distinct ripples. Using the model for auditory filter shapes presented by Glasberg and Moore (1990), which is based on measurements using a notched-noise masker (Patterson, 1976; Patterson and Nimmo-Smith, 1980), the excitation patterns in Figure 2.2 are obtained.

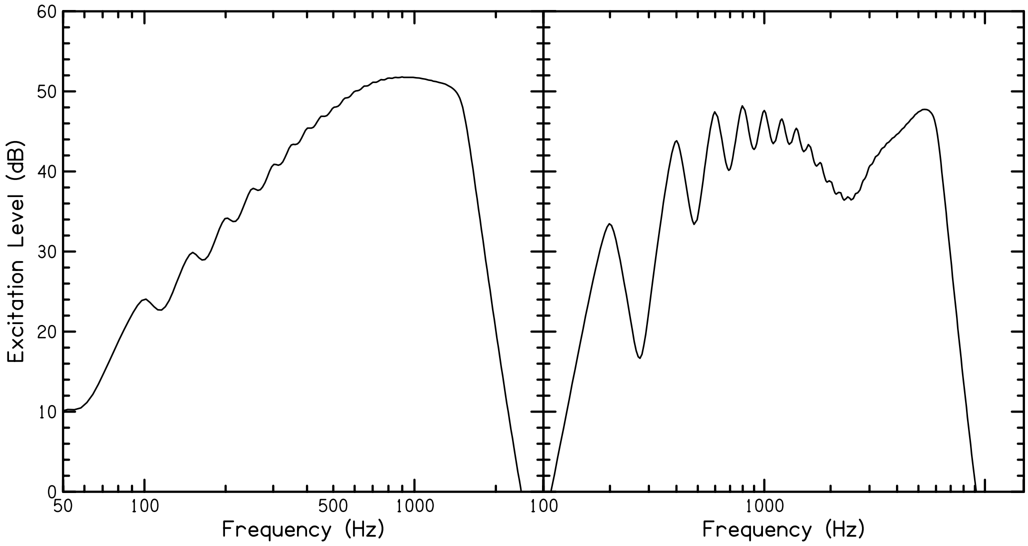


Figure 2.2: Excitation patterns for a complex tone containing the first 30 harmonics, each with a level of 50 dB SPL, generated using the response of HD580 headphones. Auditory filters were as specified by Glasberg and Moore (1990). Left panel: F0 = 50 Hz. Right panel: F0 = 200 Hz.

The right panel of Figure 2.2 shows the excitation pattern for a complex tone with an F0 of 200 Hz containing the first 30 harmonics, each with a level of 50 dB SPL. It is thought that ripples need to have a magnitude exceeding about 2 dB for them to be detectable (Moore *et al.*, 1989; Moore and Sek, 1994; Buus and Florentine, 1995). Ripple depth here is defined as the level difference between a given peak and the average of the two neighbouring troughs, after the suggestion of Ives and Patterson (2008). The two neighbouring troughs are those located on either side of the peak. In the right panel of Figure 2.2, only the lowest seven ripples are larger than 2 dB. However, it is possible that ripples with a somewhat smaller depth can be detected when they have a regular pattern and extend over a wide frequency range (Green *et al.*, 1987; Bernstein and Green, 1987). This possibility is assessed in chapters 3 and 4 of this thesis using a computer model.

For low F0s, even fewer components evoke ripples larger than 2 dB. The left panel of Figure 2.2 shows the excitation pattern for a 30-harmonic complex tone with an F0 of 50 Hz. As before, the level of each component was 50 dB SPL. Here, only the lowest five ripples are larger than 2 dB.

This approach leads to the same conclusion as that reached from studies of the ability to hear out harmonics; harmonics above the 8th are not resolved, which supports the TFS hypothesis. However, the calculation of excitation patterns is based on assumptions about the sharpness and shape of the underlying auditory filters. The auditory filters may be sharper than assumed in the model of Glasberg and Moore (1990), as demonstrated by measures of auditory filter shape obtained using forward masking (Moore and Glasberg, 1981; Oxenham and Shera, 2003; Oxenham and Simonson, 2006), which avoid effects of two-tone suppression. Suppression is the process whereby excitation at one frequency on the basilar membrane is reduced by the presence of excitation at adjacent frequencies (Moore, 2003a), and is avoided in non simultaneous masking paradigms. Therefore, the use of forward masking results in better estimates of the frequency selectivity of the auditory system, as experiments using complex tones are subject to the effects of suppression. Short-term adaptation to energy at a particular frequency enhances frequency resolution (Bacon and Jesteadt, 1987), and effects of adaptation would also be present in experiments using complex tones.

A simplified formula for the shape of an auditory filter following the roex (rounded exponential, Patterson and Nimmo-Smith, 1980) model is:

$$W(g) = (1 + pg)e^{-pg} \tag{2.2}$$

where g is the deviation from the centre frequency of the filter divided by the centre frequency, and p is a parameter determining the slope of the filter skirts, which defines the sharpness of the filter. Doubling the value of p results in a filter that is twice as sharp as the original and that has an ERB equal to half the original, because:

$$\mathrm{ERB} = (f_c * 4)/p \tag{2.3}$$

where f_c is the centre frequency of the filter. Doubling the sharpness of the auditory filters in the model of Glasberg and Moore (1990) would lead to better resolution of higher harmonics. Excitation patterns for the same tones as described above using auditory filters that are twice as sharp as normal are given in Figure 2.3. From this figure, it can be seen that the lowest 16 components of the 200-Hz F0 (right panel) and the lowest 11 components of the 50-Hz F0 (left panel) evoke ripples deeper than 2 dB.

On the other hand, the auditory filters may have broader tips than assumed in the model of Glasberg and Moore (1990), as proposed by Unoki *et al.* (2007). This would lead to poorer resolution of higher harmonics. Thus, for this reason, and due to the opposing effects of suppression and adaptation on frequency selectivity in simultaneous masking experiments described above, the exact upper limit of resolvability as estimated from excitation patterns remains in some doubt.

2.2.3 The effect of relative phase

A third approach to defining resolvability concerns effects of the relative phases of the components in a harmonic complex tone. If components are unresolved, the waveform describing the vibration of the basilar membrane is approximately equivalent to summing the waveforms of those components (ignoring nonlinear

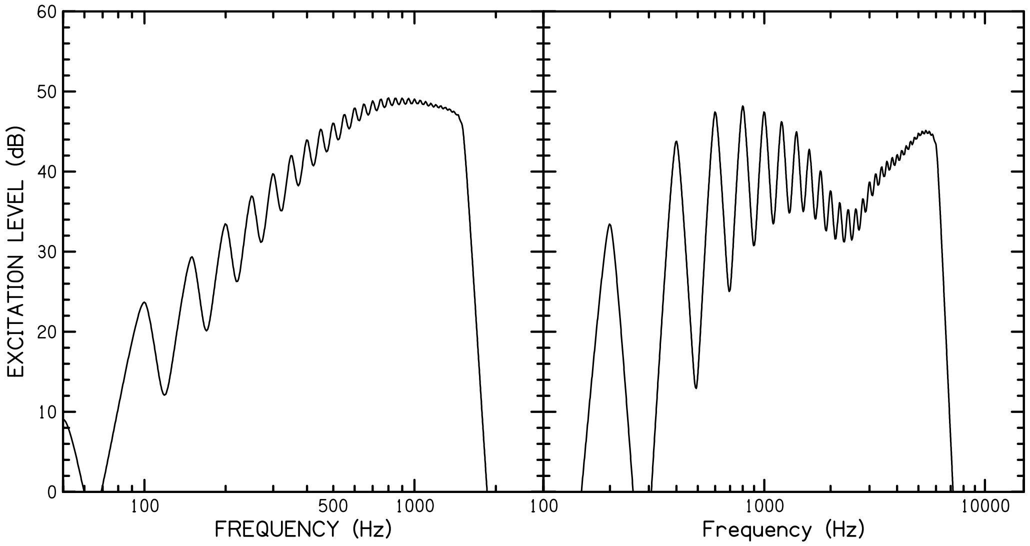


Figure 2.3: Excitation patterns for a complex tone containing the first 30 harmonics, each with a level of 50 dB SPL, generated using the response of HD580 headphones. Auditory filters were twice as sharp as specified by Glasberg and Moore (1990). Left panel: F0 = 50 Hz. Right panel: F0 = 200 Hz.

effects). When components are summed in cosine phase, there will be a large peak in the waveform at the beginning of each period to which auditory nerve neurones can phase lock accurately. This large peak at the beginning of each period is not present when components are summed in random phase. The ability to phase lock accurately to a signal improves pitch perception. It is usually assumed that phase effects do not occur when all harmonics are resolved, but phase effects may occur when two or more harmonics are unresolved (Patterson, 1973; Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994; Moore, 1997; Bernstein and Oxenham, 2003). Moore *et al.* (2006) found that F0 difference limens (F0DLs) were smaller for groups of harmonics added in cosine phase than for harmonics added in alternating phase when the lowest component present in the stimulus, N, was above the 8th. For alternating phase, successive components are alternately added in sine and cosine phase, resulting in a fairly flat envelope. The phase effect was taken by Moore *et al.* as indicating that the harmonics were not resolved. However, F0DLs remained relatively low for N in the range 8–11. Moore *et al.* argued that these low F0DLs supported the TFS hypothesis, as the phase effect suggested that the harmonics were unresolved.

Oxenham *et al.* (2009) suggested that the F0DLs measured by Moore *et al.* were affected by distortion products. Oxenham *et al.* replicated the experiment of Moore *et al.* using the same level of background threshold-equalizing noise (TEN, Moore *et al.*, 2000) as in the original study, and using a level that was 20 dB higher, which they argued would render any distortion products completely inaudible. The results with the original noise level replicated the findings of Moore *et al.*, but use of the higher noise level eliminated the phase effect for values of N for which F0DLs were low. While the presence of a phase effect indicates that at least some harmonics are unresolved, the presence of unresolved harmonics will not always result in a phase effect. Thus, all that can be concluded from these results is that they provide no positive evidence for the use of TFS from groups of unresolved harmonics.

It should be noted here that there is no difference in the low pitch perceived for complex tones added in random and cosine phase, as demonstrated by Patterson and Wightman (1976) amongst others. This lack of difference is predicted by autocorrelation models such as that of Meddis and Hewitt (1991). There is however a timbre difference between complex tones added in differing phases (Patterson, 1987; Plomp and Steeneken, 1969).

2.3 Conclusions

To summarise, if models of auditory filter shape based on simultaneous masking are accepted, it seems likely that harmonics above the 7^{th} or 8^{th} are not resolved peripherally for intermediate F0s. Calculated excitation patterns show very shallow ripples corresponding to individual harmonics above the 7^{th} for an F0 of 200 Hz. However, it is possible that the auditory filters are sharper than indicated by simultaneous masking, in which case harmonics above the 7^{th} might be partially resolved.

Direct measures of the ability to hear out partials suggest that only harmonics up to about the 8^{th} can be heard out. The only study that showed some ability to hear out any individual harmonics above this rank was shown to have been influenced by release from adaptation.

Overall, it is possible, even likely, that the use of TFS may underlie good pitch perception for complex tones containing only harmonic ranks above the 7^{th}. Chapters 4 and 5 of this thesis will present experimental evidence from human listeners suggesting that this is indeed the case.

Chapter 3

Computational modelling of frequency-shifted complex tone discrimination

3.1 The procedure of Moore and Sek (2009)

The procedure of Moore and Sek (2009a), described in the introduction, provides a potential direct estimate of TFS sensitivity by measuring thresholds for discriminating between harmonic and frequency-shifted tones. To make the stimuli, a large number of components were synthesised and then bandpass filtered. The bandpass filter had a flat region with a width that was always 5F0, and skirts that decreased at a rate of 30 dB/octave. The effect of these skirts was to reduce excitation pattern changes by avoiding the sudden presence or absence of a frequency component close to the edge frequency of the flat part of the passband as a result of the frequency shift. Excitation-pattern cues were further reduced by presenting the stimuli against a background of threshold-equalising noise (TEN, Moore *et al.*, 2000). The overall level of the tones was 20 dB SL, and the level of the TEN (in dB/ERB$_\mathrm{N}$) was set to 15 dB below this total level. Each stimulus interval consisted of four tones: either four harmonic tones (HHHH) or alternating harmonic and inharmonic tones (HIHI). Subjects were asked to indicate which interval contained the HIHI pattern. It has been shown previously that frequency-shifted tones usually have a slightly different pitch from their harmonic counterparts (Schouten, 1940a; de Boer, 1956b; Schouten *et al.*, 1962; Patterson, 1973; Moore and Moore, 2003). Hence, it is assumed that subjects performed the task by deciding in which interval the pitch of the tones went up and down; this is consistent with subjective reports of the subjects. The components were added together in random phase, resulting in waveforms whose envelope varied randomly in shape across tones. The envelope repetition rate was the same for all tones, based on the Hilbert transform. This reduced the likelihood that subjects performed the task using temporal envelope (TE) cues. For an F0 of 200 Hz, when the harmonic ranks of components in the flat part of the passband of the stimulus were 9–13, all 10 subjects tested by Moore and Sek achieved a threshold of less than 0.2F0, indicating good performance.

3.2 Excitation-pattern cues

3.2.1 Single-ripple discrimination

Although excitation-pattern cues in this task were minimised, it is still possible that there were ripples in the excitation pattern that could be compared between stimuli and that might have signalled the HIHI interval. Figure 3.1 shows the excitation patterns for examples of the average of two H tones, the first and third tones in an interval (solid line), and the average of two I tones (dashed line), the second and fourth tones in an interval, for a condition tested by Moore and Sek with the maximum possible frequency shift (0.5F0). This excitation pattern was calculated using the model described by Glasberg and Moore (1990), using input signals that were generated by Moore and Sek's software. The nominal F0 was 200 Hz, the lowest component in the flat part of the passband, N, was the 9$^{\text{th}}$, and the width of the flat part of the bandpass filter was 5F0. Distinct ripples in the excitation pattern can be seen corresponding to individual components, but the largest difference between the excitation patterns of the H and I tones is 1.35 dB. As outlined in chapter 2, it is thought that a ripple must have a magnitude exceeding 2 dB for it to be detectable (Moore *et al.*, 1989; Moore and Sek, 1994; Buus and Florentine, 1995). Therefore it is unlikely that a single ripple was used to discriminate the tones in this task. Note that the change in excitation level would be much smaller than 1.35 dB for frequency shifts corresponding to the threshold value (rather than 0.5F0), as for most conditions tested threshold was well below 0.5F0.

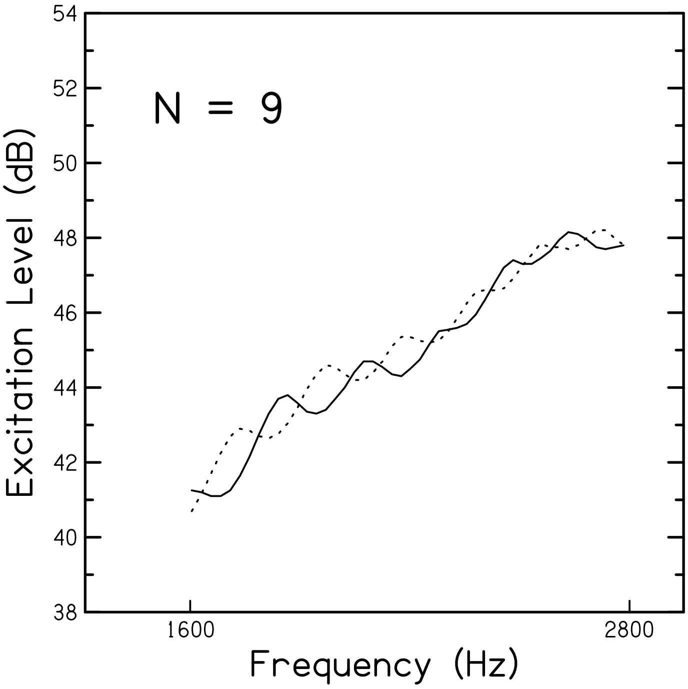


Figure 3.1: Excitation patterns for examples of the average of two H tones, the first and third tones in an interval (solid line), and the average of two I tones, the second and fourth tones in an interval (dashed line), for a condition tested by Moore and Sek with the maximum possible frequency shift (0.5F0). F0 was 200 Hz, the lowest component in the flat part of the passband, N, was the 9th, and the width of the flat part of the bandpass filter was 5F0.

More evidence that single ripples in the excitation pattern were unlikely to have been used in this task was provided by Moore and Sek themselves. Moore and Sek (2009b) measured the threshold for level discrimination of the 11th component in a complex tone where F0 was 800 Hz and N was 12. It is likely that the 11th component was the lowest (and hence potentially the most resolvable) audible component in the stimulus after passing through the bandpass filter. They found that the mean level difference at threshold was 5.3 dB, giving a corresponding change in the excitation pattern of 3.8 dB at the point where the change was largest. However, the largest difference in excitation level between the H and I tones for a frequency shift equal to the measured threshold was 0.6 dB. They concluded that it was unlikely that subjects were able to use such a small difference in excitation level to perform the task.

3.2.2 Discrimination of a regular pattern of ripples

Profile-analysis studies (Green *et al.*, 1987; Bernstein and Green, 1987) have suggested that a regularly spaced pattern of ripples in an excitation pattern, for example as evoked by a stimulus with log-spaced components in which every second component is incremented in level, is more detectable than a single ripple. Micheyl, Dai, and Oxenham (2010) have proposed that the smaller thresholds for level discrimination of such "up-down-up-down" spectra relative to a "single incremented component" could be due either to auditory filters being narrower than assumed in the model of Glasberg and Moore (1990) or to correlation of internal noise across auditory channels. These suggestions are not mutually exclusive. It could also be the case that information can be combined across frequency channels, such that the presence of more ripples lead to better discrimination. Micheyl *et al.* (2010) suggested that if subjects were using a "single largest difference"-based decision rule to discriminate the frequency shifts in Moore and Sek's (2009a) experiment, then discrimination performance would be approximately proportional to the magnitude of the difference. Using auditory filters that are $0.75 * \mathrm{ERB_N}$ in width instead of $\mathrm{ERB_N}$ increases the magnitude of the largest difference for a given frequency shift, allowing a smaller frequency shift to be discriminated. They further suggest that internal noise, caused by slight inaccuracies in neural coding, would be correlated across auditory channels. Dai and Micheyl (2010) describe how nearby stimulus components can excite partially overlapping portions of the neural population at the output of a particular auditory channel; therefore channels that respond to one component can also respond to another, with the same internal noise perturbation being represented in both. This would make it easier to detect spectral ripples that occur over a wide frequency range than over a narrow frequency range. However, better detection of ripples that extend over a wide frequency range could occur even without effects of correlated internal noise, simply on the grounds that more information generally leads to better performance.

In summary, Micheyl *et al.* (2010) proposed that performance in the task of Moore and Sek was influenced by the detection of a regular pattern of spectral ripples, and that the results did not demonstrate listeners' sensitivity to TFS above the 8th harmonic. They suggested instead that the low thresholds measured for signals containing harmonics 9–13 were due to residual excitation pattern cues being used by the subjects as a result of auditory filters being narrower than assumed in the model of Glasberg and Moore, and to the short-range correlation of the neural noise in auditory channels responding to the signals.

3.3 A computer model based on the detection of regular spectral ripples

3.3.1 Rationale

To assess whether the results obtained using the test of Moore and Sek (2009a) can be explained in terms of excitation-pattern cues (shifts in a regular pattern of ripples), a computational excitation-pattern model was developed, with two implementations. This model selected which interval was most likely to be the HIHI interval on the basis of the similarity between the excitation-pattern differences between the H and I tones and a "template" that included the ripples and was constructed from stimuli presented during the first few trials of the task. It has been suggested (Dau, 1996) that subjects might develop templates based on repeated presentations of the signals, and might base their decisions on the similarity of the representations of a specific trial to the templates. Each implementation of the model is described in detail below. The main motivation for developing the model was to compare the pattern of results predicted by the model with the pattern actually obtained using human listeners. This comparison is given in chapter 4, following presentation of the results obtained for human listeners.

The computer model essentially simulated the forced-choice task that was performed by the subjects in the procedure of Moore and Sek. The model was required to discriminate between an interval containing HHHH and one containing HIHI, where the components of the I tones had an upward frequency shift of ΔF (in Hertz) relative to the components of the H tones. The model completed 2002 trials (two trials to construct a template and 2000 trials for the experiment) for each of a range of fixed values of ΔF for each condition, and the number of times the correct interval was selected was counted for each value of ΔF. This number of trials gave results that were repeatable to within 5 %. Probit analysis (Finney, 1971) was then used to estimate the 70.7 %-correct point on the resulting psychometric function, so performance could be compared directly with human thresholds. The value of ΔF at the 70.7 %-correct point was taken to be the threshold for that condition. The signals used as input to the model were generated by the same software as used with the human listeners.

Thresholds were predicted for the same conditions as used with human listeners in each of four experiments. Full details of all conditions tested in each experiment are given in chapter 4.

3.3.2 Overview of the model

There are several strategies that a listener might use to select the interval containing the HIHI tones based solely on ripples in the excitation pattern. Here, a strategy was selected so as to be as close as possible to optimal.

For a given trial, the sequence of steps in the model was as follows:

1. A complete stimulus waveform file was generated. This contained the HIHI and HHHH tone patterns as well as the masking TEN, exactly as in the experiments using human listeners. The software was modified so that the HIHI stimulus always occurred in the first interval in the file. The generated stimuli all had a total level of 31 dB SPL.

2. The waveform file was split into eight sections, each containing a single tone with its onset and offset ramps. Each section was 9600 samples in length. The tones for each interval are denoted T1, T2, T3 and T4.

3. A fast Fourier transform with a size of 8192 using a Hamming window was performed on each tone to determine its spectrum.

4. The frequency and level of each spectral component were entered into the excitation-pattern model described by Glasberg and Moore (2000). This model estimated the excitation pattern for each tone. The excitation patterns were calculated at intervals of 0.1-ERB$_\mathrm{N}$ on the ERB$_\mathrm{N}$-number scale. Model A used Glasberg and Moore's estimate of the sharpness of the auditory filters, and model B used modified auditory filters that were twice as sharp as assumed by Glasberg and Moore.

5. A Gaussian-distributed random noise with a mean of zero and a standard deviation of σ was added to the excitation level at each frequency in 0.1-ERB$_\mathrm{N}$ steps. This simulated the effect of internal noise, corresponding to the inaccuracy in neural coding in the cochlea. This internal noise had the effect of perturbing the excitation pattern of each tone in each interval. The value of σ was adjusted for each model (A and B) so that the model's performance matched observed human performance for a baseline condition in each experiment. More information on this is given below.

6. Averaging the perturbed excitation patterns for tones T1 and T3 and for tones T2 and T4 gave EP1 and EP2 for interval 1 (the HIHI interval) and EP3 and EP4 for interval 2 (the HHHH interval). The difference in dB for each frequency step of the averaged excitation patterns in each interval (EP2−EP1 and EP4−EP3) was taken over the range $N-1$ to $N+$(passband width), giving the arrays EPdiff1 for interval 1 and EPdiff2 for interval 2. As interval 1 contained the I tones, EPdiff1 usually showed an almost regular pattern of ripples, whereas EPdiff2 usually did not.

7. Steps 1 to 6 were carried out for the first two trials, for which the value of ΔF was always 0.5F0 and it would have been clear to human listeners (in most conditions) which interval contained the I tones. The mean was then taken for each frequency point of the resulting two EPdiff1 arrays. This gave the template, EPT. It is plausible that human listeners may construct

such a template when performing this task, and may compare each observed set of tones with it in order to make a decision.

8. Steps 1 to 6 were repeated for each remaining trial (2000 in total). For each trial, EPdiff1 and EPdiff2 were each cross-correlated with EPT as a function of ΔE, the shift in ERB_N-number, giving xcorr1 and xcorr2 for intervals 1 and 2, respectively.

9. The cross-correlation function giving the largest peak at a non-zero value of ΔE was selected by the model as corresponding to the interval containing the HIHI tone. Hence, if xcorr1 was selected, the answer was deemed correct, as the first interval always contained the I tones.

Psychometric functions were measured using fixed values of ΔF of 0.1F0, 0.2F0, 0.3F0, 0.4F0, and 0.5F0. For some conditions, additional smaller values of ΔF were used. These smaller values were 0.05F0, 0.025F0 and so on, halving the value each time until there were two measured points on the psychometric function giving less than 70.7% correct.

3.3.3 Selection of the internal noise variable σ

The value of σ was adjusted such that the threshold predicted by each model matched that obtained for human listeners for one baseline condition in each experiment. The baseline condition for experiments 1 and 2 was where F0 was 200 Hz, N was 9 and the passband width was 1F0. This condition was selected as the baseline for two reasons: firstly, human listeners gave the best performance for this condition; and secondly, it is an intermediate F0 for which the 9^{th} component would be just beyond the limit of peripheral resolution and TFS processing ability might plausibly be good. Experiments 3 and 4 used different human listeners, who on the whole gave slightly higher thresholds than those who took part in experiments 1 and 2. The baseline condition for these two experiments was where F0 was 200 Hz, N was 13 and the passband width was 1F0. This condition was chosen as it was the condition common to both experiment 3 and experiment 4 for which human listeners gave the best performance.

To determine the appropriate value of σ, the model completed 5002 trials (the first two were to construct the template) for a series of values of ΔF for each of a range of values of σ. The value of σ was varied in 0.25-dB steps. Predicted thresholds estimated using probit analysis were compared with obtained human thresholds. The mean human threshold for the baseline condition for experiments 1 and 2 was 0.0375F0, and for experiments 3 and 4 it was 0.1782F0. Model A (original filter sharpness) could not reach human performance in either set of experiments even when σ was 0.00 dB. Model B required $\sigma = 1.00$ dB for experiments 1 and 2, and $\sigma = 2.00$ dB for experiments 3 and 4. The predicted threshold for the selected value of σ was within 3% of the mean human threshold for each baseline condition in all cases.

The next chapter presents the results predicted by these models and obtained from the human listeners.

Chapter 4

Evaluation of the relative importance of temporal-fine-structure and excitation-pattern cues in the discrimination of complex tones with high harmonics

4.1 Introduction

4.1.1 The resolvability and TFS hypotheses

Discrimination of the fundamental frequency (F0) of complex tones is usually good when the tones contain low harmonics, but worsens when the number of the lowest harmonic, N, increases above about 7, reaching a plateau when N is about 14 (Hoekstra and Ritsma, 1977; Houtsma and Smurzynski, 1990; Bernstein and Oxenham, 2003; Moore *et al.*, 2006). The worsening as N is increased from 7 to about 14 has been interpreted by some as resulting from a progressive reduction of the ability to resolve the components in the complex tone (Houtsma and Smurzynski, 1990; Shackleton and Carlyon, 1994). This is referred to here as the "resolvability" hypothesis. However, others have interpreted the worsening as resulting from a progressive loss of the ability to use TFS information (Moore *et al.*, 2006; Hopkins and Moore, 2007; Ives and Patterson, 2008; Moore *et al.*, 2009b; Moore and Sek, 2009a). This is referred to here as the "TFS hypothesis". The decision as to which hypothesis is more nearly correct depends on the extent to which harmonics with numbers in the range 7 to 14 are resolved, which is still a matter of debate (Plomp, 1964; Moore, 1993; Bernstein and Oxenham, 2003; Moore *et al.*, 2006, 2009b). An overview of this debate is given in chapter 2.

4.1.2 The task of Moore and Sek

The experiments described here were aimed at deciding whether complex tones whose audible harmonics all lie above the seventh are discriminated based on their TFS or upon cues related to changes in the excitation pattern resulting from (partial) resolution of the harmonics. The experiments were based on a method of measuring sensitivity to TFS described by Hopkins and Moore (2007) and later modified by Moore and Sek (2009a). Subjects were required to discriminate a harmonic complex tone (H) with a given F0 from a complex tone in which all components were shifted upwards by the same amount in Hertz, ΔF, resulting in an inharmonic tone (I). The two tones had the same envelope repetition rate (equal to F0), but had different TFS. Subjects were asked to identify which of two intervals contained tones in the format "HIHI", the other interval containing tones in the format "HHHH". It has been shown previously that frequency-shifted tones usually have a slightly different pitch from their harmonic counterparts (Schouten, 1940a; de Boer, 1956a; Schouten *et al.*, 1962; Moore, 2003b). Hence, it is assumed that subjects performed the task by deciding in which interval the pitch of the tones went up and down; this is consistent with subjective reports of the subjects. To reduce cues related to differences in the excitation patterns of the H and I tones, all tones were passed through a bandpass filter with a flat region with a width equal to an integer multiple of the F0 (which was varied), and skirts that decreased in level at a rate of 30 dB/octave. Note that in the studies of Hopkins and Moore (2007) and Moore and Sek (2009a) N was used to refer to the harmonic number corresponding to the center of the bandpass filter through which the stimuli were passed. Here, N is used to refer to the number of the lowest harmonic within the passband. To mask combination tones, and to prevent components well away from the centre frequency being audible, the stimuli were presented in a threshold-equalizing noise (TEN, Moore *et al.*, 2000).

When the lowest component in the passband, N, is the 9^{th}, performance of normal-hearing subjects in this task is usually very good for F0s in the range 100 to 400 Hz (Hopkins and Moore, 2007; Moore *et al.*, 2009b; Moore and Sek, 2009a). Performance worsens as N is increased, and the task usually cannot be performed reliably when $N = 16$ (Hopkins and Moore, 2007; Moore and Sek, 2009a). The question addressed by the experiments here was: does the worsening in performance with increasing N reflect a progressive reduction in the ability to use excitation-pattern cues, or a progressive loss of the ability to use TFS information?

4.1.3 Prediction of performance based on excitation-pattern cues

The experiments described here involved several manipulations of the stimuli that would be expected to influence the ability to use excitation-pattern cues. To predict the pattern of changes to be expected from these manipulations if changes in the excitation pattern were the sole cue used to perform the task, a computational excitation-pattern model was developed, the exact implementation of which was described in the previous chapter. Model A used Glasberg and Moore's (1990) estimate of the sharpness of auditory filters, and model B used modified auditory filters that were twice as sharp as assumed by Glasberg and Moore.

Each excitation-pattern model selected which interval was most likely to be the HIHI interval on the basis of detecting a regular pattern of differences between the excitation patterns of the tones within each interval. Both model A and model B included a source of internal noise, which was varied until performance for the model matched human performance for two baseline conditions, one for each different group of subjects. A solution was implemented that was believed to be a robust method of detecting regular differences between successive excitation patterns on the basilar membrane. It is not claimed that this is the best method for doing this: this idea is expanded upon in the general discussion section. However, this method was sensitive to regular changes in excitation patterns, and could plausibly be similar to a strategy employed by human listeners for performing the task. The main motivation was to compare the pattern of results predicted by the model with the pattern actually obtained using human listeners.

4.1.4 Temporal envelope cues

The experiments described here also addressed the issue of whether performance in discriminating the H and I tones might depend on cues in the envelopes of the signals as represented in the auditory system. In the original version of the task, as used by Hopkins and Moore (2007), most of the stimuli had components added in cosine phase. This results in a waveform with one large envelope peak per period. It is possible that the representation of the envelope in the auditory system differed slightly for the H and I tones. Hence, subjects might have discriminated the stimuli using subtle differences in temporal envelope (TE) shape. This is illustrated in Figure 4.1.

The top panel shows the waveform and Hilbert envelope of an H tone after passing through a simulated (gammatone) auditory filter with a centre frequency (CF) of 787.5 Hz. The tone had an F0 of 75 Hz, and the components were added in cosine phase. The lowest component in the passband was the 9$^{\text{th}}$, and the width of the passband was 5F0. The middle panel shows the filtered waveform and Hilbert envelope of the corresponding I tone when ΔF was 37.5 Hz (0.5F0). The bottom panel shows the two envelopes plotted on a logarithmic scale and superimposed. There is a slight difference between the envelopes of the H and I tones in the "dips" of each period, which might be used as a cue to discriminate the tones.

Hopkins and Moore (2007) re-ran a subset of their conditions using stimuli with components added in random phase, with a different selection of random phases for every stimulus. This meant that the envelope shape varied markedly and randomly from one stimulus to the next, for both the H and I tones, making envelope shape an unreliable cue for discrimination. Hopkins and Moore found

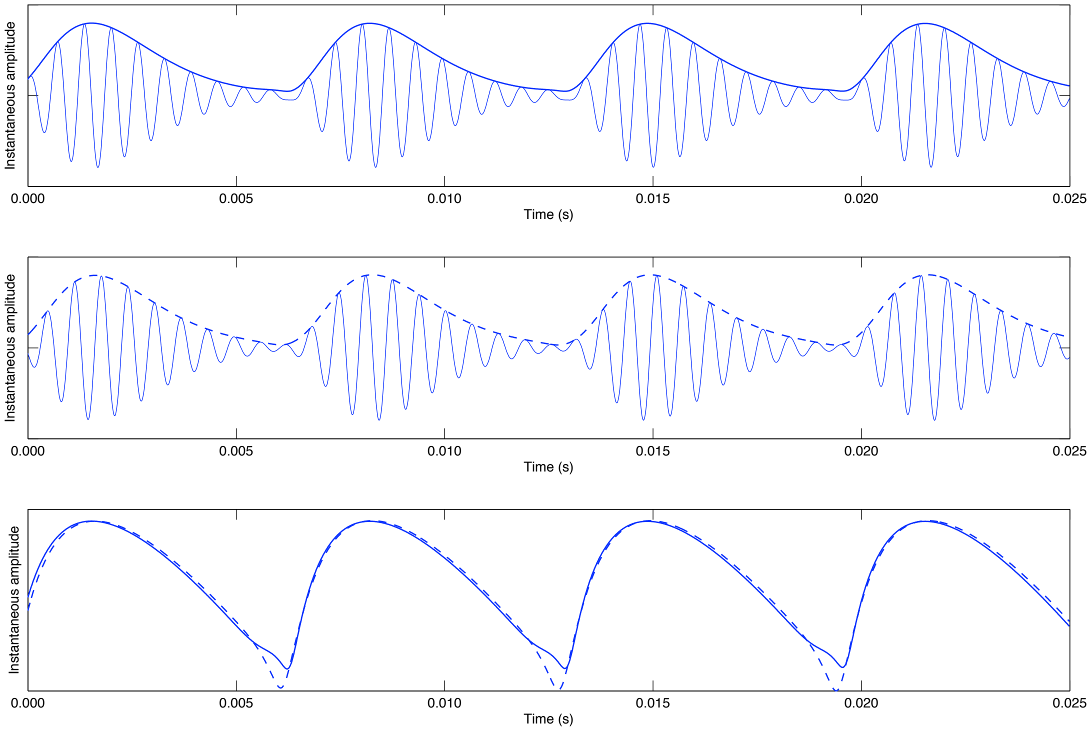


Figure 4.1: Waveforms and Hilbert envelopes of an example H tone (top panel) and I tone (middle panel) after passing through a gammatone filter. F0 was 75 Hz, N was 9, passband width was 5F0, ΔF was 0.5F0 and components were added in cosine phase.

that the task could still be performed when the components had random starting phases, a result that was confirmed by Moore and Sek (2009a). However, as far as is known, no study has directly compared performance for discrimination of the H and I tones in the same subjects using cosine- and random-phase complex tones. The predictions of models A and B would not have been affected by TE cues as the models discriminated the tones based on spectral information only. Therefore, only random phase was used for models A and B, whereas for the human listeners both random and cosine phase were used so that performance could be compared.

The changes in envelope shape between the H and I tones depend on the way the envelope is calculated. Figure 4.1 shows the Hilbert envelope of the signals, but it is possible that the envelopes would be slightly different if they were calculated another way, for example by half-wave rectification followed by lowpass filtering. Future work could establish a threshold for envelope change discrimination, which might shed light on how the envelope of a signal is represented in the auditory system.

4.1.5 Summary of conditions

Experiment 1 assessed the effect of passband width. Experiment 2 measured thresholds for a range of values of F0 and N to provide a description of TFS processing in different conditions. Experiment 3 assessed the possibility that experiment 1 had inadvertently measured sensitivity to the signal-to-TEN ratio. Experiment 4 introduced a random amplitude perturbation to each component, which was expected to disrupt any regular pattern of ripples in the excitation pattern that might be used to perform the task.

4.2 General method

Obtained human performance was compared directly with thresholds predicted by both excitation-pattern models (model A and model B).

4.2.1 Stimuli

All signals were generated using a PC and an external M-Audio Audiophile soundcard, using a 48-kHz sampling rate and 16-bit resolution. The output of the soundcard was used to drive one earpiece of a Sennheiser HD580 headset, or was used as input to a model. Hence the signals used by the models and human listeners were generated identically.

4.2.2 Human threshold estimation

The procedure used for human listeners for all experiments reported here was as described by Moore and Sek (2009a). Subjects were required to discriminate between an H tone and an I tone in which each component was shifted upwards by a given number of Hertz, ΔF. To avoid problems due to the ambiguity of the pitch of stimuli with only a few components (Schouten *et al.*, 1962), the task was designed so that subjects did not have to indicate the direction of any pitch shift between the H and I tones, but only to indicate the set of tones within which the frequency shift occurred. In a given trial, subjects were presented with two sets of four tones: either "HHHH" followed by "HIHI" or "HIHI" followed by "HHHH". Each tone lasted 200 ms, including 20-ms raised-cosine onset and offset ramps. The silent interval between the four tones within a set was 100 ms. The two sets of four tones were separated by 300 ms. The interval that contained the "HIHI" tone was varied randomly. The model or the subject indicated which of the two intervals contained "HIHI". Subjects responded by using a mouse to click one of two buttons on a computer screen, each of which flashed during the corresponding interval of the task. Subjects received visual feedback on their performance via colour flashes on the computer screen.

A background TEN was used to mask combination tones as well as components falling on the skirts of the filter. The TEN started 300 ms before the first set of four tones started, and ended 300 ms after the second set of tones had finished, again including 20-ms onset and offset ramps. The level of the TEN, specified as the level in a 1-ERB$_\mathrm{N}$-wide band centred on 1000 Hz, was set to 15 dB below the overall level of the signal. With this noise level, for $N = 9$, component $N - 1$ would probably have been above the masked threshold imposed by the TEN, component $N-2$ would have been close to masked threshold, and component $N-3$ would have been below the masked threshold. The human subjects completed an adaptive two-alternative forced-choice two-down one-up procedure (Levitt, 1971), which tracked the 70.7 %-correct point on the psychometric function.

4.2.3 Excitation-pattern models threshold estimation

The algorithm for the excitation-pattern models essentially simulated the forced-choice task that was performed by the human subjects. For each condition, each of model A and model B completed 2002 trials for each of a range of fixed values of ΔF; probit analysis (Finney, 1971) was then used to estimate the 70.7 % correct point on the resulting psychometric function.

4.3 Experiment 1: effect of varying the width of the passband

4.3.1 Rationale and conditions

In this experiment, the effect of varying the width of the passband was assessed. F0s of 75 and 200 Hz were used. The bandpass filter had a central flat region with a width of 1, 3 or 5 times F0: here "passband width" refers to the width of this flat region. The lower edge of the flat region of the passband fell at 8.5 or 12.5 times F0, so that for the H tone the number, N, of the lowest component within the flat region of the passband was 9 or 13. As the passband width was increased, the lower edge of the passband was kept fixed, and the upper edge frequency was increased. Components were added in either random or cosine phase. For the random-phase stimuli, the selection of starting phases was recalculated for every tone.

The overall level of the tones was set to be 30 dB above the absolute threshold level for a sinusoid at the frequency corresponding to N; for brevity, this will be denoted 30 dB sensation level (SL). Moore and Sek (2009a) used a level of 20 dB SL, but some subjects commented that this level was too low to be comfortable, so the level was raised for all experiments reported here. 30 dB SL is a moderate level, and would not result in the decrease in sharpness of the auditory filter at high stimulus levels described by Glasberg and Moore (1990), whilst being comfortably audible to the subjects. This is consistent with Moore and Sek's (2009a) demonstration that there was no change in estimated threshold using this method for total presentation levels between 20 and 50 dB SL.

4.3.2 Subjects

Seven subjects with absolute thresholds of 15 dB HL or below in both ears at all audiometric frequencies took part. Three were male and four were female, and their ages ranged from 22 to 31 years. All subjects had some experience playing non-keyboard musical instruments. Musically trained subjects were used to avoid long learning effects, which can occur in frequency discrimination tasks, although Moore and Sek (2009a) found no significant effect of training for this procedure. As no training effect was expected, subjects were not trained before completing the present experiment. However, after completing the last block of conditions, subjects were retested on the first condition tested to check that performance was similar to that measured initially. This was the case for all subjects.

4.3.3 Procedure for human listeners

Discrimination of the H and I tones was measured using the same two-down one-up procedure as used by Moore and Sek (2009a). The starting value of ΔF was 0.5F0, as this value results in the largest difference between the H and I tones. Before the second trial after the first turnpoint, ΔF was changed by a factor of 1.25^3; before the second trial after the second turnpoint, the factor was reduced to 1.25^2, and after this, the factor was 1.25. Eight turnpoints were obtained in a run, and the threshold was calculated as the geometric mean of the values of ΔF at the last six turnpoints.

If the value of ΔF requested by the procedure exceeded 0.5F0 more than three times during a run, the task was switched to a non-adaptive procedure, with the value of ΔF fixed at 0.5F0. Subjects completed 20 trials in this non-adaptive procedure and the percentage correct for these trials was used to estimate d′.

Each subject completed two runs for each condition. If the standard deviation of the logarithms of the values of ΔF at the last six turnpoints of a run was greater than 0.2, the results for that run were discarded and a new run was performed.

4.3.4 Absolute threshold estimation

Before measuring the threshold for discriminating the H and I tones in each condition, the absolute threshold of the subject for a sinewave with a frequency corresponding to N was measured. This was done using an adaptive two-alternative forced-choice, two-down one-up procedure that tracked the 70.7 %-correct point on the psychometric function, corresponding to $d' = 0.77$. The starting level was always 60 dB SPL. The step size was 6 dB before the first turnpoint, 4 dB before the second turnpoint, and 2 dB thereafter. The procedure terminated after six turnpoints had occurred, and the threshold was taken as the mean of the signal levels at the last four turnpoints. The overall level of the H and I tones was set to be 30 dB above the absolute threshold determined in this way.

4.3.5 Data analysis

As each subject completed two runs for each condition, where each result could be a threshold or a percent-correct score, it was necessary to convert all results into a common format for analysis. Therefore, the two results for each subject and condition were combined into a single threshold in the following way:

For a given subject and condition, if any results were a percent-correct score, they were summed and, if the result was greater than 14/20 or 25/40 (the score required for it to be considered above chance at the 0.05 level, based on the binomial distribution), it was converted to a d' value using the table presented by Hacker and Ratcliff (1979). The threshold value of ΔF that would be obtained for a d' value of 0.77 (the d' score given for 70.7 % correct, obtained by interpolating linearly between the values for 70 % and 71 %) was then extrapolated from this value, assuming that d' is proportional to ΔF. A similar method of analysis, though extrapolating d' values rather than ΔF values, was used by Hopkins and Moore (2007) and Moore and Sek (2009a). If the percent-correct score was not significantly different from chance, the threshold associated with it was set to

0.5F0, the maximum shift allowed by the procedure. Next, the square root of any threshold estimates (whether measured directly or extrapolated) was taken, the arithmetic mean of these values was calculated and the result was squared to give a single threshold result for that condition. This square-root transform was used since the data more closely fitted a normal distribution on the square-root scale.

4.3.6 Model predictions

Figure 4.2 shows the geometric mean of the two thresholds predicted by model A and model B for each condition. Note that the predictions are the same for cosine-phase and random-phase stimuli, as the model is based on the power spectra of the stimuli. The error bars show the standard error of the two estimates for each condition. The horizontal dotted line at 0.5F0 shows the maximum threshold possible in the procedure. An asterisk indicates a condition for which the psychometric function failed to cross 70.7 % correct, so the threshold was set to 0.5F0 (the maximum possible value).

Model A predicted that thresholds should increase with increasing width of the passband. Model B also predicted this, but the predicted increase was very slight. Thresholds were predicted by both models to be higher for F0 = 75 Hz than for F0 = 200 Hz, and higher for $N = 13$ than for $N = 9$. The predicted effect of F0 is mainly a consequence of the fact that the bandwidth of the auditory filter relative to the centre frequency is somewhat greater for frequencies around 750 Hz (the region of the passband for F0 = 75 Hz) than for frequencies around 2000 Hz (the region of the passband for F0 = 200 Hz). Hence, the ripples in the excitation pattern corresponding to individual components were more marked for F0 = 200 than for F0 = 75 Hz. The predicted increase in thresholds with increasing N occurs for essentially the same reason. As model B uses sharper filters, ripples in the excitation pattern were more marked than those at the same harmonic rank in the excitation patterns for model A, so model B predicted lower thresholds overall.

An internal noise variable was used to match the performance of each model to human performance in a “baseline” condition, the one for which the lowest mean

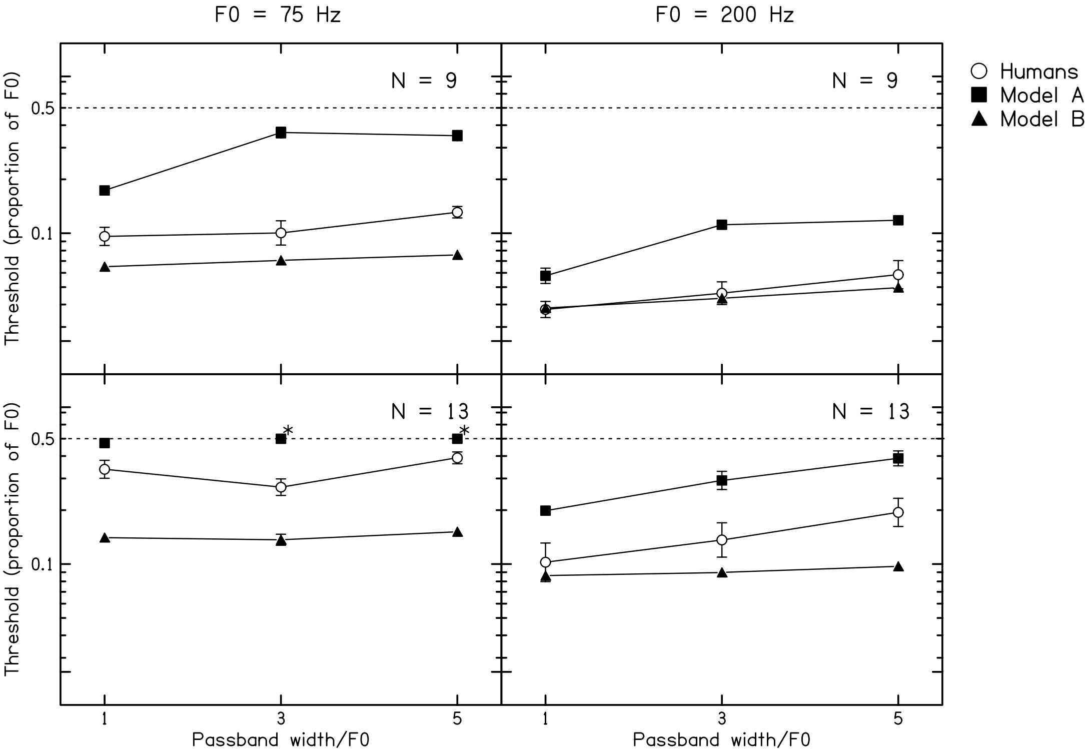


Figure 4.2: Geometric mean of predicted thresholds for model A and model B and obtained thresholds for each condition in experiment 1. Error bars for the models show the standard error of the runs, and for the humans show the standard error across subjects and phase.

threshold was obtained. For experiments 1 and 2, the baseline condition was F0 = 200 Hz, $N = 9$, and passband width = 1F0; for experiments 3 and 4, the baseline condition was F0 = 200 Hz, $N = 13$ and passband width = 1F0. The models' internal representations of the excitation pattern of each tone were perturbed by the addition of a value selected randomly from a Gaussian distribution with a mean of zero and a standard deviation of σ to each point in each excitation pattern. Details about how the level of the noise was selected for each model were given in the previous chapter. Model A predicted a higher threshold than obtained for both baseline conditions, so the internal noise variable (σ) was set to 0.00 dB for all four experiments. For model B, the level of the internal noise was 1.00 dB for experiments 1 and 2, and 2.00 dB for experiments 3 and 4.

4.3.7 Obtained data and comparison

A within-subjects analysis of variance (ANOVA) was conducted on the obtained data, using the logarithm of the single combined threshold for each subject and condition as the variate. The data were positively skewed, so a logarithmic transform was used as the residuals of the transformed data were more normally distributed than were the residuals for the untransformed data. The effects of passband width [F(2,12)=11.65; p=0.002], F0 [F(1,6)=29.08; p=0.002], and N [F(1,6)=119.55; p<0.001] were all significant. The effect of phase was not significant [F(1,6)=0.25; p=0.633]. There was a significant interaction between passband width, N and phase [F(2,12)=42; p=0.036]. The interaction between passband width and F0 just failed to reach significance [F(2,12)=3.77; p=0.054].

Figure 4.3 shows the threshold for each subject and condition. As an ANOVA conducted on the data showed no significant effect of phase, or two-way interactions with phase, the geometric mean of the random-phase and cosine-phase data is plotted for each subject, with error bars showing the standard error of the four runs (two repeats for each of two phases per subject). There were rather large individual differences in the overall level of performance, but the general pattern of the results was similar across subjects.

Performance was generally better for F0 = 200 Hz than for F0 = 75 Hz. Also, performance was better for $N = 9$ than for $N = 13$. For F0 = 200 Hz, the mean

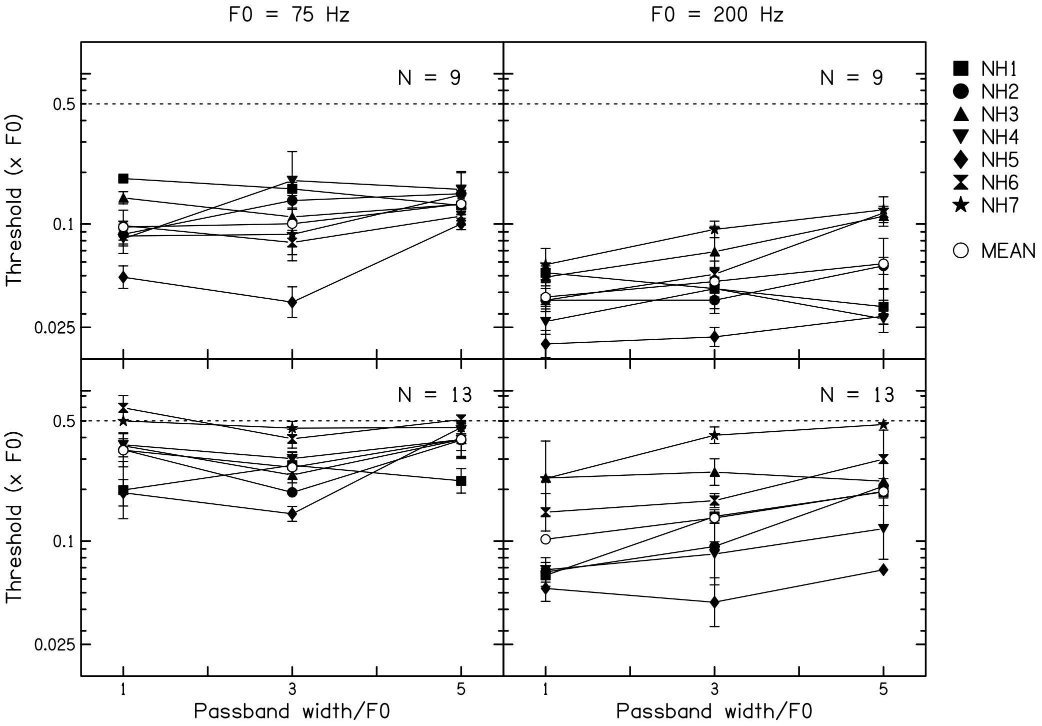


Figure 4.3: Geometric mean of four obtained thresholds for each condition and each human listener for experiment 1. Error bars show the standard error across phase.

data showed slightly worse performance when the passband width was 5F0 than when it was 3F0 or 1F0. For F0 = 75 Hz, the mean data showed no clear change in performance with bandwidth. However, the interaction between F0 and passband width just failed to reach significance, as the absolute differences involved were very small.

Figure 4.2 shows the geometric mean of the random-phase and cosine-phase thresholds for all subjects for each condition and the data predicted by each model. For the human data, the geometric mean of all thresholds for each condition is plotted, with error bars showing the standard error.

A further two ANOVAs were conducted to compare the predictions of each model with the geometric mean of the data across subjects, ANOVA 1a for the human-model A data and ANOVA 1b for the human-model B data. The variate was the single threshold estimate for each condition. For each ANOVA, the four factors were data source (human listener or model A, or human listener or model B), F0, N and passband width. The pooled variance associated with four-way interactions was used as an estimate of the residual variance.

For ANOVA 1a, there were significant effects of data source [F(1,23)=1334.08; p<0.001], F0 [F(1,23)=2030.53; p<0.001], N [F(1,23)=2708.54; p<0.001] and passband width [F(2,23)=204.66; p=0.005]. There were significant interactions between data source and N [F(1,23)=47.65; p=0.02], between F0 and N [F(1,23)=54.53; p=0.018], between data source and passband width [F(1,23)=35.62; p=0.027], between F0 and passband width [F(1,23)=25.29; p=0.038], between N and passband width [F(1,23)=19.94; p=0.048] and between data source, F0 and N [F(1,23)=62.31; p=0.016].

For ANOVA 1b, there were significant effects of data source [F(1,23)=189.84; p=0.005], F0 [F(1,23)=421.55; p=0.002], N [F(1,23)=813.42; p=0.001] and passband width [F(2,23)=25.68; p=0.037]. There were significant interactions between data source and F0 [F(1,23)=35.87; p=0.027] and between data source and N [F(1,23)=35.52; p=0.027].

The effects of F0 and N were similar to those predicted by both excitation-pattern models: thresholds were higher for F0 = 75 Hz than for F0 = 200 Hz, and higher for $N = 13$ than for $N = 9$. However, the increase in threshold with a decrease in F0 was less than predicted by model A, but more than predicted by model B; similarly, the increase in threshold with increasing N was less than predicted by model A, and more than predicted by model B. The increase in threshold with increasing passband width was less than predicted by model A, but not significantly different from that predicted by model B [F(2,23)=5.45; p=0.155].

4.3.8 Discussion

The data in general showed effects of F0, N and passband width that were similar to the effects predicted by the excitation-pattern models, though to different extents, as demonstrated by the interactions with data source in the ANOVA.

4.3.8.1 Effect of F0 and N

The differences in the excitation-pattern models' predictions for different values of F0 and N can be explained by looking at examples of the excitation patterns associated with each condition. Figure 4.4 shows the excitation patterns of example H and I tones used in experiment 1 for a shift of 0.5F0 when there were five components in the passband. The frequency region plotted ranges from $N - 1$ to N+(passband width), which is the range analysed by the models. These excitation patterns were generated using the model of Glasberg and Moore (1990). In all panels, the solid line shows the average excitation pattern of the first and third tones in each interval (harmonic), and the dashed line shows the average excitation pattern of the second and fourth tones in each interval (inharmonic). It can be seen from this figure that the differences between the excitation patterns of the H and I tones are much greater when $N = 9$ than when $N = 13$, for both F0s. This difference occurs because auditory filter bandwidths increase with increasing centre frequency. As a result, the resolution of individual harmonics decreases at higher harmonic ranks. Therefore, the ripples in the excitation pattern evoked by individual components become shallower as the harmonic rank of the components increases. If each ripple is shallower, then the overall difference (measured here by the cross correlation) between the H tones and the I tones in each interval is smaller.

It can also be seen from Figure 4.4 that the differences between the excitation patterns of the H and I tones are greater for an F0 = 200 Hz than for an F0 = 75 Hz, for both values of N. This can also be explained by the change in auditory filter bandwidth at different centre frequencies (CFs). At lower F0s, the widths of auditory filters are larger as a proportion of their CF, so a harmonic of a given rank is less well resolved than a harmonic with the same rank for a higher F0. This excitation-pattern-based explanation can also apply to the human data.

Model A predicted performance that was worse than shown by the human listeners for all conditions, despite the lack of any internal noise. Model B gave a better overall fit to the data. For F0 = 200 Hz and $N = 9$, model B gave a very good fit to the data, which is not surprising, since the internal noise used in the model was adjusted to give a good fit to the data for that condition with a passband width equal to F0. However, using that same value of internal noise, model B predicted performance better than shown by human listeners for F0 = 200 Hz and $N = 13$. For F0 = 75 Hz, model B predicted performance better than shown by human listeners for all conditions. It would have been possible to improve the predictions for F0 = 75 Hz by increasing the internal noise in the model, but this would have led to predicted performance that was worse than obtained for F0 = 200 Hz. In summary, while model B predicted some aspects of the data, the pattern of results across conditions was not predicted accurately.

If the TFS hypothesis holds, then subjects perform the task by analysing the time intervals between peaks in the TFS close to neighboring envelope maxima (Schouten *et al.*, 1962; Moore and Sek, 2009a); such intervals are referred to as $\mathrm{I}_{\mathrm{nem}}$.

Consider first the effect of N. As N increases, adjacent peaks in the TFS of the waveform become closer together in time. This means that the time intervals to be discriminated also become more similar. For example, consider discrimination of H and I tones with an F0 of 100 Hz, corresponding to an envelope period of 10 ms. Assume that the frequency shift, ΔF, is 50 Hz. For $N = 9$, the largest shift in the TFS would occur for an auditory filter centered close to 900 Hz. For the H tone, the most prominent values of $\mathrm{I}_{\mathrm{nem}}$ would be 8.89, 10 and 11.1 ms. For the I tone, the corresponding intervals would be 9.44, 10.56 and 11.67 ms. The auditory system therefore has to discriminate intervals such as 10 and 10.56 ms, which differ by 5.6 %. If N is increased to, say, 16, then the largest shift in the

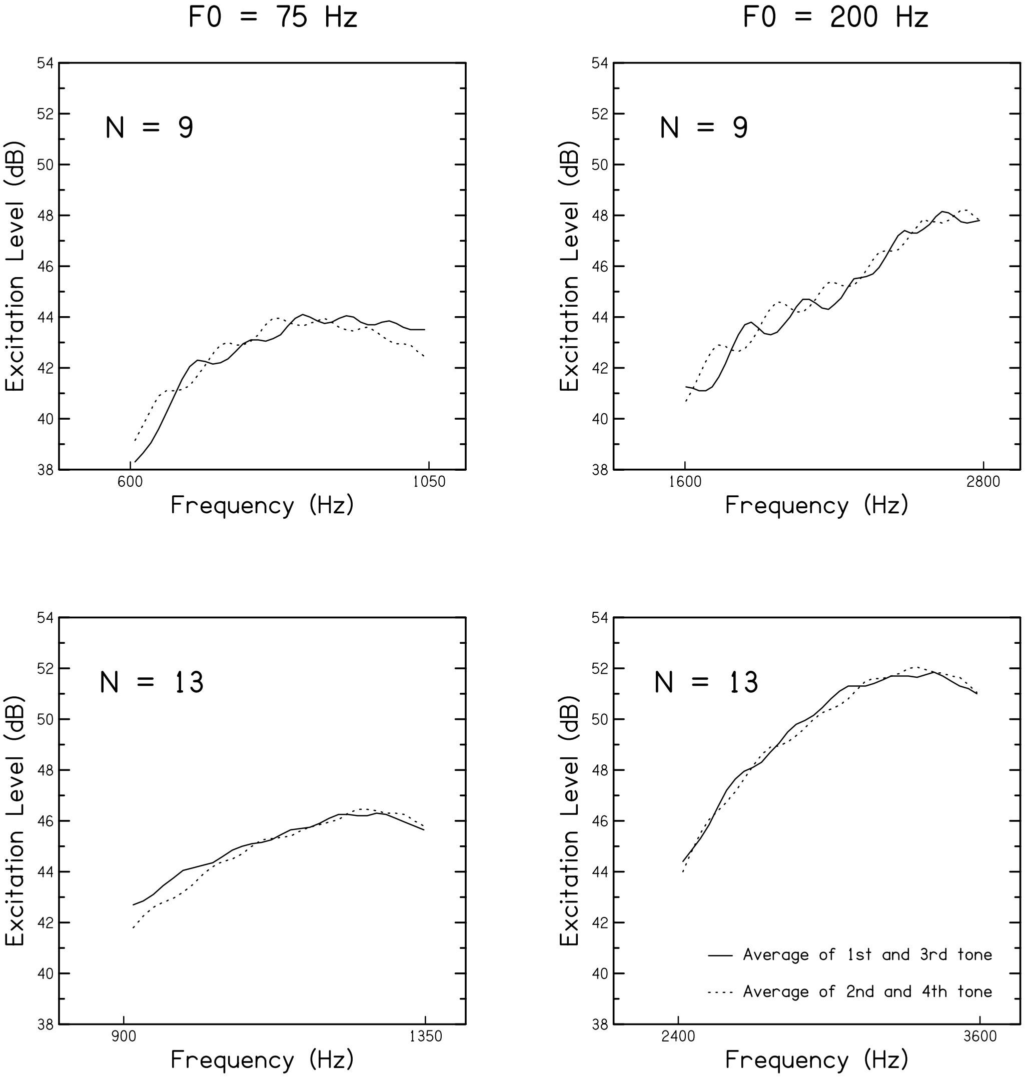


Figure 4.4: Excitation patterns of example H (solid line) and I (dotted line) tones used in experiment 1 for a shift of 0.5F0 when there were five components in the passband. F0 and N are given for each panel. Each line shows the average of the excitation patterns of two tones.

TFS would occur for an auditory filter centered close to 1600 Hz. The prominent values of $\mathrm{I}_{\mathrm{nem}}$ for the H tone become 9.38, 10 and 10.63 ms, while those for the I tone become 9.69, 10.31 and 10.94 ms. In this case, the intervals to be discriminated differ by only about 3.1 %. This could account for the worsening in human performance with increasing N. Patterson and Wightman (1976) demonstrated this effect, showing that the function relating pitch shift to frequency difference was approximately linear, and its slope decreased with an increase in N; that is, the same physical shift in frequency resulted in a smaller perceived shift in pitch when N was high than when it was low.

Consider now the effect of F0. As F0 decreases, the values of $\mathrm{I}_{\mathrm{nem}}$ become progressively longer. The auditory system may find it difficult to measure long time intervals with high accuracy (de Cheveigné and Pressnitzer, 2006), and may have some upper limit for the length of time intervals that can be measured (Krumbholz *et al.*, 2000); this could account for the higher thresholds for the lower F0. This kind of explanation is consistent with work showing that the ability to process TFS information in complex tones worsens for very low F0s (Moore *et al.*, 2009b). It is also consistent with the increase in pure-tone frequency difference limens for very low frequencies (Moore, 1973).

4.3.8.2 Effect of passband width

The human listeners and both models showed the same effect of passband width: performance worsened with increasing passband width. Although this trend was predicted by model A, the fit of the predictions to the data was not good; model A predicted that thresholds would increase more with increasing passband width than was found in the data. Model B also predicted an increase in threshold with increasing passband width, and gave a better fit to the data in most cases. However, model B predicted almost no effect of passband width for F0 = 200 Hz and $N = 13$, whereas the human data and showed a distinct increase in threshold with increasing passband width.

Although excitation-pattern differences were minimised by passing the signals through the fixed bandpass filter and by the use of the TEN, the excitation-pattern differences used by the model could have been available to the human listeners too. Could an explanation involving the use of a single ripple or a pattern of ripples in the excitation pattern of the tones in each interval underlie the effect of passband width on human performance? It is unlikely that for an F0 of 75 Hz the 8th component would be resolved, though it may have been partially resolved for an F0 of 200 Hz (assuming that the 8th component is the absolute limit of peripheral resolution for this F0). Therefore, if the humans were "optimal" listeners and were using excitation-pattern cues to perform the task, then a significant interaction between passband width and F0 would be expected, reflecting the better resolvability of harmonics with intermediate ranks on the basilar membrane for medium F0s such as 200 Hz than for low F0s such as 75 Hz. In fact, this effect just failed to reach significance [$F(2,12)=3.77$; $p=0.054$], so it is possible that the subjects in this experiment were sensitive to the partially resolved 8 th harmonic for F0 = 200 Hz, and could use information from the ripple it evoked in the excitation pattern to perform the task. However, for the excitation patterns shown in Figure 4.4, the largest single difference in excitation level between the H and I tones was for F0 = 200 Hz and N = 9, and this difference was about 1.35 dB at a frequency of about 1902 Hz, being related to the 9th component of the I tone. This difference is smaller than the purported smallest detectable change in excitation level at any single point in the excitation pattern, which is 2–3 dB (Moore *et al.*, 1989; Moore and Sek, 1994; Buus and Florentine, 1995). Although it is unlikely that a single ripple could be used to discriminate the tones in this task, it is still possible. It is also possible that the human listeners were using the regular pattern of ripples in each interval. These possibilities are assessed in experiment 4.

The TFS hypothesis would explain the effect of passband width in terms of the integration of timing information across auditory filters with different CFs. For the I tone, the TFS information is different (conflicting) at the outputs of different auditory filters, for example filters centred at the lower and upper edges of the passband. The TFS information may be easier to analyse when the conflict is reduced by decreasing the bandwidth of the stimulus. This is illustrated in Figure 4.5, which shows 10-ms samples of responses of simulated (gammatone) auditory filters (Patterson *et al.*, 1995) to an I tone with a nominal F0 of 200 Hz when ΔF was 100 Hz. The passband of the stimulus extended from 1700 to 2700 Hz (so that the 9th–13th harmonics of the H tone fell within it) and components were added in cosine phase. The top panel shows the output of a simulated auditory filter with a CF of 1900 Hz in response to this signal, and the bottom panel shows the output of a simulated auditory filter with a CF of 2500 Hz in response to the same signal. The intervals between prominent peaks in the TFS are different for the two filters. This creates a conflict of information across CFs. The wider the bandwidth of the stimulus, the more auditory filters will respond, and hence the more conflicting sets of information the auditory system will have to process. The auditory system might be able to use the output of the single auditory filter for which the TFS differs most for the H and I tones, which would be a filter with CF close to the lower edge of the passband of the stimulus, and to suppress or ignore the conflicting information from other auditory filters. However, assuming that performance in the task depends on judgments about pitch, this seems unlikely, because of the across-channel effect of pitch discrimination interference (Gockel, Carlyon, and Plack, 2004).

Although the difference in the effect of bandwidth across the two F0s was not quite significant, it still merits discussion. This effect may be related to the efficiency with which TFS information can be used in different frequency regions. Data on the frequency discrimination of pure tones (Moore, 1973; Nelson *et al.*, 1983) suggest that TFS information is most effective for mid-range frequencies, from 1000 to 2000 Hz, with reduced efficacy outside this range (Goldstein and Srulovicz, 1977; Heinz *et al.*, 2001). Although the exact upper limit of frequencies to which neurones can phase lock accurately is unknown, it is very likely that the ability to phase lock decreases above a certain frequency, with higher frequencies becoming more problematic, rather than this limit being a “hard” cut off.

Consider the case where F0 was 200 Hz and N was 13. The lower edge of the flat region of the passband fell at 2500 Hz, and the upper edge fell at 2700, 3100 or 3500 Hz for passband widths of 1F0, 3F0 and 5F0, respectively. For the greatest passband width, the most effective information would have come from an auditory filter centred close to the lower edge of the passband, both because the difference in TFS between the H and I tones would have been greatest for that filter and because that filter was centred closest to the frequency region where TFS information is processed efficiently. Decreasing the passband width

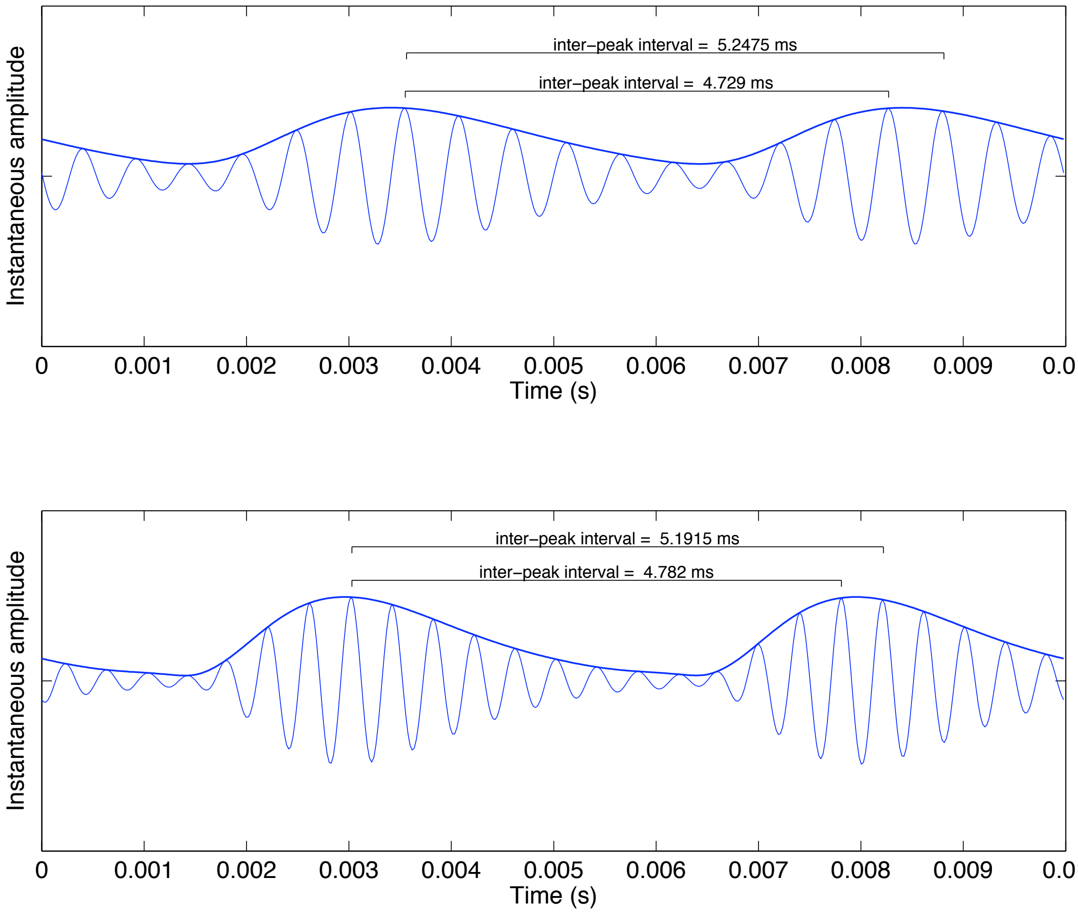


Figure 4.5: Responses of gammatone filters to a 10-ms sample of an I tone where F0 was 200 Hz, N was 9, the passband width was 5F0, ΔF was 100 Hz and components were added in cosine phase. Top panel: CF = 1900 Hz. Bottom panel: CF = 2500 Hz.

removed information that was less effective, reducing the number of channels that contained conflicting TFS information and so improving performance. Consider now the case where F0 was 75 Hz and N was 13. In this case, the lower edge of the flat region of the passband fell at 937.5 Hz, and the upper edge fell at 1012.5, 1162.5 or 1312.5 Hz for passband widths of 1F0, 3F0 and 5F0, respectively. For the largest passband width, the greatest difference in time intervals of the TFS for the H and I tones would have occurred at the output of an auditory filtered centred towards the lower edge of the passband, but the TFS information at the outputs of filter centered higher in the passband may also have been important, because these filters were centred closer to the frequency region where TFS information is used most effectively. In this case, reducing the passband width may have led to a loss of some useful TFS information, offsetting the potential beneficial effect of reducing the conflict of TFS information across channels and resulting in a smaller increase in threshold with passband width for F0 = 75 Hz than for F0 = 200 Hz.

As noted previously, the level/ERB$_\mathrm{N}$ of the TEN was set 15 dB below the total level of the complex tone. The total level of the complex tones was fixed. Hence, as the bandwidth was increased, the level of each component in the tones decreased. This meant that the signal-to-TEN ratio of individual components decreased slightly as the bandwidth was increased. It is possible that both the models and the human listeners were sensitive to this change in signal-to-noise ratio, and that this sensitivity underlay the demonstrated effect of passband width. Experiment 3 assessed this possibility directly.

4.3.9 Summary

The lowest threshold for humans occurred when F0 was 200 Hz, N was 9 and the passband width was 1F0, and this was the condition to which the models' performance was matched by changing the value of the variable representing internal noise. Model A did not achieve performance equal to that of the humans even when the internal noise was 0.00 dB. Model B required 1.00 dB of internal noise to match human performance for the above condition, and the predictions fitted the data well for the other passband widths when F0 was 200 Hz and N

was 9. However, when F0 was 75 Hz, model B predicted much lower thresholds than obtained for all passband widths and both values of N, and the predictions did not fit the data well. The same is true when comparing the thresholds for each value of N for a given F0: the predictions of model B did not fit the data well. The pattern of discrepancies with decreasing F0 and increasing N suggests that humans did not have auditory filters that were twice as sharp as in Glasberg and Moore's (1990) model. If that were the case, then the data would not have shown such a marked worsening with decreasing F0 or increasing N. However, as model A used filters as originally specified in Glasberg and Moore's model, and the level of human performance could not be reached, it is unlikely that humans operate in the same manner as these excitation pattern models.

The fact that there was no significant effect of component phase suggests that the large peak per period present in the filtered waveform of cosine-phase tones was not used to perform the task. Furthermore, this suggests that subtle differences between the envelope shapes of the random- and cosine-phase stimuli demonstrated in Figure 4.1 were also not used to perform the task.

Overall, these data suggest that humans are unlikely to be using a single spectral ripple or changes in the temporal envelope shape to discriminate the H and I tones in this experiment, although excitation-pattern model B did predict some aspects of the results. Therefore, remaining explanations include the detection of a regular pattern of ripples in the excitation patterns and the use of TFS cues.

4.4 Experiment 2: effect of varying F0 and harmonic rank

4.4.1 Rationale and conditions

To assess the plausibility of the idea that differences in TFS between the H and I tones were used to perform the task in experiment 1, the variation in the discrimination of the H and I tones for different conditions was measured. Experiment 2 provides data for a range of values of F0 and N.

F0s of 50, 75, 100, 200 and 400 Hz were used. The bandpass filter had a central flat region with a width of five times F0. The number, N, of the lowest component within the flat region of the passband was 9, 13 or 16. Components were added in either random or cosine phase, and the overall level of the tones was 30 dB SL.

4.4.2 Subjects, procedure and data analysis

The same seven subjects took part in experiment 2 as took part in experiment 1. Experiment 2 was carried out after experiment 1, so no further training was given.

The same procedure was used in experiment 2 as was used in experiment 1. The data were also analysed in the same way.

4.4.3 Model predictions

Figure 4.6 shows the threshold achieved by each implementation of the model for each condition. As before, the error bars show the standard error of the two runs for each condition. The horizontal dotted line at 0.5F0 shows the maximum threshold possible in the procedure. An asterisk indicates a condition for which the psychometric function failed to cross 70.7 % correct, so the threshold was set to 0.5F0 (the maximum possible value).

Model A predicted that thresholds should increase with decreasing F0 and with increasing N; model B also predicted this, but achieved thresholds for all conditions, whereas model A failed to achieve thresholds for F0s below 200 Hz when N was 13 and below 400 Hz when N was 16. The predicted improvement with increasing F0 was greater for model B than for model A. These predicted changes occurred for the same reasons as described for experiment 1.

4.4.4 Obtained data and comparison

A within-subjects ANOVA was conducted on the obtained data, using the logarithm of the single combined threshold for each subject and condition as the variate. A logarithmic transform was used as the residuals of the transformed data were more normally distributed than were the residuals for the

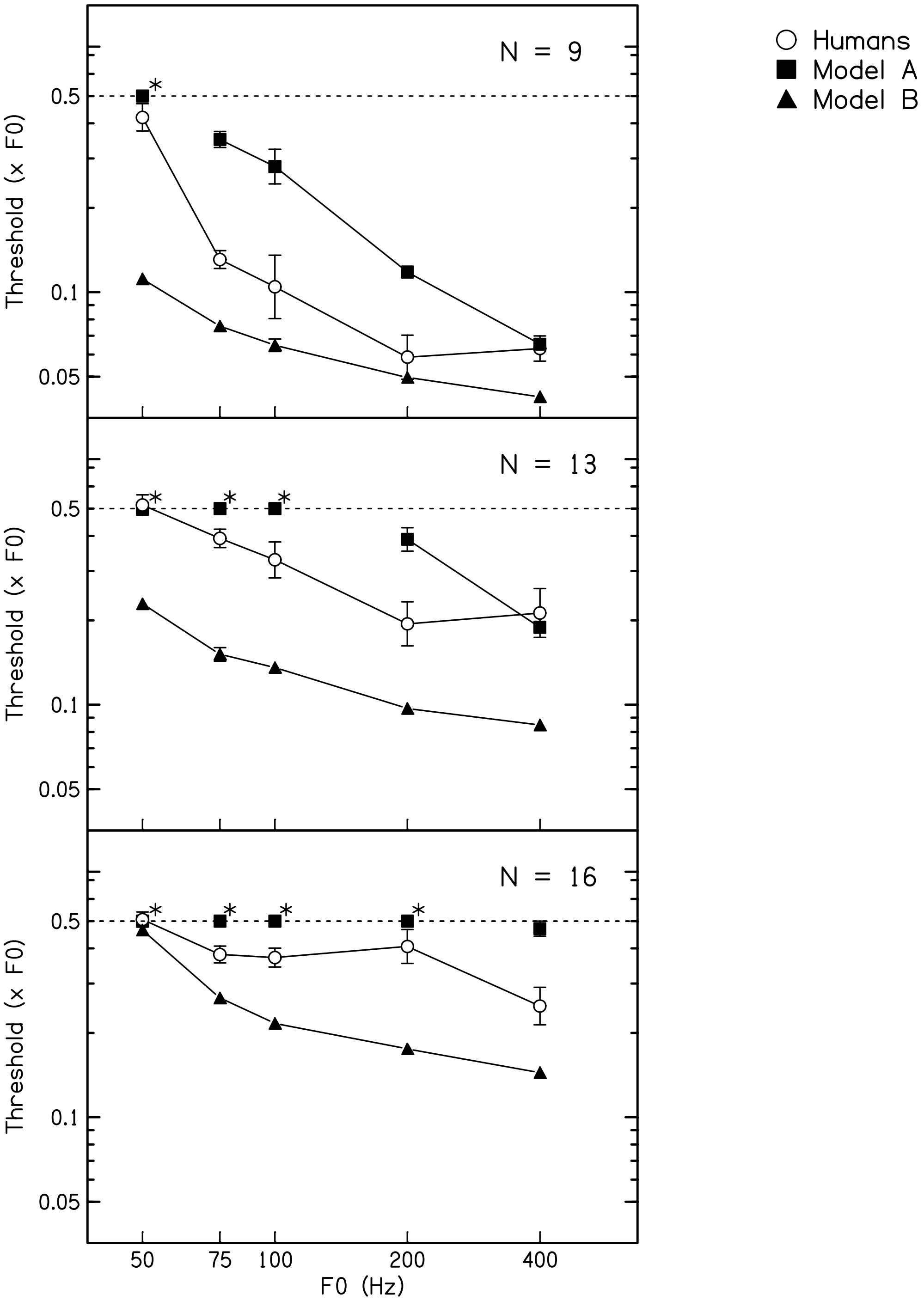


Figure 4.6: Geometric mean of predicted thresholds for model A and model B and obtained thresholds for each condition in experiment 2. Error bars for the models show the standard error of the runs, and for the humans show the standard error across subjects and phase.

untransformed data. The effects of F0 [F(4,24)=19.22; p<0.001], and N [F(2,12)=137.27; p<0.001] were both significant. The effect of phase was not significant [F(1,6)=1.48; p=0.27]. There was a significant interaction between F0 and N [F(8,48)=7.63; p<0.001].

Figure 4.7 shows the threshold for each subject and condition. As an ANOVA conducted on the data showed no significant effect of phase, or interactions with phase, the geometric mean of the random-phase and cosine-phase data is plotted for each subject, with error bars showing the standard error of the four runs. As before, there were individual differences in the overall level of performance, but the pattern of results was similar across subjects.

Performance generally improved with increasing F0 up to 200 Hz, and then flattened off. Performance worsened with increasing N, consistent with the results of experiment 1, and with earlier results (Hopkins and Moore, 2007; Moore *et al.*, 2009b). Some subjects gave thresholds close to 0.5F0 for the F0s of 50, 75 and 100 Hz for $N = 13$ and $N = 16$, but for $N = 9$ all subjects but one gave thresholds well below 0.5F0 for F0s of 75 Hz and above.

Figure 4.6 shows the geometric mean of the random-phase and cosine-phase thresholds for all subjects for each condition and data predicted by each model. For the human data, the geometric mean of all thresholds for each condition is plotted, with error bars showing the standard error.

A further two ANOVAs were conducted to compare the predictions of each model with the geometric mean of the data across subjects, ANOVA 2a for the human-model A data and ANOVA 2b for the human-model B data. The variate was the single threshold estimate for each condition. For each ANOVA, the three factors were data source (human listener or model A, or human listener or model B), F0, and N. The pooled variance associated with three-way interactions was used as an estimate of the residual variance.

For ANOVA 2a, there were significant effects of data source [F(1,29)=82.13; p<0.001], F0 [F(4,29)=11.89; p=0.002], and N [F(2,29)=16.75; p=0.001]. There was a significant interaction between data source and F0 [F(4,29)=32.74; p<0.001].

For ANOVA 2b, there were significant effects of data source [F(1,29)=8.24; p=0.021], F0 [F(4,29)=47.27; p<0.001] and N [F(2,29)=94.46; p<0.001]. There were significant interactions between data source and F0 [F(4,29)=72.06; p<0.001]

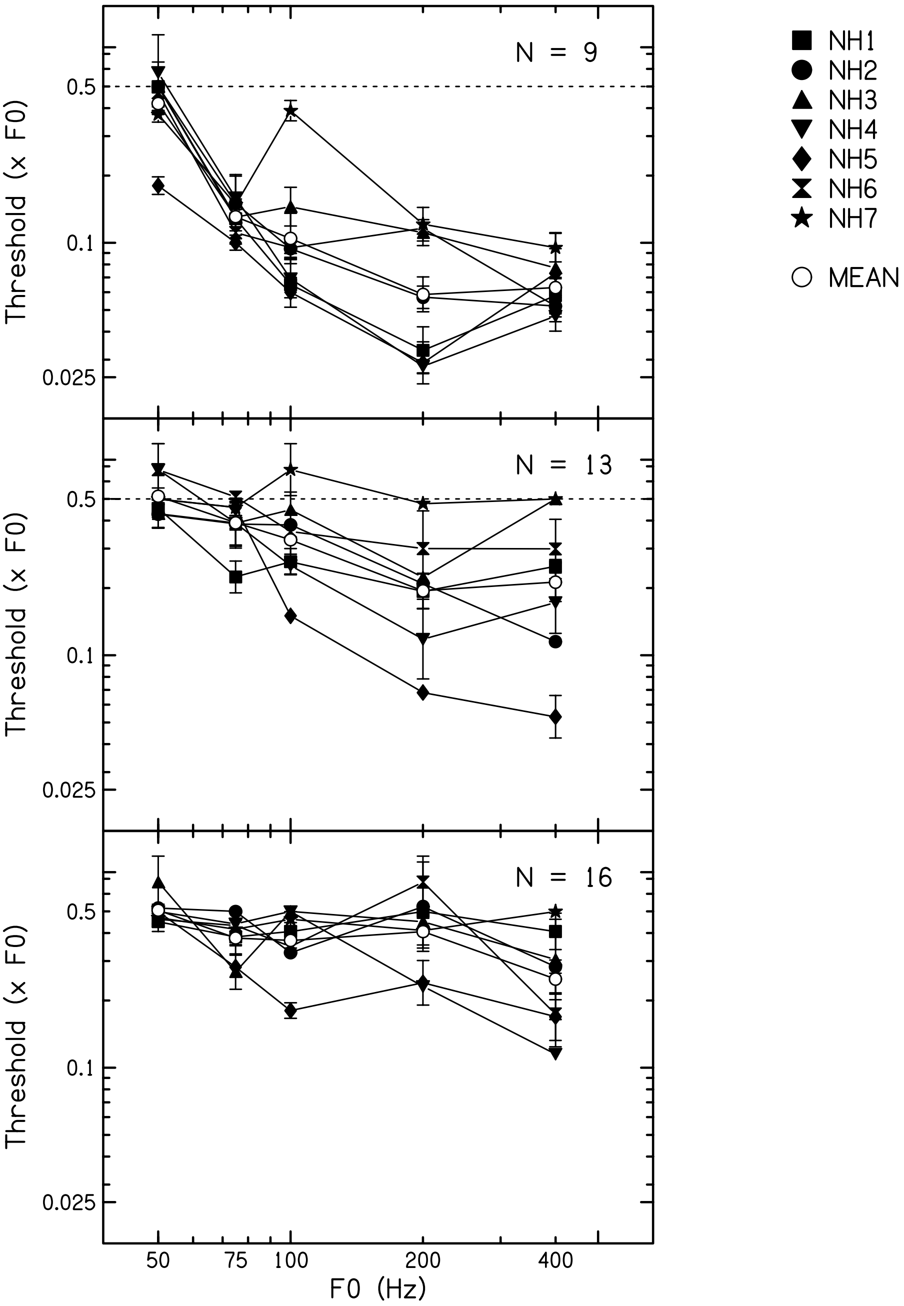


Figure 4.7: Geometric mean of four obtained thresholds for each condition and each human listener for experiment 2. Error bars show the standard error across phase.

and between data source and N [$F(2,29)=21.84$; $p<001$].

The effects of F0 and N were similar to those predicted by both excitation-pattern models: thresholds increased with decreasing F0 and with increasing N. However, model A predicted higher thresholds than obtained for all conditions, and for the F0 of 400 Hz only predicted a threshold below 0.5F0 when N was 16. For $N = 13$, Model A also predicted an improvement in performance between F0 = 200 Hz and F0 = 400 Hz, whereas human performance did not improve between these two F0s. For $N = 9$, model B predicted a smaller improvement in performance between F0 = 50 Hz and F0 = 75 Hz than was obtained, and when N was 16, model B predicted a greater improvement in performance with increasing F0 than was obtained.

4.4.5 Discussion

The data in general showed effects of F0 and N that were similar to the effects predicted by the models, although with some important differences. Whether the human listeners were using excitation-pattern cues or TFS cues to discriminate the tones in the task, the effects of F0 and N could be due to the same reasons described for experiment 1.

The data for this experiment taken alone do not shed much light on the possible use or otherwise of excitation-pattern cues by the human listeners, whether via a single ripple or a regular pattern of ripples. However, the fact that human performance for $N = 16$ did not improve with increasing F0 to the extent predicted by model B suggests that human auditory filters are not as sharp as used by this model. Similarly, model B predicted a much lower threshold than was obtained for the F0 of 50 Hz when N was 9, suggesting that the increase in filter bandwidth with decreasing CF alone did not underlie this worsening for humans if auditory filters are twice as sharp as specified by Glasberg and Moore (1990). Overall, the significant interactions between data source and F0 and data source and N in both ANOVA 2a and ANOVA 2b show that neither model predicted the same pattern of performance as the humans demonstrated in this experiment.

There was a ceiling effect in ANOVA 2a, which was imposed by model A's failure to achieve a threshold for some of the conditions; in these cases, threshold was set to 0.5F0 (the maximum possible), whereas a value of “infinity” might be more accurate, albeit less appropriate for statistical analysis. This could be the reason why no interaction between F0 and N was demonstrated in ANOVA 2a, whereas this interaction was significant for the human data taken alone.

It is notable that the human listeners could achieve good performance when F0 was 400 Hz and N was 16; in this condition, the frequency range of the flat part of the passband was 6200 Hz–8200 Hz. This is above the usual assumed limit of 5000 Hz for the use of phase-locking information (Kiang *et al.*, 1965; Palmer and Russell, 1986), although the limit in humans is not definitely known. The present results are, however, consistent with those of Moore and Sek (2009a) obtained using the same task as was employed here. They found that most listeners could perform the task for an F0 of 800 Hz when N was 12 and the lowest audible component in the H tone had a frequency of 8000 Hz. Moore and Sek argued that the results supported the idea that TFS information can be used to discriminate complex tones when all of their audible frequency components lie above 5000 Hz. The present results are consistent with this conclusion, although they do not define what the upper frequency limit for the use of TFS information may be.

Further investigation is necessary to verify that this task actually does measure sensitivity to TFS. If this turns out to be the case, then experiment 2 provides a good empirical account of how human ability to process TFS changes under different conditions.

4.5 Experiment 3: effect of varying the signal-to-noise ratio

4.5.1 Rationale and conditions

The level/ERB$_\mathrm{N}$ of the TEN in experiment 1 was 15 dB below the overall level of each stimulus. This meant that the TEN level relative to the level of each component in the tones increased with increasing passband width. The increase in threshold with increasing passband width, as observed both in the predictions and the obtained data, might have been a consequence of this. Experiment 3 assessed this possibility. The largest difference in excitation patterns for the H and I tones occurred in the frequency region around component N, the lowest component in the flat part of the passband. For a fixed overall level of the complex tone, and for the case where F0 was 200 Hz and N was 13, the level of this component was about 3 dB higher for the passband width of F0 than for the passband width of 5F0. Hence, the difference in signal-to-noise ratio for component N (2600 Hz) was 3 dB. This figure of 3 dB was confirmed using a spectrum analyser. The result was further confirmed using an iterative calculation. In this calculation, passband widths of 1F0 and 5F0 were used, and the level of each component in an H tone for F0 $=$ 200 Hz and $N = 13$ was calculated, using the same bandpass filter characteristics as used in the experiment, to achieve a total level of 32 dB SPL. For the passband width of 1F0, the level of component N was about 3 dB higher than for the passband width of 5F0.

To assess the effect that a 3-dB change in signal-to-noise ratio would have on performance, the condition with F0 $=$ 200 Hz, $N = 13$ and the passband width of 1F0 was repeated with the level of the TEN set 12 dB lower than the overall level of the stimulus rather than 15 dB lower as in experiment 1. With the higher TEN level, the signal-to-noise ratio for each component was the same as for the passband width of 5F0 in experiment 1. The condition with the original relative TEN level was repeated, as most subjects had not taken part in experiment 1. Both random-phase and cosine-phase conditions were tested, and the overall level of the tones was 30 dB SL.

4.5.2 Subjects

Seven subjects with absolute thresholds of 20 dB HL or below in both ears at all audiometric frequencies took part. Four subjects were male and three were female, and their ages ranged between 23 and 30 years. Two of the female subjects also took part in experiments 1 and 2, so they received no further training. The five naïve subjects were trained using four conditions from experiment 1, which took about half an hour to complete. Although training effects were not expected, this training was included because feedback from the subjects in experiments 1 and 2 suggested that some time to become familiar with the task would be helpful.

4.5.3 Procedure and data analysis

The procedure was the same as for experiment 1, with the following exceptions. The absolute threshold at the frequency corresponding to N was estimated twice instead of once, and the mean of the two was used to set the SL of the stimuli to 30 dB. The step size was 1.25^3 until one turnpoint occurred, was reduced to1.25^2 until the second turnpoint occurred and thereafter was 1.25. If the value of ΔF requested by the procedure exceeded 0.5F0 more than twice before the second turnpoint or at all after this, the task switched to the non-adaptive procedure. Forty trials were completed for the non-adaptive procedure rather than 20. These changes were made to reduce the number of times that a subject could complete a run and achieve a threshold measurement by chance when the tones could not actually be discriminated. Each subject completed three valid runs for each condition rather than two in order to reduce the effects of within-subject variability on the data.

The results for each subject and condition were combined into a single threshold result in the same way as for experiment 1. As there were 40 trials in the non-adaptive procedure in this experiment, performance that was significantly different from chance at the 5 % level was calculated to be >25/40, >47/80 or >69/120, using the binomial distribution.

4.5.4 Model predictions

Figure 4.8 shows the geometric mean of the two thresholds predicted by each implementation of the model for each condition. As before, the error bars show the standard error of three runs for each condition, and the horizontal dotted line at 0.5F0 shows the maximum threshold possible in the procedure.

Model A actually showed a slight improvement in performance at the higher noise level, but as can be seen from the error bars this difference was not significant, and can be ascribed to inherent variability in the signals due to the TEN. Model B showed essentially no change in performance with signal-to-noise ratio.

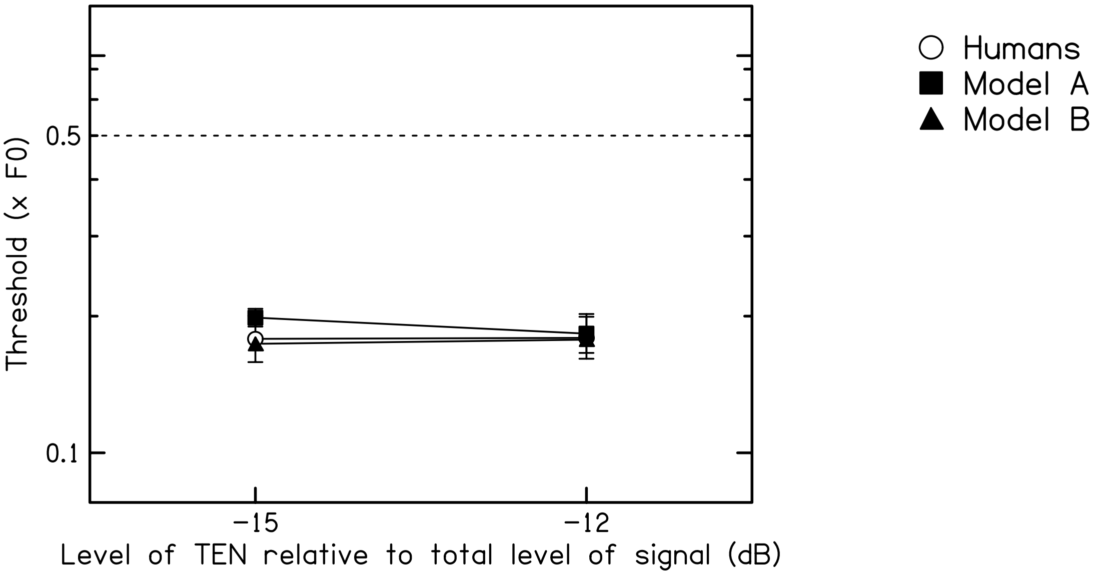


Figure 4.8: Geometric mean of predicted thresholds for model A and model B and obtained thresholds for each condition in experiment 3. Error bars for the models show the standard error of the runs, and for the humans show the standard error across subjects and phase.

4.5.5 Obtained data and comparison

A within-subjects ANOVA was conducted on the obtained data, with factors relative TEN level and phase, using the logarithm of the single combined threshold for each subject and condition as the variate. A logarithmic transform was used as the residuals of the transformed data were more normally distributed than were the residuals for the untransformed data. There were no significant effects.

Figure 4.9 shows the threshold for each subject and condition. As an ANOVA conducted on the data showed no significant effect of phase, or two-way interactions with phase, the geometric mean of the random-phase and cosine-phase data is plotted for each subject, with error bars showing the standard error of the six runs (three repeats for each of two phases per subject).

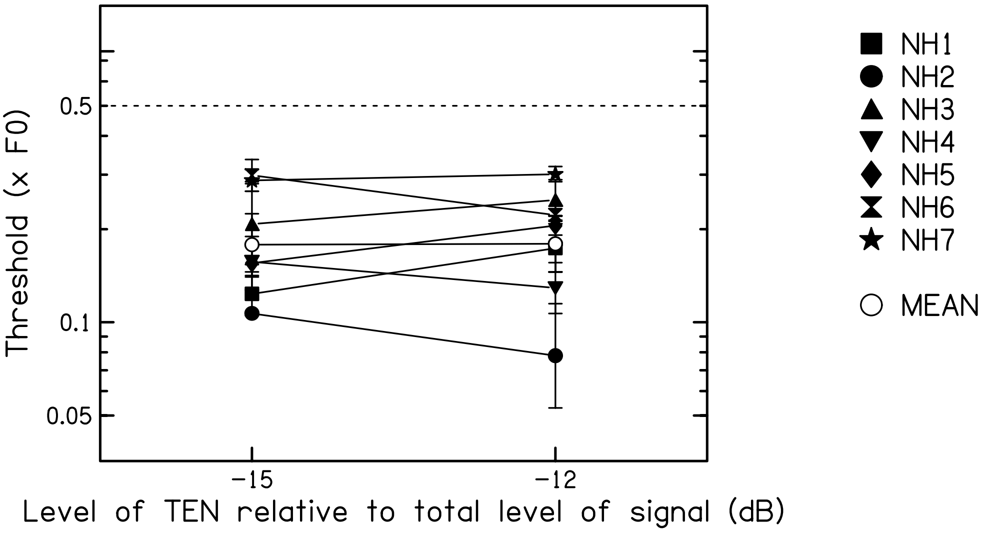


Figure 4.9: Geometric mean of six obtained thresholds for each condition and each human listener for experiment 3. Error bars show the standard error across phase.

The mean threshold was almost the same for the two relative TEN levels: 0.178F0 for the relative level of −15 dB, and 0.179F0 for the relative level of −12 dB.

Figure 4.8 shows the geometric mean of the random-phase and cosine-phase thresholds for all subjects for each condition and the data predicted by each model. For the human data, the geometric mean of all thresholds for each condition is plotted, with error bars showing the standard error.

There was no clear effect of relative TEN level either for the model or for the human listeners.

4.5.6 Discussion

The fact that there was no effect of relative TEN level shows that neither the human listeners nor the models were strongly sensitive to the signal-to-noise ratio, although an effect presumably would have been found if the ratio had been varied over a wider range. The strategy of comparing spectral ripples seems to be robust to small changes in signal-to-noise ratio, as demonstrated by the lack of effect of signal-to-noise ratio for either excitation pattern model.

In experiment 1, where F0 was 200 Hz, N was 13 and there were five components in the passband, the TEN level relative to the level of each component was -12 dB. For this case, the threshold predicted by model A was 0.3885F0, whereas in the present experiment, when there was one component in the passband, the relative TEN level was also -12 dB but the predicted threshold was 0.1832F0, which is much lower. This shows that the passband width is the important factor here rather than relative TEN level. It is not straightforward to make the same comparison for the thresholds of the human listeners or for model B, as different listeners took part experiments 1 and 3 so different values of internal noise were used for model B for these two experiments. The comparison was possible for model A because the internal noise was set to 0.00 dB for both experiments.

These data show that the improvement in performance with decreasing passband width demonstrated in experiment 1 was not due to an increase in signal-to-noise ratio. Instead, it seems that passband width *per se* affects performance. It remains to be determined whether the human listeners were using a regular pattern of spectral ripples, as were the models, or were demonstrating sensitivity to TFS.

4.6 Experiment 4: effect of perturbing component levels

4.6.1 Rationale and conditions

If the H and I tones were discriminated using excitation-pattern cues, then performance should worsen markedly if these cues are disrupted by randomly perturbing (roving) the level of each component in the complex tones. It is reasonable to assume that randomly perturbing the level of each component in each tone would disrupt both single-ripple discrimination and detection of a regular pattern of ripples. This prediction was tested in experiment 4.

The stimuli were similar to a subset of those used for experiment 1, except that a random level perturbation was applied to each component in the stimulus. The magnitude of the perturbation was chosen from a uniform distribution with range plus or minus P, where P was 0, 3 or 5 dB. The F0 was 200 Hz, the bandpass filter had a central flat region with a width of 1F0 or 5F0, and N was 13. Both random-phase and cosine-phase conditions were tested. The overall level of the tones was 30 dB SL, and the level of the TEN was −15 dB relative to the overall level of the complex.

The top row of Figure 4.10 shows an example of the averaged excitation patterns for the HIHI interval of one trial when P was 5 dB. Excitation patterns are shown as the average for two H tones (first and third tones in the interval, solid line) and for two I tones (second and fourth tones in the interval, dotted line). The bottom row shows an example of the template constructed by the model (solid line) and the difference between the excitation levels of the averaged H and I tones in the corresponding top panel (dashed line). The left-hand column shows data generated by model A, and the right-hand column shows data generated by model B. It can be seen that the excitation patterns generated by both models are much less regular than shown in previous comparable figures, due to the random level perturbation applied to each component. The ripples in the excitation pattern are deeper for model B due to the sharper auditory filters. Even though ripples are present in the excitation patterns, the ripples in the difference function (bottom row, dotted lines) between the excitation patterns for the H and I tones are much less regular than when $P = 0\,\text{dB}$ (shown in the bottom right-hand panel of Figure 4.4 for model A). This loss of regularity is reflected in a lower value for the height of the largest peak in the resulting function when the difference function is cross-correlated with the harmonic template in the model when $P = 5\,\text{dB}$ than when $P = 0\,\text{dB}$. As a result, it would be expected that both models would incorrectly select the HHHH interval more often when $P = 5\,\text{dB}$ than when $P = 0\,\text{dB}$. If human listeners are not using the regular pattern of ripples in the excitation pattern to perform the task, then human performance should not necessarily worsen greatly with increasing P; the extent of any worsening would depend on the extent to which TFS cues were disrupted by the perturbation. This is considered in more detail later.

4.6.2 Subjects, procedure and data analysis

The same seven subjects took part in experiment 4 as took part in experiment 3. Experiment 4 was carried out after experiment 3, so no further training was given.

The same procedure was used in experiment 4 as was used in experiment 3. The data were also analysed in the same way.

4.6.3 Model predictions

Figure 4.11 shows the threshold achieved by each implementation of the model for each condition. As before, the error bars show the standard error of the two runs for each condition. The horizontal dotted line at 0.5F0 shows the maximum threshold possible in the procedure. An asterisk indicates a condition for which the psychometric function failed to cross 70.7 % correct, so the threshold was set to 0.5F0 (the maximum possible value). Model A was unable to achieve a threshold when P was $3\,\text{dB}$ or $5\,\text{dB}$. Model B predicted an increase in threshold with increasing P, which, for both passband widths, was greater between $P = 3\,\text{dB}$ and $P = 5\,\text{dB}$ than between $P = 0\,\text{dB}$ and $P = 3\,\text{dB}$. Both model A and model B predicted worse performance for the larger passband width, consistent with experiment 1 and with the conclusions of experiment 3.

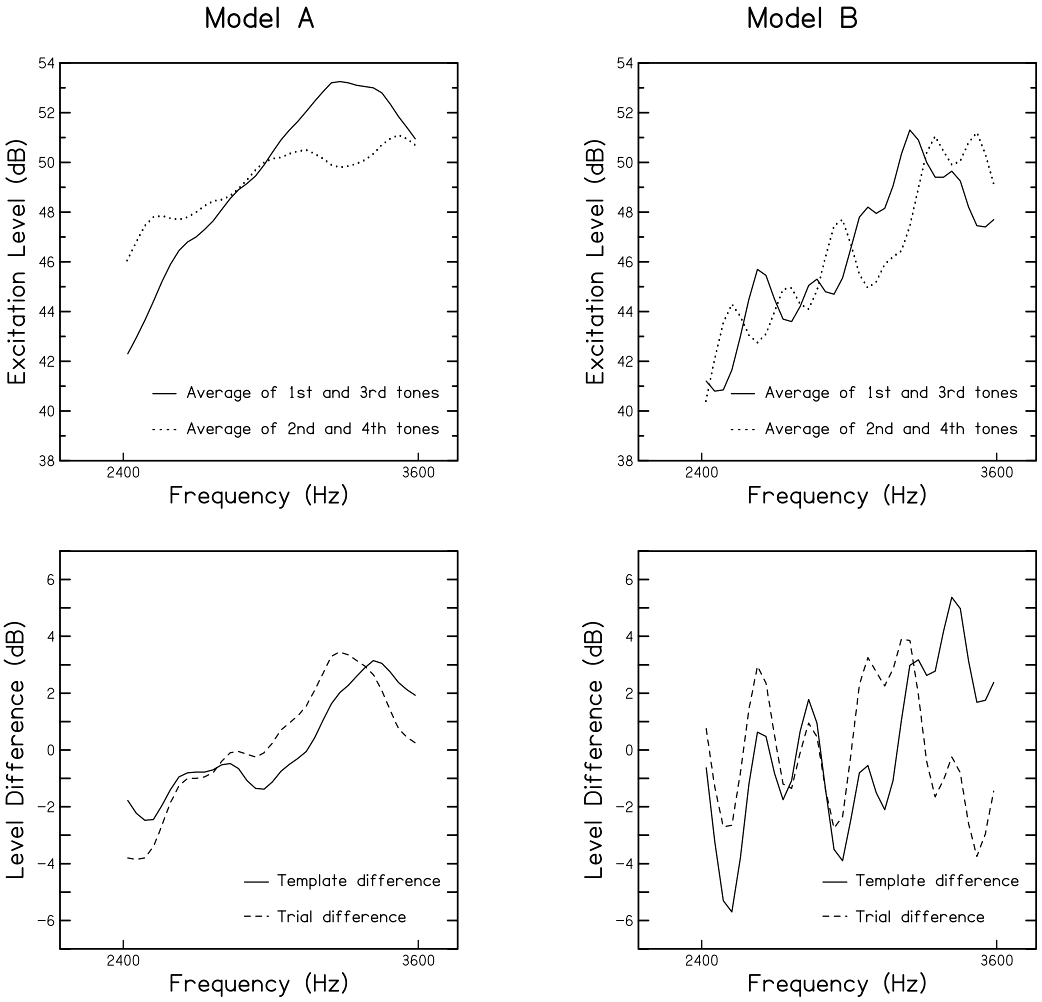


Figure 4.10: Top row: excitation patterns for the average of two H tones (first and third tones in the interval, solid line) and for two I tones (second and fourth tones in the interval, dotted line) where P was 5 dB. Bottom row: the template constructed by the model (solid line) and the difference function between the excitation levels of the averaged H and I tones in the corresponding top panel (dashed line). Left-hand column: data generated by model A. Right-hand column: data generated by model B.

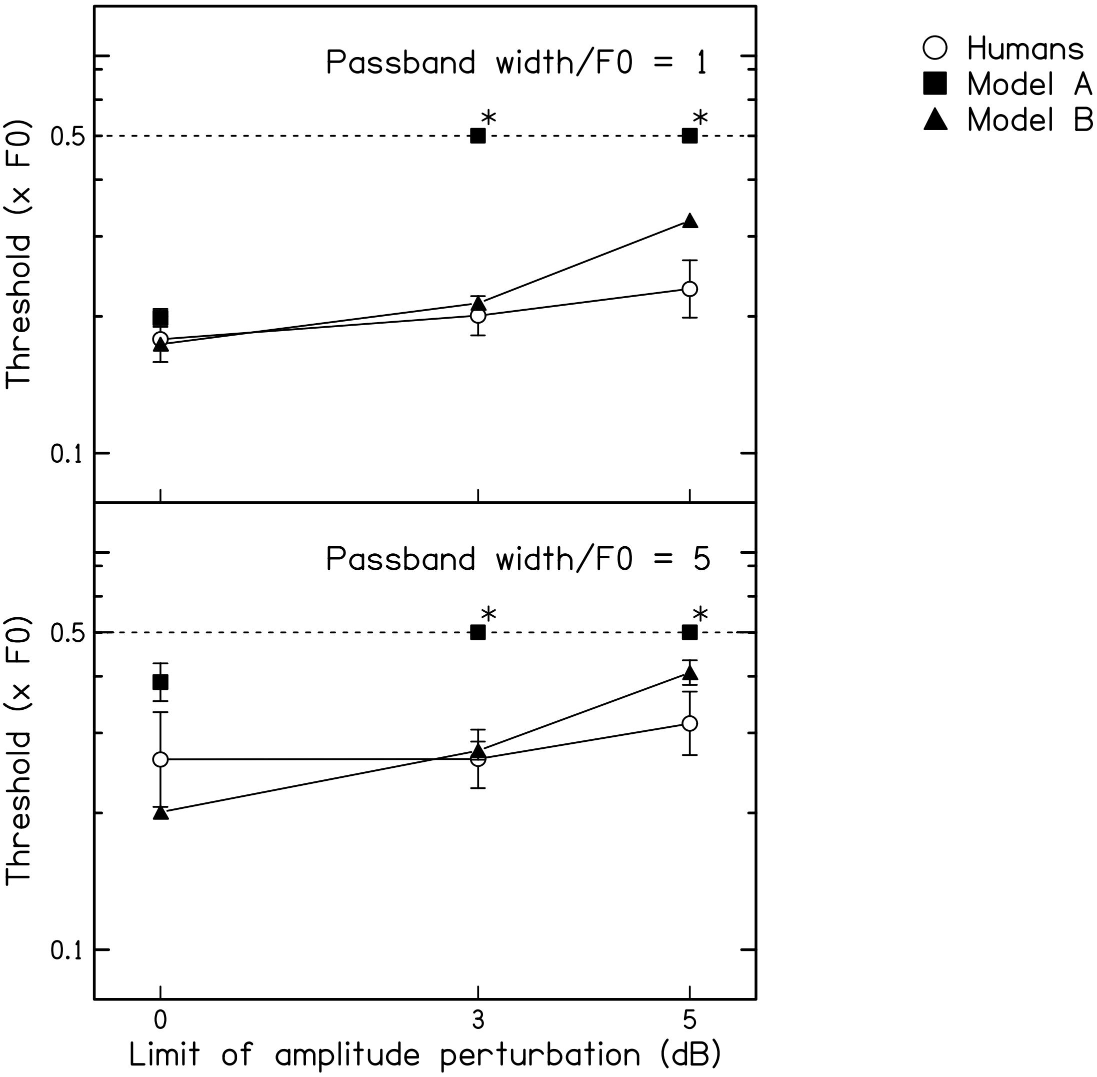


Figure 4.11: Geometric mean of predicted thresholds for model A and model B and obtained thresholds for each condition in experiment 4. Error bars for the models show the standard error of the runs, and for the humans show the standard error across subjects and phase.

4.6.4 Obtained data and comparison

A within-subjects ANOVA was conducted on the obtained data, with factors passband width, phase, and P, using the logarithm of the single combined threshold for each subject and condition as the variate. A logarithmic transform was used as the transformed data more closely fitted a normal distribution. There was a significant effect of passband width [F(1,6)=8.53; p=0.027]. There were no other significant effects or interactions.

Figure 4.12 shows the threshold for each subject and condition. As an ANOVA conducted on the data showed no significant effect of phase, or two-way interactions with phase, the geometric mean of the random-phase and cosine-phase data is plotted for each subject, with error bars showing the standard error of the six runs. As for the other experiments presented in this chapter, there were individual differences in the results, but the general pattern was consistent across all subjects.

Figure 4.11 shows the geometric mean of the random-phase and cosine-phase thresholds for all subjects for each condition and the data predicted by each model. For the human data, the geometric mean of all thresholds for each condition is plotted, with error bars showing the standard error.

A further two ANOVAs were conducted to compare the predictions of each model with the geometric mean of the data across subjects, ANOVA 4a for the human-model A data and ANOVA 4b for the human-model B data. The variate was the single threshold estimate for each condition. For each ANOVA, the three factors were data source (human listener or model A, or human listener or model B), passband width and P. The pooled variance associated with three-way interactions was used as an estimate of the residual variance.

For ANOVA 4a, there was a significant effect of data source [F(1,1)=32.34; p=0.03]. There were no other significant effects or interactions.

For ANOVA 4b, there was a significant effect of passband width [F(1,11)=65.71; p=0.015] and P [F(2,11)=62.62; p=0.016]. There was a significant interaction between data source and P [F(2,11)=15.72; p=0.06].

The effect of increasing P was much greater for both models than for the human listeners. In particular, when the passband width was 5F0, model B

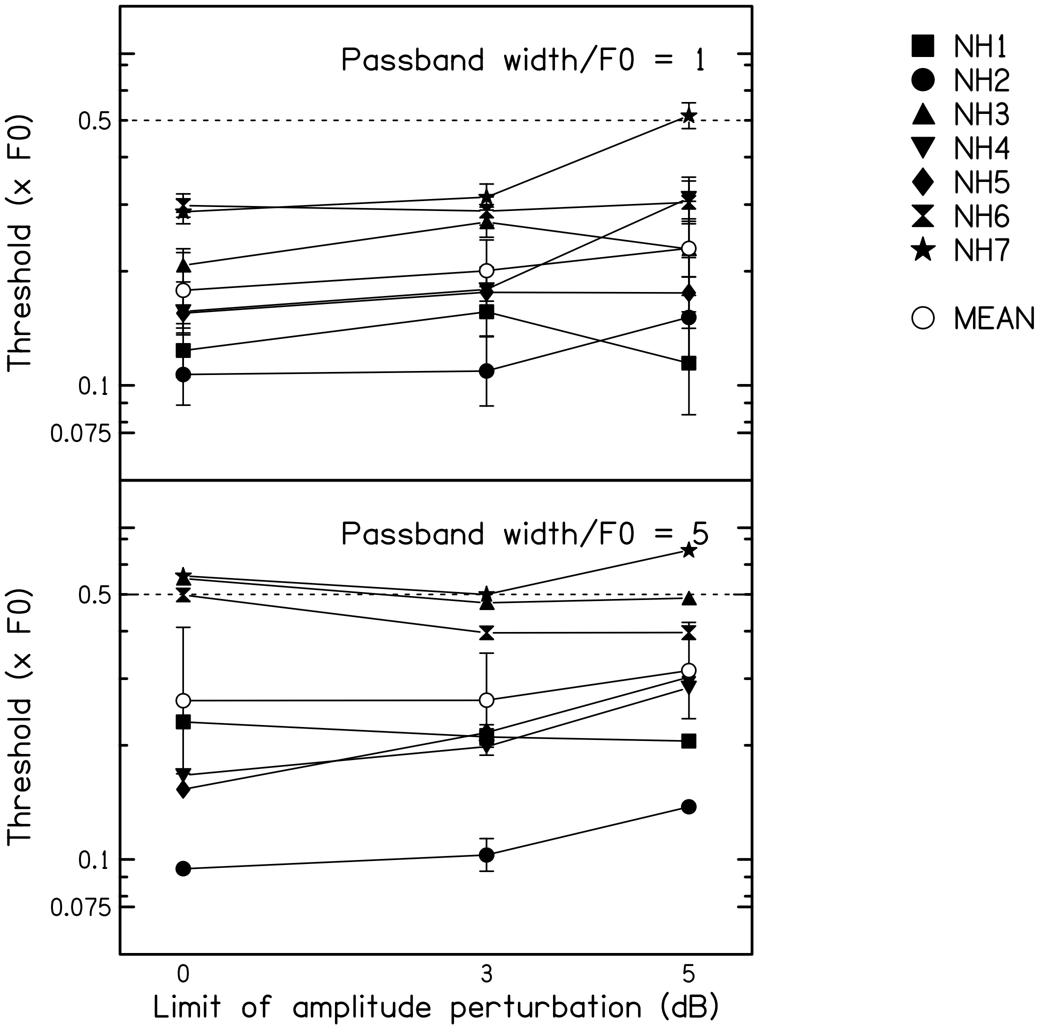


Figure 4.12: Geometric mean of six obtained thresholds for each condition and each human listener for experiment 1. Error bars show the standard error across phase.

predicted a lower threshold than obtained for $P = 0\,\text{dB}$, consistent with experiment 1, but predicted a higher threshold than obtained for $P = 5\,\text{dB}$. Model B predicted essentially a linear increase (on the log scale used) in threshold with increasing P.

4.6.5 Discussion

The fact that obtained performance was not significantly affected by the introduction of a random perturbation in level of each component strongly suggests that excitation-pattern differences between the H and I tones were not used to perform this task. Neither the most distinct ripple corresponding to the lowest component in the passband nor any overall pattern of ripples across the whole frequency range of the stimulus would have provided a reliable cue when P was 5 dB, which is why the models predicted a clear worsening of performance with increasing P. Whether human auditory filters have widths consistent with the model of Glasberg and Moore (1990) or are as much as twice as sharp, these predicted results demonstrate that if excitation pattern cues were used to discriminate the tones in this task then threshold would increase markedly with increasing P.

Adding a large perturbation to the level of all components in the tones might be expected to make the TFS vary more across both H tones and I tones, perhaps making it more difficult to detect differences in TFS between the H and I tones. Also, a component with a large amplitude might dominate the waveform at the output of auditory filters with CFs close to its frequency — particularly if neighbouring components happened to have low amplitudes — which would disrupt any measure of periodicity in these waveforms. Therefore it is presumed that performance based on TFS cues would worsen for a sufficiently high value of P. However, it seems possible that the disruption of TFS cues caused by $P = 5\,\text{dB}$ was not sufficient to prevent those cues from being used effectively.

It should be noted that three of the subjects used in this experiment showed an increase in threshold between $P = 3\,\text{dB}$ and $P = 5\,\text{dB}$. This occurred for subjects NH2, NH4 and NH7 for the passband width of F0, and NH2, NH4 and NH5 for the passband width of 5F0. In each case, the increase was quite small, and the fit of the ANOVA to the data was good. It would be interesting in the future to determine a value of P for individual subjects above which performance worsens significantly; it is possible for some subjects here, perhaps subjects NH2 and NH4, that this value of P was less than 5 dB.

4.7 General discussion

Overall, the experiments described here, particularly experiment 4, show that it is likely that subjects did not use differences in excitation patterns between the H and I tones to perform the task, indirectly supporting the TFS hypothesis. Experiments 1 and 4 showed that human performance worsened with increasing passband width, and both experiments 1 and 2 showed a worsening with increasing N and decreasing F0. The predictions of model A and model B followed the same trends as the obtained data for experiments 1 and 2, but to different extents. In particular, the worsening predicted by model B for an increase in passband width or for a decrease in F0 was much less than obtained. Experiment 3 showed that the effect of passband width was not due to a change in signal-to-noise ratio for either the human listeners or the models. Experiment 4 showed that human performance was not disrupted by adding a random level perturbation to each component in each tone, while performance predicted by both model A and model B was markedly disrupted. The results from experiment 4 in particular suggest that the human listeners were not using differences in a regular pattern of spectral ripples to discriminate the tones in this task.

In what follows, the focus is on the extent to which the pattern of results in these four experiments is consistent with the TFS hypothesis.

4.7.1 Effect of passband width

Both experiments 1 and 4 showed that human performance worsens with increasing width of the passband. The TFS hypothesis would explain this in terms of the integration of timing information across auditory filters with different CFs; for a large passband width, more auditory filters respond to the signal than for a narrow passband width. For the I tone, the TFS information is different (conflicting) at the output of different auditory filters, for example filters centred at the lower and upper edges of the passband, as demonstrated in Figure 4.5. The TFS information may be easier to analyse when the conflict is reduced by decreasing the bandwidth of the stimulus.

4.7.2 Effect of N

The results of experiments 1 and 2 showed that performance for a given F0 worsened with increasing N. This is consistent with previous results obtained using this or a similar task (de Boer, 1956a; Hopkins and Moore, 2007; Moore and Sek, 2009a; Moore *et al.*, 2009b). The worsening with increasing N is also consistent with much other work measuring F0DLs for harmonic tones with different values of N (Houtsma and Goldstein, 1972; Hoekstra and Ritsma, 1977; Houtsma and Smurzynski, 1990; Bernstein and Oxenham, 2003; Moore *et al.*, 2006; Oxenham *et al.*, 2009). The effect for both types of discrimination could occur partly because, as N increases for a fixed F0, adjacent peaks in the TFS of the waveform become closer together in time. This means that the time intervals to be discriminated also become closer together. This is consistent with the first effect of pitch shift, as described by Schouten *et al.* (1962).

4.7.3 Effect of F0

The poorer performance for low F0s is consistent with earlier work (Moore and Sek, 2009a; Moore *et al.*, 2009b) showing that performance worsens when F0 is decreased below 200 Hz, especially for N = 9. Moore and Sek found that, for N = 9, all of their normal-hearing listeners could discriminate the H and I tones when F0 was 100 Hz, but only about half of the listeners could perform the task reliably when F0 was 50 Hz. The present results are consistent with these findings, in that while some of the listeners tested had thresholds close to 0.5F0 for F0 = 50 Hz, all but one achieved thresholds well below 0.5F0 for F0 = 75 Hz and higher. The worsening in performance for low F0s is consistent with the idea that the auditory system is unable to measure long time intervals accurately (Krumbholz *et al.*, 2000; de Cheveigné and Pressnitzer, 2006; Moore and Sek, 2009a; Moore *et al.*, 2009b).

Moore and Sek (2009a) proposed the procedure used here as a measure of sensitivity to TFS that might be applied in the clinic to assess hearing-impaired listeners, and they pointed out that this required that normally hearing listeners should be able to perform the task consistently. The fact that normally hearing listeners could not consistently perform the task for F0 = 50 Hz (an effect confirmed here) meant that the task could not be used to assess sensitivity to TFS for frequencies as low as 450 Hz. As the present results include data for an F0 of 75 Hz that Moore and Sek did not, it can be seen that this task can be used to measure TFS sensitivity for frequencies down to 600 Hz.

4.7.4 Effect of phase

There was no significant main effect of phase in any of the four experiments here (though there was a significant three-way interaction involving phase for experiment 1). This suggests that listeners were not able to use the single large peak per period of the cosine-phase tones to improve performance relative to that for random-phase tones.

We described in the introduction that there might have been subtle differences in the auditory system's representation of the envelope of the H and I tones when the components were added in cosine phase (see Figure 4.1). Such a cue would not have been available for the random-phase tones, since there were large differences in envelope shape from one stimulus to the next. If listeners did use this envelope cue for the cosine-phase stimuli, then performance should have been better for these stimuli. The fact that it was not suggests that envelope cues were not being used to perform the task.

Performance of the task using TFS cues does require envelope cues to be present in general, since it is assumed that listeners perform the task by estimating time intervals between peaks in the TFS close to adjacent envelope maxima, as illustrated in Figure 4.1. If the envelopes at the outputs of the auditory filters were flat, or nearly so, the auditory system might have difficulty in determining which time intervals in the TFS to estimate in order to discriminate the H and I tones. However, even for random-phase stimuli, for which the envelope fluctuations are less prominent than for cosine-phase stimuli, simulations suggest that the waveforms at the outputs of auditory filters have distinct periodic envelope fluctuations. This could explain the fact that performance did not differ significantly across phase conditions.

4.7.5 Resolvability

The fact that no significant main effect of phase was demonstrated here does not necessarily imply that the harmonics were resolved. While the presence of a phase effect indicates that at least some harmonics are unresolved, the presence of unresolved harmonics will not always result in a phase effect (Moore *et al.*, 2006). It is possible that harmonics in a complex tone can at the same time be sufficiently unresolved that TFS cues can be used, whilst not being "unresolved enough" to produce a phase effect.

In the experiments described here, the differences between the performance of human listeners and the predictions of the computer models, particularly for experiment 4, show that it is unlikely that human listeners used information derived from excitation-pattern cues (peaks corresponding to individual components) to perform the task. Pitch derived from resolved components in a complex tone is strong and robust, whereas pitch derived from completely unresolved components, although perceivable, is weaker (Houtsma and Smurzynski, 1990; Carlyon and Shackleton, 1994; Ives and Patterson, 2008) and more ambiguous. It is probably safe to assume that the auditory system makes the best use of available cues, so if a component in the tones were even partially resolved then excitation-pattern information evoked by it might be used to perform the task. The interaction between passband width and F0 that fell just short of significance in experiment 1 could be explained by sensitivity to the 8th component when N was 9. The 8th component is unlikely to be fully resolved on the basilar membrane, but use might be made, for example, of phase-locking information reflecting its frequency in the outputs of auditory filters with CFs in that frequency region.

4.7.6 Computer modeling

The computer models used here were developed to allow prediction of the pattern of results that would be expected if the H and I tones were discriminated using only the regular ripples in their excitation patterns. It is not claimed that human listeners actually operate in exactly the same way as the models, or that this is the best method for discriminating regular changes in excitation patterns. There are many ways in which one could construct a decision rule based on excitation-pattern information. As such, no great importance should be attached to the absolute values of the thresholds predicted by the model, particularly as the level of the internal noise was arbitrarily chosen to match predictions to human thresholds in specific conditions for model B. However, we believe it is likely that the pattern of predicted results based on excitation-pattern cues would be similar regardless of the exact implementation of the model, particularly for experiment 4. Therefore, deviations between the obtained and predicted pattern of results can reasonably be taken as indicating that performance cannot be fully explained in terms of the use of excitation-pattern cues.

4.8 Conclusions

1. Overall, the experiments presented here favour the idea that human performance in Moore and Sek's (2009a) task is not based on excitation-pattern cues. In particular, experiment 4 showed that disrupting excitation-pattern cues via perturbation of the levels of components had a marked detrimental effect on the performance of the excitation-pattern models but no significant effect on the performance of human listeners. This indirectly supports the idea that human listeners used TFS cues to discriminate the tones in this task.

2. Model A did not predict thresholds that were as low as those of the human listeners. The sharper auditory filters used by model B led to better predicted performance, requiring the introduction of a variable representing internal noise to match model B's performance to human performance for two baseline conditions. However, model B predicted a smaller worsening in performance with decreasing F0 or increasing N than was actually observed. Thus, the use of sharper filters than in the model of Glasberg and Moore (1990) did not lead to accurate predictions of the pattern of results.

3. As expected, performance worsened with increasing N. This could be because the intervals that the auditory system must discriminate in order to use TFS cues become closer together with increasing harmonic rank of a stimulus.

4. Human performance in this task improved as the number of components in the stimuli was reduced, which could be due to a decrease in "TFS conflict" across auditory filters. Reducing the number of components in a complex tone simplifies the conflicting TFS information across channels and results in more effective use of TFS cues to discriminate the tones in this task.

5. There appears to be a decrease in TFS processing ability with decreasing F0. This confirms earlier results suggesting that the auditory system has difficulty in estimating long time intervals accurately. The present results show that normally hearing listeners could perform the task consistently for F0 = 75 Hz with $N = 9$, which means that the task can be used diagnostically for frequencies of 600 Hz and above.

6. Performance in the task was still good when all audible components in the tones had frequencies higher than 6000 Hz, consistent with the findings of Moore and Sek (2009b). Human listeners could achieve good performance when F0 was 400 Hz and N was 16; in this condition, the flat part of the passband extended from 6200 Hz – 8200 Hz. This is above the usual assumed limit of 5000 Hz for the use of phase locking information. However, it is possible that TFS information can still be used to discriminate such tones.

7. The lack of a phase effect confirms that listeners were not using envelope cues alone to perform this task.

Chapter 5

The contribution of resolved harmonics to pitch perception for low fundamental frequencies

5.1 Introduction

The previous two chapters described a method for measuring human sensitivity to temporal fine structure (TFS), and provided some evidence for how sensitivity changes under different conditions. This chapter presents some further, indirect, evidence for the use of TFS cues in pitch perception, preceded by a brief overview of the literature in this area.

5.1.1 Complex tone pitch perception

There has much debate spanning almost two centuries over whether our sense of musical pitch is derived from neural information related to individual components within complex tones, or from the inter-spike intervals evoked by vibration of the basilar membrane at places that respond to more than one component. The perceptual phenomenon of the "missing fundamental", first reported by Seebeck (1841), provides an important piece of evidence in this debate. It refers to the perceived pitch of a complex tone, whereby a pitch corresponding to the fundamental frequency (F0) of the tone is perceived whether or not energy at that frequency is physically present. Furthermore, the missing fundamental is still heard when a low-frequency noise, which would mask the fundamental component, is presented with the stimulus, showing that perception of the missing fundamental does not depend on there being a local peak in the excitation on the basilar membrane at the place corresponding to the frequency of the missing fundamental (Licklider, 1954). There are two main classes of theory to account for this phenomenon: pattern-recognition models, and temporal models.

5.1.1.1 Pattern-recognition models

Pattern-recognition theories propose that the perception of the pitch of complex tones relies on harmonics that are resolved by the auditory system (Terhardt, 1972a,b; Goldstein, 1973). Resolved harmonics lead to individual peaks in the pattern of vibration on the basilar membrane, allowing a clear spatial representation of the component frequencies in the cochlea. In addition, the place on the basilar membrane that responds maximally to a resolved harmonic evokes a pattern of phase locking very similar to that which would be evoked by a single sine wave at that frequency. Thus, the frequency of a harmonic may be estimated from the pattern of phase locking as well as from the position of the local peak in excitation. The estimates of the frequencies of resolved harmonics are assumed to be analysed by a central processor or pattern recogniser, which determines the low pitch from the estimates. Terhardt (1972a,b) proposed that the ear generates subharmonics of resolved partials, and then looks for a common frequency amongst these sub-harmonics that could be a candidate for the low pitch. He suggested that resolved components might be those in the frequency range between 500–1500 Hz. Another possible mechanism for the operation of the pattern recogniser was proposed by Goldstein (1973), who suggested that the central processor might assume the resolved components to represent successive components of a harmonic series, and might then calculate the fundamental component of the harmonic series which provides the best fit to the identified components.

5.1.1.2 Temporal models

Schouten (1940b) proposed a temporal theory, which relies on the interaction between unresolved harmonics in the cochlea. Due to the fact that the bandwidth of auditory filters (in Hz) increases at higher centre frequencies (CFs), the higher-ranked components within a complex tone will not be "heard out", that is, individually perceived as having distinct pitches. Instead, due to their closeness on the basilar membrane, these components interfere with each other within a single auditory filter, resulting in an overall waveform with a periodicity corresponding to the F0 of the complex tone. Schouten proposed that auditory nerve neurones phase lock to the TFS or to the temporal envelope of this wave, and that the auditory system derives pitch from the resulting inter-spike intervals. This hypothesis has since been supported by physiological data, for example in the work of Cariani and Delgutte (1996), where it was found that the pitch percept of subjects corresponded to the inter-spike intervals measured from auditory nerve neurones. Schouten named the perceived missing fundamental the "residue", implying that it is somehow the "left over" product of higher-ranking harmonics. However, for clarity, the term "low pitch" will be used to refer to the missing fundamental throughout this review, as a low pitch is also evoked by low-ranking components.

5.1.2 Peripheral resolvability

Since temporal theories of pitch perception depend on unresolved harmonics in a complex tone whereas pattern-recognition theories depend on resolved harmonics, it is necessary to describe the frequency selectivity of the auditory system quantitatively. Both direct and indirect methods have been employed in order to do this, and an overview of some relevant evidence is given in chapter 2 of this thesis. To summarise, the data are generally consistent with the idea that a component in a complex tone can be heard out when the frequency separation between it and neighbouring partials is more than 1.25 times the value of the equivalent rectangular bandwidth of the auditory filter ($\mathrm{ERB_N}$) at that frequency. Calculation using this assumption gives a limit of the 7th or 8th harmonic for an F0 of 200 Hz, and a lower limit for lower F0s, possibly as low as the 2nd or

3$^{\text{rd}}$ harmonic for an F0 of 50 Hz.

It is clear that pattern-recognition models of pitch perception, which depend on one or more components in the complex tone being resolved by the auditory system, cannot account for the fact that a low pitch can be evoked by high-ranking groups of components. For example, clear pitch percepts have been measured for stimuli containing only unresolved harmonics (Ritsma, 1967; Moore and Rosen, 1979; Houtsma and Smurzynski, 1990). This raises the question of whether the low pitch evoked by a group of unresolved harmonics is derived from the TFS of the waveform at the output of auditory filters with high CFs, or from its envelope. This question is addressed later in the review.

The series of experiments conducted by Houtsma and Smurzynski (1990) included a task where musically trained subjects were asked to name the musical intervals between the low pitches of groups of 11 high-ranking harmonics. They found that, although performance was poorer when the lowest-ranking harmonic present (N) was the 13$^{\text{th}}$ than when it was the 7$^{\text{th}}$, performance did not worsen further for higher ranks. Furthermore, performance was still significantly better than chance for $N = 25$. Although pattern-recognition models cannot account for these results, it must be noted that the low pitch evoked by a group of very high harmonics is much less clear than that evoked when resolved components are present, and is described qualitatively as differing in timbre, being more buzz-like. Houtsma and Smurzynski used the same reference tone in each trial meaning that, for each pair of tones, the first always had the same pitch (200 Hz). Therefore, subjects could have developed some kind of memory for the pitch of the reference tone, which may have enhanced interval-naming performance across all trials.

In another experiment, Houtsma and Smurzynski measured fundamental frequency difference limens (F0DLs) for the same stimuli. It was found that performance became worse as N was increased from 7 to 13, but did not become significantly worse for higher N. These experiments both show that the low pitch evoked by groups of lower-ranking harmonics is stronger than that evoked by groups of higher-ranking harmonics. It is likely that the stimuli in these two experiments contained components that were unresolved (or nearly fully unresolved) as the lowest harmonic rank included in their stimuli was 7 for F0 = 200 Hz. Therefore, it is unlikely that the worsening as N was increased from

8 to 13 was due to a progressive decrease in component resolvability, as they suggested.

It seems that pitch can be conveyed by groups of resolved harmonics, unresolved harmonics and combinations of the two. The question of whether resolved or unresolved harmonics are more important for pitch perception has clear implications for the proposed theories mentioned previously; if the harmonics that are most important for pitch perception are those that are resolved, then pattern-recognition theories would provide a better account of how the auditory system perceives pitch. Hence, we reach the concept of dominance.

5.1.3 Dominant regions for pitch perception

The work described so far was concerned with assessing whether particular groups of harmonics lead to the percept of a low pitch, and with measuring the salience of the low pitch when N was varied. An alternative method is to identify which harmonics have the greatest effect on the low pitch perceived: this is the concept of dominance, introduced by Ritsma (1967). Ritsma used stimuli in which the frequencies of one group of harmonics were multiples of a slightly higher or lower F0 than for the other harmonics in the tone. Fundamental frequencies ranged from 100 to 400 Hz. Subjects' pitch judgements were used to assess the relative influence of different groups of harmonics on the pitch of the complex. Using this method, Ritsma found that the 3rd, 4th and 5th harmonics tended to dominate the pitch sensation, as long as their combined level was at least 10 dB above threshold. Work by Plomp (1967) broadly supported this idea, and both sets of data showed a slight trend for the rank corresponding to the dominant region to decrease with increasing F0.

Later experiments attempted to further narrow down the region of dominance by assessing the influence of individual harmonics on the overall pitch of the complex. Moore, Glasberg, and Peters (1985) asked subjects to match the pitch of complex tones where one component was slightly mistuned from its correct harmonic value to the pitch of harmonic complex tones with the same number of components. They found that mistunings in each of the first 6 components had an effect on the pitch of the complex, although there was considerable variation between subjects as to which of these components most affected the pitch. In a related experiment (Moore, Glasberg, and Shailer, 1984), the threshold frequency change for partials within harmonic complex tones was measured. They found, for most subjects, that thresholds were lowest for the 3$^{\text{rd}}$, 4$^{\text{th}}$ and 5$^{\text{th}}$ harmonics of a complex tone with a F0 of 100 Hz. Thus, the dominant harmonics are the ones whose frequencies are discriminated the best. However, as only one component was mistuned at a time in these experiments, listeners' attention may have been cued towards it, which might result in that component's contribution to the overall pitch being weighted more highly than normal relative to the other components in the tones.

Gockel *et al.* (2007) carried out a similar experiment to assess the effect of duration (16, 50 or 200 ms) on FDLs (frequency difference limens) for the first seven harmonics of a complex tone with an F0 of 250 Hz. Gockel *et al.* found that FDLs for each harmonic increased, and that the harmonic rank of the dominant region increased, with decreasing duration. Their results were consistent with the hypothesis of Moore *et al.* (1984) that F0DLs for a given complex tone can be predicted from the FDLs for the components within it for the durations of 200 and 16 ms, but predicted F0DLs based on this hypothesis were consistently larger than obtained F0DLs for the duration of 50 ms. They concluded that estimates of the pitch of individual harmonics within a complex tone are not always combined optimally when the task requires discrimination of the fundamental frequency.

Overall, all the results concerning the dominant region support pattern-recognition theories; it seems clear that low-ranking, resolvable harmonics have a greater effect on the low pitch of a complex tone than high-ranking, unresolved harmonics.

5.1.4 Low F0s

There has been little published work concerning pitch perception for complex tones with very low F0s. These low F0s are interesting and relevant because the range of F0s in human speech includes low values, so understanding more about the mechanisms of pitch perception for low F0s could have implications for the hearing impaired.

Moore, Hopkins, and Cuthbertson (2009b) measured thresholds for discriminating harmonic complex tones from inharmonic tones, which were created by adding the same value in Hertz to the frequency of each component of the harmonic tone. Tones were bandpass filtered to reduce excitation-pattern differences between the tones, and the bandpass filter used had a central flat region with a width equal to 5F0 and skirts with a slope of 30 dB/octave. The harmonic rank of the central component in the flat part of the filter, N, was varied, as was the nominal F0. They found that performance for an F0 of 35 Hz was no better than chance for $N = 13$ and $N = 15$, whereas for all other F0s tested (50–400 Hz) performance was better than chance for these values of N. Performance for F0 = 35 Hz and F0 = 50 Hz was poorer than for the other, higher F0s tested, with performance improving up to F0 = 100 Hz. In a second experiment, F0DLs were measured for the same stimuli. Performance improved with increasing F0 from 35 to 100 Hz.

One explanation for the poor F0 discrimination found for low F0s is related to the resolvability of components in a complex tone. Glasberg and Moore's (1990) equation predicts that the equivalent rectangular bandwidth ($\mathrm{ERB_N}$) of auditory filters with low CFs is wider as a proportion of the CF than at higher centre frequencies, meaning that fewer harmonics would be resolved on the basilar membrane, assuming that harmonics are resolved if their frequency separation is greater than 1.25 times the $\mathrm{ERB_N}$ at that frequency. Chapter 2 presented Table 2.1 relating frequency separation to $\mathrm{ERB_N}$-number, which shows that only the first three harmonics are well resolved (for this definition) for an F0 of 50 Hz.

Another possible explanation for the worsening in pitch perception for low F0s is offered by de Cheveigné and Pressnitzer (2006), who suggested that the ability of the auditory system to measure time intervals accurately decreases as the time intervals become longer. Pressnitzer, Patterson, and Krumbholz (2001) measured the lower limit for melodic pitch using a melody-change task. They proposed that an autocorrelation model could predict their data if an upper limit of 33 ms were imposed on the time intervals that the auditory system can measure between peaks in the TFS of the waveform at the output of auditory filters centred on unresolved harmonics. These findings could explain the progressive worsening in pitch perception for low F0s demonstrated by Moore *et al.* (2009b), and the lower limit for melodic pitch.

One remaining question concerning dominance is whether the 3rd, 4th and 5th harmonics (or at least harmonics contained within the first 6) continue to dominate pitch perception for complex tones with low F0s, below 100 Hz, given that the 3rd is the highest-ranking component that is likely to be resolved on the basilar membrane for an F0 of 50 Hz. Moore and Glasberg (1988) measured F0DLs for complex tones containing either harmonics 1–5, 1–12 or 6–12 with F0s of 50, 100, 200 and 400 Hz. For F0s of 200 and 400 Hz, F0DLs were roughly equal for tones containing harmonics 1–5 and harmonics 6–12. However, for an F0 of 50 Hz (and 100 Hz for one subject), F0DLs were smallest for the tone containing harmonics 6–12, and the tones containing only harmonics 1–5 were quite poorly discriminated. This shows that higher-ranking harmonics make a more important contribution to the discrimination of complex tones with low F0s than they do to tones with higher F0s. Similarly, some unpublished data collected by Pinker (2006) implicate higher-ranking harmonics in the pitch perception of complex tones with low F0s. Pinker used stimuli comprising two groups of harmonics with different F0s, and found that harmonic ranks around 16 dominated for an F0 of 50 Hz, but that harmonic ranks around 19 dominated for an F0 of 35 Hz. The difference between the results of Moore and Glasberg (1988) and Pinker (2006) could be due to the differing nature of the tasks and the stimuli used; the important point is that the estimates of dominant harmonic ranks for low F0s are somewhat higher than for intermediate or high F0s.

5.1.5 Rationale

A number of studies have measured aspects of pitch perception associated with individual components within complex tones, whether a frequency difference limen (FDL) for an individual component (Moore *et al.*, 1984) or the effect of mistuning a single component on the overall pitch of the complex (Moore *et al.*, 1985). One underlying assumption here has been that the partials that have the smallest thresholds or greatest effect of the overall pitch of the complex are given preferential “weighting” by the auditory system, in line with the concept of dominance discussed earlier. Another assumption has been that dominant components are resolved (based for example on the studies of Ritsma (1967) and Plomp (1967) and that pitch is derived from estimates of the frequencies of individual harmonics, based on either place or temporal information, or both. This approach makes sense when the single component to be discriminated is resolved. If the dominant components were unresolved, and pitch was based on the temporal structure of the waveform evoked by a group of harmonics, such paradigms would not be appropriate, as shifting the frequency of a single harmonic would disturb the periodicity of the waveform, possibly disrupting the main cue that is used for F0 discrimination.

The experiments presented in this chapter extended the paradigm of Moore *et al.* (1985) by shifting the F0 of a group of harmonics embedded in a harmonic background rather than shifting just a single harmonic. The main intention here was to ascertain which harmonic ranks have the greatest effect on the pitch of complex tones with low F0s, for which it is likely that fewer harmonics are resolved on the basilar membrane. Comparing thresholds when components were added in random and cosine phase, and with and without the availability of pitch pulse asynchrony (PPA) cues (see below for an explanation), was intended to provide some indirect evidence as to the extent to which the harmonics were resolved and hence for the use of TFS cues for pitch perception in this paradigm.

In the experiments presented here, F0DLs were measured for a group of harmonics of varying harmonic rank and F0, rather than for a single harmonic. The stimuli consisted of a harmonic background, group A, with a fixed F0, and a group of harmonics, group B, for which the F0 was shifted either up or down by $0.5 * \Delta$F. The value of ΔF was varied adaptively to measure a threshold for each condition. The lowest harmonic rank assigned to group B, N, was varied, as was the number of harmonics assigned to group B, B. Harmonics with ranks corresponding to those in group B were absent from group A, and the harmonics adjacent to the spectral edges of B (harmonics N and $N + B$) were also omitted in order to avoid overlap of harmonics from groups A and B, which could have resulted in beating.

Criticisms of studies adopting the experimental design of Moore *et al.* (1984) have included the suggestion that mistuning a partial from its correct harmonic value draws attention to that partial, resulting in analytic listening by the subject, rather than the synthetic listening required to hear a single low pitch. The design of the stimuli in the present experiment, specifically the fact that a group of components rather than a single component was used, was intended to reduce the effect of analytic listening when synthetic listening was required. However, a form of analytic listening was still required to separate the F0 of group B from the F0 of group A. As in the experiment of Moore *et al.* (1984), it was assumed that F0DLs would be smallest when the components in group B fell in the dominant region. Furthermore, shifting the F0 of a group of components ensured that the group to be discriminated was harmonic, preserving the regular spectral structure and regular periodicity of that group. The presence of the background harmonics also prevented subjects from listening to the spectral edges of the shifted group rather than to its low pitch.

The components in the tones were added in either random or cosine phase to identify any effect of phase that would indicate that the components in the tones were unresolved. When the components in a complex tone have frequencies that fall within the passband of the same auditory filter, the waveform at the output of that filter is equivalent to the sum of those individual components at the levels they would have after any attenuation has been applied by the filter. When components are presented in cosine phase, the resulting waveform shows one large envelope peak per period, which can be used by the auditory system to enhance the accuracy of pitch estimates based on temporal cues. When components are added in random phase, the resulting waveform is much flatter and may have more than one envelope peak per period. It is generally agreed that if cosine phase results in better performance, a “phase effect”, then unresolved components are present, although unresolved components will not always result in a phase effect.

The number of components in the shifted group, B, was also varied. Discriminating the F0 of a group of unresolved components may be easier for cosine phase when more components are in the stimulus, as the waveform on the basilar membrane resulting from the interfering components will have a higher peak factor.

5.1.6 Pitch Discrimination Interference

In a paradigm where two complex tones with differing F0s are presented simultaneously, pitch discrimination of the "target" tone can be impaired by the presence of the other "distractor" tone, even if the two tones contain components occupying differing frequency regions (Gockel, Carlyon, and Plack, 2004). This effect is termed pitch discrimination interference (PDI). Gockel *et al.* measured d′ scores for the discrimination of the F0 of a complex tone with unresolved harmonics in the presence of a complex tone with a fixed F0 filtered into a lower frequency region. The effect of PDI was tuned for F0, that is, the impairment to performance was greater when the two F0s were similar, with performance improving as the distractor's F0 was moved further away in frequency from that of the target.

Gockel *et al.* measured performance for conditions where both the target and background complex tone contained either resolved or unresolved harmonics. They found that a resolved interferer produced a large PDI effect for an unresolved target, which they interpreted as evidence that there cannot be two completely independent pitch mechanisms, one for resolved harmonics and one for unresolved harmonics. Gockel *et al.* also found that the effect in the opposite direction was small, that is, an unresolved interferer produced a negligible PDI effect for a resolved target, and that the effect of PDI was much smaller when both tones contained unresolved harmonics. They interpreted these findings in terms of the greater pitch salience of a resolved complex relative to an unresolved complex.

Gockel, Carlyon, and Moore (2005) provided an alternative explanation for the reduced PDI when both target and background were unresolved. When components in a complex tone are added in the same phase relative to each other, the resulting waveform has a high peak factor; there is a single large peak, or "pitch pulse", per period. When components are added in the same phase relative to each other in a situation with two competing F0s, the pitch pulses corresponding to each F0 in different regions of the basilar membrane have differing periodicities, resulting in an asynchrony across the output of different auditory filters. This pitch pulse asynchrony (PPA) does not occur when components are added with differing phases, such as random phase, because the resulting waveform at the output of each auditory filter does not have a high peak factor, so there are no clear pitch pulses as there are for cosine phase. Carlyon and Shackleton (1994) showed that listeners were sensitive to changes in PPA. Gockel *et al.* (2005) suggested that this sensitivity would enhance performance in the condition of Gockel *et al.* (2004) where both the target and background complex tones were unresolved, which could explain why the effect of PDI was not as great in that case as it was when both complexes were resolved.

Gockel *et al.* (2005) extended these findings by showing that listeners were sensitive to the direction of the PPA, that is, that listeners were sensitive to whether the pitch pulses evoked by the target F0 were advanced or delayed in time relative to the pitch pulses evoked by the distracting F0.

Miyazono, Glasberg, and Moore (2009) used similar stimuli to those for experiment 1 in the present chapter to assess which harmonics in a complex tone were the most important for pitch perception at low F0s. They used an F0 of 50 Hz, and the values of N were the same as described for the present experiments, with either 4 or 10 components in group B. Miyazono *et al.* found that, for random phase, thresholds increased with increasing N as expected. For cosine phase, thresholds actually decreased (improved) with increasing N, which they attributed to the use by the listeners of the clear single peak per period in the waveform at the output of auditory filters centred on unresolved harmonics. An alternative explanation for this finding is the use of PPA cues when tones were added in cosine phase.

Consider, as an example, a case where the F0 of group A was 50 Hz, N was 15, B was 8 and ΔF was 0.1F0 (5 Hz). In one interval, the F0 of group B was 47.5 Hz and in the other it was 52.5 Hz. Consider the auditory filter X with a CF of 550 Hz, and the auditory filter Y with a CF of 950 Hz. Each of the filters X and Y responded to unresolved components in the stimuli. The output of filter X was dominated by components in group A, and the output of filter Y was dominated by components in group B. At the start of each stimulus, all components in both tones had cosine phase, so the pitch pulses at the output of filter X and filter Y were synchronous (zero PPA). The PPA between the outputs of filter X and filter Y would go through one complete cycle in 1/5 seconds, so the PPA would again be zero after 200 ms. When the F0 of group B was higher than that of group A, the pitch pulses at the output of filter Y occurred before those at the output of filter X, the opposite being the case when the F0 of group B was lower than the F0 of group A. Therefore, if listeners could compare the relative timing of the pitch pulses in the outputs of filter X and filter Y towards the beginning of each stimulus interval, then the interval in which the F0 of group B was higher could be identified due to its pitch pulses leading relative to the pitch pulses evoked by group A, without the pitch difference itself being perceived.

If, instead of being presented with zero PPA, the onset time of the pitch pulses in group B relative to group A were varied for each interval, then the PPA between the outputs of filters X and Y would no longer be a reliable cue for discriminating between the tones. Miyazono, Glasberg, and Moore (2010) tested this using stimuli like those of Miyazono *et al.* (2009), but with either a fixed or randomly varying pitch-pulse offset for group B. The fixed offset was either 0, 5, 10 or 20 ms, and the random offset was selected randomly from a uniform distribution with a range of 5, 10 or 20 ms, independently for each interval. The aim of the experiment was to identify whether the perceptual learning demonstrated by Miyazono and Moore (2009) for these stimuli was likely to have been based on PPA cues. The F0 used was 50 Hz. For the fixed-offset conditions, offsets of 5 ms or 10 ms resulted in significantly worse performance than offsets of 0 ms or 20 ms, suggesting that PPA cues could not be used when there was initial asynchrony between the pitch pulses evoked by groups A and B. For the random-offset conditions, performance worsened up to the range of 10 ms. Miyazono *et al.* noted that their results might have been influenced by the detection of beats and combination tones produced by the interaction of components within and across groups, but argued that the levels of any combination tones present would have been too low to have had an effect on performance. Overall, Miyazono *et al.* (2010) concluded that the perceptual learning effect demonstrated by Miyazono and Moore (2009) was partly mediated by PPA cues.

A similar manipulation was done for experiment 2 described here, where the initial PPA between groups A and B was varied randomly by phase-shifting each component in group B, while keeping the onset of groups A and B synchronous. This resulted in there being no reliable cue in the direction of the PPA.

The main question addressed in the present experiments was: which are the dominant harmonics for low F0s, as measured in a task similar to that of Moore *et al.* (1985), but shifting a group of harmonics rather than a single harmonic? The experiments were conducted under conditions where PPA cues might have been used (experiment 1) and where they probably could not be used (experiment 2). If PPA cues play a role for some conditions (*i.e.* if there are differences between the results for the two experiments for some conditions), this implies that components were unresolved both for the components in group A and for (at least some of) the components in group B.

5.2 Method

5.2.1 Subjects and training

Six subjects took part in experiment 1, and seven subjects took part in experiment 2. All subjects had absolute thresholds of 10 dB HL or below for both ears. The subjects' ages ranged between 19 and 31 years. Two of the subjects took part in both experiments: subjects 1 and 5. All but one of the subjects had some experience with playing musical instruments.

All subjects were trained on the task for two hours before the experiment began. The subjects were exposed to a range of conditions similar to those used in the experiment, and repeated these conditions twice. The initial difference in the F0 of group B was 0.05F0 for each run, and all subjects achieved thresholds much lower than this for the conditions with low values of N.

5.2.2 Stimuli

The stimuli were complex tones comprising the lowest 59 harmonics. Group B always included only harmonics within the first 30, with the remainder being assigned to group A. The upper 29 harmonics formed a "tail", with successive harmonics decreasing in level by 3 dB per component to avoid an "edge pitch" (Kohlrausch and Houtsma, 1992).

Subjects completed trials in blocks where the value of B and the phase of the components were held constant within a block, and the value of N was systematically increased from 1 to 15 in steps of 2, and then decreased back to 1 again. Therefore, a block comprised 16 runs of 50 trials each. Where random phase was used, the phase for each component was recalculated for each stimulus.

In each trial, two tone-pulses were presented sequentially, each of 500-ms duration, including 50-ms onset and offset ramps to minimise spectral splatter. The inter-stimulus interval lasted 300 ms.

The stimuli were generated digitally, using a Tucker Davis Technologies (TDT) system II. The tones were played through a 16-bit digital to analogue converter (TDT DD1) at a 50-kHz sampling rate, lowpass filtered at 10 kHz (Kemo VBF8), attenuated (TDT PA4) and presented via a headphone buffer (TDT, HB6), a manual attenuator (Hatfield 2125) and one earpiece of a Sennheiser HD580 headset. Subjects were tested in a double-walled sound-attenuating booth.

5.2.3 Procedure

An adaptive three-down one-up two-alternative forced-choice procedure was used. In a given trial, the F0 of group B was shifted upwards by $0.5 * \Delta\mathrm{F}$ for one tone (the "group-B-up" tone) and downwards by $0.5 * \Delta\mathrm{F}$ for the other (the "group-B-down" tone). The initial value of ΔF was always 0.05F0, and the step size was a factor of $2^{1/2}$ before the 5th turnpoint and $2^{1/4}$ thereafter. Subjects were asked to indicate which tone had the higher pitch by pressing the appropriate button on a keypad. Subjects received feedback on their answers via flashing lights above the buttons. Twelve turnpoints were obtained for each run, and the geometric mean and standard deviation of the values of ΔF at the last 8 were recorded.

If the value of ΔF requested by the procedure exceeded 0.1F0 more than twice in a run, then that run was terminated and performance was estimated using a different method. In this method, the value of ΔF was fixed at 0.05F0 and subjects completed 50 trials, to give a percent-correct score. If this percent-correct score was significantly different from chance at the 5 % level (more than 32 trials correct out of 50) it was converted to a d' score using the table provided by Hacker and Ratcliff (1979), from which an F0DL estimate was extrapolated. The threshold value of ΔF that would be obtained for a d' value of 1.16 (the d' score given for 79.4 % correct, obtained by interpolating linearly between the values for

79 % and 80 %) was extrapolated from the d′ value associated with the measured percent-correct score, assuming that d′ is proportional to ΔF. If the score was not significantly different from chance, the F0DL estimate for that run was set to 0.1F0 (the maximum possible). This method of combining results across methods is essentially the same as reported in chapter 4 of this thesis, and similar to that used by Moore *et al.* (2009b). If both runs for a given condition yielded a percent-correct score, and both were significantly different from chance, then the scores were summed to give a score out of 100, which was then converted to a threshold as before. All subjects completed all conditions twice, and the geometric mean of the two F0DL estimates was taken for each condition.

It was noted in pilot experiments that subjects sometimes perceived the tones "upside down", that is the group-B-down tone was reliably perceived as having a higher pitch than the group-B-up tone for some conditions. These cases were identifiable by a very low score out of 50 when the non-adaptive method was used. In these cases, stricter criteria were used to determine whether the score was significantly different from chance; scores were required to be significantly different from chance at the 1 % level, corresponding to a score of fewer than 6 correct out of 50 trials. However, if on both repeats of a particular run the tones were perceived to be upside down, the 5 % level was used for significance. As before, the F0DL estimate for a score that was not significantly different from chance was set to 0.1F0. Not all subjects demonstrated the upside-down percept, and those that did usually demonstrated it when N was 7 or greater.

5.2.4 Conditions

The lowest harmonic rank included in group B, N, was 1, 3, 5, 7, 9, 11, 13 or 15, and the number of harmonics in group B, B, was either 4 or 8. Components were added in random or cosine phase. For experiment 1, the F0s were 50 Hz and 200 Hz. For experiment 2, the F0s were 100 Hz and 200 Hz, and the initial PPA between groups A and B was selected randomly from a uniform distribution with a range of half a period to remove the cue of PPA, based on the findings of Miyazono *et al.* (2010). Experiment 2 was carried out after experiment 1; a higher F0 was used for the low-F0 condition because the performance of some subjects was slightly erratic for F0 = 50 Hz in experiment 1.

5.2.5 Results and statistics

Two threshold estimates were obtained for each subject and condition, either directly or by extrapolating from the d′ value associated with the percent-correct score achieved. The geometric mean of these two results was taken.

5.2.5.1 Experiment 1

Figure 5.1 and Figure 5.2 show the results for F0 = 50 Hz and F0 = 200 Hz, respectively. Individual panels show the results for individual subjects, with error bars showing the standard error of the two threshold estimates obtained for each condition. In each figure, the bottom right-hand panel shows the geometric mean data across subjects for each condition. Error bars here show the standard error across all 6 subjects. A within-subjects ANOVA was carried out on the data, with factors F0, N, B and phase. The variate was the logarithm of the single combined threshold for each subject and condition. This transform was used as the data were negatively skewed.

A summary table showing the results of the ANOVA is given in Table 5.3. Significant effects are shown with bold type.

There was a main effect of phase; thresholds were generally lower for cosine phase than random phase. All two-way interactions with phase were significant. For F0 = 50 Hz, there was a phase effect for all values of N, whereas for F0 = 200 Hz there was only a phase effect for values of N above about 7, though this difference across F0s was not significant as shown by the insignificant interaction between F0, N and phase shown in Table 5.3 [$F(7,35)=0.45$; $p=0.855$]. The improvement in performance for cosine phase was clearest for $N>7$ and was greater for F0 = 50 Hz than F0 = 200 Hz, due to the improvement in performance with increasing N for $B = 8$ when F0 was 50 Hz. The presence of the phase effect indicates that the harmonics on which judgments were based were likely to have been unresolved.

Thresholds generally increased with increasing N for both F0s. The data for F0 = 200 Hz showed low thresholds for N up to 7, then a worsening covering

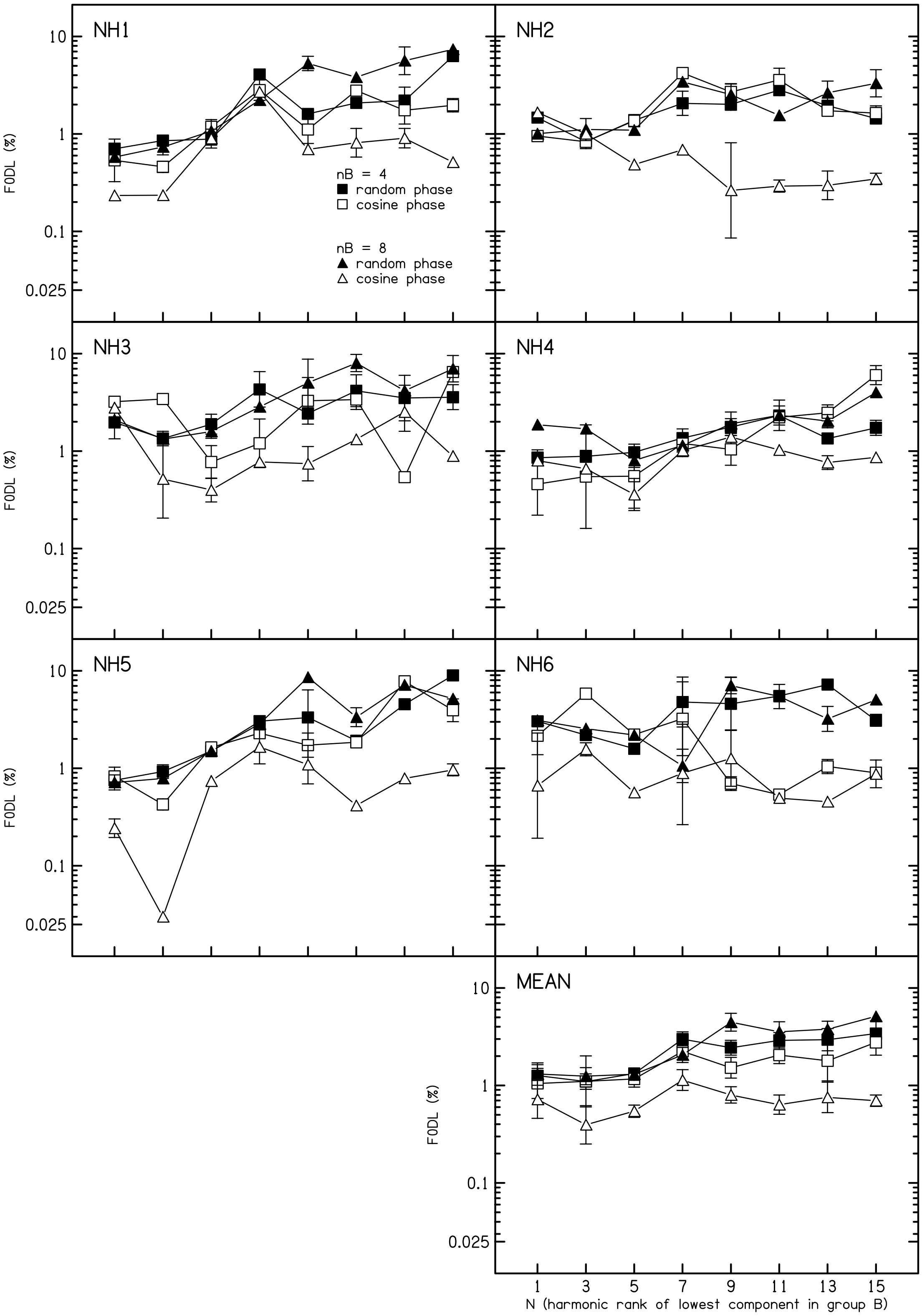


Figure 5.1: Geometric mean results for each subject for experiment 1 where F0 was 50 Hz. Error bars show the standard error of the two runs. The bottom right-hand panel shows the geometric mean and standard error across all subjects.

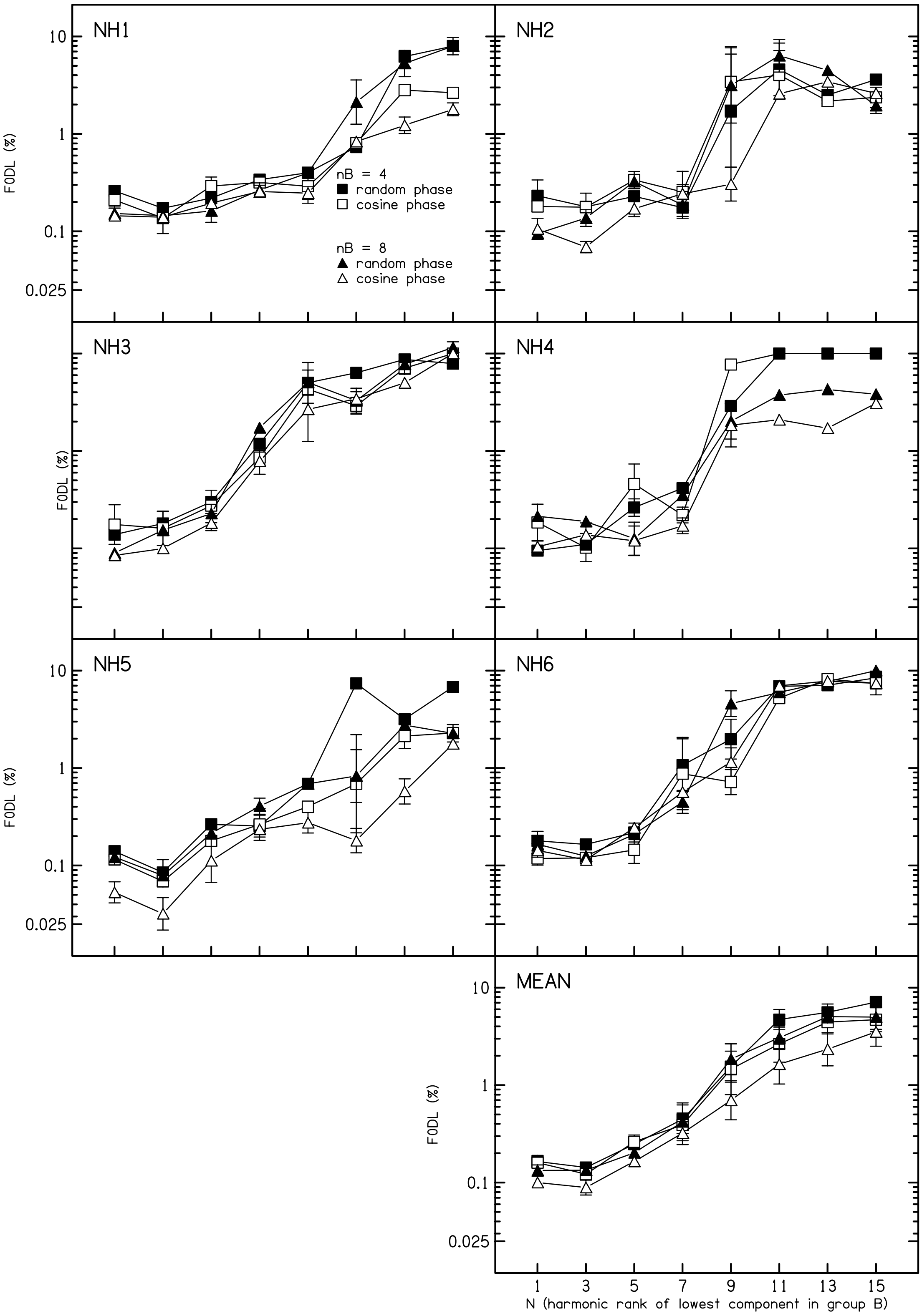


Figure 5.2: Geometric mean results for each subject for experiment 1 where F0 was 200 Hz. Error bars show the standard error of the two runs. The bottom right-hand panel shows the geometric mean and standard error across all subjects.

	Factor	Degrees of Freedom	Variance Ratio	F pr.
Main	**F0**	**1,5**	**22.07**	**0.005**
	N	**7,35**	**69.45**	**<0.001**
	B	**1,5**	**25.67**	**0.004**
	Phase	**1,5**	**41.81**	**0.001**
Two-way	**F0.N**	**7,35**	**18.18**	**<.001**
	F0.B	1,5	0.14	0.722
	N.B	7,35	0.20	0.983
	F0.Phase	1,5	24.05	0.004
	N.Phase	7,35	2.90	0.017
	B.Phase	**1,5**	**45.24**	**0.001**
Three-way	F0.N.B	7,35	0.68	0.683
	F0.N.Phase	7,35	0.88	0.881
	F0.B.Phase	**1,5**	**0.01**	**0.005**
	N.B.Phase	7,35	0.27	0.269
Four-way	F0.N.B.Phase	7,35	0.16	0.163

Table 5.3: The results of an ANOVA carried out on the data for experiment 1. Significant effects are shown with bold type. The column “F pr.” gives the p value for each effect.

nearly one order of magnitude as N was increased to 11, above which thresholds did not worsen further. For $B = 8$ and cosine phase this trend was not seen; for this condition, the worsening for N above 7 continued up to $N = 15$ with no clear plateau in the thresholds. Overall performance was best for this condition, perhaps indicating a role for PPA cues.

For F0 = 50 Hz, there was a similar pattern of results, except for $B = 8$ and cosine phase. For the other three conditions, thresholds were low for N up to 5, although these thresholds were up to an order of magnitude higher than for the corresponding conditions for F0 = 200 Hz. This difference could reflect the wider auditory filter bandwidth as a proportion of CF for low F0s, at least for low values of N, for which harmonics were likely to be resolved on the basilar membrane. It might also reflect the limited ability of the auditory system to estimate long time intervals, as discussed earlier. The overall worsening with increasing N was less marked than for F0 = 200 Hz, and a much less marked intermediate region was demonstrated.

For F0 = 50 Hz, $B = 8$ and cosine phase, thresholds were lower than for the other conditions with F0 = 50 Hz, and thresholds decreased with increasing N above 7. The data for subject NH2 showed this effect particularly clearly. The effect is presumably attributable to the use of PPA cues. Note that performance was better for the condition with $B = 8$ and cosine phase than for the other conditions for all N (although the effect for low N varied markedly across subjects), suggesting that, even for low N, there were unresolved harmonics in group B (since the use of the PPA cue requires at least some unresolved harmonics in both groups).

Thresholds were generally lower for $B = 8$ than for $B = 4$, particularly for cosine phase. This trend was particularly distinct for F0 = 50 Hz. The main effect of B is likely to be due to the fact that pitch pulses are clearer – the waveforms have a higher peak factor – at the output of an auditory filter for which multiple components interact than at the output of a filter for which fewer components interact. This might result in lower thresholds for $B = 8$ than for $B = 4$ for two reasons. Firstly, it would result in a more salient percept for pitch based on unresolved components. Secondly, it would increase the salience of the PPA cue. Data of Miyazono *et al.* (2009) collected using a similar task showed that thresholds were higher for the larger number of components in the shifted group for F0s of 35, 50 Hz, and 200 Hz, particularly for $N = 7$ and higher. In chapter 4, it was shown that thresholds worsened with increasing number of components in the stimuli when discrimination was based on TFS cues. The difference here (apart from the fixed harmonic background) is that the shifted group was harmonic, whereas in chapter 4 it was inharmonic. Adding more inharmonic information increased the conflict of TFS across auditory filters in chapter 4, but here adding more harmonic information improved performance.

5.2.5.2 Experiment 2

Figure 5.4 and Figure 5.5 show the results for F0 = 100 Hz and F0 = 200 Hz, respectively. Each panel shows the results for one subject, with error bars showing the standard error of the two threshold estimates for each condition. In each figure, the bottom right-hand panel shows the geometric mean data across subjects for each condition. Error bars here show the standard error across all 7 subjects. A within-subjects ANOVA was carried out on the data, with factors F0, N, B and phase. The variate was the logarithm of the single combined threshold for each subject and condition. This transform was used as the data were negatively skewed.

A summary table showing the results of the ANOVA is given in Table 5.6. Significant effects are shown with bold type.

Increasing N had the same detrimental effect on performance as demonstrated in experiment 1, for both F0s. Thresholds for N below 7 and 5 were low for F0 = 100 Hz and F0 = 200 Hz, respectively, showing only a slight worsening with increasing N. The thresholds for small N were lower for F0 = 200 Hz than for F0 = 100 Hz. There was a significant interaction between F0 and N, reflecting this; however, there was no significant main effect of F0, which is probably because, for high N, thresholds were similar for the two F0s.

For F0 = 200 Hz, the data show a worsening in performance of about one order of magnitude between $N = 5$ and $N = 9$, with no further worsening above this value of N. The pattern of results for all four conditions was similar, unlike for experiment 1. This is presumably because listeners were unable to use PPA

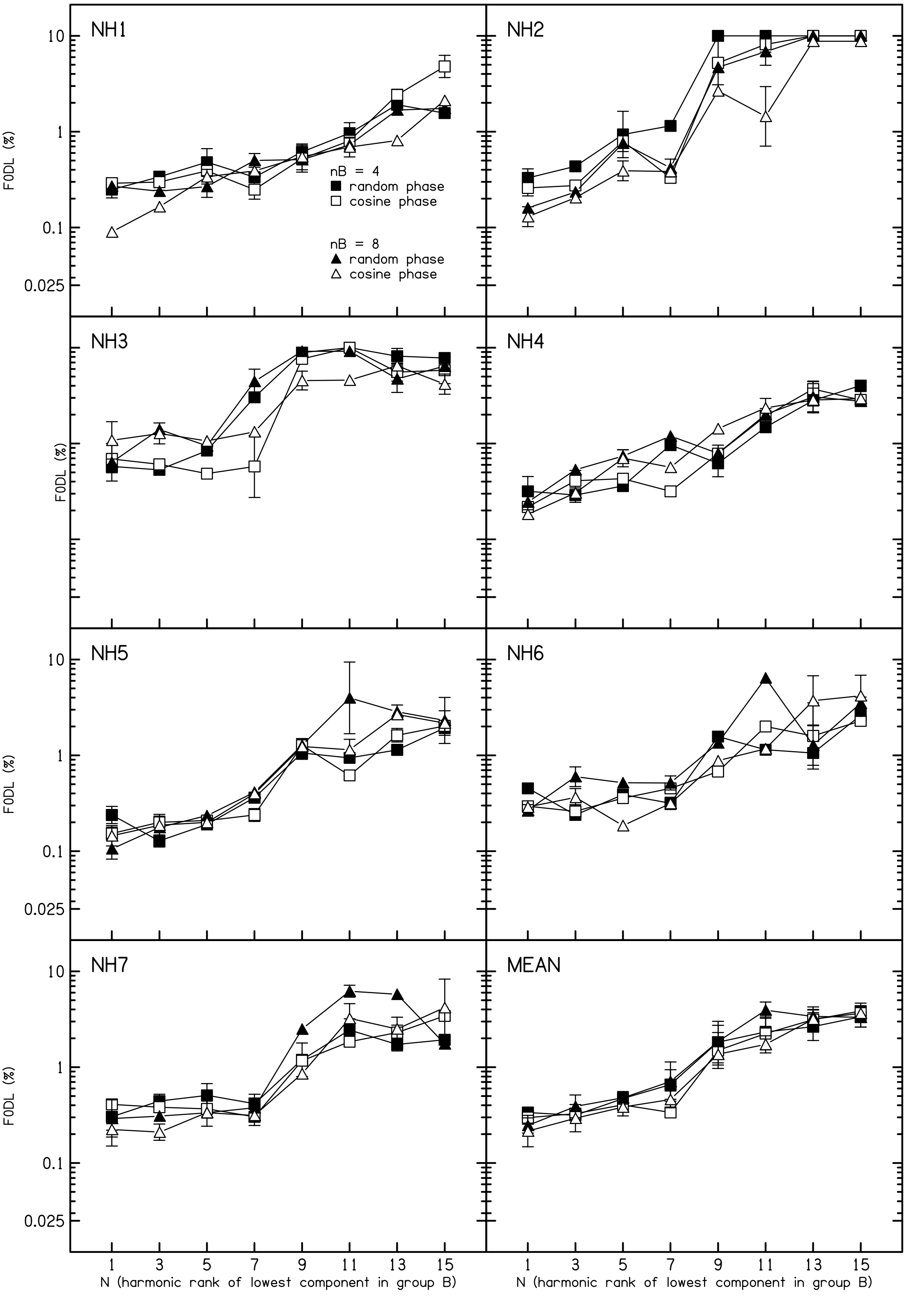


Figure 5.4: Geometric mean results for each subject for experiment 2 where F0 was 100 Hz. Error bars show the standard error of the two runs. The bottom right-hand panel shows the geometric mean and standard error across all subjects.

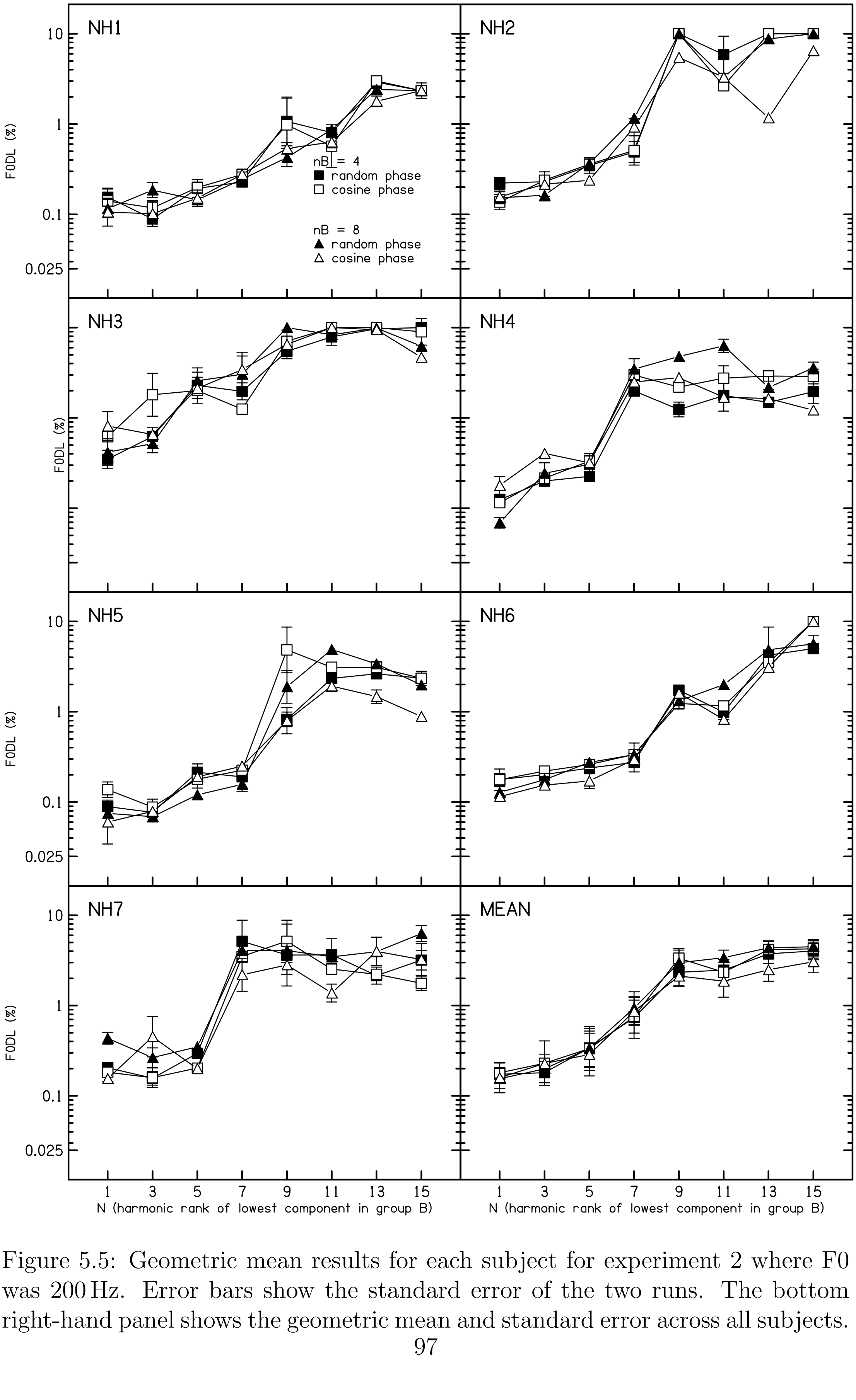


Figure 5.5: Geometric mean results for each subject for experiment 2 where F0 was 200 Hz. Error bars show the standard error of the two runs. The bottom right-hand panel shows the geometric mean and standard error across all subjects.

	Factor	Degrees of Freedom	Variance Ratio	F pr.
Main	F0	1,6	0.18	0.688
	N	**7,42**	**72.02**	**<.001**
	B	1,6	0.06	0.815
	Phase	**1,6**	**13.38**	**0.011**
Two-way	**F0.N**	**7,42**	**3.41**	**0.006**
	F0.B	1,6	0.21	0.661
	N.B	7,42	1.58	0.169
	F0.Phase	1,6	3.21	0.124
	N.Phase	7,42	2.03	0.073
	B.Phase	**1,6**	**24.45**	**0.003**
Three-way	F0.N.B	7,42	0.45	0.867
	F0.N.Phase	**7,42**	**2.62**	**0.024**
	F0.B.Phase	1,6	1.79	0.229
	N.B.Phase	7,42	1.53	0.185
Four-way	F0.N.B.Phase	7,42	0.76	0.624

Table 5.6: The results of an ANOVA carried out on the data for experiment 2. Significant effects are shown with bold type. The column “F pr.” gives the p value for each effect.

to improve performance for $B = 8$ and cosine phase.

For F0 = 100 Hz, the data show a worsening between $N = 7$ and $N = 11$, with no obvious worsening above this. The worsening is less marked than for F0 = 200 Hz, because thresholds for low values of N were higher.

In contrast to the results for experiment 1, thresholds were not lower for the condition with $B = 8$ and components added in cosine phase, except for F0 = 200 Hz and high values for N. This difference across experiments supports the idea that PPA cues were used in experiment 1 when $B = 8$ and components were added in cosine phase. The effect of component phase found in experiment 2 for high N with $B = 8$, which was confirmed by a significant interaction between F0, phase and N, presumably occurred because cosine phase led to a waveform at the output of the auditory filters with a higher peak factor than for random phase, promoting F0 discrimination based on temporal cues.

Thresholds were not significantly affected by the value of B, although there was a significant interaction between B and phase; increasing B from 4 to 8 resulted in an improvement in performance for cosine phase, as in experiment 1, and no significant change in performance for random phase.

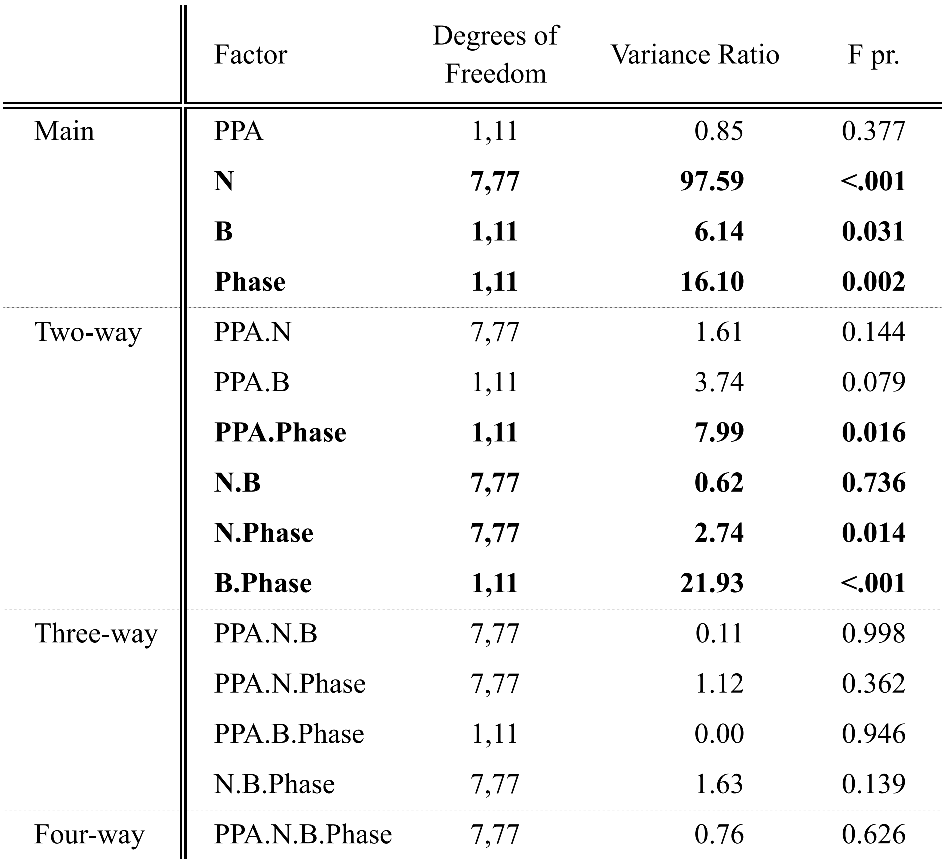
5.2.5.3 Further comparison

A further ANOVA was conducted on the data for F0 = 200 Hz from both experiments to assess the effect of removing PPA cues. PPA (present or absent) was a between-subjects factor since most subjects differed across the experiments, and all other factors were within-subjects.

The three factors were N, B, phase and PPA (present or removed), and the variate was the logarithm of the single combined threshold for each subject and condition. This transform was used as the data were negatively skewed (the combined data had a range of 0.03 to 11.58 with a mean of 2.274).

A summary table showing the results of the ANOVA is given in Table 5.7. Significant effects are shown with bold type.

There was no main effect of PPA in the ANOVA on the combined data; however, there were some significant two-way interactions with PPA. When PPA cues were present, performance was better for cosine phase, and for the higher



	Factor	Degrees of Freedom	Variance Ratio	F pr.
Main	PPA	1,11	0.85	0.377
	N	**7,77**	**97.59**	**<.001**
	B	**1,11**	**6.14**	**0.031**
	Phase	**1,11**	**16.10**	**0.002**
Two-way	PPA.N	7,77	1.61	0.144
	PPA.B	1,11	3.74	0.079
	PPA.Phase	**1,11**	**7.99**	**0.016**
	N.B	**7,77**	**0.62**	**0.736**
	N.Phase	**7,77**	**2.74**	**0.014**
	B.Phase	**1,11**	**21.93**	**<.001**
Three-way	PPA.N.B	7,77	0.11	0.998
	PPA.N.Phase	7,77	1.12	0.362
	PPA.B.Phase	1,11	0.00	0.946
	N.B.Phase	7,77	1.63	0.139
Four-way	PPA.N.B.Phase	7,77	0.76	0.626

Table 5.7: The results of an ANOVA carried out on the data for F0 = 200 Hz. Significant effects are shown with bold type..

value of B.

5.3 Discussion

Under conditions where PPA cues were not usable (random phase for experiment 1 and all conditions for experiment 2), performance worsened when N was increased above a certain value. The increase first occurred for $N = 7$ for $\mathrm{F0} = 50\,\mathrm{Hz}$ and $\mathrm{F0} = 100\,\mathrm{Hz}$ and $N = 5$ for $\mathrm{F0} = 200\,\mathrm{Hz}$. This worsening could be due to a reduction in resolution of the harmonics or to a progressive loss of the ability to use TFS information with increasing harmonic rank. Performance reached an asymptote for high N ($N =$ about 9 for $\mathrm{F0} = 50\,\mathrm{Hz}$ and $N =$ about 11 for the two higher F0s), presumably reflecting the point at which the harmonics were completely unresolved or the point at which the ability to use TFS information was completely lost. For very high N, performance was presumably based on the use of temporal envelope cues.

If the worsening in performance with increasing N were due to a reduction in resolution of the harmonics, one would expect the value of N at which the increase first occurred to increase with increasing F0, since the relative bandwidths of the auditory filters (bandwidth divided by CF) decrease with increasing CF over the range 50–500 Hz. In fact, this was not the case; the increase occurred for a lower N for $\mathrm{F0} = 200\,\mathrm{Hz}$ than for $\mathrm{F0} = 50\,\mathrm{Hz}$. For $\mathrm{F0} = 50\,\mathrm{Hz}$, one would have expected that an increase in N from 1 to 5 would result in no harmonics in group B being resolved, as described earlier, but in fact F0DLs did not increase over that range. These aspects of the results are not consistent with the idea that the worsening in performance with increasing N was due to a reduction in resolution of the harmonics.

Especially for low N, performance worsened with decreasing F0. As described earlier, this may reflect the difficulty that the auditory system has in measuring long time intervals. For $\mathrm{F0} = 50\,\mathrm{Hz}$, even if the lowest harmonics were resolved, the inter-spike intervals evoked by those harmonics would be relatively long. For example, for the second harmonic (100 Hz), the inter-spike intervals would be close to integer multiples of 10 ms. Consistent with this idea, frequency difference limens for pure tones, expressed as a proportion of baseline frequency, worsen for frequencies below about 500 Hz (Moore, 1973). It is unlikely that the higher thresholds for low N for F0 = 50 Hz were due to the inaudibility of partials at the levels used in the experiment; the fundamental component for the lowest F0 used (50 Hz) would have been close to the absolute threshold for monaural listening, as calculated using the loudness model described by Moore and Glasberg (2007), but all other components would have been at least 15 dB above the absolute threshold.

There was a main effect of phase for both experiments. For experiment 1, post hoc analyses, based on Fisher's protected least significant differences test, showed that there was a significant effect of phase for $N>1$ for F0 = 50 Hz, and for $N = 9$, 11 and 13 for F0 = 200 Hz. For experiment 2, the same analyses showed that there was a significant effect of phase for $N = 7$, 9 and 11 for F0 = 100 Hz, and only for $N = 11$ for F0 = 200 Hz. The difference in the effect of phase across experiments for F0 = 200 Hz is presumably due to the PPA cues present in the stimuli for experiment 1. PPA cues were not usable when components were added in random phase, but were usable when components were added in cosine phase and components were unresolved, which is the likely explanation for the effect of phase for high N for experiment 1. For experiment 2, where PPA cues were not present in the stimuli, there was an effect of phase for only one value of N.

It cannot be assumed that PPA can be detected for arbitrarily small time intervals, that is when the difference between the absolute timing of pitch pulses at the output of different auditory filters is very small. Therefore one might expect that the ability to use PPA cues to discriminate tones would deteriorate with an increase in F0 or absolute frequency. Such effects were demonstrated here by the higher thresholds (expressed as a percentage of F0) for $N = 11$, $N = 13$ and $N = 15$ for F0 = 200 Hz than F0 = 50 Hz in experiment 1 when B was 8 and components were added in cosine phase. For example, for F0 = 200 Hz, cosine phase, $N = 9$ and $B = 8$ in experiment 1 (PPA cues present), the mean threshold was 0.75 % (1.5 Hz). In this case, in one interval the F0 of group B was 200.75 Hz, and in the other it was 199.25 Hz. In the case where group B had the higher F0, the PPA between group A and group B would repeat every 1/0.75 seconds (1.33 seconds) if the stimuli were continuous. For the first half of this time, 650 ms, the pitch pulses evoked by group B would occur earlier in time than those evoked by group A, until after 650 ms they would occur exactly half way between those for group A. As each stimulus interval was only 500 ms in duration, for an F0 shift corresponding to threshold a whole cycle of PPA was not present in the stimuli. Consider now the case where N was 15 for the same condition, for which the mean threshold was 3.5 % (7.0 Hz). The maximum PPA between group A and group B here would have been the same ($0.5 * (1/200)$seconds $= 2.5$ ms), but it would have occurred after 141 ms, meaning that there would be several cycles of PPA in each stimulus interval. The fact that performance was worse for this latter condition, in spite of the actual asynchrony to be detected being the same, suggests that it might be easier for listeners to detect PPA if it is consistently in one direction throughout the stimulus, as in the former condition.

When F0 was 50 Hz, the ability of listeners to use PPA cues to discriminate the tones for high values of N was good, and improved with increasing harmonic rank from 7 to 15. Increasing the frequency of the lowest component in group B from 350 Hz to 750 Hz brought the frequency region covered by group B closer to the range for which frequency processing is very good (Moore, 1973). As the auditory system may have trouble measuring long time intervals accurately (de Cheveigné and Pressnitzer, 2006), it is likely that PPA cues for lower harmonic ranks were harder to use to discriminate the tones, which is why performance improved with increasing N. Another explanation could be that when the envelope of a waveform has sharp and distinct peaks, such as occurs for high N, PPA cues are easier to detect.

It is possible, as noted by Miyazono *et al.* (2010) in their report of experiments using the same stimuli, that combination tones of the form $2f_1 - f_2$ generated by the interaction of components within group B or between the groups could have affected the results. For example, take the case where F0 was 200 Hz, N was 7 and B was 8. Recall that the components $N - 1$ and $N + B$ were omitted to avoid beating cues, leaving spectral gaps. For a value of ΔF equal to the mean F0DL for this condition, about 0.4 %, the combination tone produced by components N and $N + 1$ (1400 Hz and 1600 Hz) fell at about 1198 Hz when group B was shifted down and at about 1202 Hz when it was shifted up. As the spectral gap ranged between 1000 Hz and about 1397 Hz, this shift in the frequency of the combination tone could have been audible, and could have been used to perform the task, behaving like the missing 6th harmonic. However, the level of these combination tones would have been at least 10 dB lower than that of each neighbouring component, as explained by Miyazono *et al.* based on the data of Zwicker (1981), and therefore the combination tones would have been only just audible. Hence, they would probably have had a negligible effect on F0 discrimination (Moore *et al.*, 1985).

That said, discrimination based on combination tones could provide an explanation for the upside-down percept described earlier, where some subjects reliably perceived the tone in which the F0 of group B had been shifted downwards as having the higher pitch for some conditions. For example, the combination tone of the form $2f1 - f2$ between component N (the lowest in group B) and $N - 2$ (the highest in group A) fell at about 606 Hz for the group-B-down tone and at about 594 Hz for the group-B-up tone for a shift equal to mean threshold for F0 $= 200$ Hz, $N = 7$ and $B = 8$ where PPA was absent. Even though these tones did not fall in the spectral gap between 1000 and 1394 Hz, they could have been perceived. The difference between the frequencies of these tones increases as N increases; for example, for $N = 3$ the threshold shift gives combination tones at 200.7 Hz and 199.3 Hz for the group-B-down and -up tones respectively, which is a very small difference. This could explain why the upside-down percept only occurred for $N = 7$ and higher. It is unlikely that these tones were perceived clearly due to their low levels relative to those of the physically present components in the spectral region where they fell, but it is possible that the fact that these combination tones shifted in the "wrong" direction cued attention to them in some way.

5.4 Summary

In summary, it seems likely that dominance was not governed by the resolvability of harmonics for low F0s, because performance based on harmonics 5–8 was similar to performance based on harmonics 1–4 for F0 = 50 Hz. For F0 = 200 Hz, performance for low N was good, implicating the use of resolved harmonics, then worsened progressively above $N = 7$, the limit of peripheral resolvability for this

F0 calculated in a variety of ways.

5.5 Conclusions

1. The increase in F0DLs with increasing N did not start at a lower N for F0 = 50 Hz than for F0 = 200 Hz. This is inconsistent with the idea that the increase in F0DLs with increasing N is caused by a progressive loss of resolution of harmonics.

2. PPA cues were probably used in experiment 1 for $B = 8$ and cosine phase. The lack of effect of PPA for $B = 4$ can be explained by the idea that sharp pitch pulses only occur when several cosine-phase harmonics are present within the passband of an auditory filter.

3. When PPA cues were not usable (random phase), F0DLs for F0 = 50 Hz were almost constant for N = 1-5. Thus harmonics in the ranges 1–4, 3–6, and 5–8 are roughly equally dominant. The harmonics in the range 5–8 were probably barely resolved, and certainly much less well resolved than harmonics 1–4. This suggests that resolvability is not the major factor governing dominance for low F0s.

4. Analysis of the time intervals involved suggests that a listener's ability to use PPA cues is limited by the magnitude of the asynchrony. For high F0s, even the largest possible asynchrony may be too small to be usable.

Chapter 6

The contribution of temporal fine structure information and fundamental frequency separation to speech intelligibility in a competing-speaker paradigm

6.1 Introduction

6.1.1 Grouping and fundamental frequency

When trying to follow one speaker while another is talking at the same time, a listener must solve two problems: firstly, those portions of the speech that occur simultaneously must be separated, with part of what is heard allocated to each source; secondly, such allocations must be joined together in time correctly, so the target speaker is perceived as a continuous stream. The first of these problems is one of simultaneous grouping, and the second is one of sequential grouping. When discussing cues for grouping, it is useful to distinguish between low-level cues, such as a common harmonicity or onset time, and higher-level cues, such as known properties of speech sounds (*e.g.* vowel spectral shapes), listening situations, and sentence context effects.

One robust cue for grouping is the fundamental frequency (F0) of a speaker. The F0 of a voice arises from modulation of the airflow introduced by the vocal folds in the larynx during voiced portions of speech. For normal speech, variations in F0 give rise to prosody, and can even cue meaning in tone languages such as Mandarin Chinese. Differences in the F0 between two speakers can provide a basis for separating the speech signal across both frequency (simultaneous grouping) and time (sequential grouping).

6.1.2 Simultaneous grouping

6.1.2.1 Beating

It is well established that a difference in F0 between two speakers improves intelligibility for short sounds, such as simultaneous vowels (Scheffers, 1983; Assmann and Summerfield, 1994; Rossi-Katz and Arehart, 2005). Vowel sounds are characterised by their formants, which are peaks in their spectra at particular frequencies. A change in the ratio of the frequencies of the first two formants can result in a change in perception. In a study of Scheffers (1983), pairs of 220-ms vowels were played simultaneously to listeners, and correct identification of both vowels was found to improve with increasing F0 separation up to one semitone. Performance in the same kind of task has been shown to improve even when the F0 separation is only half a semitone (Culling and Darwin, 1994).

Culling and Darwin (1994) showed that the improvement in performance for separations up to half a semitone was due to beating between components that were close together in frequency. As a result of this beating, the level of adjacent pairs of components was sometimes dominated by one component in the pair, corresponding to one speaker, and sometimes to the other, corresponding to the other speaker. In this way, listeners in the task might have identified one vowel or the other by focusing on different points in time during the signal, allowing performance to improve via “glimpses” of the acoustic features of the target speaker when the amplitude of the background speaker was momentarily low. Assman and Summerfield showed that identification of concurrent vowels improved when listeners were played four different 50-ms extracts from the signal in rapid succession compared with when they were played four identical 50-ms extracts from the signal. This finding supports the hypothesis of Culling and Darwin (1994) that listeners analyse the signal over time and are able to focus on particular temporal regions where each vowel is the most clearly represented.

For shorter (50 ms) pairs of simultaneous vowels (Culling and Darwin, 1994; Assmann and Summerfield, 1994), and for whole sentences (Brokx and Nooteboom, 1982; Bird and Darwin, 1997), this beating cue is less likely to be available, and performance improves more gradually up to a larger F0 separation. For example, Brokx and Nooteboom (1982) measured intelligibility for a monotone target sentence against a monotone competing-speaker background, and found that performance improved gradually up to a F0 separation of three semitones. They included a condition for which the F0 separation was 12 semitones (one octave), and found that intelligibility was worse than that for three semitones, presumably because even harmonics of the lower-frequency voice coincided with harmonics of the higher-frequency voice. Bird and Darwin (1997) replicated and extended this, finding that intelligibility improved for separations up to four semitones (and again between six and eight semitones), and the improvement was approximately the same when the masker was higher or lower in frequency than the target. These findings can be contrasted with the effect of F0 separation on concurrent vowel identification, for which an improvement in performance is generally not seen above two semitones (Assmann and Summerfield, 1994; Culling and Darwin, 1994; Rossi-Katz and Arehart, 2005).

Overall, the contribution of F0 differences to perceptual separation is greater for concurrent sentences than concurrent vowels. For concurrent vowels, most of the improvement in identification with increasing F0 separation is due to beating cues, which are not usable for concurrent sentences.

6.1.2.2 Across-formant grouping

Successful separation of simultaneous syllables of the target and background speech is crucial to intelligibility when concurrent sentences are heard. Other than beating, one way in which an F0 separation between speakers could improve the intelligibility of concurrent portions of speech in a sentence is via across-formant grouping. When comparing the spectra of the two speakers, the first two formants in a vowel sound from one speaker might dominate in one frequency region or set of frequency regions, while those of a vowel sound from the other speaker might dominate in different frequency regions. Within each of these regions, the representation of the signal in the cochlea will reflect the harmonic and temporal structure of the dominant speaker, and thus the target vowel will be easier to recognise than if the first two formants of the two vowel sounds occupied the same frequency regions. Broadbent and Ladefoged (1957) proposed that a common harmonic structure across different non-contiguous frequency regions might facilitate perceptual fusion by the auditory system, allowing the intact representation of a vowel sound in spite of non-useful information being present in the intervening frequency regions. A difference in F0 between competing speakers would result in a greater difference in harmonic and temporal structure between frequency regions dominated by one speaker and those by the other, resulting in more effective perceptual separation.

Across-formant grouping has been demonstrated by Darwin (1981). He used stimuli that were heard as a different phoneme depending on whether or not the second formant in a four-formant syllable had the same F0 as the other formants. Listeners heard /ru/ when the whole signal was presented, but heard /li/ when the second formant was removed or when it had a different F0 from the other formants. The /li/ percept was easier to achieve when the second formant contained resolved harmonics than when it contained only unresolved harmonics (Darwin, 1992).

Culling and Darwin (1993) found that, for concurrent vowel identification, across-formant grouping was mainly based on the first formant for small F0 separations of about 2–4 semitones, whereas larger F0 separations were required before higher formants contributed to identification performance. This is similar to findings of Rossi-Katz and Arehart (2005), who demonstrated that both normal-hearing and hearing-impaired listeners were able to use a small difference in F0 between two speakers to identify concurrent vowels when the F0 difference was represented in both low and high frequency regions, covering the first two formants of the vowels. The exact frequency regions varied depending on the average values for the first two formants in the range of vowel sounds that were tested, and the F0 separation was varied between zero and nine semitones. Rossi-Katz and Arehart also found that hearing-impaired listeners were not able to do this when the F0 difference was represented in high frequency regions alone, although normal-hearing listeners could. The hearing-impaired listeners also benefitted less than the normal-hearing listeners from increasing the F0 difference from zero to one semitone, whether the F0 difference was represented in the low frequency region, the high frequency region or both. These results suggest that across-formant grouping is a stronger cue for perceptual separation of vowel sounds when low frequency information is available in the signal. Furthermore, the results suggest that the hearing-impaired listeners were less sensitive to high frequency information that might have been used to separate concurrent signals, although they were able to group information across formant frequency regions to achieve good performance when both low- and high-frequency information were present. It is worth noting that harmonicity information is clearer for low frequency regions, as auditory filter bandwidths are narrow enough that information is not “smeared”.

The findings of Darwin (1981, 1992) and Rossi-Katz and Arehart (2005) are both consistent with data of Carlyon (1994) showing that listeners were more likely to perceive a signal made up of tones with different F0s as originating from two separate sources when both complexes contained resolved components than when they contained unresolved components, and with studies demonstrating the greater salience of pitch resulting from resolved than from unresolved groups of harmonics (Plomp, 1964; Ritsma, 1967; Houtsma and Smurzynski, 1990; Moore *et al.*, 2006).

6.1.2.3 Spatial grouping

Darwin and Hukin (2004) later carried out an experiment to identify the contribution of a common F0 to spatial grouping. They found that a common F0 was not sufficient for subjects to perceive a speech signal as originating from one source when low- and high-pass filtered versions of the signal were presented to different ears. Subjects were more likely to perceive fusion when some frequency components were common between the two ears. This suggests that F0 is not a strong cue for spatial grouping.

Interestingly, it has been shown that listeners are unable to separate simultaneous frequency components into appropriate formants of vowel sounds on the basis of ITD (inter-aural time difference) alone (Culling and Summerfield, 1995), so spatial grouping cues apply at the (high) sentence level rather than at the (low) individual word level, and are useful for sequential grouping rather than simultaneous grouping.

6.1.3 Sequential grouping

Although it is unlikely that sequential streaming of competing speakers is achieved on the basis of tracking the F0 of each speaker over time (Darwin and Hukin, 1999), it is still important to ascertain what contribution F0 differences can make to perceptual fusion of elements within a given stream or to fission of one stream from another.

Vliegen and Oxenham (1999) presented sequences of complex tone triplets (ABA ABA ABA) lasting 10 seconds, where the F0 of tone A was always 100 Hz and the F0 of tone B was varied across trials to be between one and 11 semitones higher. The tones were bandpass filtered into different frequency regions to vary the resolvability of the components. Listeners were asked to try to perceive the tones as belonging to different streams, and to indicate after each sequence presentation whether they had successfully done this. Successful perceptual separation increased with increasing F0 difference, and, importantly, performance was similar when the components in the tone sequences were resolved and when they were unresolved. This suggests that perceptual separation based on temporal (periodicity) cues about harmonicity is as strong as perceptual separation based on spectral cues about harmonicity. This result contrasts with what might be expected given the relative salience of the pitch of complex tones with resolved and unresolved components, and with the worsening in perceptual separation of simultaneous portions of speech signals with decreasing resolvability.

Vliegen, Moore, and Oxenham (1999) found a different result using a different task in which integration rather than separation led to the best performance. They used tone triplets (again with the form ABA) to measure temporal discrimination. Listeners were asked to indicate in which of two intervals the "B" tone was slightly delayed relative to the midpoint of the two "A" tones, the underlying assumption being that listeners would find the task harder when perceptual separation of the A and B tones occurred. Therefore, in contrast to the task of Vliegen and Oxenham (1999), good performance in this task required integration rather than separation. Tones A and B were either: (1) pure tones; (2) complex tones with differing F0s filtered to contain only harmonics above the tenth; or (3) complex tones with the same F0 but containing different harmonic ranks. Therefore tones A and B differed in excitation pattern and temporal pattern for types (1) and (3), and but differed only in periodicity for type (2). The difference between the tones – the difference in the frequency of the pure tones for type (1), the F0 of the complex tones for type (2) and the centre frequencies of the spectra of the complex tones for type (3) – was varied between one and 18 semitones. It was found that performance worsened (thresholds increased) as the frequency separation increased for all three types of tone but the effect was greater for types (1) and (3) than for type (2). Vliegen *et al.* interpreted this as demonstrating that, although periodicity (temporal) cues allowed sequential separation, spectral cues were much stronger, concluding that perceptual separation on the basis of periodicity alone is not "automatic and obligatory".

The difference between these two results is likely to be related to the nature of the tasks. When separation was encouraged, as in the task of Vliegen and Oxenham (1999), spectral and periodicity cues were equally strong; whereas, when integration was advantageous, as in the task of Vliegen *et al.* (1999) periodicity cues resulted in a smaller involuntary segregation effect. If periodicity cues are somehow altogether less salient than spectral cues, changes in periodicity might be easier to ignore, resulting in better performance when integration is required. In contrast, when separation is required, listeners might focus on all available changes in order to perform the task.

An experiment of Grimault *et al.* (2000) supports this interpretation. Subjects were again presented with sequences of complex tone triplets (ABA), but were asked to report whether they heard one or two streams rather than to try to hear two streams: this was a more neutral instruction than used by Vliegen and Oxenham (1999). The F0 of tone A was fixed at 88 or 250 Hz, and the F0 of tone B was varied between 88 and 352 Hz, and again the tones were bandpass filtered into different frequency regions. In contrast to Vliegen and Oxenham (1999), but consistent with the findings of Vliegen *et al.* (1999), Grimault *et al.* found less perceptual segregation for sequences where all tones contained unresolved components than for sequences where the tones contained resolved components. As for the findings of Vliegen *et al.* (1999), some segregation was reported for sequences where tones contained unresolved components and differed only in F0, showing that periodicity changes can cue segregation to some extent.

6.1.4 Resolvability and temporal fine structure

It seems that both spectral and periodicity changes between successive stimuli, as well as whether integration or segregation of streams is advantageous, can affect sequential streaming. As mentioned earlier, the presence or absence of spectral cues, that is, whether resolved harmonics are present or not present, also has an effect on the efficacy of across-formant grouping for the separation of concurrent speech signals, as well as on whether or not two simultaneous complex tones are perceived as two separate sources (Carlyon and Shackleton, 1994). Therefore, it is important to define the frequency selectivity possible within the auditory system. There has been much discussion in the pitch perception literature as to the extent of peripheral resolvability. An overview of this debate was given in chapter 2 of this thesis. To summarise briefly, it seems likely that, for intermediate F0s, no harmonics above the seventh or eighth are resolved in the cochlea. For complex tones containing components with ranks below this, pitch perception is good, and is likely to be based mainly on the resolved component(s). There is evidence that tones in which the lowest physically present harmonic has a rank between 8 and 13 can be discriminated on the basis of their temporal fine structure (TFS) alone, and that pitch perception remains good for these harmonic ranks, only worsening above this, where the usefulness of TFS cues is degraded (Moore *et al.*, 2006; Moore and Sek, 2009a)

In the tasks of Vliegen and Oxenham (1999) and Vliegen *et al.* (1999), the complex tones used in the condition where harmonics were considered to be unresolved contained harmonics above the 10$^{\text{th}}$, but it is possible that listeners in this task were using TFS cues from the interaction between the lowest harmonics in these tones to compare successive elements of the two streams. While harmonics with ranks 11 to 13 are certainly unresolved in the cochlea, TFS information may still be extracted from the signal if the absolute frequencies are not too high. Temporal fine structure information may affect either simultaneous or sequential grouping (although only the latter is relevant to the experiments mentioned earlier in this paragraph). As mentioned before, understanding one speaker in the presence of another competing speaker is a problem of both simultaneous and sequential streaming, so the role of TFS information in such competing speaker situations should be assessed quantitatively. As yet, the role for TFS in speech intelligibility remains unclear.

6.1.5 Envelope and temporal fine structure cues for speech intelligibility

Speech is a complex broadband signal. The response to a speech signal at the output of one auditory filter can be considered as the product of a slow-varying envelope and a fast-varying carrier, the TFS. The range of frequencies passed by the filter, and hence the complexity of the waveform at its output, will depend on the filter's bandwidth, which depends on its centre frequency (CF).

Many studies have attempted to identify the relative importance of the envelope and TFS for speech intelligibility for normal-hearing subjects. Some approaches to this question have involved attempts to manipulate the relative amount of envelope and TFS information in a signal. One method of separating the TFS and envelope information in a signal is to filter the signal into a number of channels with a range of CFs, and to calculate the Hilbert transform for each channel to obtain the envelope and TFS for each channel separately.

6.1.5.1 Envelope cues

To measure intelligibility resulting from envelope cues alone, "vocoder" processing can be used (Dudley, 1939; Shannon *et al.*, 1995). The signal is first filtered into a number of channels as described above. Then, the envelope is extracted for each channel and is used to modulate a sinewave carrier with a frequency equal to the CF of the channel (tone vocoder) or narrow noise band carrier with the same CF as the channel (noise vocoder). The signals in each channel are then added back together. The resulting signal contains none of the original TFS. Vocoders are often used to simulate the information provided by a cochlear implant. When a low number of channels is used, little spectral information is available in the signal, but good speech intelligibility can still be achieved for speech in quiet (Shannon *et al.*, 1995; Loizou *et al.*, 1999). Increasing the number of channels results in an improvement in intelligibility for speech in quiet and in noise, with the number of channels required for asymptotic performance varying with the material and training. It is generally accepted that using 32 channels equally spaced on an equivalent rectangular bandwidth number (ERB$_{\mathrm{N}}$-number) scale allows a good approximation of the spectral information represented in the healthy cochlea.

For normally hearing listeners, the intelligibility of vocoded speech is impaired somewhat by adding a steady masker, such as broadband noise, to the signal (Xu *et al.*, 2005), by an amount depending on the signal to noise ratio (SNR). However, intelligibility becomes very poor when instead a fluctuating masker, such as amplitude-modulated noise or a competing speaker, is added to the signal at the same average SNR (Hopkins, Moore, and Stone, 2008). This finding contrasts with the masking release usually seen when a steady masker is replaced with a fluctuating one (Peters, Moore, and Baer, 1998), and suggests that including only envelope information in the signal impairs a listener's ability to "glimpse" the signal "in the dips" of the masker, or impairs the ability to segregate the signal from the background (Hopkins *et al.*, 2008).

6.1.5.2 Temporal fine structure cues

To measure speech intelligibility mediated by TFS cues alone, the TFS extracted from each channel, as described above, can be added together across all channels (Smith *et al.*, 2002; Lorenzi *et al.*, 2006). The signal in each channel has a constant envelope, but has the same TFS as the original signal. The signal formed by combining the TFS across channels is referred to as "TFS speech". One problem with this method is that TFS and envelope information in speech are correlated; therefore, some of the "missing" envelope of TFS speech may be recovered from its TFS via cochlear filtering, particularly if a small number of channels is used in the processing. Gilbert and Lorenzi (2006) used TFS speech as input to a noise vocoder to assess the extent to which recovered envelope cues could be used. In a consonant-identification task, they found that, when a small number of channels was used, listeners could recover sufficient envelope information to achieve good performance. However, when eight or more channels were used, subjects could not identify the consonants at a level above chance. Lorenzi *et al.* (2006) found that, after extensive training, subjects could still identify consonants in TFS speech at a level that was above chance when a large number of channels was used. They concluded that TFS cues alone can convey some useful information for speech intelligibility, perhaps via glimpsing in the dips of any background noise.

Hopkins *et al.* (2008) devised a method to assess the contribution of TFS information to speech indirectly, avoiding the problem of recovered envelope cues encountered in direct measurements. TFS information was progressively removed from a signal by vocoding only some channels, and intelligibility was compared with that for the intact signal. The assumption was that the difference in intelligibility between the intact signal and the signal from which TFS information had been removed would provide a measure of the contribution of TFS. They used a tone vocoder where only some channels of the signal were vocoded. The signal was split into 32 1-ERB$_\mathrm{N}$ wide channels, and channel numbers above a "cut off" (*CO*) number were vocoded with all others left intact. Then, the signals in each channel were added back together. Listeners were asked to repeat a target sentence presented against a background of a competing speaker. Speech Reception Thresholds (SRTs) were measured for each condition: the threshold Signal-to-Background Ratio (SBR) for 50 % of key words to be identified correctly was measured using a one-down one-up adaptive procedure. Varying the *CO* number allowed Hopkins *et al.* to measure the benefit to intelligibility of increasing the number of TFS-containing channels in the signal. Hearing-impaired listeners overall benefited less than normal-hearing listeners from the addition of TFS to the signal, although the benefit for individual hearing-impaired listeners varied. This result, and similar results from other studies, for example by Lorenzi *et al.* (2006), suggests that hearing-impaired listeners are less able to use TFS information in speech to separate two speakers, which could be an important finding for the future design of hearing aid algorithms (Moore, 2008b).

6.1.6 Rationale for this set of experiments

This chapter describes two experiments involving normal-hearing listeners using the same paradigm as Hopkins *et al.* (2008), but modified to introduce a systematic F0 difference between the target and interfering speakers. The main aim was to assess any relationship between the contributions of F0 differences and TFS to intelligibility in a competing-speaker paradigm, given that it has been demonstrated that each cue alone does contribute to performance in such a task. It was expected that, if the benefit of F0 separation depends partly on temporal information derived from unresolved harmonics, then adding more TFS information should increase the benefit of F0 separation between the two speakers.

Listeners heard two simultaneous sentences, one “target” sentence and one distracting “background” sentence, and were asked to report the target sentence. Experiment 1 manipulated the number of channels containing TFS and the mean F0 of each sentence independently, preserving the original F0 contour of each speaker. A range of F0 separations up to 4 semitones was used. Experiment 2, carried out after experiment 1, used the same procedure and vocoder setup, but the F0 contour of both speakers was flattened, and F0 separations up to 8 semitones were used. The flattening manipulation was used in experiment 2 to enhance the effect of F0 separation; for experiment 1, the effect of F0 separation was found to be negligible.

6.2 Method

6.2.1 Subjects

All subjects had absolute thresholds of 20 dB HL or below for both ears at all audiometric frequencies. Fourteen subjects (seven males), aged between 21 and 28 years, took part in experiment 1, and 11 subjects (five males), aged between 22 and 28 years took part in experiment 2. Nine of the subjects took part in both experiments. All subjects were British, monolingual, had never lived outside the UK, had no strong non-English accent and had never heard the test speech material before.

6.2.2 Stimuli and Equipment

The stimuli consisted of a target speech sentence in the presence of a single competing speaker background. The speech material for training phase 1 was taken from the MRC Institute of Hearing Research Adaptive Sentence Lists (ASL) developed by MacLeod and Summerfield (1990). These sentences each contained three key words. The material for training phase 2 and for both experiments was taken from the Institute of Electrical and Electronic Engineers (IEEE) low-context sentences (Rothauser *et al.*, 1969). These sentences each contained five key words. The background speech was a randomly selected section of a passage of continuous prose with a fluctuating F0, from which pauses between sentences had been removed. The background speech began 500 ms before the target, and finished about 500 ms after the end of the target; the exact duration depended on the length of the target sentence in question. Both the target and background speakers were male speakers of southern British English.

The target speech was always presented at a level of 60 dB SPL, and the level of the background speech was varied to achieve the different Signal to Background Ratios (SBRs). The combined level of the speech and background was never allowed to exceed 81 dB SPL; hence if the SBR required was lower than -21 dB, the level of the target speech was reduced instead of the level of the background speech being increased.

In experiment 1, the mean F0 of the background speech was manipulated using STRAIGHT (Kawahara and Irino, 2004). Signal processing using STRAIGHT results in very natural-sounding speech. The target speech for experiment 1 was also analysed and resynthesised using STRAIGHT, but its F0 was not manipulated. In experiment 2, the target and background speech were individually processed using STRAIGHT so that each had a constant F0.

To vary the amount of original TFS information in the stimuli, the method of processing of Hopkins *et al.* (2008) was used. The STRAIGHT-processed target and background speech were added together at the appropriate SBR and were then filtered into 32 channels with CFs between 100 and 10000 Hz. Channel edges were regularly spaced on the $\mathrm{ERB_N}$-number scale, and each channel was 1 $\mathrm{ERB_N}$ wide. Channels up to and including a cut-off (CO) channel were left unaltered, and all channels above the *CO* channel were vocoded using a tone vocoder. The envelope of each channel above the *CO* channel was extracted using half-wave rectification and was used to modulate a sine wave carrier with a frequency equal to the CF of the channel. Each resulting signal was then re-filtered to remove sidebands introduced by the processing, so envelope fluctuations with frequencies greater than half the channel bandwidth were attenuated. Lastly, the signals from the vocoded and non-vocoded channels were then added together.

All stimuli were played using a PC and a high-quality external sound card (M-Audio Audiophile USB) with a sampling rate of 22050 Hz and 16-bit resolution, and were passed through a Mackie 1202-VLZ mixing desk. The output of the mixing desk was used to drive one earpiece of a Sennheiser HD580 headset. The subject and experimenter could communicate via microphones that were also routed via the mixing desk.

6.2.3 Conditions

In Experiment 1, the target IEEE sentences had a mean F0 of 96 Hz, and the background speech was processed to have a mean F0 that was 0, 2 or 4 semitones higher than this (96, 108 or 121 Hz) ($F0sep$ = 0, 2 or 4).

In Experiment 2, the F0 contours of both the target and the background speech were flattened. The target speech had a constant F0 of 96 Hz, and the background speech had a constant F0 that was 0, 2, 4 or 8 semitones higher than this (96, 108, 121 or 153 Hz) ($F0sep$ = 0, 2, 4 or 8).

The value of *CO* was 0, 8, 16, 24 or 32 for experiment 1, and 0, 16 or 32 for experiment 2.

6.2.4 Procedure

The procedure described here applied to the second phase of training and to the main experiments. Conditions were presented as blocks, and each block contained 20 sentences. Blocks were presented in a counterbalanced order to reduce any effects of training beyond the two training phases. After presentation of a sentence, subjects were asked to repeat back the sentence. The number of correctly reported key words out of five was recorded. SRTs were measured using an up-down procedure, whereby if a subject correctly identified more than half the key words in each sentence then the level of the background was increased (SBR went down) , and otherwise the level of the background was decreased (SBR went up). Before the second turnpoint, the step size was 4 dB, and after this the step size was 2 dB. For each condition, the first sentence was presented at a very low SBR (-10 dB for CO channels of 0 and 8, -14 dB for CO channels of 16 and 24 and -18 dB for a CO channel of 32), and was then repeated using progressively higher SBRs until the subject correctly identified three or more of the five key words. This was done to ensure that the subject did not go through a large number of sentences from each block of 20 to reach the approximate level of his/her threshold, as might have occurred had a high initial SBR been used. For subsequent trials, each sentence was presented once, at the SBR defined by the subject's responses in the adaptive procedure up until that point. Subjects were informed how many key words they had identified correctly for each trial via a computer screen.

6.2.5 Training

All subjects who took part in experiment 1 underwent training phases 1 and 2. All subjects who took part in experiment 2 underwent training phase 3. Subjects who took part in experiment 2 who had not taken part in experiment 1 underwent training phases 1 and 2 before training phase 3.

The first phase of training was intended to familiarise subjects with listening to target speech in the presence of a competing speaker at different F0 separations, and consisted of four blocks of fifteen sentences taken from the ASL material. All target speech was F0-processed using STRAIGHT to have a mean F0 of 96 Hz and was added to sections of the same background speech as was used in the main experiment, which had also been processed using STRAIGHT to have a mean F0 that was 4 (block 1), 2 (block 2) or 0 semitones (blocks 3 and 4) higher than that for the target sentences. The first three blocks of sentences were not processed further, and the fourth was tone-vocoded with $CO = 16$. Sentences were presented at a fixed SBR, which was $+4$ dB for the first block. The SBRs for the subsequent blocks were made progressively lower, until subjects could identify about 50 % of the words in the fourth block. This was intended to familiarise subjects with listening to sentences at SBRs close to their SRT. The final block was included to familiarise subjects with listening to partially tone-vocoded material as part of a non-adaptive procedure. When a subject did not correctly identify all three of the key words in the target sentence, the experimenter read the sentence aloud, and then the sentence was played to the subject again before the next sentence was presented.

The second phase of training was intended to familiarise subjects with the adaptive procedure, and to expose them to the range of the conditions used in the experiment. The processing of the speech material was the same as for experiment 1, and the procedure used was the same as described above for both experiments. However, if a subject did not correctly identify all five of the key words in an IEEE sentence, the experimenter read the sentence aloud, and then the sentence was played to the subject again before the next sentence was presented. Four blocks of twenty sentences were presented in this way. The fifth block was exactly the same as a block of the main experiments: no sentences were repeated aloud by the experimenter. This block was intended to familiarise the subjects with the adaptive procedure without feedback. For the first and second blocks, the mean F0 of the background was 4 semitones higher than that of the target speech. For the third block, the mean F0 of the background speech was 2 semitones higher than that of the target speech, and for the fourth and fifth blocks it was the same as for the target speech. The *CO* channel was 16 for the first and fourth blocks and 0 for the second, third and fifth blocks. These conditions were selected to make the task progressively harder through the second training phase.

The third phase of training was only given for experiment 2. This phase was intended to familiarise subjects with the flattened F0 contours. The processing of the speech material was the same as for experiment 2, and the procedure used was the same as described above. The F0 separation was 8 semitones for the first block, 0 semitones for the second, and 4 semitones for the third. The value of *CO* was 24 for the first block, 32 for the second, and 0 for the third. These conditions were selected to familiarise subjects with the range of conditions in experiment 2, and to make the task progressively harder through this training phase. Feedback was given for the first two blocks, but not for the third.

6.2.6 Analysis

For each subject and condition, the total number of key words correctly identified was calculated for each SBR, excluding those results corresponding to the first sentence of each block. Probit analysis (Finney, 1971) was used to estimate the 50 %-correct point on the psychometric function (the SRT) for that condition. In a small number of instances (20 cases in experiment 1 from a total of 225, and three cases in experiment 2 from a total of 144), the slope of the estimated psychometric function was not significantly different from zero; this is because performance was good even for very low SBRs. In these cases, the result for the very first sentence presented in that condition (always having a level below threshold) was included in the summary for the probit analysis, which allowed the psychometric function to be estimated more satisfactorily and produced a slope that was significantly different from zero in each case.

6.3 Results

6.3.1 Experiment 1

The results for experiment 1 are given in Figure 6.1. The bottom right-hand panel shows the arithmetic mean SRT across all subjects, and error bars show the standard errors. The other panels show the results for individual normal-hearing (NH) subjects. A within-subjects ANOVA with factors $F0sep$ and CO was conducted on the data, using the SRT for each subject and condition as the variate.

SRTs generally decreased (performance improved) with increasing CO, and this effect was significant [$F(4{,}52)=22.07$; $p<0.001$]. This result is consistent with the findings of Hopkins *et al.* (2008), namely that performance improves when a higher number of channels contain TFS information. However, the data did not show a continuous trend between $CO = 0$ and $CO = 32$ for each F0 separation. In particular, listeners seemed to perform abnormally well for the condition for $F0sep = 0$ and $CO = 24$. The mean SRT for this condition was -11.0, whereas the mean SRT for $F0sep = 0$ and $CO = 32$ was -9.0. It was expected that the SRT should decrease monotonically with increasing CO. This point will be

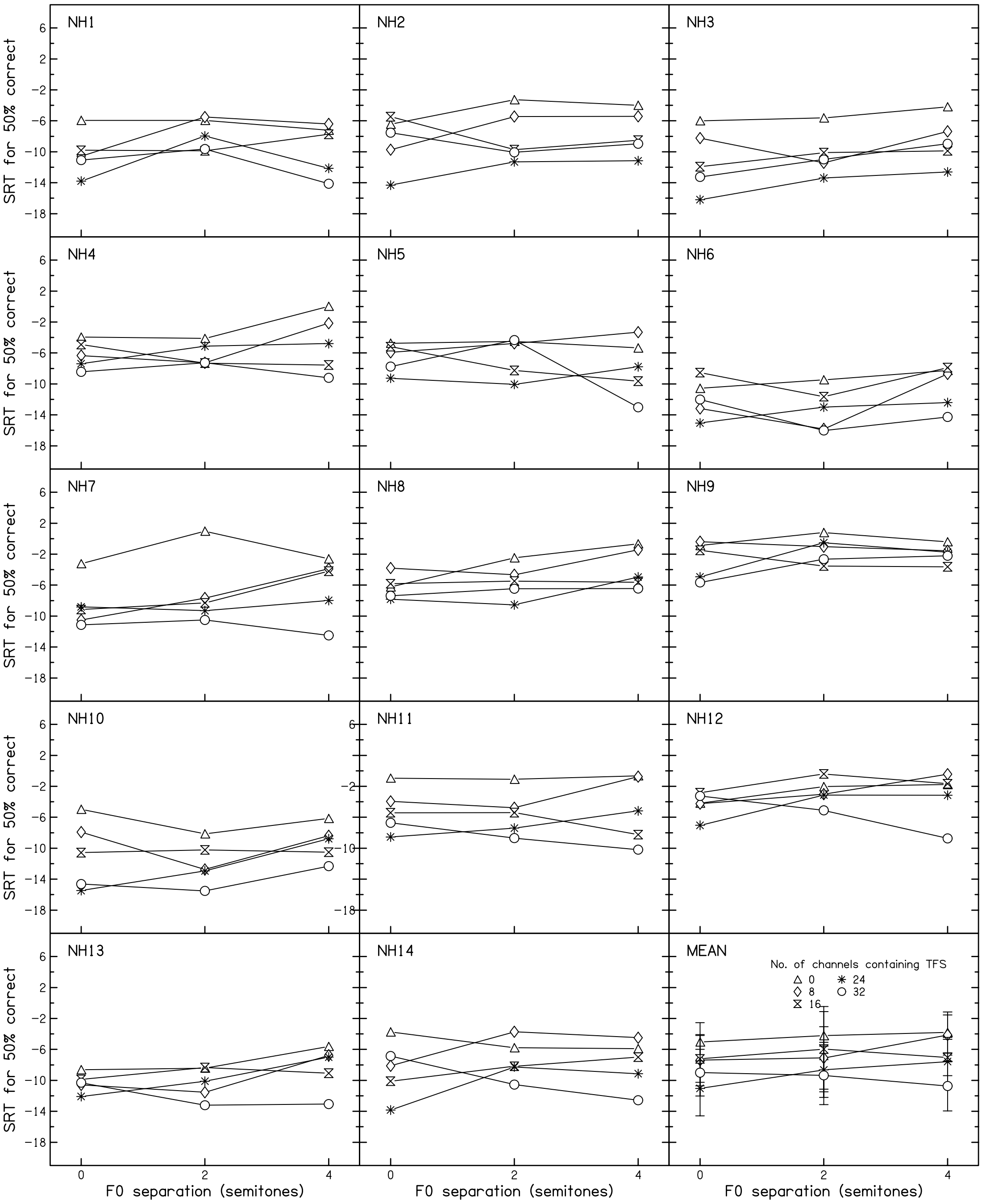


Figure 6.1: SRTs for each subject and condition in experiment 1. The bottom right-hand panel shows the arithmetic mean and standard error across subjects.

returned to later in the chapter, where an explanation and possible correction for the effect will be offered.

Rather counter-intuitively, SRTs increased (performance became worse) with increasing $F0sep$, and this effect was significant [F(2,26)=7.26; p=0.003]. Post hoc analyses, based on Fisher's protected least significant differences test, showed that the worsening in performance between the separations of zero and two semitones was significant, but that the worsening between the separations of two semitones and four semitones was not. The data for CO = 32 (intact signal) actually showed a slight improvement in SRT with increasing $F0sep$.

In spite of the overall counterintuitive effect of F0 separation, listeners gained a greater benefit of having TFS in all channels of the signal when the F0 separation was four semitones than when it was zero semitones. This interaction was significant [F(8,104)=4.21; p<0.001].

6.3.2 Experiment 2

The results for experiment 2 are given in Figure 6.2. The bottom right-hand panel shows the arithmetic mean SRT across all subjects, and error bars show the standard errors. The other panels show the results for individual subjects. A within-subjects ANOVA with factors F0 separation and CO was conducted on the data, using the SRT for each subject and condition as the variate.

SRTs decreased (performance improved) with increasing CO, and this effect was significant [F(2,20)=151.26; p<0.001]. This is consistent with the results of experiment 1 and with the findings of Hopkins *et al.* (2008), as mentioned before. The effect was roughly linear; adding TFS in all 32 channels gave approximately twice the benefit of adding TFS in the lowest 16 channels. This suggests that TFS information in high-frequency channels can contribute to speech intelligibility.

SRTs decreased (performance improved) as expected with increasing F0 separation, and this effect was also significant [F(3,30)=75.64; p<0.001]. Consistent with other studies where the benefit of F0 separation for target sentence intelligibility was measured, the benefit observed was gradual, and did not plateau above one semitone (Brokx and Nooteboom, 1982; Bird and Darwin, 1997).

There was a greater benefit of increasing CO when the F0 separation was

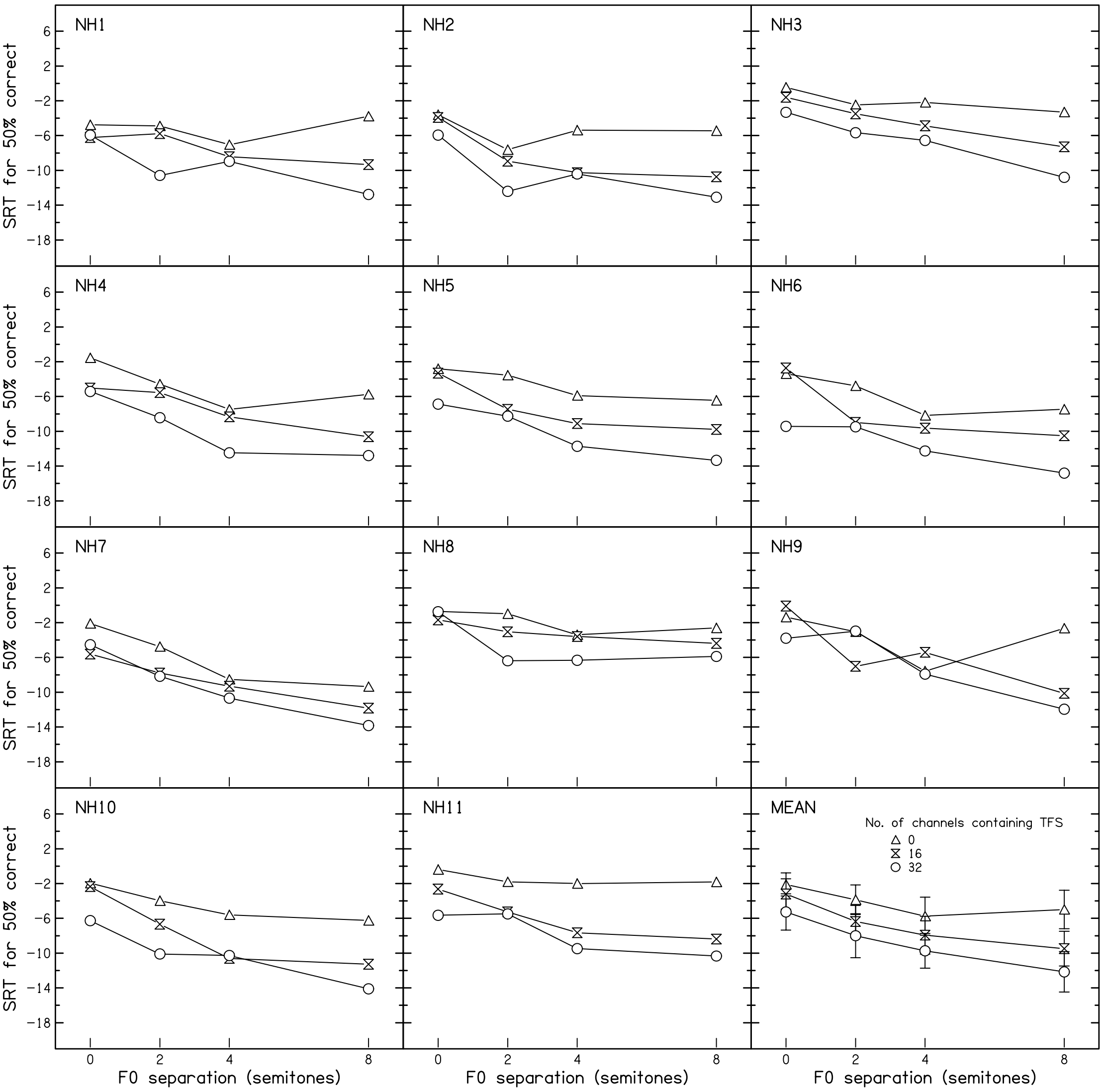


Figure 6.2: SRTs for each subject and condition in experiment 2. The bottom right-hand panel shows the arithmetic mean and standard error across subjects.

higher, confirmed by a significant interaction between F0 separation and CO [F(6,60)=6.72; p<0.001]. There was a significant benefit of increasing $F0sep$ from zero to two semitones and from two to four semitones even when the entire signal was vocoded ($CO = 0$). Increasing the number of channels that contained TFS allowed listeners to make greater use of the benefit from increasing the F0 separation between speakers.

6.3.3 Correction for experiment 1

As mentioned previously, the data for experiment 1 did not show a continuous trend between $CO = 0$ and $CO = 32$ for each F0 separation: the data were rather noisy. It would be expected that the SRT should decrease monotonically with increasing CO, as was observed for experiment 2. It is likely that this effect was not observed for experiment 1 due to the fact that the presentation order of blocks of sentences was counterbalanced to reduce the effect of training. Each subject would have heard (probably) a different section of background speech simultaneously with each sentence in a given list, but the value of $F0sep$ and CO for a given sentence list would have been the same for all subjects. This would have introduced some systematic variability in the data resulting from the inherent variability between sentence lists.

It would have been better instead to counterbalance for type of processing, so each subject would hear a given sentence block processed in a different way. This would have reduced the noise in the data. It is possible, for example, that the sentence list presented for the condition $F0sep = 0$ and $CO = 24$ was slightly easier than other lists, in spite of the fact that the IEEE sentence lists are balanced for linguistic content (Rothauser *et al.*, 1969).

The data for experiment 2 were much more regular, which is likely to be because the flattened F0 contour made the task easier, reducing the effect of the noise introduced by the incorrect counterbalancing.

One way of correcting for the effects of the incorrect counterbalancing is to fit the relationship between SRT and CO with a straight line, consistent with what was found for experiment 2. A straight line was fitted to the data for each value of $F0sep$ in experiment 1, and the equation for each line was then used to predict the SRT for each value of *CO* that would have been expected for that F0 separation.

Figure 6.3 shows the arithmetic mean SRTs across all subjects for experiment 1 using this linear fit. The fitted lines indicate that listeners received a slightly greater benefit from adding TFS when the F0 separation was greater. However, for the fitted lines, the slight improvement in SRT between $F0sep = 0$ and $F0sep = 4$ for $CO = 32$ is no longer apparent.

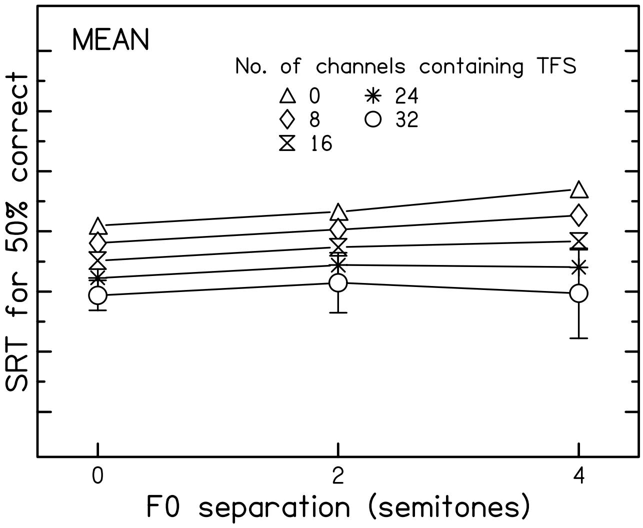


Figure 6.3: Fitted SRTs for each subject and condition in experiment 1 as a linear function of F0 separation for each value of *CO*.

6.4 Discussion

6.4.1 The effect of *CO*

It is clear from both Figure 6.2 and Figure 6.3 that performance improved as *CO* increased, that is, when more frequency channels contained the original TFS information. As spectral information is well preserved when a 32-channel vocoder is used, the benefit from having the intact signal in some channels is probably due to the availability of TFS information. Being able to use the TFS of a signal may allow good pitch perception, and give a clear sensation of the F0 of a signal (Moore *et al.*, 2006; Hopkins and Moore, 2007). In the experiments presented here, separation of concurrent vowels may have been aided by a clear pitch sensation for each vowel; this could be due to the use both of the TFS derived from individually resolved harmonics (for low values of CO) and of TFS information derived from harmonics just beyond the limits of peripheral resolution.

The improvement in performance when CO was increased from 0 to 8 in experiment 1, and from 0 to 16 for experiment 2, is likely to be due to TFS information from resolved harmonics. The frequency range of the first 16 channels was about 100 Hz to 1605 Hz, which would have allowed 10–16 harmonics of the target and background voices to be represented depending on the F0 separation. Spectral information from these harmonics would have been present when the signal was entirely vocoded, but phase-locking to the temporal pattern of individual harmonics would not have been possible when CO was 0. This improvement is consistent with studies showing that stimuli containing low frequency information (Rossi-Katz and Arehart, 2005), or tones containing resolved harmonics (Darwin, 1992; Carlyon and Shackleton, 1994), are more likely to be perceived as originating from separate sources when across-formant grouping mechanisms are used to separate the signals.

For experiment 1, Figure 6.3 shows an improvement in performance when CO was increased from 8 to 16 for each value of $F0sep$. The frequency range in which TFS was added to the signal from this increase in CO was between about 548 Hz and 1605 Hz, adding harmonic ranks of about 6 to about 16 to the target voice, and ranks of about 6 to 13 to the background voice, depending on the F0 separation. The harmonic ranks for which TFS information was added to the signal would have varied dynamically as the F0 contour rose and fell. For an F0 of 96 Hz, the limit of peripheral resolvability is likely to be around the 5th harmonic, where a resolved harmonic is defined as one for which the spacing between it and neighbouring harmonics exceeds 1.25 on the $\mathrm{ERB_N}$-number scale (Moore and Ohgushi, 1993). By this definition, the limit of peripheral resolvability for an F0 of 121 Hz (corresponding to the highest mean F0 used for the background speaker) is around the 6th harmonic. Therefore, adding TFS information from harmonics that were likely to be just at or beyond the limit of peripheral resolvability improved performance in this task.

It is interesting to note that the data for experiment 1 showed a benefit of increasing CO from 24 to 32, a frequency range of about 4102 Hz to 10000 Hz. Similarly, the data for experiment 2 showed a benefit of increasing CO from 16 to 32. For experiment 2, the relationship between SRTs and CO was approximately linear: the benefit of increasing CO from 0 to 32 was about twice as great as the benefit of increasing CO from 0 to 16 for each value of $F0sep$. These observations are consistent with a finding of Hopkins and Moore (2010b) that TFS information in speech is important over a wide range of frequencies. It is also possible that the benefit in increasing CO from 24 to 32 was due to the observation that TFS cues and envelope cues are related; recent modelling work by Kates (2011) suggests that when the TFS is removed from a signal via use of a vocoder, there is a loss in the accuracy of the envelope reproduction. Kates (2011) suggested that using a tone vocoder gives the best accuracy.

6.4.2 The effect of F0 separation

In experiment 2, where the F0 contours of both speakers were flattened, increasing the F0 separation between the target and the background speaker improved performance, as expected. It is likely that increased frequency separation allowed the formants of the vowels from each talker to be grouped appropriately (Brokx and Nooteboom, 1982; Bird and Darwin, 1997). When the F0 separation was large, formants of concurrent vowels would have fallen in more distinct spectral regions, so the differing harmonic and temporal structure of each could be extracted more easily. This finding is also broadly consistent with that of Darwin (1981), who showed that it was possible to ignore information from the second syllable of a formant if this formant had a different F0 to the first, third and fourth formants. It is unlikely that the F0 difference aided grouping due to beating cues, because this cue is less usable for long stimuli, and is dominant for small F0 separations up to about a semitone (Culling and Darwin, 1994).

For experiment 1, where the natural F0 variation of both speakers was preserved, increasing the F0 separation between the target and the background speaker did not improve performance. In fact, performance worsened slightly with increasing F0 separation. It is possible that the natural variations in F0 provided enough momentary F0 difference between the speakers that a shift in the mean F0 of the sentence did not improve intelligibility. This is consistent with the observation that the mean SRT for the condition where CO was 32 was about -9 or -10 dB for all F0 separations for experiment 1, but decreased from about -6 dB to -12 dB with increasing F0 separation for experiment 2. This observation suggests that using an even larger F0 separation might have yielded some slight improvement, but that the main reason for the lack of improvement seen was the inherent momentary F0 difference between the speakers. The F0 variation of all the sentences in the IEEE corpus was measured, and it was found that 95 % of the F0s in a given sentence were within about plus or minus 3 semitones of the mean F0 for that sentence, making the maximum F0 separation at a given moment likely to have been up to 6 semitones. Studies mentioned earlier, where speech intelligibility was measured for concurrent sentences, showed an improvement in performance over the range of zero to three (Brokx and Nooteboom, 1982) or four (Bird and Darwin, 1997) semitones of F0 separation. However, for both of these studies, both the target and background speech was processed to be monotone, as in experiment 2 here.

A second possible reason for the counterintuitive effect of $F0sep$ demonstrated in experiment 1 is that the symmetry of the effect of F0 separation was not tested. However, Bird and Darwin (1997) measured concurrent sentence intelligibility for conditions where the masker was both higher and lower in F0 than the target, and found that the effect was broadly symmetrical. One possible extension to the experiments described here would be to measure SRTs for conditions where the masker F0 is lower than the target F0, to attempt to replicate the finding of Bird and Darwin.

A third possible reason, and the most likely, for the observed worsening in performance with increasing F0 separation is the inherent variability between sentence lists. In future, it would be better to counterbalance the order of sentence lists for each condition rather than to counterbalance the order of conditions to attempt to eliminate training effects. That said, it is likely that for a wide enough range of F0 separations the noise introduced by the incorrect counterbalancing would become unimportant.

6.4.3 Interaction between CO and F0 separation

Listeners gained more benefit from increasing the F0 separation between the target and background speaker when more TFS was available in the signal. This was true both when the F0 contours of the speakers were dynamically fluctuating (experiment 1), and when they were flat (experiment 2). This is consistent with the idea that across-formant grouping based on F0 differences is helped by TFS information from higher (unresolved) harmonics. Culling and Darwin (1993) found that information from higher formants in concurrent vowel sounds contributed to identification when the F0 separation was sufficiently large, which also supports this hypothesis.

6.5 Conclusions

1. Normal hearing listeners gained a benefit to speech intelligibility in a competing-speaker task from increasing the number of channels containing TFS. This effect was observed for both dynamically fluctuating and flat F0 contours.

2. The benefit of adding channels containing TFS information to the signal was equal for low and high frequency regions when the F0 contours were flat. When the F0 contours were dynamic, there was a benefit of adding TFS information for frequencies above 4 kHz. This is consistent with the finding of Hopkins and Moore (2010b) that TFS information in speech is important over a wide range of frequencies. No clear linear relationship was seen when the F0 contours were dynamic.

3. Increasing the F0 separation between the target and background speakers improved intelligibility when the F0 contours were flat, consistent with earlier work (Brokx and Nooteboom, 1982; Bird and Darwin, 1997) but not when they were dynamic. It may be the case that the momentary F0 differences across talkers—up to about 6 semitones—introduced by the natural variations in F0 were sufficient to allow F0 cues to be exploited, resulting in no substantial improvement when the F0 separation was increased. It is

possible that increasing the range of F0 separations tested for experiment 1 would have yielded a slight improvement.

4. Listeners gained more benefit from increasing the F0 separation between the target and background speaker when more TFS was available in the signal. This is consistent with the idea that across-formant grouping mechanisms depend partly on TFS information from unresolved harmonics.

Chapter 7

Summary, Conclusions and Extensions

7.1 Summary

The aim of this thesis was to investigate the role of temporal fine structure (TFS) information in the perception of complex signals such as tones and speech.

7.1.1 Chapter 2: TFS and resolvability hypotheses

The TFS and resolvability hypotheses were described in chapter 2. These hypotheses have been offered in the literature to explain the observation that fundamental frequency difference limens (F0DLs) for complex tones increase as the number of the lowest harmonic in the stimulus, N, increases from about 7 to about 14. A model described in chapter 3, and experiments described in chapters 4 and 5, were intended to provide evidence regarding which of these hypotheses was more nearly correct.

Chapter 2 also presented an overview of different definitions of resolvability and how it can be measured, concluding that the majority of available evidence is consistent with the idea that harmonics above the 7^{th} or 8^{th} are not resolved on the basilar membrane for intermediate F0s. This conclusion was supported by data presented in chapters 4, 5 and 6.

7.1.2 Chapters 3 and 4: Excitation-pattern model and human data

Moore and Sek (2009a) developed a fast method that was intended to measure the sensitivity of a given listener to changes in TFS between successive stimuli. However, some authors have suggested that the task may be performed using excitation-pattern cues. Chapter 3 described the implementation of a computer model that was designed to predict the pattern of results that would be obtained using Moore and Sek's task if changes in the excitation pattern between the tones in each interval were the sole cue used.

In the task of Moore and Sek (2009a), subjects were required to discriminate a harmonic complex tone (H) with a given fundamental frequency (F0) from a complex tone in which all components were shifted upwards by the same amount in Hertz, ΔF, resulting in an inharmonic tone (I). The two tones had the same envelope repetition rate (equal to F0), but had different TFS. Subjects were asked to identify which of two intervals contained tones in the format "HIHI", the other interval containing tones in the format "HHHH". To reduce cues related to differences in the excitation patterns of the H and I tones, all tones were passed through a bandpass filter with a flat region with a width equal to an integer multiple of the F0 (which was varied), and skirts that decreased in level at a rate of 30 dB/octave. The stimuli were presented in a threshold-equalizing noise (TEN, Moore *et al.*, 2000).

In spite of the bandpass filter and the TEN, some residual changes in excitation pattern remained between the HIHI and HHHH intervals, and it is possible that these changes were available to the human listeners to some extent. These cues were used by the computer models. Each model selected which interval was most likely to be the HIHI interval on the basis of detecting a regular pattern of differences between the excitation patterns of the tones within each interval. Model A used Glasberg and Moore's (1990) estimate of the sharpness of auditory filters, and model B used modified auditory filters that were twice as sharp as assumed by Glasberg and Moore.

The same stimuli and method were used in chapter 4, but for human listeners rather than a computer model. This chapter compared data obtained from human listeners and data predicted by the two excitation-pattern models for four experiments, each involving a different manipulation of the stimuli that would be expected to influence the availability of useful excitation-pattern cues. Experiment 1 assessed the effect of varying the number of components in the flat region of the filter passband for two different F0s and two values of N. Experiment 2 provided data for a range of values of F0 and N. Experiment 3 tested whether a change in the signal-to-noise ratio in experiment 1 could explain the predicted and obtained results. Experiment 4 assessed the effect of applying a random level perturbation to each component in each stimulus. For all experiments, components were added in both random and cosine phase.

Experiments 1 and 4 showed that human performance worsened with increasing passband width, and both experiments 1 and 2 showed a worsening with increasing N and decreasing F0. The predictions of both model A and model B followed the same trends as the obtained data for experiments 1 and 2, but to different extents. In particular, the worsening predicted by model B for an increase in passband width or for a decrease in F0 was much less than obtained. Experiment 3 showed that the effect of passband width was not due to a change in signal-to-noise ratio for either the human listeners or the models. Experiment 4 showed that human performance was not disrupted by adding a random level perturbation to each component in each tone, while performance predicted by both model A and model B was disrupted markedly. The results of experiment 4 in particular imply that the human listeners were not using differences in the excitation patterns of the tones in each interval to perform the task, which is evidence against the resolvability hypothesis.

In light of this conclusion, if the TFS hypothesis is accepted, the data obtained from human listeners in experiment 2 provide a good account of how TFS processing changes under different conditions. Performance for each F0 worsened with increasing N. This effect could occur partly because, as N increases for a fixed F0, adjacent peaks in the TFS of the waveform become closer together in time. This means that the time intervals to be discriminated become more similar. Performance also worsened with decreasing F0, consistent with earlier work showing that TFS processing is poorer when F0 is decreased below about 200 Hz (Moore and Sek, 2009a; Moore, Hopkins, and Cuthbertson, 2009b). The worsening in performance for low F0s is consistent with the idea that the auditory system is unable to measure long time intervals accurately (de Cheveigné and Pressnitzer, 2006). TFS processing was still good when all the components in the flat part of the filter bandpass had frequencies higher than 6 kHz, which suggests that the frequency limit for phase-locking in the human auditory system may be higher than the 4-5 kHz usually assumed (Kiang *et al.*, 1965; Palmer and Russell, 1986).

If it is accepted that information from resolved harmonics is more salient than TFS or envelope information from unresolved harmonics, then it would be expected that information from resolved harmonics would be used preferentially by the auditory system to extract a pitch from a stimulus if such information were present. In chapter 4, it was shown that difference in the excitation patterns between the stimuli could not explain the pattern of results for harmonic ranks of 9 and above for an F0 of 200 Hz. As information from resolved harmonics was unlikely to have been the reason for the good performance demonstrated in chapter 4, it is implied that the 8th harmonic (the lowest audible) was unresolved, supporting the conclusion of chapter 2 that the limit of resolvability for intermediate F0s is about the 7th or 8th harmonic rank for an F0 of 200 Hz.

7.1.3 Chapter 5: Resolvability for low F0s

The experiments presented in chapter 5 supported the TFS hypothesis in an indirect way. These experiments extended the paradigm of Moore, Glasberg, and Shailer (1984) by measuring F0DLs of a group of harmonics (group B) embedded in a fixed harmonic background (group A) rather than measuring frequency DLs for a single harmonic. The main intention was to ascertain the ranks of the harmonics that had the greatest effect on the pitch of complex tones with low F0s, for which it is likely that fewer harmonics are resolved on the basilar membrane. A range of F0s and values of N, the rank of the lowest harmonic present in the shifted group, group B, were tested. Comparison of thresholds when components were added in random and cosine phase, and with and without the availability of pitch pulse asynchrony (PPA) cues, was intended to provide some indirect evidence as to the extent to which the harmonics were resolved and hence for the use of TFS cues for pitch perception in this paradigm.

Under conditions where PPA cues were not usable, performance worsened when N was increased above a certain value. The increase first occurred for about $N = 7$ for F0 = 50 Hz and F0 = 100 Hz and for about $N = 5$ for F0 = 200 Hz. If the worsening in performance with increasing N were due to a reduction in resolution of the harmonics (the resolvability hypothesis) one would expect the value of N at which the increase first occurred to increase with increasing F0, since the relative bandwidths of the auditory filters (bandwidth divided by centre frequency) decrease with increasing centre frequency over the range 50 to 500 Hz. In fact, this was not the case: the increase occurred for a lower N for F0 = 200 Hz than for F0 = 50 Hz. For F0 = 50 Hz, one would have expected that an increase in N from 1 to 5 would result in no harmonics in group B being resolved, but in fact F0DLs did not increase over that range. These aspects of the results are not consistent with the idea that the worsening in performance with increasing N was due to a reduction in resolution of the harmonics, which is evidence against the resolvability hypothesis.

Overall, performance was better for F0 = 200 Hz than for F0 = 50 Hz or 100 Hz. Especially for low N, performance worsened with decreasing F0. These findings are consistent with the changes in TFS processing with F0 demonstrated in chapter 4.

7.1.4 Chapter 6: TFS and F0 separation for speech intelligibility

The main aim of chapter 6 was to assess the contributions of F0 differences and TFS information to intelligibility in a competing-speaker paradigm. Target and background speech signals were processed to vary the mean F0, were filtered into 32 channels, and then had original TFS information removed from all, some or none of the channels using a tone vocoder.

It was found that adding TFS information to the signal improved intelligibility, even when the information added was to frequency regions corresponding to harmonic ranks well beyond the limit of peripheral resolution. Listeners gained more benefit from increasing the F0 separation between the target and background speaker when more TFS was available in the signal. This is consistent with the idea that the across-frequency mechanisms by which the formants of concurrent vowel sounds are grouped appropriately depend partly on TFS information from unresolved harmonics.

Although the experiments in chapter 6 used speech signals rather than complex tones, the conclusion that TFS information can mediate good pitch perception beyond the limit of peripheral resolution is common to chapters 4, 5 and 6.

7.2 Conclusions

Overall, the results support the idea that normally hearing human listeners are sensitive to changes in the TFS information derived from harmonics beyond the limit of peripheral resolution. Moore and Sek's (2009a) test probably provides a direct measure of the sensitivity of a given listener to TFS information at different frequencies, and could be used clinically to diagnose the early stages of sensorineural hearing loss.

The use of TFS for pitch perception is only possible when the time intervals to be discriminated are not too similar. For complex tones with F0s around 200 Hz, TFS information can reliably be used for pitch perception up to a harmonic rank of about 13. For lower F0s, such as 50 Hz, TFS information can be used up to a harmonic rank of about 9. Above these approximate limits, the intervals that the auditory system must discriminate in order to use TFS cues become sufficiently similar that these cues cannot be used.

The ability of the auditory system to use phase-locking information decreases for low F0s, suggesting that the auditory system has difficulty estimating long time intervals accurately. The usual assumed limit for the use of phase-locking information for pitch perception is about 5000 Hz, but, if the TFS hypothesis is accepted, data presented here show it is likely that TFS information can be used to discriminate complex tones when all components have frequencies higher than 6000 Hz.

For low F0s, the limiting harmonic rank for the use of TFS information is much higher than the limit of resolvability calculated in a number of different ways. The presence of a phase effect for harmonic ranks as low as 1 and 3 suggests the possibility that no harmonics in a complex tone with a low F0 are resolved, whereas pitch perception does not worsen until harmonic ranks are above about 5.

TFS information enhances the intelligibility of target speech against a competing speaker, perhaps by aiding perceptual segregation of the target and masker. The presence of TFS information where two speakers are present allows more effective use of F0 differences between the speakers to separate the concurrent sounds. TFS information in speech is important over a wide range of frequencies.

7.3 Extensions

The work presented in this thesis falls into three main categories: modelling of Moore and Sek's method for testing sensitivity to TFS, psychoacoustic experiments, and speech intelligibility experiments. Each of these areas of work has raised questions that could be addressed in the future.

7.3.1 Modelling

7.3.1.1 Excitation-pattern models

The selected implementation of the excitation-pattern models described in chapter 3 of this thesis was just one from a number of possible implementations. Earlier work attempted to use the single largest excitation-level difference or simple correlation as a measure of the similarity of the excitation patterns of tones within an interval. These approaches were discarded as they did not take account of the possibility that a regular pattern of spectral ripples might be more discriminable than a single, large ripple or general differences between the excitation patterns.

The cross-correlation approach that was used in chapter 3 was sensitive to this regular pattern, rather than to the depth of single ripples, attempting to quantify the similarity of the difference function between the tones in each interval to a “template” difference function. Other, more complicated methods could test the difference function for regularity in a different way.

One possibility might be to compare the similarity of the difference function in each interval to a sine wave, rather than to a template, assuming that when ΔF is 0.5F0 the ripples in the excitation patterns for the H and I tones are perfectly out of phase (although "spectral phase" might be a more accurate description). The excitation patterns could be corrected for the change in absolute excitation level with increasing frequency before the difference function is calculated. The Fast Fourier Transform (FFT) of the difference function could be calculated for each interval, and the frequency range of the resulting spectra could be compared, with the transform for which the range was wider corresponding to a "less sinusoidal" difference function.

Another possibility is using autocorrelation to test for self-similarity in a given difference function, assuming that a regular pattern of ripples would be more self-similar than noise.

It would be interesting to explore these methods and other possibilities in the future to see how close predictions on the basis of excitation-pattern differences can come to human performance.

7.3.1.2 Temporal models

An analysis conducted as part of this research, but which was not described in this thesis, attempted to explore whether the pitch-strength model of Ives and Patterson (2008), based on the Auditory Image Model (Patterson, 2000), could predict human performance in Moore and Sek's (2009a) task. It was found that human performance and predicted performance did not vary in the same way, but certain limitations to the attempt were accepted due to time constraints. It would be valuable in the future to revisit this work, as the question of whether human TFS processing can be accounted for by temporal models in their current form is an important one.

7.3.2 Psychoacoustics of complex tone pitch perception

The scope for extending the experiments concerning the psychoacoustics of complex tone pitch perception is probably infinite, but there are several extensions to the work presented here that would be interesting and useful to explore.

Firstly, more investigation of Moore and Sek's method, used in chapters 3 and 4, is required to confirm the conclusion that the method does indeed measure sensitivity to TFS for $N = 9$. While the work presented in this thesis goes some way towards excluding the use of excitation-pattern or temporal envelope cues to perform the task, this is negative evidence, rather than positive evidence confirming the use of TFS cues. The problem is that there is little value in making predictions about what patterns might be seen in thresholds across conditions if TFS cues are used by human listeners, as a case could be made for a variety of patterns of results. When considering other work that has attempted to provide evidence for the use of TFS cues in pitch perception, it seems that there is a fairly compelling case for TFS underlying pitch perception for harmonic ranks between about 7 and 14 (Moore *et al.*, 2006; Ives and Patterson, 2008; Moore and Sek, 2009a).

It would be interesting to collect more data from human listeners to show how large the limit of the random amplitude perturbation, P, applied to each component (chapter 4) can be before performance in the task is disrupted, for a range of values of N. Both hypotheses would predict a sharp increase in the threshold value of P for performance to be disrupted with increasing N, but the resolvability hypothesis would predict this increase to occur for a higher value of N than the TFS hypothesis. The reason for this is that the amplitude perturbation is expected to disrupt the spectral ripples evoked by resolved harmonics, so the value of P that affects performance should be very low for resolvable harmonic ranks. If the TFS hypothesis holds, harmonic ranks up to about 7 or 8 should be adversely affected, whereas if the resolvability hypothesis holds, harmonic ranks up to about 10 or 11 should be adversely affected by the amplitude perturbation. Above resolvable harmonic ranks, a greater value of P would be needed to disrupt performance. It is reasonable to assume that there is some value of P that would disrupt performance for high, unresolved harmonics beyond the limit of TFS processing abilities, but indications from the data presented here are that this value would be quite high—above plus or minus 5 dB for a harmonic rank of 13.

Another possibility is that high F0s could be tested using non-zero values of P to rule out the possibility that apparent TFS processing for frequencies higher than 5 kHz is due to some irregularity of headphone response at high frequencies.

7.3.3 Speech

The experiments presented in chapter 6 demonstrated that adding original TFS to channels of a competing-speaker signal increases intelligibility more when there is a greater F0 separation between the speakers, but that separations up to four semitones only improve performance when the competing speakers are monotone.

A wider range of F0 separations should be tested for speech with a natural contour to assess whether some benefit is seen when the applied F0 separation is larger than that present momentarily between the F0s of the two speakers.

Testing more values of the number of channels to which TFS information is added would provide a better estimate of which frequency regions contain the TFS information that is the most important.

It should also be confirmed that the effect of F0 separation is symmetrical in this paradigm. It is possible that SRTs are different for the same F0 separation when the background speech has a higher F0 than the target and when the background speech has a lower F0 than the target.

References

Assmann, P. F. and Summerfield, Q. (**1994**). “The contribution of waveform interactions to the perception of concurrent vowels”, Journal of the Acoustical Society of America **95**, 471–484. 107, 108

Bacon, S. P. and Jesteadt, W. (**1987**). “Effects of pure-tone forward masker duration on psychophysical measures of frequency selectivity”, J. Acoust. Soc. Am. **82**, 1925–1932. 12

Bernstein, J. and Green (**1987**). “The profile-analysis bandwidth”, J. Acoust. Soc. Am. **81**, 1888–1895. 12, 20

Bernstein, J. G. and Oxenham, A. J. (**2003**). “Pitch discrimination of diotic and dichotic tone complexes: harmonic resolvability or harmonic number?”, Journal of the Acoustical Society of America **113**, 3323–3334. 7, 10, 15, 26, 27, 69

Bird, J. and Darwin, C. (**1997**). “Effects of a difference in fundamental frequency in separating two sentences”, in *11th International Conference on Hearing, Grantham, UK, August, 1997*. 108, 124, 129, 130, 131

Broadbent, D. E. and Ladefoged, P. (**1957**). “On the fusion of sounds reaching different sense organs”, Journal of the Acoustical Society of America **29**, 708–710. 109

Brokx, J. P. L. and Nooteboom, S. G. (**1982**). “Intonation and the perceptual separation of simultaneous voices”, Journal of Phonetics **10**, 23–36. 108, 124, 129, 130, 131

Buus, S. and Florentine, M. (**1995**). “Sensitivity to excitation-level differences within a fixed number of channels as a function of level and frequency”, in *Advances in Hearing Research*, edited by G. A. Manley, G. M. Klump, C. Kppl, H. Fastl, and H. Oekinghaus, 401–412 (World Scientific, Singapore). 12, 18, 45

Cariani, P. A. and Delgutte, B. (**1996**). “Neural correlates of the pitch of complex tones. I. Pitch and pitch salience”, Journal of Neurophysiology **76**, 1698–1716. 3, 76

Carlyon, R. P. (**1994**). “Detecting mistuning in the presence of synchronous and asynchronous interfering sound”, Journal of the Acoustical Society of America **95**, 2622–2630. 110

Carlyon, R. P. and Shackleton, T. M. (**1994**). “Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms?”, Journal of the Acoustical Society of America **95**, 3541–3554. 71, 85, 113, 128

Culling, J. F. and Darwin, C. J. (**1993**). “Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0”, Journal of the Acoustical Society of America **93**, 3454–3467. 109, 131

Culling, J. F. and Darwin, C. J. (**1994**). “Perceptual and computational separation of simultaneous vowels: Cues arising from low-frequency beating”, Journal of the Acoustical Society of America **95**, 1559–1569. 107, 108, 129

Culling, J. F. and Summerfield, Q. (**1995**). “Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay”, Journal of the Acoustical Society of America **98**, 785–797. 111

Dai, H. and Micheyl (**2010**). “Psychophysical reverse correlation with multiple response alternatives”, Journal of Experimental Psychology **36**, 976–993. 20

Darwin, C. J. (**1981**). “Perceptual grouping of speech components differing in fundamental frequency and onset time”, Quarterly Journal of Experimental Psychology **33A**, 185–287. 109, 110, 129

Darwin, C. J. (**1992**). "Listening to two things at once", in *The Auditory Processing of Speech - From Sounds to Words*, edited by M. E. H. Schouten, 133–147 (Mouton de Gruyter, Berlin). 109, 110, 128

Darwin, C. J. and Hukin, R. W. (**1999**). "Auditory objects of attention: the role of interaural time differences", Journal of Experimental Psychology: Human Perception and Performance **25**, 617–629. 111

Darwin, C. J. and Hukin, R. W. (**2004**). "Limits to the role of a common fundamental frequency in the fusion of two sounds with different spatial cues", Journal of the Acoustical Society of America **116**, 502–506, 0001-4966 Journal Article. 110

Dau, T. (**1996**). "Modeling auditory processing of amplitude modulation", Ph.D., University of Oldenburg, Germany. 21

de Boer, E. (**1956**a). "On the "residue" in hearing", Ph.D., University of Amsterdam. 3, 27, 69

de Boer, E. (**1956**b). "Pitch of inharmonic signals", Nature **178**, 535–536. 18

de Cheveigné, A. and Pressnitzer, D. (**2006**). "The case of the missing delay lines: synthetic delays obtained by cross-channel phase interaction", Journal of the Acoustical Society of America **119**, 3908–3918. 44, 69, 80, 103, 136

Dudley, H. (**1939**). "Remaking speech", Journal of the Acoustical Society of America **11**, 169–177. 114

Finney, D. J. (**1971**). *Probit Analysis*, 3 edition (Cambridge University Press, Cambridge). 33, 122

Gilbert, G. and Lorenzi, C. (**2006**). "The ability of listeners to use recovered envelope cues from speech fine structure", Journal of the Acoustical Society of America **119**, 2438–2444. 116

Glasberg, B. R. and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data", Hearing Research **47**, 103–138. xi, 8, 10, 11, 12, 13, 14, 18, 20, 28, 34, 41, 49, 54, 67, 72, 80, 134

Glasberg, B. R. and Moore, B. C. J. (**2000**). "Frequency selectivity as a function of level and frequency measured with uniformly exciting notched noise", Journal of the Acoustical Society of America **108**, 2318–2328. 23

Gockel, H., Carlyon, R. P., and Moore, B. C. J. (**2005**). "Pitch discrimination interference: The role of pitch pulse asynchrony", Journal of the Acoustical Society of America **117**, 3860–3866. 84, 85

Gockel, H., Carlyon, R. P., and Plack, C. J. (**2004**). "Across-frequency interference effects in fundamental frequency discrimination: questioning evidence for two pitch mechanisms", Journal of the Acoustical Society of America **116**, 1092–1104, 0001-4966 Journal Article. 46, 84, 85

Gockel, H., Moore, B. C. J., Carlyon, R. P., and Plack, C. J. (**2007**). "Effect of duration on the frequency discrimination of individual partials in a complex tone and on the discrimination of fundamental frequency", Journal of the Acoustical Society of America **121**, 373–382. 79

Goldstein, J. L. (**1973**). "An optimum processor theory for the central formation of the pitch of complex tones", Journal of the Acoustical Society of America **54**, 1496–1516. 3, 75

Goldstein, J. L. and Srulovicz, P. (**1977**). "Auditory-nerve spike intervals as an adequate basis for aural frequency measurement", in *Psychophysics and Physiology of Hearing*, edited by E. F. Evans and J. P. Wilson, 337–346 (Academic Press, London). 46

Green, D. M., Onsan, Z. A., and Forrest, T. G. (**1987**). "Frequency effects in profile analysis and detecting complex spectral changes", J. Acoust. Soc. Am. **81**, 692–699. 12, 20

Grimault, N., Micheyl, C., Carlyon, R. P., Arthaud, P., and Collet, L. (**2000**). "Influence of peripheral resolvability on the perceptual segregation of harmonic complex tones differing in fundamental frequency", Journal of the Acoustical Society of America **108**, 263–271. 112

Hacker, M. J. and Ratcliff, R. (**1979**). “A revised table of d′ for m-alternative forced choice”, Percept. Psychophys. **26**, 168–170. 35, 88

Hall, J. W. and Peters, R. W. (**1981**). “Pitch for nonsimultaneous successive harmonics in quiet and noise”, J. Acoust. Soc. Am. **69**, 509–513. 4

Heinz, M. G., Colburn, H. S., and Carney, L. H. (**2001**). “Evaluating auditory performance limits: I. One-parameter discrimination using a computational model for the auditory nerve”, Neural Computation **13**, 2273–2316. 46

Hoekstra, A. and Ritsma, R. J. (**1977**). “Perceptive hearing loss and frequency selectivity”, in *Psychophysics and Physiology of Hearing*, edited by E. F. Evans and J. P. Wilson, 263–271 (Academic, London, England). 7, 8, 26, 69

Hopkins, K. and Moore, B. C. J. (**2007**). “Moderate cochlear hearing loss leads to a reduced ability to use temporal fine structure information”, Journal of the Acoustical Society of America **122**, 1055–1068. 4, 5, 7, 27, 28, 29, 35, 52, 69, 128

Hopkins, K. and Moore, B. C. J. (**2010**a). “Development of a fast method for measuring sensitivity to temporal fine structure information at low frequencies”, International Journal of Audiology **49**, 940–946. 6

Hopkins, K. and Moore, B. C. J. (**2010**b). “The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects”, Journal of the Acoustical Society of America **127**, 1595–1608. 129, 131

Hopkins, K., Moore, B. C. J., and Stone, M. A. (**2008**). “Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech”, Journal of the Acoustical Society of America **123**, 1140–1153. 5, 115, 116, 117, 118, 122, 124

Houtsma, A. J. and Goldstein, J. L. (**1972**). “The central origin of the pitch of complex tones: evidence from musical interval recognition”, Journal of the Acoustical Society of America **51**, 520–529. 4, 69

Houtsma, A. J. M. and Smurzynski, J. (**1990**). "Pitch identification and discrimination for complex tones with many harmonics", Journal of the Acoustical Society of America **87**, 304–310. 4, 7, 8, 15, 26, 69, 71, 77, 110

Ives, D. T. and Patterson, R. D. (**2008**). "Pitch strength decreases as F0 and harmonic resolution increase in complex tones composed exclusively of high harmonics", Journal of the Acoustical Society of America **123**, 2670–2679, 1520-8524 (Electronic) Journal Article Research Support, Non-U. S. Gov't. 7, 12, 27, 71, 140, 141

Kates, J. M. (**2011**). "Spectro-temporal envelope changes caused by temporal fine structure modification", Journal of the Acoustical Society of America **129**, 3981–3990. 129

Kawahara, H. and Irino, T. (**2004**). "Underlying principles of a high-quality speeach manipulation system STRAIGHT and its application to speech segregation", in *Speech segregation by humans and machines*, edited by P. Divenyi, 167–180 (Dordrecht: Kluwer Academic). 118

Kiang, N. Y.-S., Watanabe, T., Thomas, E. C., and Clark, L. F. (**1965**). *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve* (MIT Press, Cambridge, Mass.). 2, 55, 136

Kohlrausch, A. and Houtsma, A. J. M. (**1992**). "Pitch related to spectral edges of broadband signals", Philosophical Transactions of the Royal Society of London, B **336**, 375–382. 87

Krumbholz, K., Patterson, R. D., and Pressnitzer, D. (**2000**). "The lower limit of pitch as determined by rate discrimination", Journal of the Acoustical Society of America **108**, 1170–1180. 44, 69

Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics", J. Acoust. Soc. Am. **49**, 467–477. 33

Licklider, J. (**1954**). ""Periodicity" pitch and "place" pitch", J. Acoust. Soc. Am **26**, 945. 75

Loizou, P. C., Dorman, M., and Tu, Z. (**1999**). “On the number of channels needed to understand speech”, Journal of the Acoustical Society of America **106**, 2097–2103. 115

Lorenzi, C., Gilbert, G., Carn, C., Garnier, S., and Moore, B. C. J. (**2006**). “Speech perception problems of the hearing impaired reflect inability to use temporal fine structure”, Proceedings of the National Academy of Sciences USA **103**, 18866–18869. 115, 116

MacLeod, A. and Summerfield, Q. (**1990**). “A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use”, British Journal of Audiology **24**, 29–43. 118

Meddis, R. and Hewitt, M. (**1991**). “Virtual pitch and phase sensitivity of a computer model of the auditory periphery. ii: Phase sensitivity”, Journal of the Acoustical Society of America **89**, 2883–2894. 15

Micheyl, C., Dai, H., and Oxenham, A. J. (**2010**). “On the possible influence of spectral- and temporal-envelope cues in tests of sensitivity to temporal fine structure”, Journal of the Acoustical Society of America **127**, 1809–1810. 20, 21

Miyazono, H., Glasberg, B. R., and Moore, B. C. J. (**2009**). “Dominant region for pitch at low fundamental frequencies (F0): the effect of fundamental frequency, phase and temporal structures”, Acoustical Science and Technology **30**, 161–169. 85, 86, 94

Miyazono, H., Glasberg, B. R., and Moore, B. C. J. (**2010**). “Perceptual learning of fundamental frequency (F0) discrimination: Effects of F0, harmonic number, and component phases”, Journal of the Acoustical Society of America **128**, 3649–3657. 86, 89, 103

Miyazono, H. and Moore, B. C. J. (**2009**). “Perceptual learning of frequency discrimination for tones with low fundamental frequency: Learning for high but not for low harmonics”, Acoustical Science and Technology **30**, 383–386. 86

Moore, B. C. J. (**1973**). "Frequency difference limens for short-duration tones", Journal of the Acoustical Society of America **54**, 610–619. 44, 46, 102, 103

Moore, B. C. J. (**1993**). "Frequency analysis and pitch perception", in *Human Psychophysics*, edited by W. A. Yost, A. N. Popper, and R. R. Fay, 56–115 (Springer-Verlag, New York). 27

Moore, B. C. J. (**1997**). "Frequency analysis and pitch perception", in *Encyclopedia of Acoustics, Vol. 3*, edited by M. J. Crocker, 1447–1460 (Wiley, New York). 15

Moore, B. C. J. (**2003**a). *An Introduction to the Psychology of Hearing, 5th Ed.* (Emerald, Bingley, UK). 6, 12

Moore, B. C. J. (**2003**b). "Speech processing for the hearing-impaired: Successes, failures, and implications for speech mechanisms", Speech Communication **41**, 81–91. 27

Moore, B. C. J. (**2008**a). "The choice of compression speed in hearing aids: Theoretical and practical considerations, and the role of individual differences", Trends in Amplification **12**, 103–112. 2

Moore, B. C. J. (**2008**b). "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people", Journal of the Association for Research in Otolaryngology **9**, 399–406. 2, 116

Moore, B. C. J. and Ernst, S. M. (**2012**). "Frequency difference limens at high frequencies: Evidence for a transition from a temporal to a place code", Journal of the Acoustical Society of America **132**, 1542–1547. 2

Moore, B. C. J. and Glasberg, B. R. (**1981**). "Auditory filter shapes derived in simultaneous and forward masking", J. Acoust. Soc. Am. **70**, 1003–1014. 12

Moore, B. C. J. and Glasberg, B. R. (**1988**). "Effects of the relative phase of the components on the pitch discrimination of complex tones by subjects with unilateral cochlear impairments", in *Basic Issues in Hearing*, edited by H. Duifhuis, H. Wit, and J. Horst, 421–430 (Academic Press, London). 81

Moore, B. C. J., Glasberg, B. R., and Baer, T. (**1997**). “A model for the prediction of thresholds, loudness and partial loudness”, Journal of the Audio Engineering Society **45**, 224–240. 10

Moore, B. C. J., Glasberg, B. R., Flanagan, H. J., and Adams, J. (**2006**). “Frequency discrimination of complex tones; assessing the role of component resolvability and temporal fine structure”, Journal of the Acoustical Society of America **119**, 480–490. 4, 7, 10, 15, 26, 27, 69, 71, 110, 113, 128, 141

Moore, B. C. J., Glasberg, B. R., and Jepsen, M. L. (**2009**a). “Effects of pulsing of the target tone on the audibility of partials in inharmonic complex tones”, Journal of the Acoustical Society of America **125**, 3194–3204. 10

Moore, B. C. J., Glasberg, B. R., and Oxenham, A. J. J. (**2012**). “Effects of pulsing of a target tone on the ability to hear it out in different types of complex sounds”, Journal of the Acoustical Society of America **131**, 2927–2937. 10

Moore, B. C. J., Glasberg, B. R., and Peters, R. W. (**1985**). “Relative dominance of individual partials in determining the pitch of complex tones”, Journal of the Acoustical Society of America **77**, 1853–1860. 4, 78, 81, 82, 87, 104

Moore, B. C. J., Glasberg, B. R., and Shailer, M. J. (**1984**). “Frequency and intensity difference limens for harmonics within complex tones”, J. Acoust. Soc. Am. **75**, 550–561. 79, 81, 82, 83, 136

Moore, B. C. J., Hopkins, K., and Cuthbertson, S. J. (**2009**b). “Discrimination of complex tones with unresolved components using temporal fine structure information”, Journal of the Acoustical Society of America **125**, 3214–3222. 4, 7, 10, 27, 28, 44, 52, 69, 79, 80, 89, 135

Moore, B. C. J., Huss, M., Vickers, D. A., Glasberg, B. R., and Alcntara, J. I. (**2000**). “A test for the diagnosis of dead regions in the cochlea”, British Journal of Audiology **34**, 205–224. 15, 17, 28, 134

Moore, B. C. J. and Moore, G. A. (**2003**). “Discrimination of the fundamental frequency of complex tones with fixed and shifting spectral envelopes by normally hearing and hearing-impaired subjects", Hearing Research **182**, 153–163. 4, 5, 18

Moore, B. C. J. and Ohgushi, K. (**1993**). "Audibility of partials in inharmonic complex tones", Journal of the Acoustical Society of America **93**, 452–461. 8, 128

Moore, B. C. J., Oldfield, S. R., and Dooley, G. (**1989**). "Detection and discrimination of spectral peaks and notches at 1 and 8 kHz", J. Acoust. Soc. Am. **85**, 820–836. 12, 18, 45

Moore, B. C. J. and Rosen, S. M. (**1979**). "Tune recognition with reduced pitch and interval information", Q. J. Exp. Psychol. **31**, 229–240. 4, 77

Moore, B. C. J. and Sek, A. (**1994**). "Discrimination of modulation type (AM or FM) with and without background noises", Journal of the Acoustical Society of America **96**, 726–732. 12, 18, 45

Moore, B. C. J. and Sek, A. (**2009**a). "Development of a fast method for determining sensitivity to temporal fine structure", International Journal of Audiology **48**, 161–171. 2, 4, 7, 17, 20, 21, 27, 28, 31, 32, 33, 34, 35, 42, 55, 69, 72, 113, 134, 135, 138, 140, 141

Moore, B. C. J. and Sek, A. (**2009**b). "Sensitivity of the human auditory system to temporal fine structure at high frequencies", Journal of the Acoustical Society of America **125**, 3186–3193. 2, 19, 73

Nelson, D. A., Stanton, M. E., and Freyman, R. L. (**1983**). "A general equation describing frequency discrimination as a function of frequency and sensation level", Journal of the Acoustical Society of America **73**, 2117–2123. 46

Ohgushi, K. and Hatoh, T. (**1991**). "Perception of the musical pitch of high frequency tones", in *Ninth International Symposium on Hearing: Auditory Physiology and Perception*, edited by Y. Cazals, L. Demany, and K. Horner, 207–212 (Pergamon, Oxford). 2

Oxenham, A. J., Micheyl, C., and Keebler, M. V. (**2009**). “Can temporal fine structure represent the fundamental frequency of unresolved harmonics?”, Journal of the Acoustical Society of America **125**, 2189–2199, 1520-8524 (Electronic) Journal Article Research Support, N. I. H., Extramural. 4, 15, 69

Oxenham, A. J. and Shera, C. A. (**2003**). “Estimates of human cochlear tuning at low levels using forward and simultaneous masking”, Journal of the Association for Research in Otolaryngology **4**, 541–554, 1525-3961 eng Journal Article United States 100892857. 12

Oxenham, A. J. and Simonson, A. M. (**2006**). “Level dependence of auditory filters in nonsimultaneous masking as a function of frequency”, Journal of the Acoustical Society of America **119**, 444–453, 0001-4966 (Print) Journal Article. 12

Palmer, A. R. and Russell, I. J. (**1986**). “Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells”, Hearing Research **24**, 1–15. 2, 55, 136

Patterson, R. (**2000**). “Auditory images: How complex sounds are represented in the auditory system”, J. Acoust. Soc. Japan(E) **21**, 183–190. 140

Patterson, R. D. (**1973**). “The effects of relative phase and the number of components on residue pitch”, Journal of the Acoustical Society of America **53**, 1565–1572. 15, 18

Patterson, R. D. (**1976**). “Auditory filter shapes derived with noise stimuli”, Journal of the Acoustical Society of America **59**, 640–654. 10

Patterson, R. D. (**1987**). “A pulse ribbon model of peripheral auditory processing”, in *Auditory Processing of Complex Sounds*, edited by W. A. Yost and C. S. Watson, 167–179 (Erlbaum, New Jersey). 16

Patterson, R. D., Allerhand, M. H., and Gigure, C. (**1995**). “Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform”, Journal of the Acoustical Society of America **98**, 1890–1894. 45

Patterson, R. D. and Nimmo-Smith, I. (**1980**). "Off-frequency listening and auditory filter asymmetry", Journal of the Acoustical Society of America **67**, 229–245. 10, 13

Patterson, R. D. and Wightman, F. L. (**1976**). "Residue pitch as a function of component spacing", J. Acoust. Soc. Am. **59**, 1450–1459. 15, 44

Peters, R. W., Moore, B. C. J., and Baer, T. (**1998**). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people", Journal of the Acoustical Society of America **103**, 577–587. 115

Plomp, R. (**1964**). "The ear as a frequency analyzer", Journal of the Acoustical Society of America **36**, 1628–1636. 8, 27, 110

Plomp, R. (**1967**). "Pitch of complex tones", Journal of the Acoustical Society of America **41**, 1526–1533. 4, 78, 82

Plomp, R. and Steeneken, H. J. M. (**1969**). "Effect of phase on the timbre of complex tones", Journal of the Acoustical Society of America **46**, 409–421. 16

Pressnitzer, D., Patterson, R. D., and Krumbholz, K. (**2001**). "The lower limit of melodic pitch", Journal of the Acoustical Society of America **109**, 2074–2084. 80

Ritsma, R. J. (**1967**). "Frequencies dominant in the perception of the pitch of complex sounds", Journal of the Acoustical Society of America **42**, 191–198. 4, 77, 78, 82, 110

Rossi-Katz, J. A. and Arehart, K. H. (**2005**). "Effects of cochlear hearing loss on perceptual grouping cues in competing-vowel perception", Journal of the Acoustical Society of America **118**, 2588–2598, 0001-4966 (Print) Journal Article. 107, 108, 109, 110, 128

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (**1969**). "IEEE recommended practice for speech quality measurements", IEEE Transactions on Audio and Electroacoustics **17**, 225–246. 118, 126

Scheffers, M. T. M. (**1983**). "Sifting vowels: auditory pitch analysis and sound segregation", Ph.D., Groningen University, The Netherlands. 107

Schouten, J. F. (**1940**a). "The perception of pitch", Philips Technical Review **5**, 286–294. 18, 27

Schouten, J. F. (**1940**b). "The residue and the mechanism of hearing", Proceedings of the Koninklijke Nederlands Akademie van Wetenschap **43**, 991–999. 3, 76

Schouten, J. F. (**1970**). "The residue revisited", in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. F. Smoorenburg, 41–54 (Sijthoff, Leiden, The Netherlands). 3

Schouten, J. F., Ritsma, R. J., and Cardozo, B. L. (**1962**). "Pitch of the residue", J. Acoust. Soc. Am. **34**, 1418–1424. 4, 18, 27, 32, 42, 69

Seebeck, A. (**1841**). "Beobachtungen über einige Bedingungen der Entstehung von Tönen", Ann. Phys. Chem. **53**, 417–436. 74

Shackleton, T. M. and Carlyon, R. P. (**1994**). "The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination", Journal of the Acoustical Society of America **95**, 3529–3540. 4, 7, 15, 27

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues", Science **270**, 303–304. 114, 115

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception", Nature **416**, 87–90, 0028-0836 (Print) Journal Article. 115

Terhardt, E. (**1972**a). "Zur Tonhöhenwahrnehmung von Klüngen II. Ein Funktionsschemas", Acustica **26**, 187–199. 75

Terhardt, E. (**1972**b). "Zur Tonhöhenwahrnehmung von Klüngen. I. Psychoakustische Grundlagen", Acustica **26**, 173–186. 75

Terhardt, E. (**1974**). "Pitch, consonance, and harmony", Journal of the Acoustical Society of America **55**, 1061–1069. 3

Unoki, M., Miyauchi, R., and Tan, C.-T. (**2007**). "Estimates of tuning of auditory filter using simultaneous and forward masking", in *Hearing - from Sensory Processing to Perception*, edited by B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Erhery, 19–26 (Springer Verlag, Heidelberg). 13

Vliegen, J., Moore, B. C. J., and Oxenham, A. J. (**1999**). "The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task", Journal of the Acoustical Society of America **106**, 938–945. 111, 112, 113

Vliegen, J. and Oxenham, A. J. (**1999**). "Sequential stream segregation in the absence of spectral cues", Journal of the Acoustical Society of America **105**, 339–346. 111, 112, 113

Ward, W. D. (**1954**). "Subjective musical pitch", J. Acoust. Soc. Am. **26**, 369–380. 2

Xu, L., Thompson, C. S., and Pfingst, B. E. (**2005**). "Relative contributions of spectral and temporal cues for phoneme recognition", Journal of the Acoustical Society of America **117**, 3255–3267, 0001-4966 Journal Article. 115

Zwicker, E. (**1981**). "Dependence of level and phase of the ($2\delta 1-\delta 2$)-cancellation tone on frequency range, frequency difference, level of primaries, and subjects", Journal of the Acoustical Society of America **70**, 1277–1288. 104