

The Role of Differences in Fundamental Frequency between Competing Voices in a Reverberant Room



Mickael Deroche

Thesis submitted to Cardiff University
for the degree of Doctor of Philosophy

October, 2009



The Role of Differences in Fundamental Frequency between Competing Voices in a Reverberant Room



Mickael Deroche

Thesis submitted to Cardiff University
for the degree of Doctor of Philosophy

October, 2009



UMI Number: U585286

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U585286

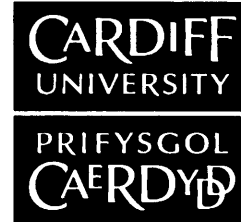
Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

**NOTICE OF SUBMISSION OF THESIS FORM:
POSTGRADUATE RESEARCH**

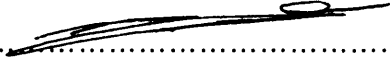


APPENDIX 1:

Specimen layout for Thesis Summary and Declaration/Statements page to be included in a Thesis

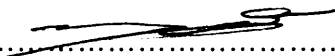
DECLARATION

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed  (candidate) Date25/01/2010...


STATEMENT 1

This thesis is being submitted in partial fulfillment of the requirements for the degree ofPhD.....(insert MCh, MD, MPhil, PhD etc, as appropriate)

Signed  (candidate) Date25/01/2010...

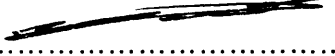
STATEMENT 2

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references.

Signed  (candidate) Date25/01/2010...


STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed  (candidate) Date25/01/2010.....

STATEMENT 4: PREVIOUSLY APPROVED BAR ON ACCESS

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loans after expiry of a bar on access previously approved by the Graduate Development Committee.

Signed  (candidate) Date25/01/2010.....

CONTENTS

Thesis Contents	iii
Acknowledgements	ix
Summary	x
Chapter I. INTRODUCTION	1
I Speech and fundamental frequency (F0)	3
II Auditory scene	4
A. Simultaneous grouping	4
B. Sequential grouping.....	5
C. Different research questions involve different tasks.....	7
D. Different tasks involve different levels of auditory attention: automatic or directed.....	8
III $\Delta F0$ effects	9
A. Discrepancies in $\Delta F0$ effects between vowels and speech	9
B. Mechanisms responsible for $\Delta F0$ effects for vowels	10
B.1 Mechanisms of harmonic selection.....	11
B.2 Mechanism of harmonic cancellation.....	13
B.3 Mechanisms that do not require F0 identification.....	14
B.4 Remaining issues in the literature on double-vowels.....	15
C. Mechanisms responsible for $\Delta F0$ effects for speech.....	16
C.1 Mechanisms common to vowels and speech.....	16
C.2 Mechanism relevant for speech.....	17
IV Speech in rooms	18
A. Degradation of speech intelligibility in rooms.....	18
B. $\Delta F0$ effect in rooms.....	18
B1. Effect of reverberation on harmonic mechanisms.....	19
B2. Effect of reverberation on sequential grouping.....	21
C. Room colouration.....	21

Chapter II. EXPLORING HARMONIC CANCELLATION.23

I Introduction	23
II Exp. 2.1 Effect of frequency region	24
A. Stimuli.....	25
B. Procedure.....	28
C. Results.....	29
D. Discussion	30
D.1 Harmonic cancellation versus stimulus uncertainty.....	30
D.2 No benefit after 3 kHz.....	31
D.3 Unexpected frequency range for harmonic cancellation.....	31
D.4 Within-channel target-to-masker ratios.....	32
III Exp. 2.2 Effect of spectral overlap	35
A. Stimuli and procedure.....	36
B. Results.....	37
C. Discussion.....	39
IV Exp. 2.3 Effect of masker's inharmonicity	41
A. Stimuli and procedure.....	41
B. Results.....	42
C. Discussion.....	43
V Exp. 2.4 Operational bandwidths	46
A. Stimuli and procedure.....	46
B. Results.....	48
C. Discussion.....	49
VI General Discussion	52
A. Harmonic cancellation and the ΔF_0 effect.....	52
B. Parallel with the partial-pitch shift.....	54
C. Towards a predictive model.....	57
VII Conclusion	58

Chapter III. A ROLE FOR GROUPING BY F0..... 60

I Introduction.....60

II Exp. 3.1 Benefit of a large $\Delta F0$ for speech and buzz interferers.....62

 A. Stimuli.....62

 B. Procedure.....64

 C. Results.....65

 D. Discussion66

III Exp. 3.2 Benefit of a small $\Delta F0$ for speech and buzz interferers.....67

 A. Stimuli and procedure.....67

 B. Results.....68

 C. Discussion.....69

 C.1 $\Delta F0$ benefit smaller with speech interferers than with buzz69

 C.2 Temporal continuity of the interferer's F0: evidence of harmonic cancellation.....69

IV General Discussion70

 A. Why does performance increase as $\Delta F0$ increases in speech segregation?70

 B. Small contribution of harmonic cancellation with speech interferers.....71

 C. $\Delta F0$ benefit and the perception of a pitch difference.....72

 D. Importance of the number of interfering voices.....73

V Conclusion74

**Chapter IV. EFFECTS OF REVERBERATION AND F0
MODULATION ON THE $\Delta F0$ BENEFIT WITH A BUZZ
INTERFERER.....75**

I Introduction.....75

II Exp. 4.1 $\Delta F0$ benefit depends on parameters of the buzz interferer.....77

 A. Stimuli.....77

 B. Procedure.....79

 C. Results.....80

III Discussion.....82

A. Harmonic cancellation fails with a modulated reverberant interferer.....	82
B. No evidence for intrinsic effect of F0 modulation.....	83
C. No evidence for harmonic enhancement.....	85
D. STI effect.....	85
E. Possible binaural effects.....	85
IV Conclusion	89

Chapter V. EFFECTS OF REVERBERATION AND F0 MODULATION ON THE $\Delta F0$ BENEFIT WITH SPEECH INTERFERERS.....90

I Introduction.....	90
II Exp. 5.1 Benefit of a small $\Delta F0$, subject to F0-modulated speech interferers.....	91
A. Stimuli and procedure.....	91
B. Results.....	92
C. Discussion.....	94
C.1 STI and interaural phase effects.....	94
C.2 Grouping based upon the degree of reverberation.....	94
C.3 Disruption of harmonic cancellation of speech interferers.....	95
C.4 Reverberation and F0 modulation of both target and interferers.....	96
D. Interim conclusion.....	98
III Exp. 5.2 Reverberation reduces the benefit of a large $\Delta F0$ for monotonized speech interferers.....	99
A. Stimuli and procedure.....	100
B. Results.....	101
C. Discussion.....	103
IV Exp. 5.3 Benefit of a large $\Delta F0$, reduced by F0 modulation and reverberation for speech interferers.....	104

A. Stimuli and procedure.....	106
B. Results.....	106
C. Discussion.....	108
C.1 Weak contribution of harmonic cancellation.....	108
C.2 Floor effect of reverberation	109
C.3 Reverberation affects sequential F0-grouping.....	109
C.4 F0 modulation effects restricted to harmonic cancellation.....	110
C.5 Additional effects of reverberation, regardless of F0.....	110
V Conclusion	111
Chapter VI. GENERAL DISCUSSION.....	113
I Harmonic cancellation in the ΔF_0 effect.....	113
A. Contribution of harmonic cancellation for double-vowels.....	114
B. Contribution of harmonic cancellation for a buzz interferer.....	115
C. Contribution of harmonic cancellation for a speech interferer.....	116
D. Is the cancellation mechanism more generally applicable to spectral templates regularly spaced in frequency?	117
II Sequential grouping in the ΔF_0 effect.....	119
A. Sequential F0-grouping only in speech-on-speech segregation.....	119
B. Sequential F0-grouping reduced by reverberation.....	119
C. Towards a better understanding of the mechanism underpinning sequential F0-grouping: parallel with spatial separation.....	121
III Pitch perception in the ΔF_0 effect.....	122
A. Resolved partials dominate the perception of pitch.....	122
B. Temporal integration required for sequential F0-grouping.....	125
C. Size of ΔF_0 required for listeners to perceive distinct pitches.....	125
IV Application to real speech.....	126
A. Real speech.....	126
A.1 Natural intonation.....	126
A.2 Harmonic cancellation on real speech.....	127
A.3 Sequential grouping on real speech.....	128

B. Multiple talkers.....	128
B.1 Harmonic cancellation of multiple F0s.....	128
B.2 Sequential grouping of multiple F0s.....	129
IV Application to realistic environments.....	129
A. Degree of reverberation of different rooms.....	129
B. Room colouration.....	130
C. Sources directionality and dummy head.....	130
D. Spatial separation.....	131
E. Future challenges towards architectural software.....	132
CONCLUSION.....	134
References.....	138
Appendix A.....	150
Appendix B.....	152

Acknowledgements

I would like to express my deepest gratitude to my supervisor Prof. John F. Culling for his exemplary guidance throughout all my PhD, for his invaluable advice, and for his patience with my unfortunate French accent. Most importantly, I am extremely grateful to John for nourishing my interests in science and inspiring me to pursue my own career in auditory research. I can only hope to maintain our collaboration in the future.

I am very thankful to Dr. Mathieu Lavandier and Mr. Sam Jelfs for their stimulating intellectual contributions during our weekly lab and social meetings. I would also like to thank Dr. Mihaela Iordanova and Mr. David Ross for their insightful discussions regarding each of our new research findings and science in general.

I extend my gratitude to the staff from the School of Psychology, for offering me an ideal work environment, which includes technical and intellectual support. Finally, I thank all the participants of my experiments and the UK EPSRC for funding this research project and providing me with the opportunity to delve into the world of psychoacoustics research, which is ultimately what I would like to continue doing.

Summary

In noisy conversations, listeners can segregate competing voices on the basis of their fundamental frequency (F0). The aim of this thesis was to investigate which mechanisms underlie this F0-segregation ability and whether this ability is affected by reverberation. This work provided evidence for a mechanism, which cancels interfering voices on the basis of their harmonic structure; a process termed harmonic cancellation. We developed a paradigm in which listeners had to detect a band of noise masked by a harmonic or inharmonic complex masker (Chapter II). Harmonic cancellation was found to be beneficial up to about 3 kHz, sensitive to a degree of inharmonicity reflected by a peak autocorrelation of 0.9 or less, and to integrate harmonic information over very large bands. In addition to harmonic cancellation, listeners may also use F0 as a sequential cue, provided that $\Delta F0$ is sufficiently large (Chapter III), in order to organise the auditory scene in the presence of several talkers; a process termed sequential F0-grouping. By manipulating the F0 of competing sources heard in anechoic or in reverberant environments, the Speech Reception Threshold (SRT) of a target voice masked by buzz (Chapter IV) or speech (Chapter V) interferers, was elevated when the interferer but not the target, was F0-modulated and especially in reverberation for the buzz interferer. These results were explained in terms of disruption of harmonic cancellation. Moreover, the benefit of an 8 semitone $\Delta F0$ was disrupted by reverberation even for monotonized sources, suggesting that reverberation is also detrimental to sequential F0-grouping. To conclude, the listener's ability to segregate voices by F0 relies on the mechanisms of harmonic cancellation and sequential F0-grouping. Both these mechanisms are likely to be disrupted in realistic situations of conversation, i.e. real speech in reverberant rooms.

Chapter I.

INTRODUCTION

A cocktail-party (Cherry, 1953) is known as a listening situation where many people talk at once, in noisy rooms. In such situations, the speech message recorded at the listener's ears is highly masked by surrounding noise or degraded. Despite these corruptions in the speech signal, humans can still keep up a conversation. Since our ears do not limit the auditory field (contrary to our eyes which for instance cannot look behind), all audible sounds ultimately enter our ears. However, when concentrating on a specific source (talker, loudspeaker, TV), the sounds in the background seem to fade away. It seems as though focusing one's attention towards the target source enables one to process only those sounds. There must be some mechanisms in the auditory system where sounds that we want to listen to are progressively selected and separated from the background. Cherry (1953) identified several discriminating cues that contribute to perceptual separation of competing voices: spatial location (voices coming from different directions), visual cues (the ability to read lips or any gesture that assists communication), speech characteristics (voice gender, average pitch, average speeds, and accents) and transitional probabilities (topic of conversation, linguistic and syntactic factors).

The present work focuses on the use of differences in speech characteristics and more particularly on the use of differences in the pitch of the voices. The pitch refers to the perceptual sensation that a sound is located somewhere on a musical scale. After 150 years of research on pitch perception, there is still some debate about how pitch is derived from the acoustical signal (Shamma and Klein, 2000; De Cheveigné, 2005). So it might be cautious simply to state that for simple harmonic

complexes, the acoustical correlate of pitch is the fundamental frequency (F_0). Hence, the present work focuses on the use of differences in fundamental frequency (ΔF_0) which in turn might cause perceptual differences in pitch between competing voices.

The first chapter describes the research context and the most influential experimental results that led to the design and the realisation of the present experiments. This literature mainly covers effects of ΔF_0 between competing vowels or voices and effects of reverberation. The second chapter examines a candidate mechanism for the ΔF_0 effect, which aims to remove interfering sounds on the basis of their harmonicity. Such a mechanism is investigated by measuring the detection of a band of noise masked by a harmonic or inharmonic masker. The third chapter uses a task of speech segregation masked by speech or pulse-train, with monotonized F_0 . The ΔF_0 benefits are measured as the difference of speech reception thresholds (SRTs) when the target speech had the same or a different F_0 than that of the masker. This third chapter draws an important distinction between the use of ΔF_0 for simultaneous or for sequential grouping mechanisms (presented in II), depending on whether pulse-train or speech is used as a masker. Chapter four and five arise from this distinction, using a speech segregation task in which the F_0 is controlled. The fourth chapter introduces reverberation to investigate whether it affects the ΔF_0 benefit for pulse-train maskers. The fifth chapter investigates whether reverberation affects the ΔF_0 benefit for speech maskers. Finally, a general discussion of all the results compared to the literature and their implications in realistic cocktail-party situations is presented in the sixth and last chapter.

I. Speech and the fundamental frequency (F0)

The human voice is generated when the lungs force the air to pass between the vibrating vocal folds, producing periodic sounds. These periodic sounds are spectrally modified by the shape of the vocal tract, controlled by the positions of articulators such as the jaw, the tongue and the lips (Stevens et al., 1953) and are heard as phonemes, such as vowels in speech. Vowels are consequently harmonic complexes with spectral peaks called formants, corresponding to resonant frequencies of the vocal tract. Consonants can be subcategorised into unvoiced consonants such as /k/, /t/ and /p/, that are produced by constrictions in the vocal tract and are not harmonic and voiced consonants such as /m/, /w/ and /r/, that are harmonic. The “voiced” parts of speech refer to the harmonic parts of speech, i.e. the vowels and voiced consonants. Note that all harmonic sounds have a fundamental frequency (F0), but a best-fitting F0 can be derived from sounds that are not necessarily harmonic, such as a frequency-shifted complex.

The formants of vowels are highly variable. First, different speakers have vocal tracts of different lengths resulting in different mean values for the formant frequencies. Second, the positions taken by the articulators to produce a given phoneme are influenced by those producing the phonemes before and after this phoneme (co-articulation). Formant frequencies also change with the place of the phoneme in the sentence and with different stresses granted to syllables or words according to their function in the sentence (O’Shaughnessy and Allen, 1983). In spite of such variability, speech is highly resistant to corruption (spectro-temporal modifications of speech) and interference (speech masked by other sounds). The auditory system manages to achieve speech comprehension in difficult listening

conditions, because there are many redundant acoustic cues in speech. The fundamental frequency is one very powerful cue that facilitates recovery of speech in adverse conditions and is the focus of interest of the present thesis.

II. The auditory scene

In most realistic listening situations, several sound sources surround us. As I write these words in my quiet office, I can hear the fan of my computer, other PhD students typing on their keyboards and some cars in the street outside. The sound waves from these separate sources add together to create a superposed mixture of sounds that the auditory system has to decipher. An auditory scene (Bregman, 1990) appears when the mechanisms of the auditory system assign the continuous whir in front of me to the fan of my computer, the dry clicks coming from my left to the keyboard typing and the low-frequency rumble to the cars in the street at my back. A cocktail-party is an especially noisy environment where it becomes challenging to make sense of the sound mixture. Auditory scene analysis refers to the set of mechanisms that separate out competing sounds and group together sounds that share common characteristics in order to attribute different components to their respective sources. Such analysis is a very complex task given that separate sources often share some common features. For instance, several sources might originate from the same azimuth relative to the listener, might have very similar spectra or similar sound levels. The mechanisms involved in the organisation of the auditory scene can be categorised into those of simultaneous or sequential grouping. Simultaneous grouping assigns simultaneous frequency components to their respective sources. Sequential grouping assigns a temporal sequence of sounds to their respective sources. Both occur in realistic situations.

A. Simultaneous grouping

Simultaneous grouping cues include onset cues, which group together components that start together, and harmonic cues, which group together components that form a harmonic series. Scheffers (1983, Chap. III) investigated the identification of vowels masked by a noise background and found that thresholds were systematically lower when the onset of the noise preceded the onset of the vowel than when the noise was only present during the presentation of the vowel. This is an example of the benefit of grouping by onset cues. Scheffers (1983) also found that identification was better for a voiced vowel than a noise-excited vowel, when masked by pink noise. This is an example of the benefit of grouping by harmonic cues or by a form of excitation, i.e. a periodic excitation as opposed to a random excitation. Grouping by F0 can be seen as a subgroup of grouping by harmonic cues and occurs when both sources have periodic excitations but different fundamental periods. In experiments where two voiced vowels are presented simultaneously (a “double-vowel”), the common onset leads the auditory system to group the two vowels, whereas differences in F0 lead it to separate them. With connected speech, simultaneous grouping by onset cues is probably much more relevant due to the dynamic nature of speech, while simultaneous grouping by F0 occurs intermittently during the voiced segments of speech.

B. Sequential grouping

When several sounds are temporally interleaved, they could be heard either as arising from a single source or from more than one source (Miller and Heise, 1950; Bregman and Campbell, 1971). The factors that influence the grouping of competing sounds into different streams have been explored intensively. Currently, the most

common view of sequential grouping is that it depends both on bottom-up and top-down influences, which grant different levels of auditory attention (section D). When it comes to speech segregation, which is the object of the present work, the ambivalence of hearing one or several voices is reflected by an ambiguity as to which voice one should attend to, resulting in a type of masking which is not necessarily related to energetic masking and is sometimes referred as informational masking (Kidd et al., 2005; Shinn-Cunningham et al., 2005). For unprocessed full-spectrum speech, these two types of masking are not mutually exclusive; target and interfering speech share the same frequency bands, so both informational and energetic masking occur. However, Kidd et al. processed the target and interfering sentences such that they occupied different frequency bands, ensuring that energetic masking would be largely absent. They found that large amounts of masking occurred for speech-on-speech configurations, not for speech-on-noise configurations. In order to be released from the attentional ambiguity or informational masking, listeners can use a variety of cues over time to group sounds into sequential streams. Periodicity or F0 (Darwin and Hukin, 2000a; Darwin et al., 2003; Drullman and Bronkhorst, 2004), signal-to-noise ratio (Brungart, 2001; Brungart et al., 2001), spatial separation (Darwin and Hukin, 2000a; Freyman et al., 2001; Hawley et al., 2004; Kidd et al., 2005, Lee and Shinn-Cunningham, 2008), priming by the target talker or onset cues (Freyman et al., 2004), vocal-tract length, sex difference and prosody (Culling and Porter, 2005; Darwin et al. 2003, Brungart et al., 2001), and even tactile cues (Drullman and Bronkhorst, 2004) have been reported throughout the literature to help listeners to form sequential streams. The present experiments used speech material that was originally recorded by the same male talker, so the acoustic cues regarding the talker identity acted to group the competing sentences while F0 cues, among others, acted to separate them.

C. Different research questions involve different tasks

Different tasks have been used depending on whether researchers were focusing on the mechanisms responsible for simultaneous or sequential grouping. Experiments on simultaneous grouping use a target sound very distinct from the masker, in order to address the question of how the manipulated cues influence the detection of the target sound or when it comes to speech, the intelligibility of target words. In contrast, most of the studies, which investigate sequential grouping, intentionally make two competing utterances very similar, so that listeners can confuse the sentence they should attend to. A typical paradigm is known as the coordinate response measure (CRM). Two sentences are made very similar and differ only in a few words, for instance of the form “Ready <call sign>, go to <colour> <number> now”. The task is to choose which of two simultaneous target words is part of the attended utterance rather than part of the unattended utterance. A specific cue, which is the object of the investigation, may help listeners to fulfil this task, provided that this cue is sufficiently strong to maintain continuity of the attended stream. Since there is a limited set of call signs, colours and numbers, the two utterances remain very similar throughout the experiment; as a consequence the intelligibility requirements of such a task are minimal. Those experiments rather address the question of how listeners decide which words belong to the attended sentence, without which speech could be intelligible but meaningless. The experiments on speech segregation, presented in this thesis, necessarily deal with both types of grouping.

D. Different tasks involve different levels of auditory attention: automatic or directed

Mechanisms responsible for simultaneous grouping are often regarded as automatic in that they do not require listeners to engage any effort other than perceiving what is presented to their ears. In contrast, sequential grouping can be influenced by the listener's attention. Listeners can, up to a certain degree, induce an attended sequence to form a separate perceptual group (Carlyon et al., 2001). To clarify the role for auditory attention for sequential grouping, it might be useful to follow the distinction proposed by Bregman (1990). On one hand, "primitive stream segregation" can be seen as a bottom-up mechanism which automatically and obligatorily responds to a set of primitive cues, particularly temporal coherence and frequency separation. On the other hand, "schema-based selection" can be seen as a top-down influence which uses the listener's knowledge of the type of sounds to select them. In the case of segregation by F0, it is not clear which conditions involve the automatic or directed attention of the listener.

The experiments of Chap. II of the present thesis use a very simple task: detection of a noise band masked by a complex tone. The target noise sounds very distinct from the complex masker, so the mechanism investigated in these experiments, thought to be involved in F0 segregation, probably requires little attention from the listener. In contrast, for speech segregation experiments, competing voices are segregated simultaneously and sequentially by F0, leaving a potential role for the listener's attention. One of the aims of the present thesis is to differentiate the use of F0 as a simultaneous cue from its use as a sequential cue, in order to discuss

whether the listener's attention plays a major role in the mechanisms of F0-segregation.

III. $\Delta F0$ effects

A. Discrepancies in $\Delta F0$ effects between vowels and speech

Brokx and Nootboom (1982) investigated the role of differences in F0 in the segregation of simultaneous speech messages. Since real voices are intonated, their F0s vary considerably and the difference in F0 (which I denote $\Delta F0$) between two voices is difficult to evaluate. In order to alleviate this issue, they resynthesized speech recordings of two voices from a linear predictive coding (LPC) analysis, so that they controlled the F0 contour of each sentence. Critically, monotonized speech (fixed F0) was an obvious choice of experimental stimulus that enabled to evaluate $\Delta F0$ precisely. With less accuracy regarding $\Delta F0$, intonated speech was also resynthesized at another mean F0. Whether the two voices were monotonized or intonated, words spoken by competing voices with different F0s (or different mean F0s for the intonated voices) were reported more accurately. The larger the difference in F0 ($\Delta F0$), the lower the percentage in errors in reporting words, except in the monotonized case when the $\Delta F0$ equalled one octave. These results led to the idea that the harmonic structure of target and interfering speech must be distinct in order to avoid perceptual fusion of the two voices. This explanation assumes a mechanism of simultaneous grouping by harmonic cues. Alternatively, when the F0s were identical or had the same mean, listeners might have performed poorly because they inadvertently switched their attention from the target speech to the interfering speech. In other words, the listeners' ability to form a target stream was disrupted. The larger

the ΔF_0 , the less likely sequential grouping is to be disrupted. That explanation assumes a mechanism of sequential grouping by F_0 .

Those two potential explanations received different amount of interest throughout the literature. The role of ΔF_0 for simultaneous grouping has mainly been investigated with simultaneous vowels. Consisting of a single phonetic segment, vowels provide no scope for sequential grouping. Sequential grouping may play a prominent role in connected speech, and has thus mainly been investigated with speech. A distinction between the effects of ΔF_0 s for vowels and for speech is further supported by distinct patterns of improvement in performance with increasing ΔF_0 . Scheffers (1983, Chap. IV) found that double-vowel identification improved sharply from 45% to 62% (on average over eight Dutch vowels) as ΔF_0 increased from 0 to 1 semitone, before it appeared to asymptote at larger ΔF_0 s. In contrast, Bird and Darwin (1998) showed that intelligibility of a target message masked by an interfering message continued to improve progressively up to 10-semitones ΔF_0 . Therefore there is a discrepancy in the improvement of performance with ΔF_0 between vowels and speech. While the results found for vowels may be informative about some mechanisms contributing to ΔF_0 effects, other mechanisms may only be involved for speech, but also contribute to ΔF_0 effects.

B. Mechanisms responsible for ΔF_0 effects for vowels

The mechanisms underlying the ΔF_0 effect have been a matter of controversy throughout the literature. Three approaches can be identified. There are mechanisms that select the partials of one or both vowels on the basis of their common harmonicity, mechanisms that identify and then remove one F_0 , and mechanisms that do not rely on identification of F_0 s at all. Finally, some results remain questionable in the literature on ΔF_0 effects for vowels.

B.1 Mechanisms of harmonic selection

Several studies have attempted to model ΔF_0 effects based on a strategy guided by the identification of competing F_0 s. Since F_0 is closely related to the perception of pitch, these models were based on those for pitch perception using the principles of harmonic sieves (place model) and of autocorrelation (place-time model). Place models (Huggins and Licklider, 1951; Houtsma and Goldstein, 1972) derive pitch determination and spectral segregation by analysing the distribution of RMS levels across the channels of an auditory filter bank whereas place-time models (Wever and Bray, 1930; Johnson, 1980) analyse the periodicities in the waveforms in each channel.

A harmonic sieve (Parsons, 1976; Scheffers, 1983) can be regarded as a bar with narrow slots spaced at the frequencies of the harmonics of a particular F_0 . The idea is to search for the two harmonic sieves which conjointly best describe the pattern of resolved harmonics in the excitation pattern of a double-vowel. Given a certain number of slots and a tolerance within which peaks in the excitation pattern are accepted or rejected by a sieve, a quality score is computed to relate to the degree of match between the sieves and the resolved peaks. The two highest scoring sieves give rise to two dominant F_0 s. The excitation pattern of the double-vowel can then be sampled separately at the harmonic series of the two dominant F_0 s to derive two spectral patterns and which can then be classified using a template-matching procedure.

The temporal analysis in the place-time models uses the autocorrelation function (ACF) of the filtered waveform at the output of channels of the filter bank (Licklider, 1951). An ACF shows a peak whenever the delay is equal to the period of a component in the filtered waveform or its integer multiples. Frequency components

lying near the centre frequency (cf) of a filter often dominate its output. As a result, the ACF can contain a series of peaks at delays of $1/cf$ and its multiples. However, these cf -related periodicities differ from one channel to the next and are eventually averaged out by pooling ACFs across the numerous channels of the filter bank. More interestingly, the ACFs also contain a series of peaks at delays of the fundamental periods and its multiples of either or both vowels, depending on the relative amplitudes, frequencies, and phases of the harmonics in the channel. These F_0 -related periodicities are reinforced by pooling ACFs across the numerous channels, since they are common in many of them, giving rise to two dominant fundamental periods. The ACF of each channel is then sampled separately at the delays corresponding to the two dominant fundamental periods to determine the degree of synchrony to those periods. Two synchrony spectra are derived by plotting the degree of synchrony as a function of the filter centre frequency and are finally classified by a template-matching procedure.

The performance of place models depends on the resolution of spectral analysis. Frequency selectivity of the peripheral auditory system estimated by Moore and Glasberg (1983) is not sufficiently fine for such models to predict accurately the data on ΔF_0 effects. Using place and place-time models in linear and non-linear versions, Assmann and Summerfield (1990) confirmed that place-time models are better than place models at predicting the data, but still failed to show progressive improvement in identification with ΔF_0 . Meddis and Hewitt (1992) showed that the gradual improvement with ΔF_0 could be obtained by including a channel separation procedure, where channels were segregated into two groups on the basis of the F_0 of the first vowel only. The second vowel can be identified from all remaining channels. They discussed that this second operation is similar to applying a bias to the

dominated set before the stimulus has ceased. It may take some time, for the auditory system to apply such bias and switch to the second set of channels. As a consequence, this process might not occur for very short stimuli. This argument was therefore consistent with a smaller improvement in identification for 50-ms than 200-ms double-vowel stimuli (Assmann and summerfield, 1990, 1994, Culling and Darwin, 1993).

B.2 Mechanism of harmonic cancellation

The idea that listeners could switch from one subset of harmonics to another by applying some sort of bias towards a dominated set led to a second class of models. When listeners are asked to report the two vowels correctly, both vowels are target and both are mutually masking each other. The question arose as to whether harmonicity of the target vowel (harmonic enhancement), or the interfering vowel (harmonic cancellation) or both, underpinned the ΔF_0 effect.

Two experiments (de Cheveigné et al., 1995; Summerfield and Culling, 1992) showed that it made no difference whether a target was harmonic or inharmonic, but performance was much better if the interfering vowel was harmonic than if it was inharmonic. In a similar approach, Lea (1992) showed that a noise-excited vowel was more accurately identified than a harmonic vowel when they were presented simultaneously. The auditory system could segregate vowels by exploiting the harmonic structure of the interfering vowel to suppress it, the remaining vowel becoming more intelligible through the removal of this interfering vowel. This idea has been formalised as the harmonic-cancellation mechanism (de Cheveigné et al. (1997a, 1997b)). The harmonicity of the target voice does not play a determinative role in such a mechanism. The improvement with ΔF_0 of the identification of weak targets (SNR up to -20 dB) was consistent with such a process, since at such a SNR,

the estimation of the target's F0 is made difficult while that of the interferer is facilitated. De Cheveigné (1993, 1997c) pointed out that such a cancellation might operate very simply in the time-domain via delayed inhibition of the neural discharge pattern within each peripheral channel; the delay of the inhibition is equal to the period of the interferer. However this model has been developed in a relative vacuum of psychophysical data to constrain its form; a deficit that the experiments presented in Chap. II are intended to help rectify.

B.3 Mechanisms that do not require F0 identification

One interesting finding was made by Assmann and Summerfield (1990). They showed that listeners identify 50- and 200-ms segments of double-vowels with similar accuracy when there was no ΔF_0 , but there was a smaller improvement in performance with ΔF_0 for the 50-ms than for the 200-ms double-vowels. Assmann and Summerfield (1994) extended their results to show that some 50-ms segments of the 200-ms double vowels were identified more accurately than others. These segments had small ΔF_0 s (0.25-1 semitones) and differed from other segments in the levels of harmonics defining the first formant, which were reinforced or cancelled by waveform interactions. Waveform interactions, which are independent of any harmonic structure, result from the beating between unresolved components producing a spectral amplitude modulation which can be beneficial to a particular vowel. Culling and Darwin (1994) attempted to test this idea by creating interleaved vowels composed of the odd harmonics of one vowel and the even harmonics of the other vowel. These interleaved double-vowels produced similar waveform interactions to normal double-vowels but were designed to disrupt any mechanism that selected or rejected components on the basis of F0. The identification of interleaved double-vowels improved from 0 to $\frac{1}{2}$ a semitone, confirming that other

mechanisms than F0-identification mechanisms contribute to the $\Delta F0$ effect, typically waveform interactions. Culling and Darwin (1994) also created a computational model based upon psychophysical measurements of auditory frequency and temporal resolution of the auditory system which exploited the waveform interactions between corresponding harmonics. Their model succeeded in predicting that identification improved with $\Delta F0$. Therefore, at least part of the $\Delta F0$ benefit for vowels is due to the exploitation of changes in spectral envelope: when there is a small $\Delta F0$, a double-vowel stimulus is more or less identifiable over time. Compared to a 50-ms segment, a 200-ms segment of double-vowels is more likely to possess a brief period where the formants of a vowel are particularly well defined. Waveform interactions thus explain why the improvement in identification is smaller for 50- than for 200-ms stimuli.

When competing vowels have different F0s, corresponding harmonics of the two vowels are misaligned in frequency and corresponding fundamental periods tend to be asynchronous in time. Summerfield and Assmann (1991) investigated these two cues and found that for harmonic misalignment to be beneficial, harmonics must be well separated in frequency (i.e. for 200-Hz F0 but not 100-Hz F0). They also found that for pitch-period asynchrony to be beneficial, the onsets of the pitch periods must be well separated in time (i.e. for 50-Hz F0 but not 100-Hz F0). The spectro-temporal resolution of the auditory system is not fine enough for those two mechanisms to account for the observed improvement in performance from 0- to 1-semitone $\Delta F0$ when the F0 was around 100 Hz. Nevertheless, those mechanisms may well play a role at other F0s.

B.4 Remaining issues in the literature on double-vowels

In Culling and Darwin (1994), identification of normal vowels was better than that of interleaved vowels, specifically at or above one semitone: waveform

interactions cannot therefore explain all $\Delta F0$ effects. In addition to waveform interactions, there must be some mechanisms dealing with common harmonicity. It remains unclear, however, over what range of $\Delta F0$ s such mechanisms appear. At small $\Delta F0$ s, the spectro-temporal resolution of the auditory system may not be sufficiently fine to discriminate two sets of harmonics.

Culling and Darwin (1993) were interested in discovering which frequency region underlies the $\Delta F0$ benefit. They synthesized vowels with a $F0$ in the region of the first formant peak, which was different from the $F0$ in the region of higher formant peaks. A $\Delta F0$ in the first formant region largely accounted for the benefit. This result may partly be accounted for by aforementioned waveform interactions because beating between corresponding components very close to each other in the first formant region can make a particular vowel more identifiable. In contrast, in the region of higher formants, beating is less likely to occur since corresponding components are more distant with each other than in the first formant region. Thus the $\Delta F0$ benefit for vowels originates primarily from the first formant region, but there are probably several mechanisms simultaneously contributing to this effect. Interestingly, across-formant inconsistencies in $F0$ reduced identification for relatively large $\Delta F0$ s (2-9 semitones). In other words, an across-formant mechanism occurs at $\Delta F0$ s above 1 semitone, which can no longer be explained by waveform interactions.

C. Mechanisms responsible for $\Delta F0$ effects for speech

C.1 Mechanisms common to vowels and speech

Among all the mechanisms that have been discussed above for vowels, which ones could be involved in the $\Delta F0$ benefit for speech? Waveform interactions are likely to be a weak explanation of the $\Delta F0$ benefit for speech. Since speech has a spectral envelope that is constantly changing over time, listeners may have little

opportunity to wait for a better glimpse to identify a particular vowel before the next phoneme replaces it. The ΔF_0 benefit for speech does not substantially increase in the range of ΔF_0 s for which waveform interactions contribute to double vowel identification (from 0 to 1 semitone), consistent with waveform interactions being of little use for segregating simultaneous speech. Mechanisms of harmonic selection and/or harmonic cancellation might be responsible for part of the benefit. However, just as for vowels, it remains unclear over what range of ΔF_0 s these harmonic mechanisms may operate.

Bird and Darwin (1998) extended the results of Culling and Darwin (1993) for vowels, to examine which frequency region underlies the ΔF_0 effect for connected speech. They resynthesized speech sentences that were filtered into different bands above and below 800 Hz. Contrary to the pattern of ΔF_0 benefit for double-vowels, they found that the ΔF_0 benefit for speech increases progressively as ΔF_0 increases without asymptoting. Just as for vowels, a ΔF_0 in the frequency region below 800 Hz was necessary for the effect to occur and across-frequency inconsistencies only had an effect at large ΔF_0 s (5-10 semitones). In contrast with Culling and Darwin (1993) however, the results of Bird and Darwin (1998) are, for the reason stated above, unlikely to be produced by waveform interactions. That is to say harmonic mechanisms must primarily use the region below 800 Hz or else they are not the main cause for the ΔF_0 benefit between competing talkers.

C.2 Mechanism relevant for speech

Both harmonic selection and harmonic cancellation are simultaneous grouping mechanisms. In addition, F_0 may also be used as a sequential cue, especially in a speech segregation task. The listeners' ability to group sounds from a target voice over time may rely on the fact that its F_0 will not radically change from one short

segment of time to the next. When the F0s of two voices intersect, this sequential grouping might be disrupted (Parsons, 1976). Using the CRM design and naturally intonated speech, Darwin et al. (2003) showed that listeners benefited progressively from ΔF_0 s up to 12 semitones. Such sequential grouping by F0 is a good candidate for the gradual improvement in identification with ΔF_0 observed for speech, and not for vowels. Chap. III and V attempted to set apart the contribution of sequential grouping by F0 from that of harmonic mechanisms in a speech segregation task.

IV. Speech in rooms

A. Degradation of speech intelligibility in rooms

The perceptual effects of reverberation have been intensively investigated in the case of the transmission of a single voice in quiet, or in simple forms of constant noise (Houtgast and Steeneken, 1985). Such conditions relate, for instance, to a single talker delivering a speech to a quiet audience in a lecture room. The Speech Transmission Index (STI) is a reliable predictor of the intelligibility of speech in such conditions. The STI is based upon the idea that a transmission channel, (e.g. a reverberant room or a phone line) attenuates the amplitude modulations of speech, resulting in impaired intelligibility.

B. ΔF_0 effect in rooms

In more complex environments, like multi-talker communication in a room, the effects of room reverberation are less well known. The present thesis considers the two possibilities that reverberation affects simultaneous and sequential grouping.

B1. Effect of reverberation on harmonic mechanisms

Culling et al. (1994) explored the robustness of spatial cues and F0 cues to simulated reverberation by using a virtual-acoustic space with controlled surface absorption. The same simulation of a reverberant room was used in the present thesis. They found that the benefit of spatial separation between sources was affected by reverberation. Lavandier and Culling (2007, 2008) investigated this result further, showing that, when reverberation increased, speech intelligibility suffered primarily from the reductions in interaural coherence of the interferer at the two ears and to a smaller extent from the loss of target intelligibility (STI effect). The effect of reverberation in the binaural domain is beyond the scope of the present thesis. Culling et al. (1994) also found that the benefit of a ΔF_0 between two competing vowels was robust to reverberation when the vowels' F0s were fixed, but impaired by reverberation when combined with a modulation of F0. They reasoned that when the F0 of a complex sound is steady, its harmonic structure remains intact as delayed copies of the direct sound are added. A mechanism, dealing with the harmonic structure of such a sound, should therefore be little affected by reverberation. However, the harmonic structure of frequency-modulated sounds is distorted by reverberation since delayed parts of the reverberant sound will have different F0s from that of the direct sound. A mechanism, dealing with the harmonic structure of such a sound, could therefore be affected by reverberation. In these double-vowel experiments, the F0 modulation of both target and masker were varied together, leaving it uncertain whether this effect was due to the modulation of the target, the interferer or both. As a consequence, Culling et al.'s (1994) results could not provide direct evidence for either of the harmonic mechanisms.

Culling et al. (2003) attempted to extend the results of Culling et al. (1994) for running speech, in which the interfering talker differed from the target talker by about 10-semitones ΔF_0 and a 15% shorter vocal tract, i.e. feminizing the interfering voice. In Exp. 1, naturally intonated speech was more affected by reverberation than monotonized speech. Two possible interpretations could explain this result. First, reverberation might affect segregation by F_0 by disrupting the harmonic structure of speech (interpretation of Culling et al., 1994). Second, prosodic information conveyed by an intonation contour (variations of F_0) assists speech intelligibility, so that natural speech is more intelligible than monotonized speech in anechoic conditions. Reverberation might affect the use of this prosodic information. In order to disentangle those two interpretations, a third type of speech stimuli was created in Exp. 2, in which the F_0 pattern was inverted from the natural intonation. Such F_0 -inverted speech has as much variation of F_0 as intonated speech but was not expected to assist speech intelligibility. The results showed that intonated speech was about equally affected by reverberation as F_0 -inverted speech and slightly more affected than monotonized speech. Therefore reverberation is unlikely to disrupt the use of prosodic information and at least part of its detrimental effect concerns harmonic mechanisms. Another factor might have been involved with speech: reverberation might affect sequential grouping by F_0 (see next section B.2). The present experiments aimed at discovering whether reverberation only disrupts harmonic mechanisms or whether reverberation also affects some sequential grouping mechanisms (Chap. IV and V).

B2. Effect of reverberation on sequential grouping

By using the CRM design, Darwin and Hukin (2000b) investigated the robustness to simulated reverberation of several sequential grouping cues: interaural time difference (ITD, difference in arrival time between the two ears) which is arguably of great use for spatial release from masking, F0, vocal-tract length and prosody. An extreme difference in vocal-tract length can lead to differences in gender, which has been shown to be a very strong cue to sequential grouping (Darwin et al., 2003, Brungart et al., 2001). They found that reverberation reduced the listener's ability to use ITDs and reduced also the ability to use a fixed F0 to group sequentially the attended monotonized voice. However, when speech was naturally intonated, the benefits of F0 continuity and vocal-tract length were very resistant to reverberation. Darwin and Hukin (2000b) did not provide any explanation why reverberation had affected sequential F0-grouping for monotonized speech. Chap. V of the present thesis describes an experiment that attempted to replicate this effect and offers a potential track for future investigation.

C. Room colouration

Colouration is the spectral response of a reverberant room which may amplify some frequencies and not others. Acoustic rays reaching each ear follow a different path in the room and bring a unique spectral envelope since frequency components of each ray have received different absorption by bouncing off the walls. Reverberation produces other spectral envelope distortions originating from the reduction in modulation and the reverberant tails. Colouration results in changes in phonemic quality potentially causing listeners to misperceive words. However, there is evidence that the auditory system adapts to room colouration. When a target word is embedded

in a longer context utterance, spoken in the same room, listeners are able to judge the phonemic quality of the reverberant target word relative to that of the context. Several studies (Watkins, 1991, 2005; Kiefe and Kluender, 2008) investigated the mechanisms of perceptual compensation for spectral envelope distortions like the colouration of a reverberant room. These compensation mechanisms fall beyond the scope of the present thesis.

However, the investigation of ΔF_0 effects requires controlling for the F_0 pattern of the speech material. When using monotonized speech, the effects of colouration cease to be random (developed in Chap. IV) and could distort thresholds. As a consequence, this work could not be performed without compensating for room colouration. It follows that a real room was not an option and constrained us to simulate reverberation.

Chapter II.

EXPLORING HARMONIC CANCELLATION

I. Introduction

In speech segregation experiments, subjects identify a single target voice better when it differs in fundamental frequency (F0) from an interfering voice than when their F0s are the same (Brokx and Nootboom, 1982; Bird and Darwin, 1998). Other experiments used synthesized vowels rather than resynthesized speech (Scheffers, 1983; Summerfield and Assmann, 1991; Culling and Darwin, 1993) and found a similar effect. However, while the $\Delta F0$ benefit for connected speech increased steadily over a wide range of $\Delta F0$ s, double-vowel identification increased sharply for very small $\Delta F0$ s and largely saturated above about one semitone.

Waveform interactions can explain part of the $\Delta F0$ effects at very small $\Delta F0$ s for vowels (Culling and Darwin, 1994; Assmann and Summerfield, 1994). However these interactions cannot explain the $\Delta F0$ effects at or above 1 semitone for vowels (Culling and Darwin, 1994) and are irrelevant for speech. Two classes of mechanisms have thus been proposed to underlie the $\Delta F0$ effects in situations where waveform interactions could no longer play a role. The first class of models assumes selection of the partials of one or both sources on the basis of their common harmonicity. Since F0 is closely related to the perception of pitch, these models were based on those for pitch perception using the principles of the “harmonic sieve” (Parsons, 1976; Scheffers, 1983; Assmann and Summerfield, 1990) and of autocorrelation (Assmann and Summerfield, 1990; Meddis and Hewitt, 1992). The second class of models, known as “harmonic cancellation”, assumes cancellation of the interfering source, leaving the information about the target source in the residue from this cancellation.

De Cheveigné (1993, 1997c) pointed out that such cancellation might operate in the time-domain via delayed inhibition of the neural discharge pattern within each peripheral channel; the delay of the inhibition is equal to the period of the interferer.

Since relatively little psychophysical data have constrained the form of the cancellation mechanism, the present series of experiments aims to characterise the nature of harmonic cancellation more precisely. To this end, we needed to know the effectiveness of the mechanism at different frequencies (Exp. 2.1), its sensitivity to different degrees of inharmonicity, possibly different at different frequencies (Exp. 2.3) and to what extent different frequency regions could be processed independently (Exp. 2.4). Exp. 2.2 examined the effect of harmonic cancellation of the masker on the target, i.e. under what circumstance the target might also be cancelled.

II. Exp. 2.1 Effect of frequency region

One of the most fundamental deficiencies in our current understanding of the harmonic cancellation mechanism is that we do not know how its effectiveness varies as a function of frequency. Culling and Darwin (1993) and Rossi-Katz and Arehart (2005) found that the F0 of the two vowels could be the same above the first formant region, or even switched across the two vowels between the first and the second formants without much effect, unless large ΔF_0 s (4-9 semitones) were employed. This suggested that a mechanism restricted to low frequencies was largely responsible for the steep improvement in identification at small ΔF_0 s of 0.25-1 semitones, supplemented by a second mechanism related to across-formant grouping occurring at large ΔF_0 s. Bird and Darwin (1998) reported similar results using resynthesized connected speech that was filtered into different bands above and below 800 Hz: a ΔF_0 below 800 Hz was necessary for a benefit to occur and across-frequency

inconsistencies only had an effect at large ΔF_0 s (5-10 semitones). Since de Cheveigné (1995, 1997a) found effects attributable to harmonic cancellation from $\frac{1}{2}$ or 1 semitone ΔF_0 , it was a good candidate for the mechanism restricted to the low frequencies. So we expected the harmonic cancellation mechanism to be restricted to or at least most efficient below 800 Hz.

A. Stimuli

To test the hypothesis that harmonic cancellation, as a segregation mechanism by F_0 , is involved in the ΔF_0 effect, raises several issues. Ideally, the paradigm should present a target harmonic complex in competition with a masker harmonic complex based on a different F_0 . However, it is difficult to discriminate the contribution of harmonic cancellation in such case owing to the role played by the aforementioned beating cues. On the other hand, in an attempt to determine the effectiveness of the mechanism at different frequencies, it was necessary to limit the frequency band over which information was conveyed to the listener, so that only that band determined performance. This constraint ruled out the use of speech as an experimental stimulus, because speech is intrinsically broadband. Therefore, the present paradigm was based on the detection of a narrow (100 Hz-wide) band of noise against a masking complex tone consisting of 60 partials. Waveform interactions were similar whether or not the masker was harmonic. Target and masker sounded very different, so listeners had no ambiguity about which sound they had to detect.

A harmonic series based on a F_0 of 100 Hz was disrupted by randomly offsetting the frequencies of each partial. The range of random offsets applied to each partial was controlled. Its distribution was rectangular to preserve the rank ordering of partials (Chalikia and Bregman, 1993; Roberts and Holmes, 2006). To this end, the range of offsets could not exceed ± 50 Hz on each partial. In order to limit the degree

to which the excitation pattern of the masker varied at the target frequency, the target was always centred on a particular masker partial of fixed frequency. The rest of the masker was either mistuned or kept harmonic (Figure 2.1).

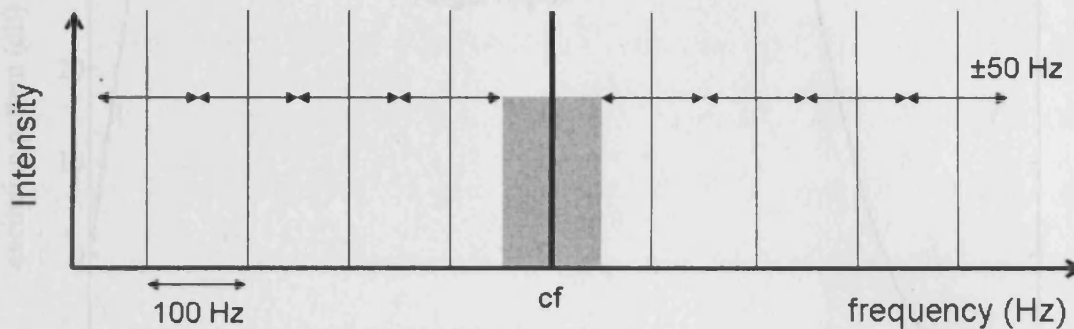


FIG. 2.1 Schematic illustration of the stimuli in Exp. 2.1. The F_0 of the complex is set to 100 Hz. The target band of noise (gray rectangle) has a width of 100 Hz and its centre frequency, cf , is varied over the experiment. The offset range of mistuned partials of the inharmonic complex is set to ± 50 Hz.

The mistuned partials had some influence on the masker's excitation pattern at the target frequency, but these influences varied randomly from trial to trial around a mean of zero and so could be neglected. On average, over a given block of trials, the harmonic and inharmonic complex maskers thus had the same excitation pattern in the frequency region of the target (Figure 2.2). As a consequence, any difference in the resulting masked detection threshold measured for the two types of masker originated from a mechanism related to the masker's harmonicity and not from a difference in target-to-masker ratio. Finally each partial was assigned a random phase to remove strong envelope modulation in the harmonic case. All stimuli were 500-ms long and gated (Hanning) by 10-ms onset and offset ramps. All stimuli were presented at 64 dB SPL.

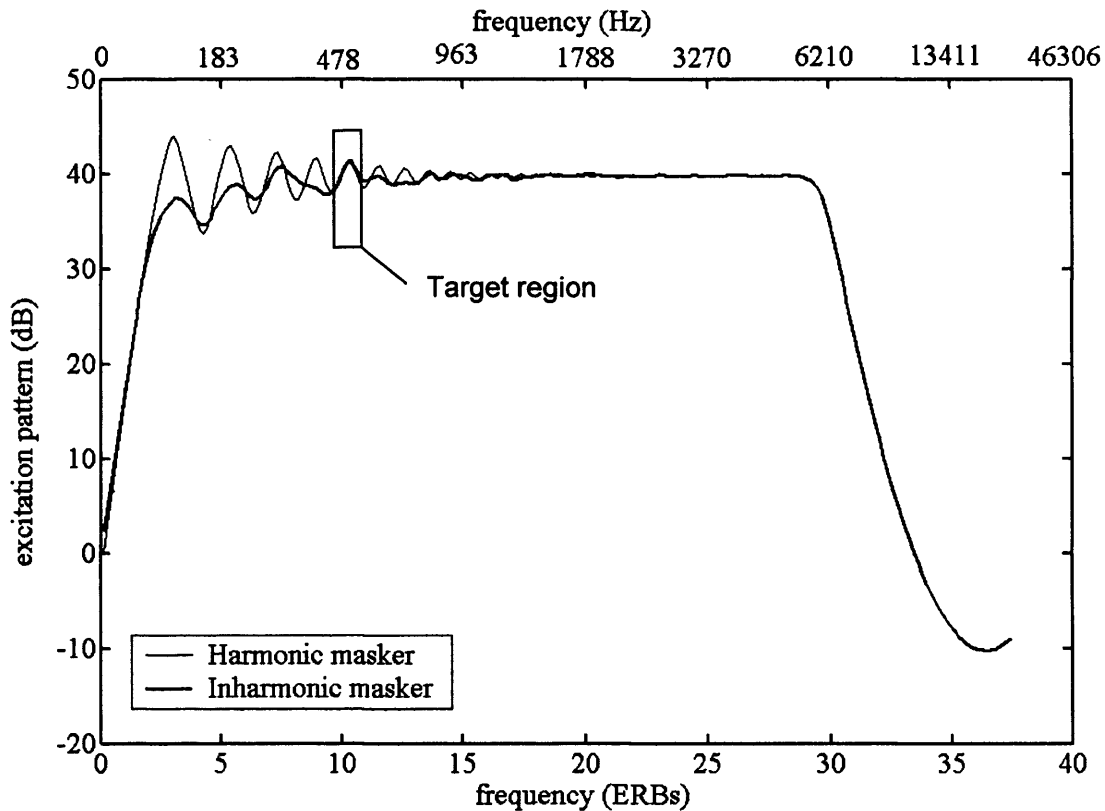


FIG. 2.2 Excitation pattern of the harmonic and inharmonic complex masker averaged over 15 stimuli. The centre frequency of the target is 500 Hz (10.3 on the ERB-rate scale), leaving the masker's partial fixed at 500 Hz whatever the harmonicity of the complex. The excitation pattern in the target region is shown to be the same with either masker.

The inharmonic maskers were generated with the maximum range of offset (± 50 Hz) to maximize the size of the expected effects. Note that there were some occasions when neighbouring partials might have been close enough together to produce low-frequency beats. Such occasions were likely to be rare and informal listening to many different tokens of the inharmonic complex did not reveal any evidence of salient beats. The target band of noise was centred at different frequencies spread over the spectrum: 100, 500, 1000, 1500, 2000, 2500, 3000, 4000, 5000 Hz. There were therefore 9 target regions \times 2 degrees of harmonicity = 18 conditions.

B. Procedure

The session consisted of 18 threshold measurements. Each measurement began with presentation of the target band of noise alone for the listener to recognize the noise to be detected during the 2-interval task. In each trial, the listener heard two intervals: one consisted of the target and the masker; the other consisted only of the masker. The listener was asked to report which interval contained the target. The same version of inharmonic complex was used for both intervals in a trial, but versions were changed between each trial in a run. Masked detection thresholds (MDTs) were measured using a 1-up/2-down adaptive threshold method. In this method, the target-to-masker ratio was initially 0 dB and each time the listener detected the target noise twice in a row, the target-to-interferer ratio decreased by 5 dB, until the listener failed to detect the target, in which case the ratio increased by 5 dB. After these two reversals, the target-to-interferer ratio varied in 2-dB steps for a further 6 reversals. Masked detection thresholds for each run were taken as the mean target-to-masker ratio derived in this way on the last six reversals. A computer monitor screen was visible outside the booth window for trial-by-trial feedback and a gamepad was inside for push-button responses. Signals were always presented diotically. They were digitally mixed, D/A converted by a 24-bit Edirol UA-20 sound card and amplified by a MTR HPA-2 Headphone Amplifier. They were presented to listeners over Sennheiser HD650 headphones in a single-walled IAC sound-attenuating booth within a sound-treated room. Five listeners each attended five one-hour sessions. The conditions were presented to each participant in a random order, counteracting possible order effects.

C. Results

Figure 2.3 presents the mean MDTs measured in Exp. 2.1. A two-factor analysis of variance (harmonicity \times centre frequency) shows a main effect of harmonicity [$F(1,4)=47.3$, $p<0.01$], namely mean MDTs were lower when the masker was harmonic than inharmonic. Mean MDTs varied as a function of the target centre frequency: main effect of centre frequency [$F(8,32)=82.6$, $p<0.001$]. The interaction was also significant [$F(8,32)=3.2$, $p<0.01$], i.e. the harmonic benefit varied as a function of the target centre frequency. Tukey pairwise comparisons indicated that the harmonic benefit was significant for the centre frequencies: 500, 1000, 1500 and 2500 Hz.

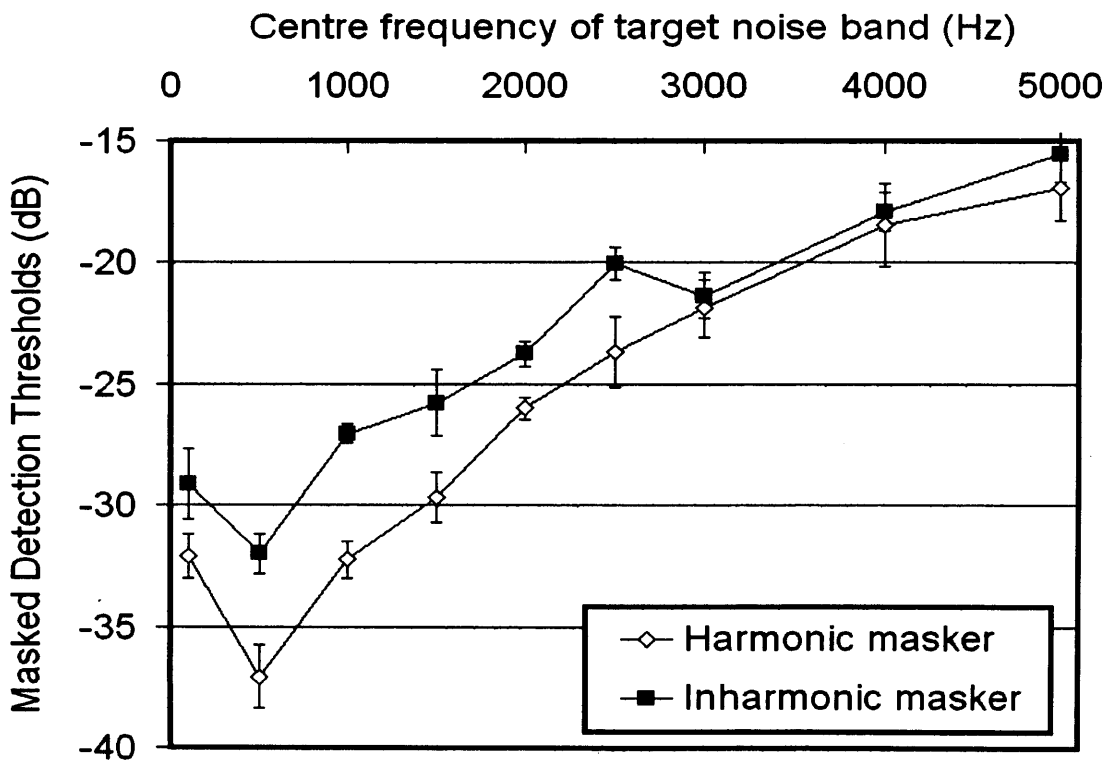


FIG. 2.3 Masked detection thresholds (dB) measured in Exp. 2.1 as a function of the centre frequency of the target band. Lower thresholds indicate better discrimination. Errors bars are ± 1 standard error of the mean.

D. Discussion

D.1. Harmonic cancellation versus stimulus uncertainty

The results of Exp. 2.1 showed that the detection of a noise band is easier when masked by a harmonic masker than an inharmonic one. This result is consistent with harmonic cancellation of the masker assisting detection of the noise band. The harmonic benefit is 3 dB for a target centre frequency of 100 Hz, about 5 dB at 0.5 and 1 kHz, and between 2 and 4 dB until 2.5 kHz. Above 3 kHz, there is no longer any harmonic benefit.

One issue must be highlighted with the present design. Compared with the harmonic masker, the versions of inharmonic masker were changed, between trials, within a run. This choice was made on one hand, to prevent the possibility of listeners learning to attend to cues peculiar to a particular randomization and on the other hand to ensure that the masking attributable to the partials either side of the noise band was, on average, the same for harmonic and inharmonic maskers. This choice might have increased the cognitive load of the task for the inharmonic case compared with the harmonic case, owing to increased stimulus uncertainty, thereby casting doubt on the grounds of the present results. The three following reasons argue that such an ambiguity did not substantially influence the results. First, the same version of inharmonic masker was used for the target-masker and single-masker intervals within a single trial, thereby minimizing the additional cognitive load. Second, the differences between the harmonic and inharmonic conditions can be as large as 5 dB. Third, comparable differences were obtained using an identification rather than detection paradigm in subsequent experimental chapters.

D.2. No benefit after 3 kHz

The sensitivity of harmonic cancellation to individual partials may be relative to the harmonic number rather than absolute frequency. Since the offset used was fixed at ± 50 Hz, the *percentage* mistunings of individual partials for high harmonic numbers is smaller than for low harmonic numbers. In other words, from 3 kHz, inharmonic partials offset at ± 50 Hz may be considered as approximately harmonic by the cancellation mechanism, resulting in no longer any difference in the MDTs between harmonic and inharmonic masker. Notwithstanding this point, phase-locking is also known to be lost above 3 kHz. The loss of temporal coding of periodicity has been thought to cause two other phenomena. First, Hartmann et al. (1990) showed that the listeners' ability to identify the pitch of a mistuned partial depends on both absolute frequency and harmonic number. They found that this ability is lost rapidly over the range 2.2 to 3 kHz. Second, Kohlrausch and Houtsma (1992) showed that the spectral edge pitch, associated with the upper edge of a flat-spectrum complex tone, largely disappears above 3 kHz. Although it is not yet clear how the loss of phase-locking would explain the present loss of harmonic benefit from 3 kHz, the parallel seems interesting to further. Other parallels between the present results and studies on the partial pitch shift are developed in the general discussion.

D.3. Unexpected frequency range for harmonic cancellation

Three studies: Culling and Darwin (1993) and Rossi-Katz and Arehart (2005) for double-vowels and Bird and Darwin (1998) for speech, found that the improvement for small ΔF_0 s largely originated from the low-frequency region and the use of common F_0 to group partials across low and high frequencies occurred for large ΔF_0 s. The present experiment however shows that harmonic cancellation is as

beneficial (if not more) above as below 800 Hz. So, if harmonic cancellation was mainly responsible for the observed $\Delta F0$ benefit, the benefit would be as large when competing voices share the same $F0$ below 800 Hz as when they share the same $F0$ above 800 Hz. When target and interferer have different $F0$ s only above 800 Hz, Culling and Darwin (1993) and Rossi-Katz and Arehart (2005) found a small $\Delta F0$ benefit and Bird and Darwin (1998) found no benefit at all, whereas the present data show that harmonic cancellation is efficient up to 2.5 kHz. Furthermore, middle and high frequencies are particularly important for speech perception (400-4000 Hz), as shown by the speech intelligibility index weighting (ANSI S3.5, 1997; Rhebergen and Versfeld, 2005). As a result, from the present data and the range of frequencies important for speech understanding, one might expect a $\Delta F0$ to be as useful above 800 Hz as below, casting serious doubts on the notion that harmonic cancellation is responsible for the effects reported by the other studies. The involvement of harmonic cancellation in the $\Delta F0$ effect is further discussed in the section VI.

D.4. Within-channel target-to-masker ratios

The difference in MDTs for the two types of masker originates from a mechanism related to the masker's harmonicity and not from a difference in target-to-masker ratio (T/M) since the excitation patterns of the two maskers were the same in the target region (Figure 2.2). However the effect of frequency on MDTs, for both harmonic and inharmonic maskers, may well result from within channel T/M ratios. The bandwidth of human cochlear filters varies with centre frequency (Moore and Glasberg, 1983). Cochlear filters at higher frequencies have a large bandwidth, where many masking partials are accepted by the filters, resulting in a reduced T/M ratio. As frequency decreases, the filters' bandwidth becomes increasingly narrow and more and more masking partials are rejected by the filters, resulting in a bigger T/M ratio.

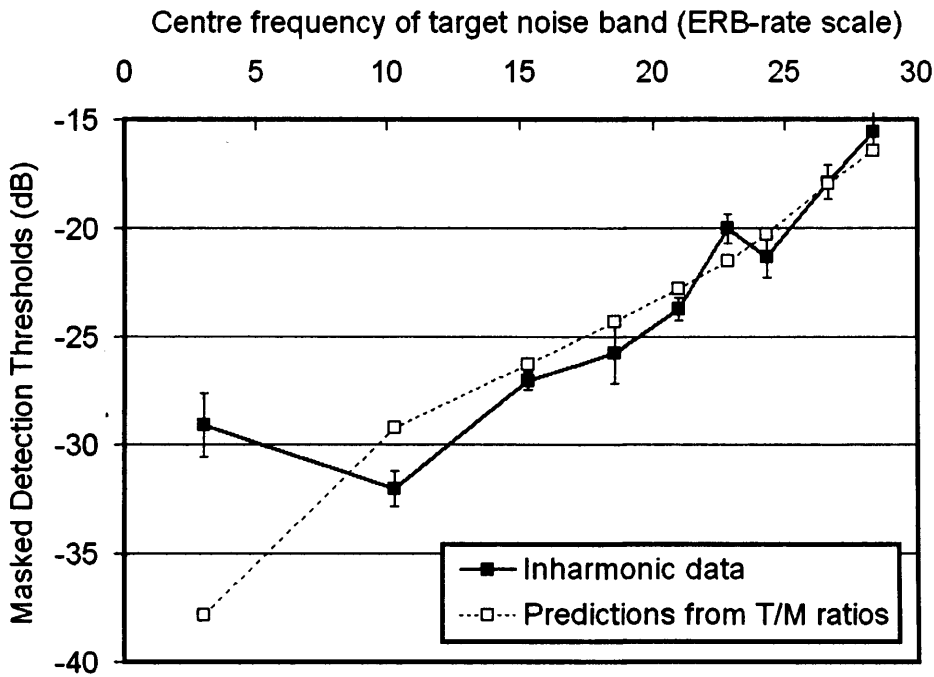


FIG. 2.4 Masked detection thresholds (dB) for the inharmonic masker measured in Exp. 2.1 (filled symbols) and predictions of masked detection thresholds (dB) from T/M ratios computed at the output of a 3-ERBs-rate range of gammatone filters around the target centre frequency (open symbols). These predictions were identical, whether the masker was harmonic or inharmonic. Since predictions were averaged over many stimuli, errors bars for the predictions were too small to be visible.

A model was used to develop the idea that the MDTs of Exp. 2.1 were dependent on these T/M ratios and attempted to predict the MDTs for the inharmonic masker, i.e. where harmonic cancellation was not involved. For each target centre frequency, target and masker signals were band-pass filtered through a gammatone filter bank consisting of 256 filters covering the whole spectrum. T/M ratios were computed at the output of each filter. The mean of the T/M ratios over ± 1.5 on an ERB-rate scale around the target centre frequency was compared to a criterion (-18 dB) chosen to fit the data but fixed across frequency. As long as the mean T/M ratio

was higher than this criterion, the target was considered as detectable and the threshold decreased. Figure 2.4 (empty symbols) shows the predictions of this algorithm for the centre frequencies used in Exp. 2.1. Plotted on an ERB-rate scale (Glasberg, and Moore, 1990), the predictions of this algorithm support the idea that MDTs are determined by target-to-masker ratios at the output of cochlear filters, located around the target region.

It is also noticeable that predicted MDTs continued to fall at 100 Hz (3 on the ERB-rate scale). The discrepancy at 100 Hz decreased the correlation coefficient between MDTs and the predictions from 0.99 to 0.87. The filter centred at 100 Hz is about 35-Hz wide, and thus contains only part of the target band, so that the T/M ratio at the target centre frequency is not as high as it is at 700 Hz where the filter is 100-Hz wide, i.e. optimised for detection of the target band. Predicted MDTs continue to fall below 700 Hz because the T/M ratios at the output of filters slightly shifted from the target centre frequency give substantially bigger T/M ratios. The data did not show a continued fall in threshold at 100 Hz, suggesting that listeners might have had difficulty exploiting the information in off-frequency channels at low frequencies. An alternative explanation is that other mechanisms are involved for the specific case, where the target band is centred on the fundamental component, which in the present experiment is at 100 Hz. In particular, the fundamental partial pitch matches the global pitch of the whole complex. The energy of the fundamental masker partial (as computed by the T/M ratio) may underestimate its ability to mask the target at that frequency because part of the masking may be caused by other phenomena. We also tested the possibility that low-frequency channels had been masked by low-frequency background noise. In this respect, our experimental auditory booth was recorded 'in silence' at the ears of a dummy head: the background noise level did not appear

sufficient to have affected the MDT. However, participants might have produced some extra noise in very low frequencies, simply by breathing or moving.

III. Exp. 2.2a and 2.2b Effect of spectral overlap

Harmonic cancellation can be seen as a comb-filter. When the target centre frequency is coincident with a masker partial, as in Exp. 2.1 represented schematically in the left part of Figure 2.5, the internal representation of the target band may be suppressed by the presence of the comb-filter, limiting the advantage produced by harmonic cancellation. When the target is located between two masker partials (right part of Figure 2.5), the internal representation of the target band would be relatively unaffected by a comb-filter and the difference in MDT for an inharmonic and harmonic masker could be larger than that observed in Exp. 2.1. The two parts of Exp. 2.2 test whether the harmonic benefit could be larger when the target band is located between two masker partials rather than coincident with one of them.

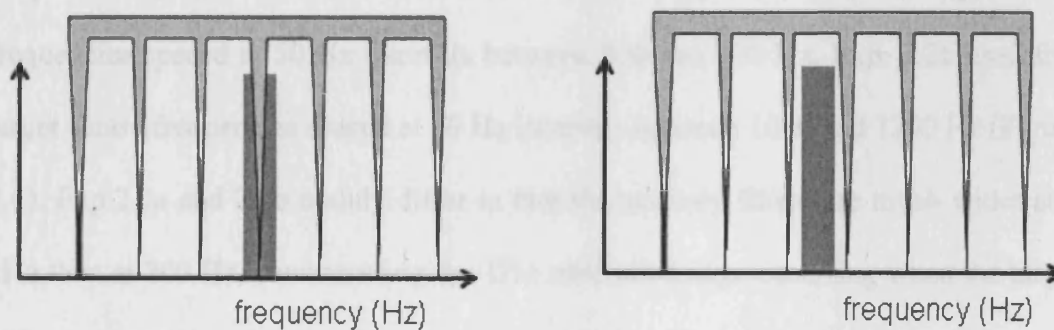


FIG. 2.5 Scheme of the comb-filtering effect of harmonic cancellation tuned to the F_0 of the masker. When the target band is centred on a masker partial, the comb-filter of the complex may affect the detection of the target (left). When the target band is located between two partials, the comb-filter no longer affects the target (right).

Regardless of the masker's harmonicity, the T/M ratios are higher for a target located between two masker partials rather than coincident with one of them, due to masking. Since detection is also a matter of T/M ratio, one would therefore expect the MDTs to drop when the target is between two masker partials. The key difference therefore, is the effect of masker harmonicity as a function of target centre frequency (i.e. an interaction).

A. Stimuli and procedure

In Exp. 2.1, the excitation pattern of the two maskers was kept the same in the target region, by fixing the masker partial coincident with the target centre frequency. In both parts of Exp. 2.2, the two partials, surrounding the target, were fixed at harmonic frequencies to hold the same masking level in the two conditions. In order to accentuate the effect of target position, the spacing between two masker partials was increased to 200 Hz, by increasing the complex's F0 to 200 Hz. The complex consisted again of 60 partials. Since the F0 of the complexes was 200 Hz, the offset range of inharmonic partials was set to ± 100 Hz. Exp. 2.2a used five target centre frequencies spaced at 50 Hz intervals between 200 and 400 Hz. Exp. 2.2b used five target centre frequencies spaced at 50 Hz intervals between 1000 and 1200 Hz (Figure 2.6). Exp. 2.2a and 2.2b mainly differ in that the auditory filters are much wider at 1 kHz than at 200 Hz, counteracting the T/M ratio advantage occurring when the target is between two masker partials. Consequently, the effect of target centre frequency might be of lesser relevance. There were therefore 5 target regions \times 2 degrees of harmonicity = 10 conditions in each experiment. Three listeners each attended six 30-min sessions in Exp. 2.2a. Five listeners each attended three 30-min sessions in Exp. 2.2b. All stimuli were presented at 64 dB SPL.

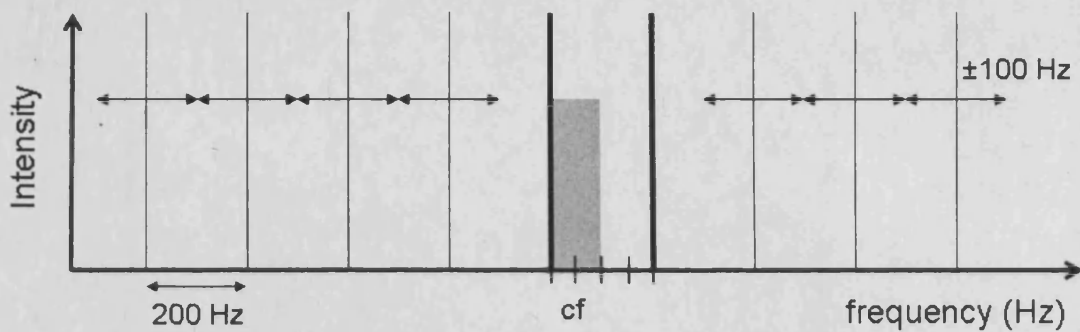


FIG. 2.6 Schematic illustration of the stimuli in Exp. 2.2b. The F0 of the complex is set to 200 Hz and the offset range of mistuned partials set to ± 100 Hz. The target band of noise (gray rectangle) has a width of 100 Hz and its centre frequency, cf, is varied over the experiment at 50 Hz intervals; here represented at 1050 Hz.

B. Results

Figure 2.7 presents the mean MDTs measured in Exp. 2.2a. A two-factor analysis of variance (harmonicity \times centre frequency) shows that mean MDTs were lower when the masker was harmonic than inharmonic: main effect of harmonicity [$F(1,2)=114.6$, $p<0.01$]. Mean MDTs varied as a function of the target centre frequency: main effect of centre frequency [$F(4,8)=7.5$, $p<0.01$], but there was no interaction [$F(4,8)=0.6$, $p>0.05$]. Tukey pairwise comparisons show that regardless of masker's harmonicity, MDTs at 300 Hz were lower than those at 200 ($q=5.2$, $p<0.05$) and 400 Hz ($q=5.9$, $p<0.05$) and MDTs at 250 Hz were lower than those at 400 Hz ($q=5.5$, $p<0.05$).

Figure 2.8 presents the mean MDTs measured in Exp. 2.2b. Mean MDTs were lower when the masker was harmonic than inharmonic: main effect of harmonicity [$F(1,4)=57.0$, $p<0.01$]. Neither mean MDTs [$F(4,16)=1.6$, $p>0.05$], nor the harmonic benefit [$F(4,16)=0.2$, $p>0.05$], significantly varied with the target centre frequency.

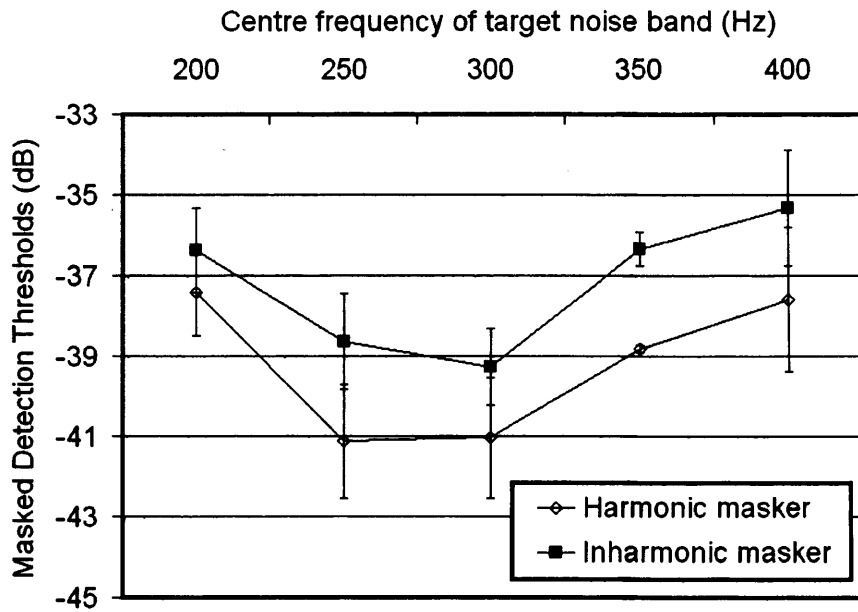


FIG. 2.7 Masked detection thresholds (dB) measured in Exp. 2.2a as a function of the centre frequency of the target band. Lower thresholds indicate better discrimination.

Errors bars are ± 1 standard error of the mean.

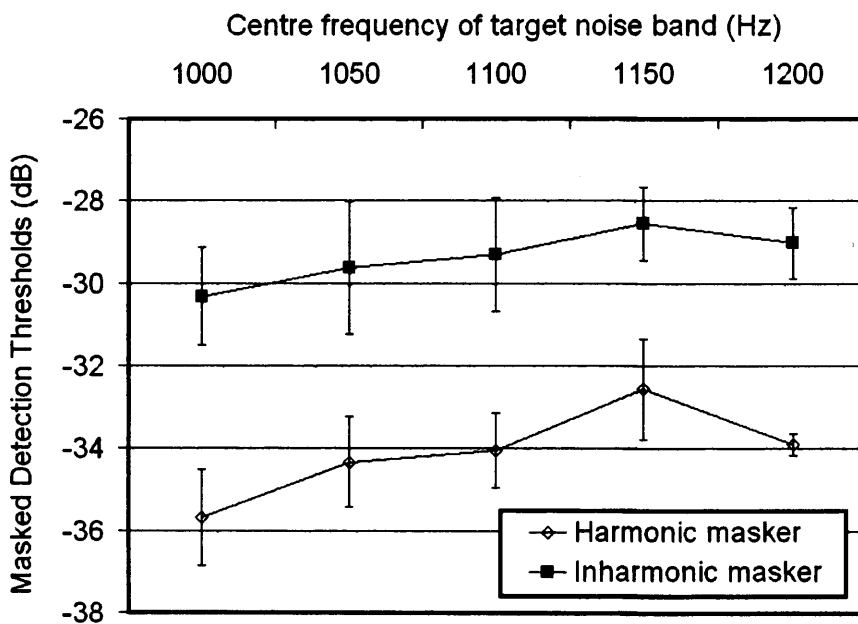


FIG. 2.8 Masked detection thresholds (dB) measured in Exp. 2.2b as a function of the centre frequency of the target band. Lower thresholds indicate better discrimination.

Errors bars are ± 1 standard error of the mean.

C. Discussion

In both these experiments, the harmonic benefit was the same whether the target was placed between two masker partials or coincident with one of them. There was no sign of an effect of harmonic cancellation on the target band. The comb-filtering of a partial located at the target centre frequency might be so narrow that its detrimental effect on target detection is negligible. The effect of harmonicity was found to be about 2 dB in the low frequency region (Exp. 2.2a) and 5 dB in the middle region (Exp. 2.2b), which was reasonably consistent with the results of Exp. 2.1. Note that in Exp. 2, two rather than one, masker partials were fixed at harmonic frequencies, shown by the thick partials in Figure 2.6. So the inharmonic complex was not as inharmonic as it was in Exp. 2.1. This change had little influence since the harmonic benefits were consistent with those of Exp. 2.1. However, all subsequent experiments were designed with the target coincident with a masker partial, since it required fixing only a single partial.

When the target falls between the 200 and 400 Hz masker partials, the cochlear filters between these two frequencies have a narrow bandwidth (Moore and Glasberg, 1983) which can reject energy from the adjacent masker partials. They consequently provide a higher T/M ratio than the filters at 200 and 400 Hz when the target is located at these frequencies. As frequency increases, filters broaden so that the advantage of placing the target between masker partials is progressively lost. Predictions of MDTs from T/M ratios were performed following the same algorithm used in Exp. 2.1, using the same criterion value, and shown by the open symbols of Figure 2.9. Figure 2.9 shows that in the 200-400 Hz region (5-9 on the ERB-rate scale), the predicted MDTs were lower when the target band was located between two masker partials than coincident with them. In the 1000-1200 Hz region (15-17 on the

ERB-rate scale), the position of the target did not have as much influence as it had in the 200-400 Hz region and predicted MDTs slightly increased with centre frequency presumably due to the widening of filters' bandwidth. In Exp. 2.2a and 2.2b combined, the correlation coefficient between MDTs and predictions was 0.99. As a conclusion, in Exp. 2.1 as in Exp. 2.2, the influence of target positions on MDTs, regardless of masker's harmonicity, is predicted by target-to-masker ratios at the output of gammatone filters around the target region.

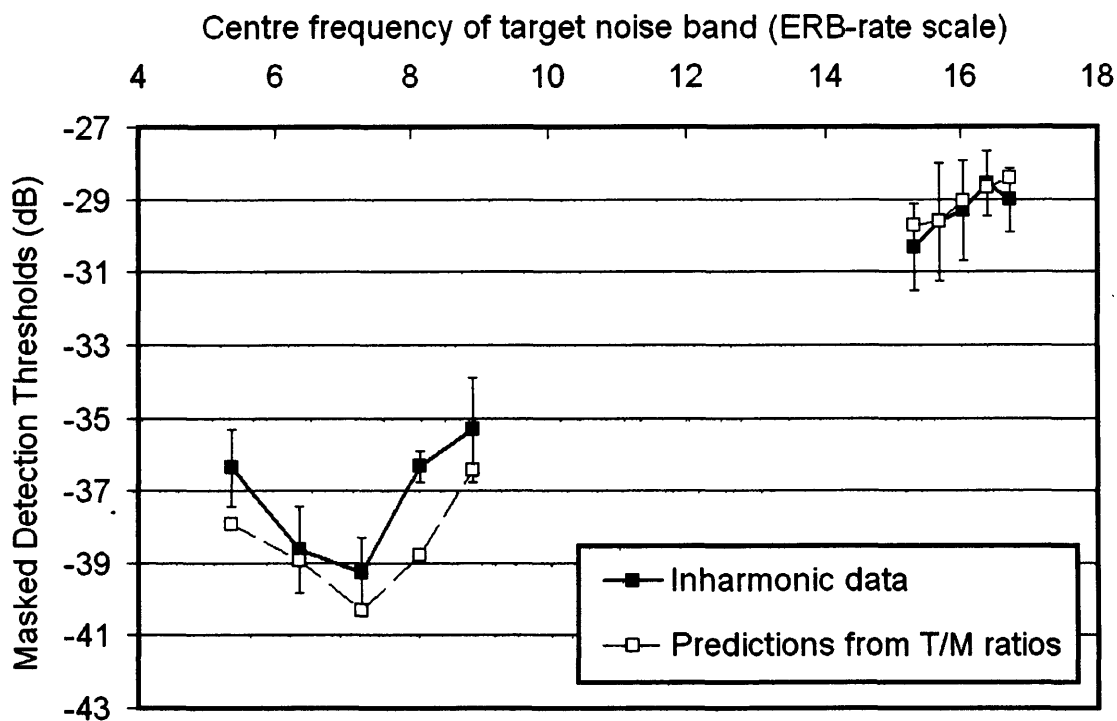


FIG. 2.9 Masked detection thresholds (dB) for the inharmonic masker measured in Exp. 2.2 (filled symbols) and predictions of masked detection thresholds (dB) from T/M ratios computed at the output of a 3-ERBs-rate range of gammatone filters around the target centre frequency (open symbols). These predictions were identical, whether the masker was harmonic or inharmonic. Since predictions were averaged over many stimuli, errors bars for the predictions were too small to be visible.

In Exp. 2.2, the spectral density of the maskers near the target band was half that in Exp. 2.1 because the maskers' F0 was increased from 100 to 200 Hz. As a consequence, the MDTs were overall shifted 3 dB lower than in Exp. 2.1; a characteristic that our algorithm also predicted.

IV. Exp. 2.3 Effect of masker's inharmonicity

How precisely harmonic must an interfering sound be for it to be cancelled by harmonic cancellation? The cancellation mechanism might be insensitive to a small degree of inharmonicity and disrupted as the masker becomes more and more inharmonic. Exp. 2.3 aimed to discover how the cancellation mechanism behaves as a function of the masker's inharmonicity.

A. Stimuli and procedure

The degree of inharmonicity can be systematically varied by altering the size of random offset applied when generating the inharmonic maskers. Six sizes of random offset were used: 0 (harmonic), ± 3 , ± 6 , ± 12 , ± 24 and ± 48 Hz, where $\pm x$ Hz refers to the maximum value in a rectangular distribution (Figure 2.10). Since the mechanism might have different sensitivity to inharmonicity at different frequencies, Exp. 2.3 used three centre frequencies 100, 1000, 2500 Hz for the target band of noise. Note that any change in sensitivity might be attributable to harmonic number rather than absolute frequency. In order to disentangle this ambiguity, further experiments would have to test different F0s. There were therefore 3 target regions \times 6 harmonicity configurations = 18 conditions. Five listeners each attended five one-hour sessions. All stimuli were presented at 64 dB SPL.

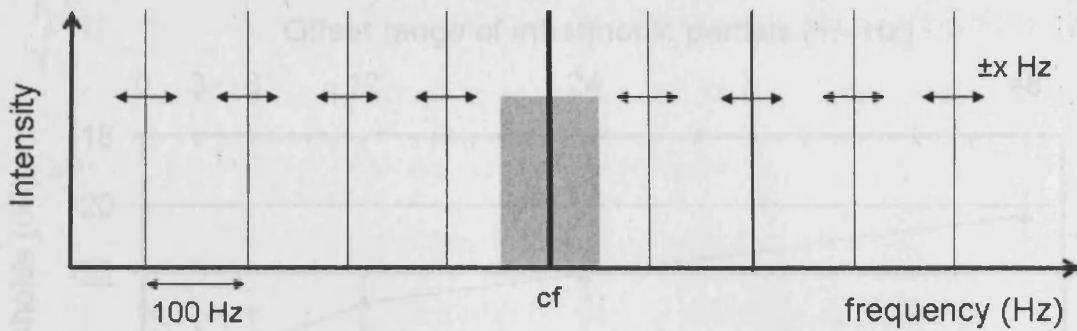


FIG. 2.10 Schematic illustration of the stimuli in Exp. 2.3. The F_0 of the complex is set to 100 Hz. The target band of noise (gray rectangle) has a width of 100 Hz and its centre frequency, cf , can be 100, 1000 or 2500 Hz. The offset range of mistuned partials of the inharmonic complex varied over the experiment.

B. Results

Figure 2.11 presents the mean SRTs measured in Exp. 2.3. A two-factor analysis of variance (centre frequency \times offset range) showed that mean MDTs varied with the target centre frequency: main effect of centre frequency [$F(2,8)=130.0$, $p<0.0001$]. Mean MDTs increased as the offset range of inharmonic partials increased: main effect of offset range [$F(5,20)=15.0$, $p<0.0001$] but there was no interaction [$F(10,40)=1.7$, $p>0.05$]. Tukey's HSD pairwise comparisons were performed between offset levels. MDTs at ± 48 Hz offset were significantly higher than at all other offsets ($q=8.3$, $q=10.7$, $q=9.4$, $q=5.9$, $q=4.7$). MDTs at ± 24 Hz offset were significantly higher than at ± 3 ($q=6.0$) and ± 6 Hz ($q=4.8$). MDTs at ± 12 Hz offset were significantly higher than at ± 3 Hz ($q=4.7$). MDTs at ± 0 , 3 and 6 Hz did not differ significantly.

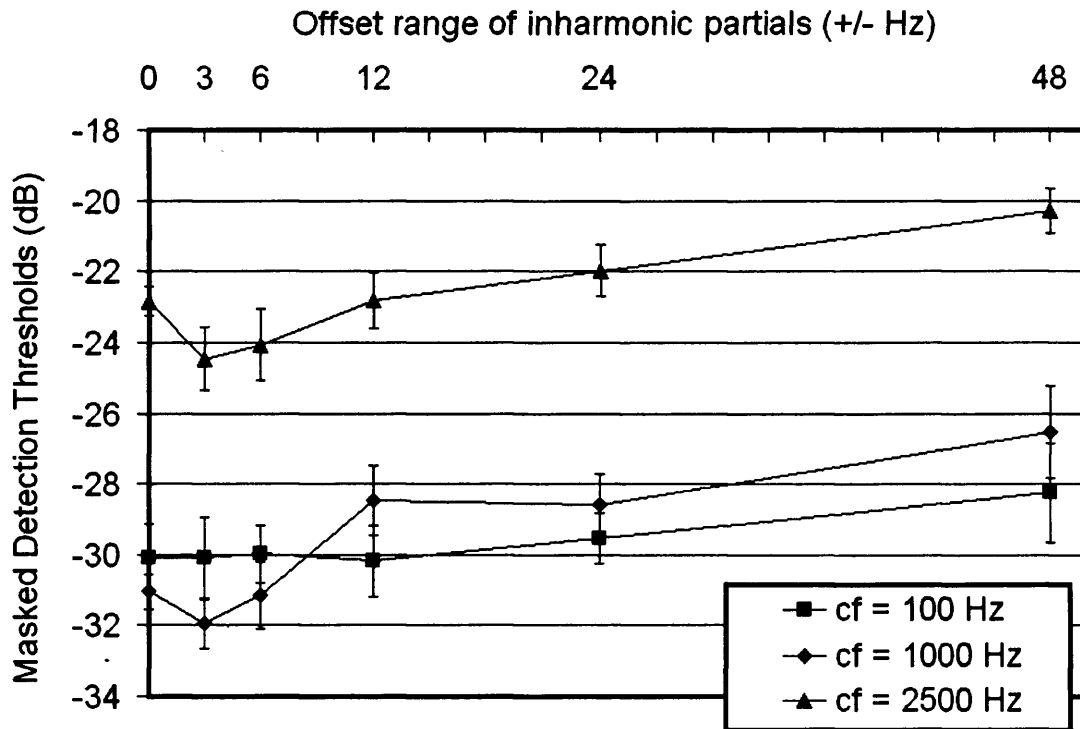


FIG. 2.11 Masked detection thresholds (dB) measured in Exp. 2.3 as a function of the offset applied to inharmonic partials in the complex masker for three target centre frequencies 100, 1000 and 2500 Hz. Lower thresholds indicate better discrimination. Errors bars are ± 1 standard error of the mean.

C. Discussion

In the present experiment, the harmonic benefit, i.e. the difference between MDTs for the purely harmonic and maximally inharmonic (± 48 Hz) complexes, was about 2 dB at 100 Hz, 5 dB at 1 kHz and 3 dB at 2.5 kHz. Those results are in agreement with the harmonic benefits found in Exp. 2.1. Following the discussion D.1 of Exp. 2.1, the extent of stimulus uncertainty over a set of runs with an inharmonic masker grows with the size of the offset range. For the reasons aforementioned and the large effects observed in the present experiment, at 1 and 2.5 kHz, the present data are likely to result from harmonic cancellation rather than an additional cognitive load.

The results of Exp. 2.3 indicated that the harmonic mechanism is insensitive to a small degree of inharmonicity, up to at least ± 6 Hz on each partial of the complex masker. The harmonicity of a complex can be disrupted in many ways. In order to compare the data across experiments, the resulting degree of harmonicity of a given complex must be assessed in a common way. Autocorrelation functions are an obvious possibility, as the largest peak at non-zero delay reflects the periodicity of a given signal. An autocorrelation peak value of one signifies that the signal is perfectly periodic, i.e. perfectly harmonic. The lower the autocorrelation peak, the less harmonic. Autocorrelation peak values were reported for each condition of Exp. 2.3, averaged across the three target centre frequencies, enabling a plot of the data as a function of this metric of harmonicity (Figure 2.12). The autocorrelation peak drops as the offset of inharmonic partials is widened. A ± 3 - and ± 6 -Hz offset give autocorrelation peak values of 0.99 and 0.98. A ± 12 -Hz offset gives an autocorrelation peak of 0.91, where MDTs significantly increased. Thus the harmonic mechanism might not cancel a complex masker whose autocorrelation peak is below about 0.9 as effectively as it cancels a purely harmonic masker. It may be interesting to relate this result to a well documented phenomenon in the literature on pitch perception: the partial-pitch shift. Such a parallel is developed in section VI.

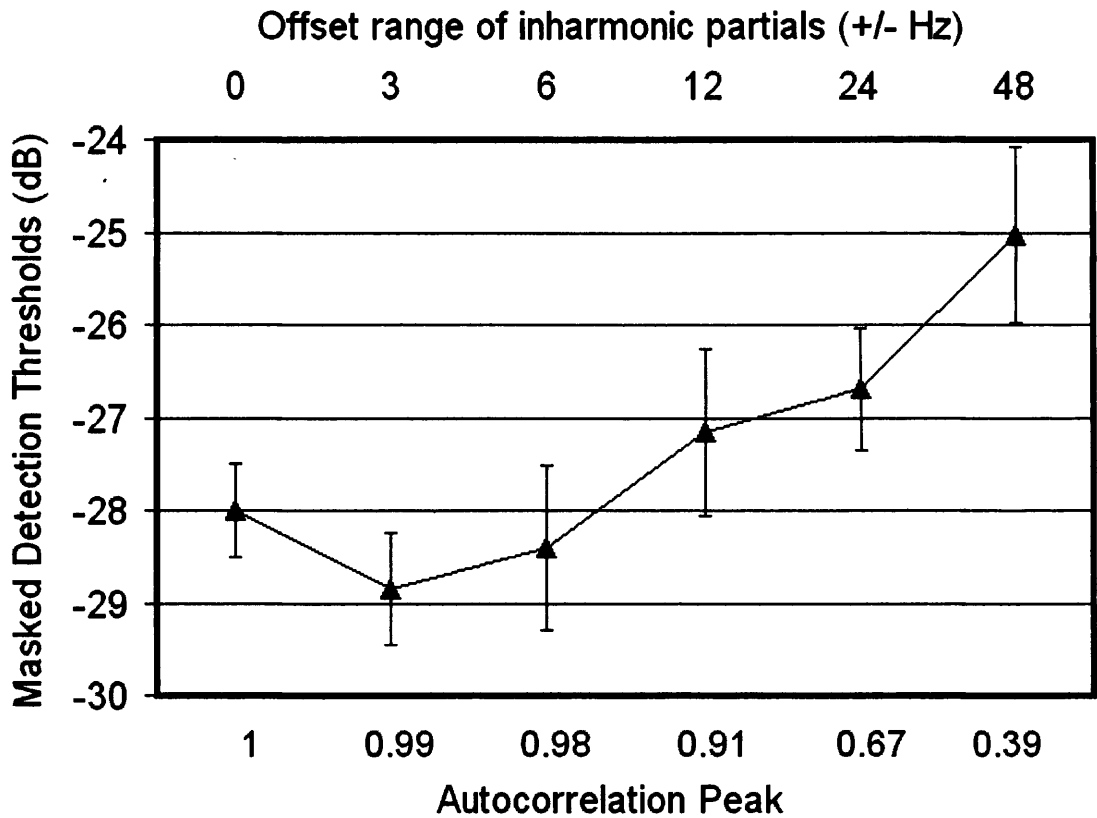


FIG. 2.12 Masked detection thresholds (dB), averaged across the three target centre frequencies, measured for each condition of the masker complexes used in Exp. 2.3, whose degree of harmonicity is reflected by the autocorrelation peak values given in the lower axis. Lower thresholds indicate better discrimination. Errors bars are ± 1 standard error of the mean of averaged data.

Autocorrelation seems a sensible tool to approach the estimation of harmonicity or inharmonicity in a given signal. The temporal window, over which the autocorrelation functions should be calculated, is however unknown. Since the harmonic or inharmonic complexes are based on a fundamental period of 10 ms, a temporal window of 12 ms has been used in the present modelling to insure that it is longer than one period of the harmonic complex. Similarly the neural comb filter of the harmonic cancellation as presented by de Cheveigné requires delay lines as long as the longest periods to cancel. There may be little physiological evidence for delay

lines that long (Meddis and Hewitt, 1991), so it is worth examining surrogates of autocorrelation. One promising alternative (Shamma and Klein, 2000; de Cheveigné and Pressnitzer, 2006) pointed out that when sounds enter the ear, vibrations of the eardrum cause pressure variations in the cochlear fluids, which in turn cause the basilar membrane to vibrate in a wave-like motion travelling from base to apex. Consequently the propagation of this travelling wave induces phase shifts of the neural response from high-to-low centre frequencies, where the pattern of the neural response can be compared for similarity. The fundamental idea is that measurement of phase delays across cochlear channels may replace absolute time intervals within a channel (autocorrelation). It remains unclear however, how this comparison might occur in the auditory system.

V. Exp. 2.4 Operational bandwidths

The fact that manipulating the harmonicity of partials, away from the target band, influences its detection threshold strongly implies an across-frequency mechanism (Roberts and Holmes, 2006). The question immediately arises of how this influence is distributed across frequency. Possibly a certain band of frequencies or a number of adjacent partials around the target band might have a predominant influence and the harmonicity of more remote partials might be of lesser relevance.

A. Stimuli and procedure

The inharmonic masker complex of Exp. 2.1 had only one partial fixed which was at the target centre frequency. In this experiment, masker complexes were generated in which different numbers of partials around the target band were kept harmonic. The frequency ranges of partials kept harmonic were ± 0 (only one partial), ± 200 , ± 400 , ± 800 , ± 1600 Hz or all 60 partials. The two extreme ranges (1 and 60

harmonic partials) produced the same complexes as those of Exp. 2.1. The offset range used in Expt. 2.4 was the same as that used in Expt. 2.1, i.e. ± 50 Hz. Since it was possible that the mechanism integrates information about harmonicity over bands of different widths at different frequencies, the present experiment used three target centre frequencies 100, 1000 and 2500 Hz. Figure 2.13 presented the design of the present experiment.

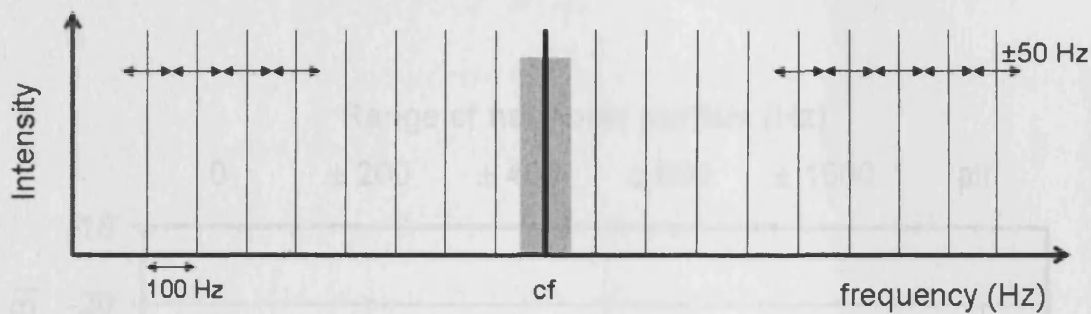


FIG. 2.13 Schematic illustration of the stimuli in Exp. 2.4. The F_0 of the complex is set to 100 Hz and the offset range of mistuned partials set to ± 50 Hz. The target band of noise (gray rectangle) has a width of 100 Hz and its centre frequency, cf , can be 100, 1000 or 2500 Hz. The range of harmonic partials, surrounding the target, varied over the experiment.

There were therefore 3 target regions \times 6 harmonicity configurations = 18 conditions. Note that there was no masker partial below 100 Hz, so that a range of harmonic partials of ± 200 Hz around a centre frequency of 100 Hz means 100-300 Hz; similarly a range of harmonic partials of ± 1600 Hz around a centre frequency of 1000 Hz means 100-2600 Hz. Five listeners each attended five one-hour sessions. All stimuli were presented at 64 dB SPL.

B. Results

Figure 2.14 presents the mean MDTs measured in Exp. 2.4. A two-factor analysis of variance (centre frequency \times harmonic range) shows that mean MDTs varied with the target centre frequency: main effect of centre frequency [$F(2,8)=236.1$, $p<0.0001$]. Mean MDTs decreased as the band of harmonic partials broadened: main effect of harmonic range [$F(5,20)=16.5$, $p<0.0001$] but there was no interaction [$F(10,40)=4.1$, $p>0.05$].

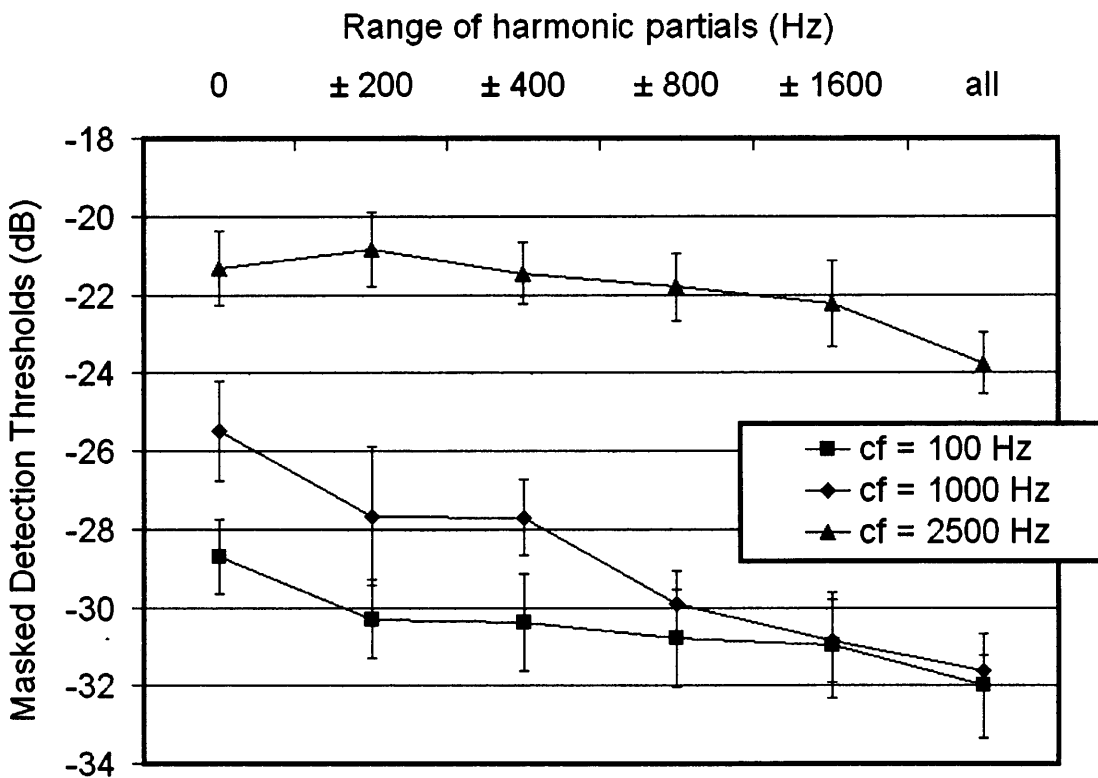


FIG. 2.14 Masked detection thresholds (dB) measured in Exp. 2.4 as a function of the bandwidth of harmonic partials in the complex masker for three target centre frequencies. Errors bars are ± 1 standard error of the mean.

Tukey's HSD pairwise comparisons revealed that mean MDTs for the purely harmonic masker were significantly lower than those for the masker with harmonic

ranges of ± 0 ($q=11.4$), ± 200 ($q=8.3$), ± 400 ($q=7.5$) and ± 800 Hz ($q=4.7$). Mean MDTs for the masker with an harmonic range of ± 1600 Hz were lower than those for the masker with an harmonic range of ± 0 ($q=8.2$) and ± 200 Hz ($q=5.0$) and mean MDTs for the masker with an harmonic range of ± 800 Hz were lower than those for the masker with an harmonic range of ± 0 Hz ($q=6.7$).

C. Discussion

In the present experiment, the harmonic benefit, i.e. the difference between MDTs for the purely harmonic and maximally inharmonic (only one harmonic partial) complexes, was about 3 dB at 100 Hz, 6 dB at 1 kHz and 3 dB at 2.5 kHz. Those results are also in agreement with the harmonic benefits found in Exp. 2.1.

Since there was no interaction with target centre frequency, the data were averaged across the three centre frequencies and plotted in Figure 2.15. MDTs decreased as the band of harmonic partials broadened. Each increase in the harmonic band decreases thresholds further, indicating that the entire spectrum is involved in the process to some extent. The purely harmonic complex gave lower MDTs than a complex containing any inharmonic partials.

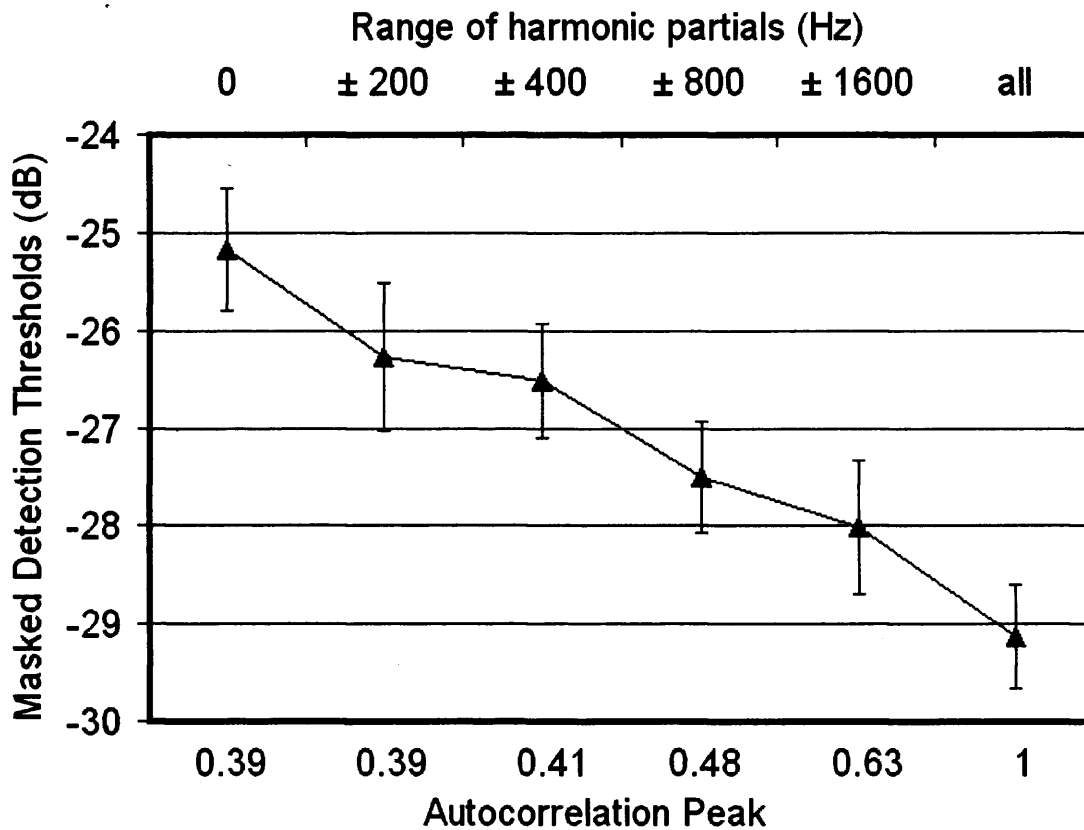


FIG. 2.15 Masked detection thresholds (dB), averaged across the three target centre frequencies, measured for each condition of the masker complexes used in Exp. 2.4, whose degree of harmonicity is reflected by the autocorrelation peak values given in the lower axis. Lower thresholds indicate better discrimination. Errors bars are ± 1 standard error of the mean of averaged data.

In order to assess whether different frequencies receive different weighting, the common metric of autocorrelation was used. The autocorrelation peak of the complex maskers of Exp. 2.4 is indicated in the lower axis of Figure 2.15. In Figure 2.16, the data of Exp. 2.3 and 2.4 are both plotted as a function of autocorrelation peak. In both remote and broadband inharmonicity cases, MDTs increased when the complexes became more inharmonic, reflected by a decrease in peak autocorrelation. However, MDTs (filled triangles of Figure 2.16) rose slowly when remote partials

became inharmonic in Exp. 2.4, owing to the harmonic benefit of the masker's cancellation in the target region. This trend suggests that nearby partials receive greater weighting than remote partials. In contrast, MDTs (empty circles of Figure 2.16) increased rapidly as the offset of inharmonic partials increased across the entire spectrum in Exp. 2.3. Thus, the cancellation mechanism seems to make a better use of the harmonicity of partials close to the target band than that of more remote partials.

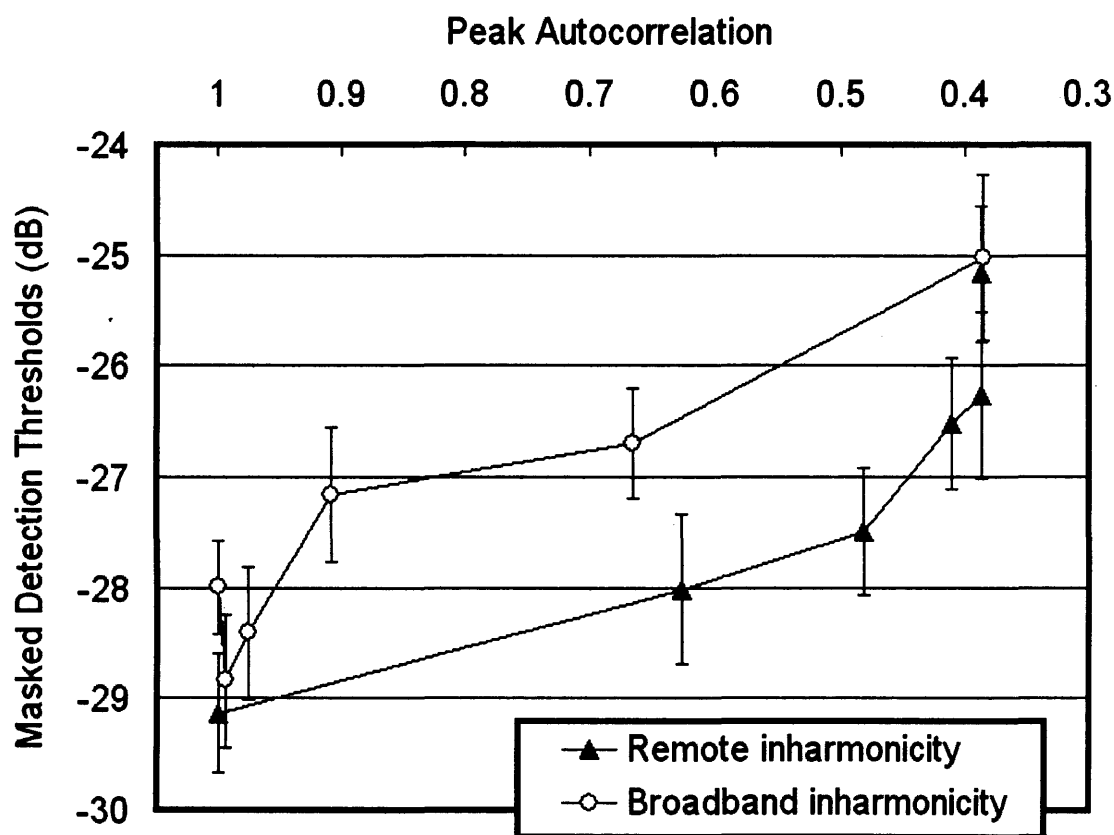


FIG. 2.16 Masked detection thresholds (dB) measured in Exp. 2.3 and 2.4, averaged across the three target centre frequencies, as a function of autocorrelation peak. In the presence of remote inharmonicity (Exp. 2.4), thresholds decreased more rapidly than did thresholds for broadband inharmonicity (Exp. 2.3), suggesting that harmonicity of close partials is more relevant than that of remote partials. Errors bars are ± 1 standard error of the mean of the averaged data.

In an attempt to determine the operational bandwidth, the complex maskers were filtered over a region around the target band, in search of a filter bandwidth at which the function relating the autocorrelation peaks to MDTs for the two types of maskers corresponded. The aim of the procedure was thus to provide a unique relation between MDTs and our metric of harmonicity, regardless of the way inharmonicity was generated. For the three centre frequencies 100, 1000 and 2500 Hz, MDTs of Exp. 2.3 and 2.4 evolved in the most similar way as a function of autocorrelation peak when the complexes were filtered by gammatone filters of 12, 10 and 12 ERBs respectively. These bandwidths correspond to about 430, 1300, and 3610 Hz. However, these values should only be taken as indicative of the approximate operational bandwidths of the cancellation mechanism across frequency, because the present experiments were not optimised for this type of analysis. Future experiments should present inharmonic complexes whose autocorrelation peak values are more regularly spaced.

VI. GENERAL DISCUSSION

A. Harmonic cancellation and the ΔF_0 effect

The present experiments provide some psychoacoustic data to characterise the mechanism of harmonic cancellation by which a target sound can be better detected through removal of harmonic interference. An important result is that harmonic cancellation is beneficial up to 2.5 kHz (Exp. 2.1). This large range of frequencies can be related to the data of three studies, for double-vowel identification (Culling and Darwin, 1993; Rossi-Katz and Arehart, 2005) and for speech (Bird and Darwin, 1998), which suggest that listeners utilize ΔF_0 cues primarily in the low-frequency region at small ΔF_0 s, but spreading to higher frequency regions at larger ΔF_0 s. To clarify how

the effective range of harmonic cancellation can fit into such a scheme, we must obtain insight into which mechanisms are involved in the ΔF_0 effects for vowels and for speech.

First, part of the ΔF_0 effect for vowels may be due to waveform interactions, which do not involve identification of F_0 s. When there is a very small ΔF_0 (0.5 semitones or less) between two vowels, corresponding low-frequency components beat, reinforcing or cancelling some parts of the combined spectrum such that each vowel can be better identified at some point in the beating cycle (Assmann and Summerfield, 1994; Culling and Darwin, 1994). Therefore, for vowels, it may not be surprising that the first formant region is largely responsible for the ΔF_0 effect since that is the region where beating primarily occurs.

Second, for ΔF_0 s larger than 0.5 semitones for vowels, the beating mechanism is irrelevant and other mechanisms must be involved. Culling and Darwin (1993) and Rossi-Katz and Arehart (2005) showed that an across-formants mechanism is involved in the ΔF_0 effect at larger ΔF_0 s (4-9 semitones). Since harmonic cancellation is efficient up to 2.5 kHz and can integrate harmonic information over the entire spectrum, it might underlie the ΔF_0 effect at larger ΔF_0 s. This interpretation also requires that the contribution of harmonic cancellation largely asymptotes from about 1 or 2 semitones, since vowels recognition performance asymptotes at these ΔF_0 s.

Third, for speech segregation, the beating mechanism may be irrelevant. The present data suggest that harmonic cancellation should only be weakly beneficial when restricted below 800 Hz compared to its entire contribution up to 2.5 kHz (Exp. 2.1). On the contrary, Bird and Darwin (1998) showed that the ΔF_0 benefits for manipulated speech, whose ΔF_0 is restricted below 800 Hz, are almost as large as for

unprocessed speech. The region below 800 Hz is also the region where harmonics are spectrally resolved in the auditory periphery and dominates the perception of pitch. Competing voices might consequently be perceived at distinct pitches as long as they have different F0s at least below 800 Hz, in which case grouping into distinct streams on the basis of the perceived pitch might still occur and largely explain the data. More interestingly, when the harmonics of two monotonized voices are swapped above 800 Hz, the pitches of both voices do not change, while most of their harmonic structure is swapped. If the cancellation mechanism is tuned to the perceived *pitch* of the masker, the harmonic structure of the target, not that of the masker, would be cancelled. This effect may be responsible for the drop in performance for large ΔF_0 s in the case of F0-swapped voices of Bird and Darwin (1998). Further research is therefore needed to understand how harmonic cancellation identifies the masker's F0 in a ΔF_0 situation between competing vowels or voices and whether the perceived pitch of the masker plays any role.

B. Parallel with the partial-pitch shift

There are reasons to believe that harmonic cancellation may belong to a more general class of cancellation by spectral templates. These reasons are developed further in the general discussion chapter (I.D). Throughout the literature on the effects of spectral templates, one effect has received a large interest among researchers: the pitch of a mistuned partial in a complex tone. When a single partial of an otherwise harmonic complex is $\pm 3\%$ mistuned, it makes a full contribution to the pitch of the complex (global pitch), i.e. is completely accepted by the template, but progressively makes a reduced contribution to the global pitch as the mistuning increases beyond $\pm 1-2\%$, up to $\pm 8\%$ where the partial is segregated from the complex and heard with its own pitch and pure-tone timbre (Moore et al., 1985; Darwin et al., 1994; Lin and

Hartmann, 1998). Hartmann et al. (1990) discovered that listeners not only perceive this mistuned partial as more salient but also perceive it systematically as more mistuned than it actually is. Such result can be observed by asking listeners to tune a pure tone to the partial pitch that is perceived: listeners exaggerate the mistuning imposed on the partial (i.e. upward pitch shifts for positive mistunings and downward shifts for negative mistunings). These mismatches are known as partial-pitch shifts. Measures across a range of mistunings have enabled to build up a pitch-shift profile for each partial tested (Hartmann and Doty, 1996). Beyond a certain degree of mistuning, which varied with partial number, these profiles typically showed saturation or super-saturation effects: the magnitude of the pitch shift peaked and then stabilized or declined. De Cheveigné (1997d) proposed that these pitch shifts could simply arise from the existence of templates. A template is a series of slots at harmonic multiples of the F0 which act to inhibit the individual pitches of in-tune partials (Brunstrom and Roberts, 1998). A mistuned partial is more likely to be segregated from the template when its frequency is represented as further away from the center of a slot. This leads to a skewed distribution in estimates of its pitch that is increasingly under-represented near the slot center. Therefore its mean pitch is displaced away from the center in both directions, resulting in pitch shifts. As the mistuning increases, the partial progressively falls outside the slot's influence. Consequently, the pitch shift saturates and then declines. Observations of pitch shifts for different partials might therefore indicate the tolerance of each slot in the template. Note that the tolerance is quite different for each slot: the fundamental component still produces a large pitch shift at $\pm 8\%$ offset.

Roberts and Holmes (2006) have had a different approach of the partial-pitch shift effect, which is more comparable to the paradigm of Exp. 2.3. They applied a

random mistuning of each partial of a complex whose F_0 was 200 Hz, with an increasing offset range, up to $\pm 40\%$ offset on each partial, and asked listeners to match the frequency of the mistuned fundamental component. The $\pm 8\%$ mistuned fundamental component was heard as 2.3% more mistuned than it was when the complex was harmonic. This result confirms that all the other in-tune partials, form a template which produces a skewed distribution of the estimates of the pitch of the fundamental, as presented above, resulting in an exaggeration of the mistuning. More interestingly, this pitch shift was only slightly affected by an inharmonic complex with $\pm 10\%$ offset on each partial, suggesting that all the other partials do not need to be exactly in-tune to form a template that has similar effect. The pitch shift almost disappeared when the complex was maximally inharmonic ($\pm 40\%$), i.e. when all the other partials are too randomly located to form any coherent spectral template. These results can be related to the results of Exp. 2.3 in that in both experiments, partials of a complex can be grouped to be perceptually suppressed, provided that the partials approximately fit a harmonic template. This grouping was found to be similarly tolerant in our noise-band detection task (up to $\pm 6\%$ offset on each partial for a 100-Hz F_0) as in their pitch matching task ($\pm 10\%$ offset on each partial for a 200-Hz F_0). Note that Roberts and Holmes (2006) also used the autocorrelation peak to estimate the degree of inharmonicity in their offset stimuli. Furthermore, they investigated the contribution of different parts of the frequency spectrum, both close to and remote from the mistuned partial. The partial adjacent to the mistuned partial made the largest contribution to the pitch shift while remote partials contributed to a smaller extent. This parallel emphasizes that common mechanisms may play a role to release from the masking of harmonic interference and to group spectral components into templates. Since single partials, at least among the first six, can be segregated out

from the complex at $\pm 2\text{-}3\%$ mistunings or even smaller (Moore et al., 1986), it is possible that the tolerance of particular slots depends on how perfectly harmonic is the rest of the template.

C. Towards a predictive model

Psychoacoustic data are still missing regarding the temporal resolution of harmonic cancellation, which is possibly different across frequency. The temporal resolution of the autocorrelation was set to 12 ms to cover at least the period of the complexes of Exp. 2.1, 3 and 4. By modulating the F0 of the complex maskers, harmonicity of the maskers is likely to be blurred once the rate of modulation exceeds the temporal resolution of the cancellation mechanism. This issue needs to be resolved in order to produce a predictive model of harmonic cancellation.

Only the voiced portions of speech are harmonic and within these portions only those where F0 moves slowly enough compared to the temporal resolution, might be cancelled by the mechanism. Once the temporal resolution of the mechanism is derived, the model must be able to discriminate parts of speech that are harmonic from parts that are inharmonic and to which degree, and thus predict how much harmonic benefit can be expected. Such a model would facilitate measurement of the contribution of harmonic cancellation in the $\Delta F0$ benefit observed with vowels or speech.

VII. CONCLUSION

Listeners were better at detecting a 100-Hz wide band of noise when it was masked by a harmonic than an inharmonic complex, while the excitation patterns of both maskers showed the same masking in the target region. Apparently, listeners were able to cancel the complex, if harmonic, to detect the target. The benefit of harmonic cancellation was 2-3 dB at 100 and 2500 Hz, about 5-6 dB at 1 kHz and nil above 3 kHz (Exp. 2.1). As frequency increases, human cochlear filters become wider; T/M ratios consequently diminish and the resulting MDTs increased.

In the 200-400 Hz region, regardless of the masker's harmonicity, MDTs decreased when the target was located between two masker partials compared to when it was coincident with one of them (Exp. 2.2a). This drop of MDTs was well predicted from T/M ratios. In the 1000-1200 Hz region, cochlear filters are too wide for T/M ratios to increase when the target is located between two partials. T/M ratios only decreased slightly with position in this frequency range and MDTs increased in correspondence. Therefore the evolution of MDTs with frequency reflected the widening of cochlear filters. Harmonic cancellation can be thought of as a comb-filter, but the data showed no indication that such a comb-filter had affected the internal representation of the target when its centre frequency coincided with a masker partial.

Autocorrelation was used to derive the degree of harmonicity in the complexes. The cancellation of the masker was substantially reduced when its harmonicity was disrupted from an autocorrelation peak of 1 to 0.9 (Exp. 2.3). The mechanism appears to integrate harmonic information over a wide band around the target centre frequency (Exp. 2.4). MDTs decreased more rapidly as a function of autocorrelation peak when

the masking partials were harmonic in the target region than away from the target. In Exp. 2.3 and 2.4, inharmonicity was generated in different ways. An attempt was made to provide a unique relation between thresholds, regardless of the way inharmonicity was generated, and autocorrelation peak within a filter of controlled bandwidth. MDTs evolved in the most similar way when the complexes were filtered over a 10 or 12 ERBs-wide region surrounding the target. Other experiments are needed to determine more accurately the width of such operational bands of harmonic cancellation.

Harmonic cancellation is thought to be involved in the benefit of a ΔF_0 between two competing vowels or two competing voices. Culling and Darwin (1993), Rossi-Katz and Arehart (2005) and Bird and Darwin (1998) showed that the low-frequency region (first formant or below 800 Hz) is largely responsible for this benefit. These similar results may have occurred for different reasons however. Beating between low-frequency components may explain the effect at small ΔF_0 s for vowels. For speech, the region below 800 Hz is not the region over which harmonic cancellation works best, but the region where harmonics are spectrally resolved in the auditory periphery, dominant for pitch perception. For speech segregation particularly, other mechanisms than harmonic cancellation, are likely to contribute to the ΔF_0 effect possibly dependent on the pitch of the competing voices that are perceived by the listeners.

Chapter III.

A ROLE FOR GROUPING BY F0

I. Introduction

The findings discussed in the previous chapter have improved our understanding of the mechanism of harmonic cancellation. We believe this mechanism to be involved in the $\Delta F0$ benefit for vowels and for speech. Throughout the literature, the effect of $\Delta F0$ has also been reported as the listener's ability to track the target's F0, in order to direct attention to the pitch of the target voice. The latter approach traditionally relates to the problem of "grouping" in which the auditory system is thought to group together sets of sound elements that come from the same sound source. A $\Delta F0$ is a potential cue to group elements originating from different sources into different auditory objects. However, it is still not understood how such a grouping mechanism might work. The aim of the present chapter was to know whether sequential F0-grouping might play a role in F0-segregation of speech or whether harmonic cancellation is the main cause of segregation.

With vowels as with speech, it is difficult to assess the contribution of harmonic cancellation because other mechanisms contribute to the $\Delta F0$ effect: waveform interactions at very small $\Delta F0$ s for competing vowels (Assmann and Summerfield, 1990, 1994; Culling and Darwin, 1994) and possibly sequential F0-grouping for competing voices. Therefore the choice of experimental stimuli was crucial to disentangle all the mechanisms involved. On one hand, the use of speech was expected to remove any confound associated with waveform interactions. On the other hand, two types of interferer were used, speech and pulse-train (non-speech), to disentangle the contribution of harmonic cancellation (harmonic benefit) from the

contribution of sequential F0-grouping in the $\Delta F0$ benefit observed experimentally. The working hypothesis was that speech and pulse-train noise were sufficiently different to be sequentially grouped separately, leaving no scope for confusion. Humans are very good at grouping and can group via very subtle cues like vocal-tract length, sentence stress or prosody. So there was little doubt that they can sequentially group speech sounds from periodic noise without the need of $\Delta F0$. Under this assumption, the benefit of a $\Delta F0$ between speech and a pulse-train noise is entirely attributable to harmonic cancellation (harmonic benefit) while the benefit of a $\Delta F0$ between competing is attributable to both harmonic cancellation and sequential F0-grouping. Typically, sequential F0-grouping occurs with speech interferers if the $\Delta F0$ benefit observed with speech interferers is larger than that observed with a pulse-train interferer, i.e. larger than the harmonic benefit. Note that two voices were used as speech interferers, in order to maximize confusion between the talkers so that listeners may need even more to use sequential F0-grouping. Nevertheless, these two voices had the same F0 pattern such that as far as harmonic cancellation is concerned, there is still a single F0 to deal with. Following this scheme, the first experiment aimed at separating the contribution of the two mechanisms for an 8-semitones $\Delta F0$. The second experiment used the same design but with 2-semitones $\Delta F0$.

Brokx and Nootboom (1982) and Bird and Darwin (1998) used monotonized speech and found that the improvement in performance (reporting target words) increased as $\Delta F0$ increased. There are two reasons for this progressive improvement: either the sequential F0-grouping or the harmonic cancellation or both became more beneficial as $\Delta F0$ increased. The comparison of the present benefits for a 2- to 8-semitones $\Delta F0$ gave some insight into the reason why the benefit increased as $\Delta F0$ increased.

II. Exp. 3.1 Benefit of a large ΔF_0 , for buzz and speech interferers

The first explanation of why performance increased progressively as ΔF_0 increased originated from sequential F_0 -grouping: the bigger the ΔF_0 , the easier it is for the listeners to perceive that target and interfering messages are spoken at a different pitch and focus on the target one. Darwin et al. (2003) examined the use of ΔF_0 for sequential grouping and found a gradual improvement in performance from 2 to 12 semitones with real speech. Therefore sequential F_0 -grouping was a good candidate for the gradual improvement in the ΔF_0 benefit.

The first experiment used monotonized speech to measure the ΔF_0 benefit found with speech and non-speech periodic interferers. If listeners do not group sequentially by F_0 with speech interferers, the benefit is expected to be similar with both types of interferer. If they do, a larger benefit is expected with speech interferers.

A. Stimuli

The corpus of target sentences comes from the Harvard Sentence List (IEEE, 1969). The recordings of voice DA, made at M.I.T. and digitized at 20 kHz with 16-bit quantization, were used as the basis of all target stimuli. The sentences have low predictability and each has five keywords which we highlight with capitals. For instance, one sentence used in the current experiment was “the PEARL was WORN in a THIN SILVER RING”. The sentences were manipulated using the Praat PSOLA speech analysis and resynthesis package, which calculated the F_0 contour for each sentence and resynthesized the sentence with a specified F_0 throughout.

Creation of a buzz interferer

The non-speech interferers were created by synthesizing harmonic complex sounds at a 20-kHz sampling rate at 110 Hz F_0 by sine-wave summation. Such a

manipulation resulted in a pulse-train. In order to increase the similarity between this pulse-train and the speech interferers, their average spectrum was made similar. Sixteen sentences were used in the present experiment, so the pulse-train was filtered with a linear-phase FIR filter designed to match the average excitation pattern of 16 monotonized sentences. This manipulation resulted in a pulse-train with an average speech spectrum, which I denote buzz. Note that in subsequent experiments using F0 modulation, another filter was used to match the average excitation pattern of 16 F0-modulated sentences. Such a buzz has constant amplitude.

Creation of a speech-modulated buzz interferer

In order to further increase the similarity between the buzz and the speech interferers, their long-term envelopes was made similar. The amplitude envelopes of the speech interferers were extracted by half-wave rectification and low-pass filtering before they were applied to the buzz. This manipulation resulted in speech-modulated buzz interferers. The dip-listening effect thereby occurred similarly with speech as with speech-modulated buzz. Two sentences were used as speech interferers. The buzz received the same temporal envelopes used for the 2-voice speech interferers. Hawley et al (2004) showed that with 2-voice speech interferers, dip-listening was almost abolished, so in this experiment dip-listening was expected to be weak.

Design

Exp. 3.1 used two types of interferer (2-voice speech-modulated buzz versus 2-voice speech) and two values of ΔF_0 (0 or 8 semitones) in all combinations. Two SRTs were measured for each condition, requiring eight blocks of ten sentences. Eighty target sentences were used, each one monotonized at two F0s of 110 Hz and 174.6 Hz giving 2×80 target stimuli. Eight interfering sentences were monotonized

at a mean F0 of 110 Hz, and then added by pair to create four 2-voice speech interferers. The buzz received the four temporal envelopes of the 2-voice speech interferers, resulting in four different speech-modulated buzz stimuli. All stimuli were presented diotically at 69 dB SPL (Appendix A).

B. Procedure

The session began with three practice runs using monaurally presented and unprocessed speech, in order to familiarize the listeners with the task. The following eight runs measured two SRTs which were averaged to give one mean SRT for each of the four different conditions. While each of the 80 target sentences was presented to every participant in the same order, the order of the conditions was rotated for successive listeners, to counterbalance effects of order and materials. Thus, all sentences contributed equally to each condition. To avoid any priming effect, no sentence was presented twice to a participant within an experiment. For the same reason, each participant could only sign up once. Since all further experiments used the same speech material, only new participants were accepted. In Exp. 3.1, eight listeners each attended a single 50-min session, resulting in two complete rotations of the conditions.

SRTs were measured using a 1-up/1-down adaptative threshold method (Plomp and Mimpen, 1979). In this method, an individual SRT measurement is made by presenting ten target sentences one after another, each one against the same interfering sentence. The target-to-interferer ratio was initially very low (-32 dB) and in the initial phase, listeners had the opportunity to listen to the first sentence a number of times, each time with a 4 dB increased target-to-interferer ratio. Listeners were instructed to attempt to type a transcript of the first sentence using a computer terminal, visible outside the booth window, when they believed that they could first

hear about half the words of the target sentence. The correct transcript was then displayed on the computer terminal, with five key words in capitals, and the listener self-marked how many key words he or she got correct. Tight scoring of keywords was used, as opposed to loose scoring (Foster et al., 1993). Subsequent target sentences were presented only once and self-marked in a similar manner; the level of the target speech was decreased by 2 dB if the listener had correctly identified three or more of the five key words or else increased by 2 dB. SRTs for a given run were taken as the mean signal-to-interferer ratio derived in this way on the last eight trials. Each run used a different interferer. Signals were digitally mixed, D/A converted using a 24-bit Edirol UA-20 sound card and amplified by a MTR HPA-2 Headphone Amplifier. They were presented to listeners over Sennheiser HD650 headphones in a single-walled IAC sound-attenuating booth within a sound-treated room.

C. Results

Figure 3.1 presents the mean SRTs measured in Exp. 3.1. A two-factor analysis of variance ($\Delta F_0 \times$ interferer type) shows that mean SRTs were lower when target and interferer had different F_0 s than when they were on the same F_0 : main effect of ΔF_0 [$F(1,7)=171.5$, $p<0.0001$]. Mean SRTs were lower with a 2-voice speech-modulated buzz than with the 2-voice speech interferer: main effect of interferer type [$F(1,7)=46.6$, $p<0.0001$]. The benefit of a ΔF_0 was significantly greater when speech interferer was used rather than speech-modulated buzz: interaction [$F(1,7)=10.8$, $p<0.05$].

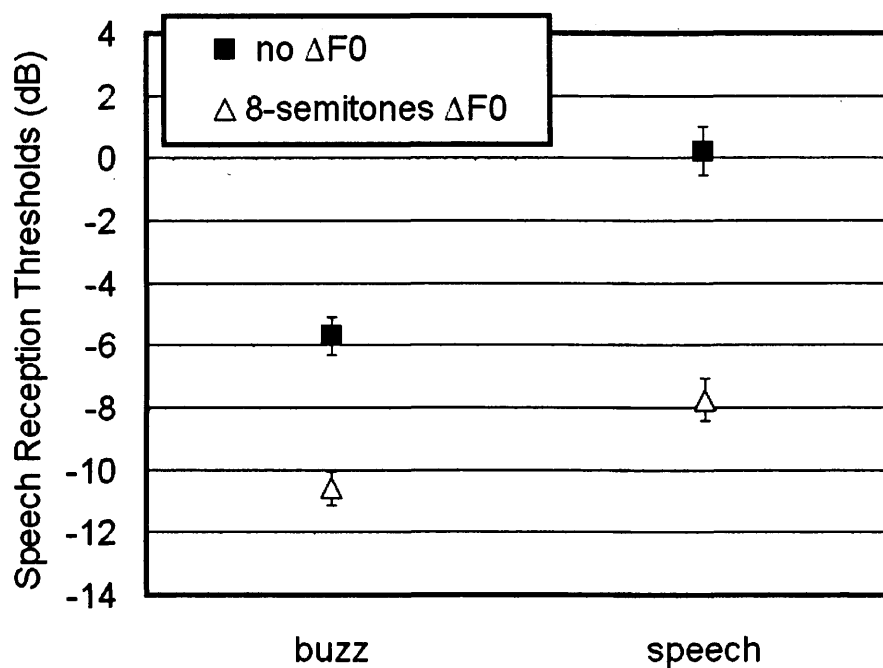


FIG 3.1 Mean speech-reception thresholds for two types of interferer (2-voice speech-modulated buzz or 2-voice speech) separated by a $\Delta F0$ of 0 (squares) or 8 (triangles) semitones. Lower thresholds indicate greater intelligibility. Error bars are ± 1 standard error of the mean.

D. Discussion

With no $\Delta F0$, mean SRT was 6 dB higher with the speech interferer than with the buzz. Presumably, listeners confused the target voice with the competing voices, whereas there was no such confusion with competing buzz. With the buzz interferer, listeners got a 5-dB benefit from an 8-semitones $\Delta F0$, which may result from the harmonic cancellation mechanism, but with a speech interferer, they got an 8-dB benefit. There was therefore 3 dB of additional benefit with speech interferers. When three talkers spoke simultaneously, listeners benefitted from a cue that could help them to differentiate the target sentence: when available, they utilized an 8-semitones $\Delta F0$ to group sounds originating from the higher-pitched voice.

It is important to highlight that the reason why the ΔF_0 benefit can be bigger with speech than non-speech interferers is that there is informational masking to release from in the first case. However, at 8-semitones ΔF_0 , intelligibility is still better when a target voice is masked by non-speech interferers than masked by other voices (Figure 3.1). Our assumption, in which speech and buzz are sufficiently different to be segregated without the need of ΔF_0 , also implies that at best the benefit of sequential grouping by F_0 offsets the informational masking caused by the presence of multiple voices. In other words, the SRT of a target voice masked by other voices cannot be lower than the SRT of the same voice masked by a buzz interferer.

III. Exp. 3.2 Benefit of a small ΔF_0 , for buzz and speech interferers

Exp. 3.1 showed that an 8-semitones ΔF_0 was large enough for listeners to get a 3-dB benefit specifically attributable to sequential F_0 -grouping. The question immediately arises of the size of ΔF_0 required for listeners to benefit from sequential F_0 -grouping. The following experiment used the same experimental design with a 2-semitones ΔF_0 to test whether F_0 -grouping appears with this lower value of ΔF_0 . An alternative explanation of the progressive increase in performance with ΔF_0 is that harmonic cancellation becomes more beneficial with larger ΔF_0 s. If so, a smaller benefit should be observed with a buzz interferer compared to that of Exp. 3.1.

A. Stimuli and procedure

Exp. 3.2 used the same two types of interferers as Exp. 3.1 and two values of ΔF_0 (0 or 2 semitones). Twelve new listeners each attended a single 50-min session. The equipment was otherwise identical to Exp. 3.1. All stimuli were presented at 69 dB SPL.

B. Results

Figure 3.2 presents the mean SRTs measured in Exp. 3.2. A two-factor analysis of variance ($\Delta F0 \times$ interferer type) shows that mean SRTs were lower when target and interferer had different $F0$ s than when they were on the same $F0$: main effect of $\Delta F0$ [$F(1,11)=67.4$, $p<0.0001$]. Mean SRTs were lower with a 2-voice speech-modulated buzz than with the 2-voice speech interferer: main effect of interferer type [$F(1,11)=135.5$, $p<0.0001$]. The benefit of a $\Delta F0$ was significantly smaller when a speech interferer was used rather than a speech-modulated buzz: interaction [$F(1,11)=5.5$, $p<0.05$].

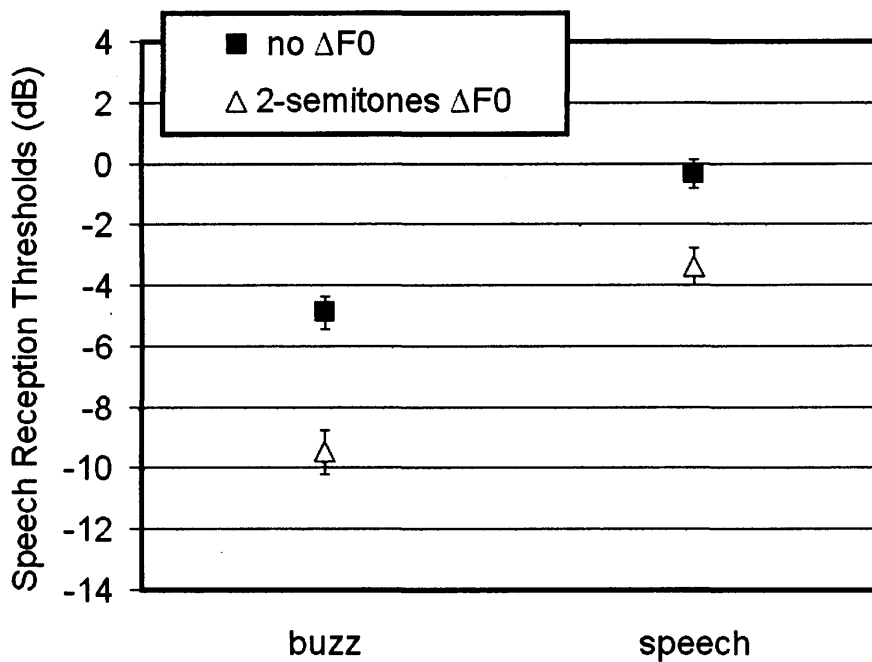


FIG 3.2 Mean speech-reception thresholds for two types of interferer (2-voice speech-modulated buzz or 2-voice speech) separated by a $\Delta F0$ of 0 (squares) or 2 (triangles) semitones. Lower thresholds indicate greater intelligibility. Error bars are ± 1 standard error of the mean.

C. Discussion

C.1 ΔF_0 benefit smaller with speech interferers than with buzz

The results of Exp. 3.2 replicated the finding of Exp. 3.1 that SRTs are higher when the target voice is masked by other voices than masked by a speech-modulated buzz. But contrary to Exp. 3.1, when three talkers spoke simultaneously, it was hard for the listeners to group sounds originating from the 2-semitones higher-pitched voice to differentiate the target sentence. Surprisingly, the benefit with speech interferers is significantly smaller than that observed with speech-modulated buzz. Had no sequential F_0 -grouping occurred at a 2-semitones ΔF_0 , one might expect the ΔF_0 benefit for speech to be of the same size as that observed for buzz. One possibility stems from the intermittent voicing of speech interferers compared to the continuous F_0 of speech-modulated buzz. In speech, there are many parts of a sentence which are unvoiced: all unvoiced fricatives and stop closures do not have a fundamental frequency. In contrast, the speech-modulated buzz has a temporally uninterrupted F_0 , except gaps due to the amplitude envelope which were scarce since the envelope chosen was that of a 2-voice speech. A temporally uninterrupted F_0 could be more completely suppressed by the harmonic cancellation than an interrupted F_0 . Thus the unvoiced portions of speech could reduce the overall benefit of harmonic cancellation.

C.2 Temporal continuity of the interferer's F_0 : evidence of harmonic cancellation

In the present experiments, both target and interferers were harmonic and so both strategies, harmonic cancellation and harmonic enhancement, could have been beneficial. However, the fact that the benefit is reduced as the temporal gaps in the

interferer's F0 increased, indicates that the mechanism responsible for the $\Delta F0$ benefit is dependent on parameters of the interferer, not the target. A strategy based on harmonic enhancement would predict the same benefit regardless of whether speech or speech-modulated buzz was used.

A test of harmonic enhancement versus harmonic cancellation would consist at manipulating the temporal continuity of respectively target and interfering speech and measuring the $\Delta F0$ benefit. This can be done by presenting two sets of sentences categorized according to their voiced/unvoiced ratio. Bird and Darwin (1998) used entirely voiced sentences which had no-stop consonants, like "a normal animal will run away" and found bigger $\Delta F0$ effects than did Brokx and Nootboom (1982) using more normal speech. So the temporal continuity of F0 may indeed have an effect on the size of the resulting benefit. Unfortunately, Bird and Darwin used continuously-voiced speech for both target and interfering speech, which does not enable us to discriminate between harmonic enhancement and cancellation.

IV. General Discussion

A. Why does performance increase as $\Delta F0$ increases in speech segregation?

There are two possible reasons why performance increases with $\Delta F0$ for speech interferers. The first reason is that as $\Delta F0$ increases, it becomes progressively easier to perceive that competing talkers speak at distinct pitches so the sequential F0-grouping mechanism becomes more beneficial. The second reason relates to an improvement in the efficiency of harmonic cancellation. From 2 to 8-semitones $\Delta F0$, Figure 3.3 showed that the benefit increased by only 0.3 dB for a buzz interferer, but by 5 dB for speech interferers. Therefore the present experiments suggested that the first explanation is dominant. Meanwhile one must be cautious with across-

experiments comparison. A proper way of designing this experiment would consist of measuring the benefit for non-speech and speech interferers for several values of $\Delta F0$ and observing over which range the increase in the respective benefits occurs.

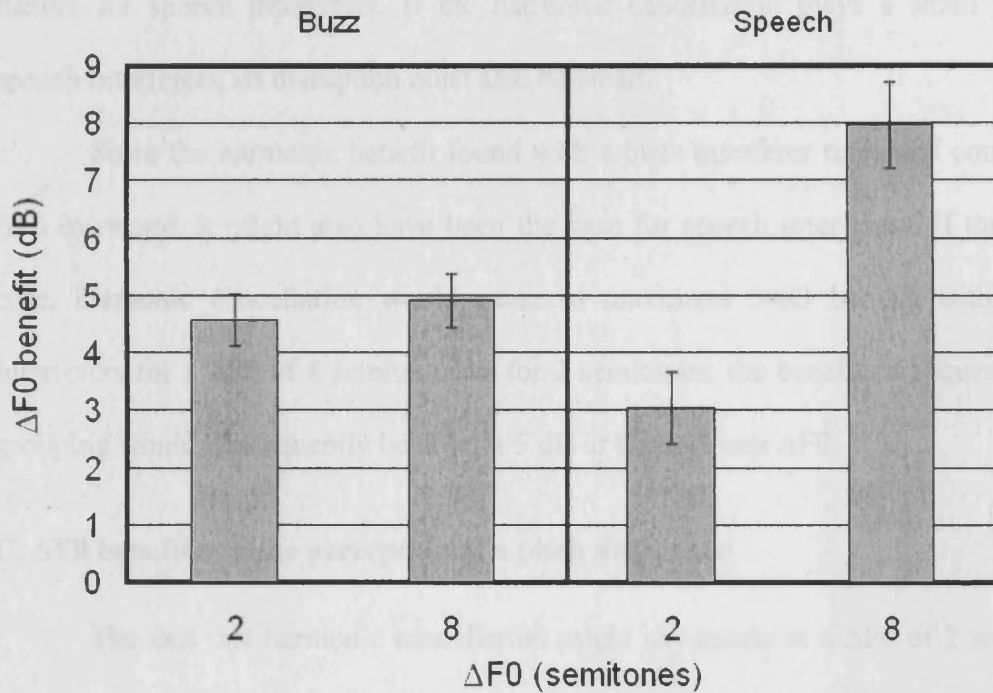


FIG 3.3 Mean benefits of a 2- and 8-semitones $\Delta F0$ for two types of interferer (2-voice speech-modulated buzz or 2-voice speech), derived from the speech-reception thresholds obtained in Exp. 3.1 and 3.2. Error bars are ± 1 standard error of the mean.

B. Small contribution of harmonic cancellation with speech interferers

The potential benefit of harmonic cancellation was smaller with speech interferers than with speech-modulated buzz, presumably due to the breaks in the voicing of speech interferers. The design of these experiments failed to isolate the contribution of harmonic cancellation in the $\Delta F0$ benefit measured with speech, because sequential $F0$ -grouping could have played a small role even with a 2-

semitones $\Delta F0$. Another way to assess the role of harmonic cancellation with speech interferers was to measure its loss, i.e. the increase in SRTs occurring when harmonic cancellation is disrupted. Chap. IV investigated which factors disrupted the cancellation mechanism for buzz interferers and Chap. V looked at those detrimental factors for speech interferers. If the harmonic cancellation plays a small role for speech interferers, its disruption must also be small.

Since the harmonic benefit found with a buzz interferer remained constant as $\Delta F0$ increased, it might also have been the case for speech interferers. If that is the case, harmonic cancellation would cause at maximum 3-dB benefit with speech interferers for a $\Delta F0$ of 8 semitones as for 2 semitones; the benefit of sequential F0-grouping would consequently be at least 5 dB at 8-semitones $\Delta F0$.

C. $\Delta F0$ benefit and the perception of a pitch difference

The fact that harmonic cancellation might asymptote at a $\Delta F0$ of 2 semitones or less, while sequential grouping by F0 increases substantially from 2 to 8 semitones suggests a distinction of the two mechanisms with respect to the size of $\Delta F0$. Sequential grouping by F0 may be more dependent on the size of $\Delta F0$ than harmonic cancellation, at least above 2 semitones. Exp. 3.2 was concerned with the size of $\Delta F0$ needed for sequential grouping by F0 to occur. Although this question did not find a clear answer for the reasons aforementioned, the contribution of sequential F0-grouping is presumably weak at 2-semitones $\Delta F0$. Consistent with this result, Darwin et al. (2003) reported that listeners started to benefit from $\Delta F0$ to selectively attend to the target sentence from a 2-semitones $\Delta F0$.

D. Importance of the number of interfering voices

Brokx and Nootboom (1982) suggested that the use of monotonized speech emphasized the effects related to perceptual fusion as opposed to perceptual tracking: in other words, emphasized harmonic cancellation as opposed to sequential F0-grouping. Bird and Darwin (1998) are likely to have emphasized harmonic cancellation even more by using continuously-voiced speech, since the harmonic benefit may be dependent on the temporal continuity of the interferer's F0. Another reason why those two studies might have focused on the harmonic mechanism more than sequential F0-grouping was the use of a single interfering voice. Even with the CRM design, the use of $\Delta F0$ for sequential grouping was reported to be weak with a single interferer (Darwin and Hukin, 2000a). In contrast, by using a 2-voice interferer, the present experiments reported substantial benefits of sequential F0-grouping. Thus, the relative contribution of harmonic cancellation and F0-grouping may change according to the number of interfering voices. When several interfering voices have different F0s, applying harmonic cancellation recursively may considerably distort the quality of the target voice. The ability to group sounds of the target voice on the basis of its F0 may also be progressively disrupted in the presence of several interfering voices, especially when intonated. The key-issue is to know how many interfering voices must be taken into account in realistic situations of conversation.

V. CONCLUSION

Because in the presence of several voices, listeners confused the target voice with two other voices, they could use a ΔF_0 to organise the auditory scene, resulting in a benefit attributable to sequential grouping by F_0 . In contrast, the auditory scene was well defined when the target voice was masked by a buzz interferer, without the need of ΔF_0 . Therefore the ΔF_0 benefit is likely to stem from harmonic cancellation for buzz interferers while it stems from harmonic cancellation and sequential F_0 -grouping for speech interferers. At large ΔF_0 s, the benefit was bigger with speech interferers than with buzz (Exp. 3.1) and smaller at small ΔF_0 s (Exp. 3.2). Besides from 2- to 8-semitones ΔF_0 , the benefit remained almost constant for buzz interferers, while it increased by 5 dB for speech interferers. These results emphasized three main points.

First, the harmonic benefit could be surprisingly small with speech interferers. A potential explanation was that harmonic cancellation only operates on the voiced portions of speech. Speech has a temporally interrupted F_0 pattern whereas a buzz has a continuous F_0 pattern, resulting in a less effective cancellation of speech interferers than of buzz interferers.

Second, the increase in the ΔF_0 benefit, as ΔF_0 increased, did occur for speech, not for buzz interferers. This led to the idea that as ΔF_0 increases, the benefit of harmonic cancellation asymptotes early, while the benefit of sequential F_0 -grouping is much more gradual.

Third, the relative contribution of harmonic cancellation and sequential F_0 -grouping might be influenced not only by the size of ΔF_0 , but also by the use of a single or multiple interferers.

Chapter IV.

EFFECTS OF REVERBERATION AND F0 MODULATION ON THE $\Delta F0$ BENEFIT WITH A BUZZ INTERFERER

I. Introduction

The previous Chapter concluded that in a segregation task of speech masked by other speech, the $\Delta F0$ benefit probably originated from the combination of a harmonic cancellation mechanism and a sequential F0-grouping mechanism. A single result suggested that harmonic cancellation rather than harmonic enhancement was involved. The benefit of a 2-semitones $\Delta F0$ (where sequential F0-grouping is limited anyway) was smaller for speech interferers than for a buzz interferer, possibly reflecting the detrimental effect of breaks in the voicing of speech. Harmonic enhancement should predict a similar harmonic benefit whether the interferer's F0 was interrupted or continuous. It was therefore essential to test what caused the $\Delta F0$ benefit with a buzz interferer, i.e. whether the harmonic benefit arose from harmonic cancellation or enhancement, before returning to speech interferers (Chap. V).

If harmonic enhancement of the target is responsible for the $\Delta F0$ benefit, then the process will fail if the target is inharmonic, whereas if harmonic cancellation of the interferer is responsible, then the process will fail if the interferer is inharmonic. There are many ways of generating inharmonicity. Some experiments of Chap. II generated inharmonic interferers by randomly offsetting some partials from a harmonic complex. At least, two other ways have been reported in the literature and occur for real speech and in a realistic environment. F0 modulation can cause inharmonicity and reverberation makes it worse as explained below.

Detrimental effect of F0 modulation

Both harmonic enhancement and cancellation must have a limited temporal resolution (see section D, general discussion of Chap. II) above which dynamic harmonic stimuli cease to be accurately defined and start to be blurred, i.e. start to be slightly inharmonic. When the rate of modulation exceeds the temporal resolution of the mechanism, the stimuli can not be enhanced or cancelled as effectively as they would be if they were not modulated. Culling et al.'s (1994) Exp. 3 measured double-vowels segregation and showed that a F0 modulation of ± 2 semitones at 5 Hz reduced the benefit of a 1-semitone $\Delta F0$ by 6 dB in anechoic conditions.

Detrimental effect of reverberation combined with F0 modulation

Reverberation adds delayed copies of the direct sound. The reflections are delayed by their path between walls of the room, so reflected sound arriving at a given time was emitted at a range of times earlier than the direct sound. Therefore if the F0 is constant over time, the reflections bring the same F0 as that of direct sound. But if the F0 varies over time, the auditory analysis at the listener's ear has to cope simultaneously with the F0 from direct sound and those from the reflections. Given that harmonicity information can be integrated over a large frequency band (Exp. 4 of Chap. II), the harmonic cancellation is likely to suffer from the presence of an interferer with several F0s. In Culling et al.'s (1994) Exp. 3, the benefit of a 1-semitone $\Delta F0$ was reduced by 10 dB in reverberant compared to anechoic conditions for a F0 modulation of ± 2 semitones at 5 Hz, while reverberation had no effect when vowels were monotonized.

Predictions

The goal of the single experiment of this Chapter (Exp. 4.1) was to test the hypotheses of harmonic enhancement and cancellation by presenting unrealistic situations where F0 modulation and reverberation of the target and the interferer were controlled factorially in all combinations. According to each theory, some predictions can be made. If the benefit was due to harmonic enhancement, then it should be disrupted primarily for a reverberant modulated target, to a smaller extent for an anechoic modulated target and it should be intact for a monotonized target (anechoic or reverberant), regardless of the interferer. If the benefit was due to harmonic cancellation, then it should be disrupted primarily for a reverberant modulated interferer, to a smaller extent for an anechoic modulated interferer and it should be intact for a monotonized interferer (anechoic or reverberant), regardless of the target.

II. Exp. 4.1 $\Delta F0$ benefit depends on parameters of the buzz interferer

A. Stimuli

The same target sentences were used from the experiments of Chap. III and were manipulated with the same Praat package. The buzz interferers were monotonized or F0-modulated so their long-term average spectrum was that of 16 sentences either monotonized or F0-modulated. Contrary to Chap. III, this buzz was not speech-modulated; it did not have any amplitude envelope, so no dip-listening could occur in this experiment. The mean F0s of the target sentences (123.5 Hz) were higher than the interferers by 2 semitones. The modulation widths of the target and interferer (0 or ± 2 semitones) were controlled orthogonally: interferer and target both monotonized, or both modulated, or interferer monotonized while target modulated, or vice versa. F0 modulation was achieved by Praat. By low-pass filtering the F0

contour, Binns (2007) found that the most important frequencies in intonated speech lied between 2 and 4 Hz. The modulation frequency was set to 5 Hz to ensure that the variations of F0 will be clearly perceived. Therefore F0 modulations of targets and interferers were always in phase with each other. All interferers were longer than all target sentences. The monotonized speech sounded like a robotic voice, whereas the frequency modulated speech sounded rather like an old man's voice.

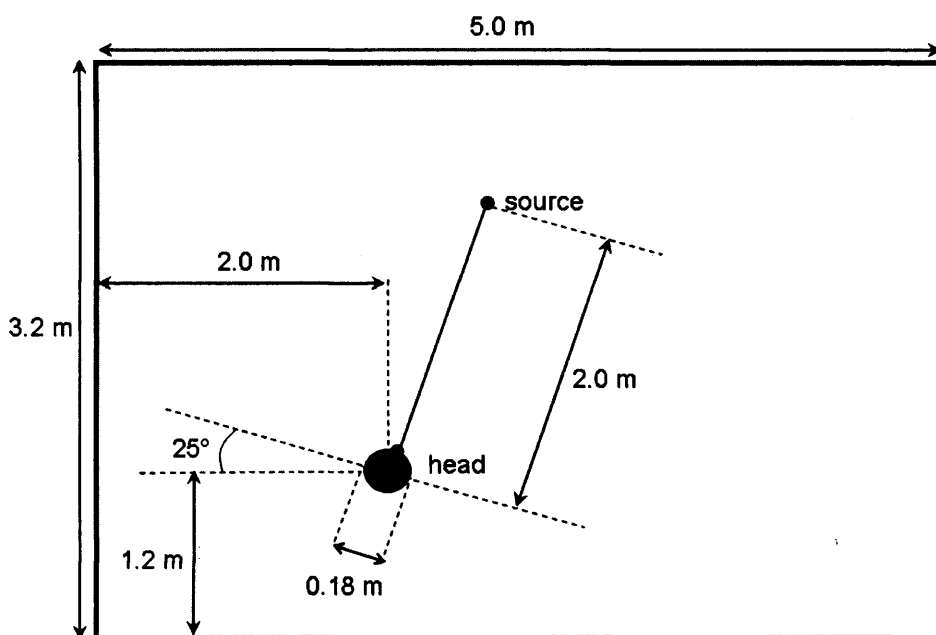


FIG. 4.1 Spatial configuration and virtual room considered in all experiments involving reverberation

Reverberation was added using the image (ray-tracing) method (Allen and Berkley, 1979; Peterson, 1986) as implemented in the |WAVE signal processing package (Culling, 1996). The virtual room and source/receiver configuration was identical to that of Culling et al. (1994). The room had dimensions 5 m long \times 3.2 m wide \times 2.5 m high and virtual sources were 2 m from the receivers (Figure 4.1). The two receivers, separated by 18 cm, were placed along an axis at 25° to the 5 m wall on

either side of a centre point located 1.2 m from the 5 m wall and 2 m from the 3.2 m wall. The receivers were modelled as omnidirectional microphones suspended in space with no head between them. Absorption coefficients for the internal surfaces of the room were all 0.3 for the reverberant room, giving a direct-to-reverberant ratio of -8.56 dB and -8.60 dB for the left-ear and right-ear impulse responses respectively (high-pass filtered above 20 Hz). For the anechoic room the coefficients were all set to 1, giving an infinite direct-to-reverberant ratio. Binaural stimuli were produced by generating the impulse responses for the two receivers in virtual space and convolving the speech samples with these two impulse responses. The degrees of reverberation on the target and the interferer were controlled orthogonally: interferer and target were both anechoic or both reverberant or the interferer was anechoic while the target was reverberant or vice versa.

The experiment had sixteen different conditions, covering two target modulations (0 versus ± 2 semitones), two interferer modulations, two target rooms (anechoic versus reverberant) and two interferer rooms. Mean ΔF_0 was constant at 2 semitones. Each of the 160 target sentences were manipulated in four conditions (2 target modulations \times 2 target rooms), creating 640 target stimuli. Four interfering buzz stimuli were created. All stimuli were presented binaurally at 69 dB SPL (see Appendix B).

B. Procedure

The same SRT procedure was used as in Chap. III. Sixteen runs measured SRTs in each of the sixteen different conditions. Sixteen new listeners attended a single 80-min session, resulting in one complete rotation of the conditions.

C. Results

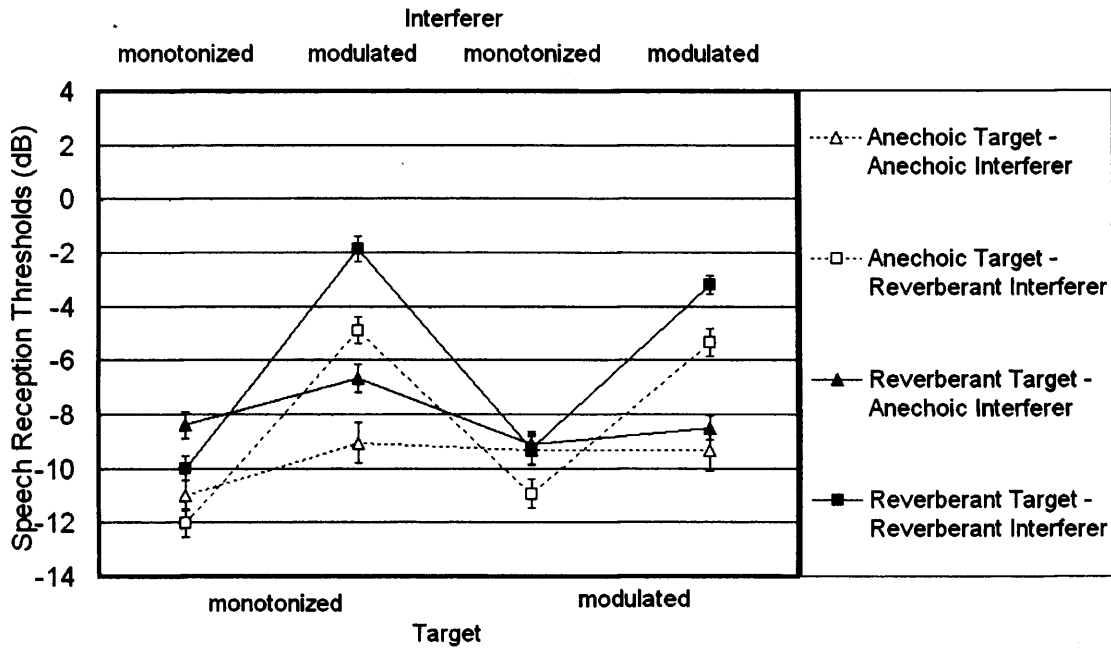


FIG. 4.2 Mean speech-reception thresholds for the conditions where the target voice and the buzz interferer were separated by a ΔF_0 of 2 semitones and modulated factorially. Reverberation was also applied factorially to the target (empty versus filled symbols) and to the interferer (triangles versus squares). Lower thresholds indicate greater intelligibility. Error bars are ± 1 standard error of the mean.

Figure 4.2 presents the mean SRTs measured in the present experiment. A four-factor analysis of variance (Target modulation \times Interferer modulation \times Target room \times Interferer room) showed no main effect of the target modulation [$F(1,15)=0.2$, $p>0.05$]. Mean SRTs were lower when the interferer was monotonized rather than modulated: main effect of interferer modulation [$F(1,15)=104.5$, $p<0.0001$]. Mean SRTs were lower when the target was anechoic than reverberant: main effect of target room [$F(1,15)=57.4$, $p<0.0001$]. Mean SRTs were also lower when the interferer was anechoic than reverberant: main effect of interferer room [$F(1,15)=36.5$, $p<0.0001$]. Figure 4.3 presents the mean SRTs averaged across target's room and modulation

(left) or across interferer's room and modulation (right) as a direct test of the predictions of harmonic cancellation and enhancement. As shown in the left part of Figure 4.3, the interferer room and interferer modulation interacted strongly [$F(1,15)=262.3, p<0.0001$]. As shown in the right part of figure 4.3, the target room and target modulation interacted [$F(1,15)=6.3, p<0.05$]. Figure 4.4 showed that across all room configurations, target modulation and interferer modulation also interacted [$F(1,15)=12.9, p<0.01$]. No other interaction was significant [$\max(F(1,15))=3.7, p>0.05$].

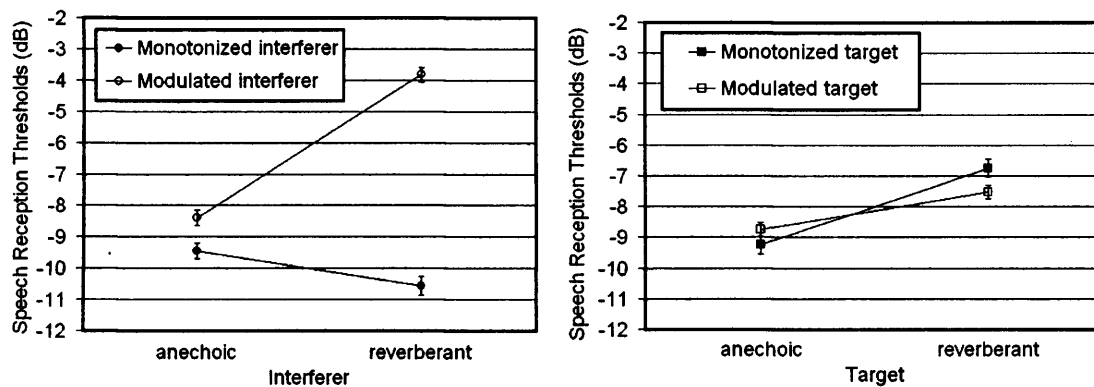


FIG. 4.3 (left) Mean speech-reception thresholds for the conditions where F0 modulation and reverberation were applied factorially to the buzz interferer, averaged across all target configurations. (right) Mean speech-reception thresholds for the conditions where F0 modulation and reverberation were applied factorially to the target speech, averaged across all interferer configurations. The target voice and the buzz interferer were separated by a mean $\Delta F0$ of 2 semitones. Lower thresholds indicate greater intelligibility. Error bars are ± 1 standard error of the mean.

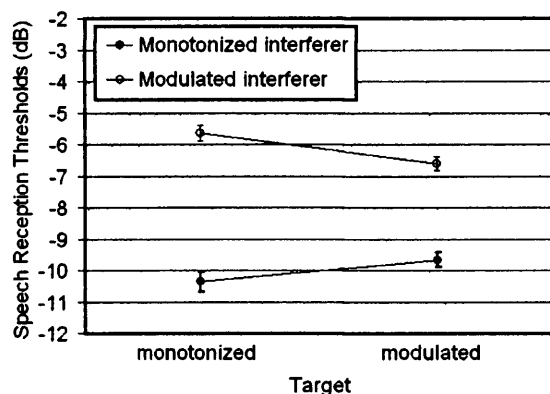


FIG. 4.4 Mean speech-reception thresholds for the conditions where F0 modulation was applied factorially to the buzz interferer and the target, averaged across all room configurations. The target voice and the buzz interferer were separated by a ΔF_0 of 2 semitones. Lower thresholds indicate greater intelligibility. Error bars are ± 1 standard error of the mean.

III. Discussion

A. Harmonic cancellation fails with a modulated reverberant interferer

Figure 4.3 enables one to directly compare the predictions of the two theories. In the left part, mean SRTs were the lowest for the monotonized interferer, increased by 1 or 2 dB for an anechoic modulated interferer and increased by 5 or 6 dB for a reverberant modulated interferer. The results are consistent with the harmonic cancellation theory. With a 5-Hz modulation frequency and a ± 2 -semitones width, the temporal resolution of the cancellation mechanism might not be fast enough to follow such a rate of modulation; the buzz's harmonicity is blurred and the buzz can not be cancelled as effectively as when it is monotonized, i.e. purely harmonic. This was a minor effect. But in reverberation, the F0 modulation provides the cancellation mechanism with several simultaneous F0s for the same interferer and the mechanism

failed to cancel the interferer. This represented a large loss of intelligibility. In other words, the F0 modulation of the interferer was much more detrimental in reverberant than in anechoic conditions.

The temporal continuity of the F0 pattern of the interferer was argued in Chap. III to vary the efficiency of harmonic cancellation. The envelopes of the speech-modulated buzz interferers stemmed from 2-voice interferers, so their F0 continuity was comparable with the completely uninterrupted buzz's F0 of the present experiment, which did not receive any envelope. Therefore the $\Delta F0$ benefit was expected to be of similar magnitude to that of Exp. 3.2 for the buzz interferer, i.e. about 5 dB. The loss of intelligibility due to the presence of a modulated reverberant buzz was about 5 or 6 dB. Therefore it was likely that the entire $\Delta F0$ benefit was lost.

B. No evidence for intrinsic effect of F0 modulation

Perceptual grouping by coherent frequency modulation has been investigated by several studies. Through prominence judgments, McAdams (1989) and Marin and McAdams (1991) showed that frequency modulation could increase the perceptual prominence of modulated sounds compared to a static background, but they failed to highlight the mechanism which would exploit coherent frequency modulation. Carlyon (1991) reported the first evidence that coherent frequency modulation does not provide a simultaneous grouping cue. Later on, Culling and Summerfield (1995) showed that whether the modulation of the target was coherent or incoherent, masked thresholds were higher for modulated than for static masker vowels. So they concluded that there is a mechanism that detects frequency modulation but is insensitive to across-frequency differences in the pattern of that modulation. In other words, common frequency modulation is not a grouping cue in itself and apparent effects of frequency modulation could be attributed to other mechanisms. The

following paragraphs argued that intelligibility did not suffer from a prominence effect due to F0 modulation and that the effects due to frequency modulation of the interferer were mediated through disruption of harmonic cancellation.

First, if frequency modulated sounds are more prominent than static sounds in a static background, then intelligibility might have suffered from the fact that the F0-modulated interferer stood out against the target, being more masking than a monotonized interferer. However, this was a minor effect when the interferer was anechoic but a large effect when it was reverberant. Interferer's modulation and interferer's room interacted strongly. If F0 modulation had any intrinsic effect, it would only concern the loss of intelligibility occurring in anechoic conditions which was a small effect.

Secondly, had listeners grouped by F0 modulation, one might have expected the target modulation to be beneficial, especially when the interferer is not modulated. The significant interaction between F0 modulation of the target and the interferer, plotted in the Figure 4.4, showed results exactly opposite to such expectancies. F0 modulation of the target was slightly detrimental when the interferer was monotonized, but slightly beneficial when the interferer was modulated. Those results did not support the idea that listeners used F0 modulation as a cue to segregate the competing sources.

Therefore the detrimental effect of F0 modulation of the interferer was attributable to a loss of harmonic cancellation, not a grouping mechanism based on common frequency modulation. The meaning of the interaction between the modulation of both target and interferer and that of the interaction between the target's modulation and target's room remained unclear. Those effects were of small magnitude.

C. No evidence for harmonic enhancement

The harmonic enhancement theory predicted that loss of intelligibility should occur when the target became inharmonic, i.e. particularly when the target was modulated in reverberation. The data showed that it was not the case: in the right part of Figure 4.3, SRTs were lower for a modulated reverberant target than for a monotonized reverberant target. Harmonicity of the target did not play any role.

D. STI effect

In all the conditions of the experiment, intelligibility suffered when target speech is subject to reverberation, resulting in about 2-dB elevation of SRTs in the present data. This sort of effect occurs even without an interferer and reflects the loss of amplitude modulation of the target due to reverberation. It is related to speech intelligibility indices, like the STI (Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985). Since the buzz was uninterrupted, this effect was not mixed with the reduction in dip-listening which results from the tails of reverberation filling the gaps of the interferer.

E. Possible binaural effects

The data were averaged across target modulation to produce Figure 4.5, which highlights the strong interaction between the interferer's room and interferer's modulation. Reverberation of the modulated interferer increased SRTs, because the cancellation mechanism failed to cancel such an interferer, as discussed in the first section. However, when reverberation was applied to the monotonized interferer, mean SRTs were slightly lower than when the monotonized interferer was anechoic. This latter effect might be a binaural effect.

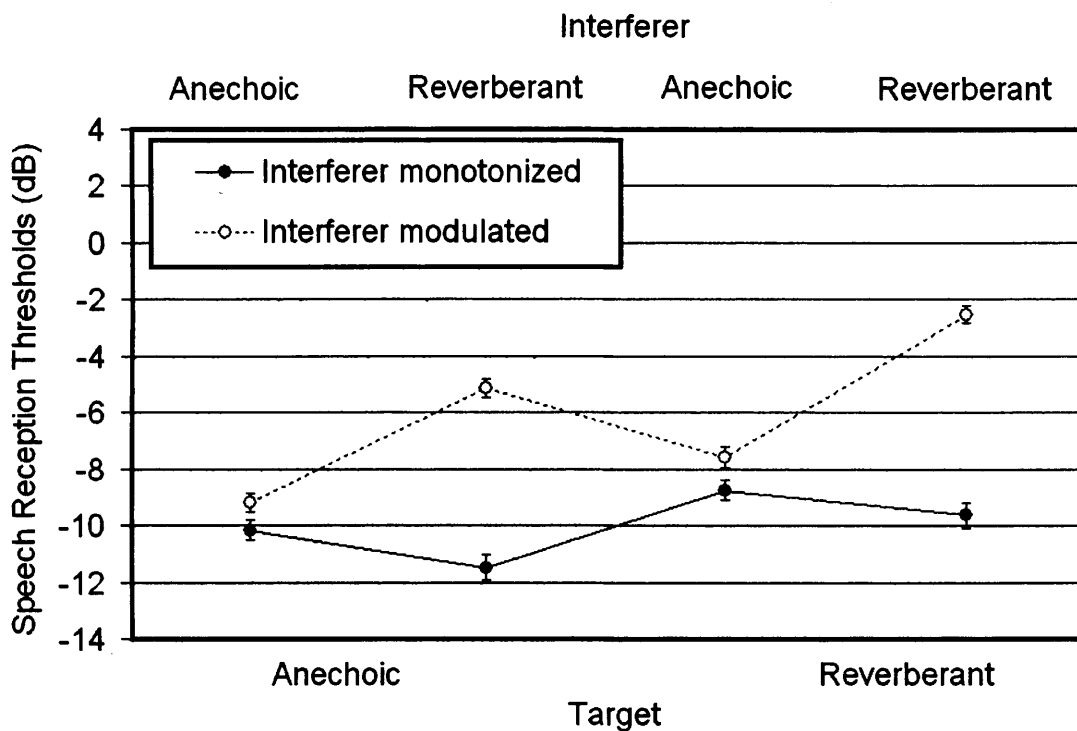


FIG. 4.5 Mean speech-reception thresholds for the conditions where the target and the buzz interferer were separated by a ΔF_0 of 2 semitones and heard independently in anechoic or reverberant conditions. The data were averaged across the two levels of the target's F_0 modulation. Lower thresholds indicate greater intelligibility. Error bars are ± 1 standard error of the mean.

Equalization-cancellation theory proposes a model for binaural unmasking, where signals at each ear are temporally aligned and subtracted one from the other (Durlach, 1963). The equalization process compensates for the interaural time delay of the masker, so that the cancellation process can remove it. The purpose is to not cancel the target along with the masker, so binaural unmasking requires the target sound to have a different interaural time delay than that of the masker. For this reason, since the target and the interferer were collocated in front of the listener, resulting in interaurally correlated sounds in anechoic conditions, the only binaural effects of the present experiment occurred when the target and the interferer were in different rooms.

Licklider (1948) showed with experiments on masking of speech by white noise that speech presented in phase at the two ears is less masked by an interaurally uncorrelated noise than by an interaurally correlated noise (and less masked by an anticorrelated masker than by an uncorrelated one). When the interferer is reverberant, it becomes slightly less correlated across the ears. This results in less binaural masking provided that the target is perfectly correlated, i.e. anechoic. This binaural configuration is similar to NuSo, occurring in Figure 4.5 for the condition presenting an anechoic target against a monotonized reverberant interferer. In this situation, the equalization-cancellation process will reduce the masking due to the reflections of the masker, leaving the direct sound of the target (i.e. the anechoic target) unaffected.

The other binaural effect occurs when the anechoic interferer (i.e. perfectly correlated) can be completely suppressed on the interaural time delay which corresponds to its direct sound, while the reflections of a reverberant target remain less affected. This binaural configuration is similar to NoSu. When the target is reverberant, it becomes slightly less correlated across the ears. By subtracting the two ears to cancel the perfectly correlated interferer, the equalization cancellation process not only will suppress the direct sound of the target but will also distort an important part of its reflections. Given that the target speech has to be understood by those distorted reflections, the binaural advantage of NoSu might be counteracted by poor intelligibility of the residual target. This can explain why the NoSu effect is not evident in the data of the Figure 4.5.

Finally, when both target and interferer were reverberant, the binaural configuration is similar to NuSu where the collocated sources have exactly the same the imperfect correlations across the ears. Cancellation of a part of the interferer necessarily cancels the same part of the target. Unexpectedly, Figure 4.5 showed that

SRTs were lower when both target and monotonized interferer were reverberant (NuSu) than when only the target was (NoSu). One possible interpretation is that it is still worth cancelling the reflections of the masker (necessarily along with those of the target), resulting in less masking, provided that intelligibility of the target has not been reduced by as much as the binaural release from masking. If speech intelligibility in a room is dominated by its direct sound, then such a binaural process might still be beneficial. The idea that the direct sound and the first reflections of a reverberant impulse response provide better intelligibility than the rest of the impulse response has been implied by the concepts of early-to-late ratios or useful-to-detrimental ratios. However, the trade between those ratios and the binaural advantage remains unclear. In any case, those binaural effects were of small magnitude and it is noticeable that they have occurred only because the position chosen for the sources was not symmetrical in the room.

IV. CONCLUSION

The single experiment of this Chapter confirmed that the SRTs increased when the interferer's F0 was modulated and increased substantially when it was modulated in reverberation. Those results were consistent with a mechanism of harmonic cancellation of the interferer. The effect of F0 modulation could be mediated by the limited temporal resolution of the mechanism, which blurs the harmonicity of dynamic interfering stimuli when their rate of modulation is too high. This was a minor effect. The presence of several F0s for the same interferer, caused by echoes, was a much larger impairment, occurring when the interferer's F0 was modulated in reverberation. The effect of F0 modulation was not related to a prominence issue or a grouping mechanism by frequency modulation, and might be only attributable to a disruption of harmonic cancellation. None of the present results fit the predictions of the harmonic enhancement theory.

Competing sources were collocated, but the listener's position was not symmetrical in the room, resulting in some small binaural effects. Given the similarity between the buzz interferers used in Chap. III and IV, the loss of harmonic cancellation was likely to represent the entire benefit. Therefore the $\Delta F0$ benefit was abolished by a modulated reverberant buzz interferer. This result was found for a given level of reverberation (direct-to-reverberant ratio of -8.6 dB). Different degrees of reverberation could substantially influence the results.

Chapter V.

EFFECTS OF REVERBERATION AND F0 MODULATION ON THE Δ F0 BENEFIT WITH SPEECH INTERFERERS

I. Introduction

Chapter IV concluded that harmonic cancellation was entirely responsible for the Δ F0 benefit with a buzz interferer. This mechanism could be disrupted to a small extent by F0 modulation of the buzz probably when the rate of modulation began to exceed its temporal resolution. To a large extent, the mechanism was no longer effective when the interferer was composed of several F0s, which occurred when F0 modulation was combined with reverberation. The question immediately arises as to whether harmonic cancellation is also responsible for the harmonic benefit with speech interferers, and whether it is disrupted in the same ways. Exp. 5.1 addressed this question.

Towards a more accurate measure of the contribution of harmonic cancellation in the Δ F0 benefit with speech interferers

Exp. 5.1 attempted to measure the contribution of harmonic cancellation for speech interferers more accurately than the measure used in Chap. III. Exp. 3.2 indeed showed that the benefit of a 2-semitones Δ F0 (where sequential F0-grouping is limited) for speech interferers was smaller than that for a buzz interferer; as a result the harmonic benefit should be smaller for speech than for buzz interferers. Another way of estimating the contribution of harmonic cancellation was to measure the increase in SRTs due to the loss of harmonic cancellation. In this respect, Exp. 4.1 showed that a combination of reverberation and F0 modulation abolished the entire

benefit of harmonic cancellation for a 2-semitones ΔF_0 , for a buzz interferer. If those conditions completely abolished the contribution of harmonic cancellation for a buzz interferer, they should also abolish that for speech interferers. Therefore, Exp. 5.1 used the same experimental conditions to measure the loss of harmonic cancellation for speech interferers.

What is sequential F0-grouping?

Chap. III concluded that with speech interferers, sequential F0-grouping contributed substantially to the ΔF_0 benefit. Very little is known about this mechanism, other than that somehow listeners can group sounds on the basis of their F0 or their *pitch* in order to avoid confusing the competing talkers. In order to increase our understanding of this mechanism, it was interesting to investigate separately the effects of reverberation and F0 modulation on a large ΔF_0 (8 semitones) where a large sequential F0-grouping benefit had been found: at least 5 dB in Exp. 3.1. The effect of reverberation on sequential F0-grouping was investigated in Exp. 5.2 and the effect of F0-modulation on sequential F0-grouping was investigated in Exp. 5.3.

II. Exp. 5.1 Benefit of a small ΔF_0 , subject to F0-modulated speech interferers

A. Stimuli and procedure

The same target and interfering sentences were used as those of Chap. III and were manipulated with the same Praat package. The present experiment was designed identically to Exp. 4.1, except that 2-speech interferers were used instead of buzz and it was a mixed within-between subjects design. Sixteen new listeners, separated in

four groups of four listeners, each attended a single 40-min session, resulting in two complete rotations of four conditions. These conditions cover two target rooms (anechoic versus reverberant) by two interferer rooms. The orthogonal manipulation of target's and interferer's modulation (0 versus ± 2 semitones) constituted two between-subjects factors, resulting in four groups of subjects. All stimuli were presented at 69 dB SPL.

B. Results

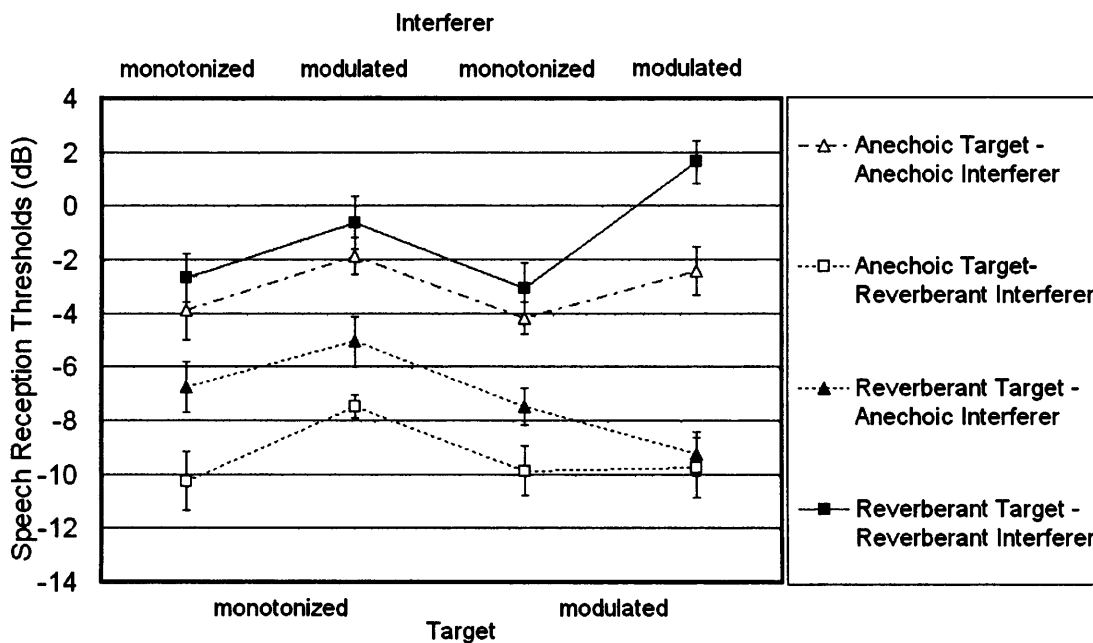


FIG. 5.1 Mean speech-reception thresholds measured in Exp. 5.1 for the conditions where the target voice and the interfering voices were separated by a ΔF_0 of 2 semitones. Reverberation was applied factorially to the target (empty versus filled symbols) and to the interferer (triangles versus squares). The four groups of listeners matched each with one of the four orthogonal configurations of F_0 -modulation of the target and the interferer. Lower thresholds indicate greater intelligibility. Error bars are ± 1 standard error of the mean.

Figure 5.1 presents the mean SRTs measured in Exp. 5.1. A four-factor analysis of variance (Target room \times Interferer room \times Target modulation \times Interferer modulation) showed that mean SRTs were lower when the target was anechoic than reverberant: main effect of target room [$F(1,12)=29.6, p<0.0001$]. There was no main effect of the interferer room [$F(1,12)=0.05, p>0.05$]. Interferer and target rooms interacted strongly [$F(1,12)=520.3, p<0.0001$]: reverberation on the interferer was beneficial for an anechoic target while detrimental for a reverberant target. There was no main effect of target modulation [$F(1,12)=1.3, p>0.05$]. The modulation of the interferer was detrimental: main effect of interferer modulation [$F(1,12)=7.0, p<0.05$]. The modulation of target and interferer only interacted when combined with room factors. Specifically, the 3-way interaction Interferer modulation \times Target room \times Interferer room was significant [$F(1,12)=12.9, p<0.01$]. The 3-way interaction Target modulation \times Target room \times Interferer room was significant [$F(1,12)=13.3, p<0.01$]. The 4-way interaction was also significant [$F(1,12)=15.5, p<0.01$]. No other interaction was significant [$\max(F(1,12))=1.2, p>0.05$].

Note that the main effects of target and interferer modulation were between-groups effects. Thus, to ensure that it did not originate from a group difference, we replicated the conditions where both target and interferer were in the same room, in a within-subjects design. We found similar results, namely the target modulation had no effect while the interferer modulation caused an increase of SRTs of about 2 dB in anechoic as in reverberant conditions. A further 1-dB impairment was found when both target and interferer were modulated in reverberation.

C. Discussion

C.1 STI and interaural phase effects

Two effects observed in Exp. 4.1 were also observed in Exp. 5.1, which caused different offsets on the SRTs of each room configuration, irrespective of the F0 modulation configurations. First, reverberation distorts the amplitude modulation of the target (STI effects). Second, interaural phase effects could be dissociated into two sub-components: a reverberant interferer was less masking than an anechoic interferer when collocated with the target (NuSo effect) and an anechoic interferer could be easily suppressed binaurally without completely suppressing the reverberant target (NoSu effect). Those effects occurred for different room configurations as following. When target and interferer shared the anechoic room, these effects did not occur. When target and interferer shared the reverberant room, SRTs were higher due to the presence of STI effects and the absence of interaural phase effects. When the target alone (not the interferer) was reverberant, detrimental STI effects were counteracted by the binaural advantage of NoSu effects. When the interferer alone was reverberant, STI effects were absent and NuSo effects provided a binaural advantage.

The binaural advantages of NuSo and NoSu effects were not expected to be very large. As a consequence, interaural phase effects might not be sufficient to explain the large release from masking observed when target and interferer were in different rooms. Another mechanism was suspected (next section C2).

C.2 Grouping based upon the degree of reverberation

Whatever the modulation of both the target and the interferer, the conditions where an anechoic target was presented with a reverberant interferer, or vice-versa,

always produced lower SRTs than when both sources were anechoic or both reverberant. In Exp. 4.1, such a release from masking did not occur, leading to the idea that this release might be related to the use of speech. Hence we reasoned that the degree of reverberation might be another cue that helps to organise the auditory scene, initially difficult in the presence of three voices. Consequently, sequential grouping by the degree of reverberation contained in the speech stimuli might be responsible for a large release from informational masking.

Given the magnitude of the effect, it is important to consider whether this effect could occur in a real environment. A similar situation occurs in a reverberant room, when a target talker is very close to the listener while an interfering talker is located away from the listener. In this situation, however, level differences between the two sources are also likely to favour the target. A less common situation occurs when a distant target is reverberant while a close interferer is anechoic, like listening to a public lecture while a nearby member of the audience is talking. Here, the target-to-masker ratio might be quite low and such a release from informational masking based upon the degree of reverberation should be of great use.

C.3 Disruption of harmonic cancellation of speech interferers

The present data confirmed the result found with buzz interferers; the modulation of the target's F0 alone had no effect, while the modulation of the interferer's F0 alone was detrimental. A ± 2 -semitones modulation at 5 Hz applied to the interferer's F0 increased SRTs by about 2 dB, reflecting a loss of the harmonic cancellation. This could occur when the temporal resolution of the harmonic cancellation is too slow to follow the rate of modulation applied to the speech interferer so that it can no longer be cancelled effectively.



More interestingly, F0 modulation of the interferer impaired the harmonic cancellation benefit to the same extent whether the speech interferer was anechoic or reverberant. This result can be explained as a ceiling effect of the disruption of harmonic cancellation (or a floor effect of reverberation). In Chap. III, the $\Delta F0$ benefit was smaller for speech than for buzz interferers, presumably because the breaks in the voicing of speech caused the cancellation mechanism to perform only intermittently. As a result the contribution of harmonic cancellation for a 2-semitones $\Delta F0$ was at most 3 dB. The present experiment intended to estimate the contribution of harmonic cancellation by measuring its disruption. Given that the same level of reverberation, modulation width, modulation frequency and $\Delta F0$ were used in both experiments, the disruption of the mechanism should have revealed the entire benefit. SRTs increased by only 2 dB in the present experiment, suggesting that the harmonic cancellation provided only a 2-dB benefit for a 2-semitones $\Delta F0$ with speech interferers. As a consequence, reverberation may have had no additional effect, because the mechanism was already abolished with an anechoic modulated speech interferer.

C.4 Reverberation and F0 modulation of both target and interferers

When both target and interferers were modulated, reverberation did have an effect, but dependent on the room configuration. The conditions where both sources shared the same room are discussed in the first following paragraph; the conditions where they were in different rooms in the second one.

SRTs were worse when both target and interferer were modulated in reverberation than when only the interferer was. At first sight, this result might not be surprising since there was more energetic masking in the first case than the latter. Indeed, when competing talkers, separated by a 2-semitones $\Delta F0$, were modulated at ± 2 semitones, the F0s of competing sources were overlapping each other. Given that

the modulation of both target and interferer's F0s were always in phase with each other and that the modulation frequency was 5 Hz, a temporal overlap between the competing F0s occurred if sufficient reverberant energy was present after a quarter of a period, i.e. 50 ms. Reverberant energy was attenuated by 10 dB after 50 ms and the tails of reverberation in the experimental stimuli were cut at 250 ms where reverberant energy was attenuated by 50 dB. Therefore, reflections from the interferers' F0, attenuated by more than 10 dB after 50 ms, were temporally overlapping with the target's F0, causing extra energetic masking. However, this interpretation strongly suffers from the fact that the same F0-overlap conditions did not lead to any worse impairment in reverberation when the interferer was a buzz in Exp. 4.1, where the same energetic masking would have occurred. As a consequence, a failure in perceptual organisation (or informational masking) may be a more appropriate explanation. This overlap between the competing F0s might have caused a perceptual fusion of the pitches which was detrimental to sequential F0-grouping, resulting in an extra 2-dB impairment in reverberation. Such perceptual overlap may not have occurred in anechoic conditions if the temporal resolution of pitch perception mechanisms is smaller than 50 ms (see section III.B of Chap VI). This interpretation is further supported by the fact that only the contribution of sequential F0-grouping was left after the entire loss of harmonic benefit by a modulated interferer.

Such a F0 overlap was no longer detrimental but rather beneficial when target and interferers were heard in different rooms. It is possible that the degree of reverberation is a strong sequential cue for grouping, so that sequential F0-grouping provides no further segregation and the F0 overlap does no longer matter. If such was the case, the F0 overlap would have no effect. The fact that it has a beneficial effect is quite puzzling; the meaning of this interaction remains unclear.

D. Interim conclusion

The present experiment confirmed that F0 modulation of the interferer, not that of the target, caused intelligibility to decrease. Those results were therefore consistent with the idea that with speech interferers, the $\Delta F0$ benefit is partially provided via harmonic cancellation not harmonic enhancement. Moreover this loss of intelligibility due to disruption of harmonic cancellation was of small magnitude (only 2 dB) supporting the idea, suggested by Exp. 3.2, that the contribution of harmonic cancellation for speech interferers is weaker than for buzz interferers. The most likely interpretation is that the breaks in the voicing of speech cause the harmonic cancellation to work only intermittently and so to be less beneficial for speech than for a buzz whose F0 is continuous. As a consequence, the contribution of harmonic cancellation is weak and can be abolished by a modulated speech interferer. Reverberation seems not to exacerbate the disruption of the cancellation mechanism any further: the entire harmonic benefit is already lost with an anechoic modulated interferer.

However, reverberation impaired intelligibility in a F0-overlap situation. The two following arguments lead to the idea that it reflects impairment in sequential F0-grouping. First, even at 2-semitones $\Delta F0$, sequential F0-grouping contributes to the benefit. Exp. 5.1 refined the contribution of harmonic cancellation to a 2 dB part of the 3-dB benefit observed in Exp. 3.2. What was left of the $\Delta F0$ benefit, after the disruption of harmonic cancellation of a modulated interferer, was the part attributable to sequential F0-grouping. Second, this F0-overlap impairment did not occur for a buzz interferer, where it was initially assumed that no sequential F0-grouping occurs. A possible explanation of this impairment is that competing pitches are perceptually fused when the competing F0s temporally overlap, from 50 ms after the direct sound.

As a result, competing talkers are heard on the same pitch and listeners confuse them more in reverberation than in anechoic conditions where competing F0s are not temporally overlapping. With a buzz interferer, incoming sounds are already segregated from another acoustic cue; sequential F0-grouping provides no further segregation, so the F0 overlap does no longer matter.

Reverberation had another effect: large releases from masking occurred when competing talkers were heard in different rooms. Those benefits presumably originated from a grouping mechanism based upon the degree of reverberation. Such releases from masking are of interest because they might occur in realistic environments, e.g. when one source is dominated by its direct sound while the other source is dominated by its reflections.

III. Exp. 5.2 Reverberation reduces the benefit of a large ΔF_0 for monotonized speech interferers

There is no reason for harmonic cancellation to fail in cancelling a monotonized interferer, even when the interferer is reverberant. Delayed reflections mix with the direct sound, but since F0 is fixed over time, the interferer retains a well defined F0. Consistent with this expectation, monotonized buzz (Exp. 4.1) or speech (Exp. 5.1) interferers separated by 2-semitones ΔF_0 , showed unimpaired thresholds in reverberation. So, as long as no F0 modulation is introduced, harmonic cancellation should be robust to reverberation. But what about sequential F0-grouping: is it also robust to reverberation? Culling et al. (2003) investigated the effect of reverberation on listeners' ability to segregate competing voices. In their Exp. 2, they used monotonized interfering speech separated by about one octave from the target voice. Relative to anechoic conditions, SRTs increased by 4 dB in reverberation, leading to

the idea that besides STI effects, reverberation might also affect the ΔF_0 benefit even for monotonized speech. In other words, reverberation might affect sequential F_0 -grouping.

By using an 8-semitones ΔF_0 , where sequential F_0 -grouping played a larger role than at 2 semitones, the possible effect of reverberation on sequential F_0 -grouping could be investigated in Exp. 5.2. To this aim, reverberation was included in the design of the experiments of Chap. III, measuring the ΔF_0 benefits for buzz and speech interferers, in which F_0 remained fixed. If reverberation has no effect on the benefits for both interferer types, then both harmonic cancellation and sequential F_0 -grouping are robust to reverberation. If reverberation does not affect the ΔF_0 benefit for a buzz, but affects that for speech interferers, then harmonic cancellation is robust to reverberation but sequential F_0 -grouping is not. If reverberation increases the ΔF_0 benefits, then we may suspect that harmonic cancellation benefits from an increased temporal continuity of the F_0 patterns due to the tails of reverberation. In any case, reverberation was not expected to reduce the ΔF_0 benefit for a monotonized buzz, because we presumed that harmonic cancellation was entirely responsible for the benefit and robust to reverberation as long as F_0 was not modulated.

A. Stimuli and procedure

Exp. 5.2 used two types of interferer (2-voice speech-modulated buzz versus 2-voice speech) and two values of ΔF_0 (0 or 8 semitones). Target and interferers were both heard in anechoic or in reverberant conditions, where the same room and reverberation characteristics were used. Therefore all anechoic stimuli were diotic and all reverberant stimuli were dichotic. All sentences were monotonized and the F_0 profile of the buzz was constant, so that interferers always had a mean F_0 of 110 Hz. Target sentences had a mean F_0 of 110 or 174.6 Hz (8 semitones). Note that changing

the F0 of the interferer while the target would remain at 110 Hz was expected not to influence the results. Indeed Bird and Darwin (1998) had tested a symmetrical arrangement of F0s of the target and interfering speech and found very similar results for both directions of $\Delta F0$. De Cheveigné et al. (1997a) also found no influence of the direction of $\Delta F0$ in double-vowels experiments. Thirty-two new listeners attended a single 50-min session. All stimuli were presented at 69 dB SPL

B. Results

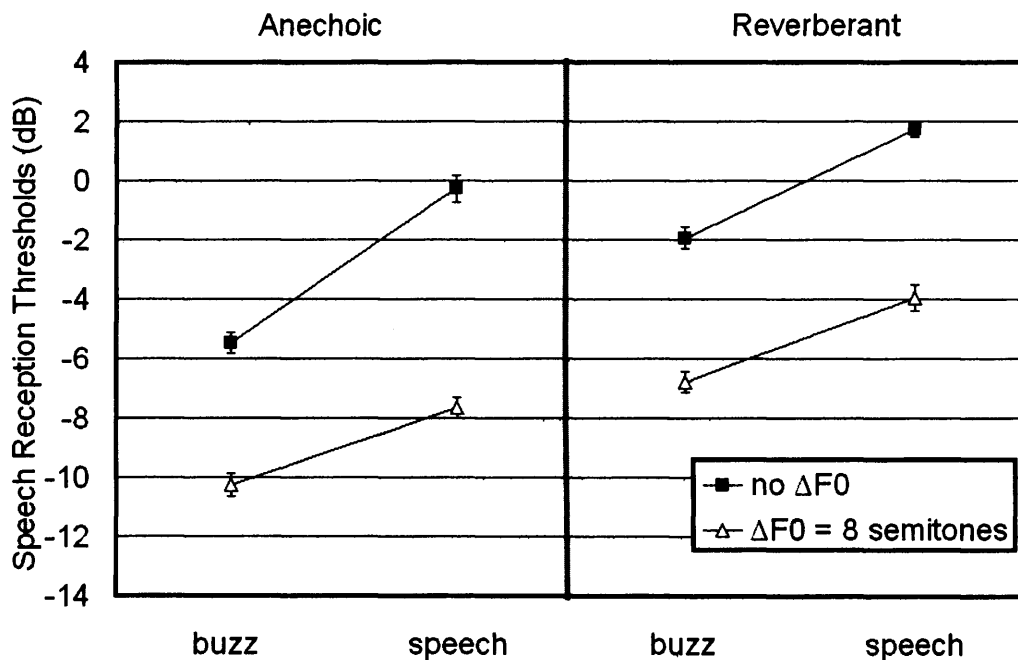


FIG. 5.2 Mean speech-reception thresholds in anechoic (left) and reverberant (right) conditions, for two types of interferer (speech-modulated buzz versus speech) separated by a $\Delta F0$ of 0 (squares) or 8 (triangles) semitones. Lower thresholds indicate greater intelligibility. Error bars are ± 1 standard error of the mean.

Figure 5.2 presents the mean SRTs measured in Exp. 5.2. A three-factor analysis of variance (room \times $\Delta F0$ \times interferer type) shows that mean SRTs were lower when the sources were heard in anechoic rather than in reverberant conditions: main effect of

room [$F(1,31)=139.2, p<0.0001$]. Mean SRTs were lower when they had different F0s than when they were on the same F0: main effect of $\Delta F0$ [$F(1,31)=534.5, p<0.0001$]. Mean SRTs were lower with a 2-voice speech-modulated buzz than with the 2-voice speech interferer: main effect of interferer type [$F(1,31)=227.6, p<0.0001$]. In Figure 5.3, the $\Delta F0$ benefit was significantly bigger when speech interferer was used rather than buzz: interaction $\Delta F0 \times$ interferer type [$F(1,31)=7.3, p<0.05$]. The 3-way interaction was also significant [$F(1,31)=4.4, p<0.05$]: the increase of $\Delta F0$ benefit with speech interferer rather than buzz was significantly bigger in anechoic than in reverberant conditions. No other interaction was significant [$\max(F(1,31))=2.8, p>0.05$].

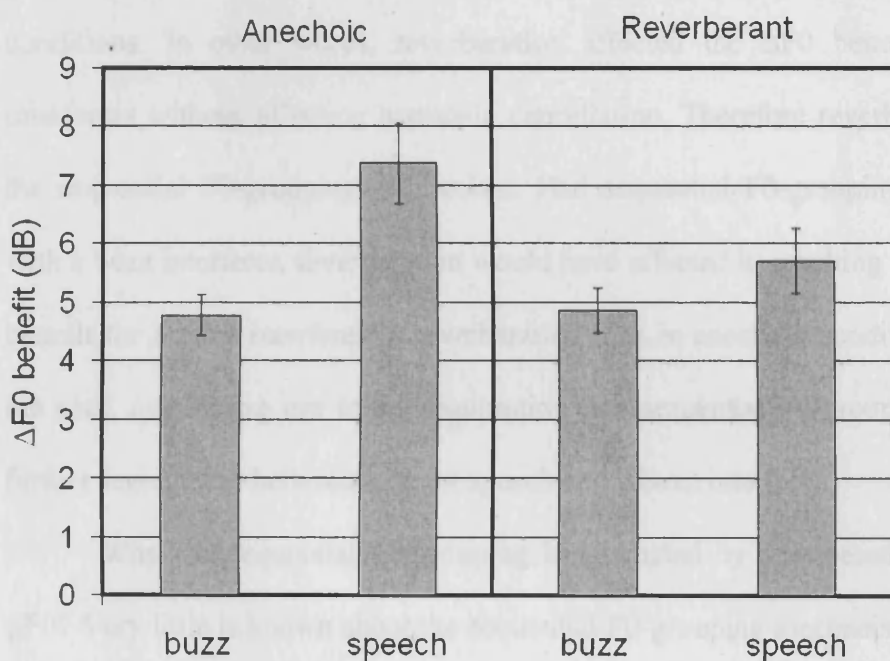


FIG. 5.3 Mean benefits of an 8-semitones $\Delta F0$, measured in anechoic (left) and reverberant (right) conditions, for a monotonized target voice masked by a monotonized buzz or 2 monotonized interfering voices. Error bars are ± 1 standard error of the mean.

C. Discussion

In anechoic as in reverberant conditions, listeners got a 5-dB benefit from a ΔF_0 of 8 semitones with the speech-modulated buzz interferer. This benefit presumably resulted from the harmonic cancellation of the buzz and was not impaired by reverberation. This confirmed that the cancellation mechanism was robust to reverberation provided that the buzz's F_0 was not modulated.

When three talkers spoke simultaneously, listeners confused the target voice with the two interfering voices, resulting in higher SRTs than when they listened to a single voice masked by a buzz. When an 8-semitones ΔF_0 was available, listeners could group sounds from the higher-pitched voice and obtained a greater ΔF_0 benefit. However, the resulting ΔF_0 benefit was smaller in reverberant than in anechoic conditions. In other words, reverberation affected the ΔF_0 benefit with speech interferers without affecting harmonic cancellation. Therefore reverberation affected the sequential F_0 -grouping mechanism. Had sequential F_0 -grouping been involved with a buzz interferer, reverberation would have affected it, resulting in a smaller ΔF_0 benefit for a buzz interferer in reverberation than in anechoic conditions. It was not the case, confirming our initial assumption that sequential F_0 -grouping provides no further segregation between a target speech and a buzz interferer.

Why can sequential F_0 -grouping be disrupted by reverberation with a fixed ΔF_0 ? Very little is known about the sequential F_0 -grouping mechanism, which makes it difficult to understand why it is subject to some factors like reverberation. It is a mechanism that organises the auditory scene by grouping sounds that are perceived as belonging to the same source, in particular by virtue of its pitch. When listeners hear that the target voice is spoken at a distinct pitch from those of the interferers, they manage to avoid confusing the competing talkers, by “directing their attention” to the

pitch of the target voice. Thus, it is possible that sequential F0-grouping acts at a higher stage of auditory processing, related to more cognitive abilities. Due to the STI effect, speech intelligibility already suffers from reverberation, resulting in more difficult listening conditions. It seems sensible to think that if understanding speech is more difficult in reverberation, a speech reception task is cognitively more demanding in reverberant than in anechoic conditions. Therefore the only speculative explanation I can offer is that reverberation produces increased cognitive demands, resulting in restricted cognitive abilities. The credibility of such an interpretation relies on whether or not sequential F0-grouping reflects a high-level mechanism in auditory processing.

IV. Exp. 5.3 Benefit of a large ΔF_0 , reduced by F0 modulation and reverberation for speech interferers

In Chap. III, we argued that the harmonic benefit for speech was constant as ΔF_0 increased because the ΔF_0 benefit for buzz interferers was the same at 2 and 8 semitones. In other words, we assumed that harmonic cancellation works in the same way for buzz and speech interferers. But this assumption might not be verified, leading to an alternative interpretation. It is conceivable that for speech interferers, the competing sources are so similar that harmonic cancellation requires feedback from the listener's attention indicating which F0 is that of the interferer. In contrast, harmonic cancellation might automatically select the F0 of the buzz by using another cue than F0, e.g. level-difference cues. Such a difference in the estimation of the interferer's F0 might explain why the harmonic benefit asymptotes at 2 semitones for a buzz interferer and is more progressive for speech interferers. If that is the case, it

would be even more arduous to differentiate sequential F0-grouping from harmonic cancellation in a speech segregation task.

Exp. 5.3 used an 8-semitones ΔF_0 , reverberation and F0 modulation, where different predictions can be made from our interpretation or the alternative one. Over the 8-dB benefit found in Exp. 3.1 for an 8-semitones ΔF_0 , our interpretation is that the harmonic benefit remains about 2 or 3 dB, and consequently its disruption remains the same at 2- and 8-semitones ΔF_0 . Moreover, because the harmonic benefit is weak with speech interferers, it is already completely abolished for an anechoic modulated interferer and there is therefore little scope for reverberation to have an effect when combined with F0 modulation. In other words, there should still be a ceiling effect of the disruption of harmonic cancellation. Our predictions are therefore that the detrimental effect of F0 modulation should only be 2 or 3 dB and reverberation should not interact with F0 modulation. In contrast, the alternative interpretation is that the harmonic benefit has substantially increased from 2- to 8-semitones ΔF_0 with speech interferers, leaving more scope for reverberation to have an effect when combined with F0 modulation. This interpretation predicts that SRTs should be impaired by F0 modulation, but significantly more in reverberation.

The alternative interpretation attenuates the role of sequential F0-grouping (because harmonic cancellation would play a bigger role), but does not deny it. Therefore the detrimental effect of reverberation on sequential F0-grouping, found in Exp. 5.2, does not favour either of the two interpretations. It should only be replicated: the ΔF_0 benefit should be smaller in reverberant than in anechoic conditions, critically when speech is monotonized.

A. Stimuli and procedure

The same speech material was used. The interferers were 2-voice interferers. In contrast with Exp. 5.1, target and interferers were modulated together by, ± 0 , ± 1 or ± 2 semitones. The interferers shared the same F_0 , and the target's F_0 was either the same or 8 semitones higher. Both target and interferer were heard in the anechoic or reverberant room used previously. The modulation frequency was again 5 Hz. Twenty-four new listeners attended a single 60-min session, resulting in two complete rotations of twelve conditions. These conditions cover two rooms (anechoic versus reverberant) by two ΔF_0 s (0 or 8 semitones) by three modulation widths (± 0 , ± 1 or ± 2 semitones). All anechoic stimuli were diotic and all reverberant stimuli were dichotic and all of them were presented at 69 dB SPL.

B. Results

Figure 5.4 presents the mean SRTs measured in Exp. 5.3. A three-factor analysis of variance (room \times ΔF_0 \times F_0 modulation) shows that SRTs were lower in anechoic than in reverberant conditions: main effect of room [$F(1,23)=130.1$, $p<0.0001$]. Mean SRTs were lower when target and interferers had different F_0 s than when they were on the same F_0 : main effect of ΔF_0 [$F(1,23)=297.3$, $p<0.0001$]. Mean SRTs varied with modulation width: main effect of F_0 modulation [$F(2,46)=29.5$, $p<0.0001$]. The ΔF_0 benefit was smaller in reverberant than in anechoic conditions: interaction room \times ΔF_0 [$F(1,23)=44.0$, $p<0.0001$]. Across the two rooms, SRTs increased due to F_0 modulation only when there was a ΔF_0 : interaction $\Delta F_0 \times F_0$ modulation [$F(2,46)=25.2$, $p<0.0001$], but there was no 3-way interaction. No other interaction was significant [$\max(F(2,46))=0.6$, $p>0.05$].

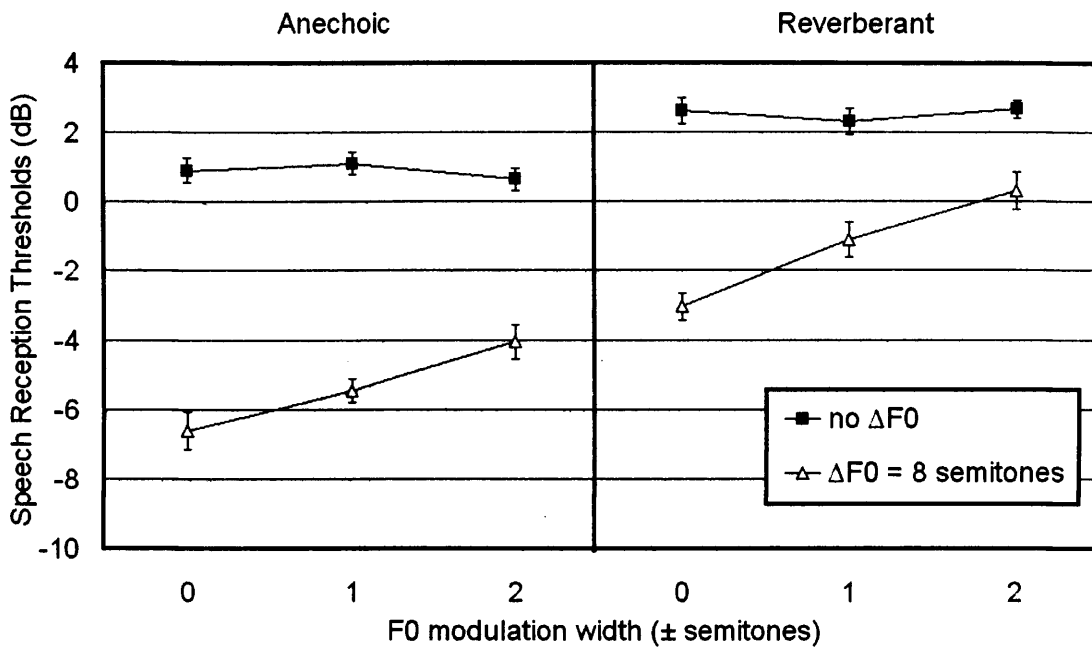


FIG. 5.4 Mean speech-reception thresholds measured in Exp. 5.3, in anechoic (left) and reverberant (right) conditions, for three competing sentences (one target, two interferers) separated by a $\Delta F0$ of 0 (squares) or 8 (triangles) semitones and for three different modulation widths (0, ± 1 , ± 2 semitones). Lower thresholds indicate greater intelligibility. Error bars are ± 1 standard error of the mean.

Figure 5.5 presents the same data by plotting the $\Delta F0$ benefits across rooms and modulation widths. Tukey pairwise comparisons revealed that F0 modulation reduced the $\Delta F0$ benefit for smaller modulation widths in the reverberant than in the anechoic room. Indeed in the anechoic room, the $\Delta F0$ benefit for a ± 2 -semitones modulation was significantly smaller than that for ± 0 ($q=6.5$) or ± 1 semitone ($q=4.3$), whereas in the reverberant room, the $\Delta F0$ benefit for a ± 0 -semitone modulation was significantly bigger than that for ± 1 ($q=5.2$) or ± 2 semitones ($q=7.7$). Put another way, the benefit at ± 0 and ± 1 semitone modulation was not significantly different in anechoic conditions whereas the benefit at ± 1 and ± 2 semitones modulation was not significantly different in reverberant conditions.

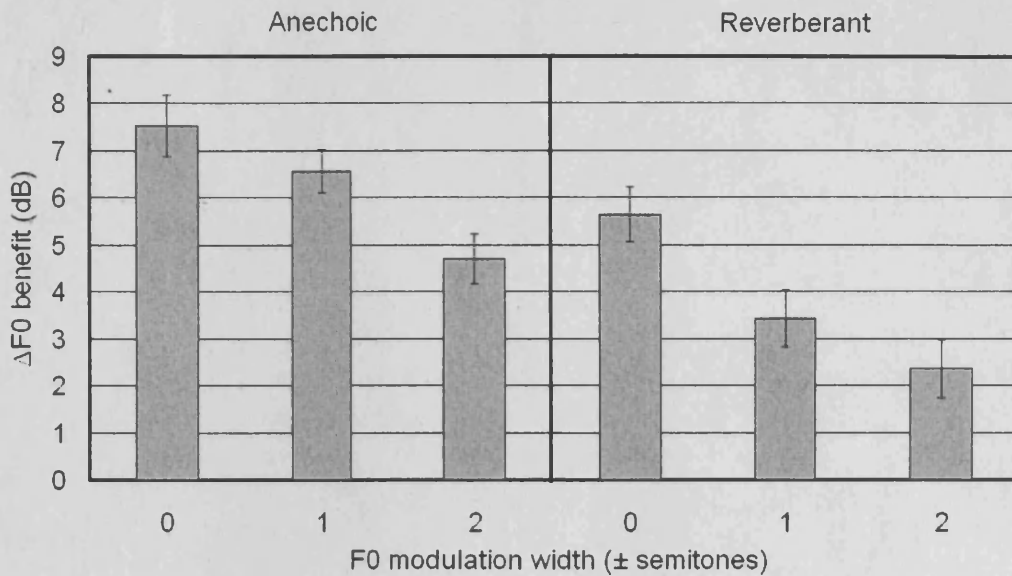


FIG. 5.5 Mean $\Delta F0$ benefits measured in anechoic (left) and reverberant (right) conditions, between one target voice and two interfering voices separated by an 8-semitones $\Delta F0$ and for three different modulation widths (0, ± 1 , ± 2 semitones). Error bars are ± 1 standard error of the mean.

C. Discussion

This experiment shows three key results which are further developed. First, harmonic cancellation provides a weak contribution to the $\Delta F0$ benefit, confirming the conclusion of Chap. III. Second, there is a floor effect of reverberation with modulated speech, confirming the discussion of Exp. 5.1. Third, reverberation affects sequential F0-grouping, replicating the effect observed in Exp. 5.2.

C.1 Weak contribution of harmonic cancellation

A ± 2 -semitones F0 modulation impairs SRTs by about 3 dB for an 8-semitones $\Delta F0$, similar to the impairment observed in Exp. 5.1 for a 2-semitones $\Delta F0$ (Exp. 5.1). Exp. 4.1 and Exp. 5.1 have shown that this impairment reflects the

disruption of harmonic cancellation, because F0 modulation of the interferer, not the target, is detrimental. Consequently, harmonic cancellation must provide only a weak and constant benefit, at 2- and 8-semitones $\Delta F0$ for speech interferers; what we previously knew for a buzz interferer (Chap. III).

C.2 Floor effect of reverberation

Figure 5.5 showed that a F0 modulation of ± 1 semitone reduced the benefit by 1 dB in anechoic conditions, but 2 dB in reverberant conditions. The temporal resolution of the harmonic cancellation was slightly sluggish to follow the rate of modulation of the interferer, modulated by ± 1 semitone. As a result, the modulated interferer was cancelled less effectively than a monotonized interferer; the harmonic benefit being not completely lost yet, reverberation exacerbated the disruption of the cancellation mechanism. Tukey pairwise comparisons confirmed that the $\Delta F0$ benefit was reduced for smaller modulation widths in reverberant than in anechoic conditions. In other words, as the interfering voice becomes modulated, harmonic cancellation is disrupted sooner in reverberant than in anechoic conditions. When the speech interferer was modulated by ± 2 semitones, the rate of modulation was too quick for the mechanism that no longer cancelled the anechoic interferer and so the effect of reverberation could not occur. Therefore, reverberation could potentially disrupt the cancellation mechanism provided that the mechanism still provides a benefit. A F0 modulation of ± 2 semitones at 5 Hz prevents the mechanism from being beneficial, so reverberation has a floor effect.

C.3 Reverberation affects sequential F0-grouping

The $\Delta F0$ benefit was smaller in reverberant than in anechoic conditions, whatever the modulation width. Crucially, when voices were monotonized, i.e. when

harmonic cancellation was not challenged, the $\Delta F0$ benefit was about 2 dB smaller in reverberant than in anechoic conditions. When speech was ± 2 -semitones modulated, the harmonic benefit was presumably lost; the $\Delta F0$ benefit was again about 2 dB smaller in reverberant than in anechoic conditions. This replicated and extended the result of Exp. 5.2, namely reverberation affects the sequential F0-grouping, regardless of F0 modulation.

C.4 F0 modulation effects restricted to harmonic cancellation

In anechoic as in reverberant conditions, F0 modulation had no effect, when there was no $\Delta F0$. This confirmed that monotonized and F0-modulated speech is equally intelligible. When there was a $\Delta F0$, the detrimental effect of F0 modulation corresponded in magnitude with the contribution of harmonic cancellation, suggesting that F0 modulation affects only harmonic cancellation.

C.5 Additional effects of reverberation, regardless of F0

SRTs increased in reverberation, due to the degradation of the amplitude modulations of the target (STI effects). The use of a 2-voice interferer considerably reduced the role of dip-listening, so the reduction in dip-listening due to the tails of reverberation was likely to be a weak effect. These effects were not the focus of interest of the present experiments, but occurred as additional effects of reverberation.

V. Conclusion

Three experiments investigated the effects of reverberation and F0 modulation on the $\Delta F0$ benefit found with speech interferers. As with a buzz interferer, F0 modulation of a speech interferer in anechoic conditions led to a small impairment in SRTs and the F0 modulation of the target had no effect (Exp. 5.1). This effect could be explained by a limited temporal resolution of the cancellation mechanism beyond which a rapid F0 modulation blurs the harmonicity of the interferer which becomes less subject to cancellation.

As shown for a buzz interferer in Chap. IV, reverberation potentially disrupts harmonic cancellation when combined with F0 modulation, provided that this mechanism is at all beneficial. But because with speech interferers, the cancellation mechanism seems weak (perhaps because it only works on the voiced parts of speech) and is already abolished by a rapid F0 modulation, reverberation had a floor effect. At either 2-semitones (Exp. 5.1) or 8-semitones $\Delta F0$ (Exp. 5.3), reverberation did not interact with F0 modulation and the loss of intelligibility due to a modulated interferer saturated at 2 or 3 dB. The introduction of a modulation of ± 1 semitone (Exp. 5.3) supported the proposal that as long as there is some harmonic benefit, it seems more affected by a modulated interferer in reverberant than in anechoic conditions. But at ± 2 -semitones, F0 modulation reduced the $\Delta F0$ benefit similarly in both rooms.

Reverberation did reduce by about 2 dB the benefit of an 8-semitones $\Delta F0$ for speech interferers, when harmonic cancellation was not challenged (monotonized speech in Exp. 5.2 and Exp. 5.3) as well as when harmonic cancellation was abolished (± 2 -semitones modulated speech in Exp. 5.3). We reasoned that the only way to

reduce the ΔF_0 benefit, regardless of harmonic cancellation, is to affect sequential F_0 -grouping. Therefore reverberation must affect the sequential F_0 -grouping mechanism. At the moment, it is hard to understand how this effect is mediated. Reverberation might produce some increase in cognitive demands, reducing the ability to focus ones attention on the pitch of a target voice.

In Exp. 5.1, reverberation caused additional impairment when the target and the interfering voices were both modulated. This effect presumably reflects another impairment of sequential F_0 -grouping because the harmonic benefit was likely eliminated so only the contribution of sequential F_0 -grouping was left; it did not occur for a buzz interferer.

A different degree of reverberation can also be used as a grouping cue to segregate the target from the interferer and get a large release from masking. Such releases did not occur for a buzz interferer (Chap IV), suggesting that they occur because the auditory scene is difficult to organise when speech interferers are used. In the presence of three voices, sequential grouping can use a difference in F_0 but also a difference in the degree of reverberation to group sounds into coherent streams. Several sequential cues might have different relevance, depending on the clarity of the scene that they can provide.

Chapter VI.

GENERAL DISCUSSION

The present thesis has enhanced our understanding of the mechanism of harmonic cancellation and provided evidence that this mechanism takes place in speech segregation by cancelling interfering voices on the basis of their harmonic structure. In addition, speech segregation likely depends also on the sequential grouping of components of each voice on the basis of their F0. This second mechanism is less well understood but appears to play a substantial role with speech interferers. It is possible that this second mechanism requires listeners to consciously perceive that two voices are spoken with distinct pitches and attend to one of them, i.e. it might reflect a sequential grouping by pitch rather than by F0 per se. As a consequence the mechanism responsible for pitch perception may also be involved in the segregation of voices by F0. Therefore the present chapter discusses the contribution firstly of harmonic cancellation, secondly of sequential grouping by F0 and thirdly of pitch perception in the $\Delta F0$ effect observed experimentally and in the literature, before generalising in a fourth and fifth part to real speech and realistic environments.

I. Harmonic cancellation in the $\Delta F0$ effect

Two results of Chap. II are most important for the understanding of the mechanism of harmonic cancellation and give valuable insight into its involvement in the $\Delta F0$ effect observed experimentally for both vowels and speech. First, MDTs were lower for a harmonic than an inharmonic masker when the target band was located in the region 0.5-2.5 kHz (Exp. 2.1). Although auditory nerves cannot phase-lock at a

rate above about 3 kHz, the auditory nerves located at the centre frequencies above 3 kHz can still phase-lock at a rate of the F0 of the masker. So it is not trivial why these high-frequency neurons cannot be inhibited at the masker's fundamental period. It may be interesting to further examine the parallel with two other phenomena: the pitch of a mistuned partial (Hartmann et al., 1990) and the spectral edge pitch (Kohlrausch and Houtsma, 1992). Both phenomena showed that thresholds for harmonic and inharmonic complexes converge between 2.5 and 3 kHz; result which has been attributed to a loss of phase-locking. Second, the mechanism integrates harmonic information over a wide band to cancel masker partials in the target region (Exp. 2.4). Therefore harmonic cancellation is clearly an across-frequency mechanism (Roberts and Holmes, 2006) that cancels frequency components up to 2.5 kHz. In the light of such a mechanism, we can refine its contribution to the $\Delta F0$ effect in double-vowels identification, and intelligibility of target speech masked by a buzz or speech interferers.

A. Contribution of harmonic cancellation for double-vowels

When competing vowels have different F0s, several mechanisms are at work, their relative roles being dependent on the size of $\Delta F0$. Pitch-period asynchrony could play a role when both sources have very low F0s (Summerfield and Assmann, 1991). Part of the $\Delta F0$ benefit for vowels is due to waveform interactions resulting from the beating between unresolved components which produce a spectral amplitude modulation beneficial to vowel identification (Assmann and Summerfield, 1994; Culling and Darwin, 1994). This latter mechanism can only occur for small $\Delta F0$ s, less than one semitone. It seems to explain two major experimental results: the improvement in identification is smaller for 50- than for 200-ms stimuli (Assmann and Summerfield, 1990) and the fact that the $\Delta F0$ benefit originates primarily from

the first formant region (Culling and Darwin, 1993; Rossi-Katz and Arehart, 2005). Given that low-frequency beating can be predominant for those small ΔF_0 s, it is difficult to evaluate the contribution of harmonic cancellation for such ΔF_0 s.

At one semitone and above, Culling and Darwin (1994) showed that identification of “normal” vowels was significantly better than that of “interleaved” vowels, indicating that beating could not explain the full benefit of this ΔF_0 . Harmonic cancellation is a good candidate to underpin the benefit at those ΔF_0 s. As presented in the next sections B and C, harmonic cancellation might saturate at 2 semitones or less, which potentially explains why the ΔF_0 benefit asymptotes above 1 semitone for double-vowel identification. This interpretation is also consistent with the results of Culling and Darwin (1993) and Rossi-Katz and Arehart (2005) showing that across-formant inconsistencies in F_0 reduced identification for relatively large ΔF_0 s (2-9 semitones). Since harmonic cancellation is an across-frequency mechanism, across-formant inconsistencies may be indicative of the presence of harmonic cancellation: they appeared to have an effect only for ΔF_0 s above 1 semitone, where the beating mechanism is no longer of use.

B. Contribution of harmonic cancellation for a buzz interferer

The results of Chap. IV showed that SRTs were disrupted to a small extent when a modulated anechoic buzz interferer was used and to a greater extent when a modulated reverberant buzz interferer was used. Those effects could only occur if the mechanism was tuned to the interferer’s F_0 , and were thus consistent with harmonic cancellation of the interferer. Because this impairment in SRTs corresponded in magnitude to the size of the ΔF_0 benefit, we reasoned that harmonic cancellation is entirely responsible for the ΔF_0 benefit between target speech and a buzz interferer.

This reasoning was then confirmed by the absence of sequential F0-grouping for a buzz interferer (section II.A). The ΔF_0 benefit for a buzz interferer was almost constant at 2 and 8 semitones, suggesting that harmonic cancellation of a buzz interferer saturates at 2 semitones or less.

The investigation of the temporal resolution of the mechanism is required not only to improve a model of harmonic cancellation but also to test our interpretation that F0-modulation of an anechoic interferer disrupted the mechanism because the rate of modulation exceeded its temporal resolution. One simple way to explore further the temporal resolution of harmonic cancellation is to change the rate of modulation of the interferer's F0 contour and observe at which rate the mechanism is disrupted.

C. Contribution of harmonic cancellation for a speech interferer

Harmonic cancellation operates only on the voiced parts of speech, so that the more continuous the F0, the more effective the cancellation. For this reason, the ΔF_0 benefit was smaller for speech interferers than for a buzz interferer with a continuous F0 (Exp. 3.2). It would be interesting to test whether a set of intermittently voiced sentences would be more robust to cancellation than a set of continuously voiced sentences. F0 modulation of the speech interferers disrupt harmonic cancellation and had the same detrimental effect at 2 as at 8 semitones: a 2 or 3-dB impairment. Therefore, the contribution of harmonic cancellation is relatively weak with speech interferers and also saturates at 2 semitones or less.

D. Is the cancellation mechanism more generally applicable to spectral templates regularly spaced in frequency?

There are types of inharmonic stimuli, not used in the present experiments, which may be particularly interesting for further study: frequency-shifted complexes (a fixed offset is added to the frequency of each partial of a harmonic series) and spectrally-stretched complexes (a cumulative increment is added to the frequency spacing of the partials with increasing partial number). Such complexes are not periodic but still present some form of spectral regularity. For reasons detailed below, the question arises as whether or not the cancellation mechanism can cancel such regular inharmonic complexes and would be directly testable by the present paradigm. Such experiments will give major insights into the process of perceptual suppression, particularly whether it operates in the time or frequency domain. Typically it is sensible to regard harmonic cancellation as a temporal mechanism if it only deals with purely periodic signals. The possible findings that some inharmonic complexes could be cancelled similarly than harmonic ones will probably cast difficulties for temporal mechanisms. In contrast, cancellation of a spectral template regularly spaced in frequency might conciliate more easily with such findings.

There are two reasons to think that regular inharmonic complexes could be segregated in the same way than harmonic complexes are. First, Roberts et al. (2010) have used concurrent sentences with harmonic or frequency-shifted excitation: for both types of speech, they found a similar increase in performance with ΔF_0 . Since we believe harmonic cancellation to be involved in the ΔF_0 benefit, we must conclude that this cancellation mechanism also operates on regular inharmonic stimuli. Second, studies on the mechanisms underlying the perception of pitch have well established the concept of harmonic sieves: when a harmonic complex consists of many partials,

one hears only the first partial, the fundamental frequency. This effect occurs presumably because the partials of the harmonic series are internally suppressed. This perceptual suppression can be observed when trying to tune a pure tone to match a probe in a spectral gap, embedded by a harmonic complex, listeners made more mistakes when the probe matched harmonic positions than other positions in the spectrum (Brunstrom and Roberts, 1998), because it is more difficult to perceive the frequency of a partial when it falls into this harmonic template. More interestingly, the same effect can be found using regular inharmonic complexes, where the partials composing such complexes are not multiple integers of the perceived pitch of the complex. Listeners made more mistakes when the probe matched the positions of the missing partials of the template than the positions of the multiple integers of the perceived pitch or other positions (Roberts and Brunstrom, 1998; Brunstrom and Roberts, 2000, Roberts and Brunstrom, 2001). These results suggested that the perceptual suppression of the partials in a complex conforms to a mechanism related to the regularity of the template in frequency rather than a mechanism related to computation of the pitch (Roberts and Bregman, 1991; Roberts and Bailey, 1996; Roberts and Brunstrom, 2003; Roberts, 2005). Such a scheme draws a clear distinction between the mechanisms of segregation by a spectral template (to which harmonic cancellation, as explored in the present experiments, may belong) and the mechanisms of pitch perception. Regular inharmonic complexes could be segregated in the same way that harmonic complexes are, even though they produce different pitch percepts.

II. Sequential grouping in the ΔF_0 effect

A. Sequential F_0 -grouping only in speech-on-speech segregation

It was initially assumed that speech and buzz were sufficiently different for grouping to separate the two sources sequentially, without the need for a ΔF_0 . Two results suggest that as expected, there was no additional sequential grouping by F_0 for a buzz interferer. First, the experiments of Chap. III showed that the ΔF_0 benefit did not increase from 2- to 8-semitones ΔF_0 with buzz interferers, whereas the ΔF_0 benefit increased by 5 dB as ΔF_0 increased for speech interferers. This discrepancy, in the way the benefit evolves as ΔF_0 increases, suggests the presence of an additional mechanism (other than harmonic cancellation) only for speech interferers. Second, Exp. 5.2 and 5.3 showed that reverberation affects the ΔF_0 benefit when the F_0 profile of the speech was monotonized, which cannot be explained by a disruption of harmonic cancellation. When the F_0 profile of the buzz was monotonized in Exp. 4.1, the same level of reverberation had no effect. Had this second mechanism been present with a buzz interferer, reverberation would have had an effect.

B. Sequential F_0 -grouping reduced by reverberation

The reduction in the ΔF_0 benefit by reverberation when the F_0 profile of the interfering speech was monotonized (Exp. 5.2 and 5.3) must occur because part of the ΔF_0 benefit is caused by a mechanism other than harmonic cancellation. Sequential grouping by F_0 was a good candidate for this other mechanism since it appears in difficult segregation tasks like speech-on-speech masking and is progressively more beneficial as ΔF_0 increases.

The configuration where the target and interfering voices were both ± 2 -semitones modulated while separated by 2-semitones ΔF_0 , created a F_0 overlap between competing voices, which led to a 2-dB impairment in reverberation. This impairment could not be explained in terms of energetic masking because the same F_0 -overlap situation did not lead to a greater impairment for a buzz interferer. Therefore the overlap between competing F_0 s probably reduced the ability to use F_0 as a sequential grouping cue to organise the auditory scene. The absence of F_0 -overlap impairment with a buzz interferer further confirmed our assumption that speech and buzz were sufficiently segregated not to suffer from a F_0 overlap; i.e. sequential F_0 -grouping does not provide further segregation.

It remains unclear why reverberation affects sequential grouping by F_0 . Ultimately, the speech segregation task, used in the present experiments, requires listeners to store temporarily the target message. Presumably the working memory has a limited processing capacity, so that the more resources that are allocated to word identification, the fewer resources are left for storage (Kjellberg, 2004). Listeners recall a smaller number of words spoken in a background noise rather than in quiet, while their intelligibility was perfect (Kjellberg et al., 2007). Similarly listeners recall a smaller number of words spoken in reverberation than in anechoic conditions, while their intelligibility was perfect (Kjellberg, personal communication). Such a trade might apply between intelligibility and auditory attention: the more intelligible the speech, the easier it is to direct ones attention to its pitch. Reverberation impairs speech intelligibility which in turn is costly to sequential F_0 -grouping. A better understanding of sequential grouping is needed to know how best to investigate such effects.

C. Towards a better understanding of the mechanism underpinning sequential F0-grouping: parallel with spatial separation

When voices compete to be heard, a ΔF_0 seems to facilitate the sequential grouping of each voice, as though listeners could direct their attention to the pitch of the target voice. Although listeners seem able, to a certain degree, to form a separate perceptual group by focusing their attention on a specific sound (Carlyon et al., 2001), it is still unclear what the listener's attention represents. How could auditory attention favour or constrain certain mechanisms? In the case of the ΔF_0 effect, how could consciously perceiving distinct pitches help to improve performance? The concepts of selective attention or auditory awareness are quite ill-defined, leading to interpretations (like that presented in the previous section) which are speculative. It may be instructive to draw a parallel with spatial separation to further our understanding of such concepts.

Spatial separation is a cue (other than F_0) that seems to engage different mechanisms depending on whether it is used for simultaneous or sequential grouping. When target and interferer come from different directions, the target is easier to understand than when they come from the same direction. This effect is thought to be due to better-ear listening and binaural unmasking. In better-ear listening, the auditory system chooses the ear with the best signal-to-noise ratio. Mechanisms of binaural unmasking use interaural time differences (ITDs), according to two dominant theories: either sounds entering each ear are cross-correlated so that a reduction from unity of interaural coherence can be interpreted as evidence of a signal (Durlach et al. 1986, Culling and Colburn, 2001); or sounds entering each ear are aligned temporally (equalized) and then subtracted one from another (cancellation), known as

Equalization-Cancellation theory (Durlach, 1963, Culling, 2007, Lavandier and Culling, 2007, 2008). However, Freyman et al. (2001) used the precedence effect to create the illusion of a spatial separation while one masker is still spatially collocated with the target. Both better-ear listening and binaural unmasking (whatever the model) predict no spatial benefit for such situation. On the contrary, speech recognition improves with perceived separation of speech-on-speech, but not of speech-on-noise. Somehow a perceived difference in spatial location is used by the auditory system to release from a form of masking that is not energetic and only occurs with speech interferers. If we could find an equivalent of the precedence effect for F0 cues, it would become possible to examine whether a perceived (i.e. illusory) $\Delta F0$ may cause a benefit while harmonic structures of competing voices are identical. Bird and Darwin (1998) may have created this type of speech stimulus, when target and interferer had different F0s up to 800 Hz and the same F0 above. Because the region below 800 Hz dominates the perception of pitch (see next section III.A), listeners might have perceived competing voices to be spoken at distinct pitches while most of their harmonic structure was in fact common. In my opinion, this type of speech stimulus is especially interesting to examine sequential grouping by F0, in the same way than the precedence effect is interesting to examine sequential grouping by perceived spatial separation (that Freyman et al. (2001) term spatial release from informational masking).

III. Pitch perception in the $\Delta F0$ effect

A. Resolved partials dominate the perception of pitch

The fact that the $\Delta F0$ benefit originated from the first-formant region in double-vowels identification (Culling and Darwin, 1993; Rossi-Katz and Arehart,

2005) might simply be due to waveform interactions and not be relevant to mechanisms of segregation by F0. Speech, however, has a spectral envelope that is constantly changing and the waveform interactions explanation becomes very questionable in this case. It is surprising therefore that Bird and Darwin (1998) replicated these findings using monotonized speech. A critical result of Exp. 1 is that harmonic cancellation is beneficial up to 2.5 kHz. So it only makes sense that the region below 800 Hz primarily causes the $\Delta F0$ effect if this effect is attributable to sequential F0-grouping and not harmonic cancellation. The region below 800 Hz is the region where harmonics are spectrally resolved in the auditory periphery and dominate the perception of pitch (Carlyon and Shackleton, 1994; Shackleton and Carlyon, 1994; Carlyon, 1998, Gockel et al., 2005; Ives and Patterson, 2008). Although we do not yet have a clear understanding of sequential grouping by F0, it is likely to be related to the pitches perceived by the listener. In Bird and Darwin (1998), the across-frequency inconsistencies only have an effect at 5- and 10-semitones $\Delta F0$ s for speech. One interpretation is that this across-frequency mechanism is harmonic cancellation, which suggests a new reading of the results of Bird and Darwin (1998) as the following.

- When two monotonized voices are based on different F0s: sequential grouping is progressively more beneficial as $\Delta F0$ increases and harmonic cancellation contributes to the benefit for 5- and 10-semitones $\Delta F0$.
- Two voices, based on different F0s but having the same F0 above 800 Hz, create a form of ' $\Delta F0$ illusion'. Low-ordered resolved harmonics may still produce distinct pitches, so that sequential grouping remains progressively beneficial as $\Delta F0$ increases. In contrast, the contribution of harmonic cancellation is weakened since it is restricted to the region below 800 Hz. The data show that

compared to unprocessed voices, speech recognition is only slightly reduced for 5- and 10-semitones ΔF_0 below 800 Hz, suggesting that sequential grouping is the main cause of the ΔF_0 effect for speech.

- Two voices, based on different F_0 s but having the same F_0 below 800 Hz, create a harmonic separation without a perceived pitch difference. Low-ordered resolved harmonics are identical so that both voices may be fused on the same pitch. As a result, sequential grouping rather groups together the two voices instead of separating them, offsetting a potential benefit of harmonic cancellation over most of the spectrum. The data showed that no benefit occurred, suggesting that sequential grouping is necessary for the ΔF_0 effect to occur.
- When the two voices swapped their F_0 s above 800 Hz, low-ordered resolved harmonics may still produce distinct pitches but the masker's pitch perceived by the listener mainly matches with the harmonic structure of the target and vice-versa above 800 Hz. If the mechanism is tuned to the perceived masker's pitch, the harmonic structure of the target (not that of the masker) is cancelled above 800 Hz. In the presence of such F_0 -swapped voices, harmonic cancellation might be detrimental to speech recognition. The data showed that speech recognition increased from 0- to 2-semitones ΔF_0 and dropped at 5- and 10-semitones ΔF_0 .

Thus Culling and Darwin (1993), Rossi-Katz and Arehart (2005) and Bird and Darwin (1998) may have produced similar results with vowels and speech but for different reasons. For double-vowels, low frequencies may underlie waveform interactions for very small ΔF_0 s. For speech, resolved harmonics dominate the percept of pitch, which may interact with both harmonic cancellation and sequential

F0-grouping. Harmonic cancellation is likely to have occurred in both studies, but for larger ΔF_0 s where across-frequency inconsistencies appeared to have an effect.

B. Temporal integration required for sequential F0-grouping

The F0-overlap impairment for speech in reverberation, presented in II.B., reflects a loss of sequential F0-grouping. When competing voices had overlapping F0s, at 5-Hz modulation frequency, the competing F0s were often separated by only a quarter of a period, i.e. 50 ms. The tails of reverberation (longer than 50 ms) caused a temporal overlap between the competing F0s, which fused the competing pitches together, resulting in a loss of sequential F0-grouping. This loss did not occur for speech in anechoic conditions (Exp. 5.1), suggesting that modulated pitch contours are still perceived as separated when they are distant by 50 ms in anechoic conditions. The temporal integration of F0 must then be less than 50 ms otherwise the competing pitches would have been perceived as fused in anechoic conditions as well.

C. Size of ΔF_0 required for listeners to perceive distinct pitches

If the presence of distinct pitches enables listeners to group words from a target sentence sequentially on the basis of their F0, one question arises: as from which ΔF_0 do listeners start perceiving that competing sounds are spoken on distinct pitches. Assmann and Paschall (1998) used a matching task to obtain judgments of the pitches evoked by double-vowels. For ΔF_0 s up to 2 semitones, listeners could only match a single pitch whereas they were able to match two distinct pitches at 4 semitones. Their results contrast with the idea that competing pitches can be perceived as distinct for small ΔF_0 s. However, such an interpretation ignores the contribution of other factors. For instance for vowels, onset cues group the vowels together so the F0s of simultaneous vowels must be set further apart for listeners to

perceive them on distinct pitches. In contrast for connected speech, onset cues facilitate the separation of competing words so that a smaller ΔF_0 may be sufficient for sequential grouping by F_0 to take place. Consistent with this idea, Rasch (1978) found that onset asynchronies as small as 10 ms between temporally overlapping tones facilitated the segregation of these two tones. Therefore, when ΔF_0 is consistent with other cues, typically temporal cues, listeners can probably segregate sounds that are separated by smaller ΔF_0 s.

IV. Application to real speech

A. Real speech

A.1 Natural intonation

One major characteristic of real speech was absent in our speech stimuli: natural variations of F_0 . In cooperation with intensity and duration variations, prosodic information conveyed by an intonated F_0 contour helps to emphasize the stresses or de-emphasize the end of words in an utterance (Fry, 1955; Cooper and Sorensen, 1981; Freeman, 1982). Consistent with the idea that prosody aids intelligibility, Culling et al. (2003) showed that, against speech interferers, natural speech is more intelligible than monotonized speech. Therefore one might expect the use of real speech to lower SRTs for all conditions.

Interestingly, Peng et al. (2009) have found that normal-hearing listeners can still recognise intonation when the F_0 contours and intensity patterns are conflicting with each other, but not for spectrally degraded stimuli (similarly to those experienced by cochlear implanted listeners). These results suggest that the coherence of prosodic

cues (F0 contours and intensity patterns) may be especially useful for difficult listening situations.

A.2 Harmonic cancellation on real speech

How would the use of real speech affect harmonic cancellation? Manipulating the F0 contour of speech does not modify its rhythm or its amplitude envelope. The unvoiced parts of our speech stimuli (monotonized or F0-modulated) occur in the same way as with real speech. Therefore, the contribution of harmonic cancellation may also be weak with real speech. Moreover, harmonic cancellation is disrupted when the interferer's F0 is modulated (Exp. 5.1 and 5.3). The human voice varies rapidly in F0 (up to 5 oct/s) over a full octave during normally intonated speech (O'Shaunessy and Allen, 1983). On one hand, the temporal resolution of harmonic cancellation might be too sluggish to follow the rate of modulation of some speech segments. On the other hand, the fast variations in F0 of real speech create many instantaneous ΔF_0 s between competing voices. Those instantaneous ΔF_0 s might lead to unexpected benefits of harmonic cancellation. It is difficult at the moment to evaluate more accurately the contribution of harmonic cancellation for real speech.

As discussed in Chap. II, it might also be interesting to further develop the parallel between the perceptual suppression of the partials in a harmonic sieve and harmonic cancellation as examined in the present tasks. Given the importance of such spectral sieves in sounds segregation, it is wise to conclude that the impact of harmonic cancellation even for naturally intonated speech should not be dismissed.

A.3 Sequential grouping on real speech

How would the use of real speech affect sequential F0-grouping? Darwin and Hukin (2000) used the CRM design and found that a fixed 4-semitone $\Delta F0$ or natural prosodic cues (necessarily accompanied by instantaneous $\Delta F0$ s) helped listeners to group the target sentence sequentially but were not strong enough to override an inconsistent spatial separation. When a difference in vocal-tract size was combined with $\Delta F0$ s and prosody, listeners were much less influenced by an inconsistent spatial separation. Using the same design, Darwin et al. (2003) showed that the benefit of $\Delta F0$ s combined with differences in vocal-tract size for sequential grouping was bigger than the summed individual benefits. Therefore sequential grouping might be especially robust or beneficial when several sequential cues are consistent with each other (super-additivity). In real situations of conversation, different talkers have different F0s combined with different vocal-tract lengths and combined with different voice characteristics, so that sequential grouping is probably reinforced and its benefits enhanced.

B. Multiple talkers

B.1 Harmonic cancellation of multiple F0s

It is likely that the auditory system is incapable of making more than one cancellation of F0. Hawley et al. (2004) used one, two or three interferers that were either voiced (speech and reversed-speech) or unvoiced (noise and modulated noise). In the case of multiple voiced interferers, each interferer had a different F0. They found that SRTs were lower for a single voiced interferer than a single unvoiced interferer, presumably due to the benefit of harmonic cancellation. However SRTs were higher for 2 or 3 voiced interferers than 2 or 3 unvoiced interferers. Although

informational masking could partly explain this pattern of data, it is likely that multiple rounds of cancellation either cannot occur or distort drastically the target waveform by comb-filtering effects, resulting in poor target intelligibility. The fact that harmonic cancellation can integrate harmonic information over very large bands (up to 2.5 kHz), is another indication that cancellation of multiple F0s is ultimately detrimental to the target. One approach could consist in applying multiple comb-filters to the target, using a tolerance of about ± 12 Hz on each partial (Exp. 2.3), and observe for which round the target becomes unintelligible.

B.2 Sequential grouping of multiple F0s

Are listeners able to group sounds sequentially into more than two streams on the basis of three distinct pitches? Musicians certainly appear to be and any listener is often performing this kind of task to segregate several voices. However, the dependency of such abilities on the presence of other grouping cues is an important factor to consider. At the moment, it is difficult to estimate the robustness of the listener's ability to track a particular voice on the basis of its pitch, because other cues (onset, timbre, voice characteristics) may also be used. Further research in this issue should closely examine mechanisms of pitch perception which still remain a matter of debate.

V. Application to realistic environments

A. Degree of reverberation of different rooms

The simulated reverberant room was designed with an absorption coefficient of 0.3 on each internal surface. Since F0-modulation of speech interferers abolished the benefit of harmonic cancellation (Exp. 5.1 and Exp. 5.3), the additional effect of

reverberation on harmonic cancellation was at floor. As long as reverberation is at floor, it may not be worth examining the effects of different degrees of reverberation.

In contrast, regarding sequential F0-grouping, Darwin and Hukin (2000b, Exp. 1) showed that sequential F0-grouping was increasingly impaired as reverberation increased. In the present thesis, we used a single reverberant environment, so we could not test the idea that sequential grouping would be more impaired with a higher level of reverberation. Given that the mechanism underlying sequential grouping is not well understood, it might give some interesting perspectives to challenge the mechanism with different types of reverberation, for instance from different sizes of room.

B. Room colouration

In the present experiments, reverberant stimuli were filtered to have the same excitation pattern as their corresponding anechoic stimuli. This process enabled the suppression of spectral colouration produced by the room. The harmonic structure of a particular source may be amplified or attenuated by the room, distorting signal-to-noise ratios and consequently distorting ΔF_0 effects, especially when using monotonized speech where harmonic frequencies were fixed. With naturally intonated speech, some syllables or words might be amplified or attenuated but randomly with respect to the target or the interferer and in consequence, the effects of room colouration are thus likely to be neutral.

C. Sources directionality and dummy head

Other aspects of realistic environments would influence the results. The directionality of sources and receivers would increase the direct-to-reverberant ratio as long as sources are facing the receiver. In most listening situations, the target

speaker would probably be facing the receiver, but the interfering speakers would not. Consequently, the direct-to-reverberant ratio of the target would increase, but that of the interferers would decrease. Since sequential F0-grouping is affected by reverberation (Exp. 5.2 and 5.3), the question arises as whether it would be more or less robust with directional sources. The answer to this question relies on examining whether sequential F0-grouping is primarily disrupted by reverberation on the target or on the interferer or on both. Such an experiment would also give a further understanding of this impairment.

The presence of a head leads to spectral differences between competing sources in the same way that room colouration does, especially problematic for our monotonized stimuli. Since these effects would not appear for real speech, it was pointless to use a head; we would have had to decolour the head as well. The presence of a head also produces head-shadow effects with spatially separated sources, which results in changes in target-to-masker ratios. But there is no head-shadow when sources are co-located in front of the listener; so once again, the simulation of a head was pointless.

D. Spatial separation

In realistic rooms, competing sources are spatially separated so better-ear listening and binaural advantages must be taken into account. It is noteworthy to underline the fact that studies on the effect of a $\Delta F0$ have often involved monaural or diotic listening, since $\Delta F0$ effects are generally thought to be monaural. Consistently in this study, collocated sources were used to prevent any benefit from different positions mixing with the benefit of $\Delta F0$. It appeared not to be sufficient, as interaural phase effects still emerged due to the asymmetrical position of the sources in the reverberant room, decorrelating the signals across the two ears (Exp. 4.1 and 5.1). If

the ΔF_0 effects are indeed purely monaural, then the present results could be applicable to spatially separated sources. However, in a “double-vowel” identification task, Shackleton et al. (1994) then Culling et al (1994) found that binaural differences due to spatial separation enhance accuracy of identification only when there is also a ΔF_0 . Later Hawley et al. (2004) obtained larger ΔF_0 effects (if defined as the difference between the reversed-speech and modulated noise conditions) with binaural than monaural stimuli, using only one interferer. The cause of those interactions remains unclear. Thus some precautions are required before claiming the present results are true for spatially separated sources.

E. Future challenges towards architectural software

We are at the very first steps of a model that would predict the benefits of harmonic cancellation. Taking into account sequential grouping by F_0 might be harder to model if such an ability is a higher-level mechanism in the auditory system. So, we may be quite far from being able to predict ΔF_0 advantages in adverse listening conditions. In the binaural domain, Lavandier and Culling (2008) developed a model of speech understanding in noise and reverberation based upon a development of equalization-cancellation theory and managed to predict speech intelligibility fairly well in a wide range of spatial configurations, in different rooms, (extended to several noise interferers at different positions in a forthcoming paper). Extending this model to modulated noise interferers should not be problematic from the binaural point of view, because dip-listening is probably based upon monaural cues. Once the mechanisms responsible for ΔF_0 effects have received as much attention as those in the binaural domain, the picture may become much clearer and hopefully some modelling will be available for the monaural domain as well.

Architectural software is now able to generate impulse responses for very realistic rooms with complex shapes and detailed furniture. So, future challenges will not concern the genuineness of the rooms, but the understanding of the psychophysical and physiological mechanisms of auditory perception.

CONCLUSION

The aim of this study has been to investigate the impact on intelligibility of differences in fundamental frequency between competing voices in a reverberant room. Since the ΔF_0 effect, on its own, represents a complex topic of research, conclusions are presented separately from the effects of reverberation.

A ΔF_0 was proposed by Cherry (1953) as one of the main acoustic cues that may facilitate speech recognition in adverse listening conditions, like a cocktail-party situation. After almost 60 years of hearing research, the mechanisms underlying these effects still remain a matter of debate. The present thesis attempted to clarify the different mechanisms involved in this effect. What have we learned?

- A mechanism of harmonic cancellation (de Cheveigné et al., 1995, 1997a, 1997b) explains why a 100-Hz wide band of noise is better detected when masked by a harmonic masker than an inharmonic masker. This mechanism can cancel maskers' partials up to 2.5 kHz. Harmonic cancellation is insensitive to a degree of inharmonicity reflected by a peak autocorrelation value of 0.97, but sensitive to a degree of inharmonicity reflected by a peak autocorrelation value of 0.9 or less. Harmonic cancellation integrates harmonic information over very large bands, highlighting the across-frequency nature of this mechanism. As well as the masker's harmonicity, detection depends on target-to-masker ratios at the output of a range of cochlear filters surrounding the target.
- The ΔF_0 benefit observed when target speech was masked by a temporally uninterrupted buzz interferer originates entirely from harmonic cancellation. This benefit did not increase from 2- to 8-semitones ΔF_0 , suggesting that the

contribution of harmonic cancellation quickly saturates at ΔF_0 s of 2 semitones or less.

- The ΔF_0 benefit observed when target speech was masked by speech interferers originates from harmonic cancellation and a sequential grouping mechanism. Because harmonic cancellation operates only on the voiced portions of speech, its contribution is weaker for speech interferers than for a buzz interferer. Contrary to harmonic cancellation, sequential grouping contributes increasingly to the benefit as ΔF_0 increases.
- It has been shown that the ΔF_0 benefit originates largely from low frequencies and we have shown that harmonic cancellation is not restricted to low frequencies. As a consequence, harmonic cancellation may not be the primary cause of the ΔF_0 effect, or at least not for small ΔF_0 s. The mechanisms, most appropriate to explain the ΔF_0 effect between vowels, are beating for ΔF_0 s up to about $\frac{1}{2}$ a semitone and harmonic cancellation for larger ΔF_0 s. Harmonic cancellation quickly saturates, consistent with the fact that ΔF_0 effect saturates at 1 or 2 semitones for double-vowels. The mechanisms, most appropriate to explain the ΔF_0 effect between voices, are sequential grouping by F_0 and harmonic cancellation for larger ΔF_0 s.

The effects of reverberation had been intensively investigated in the case of the transmission of a single voice in quiet, but were less well known in multi-talker communication. The present thesis clarified that reverberation affects both mechanisms responsible for the F_0 -segregation of speech. What have we learned?

- When the interferer's F0 is fixed, reverberation does not affect harmonic cancellation.
- When the interferer's F0 is modulated, the contribution of harmonic cancellation is disrupted to a small extent presumably because the mechanism is sluggish. If the contribution of harmonic cancellation is initially large (for example for a buzz interferer whose F0 is continuous), then reverberation may exacerbate its disruption due to the presence of several F0s for the same interferer, owing to temporal smearing. If the contribution of harmonic cancellation is initially weak (for speech interferers whose F0s are temporally interrupted), then the detrimental effect of reverberation may already be at floor.
- Whether the interferer's F0 is fixed or modulated, reverberation affects sequential grouping by F0, provided that sequential grouping is substantially beneficial, i.e. for a large $\Delta F0$.

Prospects for further research might focus on the following issues.

- Is harmonic cancellation still effective with frequency-shifted or spectrally-stretched complexes? The answer might give interesting ideas as to whether the cancellation mechanism operates in the time or the frequency domain.
- Harmonic cancellation is ineffective above 3 kHz, where phase-locking is known to be lost in auditory processing. The relation between these two experimental observations might give interesting ideas as to how neurons coding for the masker's partials are inhibited.
- What is the relation between the perceived pitch of the masker and harmonic cancellation?

- How are the effects of harmonic cancellation and sequential grouping influenced by the number of talkers and their spatial location?
- What is the influence of auditory attention on harmonic cancellation and on sequential F0-grouping?
- Are harmonic cancellation and sequential F0-grouping compromised to the same extent in hearing-impaired listeners?

References

- Allen, J. B. and Berkley, D. A. (1979). "Image method for efficiently simulating small-room acoustics", *J. Acoust. Soc. Am.* **65**, 943-950.
- American National Standards Institute (1997). "Methods for Calculation of the Speech Intelligibility Index".
- Assmann, P. F., Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies", *J. Acoust. Soc. Am.* **88**, 680-697.
- Assmann, P. F., Summerfield, Q. (1994). "The contribution of waveform interactions to the perception of concurrent vowels", *J. Acoust. Soc. Am.* **95**, 471-484.
- Assmann, P. F., Paschall, D. D. (1998). "Pitches of concurrent vowels", *J. Acoust. Soc. Am.* **103**, 1150-1160.
- Binns, C. (2007). "The role of prosodic cues in speech intelligibility". PhD. Thesis, Cardiff University, UK.
- Bird, J., and Darwin, C.J. (1998). "Effects of a difference in fundamental frequency in separating two sentences", in *Psychophysical and Physiological Advances in Hearing*, edited by A.R. Palmer, A. Rees, A.Q. Summerfield, and R. Meddis (Whurr, London), pp. 263-269.
- Bregman, A. S., and Campbell, J. (1971). "Primary auditory stream segregation and perception of order in rapid sequences of tones", *J. Exp. Psychol.* **89**, 244-249.
- Bregman, A. S. (1990). "Auditory Scene Analysis: The perceptual organization of sound", MIT, Cambridge, MA.
- Brokx, J., Nootboom, S. (1982). "Intonation and the perceptual separation of simultaneous voices", *Journal of Phonetics.* **10**, 23-36.

- Brungart, D. (2001). "Informational and energetic masking effects in the perception of two simultaneous talkers", *J. Acoust. Soc. Am.* **109**, 1101-1109.
- Brungart, D., Simpson, B., Ericson, M. and Scott, K. (2001). "Informational and energetic masking effects in the perception of multiple simultaneous talkers", *J. Acoust. Soc. Am.* **110**, 2527-2538.
- Brunstrom, J. M., and Roberts, B. (1998). "Profiling the perceptual suppression of partials in periodic complex tones: Further evidence for a harmonic template", *J. Acoust. Soc. Am.* **104**, 3511-3519.
- Brunstrom, J. M., and Roberts, B. (2000). "Separate mechanisms govern the selection of spectral components for perceptual fusion and for the computation of global pitch", *J. Acoust. Soc. Am.* **107**, 1566-1577.
- Carlyon, R. P. (1991). "Discriminating between coherent and incoherent frequency modulation of complex tones," *J. Acoust. Soc. Am.* **89**, 329-340.
- Carlyon, R. P., and Shackleton, T. M. (1994). "Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms?", *J. Acoust. Soc. Am.* **95**, 3541-3554.
- Carlyon, R. P. (1998). "Comments on "A unitary model of pitch perception"", *J. Acoust. Soc. Am.* **104**, 1118-1121.
- Carlyon, R. P., Cusack, R., Foxtton, J. M., and Robertson, I. H. (2001). "Effects of attention and unilateral neglect on auditory stream segregation," *J. Exp. Psychol.: Human Percept. Perform.* **27**, 115-127.
- Cherry, E. C. (1953). "Some experiments on the recognition of speech with one and two ears", *J. Acoust. Soc. Am.* **25**, 975-979.

- Chalikia, M. H. and Bregman, A. S. (1989). "The perceptual separation of simultaneous auditory signals: Pulse train segregation and vowel segregation", *Perception and Psychophysics*. **46**, 487-496.
- Chalikia, M. H. and Bregman, A. S. (1993). "The perceptual segregation of simultaneous vowels with harmonic, shifted, or random components," *Perception and Psychophysics*. **53**, 125-133.
- de Cheveigné, A. (1993). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing", *J. Acoust. Soc. Am.* **93**, 3271-3290.
- de Cheveigné, A., McAdams, S., Laroche, J. and Rosenberg, M. (1995). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement", *J. Acoust. Soc. Am.* **97**, 3736-3748.
- de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (1997a). "Concurrent vowel segregation. I. Effects of relative amplitude and F0 difference", *J. Acoust. Soc. Am.* **101**, 2839-2847.
- de Cheveigné, A., McAdams, S., and Marin, C. (1997b). "Concurrent vowel segregation. II. Effects of phase, harmonicity and task", *J. Acoust. Soc. Am.* **101**, 2848-2856.
- de Cheveigné, A. (1997c). "Concurrent vowel segregation. III. A neural model of harmonic interference cancellation", *J. Acoust. Soc. Am.* **101**, 2857 -2865.
- de Cheveigné, A. (1997d). "Harmonic fusion and pitch shifts of mistuned partials", *J. Acoust. Soc. Am.* **102**, 1083-1087.
- de Cheveigné, A. (2005). "Pitch perception models", in *Pitch*, edited by C. Plack and A. Oxenham Springe-Verlag, New-York, pp. 469-233.

- de Cheveigné, A., and Pressnitzer, D. (2006). "The case of the missing delay lines: Synthetic delays obtained by cross-channel phase interaction", *J. Acoust. Soc. Am.* **119**, 3908-3918.
- Cooper, W. E. and Sorensen, J. M. (1981). "Fundamental frequency in sentence production", Springer-Verlag, New-York.
- Culling, J. F. (1996). "Signal processing software for teaching and research for psychoacoustics under UNIX and X windows", *Behav. Res. Methods Instrum. Comput.* **28**, 376-382.
- Culling, J. F. (2007). "Evidence specifically favoring the equalization-cancellation theory of binaural unmasking", *J. Acoust. Soc. Am.* **122**, 2803-2813.
- Culling, J. F., Colburn, H. S., and Spurchise, M. (2001). "Interaural correlation sensitivity", *J. Acoust. Soc. Am.* **110**, 1020-1029.
- Culling, J. F. and Darwin, C. J. (1993). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by f_0 ", *J. Acoust. Soc. Am.* **93**, 3454-3467.
- Culling, J. F., and Darwin, C. J. (1994). "Perceptual and computational separation of simultaneous vowels: Cues arising from low frequency beating", *J. Acoust. Soc. Am.* **95**, 1559-1569.
- Culling, J. F., Hodder, K., and Toh, C. (2003). "Effects of reverberation on perceptual segregation of competing voices", *J. Acoust. Soc. Am.* **114**, 2871-2876.
- Culling, J. F., and Porter, J. (2005). "Effects of differences in the accent and gender of interfering voices on speech segregation", *Auditory Signal Processing. Physiology, Psychoacoustics and Models*, written by D. Pressnitzer, A. de Cheveigné, S. Mc Adams, L. Collet.

- Culling, J. F., Summerfield, Q., and Marshall, D. (1994). "Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels", *Speech Commun.* **14**, 71-95.
- Culling, J. F., and Summerfield, Q. (1995). "The role of frequency modulation in the perceptual segregation of concurrent vowels", *J. Acoust. Soc. Am.* **98**, 837-846.
- Darwin, C., Ciocca, V. and Sandell, G. (1994). "Effects of frequency and amplitude modulation on the pitch of a complex tone with a mistuned harmonic", *J. Acoust. Soc. Am.* **95**, 2631-2636.
- Darwin, C. J. and Hukin, R. W. (2000a). "Effectiveness of spatial cues, prosody and talker characteristics in selective attention", *J. Acoust. Soc. Am.* **107**, 970-977.
- Darwin, C. J. and Hukin, R. W. (2000b). "Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention", *J. Acoust. Soc. Am.* **108**, 335-342.
- Darwin, C.J., Brungart, D.S. and Simpson, B.D. (2003). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers", *J. Acoust. Soc. Am.* **114**, 2913-2922.
- Drullman, R. and Bronkhorst, A. (2004). "Speech perception and talker segregation: Effects of level, pitch, and tactile support with multiple simultaneous talkers", *J. Acoust. Soc. Am.* **116**, 3090-3098.
- Durlach, N. I. (1963). "Equalization and cancellation theory of binaural masking-level differences", *J. Acoust. Soc. Am.* **35**, 1206-1218.
- Durlach, N. I., Gabriel, K. J., Colburn, H. S., and Trahiotis, C. (1986). "Interaural correlation discrimination: II. Relation to binaural unmasking", *J. Acoust. Soc. Am.* **79**, 1548-1557.

- Foster, J. R., Summerfield, A. Q., Marshall, D. H., Palmer, L., Ball, V., and Rosen, S. (1993). "Lip-reading the BKB sentence lists: Corrections for list and practice effects," *Brit. J. Audiol.* **27**, 233-246.
- Freeman, F. J. (1982). "Prosody in perception, production, and pathologies.," *Speech, Language and Hearing: Pathologies of Speech and Language*. Vol. 2. W. B. Saunders, Philadelphia, PA.
- Freyman, R., Balakrishnan, U., Helfer, K. (2001). "Spatial release from informational masking in speech recognition", *J. Acoust. Soc. Am.* **109**, 2112-2122.
- Freyman, R., Balakrishnan, U., Helfer, K. (2004). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition", *J. Acoust. Soc. Am.* **115**, 2246-2256.
- Fry, D. B. (1955). "Duration and intensity as physical correlates of linguistic stress", *J. Acoust. Soc. Am.* **27**, 765-768.
- Glasberg, B. R., and Moore, B. C. J. (1990). "Derivation of auditory filter shapes from notched-noise data", *Hear. Res.* **47**, 103-138.
- Gockel, H., Carlyon, R. P., and Plack, C. J. (2005). "Dominance region for pitch: Effects of duration and dichotic presentation", *J. Acoust. Soc. Am.* **117**, 1326-1336.
- Hartmann, W. M., McAdams, S., and Smith, B. K. (1990). "Hearing a mistuned harmonic in an otherwise periodic complex tone", *J. Acoust. Soc. Am.* **88**, 1712-1724.
- Hartmann, W. M., and Johnson, D. (1991). "Stream segregation and peripheral channeling", *Music Percept.* **9**, 155-183.
- Hartmann, W. M., and Doty S. L. (1996). "On the pitches of the components of a complex tone", *J. Acoust. Soc. Am.* **99**, 567-578.

- Hawley, M., Litovsky, R., and Culling, J. (2004). "The benefit of binaural hearing in a cocktail party: effect of location and type of interferer", *J. Acoust. Soc. Am.* **115**, 833-843.
- Houtgast, T. and Steeneken, H. (1985). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria", *J. Acoust. Soc. Am.* **77**, 1069-1077.
- Houtsma, A. and Goldstein, J. (1972). "The central origin of the pitch of complex tones: evidence from musical interval recognition", *J. Acoust. Soc. Am.* **51**, 520-529.
- Huggins, W. and Licklider, J. (1951). "Place mechanisms of auditory frequency analysis", *J. Acoust. Soc. Am.* **23**, 290-299.
- IEEE (1969). "IEEE recommended practise for speech quality measurements", *IEEE Trans. Audio Electroacoust.* **17**, 227-246.
- 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", *J. Acoust. Soc. Am.* **123**, 2670-2679.
- Johnson, D. H. (1980). "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones", *J. Acoust. Soc. Am.* **68**, 1115-1122.
- Kidd, G., Mason, C., Brughera, A., and Hartmann, W. M. (2005). "The role of reverberation in release from masking due to spatial separation of sources for speech identification", *Acta Acustica united with Acustica* **91**, 526-535.
- Kiefte, M. and Kluender, K., R. (2008). "Absorption of reliable spectral characteristics in auditory perception", *J. Acoust. Soc. Am.* **123**, 366-376.

- Kjellberg, A. (2004). "Effects of reverberation time on the cognitive load in speech communication: theoretical considerations", *Noise and Health* **7**, 11-22.
- Kjellberg, A., Ljung, R., and Hallman, D. (2007). "Recall of words heard in noise", *Applied Cognitive Psychology*, Wiley InterScience.
- Kohlrausch, A., and Houtsma, A. J. M. (1992). "Pitch related to spectral edges of broadband signals," *Phil. Trans. R. Soc. Lond. B*, **336**, 375-382.
- Lavandier, M. and Culling, J. F. (2007). "Speech segregation in rooms: effects of reverberation on both target and interferer", *J. Acoust. Soc. Am.* **122**, 1713-1723.
- Lavandier, M. and Culling, J. F. (2008). "Speech segregation in rooms: Monaural, binaural and interacting effects of reverberation on target and interferer", *J. Acoust. Soc. Am.* **123**, 2237-2248.
- Lea, A. (1992). "Auditory models of vowel perception", Doctoral dissertation, University of Nottingham, UK.
- Lee, A., K. C. and Shinn-Cunningham, B. G. (2008). "Effects of reverberant spatial cues on attention-dependent object formation", *J. of the Association for Research in Otolaryngology*. **9**, 150-160.
- Licklider, J. C. R. (1948). "The influence of interaural phase relations upon the masking of speech by white noise", *J. Acoust. Soc. Am.* **20**, 150-159.
- Licklider, J. C. R. (1951). "A duplex theory of pitch perception," *Experientia* **7**, 128-134.
- Lin, J.-Y., Hartmann, W. M. (1998). "The pitch of a mistuned harmonic: Evidence for a template model", *J. Acoust. Soc. Am.* **103**, 2608-2617.
- Marin, C., Mc Adams, S. (1991). "Segregation of concurrent sounds. II: Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width", *J. Acoust. Soc. Am.* **89**, 341-351.

- Mc Adams, S. (1989). "Segregation of concurrent sounds. I: Effects of frequency modulation coherence", *J. Acoust. Soc. Am.* **86**, 2148-2159.
- Meddis, R., Hewitt, M.J. (1991). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. II: phase sensitivity", *J. Acoust. Soc. Am.* **89**, 2883-2894.
- Meddis, R., Hewitt, M.J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies", *J. Acoust. Soc. Am.* **91**, 233-245.
- Miller, G. A., and Heise, G. A. (1950). "The trill threshold", *J. Acoust. Soc. Am.* **22**, 637-638.
- Moore, B. C. J. and Glasberg, B. R. (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", *J. Acoust. Soc. Am.* **74**, 750-753.
- Moore, B. C. J., Glasberg, B. R., and Peters, R. W. (1985). "Relative dominance of individual partials in determining the pitch of complex tones", *J. Acoust. Soc. Am.* **77**, 1853-1860.
- Moore, B. C. J., Glasberg, B. R., and Peters, R. W. (1986). "Thresholds for hearing mistuned partials as separate tones in harmonic complexes", *J. Acoust. Soc. Am.* **80**, 479-483.
- O'Shaunessy, D., and Allen, J. (1983). "Linguistic modality effects on fundamental frequency in speech", *J. Acoust. Soc. Am.* **74**, 1155-1171.
- Parsons, T. W. (1976). "Separation of speech from interfering speech by means of harmonic selection", *J. Acoust. Soc. Am.* **60**, 911-918.
- Peterson, P. M. (1986). "Simulating the response of multiple to a single source in a reverberant room", *J. Acoust. Soc. Am.* **80**, 1527-1529.

- Peng, S.C., Lu, N. and Chatterjee, M. (2009). "Effects of cooperating and conflicting cues on speech intonation recognition by cochlear implant users and normal hearing listeners". *Audiol. & Neurotol.* **14**, 327-337.
- Plomp, R., and Mimpen, A. M. (1979). "Speech-reception threshold for sentences as a function of age and noise level", *J. Acoust. Soc. Am.* **66**, 1333-1342.
- Rasch, R. A. (1978). "The perception of simultaneous notes such as in polyphonic music," *Acustica.* **40**, 21-33.
- Roberts, B., and Bregman, A. S. (1991). "Effects of the pattern of spectral spacing on the perceptual fusion of harmonics," *J. Acoust. Soc. Am.* **90**, 3050-3060.
- Roberts, B., and Bailey, P. J. (1996). "Spectral regularity as a factor distinct from harmonic relations in auditory grouping," *J. Exp. Psychol.: Human Percept. Perform.* **22**, 604-614.
- Roberts, B. (1998). "Effects of spectral pattern on the perceptual salience of partials in harmonic and frequency-shifted complex tones: A performance measure", *J. Acoust. Soc. Am.* **103**, 3588-3596.
- Roberts, B., and Brunstrom, J. M. (1998). "Perceptual segregation and pitch shifts of mistuned components in harmonic complexes and in regular inharmonic complexes", *J. Acoust. Soc. Am.* **104**, 2326-2338.
- Roberts, B., and Brunstrom, J. M. (2001). "Perceptual fusion and fragmentation of complex tones made inharmonic by applying different degrees of frequency shift and spectral stretch," *J. Acoust. Soc. Am.* **110**, 2479-2490.
- Roberts, B., Glasberg, B. R., and Moore, B. C. J. (2002). "Primitive stream segregation of tone sequences without differences in fundamental frequency or passband", *J. Acoust. Soc. Am.* **112**, 2074-2085.

- Roberts, B., and Brunstrom, J. M. (2003). "Spectral pattern, harmonic relations, and the perceptual grouping of low-numbered components", *J. Acoust. Soc. Am.* **114**, 2118-2134.
- Roberts, B. (2005). "Spectral pattern, grouping, and the pitch of complex tones and their components", *Acta Acustica United Acustica* **91**, 945-957.
- Roberts, B., and Holmes, S. D. (2006). "Grouping and the pitch of a mistuned fundamental component: Effects of applying simultaneous multiple mistunings to the other harmonics", *Hearing Research* **222**, 79-88.
- Roberts, B., Holmes, S. D., Darwin, C. J., and Brown, G. J. (2010). "Perception of concurrent sentences with harmonic or frequency-shifted voiced excitation: Performance of human listeners and of computational models based on autocorrelation", in *Proceedings of the XVth International Symposium on Hearing*, Salamanca, Spain, June 2009, edited by E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis (Springer-Verlag, Berlin), in press.
- Rossi-Katz, J. A. and Arehart, K. H. (2005). "Effects of cochlear hearing loss on perceptual grouping cues in competing-vowel perception", *J. Acoust. Soc. Am.* **118**, 2588-2598.
- Scheffers, M. T. M. (1983). "Sifting vowels: Auditory pitch analysis and sound segregation". PhD. Thesis, Rijksuniversiteit Groningen, The Netherlands.
- Shackleton, T., M. and Carlyon, R. P. (1994). "The role of resolved and unresolved harmonics in pitch perception and frequency modulation", *J. Acoust. Soc. Am.* **95**, 3529-3540.
- Shackleton, T. M., Meddis, R., Hewitt, M. (1994). "The role of binaural and fundamental frequency difference cues in the identification of concurrently presented vowels", *Quart. J. Exp. Psychol. A* **47A**, 545-563.

- Shamma, S. and Klein, D. (2000). "The case of the missing pitch templates: How harmonic templates emerge in the early auditory system", *J. Acoust. Soc. Am.* **107**, 2631-2644.
- Shinn-Cunningham, B., Ihlefeld, A., Satyavarta, and Larson, E. (2005). "Bottom-up and top-down influences on spatial unmasking", *Acta Acustica united with Acustica* **91**, 967-979.
- Steeneken, H. and Houtgast, T. (1980). "A physical method for measuring speech-transmission quality", *J. Acoust. Soc. Am.* **67**, 318-326.
- Stevens, K. N., Kasowski, S. and Fant, C. G. M. (1953). "An electrical analog of the vocal tract", *J. Acoust. Soc. Am.* **25**, 734-742.
- Summerfield, Q. and Assmann, P. F. (1991). "Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony", *J. Acoust. Soc. Am.* **89**, 1364-1377.
- Summerfield, Q. and Culling, J. (1992). "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency", *J. Acoust. Soc. Am.* **92**, 2317 (A).
- Watkins, A., J. (1991). "Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion", *J. Acoust. Soc. Am.* **90**, 2942-2955.
- Watkins, A. J. (2005). "Perceptual compensation for effects of reverberation in speech identification", *J. Acoust. Soc. Am.* **118**, 249-262.
- Wever, E. G., and Bray, C. W. (1930). "The nature of acoustic response: the relation between sound frequency and frequency of impulses in the auditory nerve", *Journal of experimental psychology* **13**, 373-387.

Appendix A

All source sentences (originally recorded by the MIT talker) were at the same RMS level. The F0 manipulations were performed by the Praat PSOLA speech analysis and resynthesis package which introduced a small variation in RMS level. Table A.1 shows that the main RMS level difference is introduced by setting different values of mean F0, whereas the width of F0 modulation has no mean effect. The effect of F0 follows a monotonic curve: the higher the mean F0, the higher the RMS level. Adding F0-manipulated interfering sentences to each other to create the 2-voice interferer caused the resulting RMS level to be increased by an average greater than 3 dB (Table A.2). Sentences with the same F0 manipulations apparently yielded a phase spectrum more similar to each other than the original sentences. As a result when added together, their interaction was not completely random. Thus after the manipulations of F0 and the formation of the two-voice interferers, an initial RMS equalization was performed by multiplying the signal amplitude by a correcting factor.

F0 Re: 110Hz (semitones)	Monotonized F0	Modulated F0 ($\pm 1s.$)	Modulated F0 ($\pm 2s.$)
0	68.48	68.48	68.47
1	68.67	68.67	68.65
2	68.86	68.85	68.84
3	69.03	69.03	69.01
4	69.19	69.18	69.16
5	69.31	69.31	69.31
6	69.44	69.45	69.44
7	69.57	69.57	69.56
8	69.69	69.69	69.68
9	69.79	69.79	69.78
10	69.88	69.87	69.87
11	69.96	69.96	69.96
12	70.05	70.05	70.04

TABLE A.1 RMS levels in dB, averaged over 80 original sentences whose mean RMS=69.3 dB, after different manipulations by the software Praat.

original sentences	69.30 (0.06)
adding 2 different sentences	72.24 (0.09)
adding 2 F0-monotonized sentences	73.28 (0.25)
adding 2 F0-modulated sentences	73.21 (0.28)
adding 2 identical sentences	75.32 (0.06)

TABLE A.2 RMS levels in dB with standard deviation changing with adding different sentences.

Appendix B

Appendix A described the changes in RMS levels due to the Praat manipulations of F0 and the addition of two interferers. A further change in RMS level was produced by the acoustic response of the reverberant room which amplified some frequencies and not others, producing a spectral colouration plotted on the first graph of Figure B1. The middle and high frequency-domain of the spectrum was affected by this spectral colouration. Since this is the frequency range most significant for speech intelligibility, it was therefore necessary to equalize the spectra. Since the convolution with the anechoic room impulse response did not produce any RMS level difference, we used a filter that compensated for the colouration produced by the reverberant room. The colouration being slightly different for left and right ears, we used two compensating filters, one for each ear. The excitation patterns of both the anechoic and reverberant sentences were used to create this compensating filter. We used the Matlab-function `fir2` to design a finite impulse response (FIR) filter with 5000 coefficients whose frequency response is the difference between the excitation patterns of the reverberant sentence and that of the respective anechoic sentence. We then applied this filter to the reverberant sentence and compensated the delay induced by the convolution with the filter. This process was applied independently to the left and right ear before recombining them with each other to get back binaural stimuli.

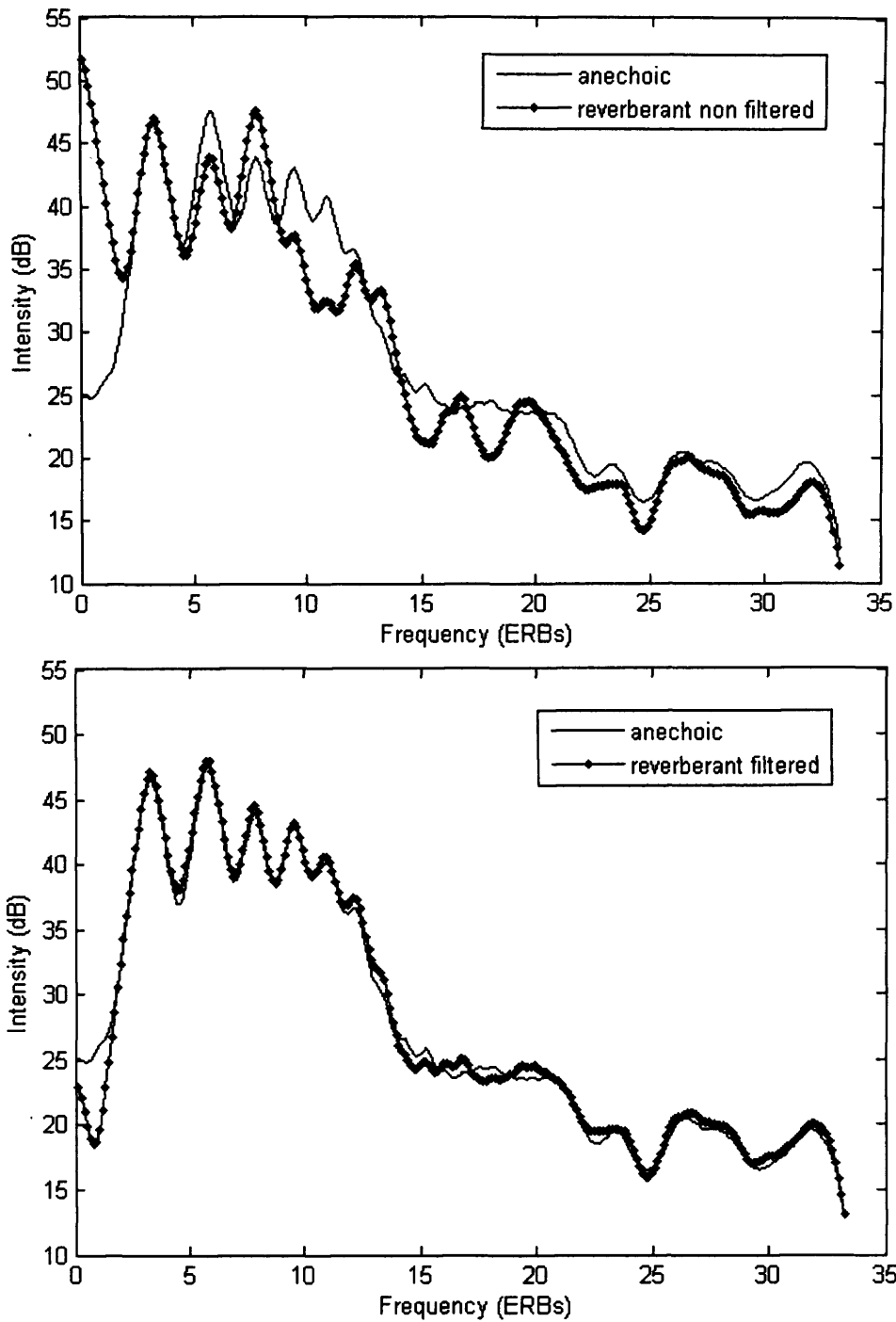


FIG. B1 (up) Left-ear excitation patterns of an interfering sentence after convolution with the anechoic or reverberant room impulse responses before filtering. (down) Left-ear excitation patterns of the same two sentences, once the reverberant one has been filtered to have the same excitation pattern as the anechoic one.