

SPECTRAL FUSION, SPECTRAL PARSING  
AND THE  
FORMATION OF AUDITORY IMAGES

A DISSERTATION  
SUBMITTED TO THE PROGRAM IN  
HEARING AND SPEECH SCIENCES  
AND THE COMMITTEE ON GRADUATE STUDIES  
OF STANFORD UNIVERSITY  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY

By  
Stephen McAdams

May 1984



Center for Computer Research in Music and Acoustics

May 1984

Department of Music  
Report No. STAN-M-22

**SPECTRAL FUSION, SPECTRAL PARSING  
AND THE FORMATION OF AUDITORY IMAGES**

**by**

**Stephen McAdams**

**Research sponsored by  
System Development Foundation**

**CCRMA  
DEPARTMENT OF MUSIC  
Stanford University  
Stanford, California 94305**





Department of Music  
Report No. STAN-M-22

## SPECTRAL FUSION, SPECTRAL PARSING AND THE FORMATION OF AUDITORY IMAGES

by

Stephen McAdams

An important perceptual aspect of the formation of auditory images evoked by acoustic phenomena is the distinguishing of different sound sources. In order to be able to form images of sound sources, the auditory system must be able to perceptually fuse the concurrent elements that come from the same source and separate the elements that come from different sources. The auditory image is a psychological representation of a sound entity exhibiting an internal coherence in its acoustic behavior. The problem is: 1) to search for a definition of what constitutes auditory coherence from a psychological standpoint, 2) to understand its relation to the behavioral coherence of the physical world, and 3) to elaborate the knowledge structures and psychological processes underlying the perceptual organization of complex acoustic situations.

The acoustic cues that contribute to the formation and distinction of multiple, simultaneous source images which are investigated include the harmonicity of the frequency content, the coherence of low-frequency frequency modulation, and the stability and/or recognizability of spectral form when coupled with frequency modulation. Listeners were asked to compare sounds within which these acoustic dimensions were varied and to report differences in the perceived number of multiplicity of the sources, or to identify particular sources embedded in a complex acoustic background. The experimental results show that: 1) Frequency modulation coherence can be defined as a modulation maintaining constant ratios among the component frequencies. 2) The auditory system is acutely sensitive to incoherence of random frequency modulation on adjacent frequency components to harmonic sources and can detect incoherence at modulation widths of less than 0.05% for partials within a critical bandwidth. The incoherence detection threshold is 5 times greater for inharmonic sounds, suggesting that the acuity lies in auditory temporal mechanisms. 3) The perception of the unity of complex spectral structures is sensitive to the coupling of frequency modulation with an amplitude modulation on each component that defines a relatively constant spectral envelope. 4) This tracing of the spectral envelope by the frequency components is a strong cue for the separation of familiar spectral forms, like vowels, from a multi-source complex. 5) There are indications that source image formation processes are independent of the derivation of some source qualities, such as identity of vowels, whereas other qualities, such as pitch and timbre (tone color), are directly related to how the acoustic information is parsed into sources. A proposition of the necessary elements for a theory of auditory image formation is discussed in terms of the experimental results.

*This thesis was submitted to the Department of Hearing and Speech Sciences and the Committee on Graduate Studies of Stanford University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.*

*This research was supported by the System Development Foundation under Grant SDF #345. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of Stanford University, any agency of the U. S. Government, or of sponsoring foundations.*

**© Copyright 1984  
by  
Stephen Edward McAdams**

to  
Pa &  
Gloria

## PREFACE

This dissertation is the culmination of an extended detour from music through the domains of psychology and neuroscience that I began some 10 years ago. Experiencing a certain frustration with what was being touted as theory of music, and in particular finding a great deal of twentieth century music drifting far from music as *heard*, my personal interests led me on a search for some less arbitrary approach to an understanding of the many aspects of music. Most importantly, this odyssey was a search for a theory of the listener faced with the artifice of a constructed sound field. What might possibly be the nature of this relation? How does music influence a listener? How does a listener influence a music? Ultimately the search is for an understanding of the relation between sound and life itself. So I set out in the directions of mind and nervous system to approach this problem. It needs be said that I did not see this path by myself. My first true mentor was a man of great insight and vision, who translated my musical desires into the questions posed above, or rather who showed me what I was really asking. Accordingly, my first gratitude must be extended to Carlton Sloan.

There followed then a series of mentors in the academic environments of McGill, Northwestern and Stanford Universities. Al Bregman, of McGill University, should be credited for teaching me how to ask questions and therefore to think clearly. He also gave me the freedom to pursue the more musical aspects of research on auditory organization. His influence and example continues to be stimulating. Fred Wightman kindly took me under his wing in the Psychoacoustics Research Lab at Northwestern, gave me a year to dive unhindered into the profundities of "hard-core" psychoacoustics, and thus helped with the fine-tuning of my comprehension of experimental methodology. Peter Dallos, by way of his courses on auditory physiology, opened up the marvelous world of neurophysiology for me, instilling an enthusiasm that has not diminished, and more importantly showed me what a truly great pedagogue is; I have

yet to witness a greater teacher. When it became obvious that my musical interests could not really be nourished at Northwestern, these two gentlemen were very kind in letting me move on to finish my studies at Stanford, with their blessings.

At Stanford, where I came to roost for a time and to discover the problems exposed in this dissertation, my beloved advisor Earl Schubert then allowed me a freedom that is relatively unheard of, as well as demonstrating an unbounded enthusiasm for my searchings and researches. His encyclopedic knowledge of auditory theory and experimentation were one of my most prized resources on the Stanford campus. And, without doubt, beyond the academic exigencies and unfailing support that he took care of, the friendship he offered is still very precious to me. My deepest gratitude and love are extended to Earl. I owe as well much thanks to John Chowning, whose musical vision provided the very seed of this dissertation. In addition, the stimulating and nourishing environment that John established at CCRMA initiated a passion for the possibilities of computer music that has landed me where I am today. I would also like to thank Roger Shepard for many interesting discussions and for pushing me somewhat more in the direction of cognition than I would have been inclined to go on my own at this stage.

Finally, I flew the coup and did the majority of my research at IRCAM, where I was initially invited by that generator of mad whirlwinds of ideas, that are always more than one can assimilate at any given time, none other than David Wessel. David's friendship, mentorship and constant barrage of new and interesting applications of my ideas were an essential force in the directions that this research took, and in the way it is now transforming itself into more musically oriented research problems. Were it not also for David's having invited Bill Hartmann from Michigan State University to spend a year at IRCAM, this dissertation might not have seen its end even as late as it did. Bill had the insight to notice that I had too many ideas to get any done in a reasonable time and took upon his shoulders the office of advisor *ex officio* gently goading me into crystallizing my project into something realizable within a fixed amount of time. I can't offer enough thanks for the friendship and the many long, long hours of discussion and listening that Bill graciously offered with his typically thorough and uncompromising manner. Bill was particularly helpful in the clarification of my ideas about within- and cross-channel mechanism in Chapter 3.

I would like to thank as well the groups of people who have contributed enormously in terms of moral, intellectual and technical support. First is the group of people at CCRMA, whom I would like to thank one and all for getting me going in computer music. In particular, special thanks go to Mike McNabb, Kip Sheeline, Betsy Cohen, Eduardo Castro-Sierra, Andy Schloss, Patte Wood, Bill Schottstaedt, David Jaffe, Chris Chafe and Andy Moorer. At IRCAM, thanks are also due for linguistic support (and much patience on the part of my dear colleagues there) while I tried to learn French. There are many ways this unusual collection of people has contributed to my research and mental well-being during the process. Among these I give special thanks to Xavier Rodet, Jean-Baptiste Barrière, Tod Machover, Yves Potard, Marco Stroppa, Philippe Manoury, Thierry Lancino, Florence Quilliard, Andrew Gerszo, Marc Battier, Adrian Freed, Michele Dell-Prane and Bennett Smith. David and Tod were very helpful in making sure that I had financial support during the period of doing research. Bill Hartmann constructed all of the response boxes and Bennett Smith wrote all of the experiment-running subroutines. In general, the interdisciplinary environment of IRCAM served as both an idea resource and critical sounding board for my own theories and ideas. Many good ideas and useful programs were generated by Clarence Barlow during his stay at IRCAM, for which I am thankful. Additional thanks need be given to Laurent Demany and Dominique Lépine of Université René Descartes (Paris V). Laurent, in particular, did a very devastating critique of Chapter 5 which made me change my thinking about the relations between source grouping processes and image qualities. Over the last few years three Dutch colleagues have also given me much encouragement and good ideas. I would like to thank Leon van Noorden, Rudolf Rasch and Adrian Houtsma.

While doing graduate study and research I have been supported by the National Science Foundation (as a Graduate Fellow, 1978-1981) and received a research grant from the Minister of Foreign Affairs of the French Government (CROUS, 1982-1983). I would like to thank both of these organisms for their support.

There have also been some very special friends in my graduate life whose gentle spirits, love and support have often been most needed. I would like to express my love and gratitude to Christopher Gaynor, Shirley Shakes, Clair Lüdenbach and Pilu Lydlow. And most special of all, I wish to express my deepest love to my companion Wendy Lindbergh, who sacrificed a great deal of things for me during this last year of writing. Her continuing emotional, intellectual and spiritual support held my head

above the dark waters many times.

There have been many times in my educational career when my dear parents have helped me both morally and financially. I have dedicated this work to them, to which I add an avowal of much love and respect. My father gave up this same opportunity for himself when I was a boy, in order to be around more as a father, and I can never express enough gratitude or love for that sacrifice.

The pilot studies and intermediate demonstrations that preceded this dissertation have been reported in McAdams (1980, 1981, 1982b) and McAdams & Wessel (1981). This research itself has been reported in McAdams (1982a, 1983a, 1983b, 1984). I have taken the liberty of using text (with modifications) from my own work to a certain extent in the Prologue, Epilogue and Chapters 1 and 6. Portions of McAdams (1982b) appear in sections 1.3, 1.4, 1.7.2, 1.7.4 and the Epilogue and are included with the kind permission of Plenum Publishing Corporation, New York. Portions of McAdams (1984) appear in the Prologue, sections 1.3, 1.7.5, 6.1, 6.3, and the Epilogue and are included with the kind permission of North Holland Publishing Company, Amsterdam.

S.M.

Paris

7 May 1984

## CONTENTS

PREFACE.....	vii
TABLE OF CONTENTS.....	xi
LIST OF TABLES .....	xviii
LIST OF FIGURES .....	xxiii
 PROLOGUE .....	 1
 CHAPTER 1: Research Problems on Source Image Formation.....	 4
1.1 Introduction.....	4
1.2 Paradigmatic Framework.....	6
1.2.1 Direct Perception.....	6
1.2.2 Feature Extraction.....	7
1.2.3 Hypothesis-testing .....	7
1.2.4 Gestalts .....	8
1.2.5 Neisser's Synthesis: The Perceptual Cycle.....	8
1.3 The Auditory Image Metaphor.....	11
1.4 The Forming and Distinguishing of Auditory Images .....	17
1.5 Perceptual Fusion.....	21
1.6 Derivation of Image Qualities.....	27
1.6.1 Pitch Perception .....	28
1.6.2 Timbre and Vowel Quality Perception .....	29
1.7 Cues for Simultaneous Grouping .....	40
1.7.1 (Apparent) Spatial Location.....	41
1.7.2 Harmonicity of Spectral Content .....	41
1.7.3 Pitch Separation .....	43
1.7.3.1 Degree of Spectral Overlap .....	44
1.7.3.2 Harmonic Coincidence.....	44
1.7.4 Frequency Modulation Coherence.....	45
1.7.5 Amplitude Modulation Coherence .....	47
1.7.5.1 Onset Synchrony.....	48
1.7.5.2 Amplitude Fluctuations .....	50



1.7.6 Resonance Structure Stability and Recognition of Spectral Form.....	52
1.8 Problems to be Addressed.....	55
CHAPTER 2: Harmonicity-preserving Frequency Modulation and Spectral Fusion.....	
2.1 Introduction.....	61
2.2 EXPERIMENT 1: Effects of sub-audio frequency modulation maintaining constant frequency differences and constant frequency ratios on perceived source image multiplicity.....	62
2.2.1 Stimuli.....	62
2.2.2 Method.....	65
2.2.3 Results.....	67
2.2.4 Discussion.....	70
2.3 EXPERIMENT 2: Perceptual scaling of perceived fusion or multiplicity for harmonic tones with different spectral envelope shapes, frequency modulation waveforms, modulation widths and modulation type. ....	75
2.3.1 Stimuli.....	75
2.3.2 Method.....	75
2.3.3 Results.....	77
2.3.4 Discussion.....	79
2.4 EXPERIMENTS 3 - 5: Corollary Studies to Experiment 1 .....	81
2.4.1 Stimuli and Subjects .....	82
2.4.2 Method and Results .....	82
2.4.2.1 Experiment 3: Source multiplicity judgments on CR/CD tone pairs with very small differences in $F_0$ modulation width. ....	82
2.4.2.2 Experiment 4: Modulation width judgments on the vibrato stimulus pairs from Experiment 1.....	85
2.4.2.3 Experiment 5: Modulation width judgments on the vibrato stimulus pairs from Experiment 3.....	87

2.4.3 Discussion.....	88
2.4.3.1 Comparison of the experiments .....	88
2.4.3.2 Subjects impressions of the stimuli and judgments.....	90
2.5 General Discussion and Summary .....	92
CHAPTER 3: Within-channel and Cross-channel Contributions to Multiple Source Perception.....	94
3.1 Introduction.....	94
3.1.1 Within channel information .....	94
3.1.2 Cross-channel information .....	98
3.2 <b>EXPERIMENT 6:</b> Effects of the frequency modulation incoherence, harmonicity and intensity on multiple source perception.....	100
3.2.1 Stimuli.....	100
3.2.2 Method.....	106
3.2.3 Results.....	107
3.2.3.1 Effects of rms deviation of modulation .....	115
3.2.3.2 Effect of the number of the partial being modulated incoherently .....	117
3.2.3.3 Effect of intensity .....	120
3.2.3.4 Effect of harmonicity (phase synchrony of adjacent partials).....	120
3.2.4 Discussion.....	
3.2.4.1 Nature of the Experimental Task and Subjective Impressions of the Stimuli .....	
3.2.4.2 Within- and Cross-channel Mechanisms of Incoherence Detection.....	
3.3 Summary .....	
CHAPTER 4: Fixed Spectral Structure and Spectral Fusion .....	138
4.1 Introduction.....	138

4.2	<b>EXPERIMENT 7: Effects of fixed spectral envelope on perceived source multiplicity.</b>	138
4.2.1	Stimuli	138
4.2.2	Method	139
4.2.3	Results	141
4.2.3.1	Effects of spectral envelope	141
4.2.3.2	Effects of modulation waveform	144
4.2.4	Discussion	145
4.3	Summary	148
CHAPTER 5: Perceptual Identity and the Distinction of Source Images		150
5.1	Introduction	150
5.2	<b>EXPERIMENT 8: Effects of sub-audio frequency modulation and fixed resonance structure on the perceived prominence of vowel sources embedded in a complex (multi-source) spectrum.</b>	152
5.2.1	Pre-test	153
5.2.1.1	Stimuli	154
5.2.1.2	Method and Results	158
5.2.2	Stimuli	159
5.2.3	Method	161
5.2.4	Results	163
5.2.4.1	Effect of modulation state of non-target vowels on prominence ratings of the target vowel.	164
5.2.4.2	Effect of permutation of non-target vowels on prominence ratings of the target vowel.	167
5.2.4.3	Regrouping of the data	169
5.2.4.4	Effect of the modulation state of the target vowel.	170
5.2.4.5	Effects of pitch position of the target vowel.	170
5.2.5	Discussion	172
5.2.5.1	Pitch Position and Modulation State of Non-target Vowels	172

5.2.5.2 Modulation of the Target Vowel and Coherence with Non- targets .....	173
5.2.5.3 Spectral Form Stability .....	174
5.2.5.4 Pitch Position of the Target Vowel .....	175
5.2.5.5 Perceived Pitch of the Vowel Sources .....	176
CHAPTER 6: The Auditory Image: A Metaphor for Musical and Psychological Research on Auditory Organization .....	180
6.1 Sequential Organization .....	180
6.1.1 Frequency Separation .....	182
6.1.2 Amplitude Differences .....	185
6.1.3 Spectral Form and Content .....	186
6.1.4 Spectral Continuity and Sequential Organization .....	187
6.2 Simultaneous Organization .....	189
6.2.1 Frequency Modulation Coherence and Harmonicity .....	189
6.2.2 Spectral Form .....	193
6.2.3 Coherent Behavior of Sound Objects and Simultaneous Organiza- tion .....	197
6.3 Interactions Between Sequential and Simultaneous Organization .....	201
6.4 Toward a Theory of Auditory Image Formation and Source Percep- tion .....	203
6.4.1 Behavioral Coherence of Sound Objects .....	204
6.4.2 The Processes of Auditory Organization .....	205
6.4.3 Derivation of Image Qualities .....	206
6.4.4 Attentional Processes .....	207
EPILOGUE .....	209

APPENDIX A: Synthesis Procedures and Sound Presentation System .....	214
A.1 Additive Synthesis Procedures .....	214
A.1.1 Amplitude Control .....	216
A.1.1.1 Global Amplitude Function .....	216
A.1.1.2 Component Amplitude Determined by Spectral Form .....	217
A.1.2 Frequency Control .....	218
A.2 Time-domain Formant-wave-function Synthesis Procedure .....	220
A.3 Experimental Equipment and Sound Presentation .....	222
 APPENDIX B: Analysis and Synthesis of Jitter Functions .....	223
B.1 Introduction .....	223
B.2 Jitter Analysis .....	224
B.2.1 Preparation of Sounds .....	224
B.2.2 Characterization of the Jitter Function .....	226
B.3 Synthesis of Jitter Functions .....	232
 APPENDIX C: (Experiment 9) Loudness and Modulation Width Matching for Experiment 1 Stimuli .....	238
C.1 Modulation Width Matching .....	238
C.1.1 Stimuli and Method .....	238
C.1.2 Results and Discussion .....	239
C.2 Loudness Matching .....	240
C.2.1 Stimuli and Method .....	240
C.2.2 Results and Discussion .....	241
 APPENDIX D: (Experiment 10) Frequency modulation detection of periodic and aperiodic waveforms imposed on complex harmonic tones .....	242

D.1 Stimuli.....	242
D.2 Method.....	243
D.3 Results.....	245
APPENDIX E: Data Tables .....	246
APPENDIX F: Pilot Studies on Spectral Fusion and Spectral Parsing .....	262
F.1 Informal Preliminary Studies .....	262
F.1.1 Fusion Investigations .....	262
F.1.2 Parsing Investigations.....	264
F.2 Formal Pilot Studies.....	266
F.2.1 Experiment A.....	267
F.2.1.1 Stimuli.....	267
F.2.1.2 Method.....	269
F.2.1.3 Results and Discussion.....	270
F.2.2 Experiment B.....	276
F.2.2.1 Stimuli.....	276
F.2.2.2 Method.....	277
F.2.2.3 Results and Discussion.....	278
F.2.3 Experiment C.....	281
F.2.3.1 Stimuli and Method .....	283
F.2.3.2 Results and Discussion.....	285
APPENDIX G: Description of Taped Examples .....	286
REFERENCES.....	294

## LIST OF TABLES

TABLE 2.1. Rms deviation of frequency modulation used in Experiments 1 and 2. ....	63
TABLE 2.2. Data summary for Experiment 1. Each cell value is the mean across Ss of the proportion of choices of <i>CD</i> tones. ....	68
TABLE 2.3. Comparison between rms and peak deviation values for vibrato and jitter waveforms. ....	72
TABLE 2.4. Rms deviation of the fundamental frequencies of <i>CR</i> and <i>CD</i> tones. ....	74
TABLE 2.5. Experiment 2 data summary. One-dimensional scaling solution for judgments on the relative degree of fusion for vibrato and jitter stimuli. ....	79
TABLE 2.6. Rms deviation (cents) for <i>CR</i> and <i>CD</i> tones, matched as closely as possible with existing stimuli. ....	83
TABLE 2.7. Modulation widths on $F_0$ in Experiments 1,3,4,5. ....	84

TABLE 2.6. Relative difference of modulation widths ( $\Delta f_{rms}$ in Hz) across CR/ CD pairs in Experiments 1,3,4,5. ....	85
TABLE 3.1. Component frequencies, Bark measures and distance between components in Barks for harmonic and inharmonic stimuli. ....	101
TABLE 3.2. The 5 rms deviations of modulation used for harmonic and inhar- monic stimuli. ....	105
TABLE 3.3. Experiment 6 data summary. Source multiplicity thresholds for individual subjects measured from 71% points on cubic spline curves fitted to data points. ....	113
TABLE 3.4. Group SMTs across subjects expressed as cents <sub>rms</sub> , $\Delta f_{rms} / \bar{f}$ and $\Delta P_{rms}$ (change in period of incoherent component that is just noticeable as yielding multiple sources. ....	118
TABLE 3.5. Loudnesses (in phons) of individual partials for harmonic tones. .....	123
TABLE 3.6. Summary of subjects' descriptions of perceived effects of intro- ducing an incoherently modulating partial into a complex tone, for the 3 main conditions (H75, H50, I75). ....	124
TABLE 3.7. Transformation of Shower & Biddulph's (1931) data for com- parison with those from Experiments 1, 6 & 10. ....	132



TABLE 4.1. Experiment 7 data summary. Means and unbiased standard deviations (in parentheses) across subjects for each spectral envelope, modulation waveform and rms deviation. ....	143
TABLE 4.2. Overall means and pooled standard deviations across subjects and modulation waveforms for Experiment 7. ....	145
TABLE 5.1. Parameters for vowel synthesis with the program CHANT. ....	153
TABLE 5.2. Rms amplitudes (dB re: the most intense stimulus) of vowel stimuli at 3 pitches both with and without modulation. ....	158
TABLE 5.3. Modulation state of the vowels under different figure-ground combinations.....	160
TABLE 5.4. Statistically significant comparisons of means for permutations within a given pitch position of the target vowel. ....	167
TABLE 5.5. Modulation forms for each vowel under different figure-ground combinations.....	168
TABLE 5.6. Means and standard deviations for data pooled according to target vowel's pitch position and modulation state for each target vowel. ....	169

TABLE 5.7. Values of the $t$ -statistic for comparisons within pitch position for means collected into pitch and modulation state of the target vowel. ....	171
TABLE 5.8. Differences between means for modulated and unmodulated target vowels at each pitch. ....	171
TABLE C.1. Experiment 9 data summary for modulation width matches.....	239
TABLE C.2. Experiment 9 data summary for loudness matches. ....	241
TABLE D.1. Experiment 10 data summary. Cell values represent the 71% MDT for a complex harmonic carrier determined with an adaptive tracking procedure. ....	244
TABLE E.1. Experiment 1 data summary. Each value represents the proportion of times in 50 2IFC judgments the subject chose the <i>CD</i> tone as yielding more sources than the <i>CR</i> tone. ....	247
TABLE E.2. Experiment 3 data summary. Each cell value represents the proportion of 25 2IFC judgments where the constant frequency difference tone was chosen as yielding more sources than the constant frequency ratio tone. ....	249
TABLE E.3. Experiment 4 data summary. Each cell value represents the proportion of 30 2IFC judgments where the constant frequency difference tone was chosen as having a larger modulation width on $F_0$ than the constant frequency ratio tone. ....	250

TABLE E.4. Experiment 5 data summary. Each cell value represents the proportion of 25 2IFC judgments where the constant frequency difference tone was chosen as having a larger modulation on $F_0$ than the constant frequency ratio tone. ....	251
TABLE E.5. Experiment 6 data summary. Individual data are the proportion of times in 30 2IFC judgments that the tone with incoherent modulation on one partial was chosen as having more sources. ....	252
TABLE E.6. Experiment 7 data summary. Each value represents the proportion of times in 50 2IFC judgments the subject chose the <i>CCA</i> tone as yielding more sources than the <i>CSE</i> tone. ....	258
TABLE E.7. Experiment 8 data summary. For each stimulus condition a separate vowel prominence rating (0 - 100) was made for each of the target vowels /a/, /o/ and /i/. The data represent the means ( $\bar{x}$ ) and unbiased standard deviations ( $\sigma$ ) of normalized data across 10 Ss. ....	260

## LIST OF FIGURES

Figure 1.1. Schematic representation of the stimuli used by Bregman & Pinker (1978) and two common perceptual results. ....	23
Figure 1.2. Schematic representation of the stimuli used by Cutting (1976) and the most common perceptual results. ....	25
Figure 1.3. The spectral form created by a 3-formant resonance structure is represented by the dotted line. As the harmonics are modulated in frequency, their respective amplitudes fluctuate as a function of the spectral envelope. ....	32
Figure 1.4. The vowel /a/ is plotted with a high fundamental frequency where there are few harmonics present. Without modulation (a), the inferred spectral form (dashed line) would be very different from the actual spectral form (dotted line) With modulation (b), the spectral slopes give a much clearer indication of the spectral form. ....	35
Figure 1.5. Two spectral envelopes used by Rodet (1983). The second harmonic falls at the intersection of the two possible $F_2$ 's. ....	38
Figure 1.6. Enlargement of the region of the second harmonic's spectral slopes from Figure 1.5 (from Rodet, 1983).....	39

Figure 2.1. Exaggerated spectrographic diagram of <i>CR</i> and <i>CD</i> modulations plotted on a log frequency scale for the first 8 harmonics.....	64
Figure 2.2. Experiment 1 data summary. Each graph shows the proportion of times the constant-difference tone was chosen as having more sources than the constant-ratio tone as a function of the rms deviation (expressed in cents). .....	66
Figure 2.3. Experiment 1 data summary. The proportion of times the <i>CD</i> tone was chosen as yielding more sources is plotted as a function of the rms deviation of modulation. Data points are averaged over <i>Ss</i> , spectral envelope and modulation waveform. ....	70
Figure 2.4. Experiment 1 data summary. Means within modulation deviation width and modulation waveform are plotted as functions of (a) rms deviation and (b) peak deviation. Data are averaged across all <i>Ss</i> and spectral envelope conditions. ....	73
Figure 2.5. Experiment 2 data summary. One-dimensional scaling solutions for judgments on the relative degree of fusion for vibrato and jitter stimuli plotted as a function of rms deviation of modulation. The scale is interpreted as the degree of perceived fusion or perceived source multiplicity. ....	78
Figure 2.6. Experiment 3 data summary. The proportion of times the <i>CD</i> tone was chosen as yielding more sources is plotted as a function of rms deviation of modulation (the modulation width on $F_0$ in the <i>CR</i> tone for each stimulus pair). ....	84

Figure 2.7. Experiment 4 data summary. The proportion of times the <i>CD</i> tone was chosen as having a larger modulation width on the $F_0$ than that on the <i>CR</i> tone is plotted as a function of the modulation width of the <i>CR</i> $F_0$ . .....	86
Figure 2.8. Experiment 5 data summary. The proportion of times the <i>CD</i> tone was chosen as having a larger modulation width on the $F_0$ is plotted as a function of <i>CR</i> modulation width. ....	87
Figure 2.9. Data from Experiments 4 and 5 (modulation width judgments) plotted as functions of the difference in modulation width ( $CD - CR$ ) of the $F_0$ 's on the two tones in a trial. ....	89
Figure 2.10. Individual data for each subject for Experiments 1 (S1 and S3 only), 3, 4 and 5 are plotted. ....	91
Figure 3.1. Plotted here is the distance in Barks to the next nearest partial of a 16-component complex tone. Both harmonic and inharmonic tones are plotted for comparison. ....	102
Figure 3.2. Plotted here are 32 msec segments (approximately 7 cycles of $f_1$ ) from the waveforms of the unmodulated harmonic and inharmonic tones used in Experiment 6. ....	104
Figure 3.3. Experiment 6 data summary for Subject 1. The proportion of incoherent tone choices is plotted as a function of rms deviation of modulation (in cents). ....	108

Figure 2.7. Experiment 4 data summary. The proportion of times the <i>CD</i> tone was chosen as having a larger modulation width on the $F_0$ than that on the <i>CR</i> tone is plotted as a function of the modulation width of the <i>CR</i> $F_0$ . .....	86
Figure 2.8. Experiment 5 data summary. The proportion of times the <i>CD</i> tone was chosen as having a larger modulation width on the $F_0$ is plotted as a function of <i>CR</i> modulation width. ....	87
Figure 2.9. Data from Experiments 4 and 5 (modulation width judgments) plotted as functions of the difference in modulation width ( $CD - CR$ ) of the $F_0$ 's on the two tones in a trial. ....	89
Figure 2.10. Individual data for each subject for Experiments 1 (S1 and S3 only), 3, 4 and 5 are plotted. ....	91
Figure 3.1. Plotted here is the distance in Barks to the next nearest partial of a 16-component complex tone. Both harmonic and inharmonic tones are plotted for comparison. ....	102
Figure 3.2. Plotted here are 32 msec segments (approximately 7 cycles of $f_1$ ) from the waveforms of the unmodulated harmonic and inharmonic tones used in Experiment 6. ....	104
Figure 3.3. Experiment 6 data summary for Subject 1. The proportion of incoherent tone choices is plotted as a function of rms deviation of modulation (in cents). ....	108

Figure 3.4. Experiment 6 data summary for Subject 2. ....	109
Figure 3.5. Experiment 6 data summary for Subject 3. ....	110
Figure 3.6. Experiment 6 data summary for Subject 4. ....	111
Figure 3.7. Mean data for Experiment 6 averaged across 4 subjects. ....	112
Figure 3.8. Source multiplicity thresholds (SMTs) for subject 1. The SMT (in cents rms deviation) is plotted as a function of the partial number of the frequency component receiving incoherent modulation. ....	114
Figure 3.9. SMTs and complex tone MDTs for subject 2. ....	115
Figure 3.10. SMTs and complex tone MDTs for subject 3. ....	116
Figure 3.11. SMTs and complex tone MDTs for subject 4. ....	117
Figure 3.12. Individual SMTs for 4 subjects plotted as a function of partial number. ....	119



Figure 3.13. Highly exaggerated schematic diagram of the frequency modulation patterns of two coherently modulated components on either side of an incoherently modulated component. The components $f_8$ and $f_{10}$ were modulated with jitter function $J_1$ , and component $f_9$ was modulated with $J_2$ . .....	127
Figure 3.14. Comparison of SMTs with MDTs for sinusoidal and complex carriers. ....	133
Figure 4.1. Schematic illustration of the effect of either tracing ( <i>CSE</i> ) or of modulating ( <i>CCA</i> ) the spectral envelopes of the vowel /a/ or the $-6$ dB/oct slope. ....	140
Figure 4.2. Experiment 7 data summary. The proportion of CCA tones chosen as yielding more sources is plotted as a function of rms deviation of modulation for 2 modulation waveforms and 2 spectral envelopes. ....	142
Figure 4.3. Experiment 7 data summary. Means and pooled standard deviations ( $\pm\sigma_p$ indicated by vertical bar) across subjects and modulation waveform are plotted as a function of rms deviation. ....	144
Figure 4.4. Superimposed transfer functions extracted from 10 successive periods of the vowel /a/ sung by a soprano. ....	147
Figure 5.1. Spectra from the steady state portion of the unmodulated vowel /a/ with $F_0$ at $C_3$ , $F_3$ , $Bb_3$ . ....	155

Figure 5.2. Spectra from the steady state portion of the unmodulated vowel /o/ with $F_0$ at $C_3$ , $F_3$ , $Bb_3$ .....	156
Figure 5.3. Spectra from the steady state portion of the unmodulated vowel /i/ with $F_0$ at $C_3$ , $F_3$ , $Bb_3$ .....	157
Figure 5.4. Experiment 8 data summary for prominence judgments on the vowel /a/. .....	164
Figure 5.5. Experiment 8 data summary for prominence judgments on the vowel /o/. .....	165
Figure 5.6. Experiment 8 data summary for prominence judgments on the vowel /i/. .....	166
Figure 5.7. Summary for data collected under pitch and modulation state of the target vowel. ....	170
Figure 6.1. The tones of parts of two common nursery rhyme melodies are interleaved. In (a) the frequency ranges of the two melodies are similar and the sequence is heard as one, unfamiliar melody. In (b) the frequency ranges of the melodies are non-overlapping and each melody is heard independently. ....	183
Figure 6.2. Illustration of the different percepts resulting from the alternation of two sinusoidal tones of identical frequency and duration. ....	184

Figure 6.3. Effect of differences in spectral composition on sequential organization. In (a) all tones are sinusoidal and the frequency separation between them is adjusted so that a single stream percept may be heard, as indicated by the dotted lines. In (b) the third harmonic is added to one pair and the spectral difference causes two streams to form, each with a different timbre. ....	185
Figure 6.4. When all of these tones are played by the same instrument, ascending pitch triplets are heard (solid lines). But when two instruments with very different spectral forms each play the X's and O's, respectively, descending triplets are heard (dotted lines). ....	186
Figure 6.5. Complex spectrum resulting from several, unmodulated sustaining harmonic sources. ....	194
Figure 6.6. Spectral forms extracted by the auditory system when the individual sources are modulated, thereby increasing the information pertaining to the number and nature of the sources. ....	196
Figure 6.7. The amplitudes of each harmonic are modulated randomly above and below their central value, that is defined by the vowel spectral envelope. The modulation pattern on each harmonic is independent of that on any other harmonic. ....	198
Figure 6.8. The parsing of even and odd harmonics of an oboe sound. Initially, all harmonics are modulated coherently. Then the even harmonics are slowly decorrelated from the odd harmonics until they have an independent modulation pattern. ....	199

Figure A.1. Schematic of sound synthesis procedure for generation of 16-component complex tones. ....	215
Figure A.2. The $-6$ dB/oct spectral envelope. The flat portion below 200 Hz never affects the amplitude of any stimulus component. ....	218
Figure A.3. The vowel /a/ spectral envelope from a singing male voice. ....	219
Figure A.4. Structure of a parallel formant-wave-function (FWF) synthesizer. ....	221
Figure B.1. Determination of the time of the $i^{\text{th}}$ positive-going zero-crossing, $t_z[i]$ , from samples flanking that moment. ....	225
Figure B.2. Jitter data for flute playing an $Eb_4$ at $mf$ . ....	227
Figure B.3. Jitter data for clarinet playing an $Eb_4$ at $mf$ . ....	228
Figure B.4. Jitter data for trombone playing an $Eb_4$ at $mf$ . ....	229
Figure B.5. Jitter data for trombone playing an $Eb_4$ at $ff$ . ....	230
Figure B.6. Schematic diagram of a typical spectrum for the jitter function found on the fundamental frequency of a professional male singer singing a steady note without vibrato. ....	231

Figure B.7. Flow diagram of the process of jitter function synthesis and analysis.....	233
Figure B.8. Characterization of jitter waveform $J_1$ . ....	235
Figure B.9. Characterization of jitter waveform $J_2$ . ....	236
Figure F.1. Experiment A data summary. Average number of sources reported for (a) 4 types of spectral envelope, (b) 3 types of spectral content, and (c) 3 degrees of correlation between modulating functions of two simultaneously-present spectral subsets. ....	270
Figure F.2. Experiment A data summary: interaction between the effects of spectral envelope and correlation. ....	272
Figure F.3. Experiment A data summary: interaction between the effects of spectral content and correlation. ....	273
Figure F.4. Experiment A data summary. Total number of <i>many</i> responses collected across subjects and repetitions.....	274
Figure F.5. Experiment B data summary. (a) Proportion responses within response categories averaged across spectral content, spectral envelope and subjects. (b) Proportion responses averaged across change and no-change tone pairs. ....	278

Figure F.6. Experiment B data summary. Proportion responses within response category averaged across spectral envelope and sub- jects. ....	279
Figure F.7. Experiment B data summary. Proportion responses within response categories averaged across spectral content and sub- jects. ....	281
Figure F.8. Experiment B data summary. Proportion responses within response categories averaged across subjects for three spec- tral content types with a vowel /a/ spectral envelope. ....	282
Figure F.9. Experiment C data summary. (a) Proportion responses within response categories averaged across spectral content, spectral envelope and subjects for each of the four tone pair combina- tions. (b) Proportion responses averaged across change and no-change tone pairs. ....	283
Figure F.10. Experiment C data summary. Proportion responses within response categories averaged across spectral envelope and subjects. ....	284
Figure F.11. Experiment C data summary. Proportion responses averaged across spectral content and subjects. ....	285
Figure G.1. Notation of <i>Original Melodies</i> for Taped Example 5 .....	288
Figure G.2. Notation of <i>Cross Melodies</i> for Taped Example 5.....	289

~ I ~

Image  
is everything  
imagination  
environment  
to compare with  
an active part  
reverberation  
sets in motion  
places us  
resonances  
re-minders  
takes root  
makes us  
defies  
measurement  
the realms of  
in the sense  
if we believe  
potentially there  
array  
addressing  
fusion  
simultaneous images  
reify a world  
this here  
when we do  
listening to  
virtual sources  
a violin from a cello  
what it is that allows  
in some way  
in which  
confused  
the listener decides  
in that realm  
involved with  
the area  
addressed.

$\sim \Pi \sim$ 

The listener drawn into  
realms  
of  
what that area looks at  
changing over time  
ongoing  
involved  
to keep up with  
changing



~ III ~

What goes on  
belongs  
across  
a sequence of  
order  
the relation of an event  
alternating tones  
against one another  
anchoring  
it's easier to tell  
it went by  
interlocking taking place  
a small amount of time  
marking time  
the interpretation shifts  
something you said  
I wonder why  
shifts over time  
within these  
a new interpretation of  
relations  
entered into  
gathered across time  
going on.

— Christopher Gaynor  
April 1981

## PROLOGUE

Imagine that you are walking blindfolded through the streets of a city. What do you hear? A combination of chugging and whirring metal and the popping of rubber on cobble stones is heard as a passing car. A rhythmic clicking of toe nails and jangling of small metal medallions is heard as a dog trotting by. A small herd of children goes giggling and screaming by on bicycles. You walk past a jack-hammer that is pounding the street with metal and your ears with painful pressure waves. Do we merely hear these sources as a collection of "sound events" (Julesz, 1971; p. 50)? Or do we hear each of these complex sound constellations as an "object"? I would opt for the latter claim. I don't just hear a jangling and clicking. I also hear a trotting dog with a well adorned collar. There is a certain coherence in the collective behavior of these events that I have learned and which allows (even induces) me to group them into the *auditory image* of the dog or the jack-hammer or the herd of children.

As organisms functioning in a not always so hospitable environment, it is important that our auditory systems - *as well as* our visual systems - be able to *objectify*, or *reify*, the elements of that environment. That is, we must be able to parse, or separate, the complex acoustic array into its many sources of sound if we are to be able, on one hand, to separate dangerous from innocuous or friendly objects, and on the other hand, to pay attention to a source in order to extract meaningful information from its emanations. In fact, the auditory system is so biased toward this parsing behavior that we have great difficulty hearing the sound environment as other than filled with objects. This is like trying to look at a landscape and seeing only patterns of colored light instead of trees, flowers, mountains, clouds, etc.

But now let us move to the world of sound artifice and enter (still blindfolded) a concert hall, where a full symphony orchestra is playing. What do you hear? At one level you probably hear the sound objects making up the orchestra: trumpet, violin, flute, tympani, contrabassoon, etc. Under many conditions you can "hear out" these various instruments whether they are playing melodically or in chords (though less so in the latter case depending on the voicing of the chord). One set of cues that is useful in separating the instruments is associated with their occupying different

positions in space. This certainly facilitates the auditory system's task. But imagine that the same orchestra is recorded with a microphone and then replayed over a single speaker. Now there is a *single physical source* emitting a very complex waveform. What do you hear? It is still relatively easy to hear out trumpets, violins, etc., though there is certainly a loss of acuity in denser orchestrations. Somehow we are able to parse the single physical source into multiple *virtual source images* and to selectively focus on their separate behaviors.

This is only one level of "grouping" or "parsing" of a musical sound environment. If three or more instruments play different pitches simultaneously, these events may be heard as a group. The composite would be experienced as a chord having a certain functional quality in a sequence of other chords. The single chord may, in some sense, be conceived as an object, as might the sequence of chords defining a certain harmonic progression. The harmonic functioning of any of these chords depends on the component pitches being taken perceptually as a group. A chord can also be perceptually "collected" from a sequence of pitches across time as with arpeggios. One might hear several groups of instruments that are blocked into differently textured organizations, e.g. rapid staccato winds against rapid legato arpeggios in the strings and a unison choral melody line. Here the "objects" would be accumulated by attending to a certain playing characteristic or movement as well as to various timbral characteristics.

The point is that many different levels of organization are possible and even desirable in a musical composition. One is less interested in hearing the *physical objects* (the instruments) than the *musical objects* (melodies, chords, fused composite timbres, group textures, etc.). Nevertheless, any listener brings into the musical situation all of the "perceptual baggage" acquired from ordinary in-the-world perceiving. And this will certainly influence the way the music is listened to and organized by the listener.

Assuming an interest on the part of the composer in the volitional act of perceptual organizing that may take place within each listener, one might ask the following questions about musical perception:

1. What might possibly be selectively perceived as a musical image? (By implication, what are the limits of musical attention?)
2. What processes can we conceive as being involved in the act of auditory organization?
3. What cues would a composer or performer need to be aware of to effect the grouping of many physical objects into a single musical image, or, in the case of music synthesis by computer, to effect the parsing of a single musical image into many?

Given that musical perception uses the same "bio-ware" as everyday perception, an understanding of these processes may help illuminate the questions posed above. Such is the aim of this dissertation.

## CHAPTER 1

### Research Problems on Source Image Formation

#### 1.1 Introduction

The auditory system participates in the forming of images evoked by acoustic phenomena in the world around us. An important aspect of the imaging process is the distinguishing of different sound sources. In order to be able to form images of sounds in the environment the auditory system must be able to decide which sound elements belong together, or come from the same source, and which elements come from different sources. This dissertation will address some of the issues involved both with the perceptual fusion of concurrent sound elements into a single source image and with the separation and distinguishing of different simultaneous source images.

In the everyday world, meaningful subsets of information generally come from a single source. And most often this source is not the only object producing sound in the environment. If the collection of acoustic elements emanating from the target source cannot be collected as a group and selectively perceived, it is very difficult to extract meaning from its emanations.

The problem of the psychologist and hearing theorist is to elaborate the psychological processes that allow this remarkable capacity of the human auditory system and also to understand the reasons for the limitations and tendencies of its performance. Julesz & Hirsh (1972) described three classes of information<sup>1</sup> provided

- 
1. This difficult term will always be used in this thesis in its most large and common-sense meaning of "that which carries a message." It is interesting to note that a widely used introductory textbook on the human information

through the senses (in terms of their "goal or purpose" for the perceiver):

1. Perception gives information on the basis of which the perceiver can know the presence of and recognize *objects* in his environment.
2. Perception can part from the object per se and be concentrated on received information in the form of *signs*, as in the case of communication, whether through language or art forms.
3. Perception serves as a mechanism for building *concepts*, not only modality-specific ones but general dimensions of experience like space and time, those basic aspects that form the matrix on which all perception is laid. (p. 290) [their emphasis]

Certainly in the case of art (and much of the material of this dissertation will be directed toward music perception; see Chapter 6) there is interest in understanding the processes that organize the sensory world. In particular, how do these processes operate with respect to ambiguous sensory information in order to beckon one's viewers or listeners beyond the boundaries of patterns of perceiving that have been established by experience in the world? In "normal" perception, source grouping and meaning extraction processes have the same target subgroups of information in the environment. That is, we usually receive meaningful messages from an integrated source. However, in art (or the psychology lab) these processes can be subverted; the different sub-processes that are involved can be made to diverge or conflict with respect to their conclusions about how the world is currently organized and what the meaning of that current state is. While this may be desirable in art, it is often a pit-fall in science. In presenting stimuli that are poor with respect to the richness of the sounds and their context in the environment, the laboratory situation often fails to evoke perceptual reactions that normally depend strongly on the total sensory context, and which are thus coded in terms of properties of, or behavior of, the sources themselves rather than being elicitable in the abstract (Schubert & Nixon, 1970). This caution urges one to consider carefully the framework within which one is

---

processing approach to psychology (Lindsay & Norman, 1972) neither defines the term nor lists it in the subject index. The first use of the term is in the sentence: "Let us start by examining how sensory information gets interpreted." (p. 7)

conducting one's experiments and to consider the relation between that situation and the "normal" world.

## 1.2 Paradigmatic Framework

The most viable framework for interpreting the data and evaluating the conclusions of experimenters in this domain is that of Neisser (1976). This view of perception is a synthesis of three classical theories of perception: direct perception, information processing and hypothesis-testing. These theories will be described very briefly below though these descriptions can hardly do justice to the fundamental arguments among them.

### 1.2.1 *Direct Perception*

This paradigm was proposed by J.J. Gibson (1966) primarily in relation to visual perception. Perception is not based on having sensations and then interpreting or organizing the "data of the senses." Rather, it is based on attention to the information in the ambient environmental light or sound, etc. Gibson proposed the notion of the "ambient array", which is a relational array or structure in the environment perceived from a "point of observation" that is not necessarily stationary (J.J. Gibson, 1974). The ambient array constitutes stimulus information, while the ambient light or sound constitutes the stimulus energy. The array is relatively invariant and independent of the observer. Information for perceiving a layout of objects is obtained by noticing what in the array is invariant under changes produced by the exploring movements of the observer.

It is certainly true that the ambient array provides accurate information about the environment that the perceiver must necessarily pick up. And we owe a great deal to Gibson for progressively making psychological experimentation cognizant of the problem of accurately specifying the nature of the stimulus. However, this view of perception, which is essentially "passive" since there is no activity of information processing, construction, interpretation or inference on the part of the observer, has great difficulty explaining all kinds of visual and auditory illusions that abound in art and music (Gregory, 1974) as well as many aspects of normal perceptual behavior. As Gregory (1981) notes, much new knowledge discovered by the disciples of Gibson has, ironically, proved very useful to people working in artificial intelligence, "who need to

know just which features of optical images are significant for scene analysis and object recognition by computer programs - though the computer programs provide just the kinds of activities that Gibson rejects for human perception" (p. 377).

### 1.2.2 *Feature Extraction*

The basic doctrine of perception as the extraction of features (cf. Lindsay & Norman, 1972) is that information from the environment is transduced by the sensory systems. This information is processed by specific mechanisms (feature detectors) which initiate neural messages in response to specific features of the information (i.e. features of the retinal "image" or cochlear "image", etc.). Information about such features gets passed on to higher processing centers for more complex processing such as sorting and comparing with previously stored information. Eventually this chain of processes results in perceptual experience in consciousness. Many aspects of perception can be accounted for by this model such as selective response of neural systems to orientations, colors, movements, spatial location, etc.

Although it seems highly likely that complex mechanisms in the brain are involved with the processing of sensory information, there are many aspects of the normal perceptual functioning of humans that cannot be dealt with, such as selective perception (different people notice different things of the same real situation), source separation (some portions of retinal or cochlear "images" belong to one object and not another), perception of meaning of events rather than detectable surface features, etc. Neisser (1967) suggested that some of these critiques can be ameliorated by proposing a 2-stage process where features are detected and analyzed in a pre-attentive stage that is followed by an act of construction in which the volition of the perceiver plays a role. This would necessarily be constrained by the kind and quality of sensory information received from the environment.

### 1.2.3 *Hypothesis-testing*

This paradigm of perception proposes that the act of perception is one of modeling (or forming hypotheses about) the behavior of the world and confirming them based on incoming sensory information. This notion dates at least from Helmholtz (1867) who proposed a process of Unconscious Inference underlying perception. A percept, then, is an unconscious conclusion resulting from these



inferences which are acquired through experience with the world. Craik (1943) proposed that the brain actually models aspects of reality. This kind of paradigm has proved the most useful to researchers in artificial intelligence concerned with pattern recognition, particularly in time-varying situations (cf. Sowa, 1984). As Bregman & Mills (1982) remarked, if percepts can be modeled then the problem of updating the world situation at each instant can be reduced to a simpler problem of performing certain kinds of transforms on the percepts ("percepts" being discernible objects in the environment in this case), rather than recomputing the whole situation at each instant.

#### 1.2.4 *Gestalts*

The *Gestalt* tradition (cf. K hler, 1929; Koffka, 1935) has some things to offer even though it was not considered explicitly by Neisser in his synthesis. The fundamental theoretical tenet of this school of perception (often neglected when their principles of perceptual organization are discussed these days) has been completely rejected; namely, that external forms are represented by corresponding shaped brain traces. This kind of first-order isomorphism is obviously incorrect given current knowledge of brain physiology. What is interesting, however, is what the Gestaltists proposed is a set of "Laws of Organization" (believed by them to be innate and *not* learned) upon which they claimed perceptual organization is based. These guiding principles are often quoted as being useful to many researchers interested in perceptual organization in vision and audition. But they do not suffice as explanations in themselves for the processes underlying organizations which take on these forms.

#### 1.2.5 *Neisser's Synthesis: The Perceptual Cycle*

An important point made by Neisser (1976) is that perception and cognition are not just operations in the head but transactions with the world. And these transactions not only *in*form the perceiver but *trans*form him or her as well. "Each of us is created by the cognitive acts in which he engages." (p. 11) Thus, our normal activity is a continual organizing of the form and meaning of the world. Perception is proposed as a cycle involving (a) the information to be picked up in the environment, (b) the exploring organism and (c) the knowledge the organism has about the way the world generally behaves. The organism operates from schemata (organized knowledge) of the world which direct perceptual exploration which samples the

available information in the environment which in turn stimulates modification of currently active schemata or calls into activity previously stored schemata. The schema itself is not a percept. It is more an anticipation or perceptual readiness for what is coming. As such it is the medium by which the past affects the future. But this influence on perception of past experience is not an adding of information from memory to stimulus information. Rather, existing schemata that were formed by experience *determine* what is most likely to be picked up. Perceptual learning, then, is a matter of *differentiation* rather than *enrichment* (Gibson & Gibson, 1955; E.J. Gibson, 1969).

Though this view might logically seem to imply that we cannot perceive that which we cannot anticipate, Neisser cautions:

Perception does not merely serve to confirm pre-existing assumptions, but to provide organisms with new information. Although this is true, it is also true that without some pre-existing structure, no information could be acquired at all. There is a dialectical contradiction between these two requirements: we cannot perceive *unless* we anticipate, but we must not see *only* what we anticipate. . . . Although a perceiver always has at least some (more or less specific) anticipations before he begins to pick up information about a given object, they can be corrected as well as sharpened in the course of looking. (p. 43) [his emphasis]

Often one finds in perceptual experimentation that subjects tend to perceive only what they expect to perceive even though other possible interpretations of the sensory information are possible. Also, in changing stimulus situations, researchers often find what are called hysteresis effects: the point at which a percept changes with stimulus change depends on whether the stimulus parameter is increasing or decreasing, for example. This is felt to be related to a perceptual "set" or bias on the part of the perceiver. The notion of the schema easily deals with such problems of interaction between perceiver and stimulus:

If the environment is rich enough to support more than one alternate view (and it usually is) expectations can have cumulative effects on what is perceived that are virtually irreversible until the environment

changes. But environments do change, and thus loosen the grip on old ways of seeing. The interplay between schema and situation means that neither determines the course of perception alone. (p. 44)

The paradigm of the perceptual cycle gathers together many important aspects of perceptual behavior:

1. Perception is inherently selective: the schema functions as a format for information pickup, and information not fitting such a format goes unused or unnoticed.
2. There are organizational tendencies in perception such that perception of the world is, for the most part, accurate: some schemata may be innate (it appears that new-born infants see and hear objects) or may be acquired from interaction with the world (learning). By way of such interaction the schemata come to reflect the laws and numerous regularities of the world. Sensory systems are adapted to exploit these laws and regularities in their organization and interpretation of sensory information, and further, they are constrained to prefer the interpretation that is most credible, given the current sensory input and a knowledge of the world's behavior embedded in schemata (cf. Hoffman, 1983).
3. In many situations where much sensory information is lost (due to occlusion or masking, for example), perception is relatively accurate nonetheless: if schemata are modeling the world and anticipating its behavior, lost information can be reconstructed according to the schemata in conjunction with an evaluation of the validity of the reconstruction in light of sensory information that *does* get through. When the schema predicts falsely, the perceiver may respond falsely, but generally responds as though he or she had actually perceived the missing information.
4. Highly improbable objects tend to be less readily perceived than probable objects of a similar nature: schemata reflect one's past experience with the world and thus also reflect, by their availability and richness, probabilities of encountering certain situations in the world.

Many other aspects of perception are, of course, unmentioned here, but the attempt is more to describe a framework within which to evaluate the notion of the auditory image as a predictive metaphor and with which to conceive and evaluate experiments on the perception of auditory sources in natural and not-so-natural contexts (such as in a computer music concert or a psychology experiment, for example). The discussion below on the notion of the auditory image will be brief (more to introduce the main aspects of the notion). A more in-depth evaluation will be conducted in Chapter 6.

### 1.3 The Auditory Image Metaphor

One of the main aims of the research project for which this dissertation is serving as a starting point is an understanding of the richness and complexity of music perception, in addition to the marvels of "ordinary" auditory perception of such *simple* stimuli as speech (for example!). It is important where music and psychology meet to develop metaphors for communication and cross-fertilization. In the search for a metaphor that embodies the combined aspects of auditory "impressions" from perception, memory and imagination, the notion of the *auditory source image* has proven fruitful in describing the results of auditory organizational processes to composers, musicians and psychologists. In particular, and directed toward musical interests, this metaphor has allowed the development of a common language for talking about the role of perception in musical processes that are to be embodied in compositions. The work to be discussed in this dissertation has been limited to the study of images deriving from sound stimulation, but many composers with whom I have worked find the metaphor and the delineation of its properties and implications useful for the imagining of musical possibilities at both conceptual *and* perceptual levels.

To summarize briefly, *the auditory image is a psychological representation of a sound entity exhibiting an internal coherence in its acoustic behavior*. The notion of coherence is necessary, if a bit general at this point. Since any natural and interesting sound event has a complex spectrum evolving through time, often involving noisy as well as periodic and quasi-periodic portions, it is important to consider the conditions under which these acoustically disparate portions *cohere* as a single entity. For example, many physical sources are quite complex acoustically and some even involve multiple sources of sound. But each of these can be perceived as a whole, as a single image. Certainly we could listen separately to some metal medallions at the

same time as the clicking nails of each of four different feet on a sidewalk. But the temporal nature of the pattern *as a whole* in conjunction with schemata for organizing it gives us the coherent auditory image of a domesticated, trotting dog. Human speech, as well, is a combination of noise and periodic sound sources (and even tongue clicks in the language of some African bush tribes) that are all integrated into one coherent sound stream that carries meaning. It is interesting to hear the African bush language because for my American English ears the tongue clicks are heard as a separate source and are not integrated (i.e. fused) with the other sound sources, whereas for the native speakers, these clicks modify the phonemic nature of the other sounds present.

One sense of the experimental question being posed is "What cues are associated with the formation of auditory images?" Some images have unitary and unequivocal perceptual attributes regardless of the number of individual spectral components of which they are composed. These are strongly fused images. Others have dispersed or equivocal attributes, such as the constellation of pitches evoked by the sound of a church bell. But the pitches in this constellation still seem to "belong together" and can be perceived or conceived as a single image. Imaging proceeds as a presentation to the conscious mind of this collection of the parts of a sounding body distributed across time and frequency.

"Belongingness" is a conceptual tool originally invoked by the Gestalt psychologists (cf. Köhler, 1929; Koffka, 1935). It is used here to name a family of rules of relations that any given sensory/perceptual system uses to group things into functional units. However, these are not rigid rules that box the sensory world into non-mutable objects. The domain of the artist and composer is one that challenges the predominant sensory patterns and evokes (among other things) the conscious transformation of perception by directing or beckoning one's attentional focus to different levels of form and structure in the work. What may at one moment be an "object" of focus for a listener may at another moment be an element collected into a *composite image*, wherein the "object" loses its identity but contributes to the quality of the more embracing image.

Here, at the outset, I have introduced what I consider to be the most powerful asset of the metaphor. It allows for a hierarchical or multi-leveled approach to auditory organization. We can consider a single trumpet tone as an image and speak of its

properties as a tone, e.g. pitch, brightness, loudness. We can consider a whole sequence of trumpet tones as an image and speak of its properties as a melody *and* of the functional properties of the articulation of individual tones as parts of the melody. We can consider a collection of brass tones, many occurring simultaneously, others in succession, as an image and speak of the properties of a brass choir as an ensemble or of the properties of a particular piece written for brass choir with harmony, polyphony, rhythm, force, *panache*, etc. All of this is to say that the metaphor allows the development and application of a broad set of criteria for musical coherence to be applied to music that permits both grouping and parsing of sound events into multi-tiered musical images.

In essence, at this stage of understanding, the problem is:

1. to search for a definition (or at least a circumscription) of what constitutes auditory coherence ("belongingness") from a psychological standpoint,
2. to understand its relation to the behavioral coherence of the physical world, and
3. to try to elaborate the knowledge structures and psychological processes underlying perceptual organizations of complex acoustic situations.

As Polanyi (1966) observes:

Because our body is involved in the perception of objects, it participates thereby in our knowing of all other things outside. Moreover, we keep expanding our body into the world, by assimilating to it sets of particulars which we *integrate into reasonable entities*. Thus do we form, intellectually and practically, an interpreted universe populated by entities, the particulars of which we have interiorized for the sake of comprehending their meaning in the shape of coherent entities. (p. 29)  
[my emphasis]

As mentioned, I have found the unifying metaphor of the auditory image to be useful in organizing thought in this direction. What needs to be considered at this point is the relation of the "image" to the previously described paradigm of perception, with

additional consideration of its functional validity as a psychological construct relating to the perception, memory and imagination of actual and virtual sound entities.

I would like to digress for a moment into an area of cognitive science where many of the aspects of Neisser's paradigm have been formalized, though I think there are several aspects of the formalization that depart considerably from Neisser's view. Most notably, I will draw much of the following material from a book on "conceptual structures" by J.F. Sowa (1984). The obvious shortcoming of this book is its lack of consideration of the structure of the ambient environmental array, whereas it is quite strong on notions of information processing in humans and machines and in its consideration of the involvement of schemata in cognition and perception.

In Sowa's view, memory is represented as a database that is a model of the evolving physical world. At any given moment, the state of the model represents the *knowledge* that has been acquired from the world. This implies already that there is an important structuring of the database. The process of perception, according to this framework, may be summarized as follows:

1. The sensory icon (e.g. retinal or cochlear "image") presents a partial or momentary view which lasts long enough to permit a continuity of perception.
2. Perception constructs a model from the incomplete views to have a complete situation.
3. A schema integrates the icons into stable images.
4. Conflicting schemata generate errors and illusions.
5. Perception tends to be top-down; that is, large-scale schemata are activated where possible so the most global percept is the first to occur; but this depends on the complexity of and familiarity with the object or situation.
6. Multiple levels of perception help deal with novelty and help process complex structures.
7. In cases of complexity, a more bottom-up approach may be used wherein the total object is constructed from lower-level percepts.
8. The interpretation of sensory input depends on the stock of percepts and

schemata.

There are several assumptions implied here, among which:

1. Percepts are pre-fabricated building blocks derived from experience.
2. A schema is a pattern for assembling perceptual units or other schemata into larger structures or unitary wholes.
3. These schemata can operate on various levels to discern structures in the sensory information (Sowa actually proposes that the schema *gives* the structure to the perceiver).

These larger structures are called conceptual structures and are made up of concepts and conceptual relations. The relation between a percept, a concept and an image is formalized as follows (p. 73):

For every percept  $p$ , there is a concept  $c$ , called the interpretation of  $p$ .  
The percept  $p$  is called the *image* of  $c$ . Some concepts have no image.

- a *concrete* concept has an image
- an *abstract* concept has no image
- the image of the interpretation of a percept  $p$  is identical to  $p$
- entities recognized by the image of  $c$  are called *instances* of  $c$ .

So there is an identity relation between a percept and its image, where the image notion serves as a kind of bridge between the percept and its interpretation (the concept or schema).

Evidence that images can be transformed in mental operations supports the notion that they are derived from models or regenerated from some kind of representation like schemata (cf. Kosslyn, 1980; Shepard & Cooper, 1982). Sowa proposes (as was also proposed by Neisser) that images serve as anticipations (or as perceptual readiesses) with ready-made percepts. This is supported by evidence that



1. familiar forms are matched by ready-made percepts; previous assemblies stored in long-term memory can be recognized very quickly, and do not need to be reconstructed from low-level percepts.
2. unfamiliar forms are reconstructed from percepts for their parts.

With this notion of the image as a reconstruction from conceptual structures or schemata, it is easy to describe the relation between perceived and imagined or recalled images. Internal images have the same nature as (though they are not identical to) sensory icons, and consciousness allows the brain to analyze and reinterpret an internal image using the same perceptual mechanisms used for sensory input. Thus, where "perception is a cyclic activity that includes an anticipatory phase; imagery is an anticipation occurring alone" (Neisser, 1976, p. 147).

If we then consider the temporal nature of auditory perception, these anticipatory schemata (a notion developed by Selz, 1913, 1922) must have some kind of temporal ordering in their structure which constrains the construction of auditory images from them, or the recognition of auditory events and objects by them. The important implication here is that these auditory images require a structural coherence. Since the schemata we are presuming to underlie them are ordered structures, perceptual grouping processes define the constraints on these structured relations. It should be mentioned that in Sowa's development of the above thoughts, there is never any mention of the processes (much less the cues) involved in deciding which elements are assembled from a complex environment into images. It seems to be more or less taken for granted that they *are* assembled and are assembled in the appropriate manner (though there is some discussion of the problems of speech perception in this respect). One aim of this dissertation is to delineate the nature of such grouping processes and the acoustic cues that are the information used to assemble images according to schemata, whether they be innate or derived from previous experience. The structure of these schemata would be required to reflect the criteria for reasonable behavior of acoustic sources.

#### 1.4 The Forming and Distinguishing of Auditory Images

Nearly all of the sounds we encounter in the world can be analyzed into many frequency components (or *partials*) which vary in frequency and amplitude over time. The auditory periphery performs such an analysis within certain limits of temporal and spectral resolution. However, we normally perceive such a complex sound "as a whole" rather than as many parts. We might say, then, that ordinary listening is *synthetic* rather than *analytic*, in the sense that it groups things together.

Why might it be useful for the auditory system to behave like this? One aspect of perception that is important for making one's way about in the world is the organization of that world into meaningful objects. Many types of sounds arise from objects we encounter repeatedly over the course of our lives. And most of the biologically relevant sources of acoustic information we encounter are physical systems (though modern times have necessitated the acquisition of life-protecting, electronic signal-producing systems such as the air raid siren). That is, the way the components of a source signal evolve individually and the way they maintain certain relations remains reasonably constant from one occurrence to the next. For example, forced-vibration systems such as the voice and most musical instruments each have predictable resonances in their spectra and all have series of partials that very closely approximate the harmonic series. We then categorize and recognize all of the constituent parts together, often even naming them as a group, such as an oboe tone, my father's voice, the word "tone". This kind of categorization results in a reduction of the amount of stored information that is necessary to represent the source in memory. These physical systems will behave under certain constraints which yield predictable patterns (Huggins, 1953; Schubert, 1975).

Consider also the processes by which we separate, or parse, two or more sources that are present simultaneously. In his seminal work on auditory psychophysiology, Helmholtz (1877/1885) noted that

... when several sonorous bodies in the surrounding atmosphere simultaneously excite different systems of waves of sound, the changes of density of the air, and the displacements and velocities of the particles of the air within the passages of the ear, are each equal to the algebraical sum of the corresponding changes of density, displacements and

velocities, which each system of waves would have separately produced, if it had acted independently. (p. 28)

He suggested that the problem becomes one of determining the means possessed by our sense organs to analyze the composite whole into its original constituents. Take the typical example of listening to a monophonic recording of a symphony orchestra evoked in the Prologue. There is a single pressure wave emanating from a single source of sound: the loudspeaker. Although the distinction is not as good as it would be were you sitting in the concert hall, it is still easily possible to separately hear many of the instruments playing simultaneously, even though all of the localization cues that can normally be used to aid in forming separate source images are missing. Much research has investigated the nature of auditory grouping processes responsible for the parsing of rapid sequences of sounds into auditory "streams", where a *stream* is the image of a sequence of sounds from a real or virtual source.<sup>2</sup> This dissertation primarily addresses the cues that play a role in the separation of simultaneous sources. This is a phenomenon that is understood only incompletely. No man-made analysis system has succeeded in parsing more than two simple sources and yet the auditory system is remarkably sensitive and accurate in this respect. It does have its limits, however. Listen to the large sound masses played by the strings in Ligeti's *Atmosphere* or Penderecki's *Threnody to the Victims of Hiroshima* and ask yourself how many individual instruments are playing simultaneously in that section. Certainly you can identify that there are "many", but "how many" is difficult to determine because the sounds are all so closely related that they obscure one another and are not individually distinguishable.

To clarify a bit the problem posed to the auditory system, imagine that you are listening to two speech streams at once and trying to extract the meaning of one of the messages. As the auditory system performs its limited frequency analysis, it must then decide which components belong to which source. Any two frequency components may or may not derive from the same source. Decoding the speech signal involves selecting among the components that are present and grouping some of them to define a voice. If these sounds fuse together into a whole, they will lose the qualities of the original voices; whereas if they can be perceptually separated, they

---

2. See McAdams & Bregman (1979) for a review of research on sequential auditory organization; the main themes of that paper will also be summarized in Chapter 6.

can be separately recognized. Bregman, Abramson & Darwin (1983) suggested that part of this task could be done by competing speech sound recognizers trying to match and select target properties from the incoming mixture (presumably being guided by semantic schemata and schemata more oriented toward voice behavior that are trying to anticipate the next "move" by the incoming message). If, however, the target properties were the same for two or more possible interpretations, the choice of which interpretation fit the acoustic situation best might be made more simple if a decision could be made as to which elements belonged to which source. If there were source component grouping processes independent of speech sound recognition processes, the two kinds of input to the total organization might enhance the possibilities of extracting a meaningful message. One aspect of grouping already apparent in the above proposal is that *multiple "levels" of processing* may be simultaneously contributing to a given interpretation of the behavior of a target source. Here the different levels would correspond to levels of the auditory nervous system which are apparently dedicated to greater degrees of complexity the nearer they are to the auditory cortex (cf. Evans, 1971).

A different kind of multi-leveled process that fits neatly into a schema-oriented paradigm of source perception concerns the *multiple, hierarchically-organized levels of structure perception* that are possible in listening to music, particularly many-sourced musical signals such as an orchestra. In this situation one can attend to a single melody line (generally a single physical source), or to the qualities of a contrapuntal composition such as harmonic progression and passing of musical material between lines, or in dense orchestrations to conflicting harmonic developments between different subgroups as in a double fugue, or even to large-scale qualities such as texture and overall spectral balance of the instrumentation. All of these can be going on at the same time and the perception of each depends on one's taking the relevant components as a group. What I am intimating here is that the several qualities described above are properties that emerge as a function of the elements being grouped together. This notion of emergent properties of groups is taken (in a larger sense) from Bregman's notion that perceived qualities are assigned to sources based on the grouping of elements into the source (Bregman & Pinker, 1978). This argument risks being circular if one insists on "source" instead of "group", more generally, and even runs into contradictions in the dichotic speech perception literature, which will be discussed in the next section.

Another property of grouping processes is that they seem to be *heterarchical* in certain respects (probably within a given level of structure). This means that there are many different criteria for grouping decisions and they may not always converge on the same solution. In some situations, one criterion may have stronger evidence than other conflicting ones and its "proposition" for a grouping solution may win out. With a small shift of attention on the part of the perceiver, this balance may be shifted as well and another interpretation may result. In cases of true ambiguities, either illusions occur or the attentional focus of the perceiver plays a very strong role.

Conceiving of the process as weighing evidences in a decision-making situation points to its *heuristic* nature. Perception is being taken as a process of modeling of the world on the basis of the available sensory data and previously stored schemata. The model or schema is created by the composition of a number of basic concepts which we can rearrange to form these models. In ambiguous or polyvalent situations the sensory data can support alternate schemata or interpretations, and the most credible or "correct" model would be considered to account for the greatest range of currently available data. In general, perception tends toward the simplest (most global) interpretation until conflicting evidence accumulates (Bregman, 1977, 1978a,b, 1980).

This introduction to the proposed nature of grouping processes is admittedly general, but I think it important to consider these generalities in order to frame the subsequent, more specific considerations. So far I have proposed that grouping processes are heuristic, heterarchical and multi-leveled and that the level of grouping that is currently active or to which one is attending (and thus to a certain extent effecting) determines the perceived emergent qualities. It will become apparent in reviewing the literature on this domain that certain groupings are difficult to affect with attention, such as groupings related to speech sounds. Let us consider more specifically certain kinds of grouping that are related to perceptual fusion.

### 1.5 Perceptual Fusion

As one reads the literature on things denoted by "fusion" one finds that this word is used in many ways, as Cutting (1976) remarked. Cutting himself delineated six types of fusion related to dichotic speech stimuli. The most common features among these were the facts that in most situations the individual elements that were fused were no longer separately audible (though this was not always true) and that their combination gave rise to some new quality that was not perceptible when the isolated elements were separately presented, i.e. an emergent property of the new group became apparent.

This disappearance of individual elements "in the service of the whole" reminds one of Helmholtz' (1877/1885) notion of analytic and synthetic perception. He apparently derived these notions while trying to develop his ability to "hear out" or perceptually analyze the separate components of harmonic tones. In alternately attenuating the tones of two bottles tuned an octave apart, he could get to a point of being able to hear out the both when present simultaneously. However, after letting them both play for awhile, "by degrees, as my recollection of the sound of the isolated upper tone died away, it seemed to become more and more indistinct and weak, while the lower tone appeared to become stronger" and acquired the timbre of the fused combination, which he remarked to be different from the individual timbres of the separate bottles. Upon hearing this he concluded:

We then become aware that two different kinds or grades must be distinguished in our becoming conscious of a sensation. The lower grade of this consciousness, is that where the influence of the sensation in question makes itself felt only in the conceptions we form of external things and processes, and assists in determining them. This can take place without our needing or indeed being able to ascertain to what particular part of our sensations we owe this or that relation of our perceptions. In this case we will say that the impression of the sensation in question is *perceived synthetically*. The second and higher grade is when we immediately distinguish the sensation in question as an existing part of the sum of the sensations excited in us. We will say then that the sensation is *perceived analytically*.<sup>3</sup> (p. 62) [his emphasis]

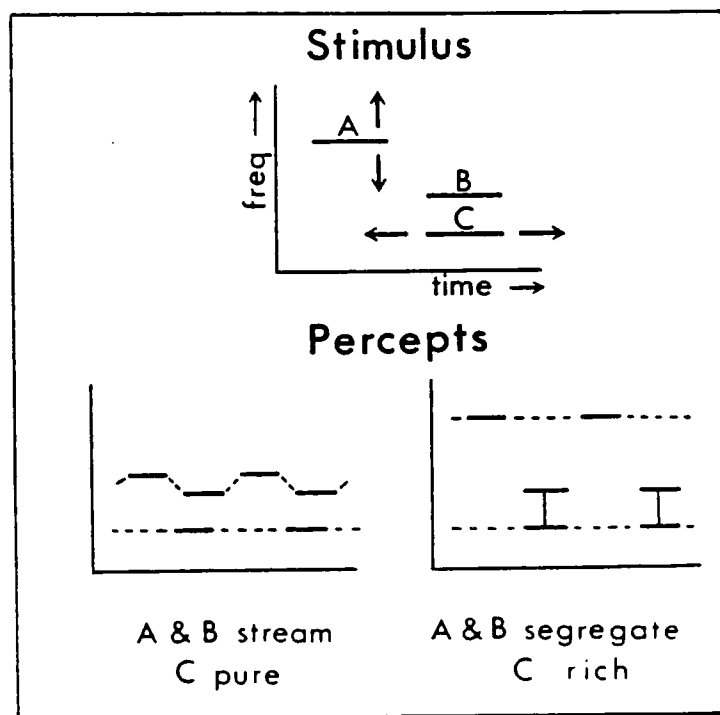
Helmholtz further observed that while, in general, the upper partials of instrument tones are very difficult to hear, one can direct one's attention to certain of these partials in the sounds of plucked strings and sirens. It has been established that the human ear is capable, with sufficient training, of hearing out the lower individual partials (up to about 5-7) of a sustained, unmodulated harmonic or inharmonic tone (Plomp, 1964; Plomp & Mimpen, 1968). This analyzability of steady-state tones may be one aspect of the unnaturalness reported for resynthesized voice and instrument tones which do not include vibrato (periodic) or jitter (aperiodic) modulations in their component frequencies (Kersta, Bricker & David, 1960; Sapozhkov, 1973; Grey, 1977; Grey & Moorer, 1977; Chowning, 1980; McNabb, 1981). Aside from the richer dynamic quality given musical sounds by the addition of frequency modulation to the harmonics, it seems possible that their fusion, i.e. the *inability* to hear the complex tone as *compound*, may also be a factor in perceived naturalness. Chowning (1980) reports that with synthesized voices "it is striking that the tone only *fuses* and becomes a unitary percept with the addition of the pitch fluctuation . . ."

There is evidence that other perceptual properties or sound qualities, such as phonemic identity and timbral quality, do not arise unless all of the acoustic elements necessary to give rise to this quality are grouped together. As mentioned this grouping usually results in a perceptual fusion where the individual elements lose their qualities or identities as such, but collectively give rise to something new. Other times (under certain laboratory conditions such as dichotic listening) certain elements may both contribute to a new emergent property of the group, and, *at the same time*, maintain an identity of their own. This latter phenomenon has been called duplex perception in that a simple stimulus element simultaneously contributes to two different percepts, considered to be at different levels of processing, e.g. "phonetic" vs. "auditory" perception (Liberman & Studdert-Kennedy, 1978; Isenberg & Liberman, 1979).

Bregman & Pinker (1978) demonstrated the former kind of grouping. They presented a stimulus in which a pure tone, *A*, alternated with a complex tone composed of two pure tones, *B* and *C* with *C* lower in frequency (see Figure 1.1). This pair was repeated cyclically. The frequency separation between the sequential

- 
3. Presumably what is "lower" here is more "normal" and what is "higher" requires the development of special perceptual skills of differentiation.

components *A* and *B* and the temporal synchrony between the simultaneous components *B* and *C* were varied in the experiment (though they were constant for a given presentation). Subjects were asked to judge the extent to which tones *A* and *B* formed a single or separate sequential "streams" and to judge the relative richness of the timbre of tone *C*. The results showed that when *A* and *B* were close in frequency, and *B* and *C* were asynchronous, *A* and *B* were judged more often as forming a single stream while *C* was judged as being more pure (percept indicated on the left in Figure 1.1). Conversely, when *A* and *B* were distant in frequency and *B* and *C* were synchronous, *A* and *B* were judged more often to be in separate streams and *C* was judged to



**Figure 1.1.** Schematic representation of the stimuli used by Bregman & Pinker (1978) and two common perceptual results. Each horizontal line represents a sinusoidal frequency component. The dashed lines represent perceived sequential organization and the vertical solid lines represent perceptual fusion.

be richer in timbre (percept indicated on the right in Figure 1.1). The experimenters



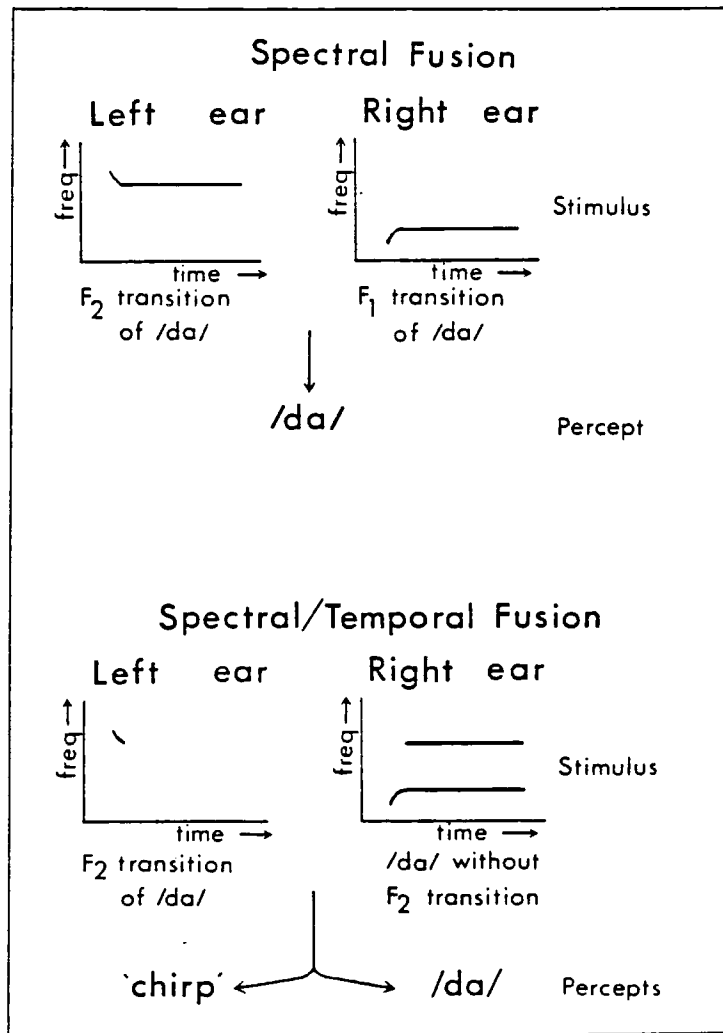
interpreted this as indicating that the degree to which the tone *C* was perceived as being rich was determined by the degree to which it was fused into a simultaneous organization with *B*. This was inversely related to the degree to which tone *B* formed a sequential organization with tone *A*. They concluded that

1. an acoustic element cannot be a member of two organizations at once, i.e. *B* cannot belong to a stream with *A* *and* be part of a fused tone with *C*, and
2. the richness of tone *C* was dependent not merely on the degree of temporal overlap with tone *B* but also on the extent to which *B* was grouped or fused with *C*.

Cutting (1976) demonstrated both types of perceptual grouping. He called them "spectral fusion" and "spectral/temporal fusion," respectively. These are represented schematically in Figure 1.2. Both were originally dichotic phenomena which Rand (1974) showed to exist for monaural listening as well. In the case of "spectral fusion" (originally investigated by Broadbent (1955) and Broadbent & Ladefoged (1957)), Cutting presented the first two formants of the consonant-vowel syllable /da/ to separate ears over headphones. If the temporal onset synchrony relations were appropriately adjusted and the fundamental frequencies were identical, Ss reported hearing a /da/ syllable 85% of the time. This stimulus was also judged as being one source 60% of the time. When onset times were desynchronized by as little as 20 msec, the choice of /da/ over /ba/ or /ga/ dropped to 75%. When the fundamental frequencies of the two formants were separated by 2 Hz, Ss reported hearing more than one source 98% of the time, but separations as much as 80 Hz had no effect on the identifiability of the /da/ syllable! When the individual channels are presented in isolation, non-speech sounds are heard and are accurately lateralized to one ear or the other.

Broadbent & Ladefoged (1957) noted that when the fused percept /da/ results, subjects are unable to say which formant is coming to which ear. Thus, this result is similar to that of Bregman & Pinker in that the components give rise to their emergent qualities when they are fused as a whole, and are no longer available to analytic perception as individual elements. In the Cutting experiment, however, there seems to be some difference or even independence between processes underlying source identification (at least for vowels) and source multiplicity judgments here. No

mention was made of whether or not the multiple sources were *all* perceived as /da/; subjects were asked simply to make a single choice of B, D or G in the identification



**Figure 1.2.** Schematic representation of the stimuli used by Cutting (1976) and the most common perceptual results. The lines represent the frequency trajectories of the first ( $F_1$ ) and second ( $F_2$ ) speech formants.

experiment, or to respond with 1 or 2 in the "number of sources" experiment. There is evidence here, though, for multiple perceptual decisions possibly being made on the same stimulus set. When asked to identify the stimulus as /ba/, /da/ or /ga/, the subject can ignore the spectral content (coming from two separate fundamentals

and thus giving rise to different pitches)<sup>4</sup> and pay attention preferentially to the evolving composite spectral form derived from both ears to decide if it is most like a /ba/, /da/ or /ga/ syllable. When asked to decide on the number of sound items present, the subject may then take into account such things as the presence of multiple pitches and the fact that separate spectral regions are arriving in different ears.<sup>5</sup> This is a case where separate processes arrive at independent and irreconcilable solutions. One hears a fused /da/ and presumably cannot hear the individual *formants* and yet one hears two *pitches* and judges that there are two sources.

The /da/ stimulus for "spectral/temporal fusion" (first investigated by Mattingly, Liberman, Syrdal & Halwes, 1971) consisted of the first formant and the steady-state portion of the second formant being presented to one ear, while the second formant transition was presented to the opposite ear. When the formant transition was not presented, subjects reported hearing a /ba/ 85% of the time. When temporal synchrony relations were appropriate (within 10 msec of normal relations), Ss reported hearing a /da/ in the ear with the steady-state signal plus a non-speech sound ("chirp") in the ear with the second formant transition. As the transition segment was moved out of synchrony by about 10 msec, /da/ identification dropped from 81% to below 75%. At least 85% of the "number of sources" judgments on this stimulus recorded 2 items, regardless of the difference or similarity of the  $F_0$ 's of the two formants.

Cutting proposed that the source identity changes because the appropriate acoustic elements are fused into a single unit. In this case the transition segment did not fuse with the contralateral sound to the extent that it lost its own status and identity as a separate event, but it certainly did fuse to the extent that it transformed the perceived identity of the contralateral sound.

4. The second formant of the /a/ in Cutting's stimulus was centered on 1620 Hz, i.e. on about the 16<sup>th</sup> harmonic of a 100 Hz fundamental frequency ( $F_0$ ). One would expect this to give a rather weak pitch sensation since these harmonics are well beyond the dominance region of partials contributing strongly to perception of a "missing"  $F_0$ , or virtual pitch (Plomp, 1967; Ritsma, 1967). However, these components still fall within the existence region of partials able to give rise to a virtual pitch (Ritsma, 1962, 1963).
5. Note that even in the standard stimulus with appropriate onset synchrony and an identical  $F_0$ , 40% of the subjects' responses judged this stimulus to be composed of 2 items!

Again we have irreconcilable solutions between speech sound recognition processes and some other process. In the case of "spectral fusion", localization of the two formants was overridden by the speech process, while pitch processing was independent. In "spectral/temporal fusion", the speech process was influenced by the spectral form of the signal in the contralateral ear, but was unaffected by its location or its pitch. Obviously, these are very unusual sounds to be making judgments on, but the results do point again to the heterarchical nature of grouping processes and to a certain fallibility of the processes that coordinate the final organization.

A similar study on more musical stimuli was conducted by Pastore, Schmuckler, Rosenblum & Szczesiul (1983). In their experiment two tones at a musical interval of a perfect 5<sup>th</sup> (e.g. C and G) were presented to one ear, while a tone was presented to the other ear that was either a major or a minor 3<sup>rd</sup> above the low tone (e.g. E or Eb). Subjects appeared to be able to make judgments on the harmonic nature of the chord (major or minor triad) and still hear the separate tone in the other ear. This result (though still an example of duplex perception) is less surprising than the dichotic speech result since we would consider the extraction of the quality of a chord to be related to a higher level grouping than the extraction of the pitch or the separation of a single tone. A higher-level grouping does not preclude membership in a lower-level subgroup, when there is no organizational conflict – as in this case.<sup>6</sup> That such may be the case in the dichotic speech examples requires a more careful consideration of the process by which qualities of images are derived and how this relates to grouping decisions.

### 1.6 Derivation of Image Qualities

The evidence discussed above seems to indicate an independence between processes that perform grouping operations and those that derive perceptual qualities such as pitch, timbre and phoneme identity. Since a vast amount of research has been published on pitch perception and since the pitch of harmonic tones is rather simple, that area will be treated only summarily. A more thorough treatment of

---

6. This proposes a modification of Bregman & Pinker's (1978) claim that a single element cannot be a member of two organizations at once. Perhaps this is true if the organizations are operating at the same level of processing and are mutually exclusive. But if one organization can logically be a subset of another, no conflict would arise and both may obtain.

timbre and vowel perception will follow, particularly as certain aspects of that domain will bear heavily on experiments to be reported in succeeding chapters.

### 1.6.1 *Pitch Perception*

Modern pitch theories (Goldstein, 1973; Wightman, 1973; Terhardt, 1974) favor the recognition of regularity of spectral pattern as a basis for the perception of a single pitch – particularly a regularity in agreement with the harmonic series found in most musical and many significant environmental sound sources such as the human voice. A number of experiments on tone complexes that were designed to depart systematically from the harmonic series (beginning with de Boer, 1956) tend to support this view; within limits, the perceived pitch moves toward the best harmonic compromise, and the greater the departure from harmonicity, the weaker and more equivocal the pitch response. For example, studies of the fusion of inharmonic complex tones of the form  $f_n = n^s F_c$  (where  $n$  is the partial number and  $f_n$  is its frequency) indicate that perceived (judged) fusion decreases in a monotonic, and seemingly linear fashion as  $s$  departs from a value of 1.0 up to 1.07 and down to 0.93 (Cohen, 1979, 1980).

But a departure from a harmonic *spectral* pattern is only one way of describing changes in various inharmonic series. From a temporal view, one aspect that is lost is the presence of periodicity. The prevalence of time intervals related by integer sub-multiples is an important concept in the operation of the volley (rotation) theory of neural following (Wever, 1949). This may also be an important aspect of the auditory response to harmonic complex tones.

To the extent that judgments on the number of sources can be made on the number of perceived pitches, we would expect simple harmonic series to be judged as single sources most of the time. Multiple source judgments would result from inharmonic series and from the presence of multiple harmonic series (see also sections 1.7.2 and 1.7.3 below).

### 1.6.2 *Timbre and Vowel Quality Perception*

Another quality important for musical and speech sources is that dimension of timbre (or tone color) derived from the spectral form or spectral envelope. This form is most often due to the resonance structure of the source, i.e the ensemble of resonant cavities following the acoustic excitation in the sound producing system. These cavities filter the original acoustic input waveform in fairly predictable ways. The more of these cavities there are, either in series or in parallel, the more complex the spectral form tends to be. In music this spectral form is associated with different aspects of tone color such as "brightness" or "sharpness" (von Bismarck, 1974; Grey, 1975, 1977; Ehresman, 1977; Ehresman & Wessel, 1978; Grey & Gordon, 1978; Wessel, 1979, 1983). In speech, spectral form gives rise to vowel qualities and certain consonants.

According to Plomp (1970), the first experimental demonstration that the timbre differences between vowels are determined by the formant peaks in the amplitude pattern was reported by Willis (1830). These results were confirmed and extended to other spectral forms by Helmholtz (1859, 1877/1885). Most experimenters of that epoch agreed that timbre was related to constant spectral form rather than constant amplitude ratios among the harmonics (Donders, 1864; Grassman, 1877; Helmholtz, 1877/1885). More recent evidence from Green and colleagues has demonstrated that subjects are capable of remembering a simple spectral form and comparing it with another on a successive trial to discern whether one of the partials was augmented or diminished in intensity. This capability was independent of large changes in overall level between the observation intervals of a given trial and was independent of the duration of silence (at least up to 8 sec) between those intervals (Spiegel, Picardi & Green, 1981; Spiegel & Green, 1982; Green, Kidd & Picardi, 1983; Green & Kidd, 1983; Green, Kidd & Mason, 1983; Green & Mason, 1983).

The notion that a relatively constant spectral envelope yields constant timbre was demonstrated by Plomp & Steenecken (1971) for non-vowel sounds. They showed that harmonic sounds with different pitches and the same spectral envelope were perceived as more similar than sounds with different pitches and constant amplitude relations between harmonics. In the former group of tones, the amplitudes of individual components changed with a change in  $F_0$  while the overall spectral form remained constant. In the latter group of tones, the amplitudes of the components

remained constant with changes in  $F_0$ , which distorts the spectral envelope. Even though spectral form seems to play a rather minor role in contributing to the *identity* of musical instruments,<sup>7</sup> its contribution is essential for the identification of vowel sounds.

Some investigators have suggested that the absolute position of formant (resonant) peaks is important for the identification of vowels (cf. Stumpf, 1926; Fairbanks & Grubb, 1961). However, other evidence demonstrates that vowel identification changes little when all formant frequencies are transposed upward or downward in frequency by the same percentage, provided this shift is not too great (Potter & Steinberg, 1950; Peterson & Barney, 1952; Miller, 1953; Fant, 1959; Stevens & House, 1972). In these cases, the frequency intervals between formant peaks would remain constant and it would thus be their interval relations that were most important for identification.<sup>8</sup> This notion has been supported by the work of Sapozhkov (1973) who emphasizes the perceptual importance for speech perception of the *formants themselves* versus the *overall spectral form*. Scheffers (1983) demonstrated that the identifiability of synthesized vowels in noise depended on the detectability of the first two formants. He proposed a model of vowel identification based on formant frequency template matching, which performed reasonably well for single synthesized vowels.

There exists, in light of these data, the problem of explaining how the complex change of spectral form with pitch and intensity that is found in voice and musical instruments still yields a constant identity. In fact, these changes are necessary to maintain identity. Sundberg (1975, 1978, 1982) mapped the trajectories of the first

- 
7. It is currently believed by many experimenters that the "signatures" of musical instruments are most closely linked with temporal fluctuations in the components, particularly during the attack portion (first 60 msec or so) of sounds produced by these instruments (cf. Berger, 1964; Saldanha & Corso, 1964; Wedin & Goude, 1972; Grey & Moorer, 1977).
  8. Of course, in all of these studies the tests were on more or less isolated vowels which is a much different condition than recognition of vowels in context. Anyone who remembers Alvin and the Chipmunks is reminded that within a context of continuous speech or song, voice signals can be transposed as much as an octave while maintaining their intelligibility. In contrast to this, though, are the problems reported with speech perception in a helium environment where a direct transposition of the frequency spectrum also occurs.

four formants with changes in  $F_0$  for 4 vowels in a professional soprano. When  $F_0$  began to pass the first formant frequency,  $F_1$ , this formant began to track the  $F_0$ .<sup>9</sup> The other formants also changed systematically with pitch for  $F_0$ 's greater than about 300 - 400 Hz. Another good example of significant change in spectral form with pitch is the clarinet. For this instrument, the changes are so dramatic that the different registers have completely different timbral characteristics and have been given different names by musicians. And yet one still, for the most part, recognizes a clarinet as such, regardless of the register it plays in. This raises questions concerning the relation of source identification processes to higher-order acoustic invariances (cf. Gibson, 1966) or to the learning of a complex constellation of characteristics associated by experience with a source. These notions will be discussed in Chapter 6.

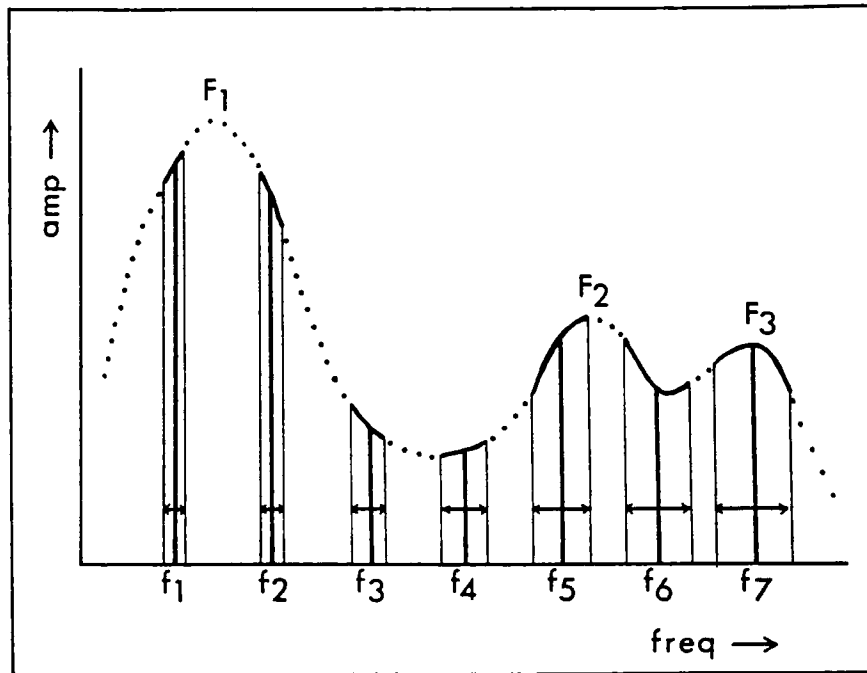
In spite of these systematic changes of resonance structure with pitch register, the resonant frequencies of these structures tend to change relatively slowly with respect to the rates of modulation found in vibrato and jitter in musical sounds (cf. Bjørklund, 1961 for voice). Rodet (1982) has developed a technique for the determination of vocal formant structures which depends on the coupled amplitude and frequency modulations defining local slopes of the spectral envelope.<sup>10</sup> In the singing voices measured, these FM waveforms are provided by vibrato and natural jitter. This technique is particularly useful for high pitched sounds where there are not enough frequency components to accurately define the spectral envelope. Very convincing voice syntheses have been obtained based on these analyses (Rodet, 1980b; Rodet & Bennett, 1980; Bennett, 1981).

Synthesis techniques where the spectral envelope moves with the vibrato and jitter have also generated acceptable results (see Chowning, 1980, 1982, for FM synthesis, and McNabb, 1981 for wavetable synthesis). With these latter techniques, however, there are limits to the acceptability of large modulation widths and of pitch glissandi, since the spectral envelopes are grossly distorted by the modulation, the amplitudes remaining constant with frequency movement.<sup>11</sup>

9. It should be noted that this is not the case in normal speech where  $F_1$  changes in a vowel-dependent, rather than pitch-dependent, manner.

10. Lewis (1936) also showed how vibrato effects a tracing of the singer's formants by measuring a coupled amplitude-frequency modulation that was a function of the formant structure.



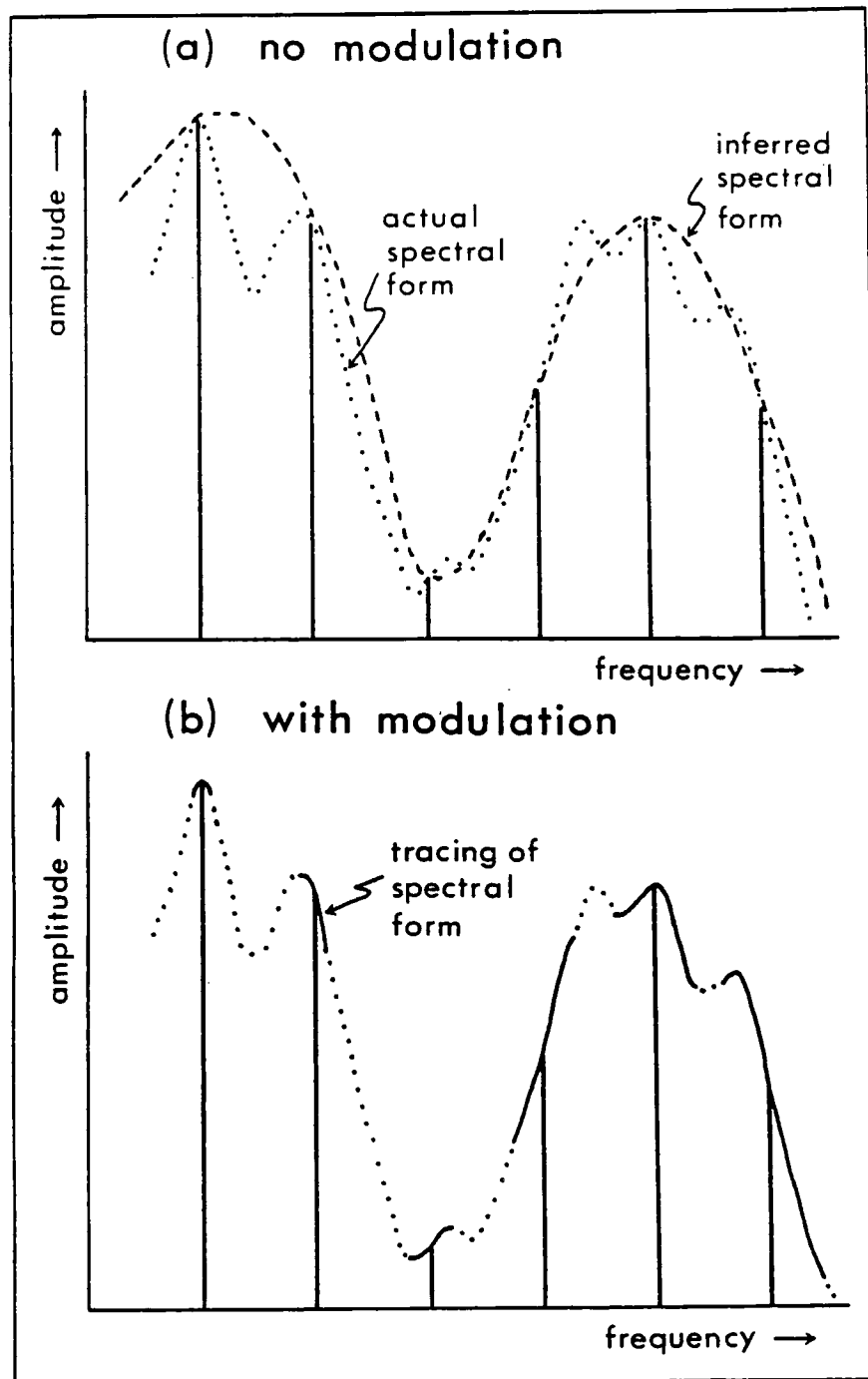


**Figure 1.3.** The spectral form created by a 3-formant resonance structure is represented by the dotted line. As the harmonics are modulated in frequency, their respective amplitudes fluctuate as a function of the spectral envelope. This is indicated by the solid portions on the dotted line. Note that these trace out portions of the formant shapes and, in the case of  $f_1$  and  $f_2$ , these shapes 'point' toward the formant peak. [from McAdams (1984)]

11. In FM synthesis of voices, Chowning (1973, 1980) uses a 3-formant model of the singing voice. One carrier frequency is placed at the harmonic nearest the formant frequency for each of the first 3 formants. These carriers are then modulated by the same modulation frequency which is set equal to the  $F_0$ . The modulation indices are used to separately control the bandwidth of each formant. Both modulation and carrier frequencies are modulated synchronously in frequency to obtain vibrato and jitter. In wavetable synthesis (cf. Moorer, 1978), one period of a complex harmonic waveform is stored in a "wave table". Then this table is read at a speed that is determined by the  $F_0$  desired. This speed can be varied over time to obtain vibrato and jitter.

The superiority (with respect to naturalness and flexibility of use) of syntheses maintaining constant spectral envelope suggests that this constancy may be important in different aspects of source perception. For example, one may hypothesize that one cue for the perceptual invariance of a resonant source is the tracing of its spectral envelope by its coherently modulating frequency components. In Figure 1.3 the horizontal axis represents linear frequency and the vertical axis, amplitude. There are three formants (bumps in the curve) represented here. Notice that for a given frequency excursion of the fundamental frequency, there are progressively greater excursions for the higher harmonics. This is due to the linear frequency scale used in the diagram. Each harmonic is moving a constant percentage lower and higher, so even though the excursion at higher harmonics is greater when measured on a linear scale, it still maintains a constant ratio distance from all of the other harmonics. The overall form of the resonance structure is indicated by dotted lines. The amplitude by frequency trajectories of each partial are indicated by solid lines. This tracing of the spectral form may serve to reduce the ambiguity concerning the actual resonance structure of the source.

If this hypothesis were true, we would expect the following result. If a source with a complex resonance structure had a  $F_0$  that was high enough so that few partials fell within each resonance region, there would be a great deal of ambiguity about the identity of that resonance structure. By adding some kind of low-frequency FM which caused the spectral envelope to be traced by the partials, this ambiguity would be reduced, and the ease and accuracy of identification of the source would increase. In Figure 1.4 another spectral form is plotted. This corresponds to the vowel /a/. The fundamental frequency is quite high here so that not very many harmonics fall into each formant region. In this case the formant structure is not very well defined and accordingly, the perception of the vowel sound would be very weak if at all existent. However, when the spectral components are made to modulate in frequency, by jitter, vibrato or intonational movement, their amplitudes trace the spectral envelope and the auditory system then has access to the *slopes* of the formants around each partial. This adds important (even essential) information which the system can use to identify the nature of the source. So one important function of frequency modulation is to reduce the ambiguity of the nature of the resonance structure defining the source.



**Figure 1.4.** The vowel /a/ is plotted with a high fundamental frequency where there are few harmonics present. Without modulation (a), the inferred spectral form (dashed line) would be very different from the actual spectral form (dotted line). With modulation (b), the spectral slopes give a much clearer indication of the spectral form. [from McAdams (1984)].

This seems almost intuitively obvious but there are claims both for and against this idea in the literature on vowel perception. Carlson, Fant & Granström (1975) claimed that introducing an intonation contour ( $F_0$  or pitch glide) with a maximum deviation of 4% (68 cents peak deviation) from the mean frequency added a slight uncertainty in the vowel identification decision. There was less uncertainty with a steady  $F_0$  stimulus. It is difficult to discern from their paper what relation this has to the problem posed above since they were changing both  $F_0$  center frequency (between 100 - 160 Hz) and  $F_1$  (first formant frequency; between 250 - 350 Hz) between the boundaries for the Swedish vowels /i/ and /e/. It is not clear from the paper what the judgment is in their experiment. If we presume it is to select one of the two vowels, then the pitch glide condition actually gives better performance at some combinations of  $F_0$  and  $F_1$  than does the steady-state stimulus (see Fig. 7, p. 81, in their paper). In any event, there are not enough data here to draw an unambiguous conclusion (even about the ambiguous nature of synthetic vowel perception).

Sundberg (1977) claims that vibrato has little, or even a detrimental, effect on identification of sung vowels synthesized on the basis of data from a professional soprano. Sundberg even speculates on the basis of this claim that "a singer may in practice profit systematically from this effect of the vibrato [obscuring perception of formant frequencies] so as to reduce the perceptibility of her deviations from the formant frequencies of normal speech." (p. 265) This finding seems so anti-intuitive as to deserve closer inspection.

There are several problems with his study as concerns the synthesis of stimuli and the analyses of the listeners responses. For one thing, the formant synthesis data were derived from tones sung with a  $F_0$  of approximately 262 Hz, while the synthesized tones had  $F_0$ 's of 300 - 1000 Hz but the same spectral form as the 262 Hz tone. As mentioned previously, it is Sundberg's own contention that formants must move with pitch register to maintain vowel identity.

Secondly, subjects were presented 6 different vowel stimuli, and allowed to choose among 12 vowels. Sundberg hypothesized that in making an identification, the actual stimulus is compared to some internal representation of known vowels and that the

response is the best match. To quantify the responses he chose the formant frequency data of Fant (1973) for the 12 possible response vowels. A measure of the "scatter" of responses for a given stimulus vowel was calculated as follows, based on the theoretical frequencies for the first three formants of the response vowels:

1. the formant frequencies ( $\bar{M}_k$ ) of the "average response vowel", expressed in Mels,<sup>12</sup> were calculated as the average of a given formant across all responses,

$$\bar{M}_k = \frac{1}{n} \sum_{i=1}^n M_{kRi} \quad (1.1)$$

where  $M_{kRi}$  = the  $k^{\text{th}}$  theoretical internal formant frequency for response vowel  $R$  given in judgment  $i$  (from Fant's data).

2. then a "scatter" statistic ( $D$ ) was calculated as the average distance (in a 3-D Euclidean space) of each response vowel from the average response vowel:

$$D = \frac{1}{n} \sum_{i=1}^n [(\bar{M}_1 - M_{1Ri})^2 + (\bar{M}_2 - M_{2Ri})^2 + (\bar{M}_3 - M_{3Ri})^2]^{\frac{1}{2}} \quad (1.2)$$

Using this measure, Sundberg shows that an increase in  $F_0$  is accompanied by increasing  $D$ . He interprets this as identification becoming more ambiguous or difficult. This would, of course, be expected *a priori* given that the formants are not changing naturally. This would also be expected for non-modulating stimuli given that formant definition is worse at higher  $F_0$ 's. No systematic difference was found between vibrato and steady stimuli, though responses to vibrato stimuli tended to be more scattered across subjects than were those for steady stimuli. There is no indication of the degree of scatter *within* subject's data.

The problem with this measure is that it assumes all subjects have the same internal reference parameters for the vowels. Also, the references are presumed to be constant with changing  $F_0$ . Furthermore, the relations between stimulus and response data are far from being obvious. Using the formula for calculating the "technical" Mel of Fant (1959), i.e.  $M = 1000 \log_2(1 + F/1000)$ ,<sup>13</sup> I have calculated

12. It seems a bit odd that formant frequencies should be expressed in Mels, the unit of a scale derived from the *pitch* of sinusoids.

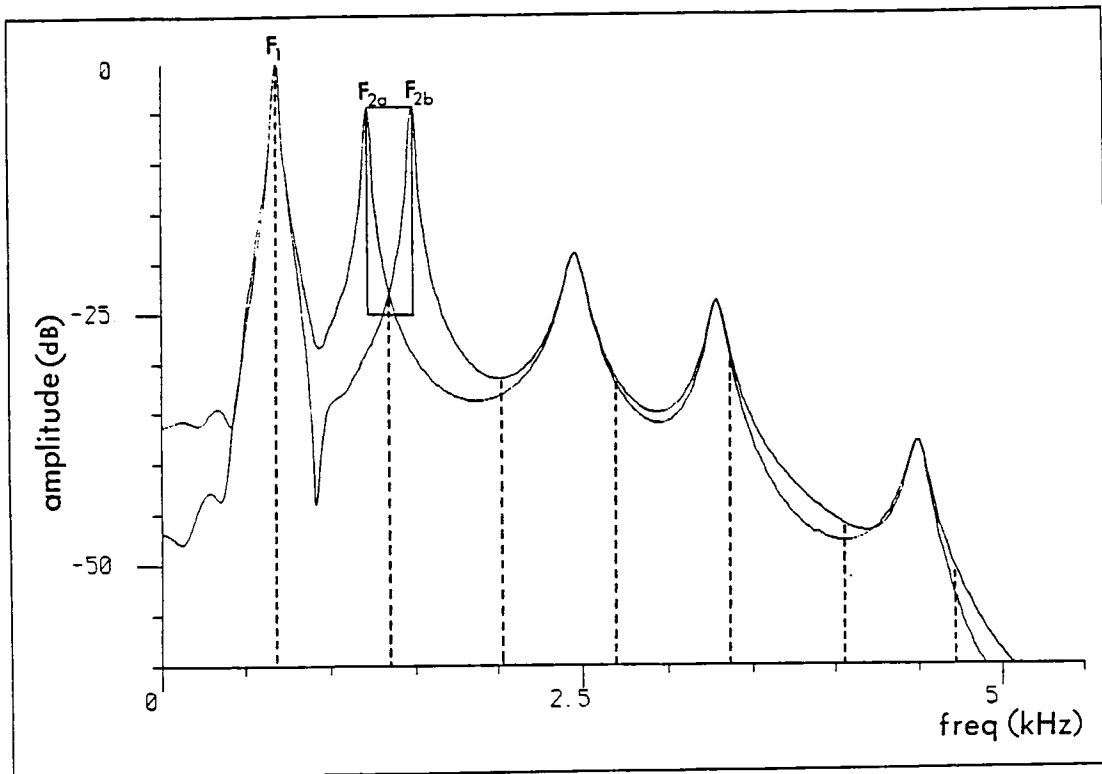
the formant frequencies for Sundberg's stimulus vowels and for the response vowels, and then calculated the statistic  $D$  for the "perceptual distance" between a given stimulus vowel and each of the 12 response vowels. For 4 of the 6 stimulus vowels, *at least one response vowel with a different name is closer*, i.e. has smaller  $D$ , *than is the response vowel with the same name*. For stimulus /u/, the response /o/ has smaller  $D$  than response /u/. For stimulus /e/, response vowels /y/, /ae/, /ε/, /ʌ/ and /ø/ had smaller  $D$  than response /u/. For stimulus /i/, response vowels /e/, /y/, /ae/, /ε/, /ʌ/, /ø/, and /oe/ had smaller  $D$  than response /i/. For stimulus /y/, response vowels /ʌ/ and /ø/ had smaller  $D$  than response /y/.

Therefore, I would conclude that these results are highly questionable on the grounds mentioned, particularly the inadequacy of the stimuli and the obscure relation between the stimulus parameters and the theoretical response parameters. As concerns the stimuli, Sundberg himself notes that "it may be argued that the stimuli used in our experiment are typical neither of singing, nor of speech, and hence the subject's reactions have little relevance to the practical situation." (p. 264) He then goes on to cite the work of Stumpf (1926) and to attempt a comparison of their respective data. He converted Stumpf's confusion data to scatter measures and then noted that data from his synthesized stimuli fell between the data for untrained and trained sopranos collected by Stumpf. One thing to note is that there is a large separation between the scatter of identification measures for trained and untrained sopranos' vowels. Trained sopranos' vowels have relatively small scatter (less ambiguity of identification) compared to untrained sopranos' vowels. This may be attributed to several factors, but we are reminded of the result of Bjørklund (1961) who demonstrated that vibrato in untrained sopranos is very small (almost nonexistent) while that in professional sopranos is quite audible and regular (often larger than the vibrato widths actually used by Sundberg). One is tempted to draw a conclusion here concerning the role of vibrato in decreasing ambiguity of vowel identity in Stumpf's data, in contradiction to Sundberg's claim. However, we have no way of knowing what the stimuli used by Stumpf actually were (since they were specious sound entities sung by living sopranos and not produced by electronic means).

---

13. See Fant (1971) for a discussion of the relevance for voice perception of this measure of a vowel spectrum.

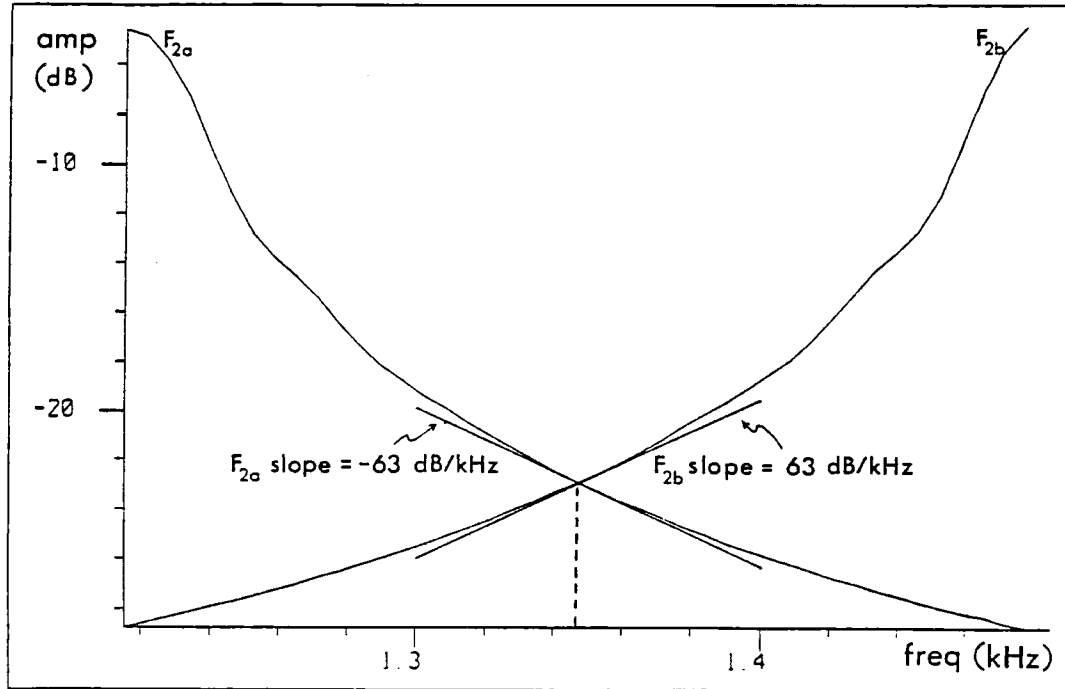
Rodet (1983) has demonstrated very clearly that vibrato and jitter can serve to reduce perceptual ambiguity concerning vowel identity. In Figure 1.5 are shown 2 spectral envelopes with 5 formant peaks each. The primary difference between the two is the location of the second formant. Rodet synthesized stimuli with a  $F_0$  (680 Hz) such that a harmonic fell exactly between the two  $F_2$  peaks. He synthesized, for



**Figure 1.5.** Two spectral envelopes used by Rodet (1983). The second harmonic falls at the intersection of the two possible  $F_2$ 's. For  $F_{2a}$  its amplitude slope with frequency modulation is negative. For  $F_{2b}$ , it is positive. Figure 1.6 shows an enlargement of the box.

each spectral form, one stimulus with vibrato and one steady stimulus. Subjects were unable to distinguish between the steady stimuli, but easily distinguished between the modulating stimuli when the modulation width was 2%. For these latter, the only difference is the sign of the amplitude slope of one partial (see Figure 1.6 for a close-up of the spectral envelope in the region of this partial). The perceptual effect with

modulation is one of a slight, but easily discernible change in vowel quality. This is rather strong supporting evidence for the notion that vibrato can play a role in reducing vowel ambiguity.



**Figure 1.6.** Enlargement of the region of the second harmonic's spectral slopes from Figure 1.5 (from Rodet, 1983).

The studies cited in this section may be summarized as follows:

1. Pitch extraction seems related to spectral fine-structure and gives the least equivocal perception in the presence of a harmonic series, though the regular periodicity of a harmonic signal may also play a role. These imply mechanisms of harmonic spectral template and periodicity detectors.
2. Timbre as tone color and vowel (and many aspects of consonant) perception is related to the spectral form of a signal, though aspects of vowel identity may be associated more closely with relations among formant peaks than to overall spectral form. These imply mechanisms for the extraction and enhancement



of characteristics of spectral form.

3. At higher pitches, jitter and vibrato may play a role in enhancing the information relating to spectral form and thus reduce ambiguity about vowel identity. It is unknown whether the same would hold for non-speech spectral forms. This implies a mechanism for the detection of either damping rate or of spectral slope detection in contribution to extraction of spectral form

Let us now move on to consider the cues known to be involved with simultaneous grouping processes. In Chapter 6 I will compare the differences between grouping processes and image quality extraction processes in light of the experimental data to be reported in Chapters 2 - 5.

### 1.7 Cues for Simultaneous Grouping

A reflection on the nature of the sources that are significant in our acoustic environment and of our relation to them has led me to consider the following set of cues that seem to contribute to the formation and separation of multiple, simultaneous source images. Certainly other cues may be involved to some extent, but I believe these are the most efficacious cues. Not all of these will be explicitly included in the experiments to follow, but they are described nonetheless for the sake of completeness and also to contribute to the picture of auditory organization to be developed in Chapter 6. They are:

1. (apparent) spatial location
2. harmonicity
3. separation of pitches
4. coherence of low-frequency frequency modulation
5. coherence of low-frequency amplitude modulation
6. stability and/or recognizability of spectral form when coupled with frequency modulation.

These will be considered briefly in turn.

### 1.7.1 (*Apparent*) *Spatial Location*

Sounds in the environment arrive at the two ears with small disparities of time of arrival, intensity and small spectral changes produced by the pinnae. These disparities are well correlated with the position in space the sound was emanating from. As one moves, or as the source moves, in the environment, these disparities change accordingly. In fact, we tend to be much more sensitive to changing conditions than steady ones and can localize sound sources better as a result.

In noisy, multi-source environments the fact that a source comes from a particular place can be used to attend selectively to that source, to the partial exclusion of information from other sources (Cherry, 1953). The improved detection of a signal in noise using two ears over one ear has been studied in classical psychoacoustics and is called the *masking level difference* (cf. Durlach, 1972; Jeffress, 1972).

To a certain extent, the time and intensity disparities can be adjusted in speakers and over headphones to change the apparent location of a sound object. But there are some rather narrow limits to the time differences that can be used and still result in a fused image. It is quite possible to create unusual effects under artificial conditions of headphone listening as is evidenced by many dichotic listening experiments, where organizational and localization paradoxes often arise (cf. Deutsch, 1975; Cutting, 1976).

In general, spatial localization processing is a global process that operates on (more or less) coherent information arriving at the two ears. These signals are fused into a single image and their disparities are translated into a spatial property of the source.

### 1.7.2 *Harmonicity of Spectral Content*

There is a great deal of psychoacoustic and physiological research which indicates that the auditory system is biased toward the processing of harmonic, as opposed to inharmonic sounds. Psychoacoustic pitch research reveals that the most unitary and unequivocal pitch sensation results from harmonic complexes (de Boer, 1976). As previously observed, many of the most relevant sources we deal with are harmonic, so it would not be surprising to find that the auditory system is biased

toward interpreting harmonic signals as representing single sources and that inharmonic signals might confound this interpretive mechanism in some predictable way. Two of the three pitch processing models currently in vogue (Goldstein, 1973; Terhardt, 1974) invoke a hypothetical *harmonic template-matching mechanism* assumed to operate somewhere in the central auditory nervous system. (Wightman's, 1973, model is based on an autocorrelation mechanism, which, nevertheless, gives very similar results in many cases.) The output corresponding to a given template would represent a given pitch and the magnitude of its output would correspond to the relative strength of the pitch sensation. The important property of such a template is that a harmonic signal best matches it and creates the least ambiguous pitch sensation. However, an inharmonic signal might partially match to several templates, thereby creating multiple pitch sensations of various strengths depending on the degree to which each match was made. There is some physiological evidence for the existence of such harmonic signal identifiers (Katsuki, 1961; Keidel, 1974).

The harmonicity of a signal is a strong factor in the perceived fusion of a tone complex. Harmonic tones fuse more readily than inharmonic tones under similar conditions and the degree to which inharmonic tones *do* fuse is partially dependent upon their spectral content. Different types of inharmonic signals have been experimented with concerning their effects on pitch perception, fusion and the perception of musical harmony (Cohen, 1980; Mathews & Pierce, 1980; Slaymaker, 1970). The interest in the types of inharmonicity used by these researchers is that they represent regular, predictable transformations of the harmonic spectral pattern and yet they do not exhibit the same property of having phase-locked partials as with harmonic partials. This points again to a certain uniqueness of the harmonic series and has implications for the parallel processing of spectral and temporal representations of the acoustic environment in the auditory nervous system. With harmonic signals, there would be a concurrence between a spectral pattern recognition mechanism and a temporal periodicity detection mechanism. But with inharmonic signals, these two kinds of processing mechanisms would provide disparate results to some sort of processor that tried to combine their respective outputs. This may be partly responsible for the equivocal response elicited by these sounds with respect to their perceived pitch and fusion. In general, as a sound is transformed to be less like the purely harmonic case, there is a decrease in the perceived fusion.

A special case of an inharmonic complex is the presence of multiple harmonic series. Under some conditions this results in the percept of a distinct number of pitches equal to the number of separate harmonic series (Cutting, 1976; Brokx & Nooteboom, 1982; Houtsma, 1983; Scheffers, 1983). Often times though, if this is the only cue for source separation, one hears a complicated tone mass without a discernible number of pitches (McAdams, 1982b; Houtsma, 1983). These effects seem related to factors of pitch separation, spectral overlap and harmonic coincidence. There is some evidence though that in the case of competing speech signals pitch differences (due to separately extracting the harmonic subgroups) can aid in source separation (Darwin, 1981; Bregman, Abramson & Darwin, 1983; Scheffers, 1983).

One would expect the processing of harmonicity as spectral pattern to be a global and central process (Houtsma & Goldstein, 1972; Goldstein, 1978). The processing of harmonicity as periodicity could operate at either a local level (extractable from auditory fibers being stimulated by several adjacent harmonics, and thus carrying the fundamental period in its temporal discharge pattern) or at a more global level if some mechanism were involved with detecting different auditory channels that were being stimulated by the same periodicity (Darwin, 1981; Scheffers, 1983).

### 1.7.3 *Pitch Separation*

This category is probably less a cue to source separation than a family of cases which limit source separation to different degrees. As one varies the pitch separation of two sources (and thus the separation of the fundamental frequencies for harmonic sources), one varies the degree of spectral overlap on a more global level since spectral forms of sources tend to change with pitch range (cf. Sundberg, 1975). One also varies the degree of harmonic coincidence that can be considered a kind of spectral interference at a more local level.

Some investigators *have* found simple effects of fundamental frequency separation. Stumpf (1890) claimed that two harmonic sounds tend to fuse when close in pitch and lose their characteristic qualities, although they may be perceptually separated and recognized individually when different in pitch. Scheffers (1983) found that source separability (judged by vowel identification) improved up to a 1 semit (6%) difference in  $F_0$ 's but did not improve beyond that. Brokx & Nooteboom (1982) found improvement (judged by errors in reproduction of vocal utterances) up to a 3 semit

(18%) difference in  $F_0$ 's. They proposed that the fusion of sources at close pitches explained their results: when competing speech messages are close in pitch, they more easily fuse; when the  $F_0$ 's are separated, the recognition process is not as inhibited and the listener can make a response.

#### 1.7.3.1 *Degree of Spectral Overlap*

This factor is less directly related to the perceived pitch, as noted. Scheffers (1983) found an appreciable effect of spectral overlap on the perception of simultaneous vowels. The greater the spectral overlap, the greater the possible masking and confusing of essential spectral features for vowel recognition. Houtsma (1983) investigated the ability of listeners to separate the virtual pitches (missing  $F_0$ 's) of two simultaneous 3-component harmonic tones. Performance on a musical interval recognition task was best when there was a total separation of the spectra.

#### 1.7.3.2 *Harmonic Coincidence*

This factor is the degree to which harmonics of different sources coincide in frequency. For perfectly steady sources, one would expect that perfect coincidence (identical pitch) would prohibit the detection of more than one source. One might also expect that harmonics that fall very close to one another would create beats and other kinds of temporal interference that would obscure their separation and allocation to separate sources (given this were an entirely spectrally-based process). However, Rasch (1978) found no major effect of coincidence on source separation (as measured in a masking experiment) except for some small phase effects when one  $F_0$  was very near a multiple of another  $F_0$ . Houtsma (1983), to the contrary, found that performance on his musical interval recognition task was good when there was harmonic coincidence and was bad when harmonics were close and interfering. These effects disappeared for dichotic listening indicating that the limits are peripheral spectral and, perhaps, temporal resolution limits on a local scale.

#### 1.7.4 Frequency Modulation Coherence

All natural, sustained-vibration sounds contain small-bandwidth random fluctuations in the frequencies of their components. These have been found for voice (Lieberman, 1961; Flanagan, 1972; Kersta *et.al.*, 1960; Rodet, 1982) and musical instruments (Cardozo & van Noorden, 1968; Grey & Moorer, 1977; MacIntyre, Schumacher & Woodhouse, 1981, 1982; see also Appendix B). There has not been much research directed toward determining the relative coherence of modulations among partials, perhaps due to the intractability of the analysis problem (though some developments in a new phase vocoder are underway by Dolson (1983). Some evidence that partial tone modulations are not perfectly correlated has been reported for voice (Björklund, 1961) and instruments (Grey & Moorer, 1977), but the precision of the analysis techniques in these studies may be questioned given that these analyses were not pitch synchronous. Conversely, there is evidence that with vibrato in violin tones, all of the harmonics more or less follow the same frequency variation such that the frequency excursions are proportional to the components' frequencies, i.e. harmonicity is maintained (Fletcher & Sanders, 1967). A theoretical consideration of the behavior of the frequency series of a forced-vibration sound, would lead us to believe that any perturbation of  $F_0$  would be imparted to all of its harmonics. Such a signal could be expressed (within the confines of Fourier-based thinking) as

$$S(t) = \sum_{n=1}^{\infty} A_n \sin (2\pi n F_0 t + n \int_0^t \text{Mod}(t') dt' + \varphi_n) \quad (1.3)$$

where  $n$  is the harmonic number,  $A_n$  is the amplitude of harmonic  $n$ ,  $\text{Mod}(t')$  is the modulating waveform representing the frequency perturbation, and  $\varphi_n$  is the starting phase of harmonic  $n$ .

As remarked before, the fact that the components move in parallel, maintaining constant frequency ratios, is an important cue in recognizing the behavior of many natural sources. It is worth noting that the initial mapping of the frequency spectrum into the auditory system via the basilar membrane roughly corresponds to a logarithmic scale. This means that constant ratios maintain constant distances along the basilar membrane. Further, it has been shown repeatedly that this "spatial" organization of the frequency domain in the auditory system is maintained (to some extent) as far as primary auditory cortex.<sup>14</sup> There is a marked regularity and richness of

connections in the anatomical organization of many of the higher processing centers in the central auditory nervous system. It is easy to imagine that there are mechanisms that would respond to a regular and coordinated pattern of activity distributed over an array of cells and fibers in this system as the neural information proceeds from the periphery and branches out to many of these centers.

There are several experimental results that support the notion that frequency modulation coherence contributes to perceptual fusion, aside from those mentioned in section 1.5 in connection with increased naturalness of an instrument or voice sound. Nordmark (1976) used a stimulus similar to the "pitch sweep" stimulus of Thurlow & Small (1955) where two pulse trains were presented slightly out of phase with respect to their pulse rates. This normally gives a pitch corresponding to the shorter period between consecutive pulses from separate trains. Nordmark presented each pulse train to a separate ear and coherently modulated the interpulse intervals of each train with a jitter function. This gives the longer pitch (associated with the lower period) and localizes toward the first pulse in the period giving the pitch. This indicates that the coherent modulation creates a bias toward a more globally-oriented pitch perception. He also showed that high frequency tones that cannot normally be lateralized with time differences *can* be lateralized when their frequencies are jittered. Blauert (1981) found similar qualitative results but at much higher jitter width thresholds than Nordmark. Somehow the fusion of the two tones due to the presence of the coherent jitter modulation allows their time disparity to be evaluated.

Charbonneau (1981) found, in resynthesizing musical instrument tones, that if the frequencies were kept constant, it was easy to discriminate this tone from the original. However, if all of the apparently slightly incoherent jitter functions on each of the harmonics were replaced with a single modulation function, this tone could not be discriminated from the original. It seems that the unmodulated tone is most likely more perceptually analyzable than the modulated versions which are perceived as being more fused. Also, this means that whatever minor inter-partial incoherences are present in the frequency movements in instrument tones are probably negligible.

- 
14. See for example: Evans (1975) - auditory nerve and cochlear nucleus; Guinan, Norris & Guinan (1972) - superior olivary complex; Roth, Aitken, Andersen & Merzenich (1978) - inferior colliculus; Aitkin & Webster (1971) - medial geniculate body; Merzenich, Knight & Roth (1975) - auditory cortex.

Brokx & Nootboom (1982) found better performance in vocal utterance reproduction in the presence of a competing speech stream for real monotone voices than for resynthesized voices with perfectly steady frequencies. They found no difference between real voices spoken either monotone or with normal intonation. I suspect that the natural jitter present in these voices may aid in fusing the image and in distinguishing it from the competing speech stream.

There is other evidence of the contribution of frequency modulation incoherence to source separation. Rasch (1978) presented two simultaneous harmonic complexes with different  $F_0$ 's. The level of the higher complex was adjusted to determine the threshold at which it was masked by the lower complex. When the higher complex had a 5 Hz, 4% vibrato imposed on it, its masked threshold was 17.5 dB *below* the threshold obtained when it was not modulated. The lower complex was never modulated. Helmholtz (1877/1885), as well, proposed that pitch movement, when not parallel, helps separate different sources in a polyphonic context.

I have determined in pilot studies (McAdams 1980, 1982b, App. F) that changes in the number of reported source images and in the noticeable pitches and timbres present in a complex tone result from using different modulation waveforms on separate sub-groups of components that are embedded in the complex spectrum if each modulation maintains the ratios between the components of its sub-group. This occurs for both harmonic and inharmonic stimuli. This is evidence that ratio-preserving FM may be one of those "circumstances which assist us first in separating the musical tones arising from different sources, and secondly, *in keeping together* [i.e. fusing into a single unified image] *the partial tones of each separate source*," [italics mine] (Helmholtz 1877/1885, p. 59).

#### 1.7.5 Amplitude Modulation Coherence

As amplitude modulation here I include all of the low-frequency modulations of amplitude we normally associate with the amplitude envelope of a sound such as attack and decay functions and various global fluctuations in the intensity of all components of a given sound source. Two aspects of amplitude behavior are important for source grouping decisions and in particular for perceptual fusion. These include onset synchrony of frequency components and global amplitude fluctuations across these components in sustained tones.



1.7.5.1 *Onset Synchrony*

The onset of a given source and the distinction of its onset from those of other sources is an important cue for the fusion of the components of that source, and for the parsing of separate source images. As Helmholtz (1877/1885) observed about fusion:

When a compound tone commences to sound, all its partial tone commence with the same comparative strength; when it swells, all of them generally swell uniformly; when it ceases, all cease simultaneously. Hence no opportunity is generally given for hearing them separately and independently. (p. 60)

Modern research would modify this statement, but there *are* characteristic onset patterns for individual instruments that maintain certain relations among the growth and decay patterns of individual partials. Here again we have some sort of "parallel" action of several components being used as a criterion for their belonging together and for their possibly arising from the same source. Cohen (1980) has shown that a common, synchronous exponential amplitude envelope can be used to successfully fuse tone complexes of *inharmonic* partials, as well.<sup>15</sup>

It would seem that if such a criterion were used, the asynchronous onsets of subsets of components, which belong to separate sources, might provide sufficient information to parse them appropriately. Let us turn to Helmholtz again, who states that

... when one musical tone is heard for some time before being joined by the second, and then the second continues after the first has ceased, the separation in sound is facilitated by the succession of time. We have already heard the first musical tone by itself, and hence know immediately that we have to deduct from the compound effect for the effect of this first tone. (p. 59)

---

15. An obvious conclusion within the framework of world modeling to be drawn from this result is that inharmonic sound sources are generally those that are plucked or struck. Both of these kinds of excitation generate more or less exponential amplitude envelopes.

It has been verified experimentally that an asynchrony in the onset of the partials in certain two- or three-component steady-state tones decreases the degree to which they fuse into a single image. In the Bregman & Pinker (1978) stimulus illustrated in Figure 1.1, asynchrony values between tones *B* and *C* of 0, 29 and 58 msec were used. One result of the study was that an increase in the asynchrony of the components in the complex tone was accompanied by a decrease in the tendency for those components to fuse. Dannenbring & Bregman (1978) measured the tendency for asynchronous components in three-tone complexes to segregate. Asynchronies of 0, 35 and 69 msec were used. Again, with greater asynchrony, more segregation and less fusion were perceived. Rasch (1978) measured masking thresholds of the higher component in asynchronous two-component tones and found less masking and greater ease in the perception of individual components with greater asynchrony. For an asynchrony of 30 msec, the threshold was as much as 40 dB lower than in the synchronous case. These low thresholds appeared to be independent of non-temporal features of the tones and were thus ascribed to asynchrony.

In a slightly different situation, Kubovy & Jordan (1979) demonstrated an effect that can be related to tone onset asynchrony. They presented a steady harmonic complex tone. At regular time intervals the phase of all components were instantly reset to 0 except for one partial whose phase was set to some other value. If this partial's phase was different from the rest by at least 30 degrees, that partial became audible as a separate pure tone. In their analysis of this phenomenon, Kubovy & Jordan proposed that the ear has a transfer characteristic that is compressive and non-linear. This kind of transfer characteristic can transform a phase disparity into a power disparity and what then results on successive comparisons of the segments with different dephased partials is a sudden increase in the power of one component, making it separately audible. We may consider that a sudden increase in power of one component is interpreted by the auditory system as the onset of another partial at that frequency.

Helmholtz (1877/1885) used a similar effect to train himself to hear out partials in a harmonic complex. By attenuating and then increasing the intensity of a given partial, he found that he could better direct his attention to it in the complex. But, as it remained unchanged, it again faded (fused) after awhile back into the complex. Here again, it is the increase in intensity that may be interpreted as an onset of that partial.

For more complex stimuli, Cutting (1976) investigated the effect of onset asynchrony on the perceptual fusion of various kinds of dichotic speech stimuli and found that with very small asynchronies (often less than 10 msec) separate formants or elements of a speech signal no longer fused to give the emergent quality of a particular consonant. Often, in fact, these stimuli lost their speech-like quality altogether when the temporal synchrony relations were not adjusted properly.

Grey & Moorer (1977) have found that there are small asynchronies in the onsets of different partials in musical instrument tones. This might seem to contradict the evidence above. However, these asynchronies are generally less than about 20 msec and there is often a significant amount of onset noise that might mask any minor differences in relative time of onset. It seems that variation in the relative onset times of individual harmonics within this small time period affects the perceived quality of the attack characteristic as a whole while the tone remains fused.

Tone onset asynchrony is a useful technique in musical practice for distinguishing certain "voices", and it is obvious that this cue is used with great versatility by many jazz and classical soloists. Rasch (1979) described how asynchronization allows for increased perception of individual voices in performed ensemble music, which also may be used in "multi-voiced" instruments such as guitar and piano. Across these studies, asynchrony values in the range of 30 - 70 msec have been found to be effective in source parsing.

#### 1.7.5.2 *Amplitude Fluctuations*

We can turn, as usual, to Helmholtz for some intuitions about the role of amplitude fluctuation in source separation:

The tones of bowed instruments are distinguished by their extreme mobility, but when either the player or the instrument is not unusually perfect they are interrupted by little, very short pauses, producing in the ear the sensation of scraping. . . . When, then, such instruments are sounded together there are generally points of time when one or the other is predominant, and it is consequently distinguished by the ear. (p. 59)

Some research efforts into the coherence of amplitude fluctuations have shown no effect on either separability or fusion (Scheffers, 1983). However, this study used rather broadband stimuli and may thus be violating the requirement that the fluctuations be relatively low-frequency in order to be trackable by the auditory system.

Békésy (1963) found that two pure tones of 750 and 800 Hz are usually heard separately when presented to the two ears. If they are sinusoidally amplitude modulated in phase at 8 Hz they fuse into a single sound image and are localized toward the higher frequency tone. Here is some evidence that an amplitude modulation coherence mechanism is global in nature and centrally located.

Moore (1982) suggests that the perceptual grouping of signal components in a noise background occurs when they are modulated in a coherent way. It is this coherent modulation that allows the components to be differentiated from the noise background which has temporal fluctuations unlike those of the signal.

In the reverse situation from this, Hall & Fernandes (1983) and Hall, Haggard & Fernandes (1983) have shown that there is a release from masking of a tonal signal by a noise masker when the masker is amplitude modulated at low modulation frequencies ( $< 100$  Hz) and is of a large enough bandwidth to cover at least 2 critical bandwidths. The implication in these results is that when there is coherent amplitude modulation in several auditory channels separated by more than a critical bandwidth, the auditory system can more easily separate what is signal and what is masker on the basis of the masker fluctuation in channels not containing the signal.<sup>16</sup> These authors also cite work by Darwin (1983) who has purportedly isolated common amplitude trajectory as an important factor for inclusion of a given harmonic partial in a speech pattern or otherwise. This conclusion is supported to a certain extent (though

16. These results can be compared with a negative result reported by Schubert & Nixon (1970), who found no discriminability between coherently or incoherently modulated sinusoidal carriers, where the modulating waveforms were narrow-band noise (75 Hz or 300 Hz). The lack of discriminability occurred for both sub- and supra-critical band spacing of the two carrier frequencies and for sub- and supra-critical band noise bandwidths. The difference between these studies are that Hall *et al* used a noise band carrier that was then amplitude modulated, but Schubert & Nixon used a pair of sinusoids. Also, Hall *et al* measured the masking threshold of a sine tone in the noise masker whereas Schubert & Nixon asked subjects to discriminate whether the two sine carriers were modulated by the same or independent noise modulators.

the effects reported are somewhat weak) by Bregman, Abramson & Darwin (1983).

The major implication of all of these studies is that amplitude coherence detection is a global process operating across several auditory channels. Both amplitude and frequency modulation coherence are temporal cues. As Helmholtz (1877/1885) has already stressed, simultaneous distinction processes have a temporal nature. This has been more recently echoed by Bregman (1982).

#### 1.7.6 *Resonance Structure Stability and Recognition of Spectral Form*

This factor has not been investigated specifically as a cue for grouping or fusion. The contribution of spectral form to aspects of timbre and phoneme (more specifically, vowel) perception were discussed above. In preliminary exploratory work (McAdams, 1982b; App. F) it seemed that tones with familiar spectral shapes, such as those from voices, tended to have a more fused or unified nature than unfamiliar laboratory shapes such as a flat spectrum where all components have equal amplitudes. If we assume that fusion implies the auditory system has decided that "this constitutes a reasonable source", then these results suggest that one of the criteria for fusion is that the spectral envelope be of a class of previously encountered, or at least plausible, spectral forms. It seems possible that the voice is a special case in this respect since the spectral shape is crucial to the identification of vowel sounds.

Huggins (1952, 1953) has suggested that the auditory system stores aspects of the structure of a physical source, which in the case of resonant sources, would be closely related to the behavior of the spectral form. Similarly, on the basis of surprisingly good identification of synthetic vowels in noise and in the presence of other vowels, Scheffers (1983) proposed (following suggestions by Klatt, 1980, 1982) that vowels are recognized by a kind of spectral template that "looks" at the frequencies of formant peaks only and not at the behavior of the spectral form in the regions between the peaks or at their relative levels. One implication of this would be that a spectral form that reasonably approximated such a template could stimulate the perception of a vowel quality. However, this also implies that in a situation with several similar and simultaneous vowels, the vowel recognizers might have problems separating out the relevant information for each vowel. In such a case, other cues would be necessary for a successful parsing of the vowels. Both Darwin (1981) and Scheffers (1983) have proposed that pitch (or harmonicity or periodicity) may be used in this case, as

mentioned in a previous section. According to Scheffers, the listener can apparently group those formants within which the harmonics belong to the same  $F_0$ , or decide that separate formants with harmonics not related to the same  $F_0$  belong to another vowel. There is also an implication here that the processes of vowel recognition and harmonicity or pitch detection are independent.

This independence with respect to source groupings is apparent, as well, in the data of Cutting (1976), as discussed above. In certain of his experiments a single phonemic identity as well as a multiple pitch was obtained with a given stimulus configuration. This may seem to contradict the conclusion of Darwin and Scheffers. However, they specifically state that the harmonicity effect comes into play as an aid in separating the speech signals *for recognition* only when the recognition processes cannot perform the task themselves. In Cutting's stimuli there may be two pitches but the separate pieces still all contribute to the same resulting (spectral form) interpretation. One possible conclusion is that the processes of spectral form perception and recognition are independent of source grouping processes.<sup>17</sup> If this is the case, one might challenge Bregman's notion that source qualities are derived from the properties of groups *after* the grouping processes have done their work.

The stimuli for which this seems most often to be the case are speech-like sounds. For other qualities such as tone color and pitch, the relation seems to be a dependent one (though one might argue, and some certainly have, that the dependence is of grouping on quality). The data of Bregman & Pinker (1978) cited above indicated that the complex tone composed of tones *B* and *C* was not perceived as being rich in timbre *unless the components were grouped and fused as a whole*, i.e. their concurrence was not enough to generate the rich timbre - they had also to be *considered as a group* before the timbre arose.

Similarly, in experiments on "profile analysis" (alias "spectral form perception") by Green & Kidd (1983), if the key part of the spectral form (whose change was to be detected across intervals in a 2IFC task) was placed in the opposite ear, subjects were unable to use the form as a whole to compare across the intervals. They easily performed this task when the stimuli were presented integrally to both ears. One might

---

17. It has been shown that spectral form processes related to timbre perception are relatively independent of the processes of perception of other qualities such as pitch (Plomp & Steenecken, 1971; Miller & Carterette, 1975).

interpret this as indicating that the separate parts are being localized differently and not considered as components of the same source and subjects are thus unable to judge differences on the complete form. In this case the task is no longer one of detection in change of spectral form but of simple detection in change of the level of a sinusoid in one ear.

One possible criticism against these studies is that their stimuli are so simple as not to engage normal auditory reactions since the behavior of the putative sources is far from being like those encountered in the "real" world. But some pilot studies that were performed (App. F) suggested that such effects can be obtained for both pitch and timbre. Let me describe a demonstration example. A harmonic tone with vowel /a/ quality was synthesized such that all of the harmonics initially had the same frequency modulation pattern (a combined vibrato and jitter). About halfway through the tone, the modulation pattern on the even harmonics was gradually changed to a pattern with an independent jitter and a vibrato of a different rate. At this point the even and the odd harmonics are modulating independently of one another but each sub-group is maintaining its own coherence of modulation.

The perceptual result with harmonic stimuli is striking. The initial percept is one of a singing vowel /a/ with vibrato and a distinct pitch. At a point approximately halfway into the stimulus, a "new" voice enters one octave above the original, also singing something not far removed from an /a/. This occurs regardless of whether the odd or even harmonics undergo the transition. This is an intriguing and seemingly paradoxical percept. A new source image is formed whose pitch and timbre derive from the even harmonic subset; however, *the timbre of the odd subset is unaffected* though one would have expected it to acquire the more "hollow" timbre normally associated with spectra with only odd harmonics. It continues unperturbed while a "new" voice "joins" it at the octave. Note that no new components have been added. The existing ones are merely parsed differently due to independent modulation functions superimposed on the separate spectral subsets. So even though half of its harmonics are parsed into a separate source and assigned a pitch based on the spectral subset composed of the even harmonics, the contribution of those harmonics is not subtracted from the timbre of the original tone.

An example similar to this was created by Roger Reynolds and Thierry Lancino at IRCAM for Reynolds' composition *Archipelago*. They used an oboe tone; however the even and odd harmonics were sent to separate speakers. Initially one hears an oboe sound centered between the speakers when the modulations are identical. As they become independent, the oboe image splits into two images of a soprano at the octave in one speaker and a clarinet-like sound at the original pitch in the speaker with the odd partials. Here, the modulation coherence initially overrides the spatial separation of the harmonics. As the modulations separate, the images move to the locations from which their respective spectra are emanating. In this case, the odd harmonics have a timbre that corresponds more closely to the actual spectrum of the source. Note that the previous example was monophonic. So when there was a separation of the modulations, there was not spatial movement of the images. It seems entirely possible that the auditory system interprets the situation as the arrival of a new voice at the octave which is then "sitting on top of" (and thus masking) the even harmonics of the original source image. If they are being masked, then they are really still there and according to this interpretation (or world model) the timbre should remain the same.

In these cases, the frequency modulation coherence was the strongest organizing factor. And the perceived qualities depended on *how* the spectrum was parsed into subgroups as well as *what* the auditory system believed the world was doing. This all fits within the framework described in the earlier sections. But these results appear to stand in contradistinction to those of Cutting, primarily because he did not report the identity, pitch and location of all perceived sources. We cannot therefore know the extent to which they were actually independent. Obviously, there are questions that need to be addressed.

### 1.8 Problems to be Addressed

Several issues have been raised in this introduction that need to be addressed experimentally and theoretically. Not all of them can be addressed here; some are many years away from being clarified enough to ask the right questions. However, a start can be made.



Concerning cues that contribute to simultaneous source image formation and separation, I will consider harmonicity, frequency modulation coherence and spectral form. Theoretically, more consideration needs to be given to the nature of local and global mechanisms involved in source image organization and the extraction of perceived qualities, as well as to the problem of the relation of grouping processes to quality derivation processes.

**Chapter 2** will address issues of frequency modulation coherence and harmonicity. It has been proposed that coherent frequency modulation maintains constant frequency ratios. One property of this kind of correlation among frequency motions of partials is that all motions are in the same direction for each partial. It seems possible that if *constant frequency differences* among partials were maintained, which *also* yield similar directions of motion of the partials, that this might also aid fusion if it were only directional "common fate" (Köhler, 1929) that determined grouping. This would represent a kind of dynamic version of the classical "pitch shift" stimulus (cf. de Boer, 1976). The difference between *frequency-difference-preserving modulation* and a *frequency-ratio-preserving modulation* is that the latter maintains constant distance between partials on a log frequency scale, while the former maintains constant distance on a linear frequency scale. Also, the latter maintains harmonicity and the shape of the signal waveform within a single period for harmonic stimuli while the former moves in and out of harmonicity deforming the signal waveform.

From the available physiological evidence, which illustrates that the basilar membrane is organized roughly according to a log frequency scale, we might suspect that this scale has some special properties with respect to processing by the auditory nervous system. Certainly perception of constant pitch intervals is related to this scale. As concerns fusion, Bregman, McAdams & Halpern (1978) have shown that constant difference modulation yields tone complexes which are much less fused than those with constant ratio modulation when the modulation form is an exponential frequency glide. Experiments 1-5 will examine these relations for vibrato and jitter modulation in harmonic tones with various types of spectral envelopes.

**Chapter 3** will address the contribution of frequency modulation incoherence and harmonicity to the distinction of multiple sound sources, and will investigate the nature of local and global mechanisms involved in this process.

If we consider the behavior of frequency components in sustained-tone forced-vibration systems (e.g. wind and bowed string instruments and voice), we find that there is a strong correlation in the random and/or periodic frequency modulation patterns among the components. Essentially, perturbations of the fundamental frequency are imparted proportionally to its harmonics.<sup>18</sup> One expects that these perturbations would be independent from one sound source to the next. It seems plausible that the auditory system may use two facets of this kind of information to form images of sound sources and to distinguish concurrent sources.

1. Since the FM on harmonics of a single source are relatively *coherent* (i.e. partials vary in frequency such as to maintain, more or less, their harmonic ratios), some mechanism may exist which is capable of grouping together partials that vary similarly in frequency.
2. And since the FM on partials of different sources are independent and thus *incoherent*, some mechanism may operate to signal the presence of multiple sources by detecting incoherent modulation on different spectral components.

Schubert & Nixon (1970) suggested that

... in our immediate classification of the sounds of continuous speech, in our easy identification of any one of a large number of familiar talkers even over narrow-band (telephone) transmission systems, in the recognition of fine temporal nuance in musical performance, and particularly in our ability to separate simultaneously-present, broad-band sound sources, such as the instruments of an ensemble or competing talkers, there is convincing evidence that the system must either include an analyzer for direct coding of the original broad-band waveform *or must routinely coordinate internally-derived temporal patterns from*

---

18. That the modulations on the several harmonics are not perfectly correlated is implied in the data of Bjørklund (1961) for voice and Grey & Moorer (1977) for musical instruments. However, Charbonneau (1981) has demonstrated that if the small degree of incoherence among the FM patterns of the harmonics in resynthesized instrument sounds is removed and replaced with a perfectly coherent modulation, most subjects are unable to detect a difference. This indicates that the amount of incoherence present in these sounds is perceptually negligible.

*different spectral locations in the cochlea.* [my emphasis]

... in general, for a sufficiently diverse analysis of the incoming waveforms, the most versatile analyzer "reading" the cochlear output would be one comprising the critical-band channel, and its subsequent spectrally-oriented analyzers plus a "straight-through" channel primarily concerned with preserving all the timing information that survives the comparatively broad mechanical filtering. (p. 1)

Of course there is much evidence against the existence of any "straight-through" channel, as Schubert & Nixon remark, while the effects of some kind of band-pass "auditory filter" are ubiquitous in neurophysiological and psychoacoustic results. Most of these effects point to critical band filtering with bandwidths that are quite a bit narrower than the measured bandwidth of mechanical filtering in the cochlea. But there is both physiological (Brugge, Anderson, Hind & Rose, 1969) and psychoacoustic (Plomp, 1966, 1976) evidence that frequencies at distances much greater than the critical bandwidth can create patterns of stimulation that interfere with one another in the cochlea. This interference induces a more complex temporal response in the regions of interference than would be obtained from stimulation by a single sinusoidal signal. Nevertheless, the further the separation between the frequency components, the less the degree of interaction in the cochlea, until, eventually, any interference that is occurring is either masked or is negligible with respect to the more forceful stimulation near the peaks of excitation.

Given the interaction of stimulation by different spectral components within frequency-specific auditory nerve fibers *and* the limited extent to which such interactions can take place at greater frequency differences, it seems reasonable to postulate two types of mechanism involved with extracting information about source behavior on the basis of frequency modulation coherence. One mechanism would operate on the regularity or change in behavior of the temporal pattern of nerve firing *within a given auditory channel*. Another mechanism would make comparisons of temporal behavior *across auditory channels*.<sup>19</sup> Drawing again from Schubert &

19. Something similar to this classification was proposed by Goldstein (1966). He proposed a "place-intensity" perception based on within channel timing information and "place-synchrony" perception based on cross-channel timing information.

Nixon with respect to this latter notion, "this facet of auditory analysis has come in for very little specific discussion in the history of auditory perception, possibly because of our preoccupation with problems of frequency resolution *rather than resynthesis*; . . . " Certainly both notions of synthesis and analysis are important for the formation and distinction of auditory source images. Experiment 6 will examine these questions.

**Chapter 4** will address the notion of spectral envelope stability and its contribution to the perception of a fused auditory image. If one were to impose a frequency modulation on the components of a complex tone, and the initial amplitude relations among the partials were maintained instead of following the spectral envelope, we would expect that this tone would acquire an unstable identity at larger modulation widths. This, of course, would be more true for complex spectral envelopes than for very simple ones since the spectral deformations would be greater and probably more audible. This unlikely movement of formants may cause them to be parsed, or separated perceptually, from the rest of the tone, if our perception is more oriented toward formants than to overall spectral form as suggested by Sapozhkov (1973). Experiment 7 will test the hypothesis that a stable spectral envelope contributes to unified auditory source image perception and that a more complex spectral envelope is more sensitive to spectral envelope stability than is a simpler one.

**Chapter 5** will address the perception of multiple vowel sounds in order to discern the relative contributions of harmonicity, global spectral overlap, frequency modulation coherence and stable and recognizable spectral forms to source image separation and identification. Few of the multiple vowel perception studies reported to date have included the frequency modulation aspect (except Brokx & Nootboom, 1982). The sources used in this study will be sung vowels rather than spoken ones. In creating a complex situation for the auditory system, the hope is to find the limits and tendencies of the different cues that might possibly be contributing to multiple sound source perception. Also, as is often the case in the "normal" world, the task will be one of discerning the presence of a known source amongst other competing sources. Another aim is that Experiment 8 will shed some light on some of the apparent discrepancies with respect to the relation between source grouping and quality perception processes.

Finally, in **Chapter 6** these data will be evaluated in terms of the framework presented in the earlier sections and integrated with some thoughts about the implications of these processes for an understanding of the perception of complex musical structures.

## CHAPTER 2

### Harmonicity-preserving Frequency Modulation and Spectral Fusion

#### 2.1 Introduction

It was proposed in the previous chapter that *coherence* of sub-audio frequency modulation among partials arising from the same sound source is an important cue contributing to the perceptual grouping of those components. This chapter will extend the work done by Bregman, McAdams & Halpern (1978) who used frequency glides. Those studies demonstrated that harmonic complex tones which had frequency sweeps applied to the components were perceived as more fused when the harmonic ratios were maintained than when the frequency differences between the components were maintained.

The purpose of the following experiments is to compare these two types of modulation for the differences in perceived multiplicity of source images they engender under varying conditions of spectral envelope shape and amount of modulation for periodic and aperiodic modulating functions. It is hypothesized that modulation not maintaining harmonicity will be perceived as yielding more sources (or as being less fused) than modulation that does maintain harmonicity. The first experiment uses a two-interval forced-choice (2IFC) task where subjects are to choose which modulation type has more sources or elements present, i.e. which tone is more dispersed or less fused. This allows the construction of a curve relating perceived source multiplicity (and by implication perceived fusion) to rms deviation of the modulation for each spectral envelope shape with each modulation waveform. The second experiment uses a multi-dimensional scaling (MDS) procedure where subjects rate the relative dissimilarity in fusedness or multiplicity between all pairs of sounds in the stimulus

set. This potentially allows a comparison of the fusion of all stimuli to see if there are relative effects of rms deviation, modulation type and spectral envelope that would not show up with the other procedure. Experiments 3 - 5 were designed to investigate a possible confounding effect in the first two experiments.

## 2.2 EXPERIMENT 1: Effects of sub-audio frequency modulation maintaining constant frequency differences and constant frequency ratios on perceived source image multiplicity.

### 2.2.1 Stimuli

All tones were synthesized with 16 harmonics of a 220 Hz  $F_0$ . Each tone was 1.5 sec in duration with 100 msec raised cosine ramps. Three spectral envelopes were used: flat, -6 dB/oct, and vowel /a/. These were imposed on the complex tones such that the amplitude of any frequency component traced the spectral envelope when being modulated in frequency. These envelopes were stored as table-lookup transfer functions and addressed with the instantaneous frequency of the partial at each sampling interval. Implementation of this synthesis procedure is described in Appendix A. Two modulation waveforms were used: periodic (vibrato) and aperiodic (jitter). The vibrato was a 6.5 Hz sinusoidal signal. The synthesis and characterization of the fixed jitter waveform,  $J_1(t)$ , are described in detail in Appendix B. This waveform has a predominantly low-frequency spectral content with two frequency bands. The higher band (30 - 150 Hz) is approximately 40 - 45 dB lower in amplitude than the lower band (0 - 30 Hz). Any components greater than 150 Hz are more than 80 dB below the 30 Hz band. Also, the amplitude distribution of this waveform is symmetrical about 0 over the duration of the signal, 1.5 sec. Five values of rms deviation<sup>1</sup> of the modulation were used. These values, expressed as both cents (1cent = 1/100<sup>th</sup> of a semitone) and as  $\Delta f / \bar{f}$ , are listed in Table 2.1. The relation between the cents measure and  $\Delta f / \bar{f}$  is expressed

---

1. An rms measure of modulation excursion was used instead of a peak measure since Klein & Hartmann (1979) found this measure better at relating vibrato width perception across various modulation waveforms. One interest in the present study is to compare a periodic and aperiodic modulation and the measure used would significantly affect the comparison.

**TABLE 2.1.** Rms deviation of frequency modulation used in Experiments 1 and 2. Values are expressed both in cents and as  $\Delta f / \bar{f}$ .

cents	7	14	28	42	56
$\frac{\Delta f}{\bar{f}}$	0.00405	0.00812	0.01630	0.02456	0.03288

$$\frac{\Delta f}{\bar{f}} = 2^{\frac{\text{cents}}{1200}} - 1. \quad (2.1)$$

Finally, the two modulation types were used: modulation maintaining constant frequency ratios, and thus harmonicity (*CR*), and modulation maintaining constant frequency differences (*CD*) among the 16 partials. The resulting signals are described as follows:

$$S_{CR}(t) = \sum_{n=1}^{16} A(f_{ni}) \sin(2\pi n F_0 t + n \psi \int_0^t \text{Mod}(t') dt'), \quad (2.2)$$

and

$$S_{CD}(t) = \sum_{n=1}^{16} A(f_{ni}) \sin(2\pi n F_0 t + k \psi \int_0^t \text{Mod}(t') dt'), \quad (2.3)$$

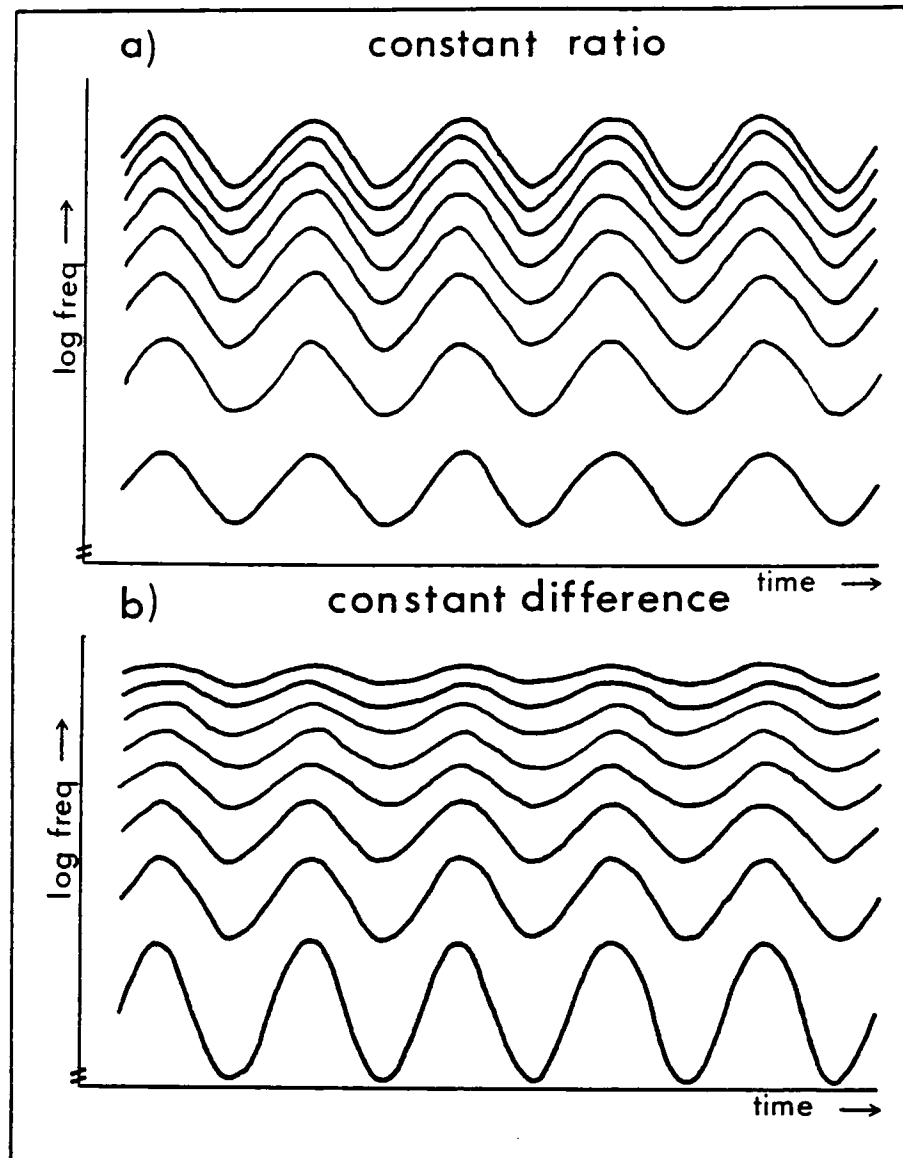
where,  $A(f_{ni})$  is the instantaneous amplitude of partial  $n$  dependent on the partial's instantaneous frequency,  $f_{ni}$ ;  $k$  is the constant rms frequency deviation factor for *CD* tones;  $\psi$  is the desired rms deviation,  $D_{rms}$ , divided by the actual rms deviation,  $A_{rms}$ , of  $\text{Mod}(t)$  (see Eq. A.4, App.A). This latter parameter represents the proportion of rms deviation from the partial's center frequency, e.g. for a sinusoidal vibrato with a peak amplitude of 1,  $A_{rms} = 0.707$ . If, for example, an rms deviation of 5 cents is desired, and  $A_{rms}$  is the rms amplitude of  $\text{Mod}(t)$ , then

$$\frac{\Delta f_{rms}}{\bar{f}} = 2^{\frac{\text{cents}_{rms}}{1200}} - 1 = 0.00289, \quad (2.4)$$

and

$$D_{rms} = \frac{\Delta f_{rms}}{\bar{f}} \cdot \frac{1}{A_{rms}}. \quad (2.5)$$





**Figure 2.1.** Exaggerated spectrographic diagram of *CR* and *CD* modulations plotted on a log frequency scale for the first 8 harmonics.

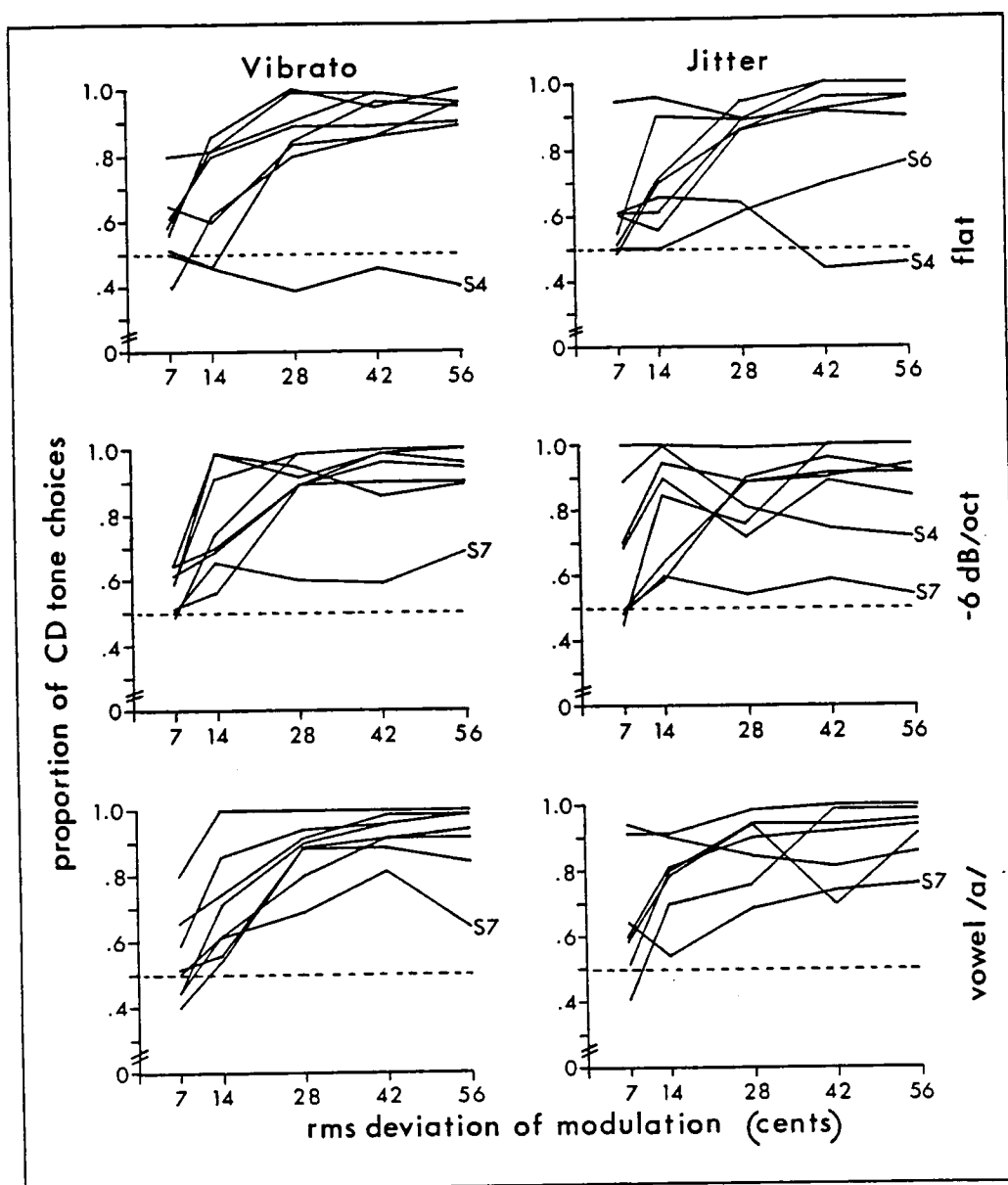
Dividing by  $A_{rms}$  normalizes the rms amplitude of the modulating waveform to 1, which is then scaled by  $\Delta f_{rms} / \bar{f}$ . Note that the second term within the sin function (Eq. 2.2) yields values that modulate around  $f_n$  by an amount that is proportional to  $f_n$ . This assures the maintenance of frequency ratios.

For *CR* tones, the rms deviation for each partial,  $f_n$ , is  $\Delta f_n = f_n \psi$  for all  $n$ . For *CD* tones,  $\Delta f_n = F_0 k \psi$ . Here,  $\Delta f_n$  is independent of  $f_n$  and proportional to  $F_0$ . The value  $k$  represents the harmonic number (not necessarily integer) that would have the same rms deviation in both *CR* and *CD* tones. For example, with  $k = 1$  and  $D_{rms} = 14$ cents, the rms deviations on the fundamentals are equal, but the maximum frequency excursion for the 16<sup>th</sup> harmonic is 20.2 Hz for *CR* and only 1.3 Hz for *CD*. Note that the only times when  $CD(t)$  is strictly harmonic are those instances when  $Mod(t) = 0$ . A spectrographic diagram illustrating the effect of these two types of modulation on a log frequency scale is shown in Figure 2.1.

The value of  $k$  has a strong effect on the perceived modulation width of the *CD* tones. An attempt was made to equalize as much as possible the perceived modulation widths and loudnesses for the entire stimulus set. The matching studies are reported in Appendix C. From the modulation width matchings, a  $k$  value of 1.9 was chosen for Experiments 1 and 2. After loudness matching the stimuli were presented over headphones at approximately 75 dbA in a sound treated room (see Appendix A).

### 2.2.2 Method

In each trial, one *CR* tone and one *CD* tone were presented in succession and in counterbalanced order. The observation intervals were marked by differently colored lights on a 2-button response box. They were separated by a 500 msec silence. Both tones had the same spectral envelope, the same modulating waveform and the same value of  $\psi$  (eqs. 2.2 and 2.3) in any given trial. The subject was instructed to choose which of the two tones seemed to have more sources in it, potentially derived from more sources, or was perceptually more analyzable into separate sounds, i.e. split apart into two or more distinguishable elements. The choice was to be indicated by pressing the appropriate button on the 2-button box. Once the subject responded there was an additional 500 msec silence before the presentation of the next trial. Since one of the most prominent perceptual effects at larger rms deviations for *CD* tones is the apparent separation or independence of the fundamental from the rest of the complex, subjects were advised to focus their attention in the region of the lowest pitch. Experimental instructions were presented in either English (5 Ss) or French (6 Ss) as the subject desired.



**Figure 2.2.** Experiment 1 data summary. Each graph shows the proportion of times the constant-difference tone was chosen as having more sources than the constant-ratio tone as a function of the rms deviation (expressed in cents). Within each graph a separate function is shown for each subject. The graphs on the right are for vibrato stimuli and those on the left are for jitter. The three spectral envelopes are ordered from the top: flat, -6dB/oct and vowel /a/, respectively.

Stimuli were blocked according to modulation waveform. Each run consisted of one such block with 150 comparisons: (3 spectral envelopes)  $\times$  (5 rms deviations)  $\times$  (10 repetitions of each pair). Five blocks of vibrato stimuli and 5 blocks of jitter stimuli were presented to each subject in counterbalanced order. In all, 50 repetitions of each stimulus pair were presented. Each experimental session consisted of 2 - 4 runs. Data consisted of the proportion of times the *CD* tone was chosen as having more sources. The greater the effect of the modulation type, the higher this value would be.

Eleven subjects participated in the experiment and were paid for their time. None reported having any serious hearing problems. The data for two were thrown out since they appeared completely random after 4 runs. Upon questioning, these subjects reported that they could not discern any difference in multiplicity between the two sounds in a pair. One subject did not finish the experiment, so his data are not included either. Therefore, complete data were collected for 8 subjects.

### 2.2.3 Results

The data for 8 subjects and the means and unbiased standard deviations across Ss are listed in Table E.1 (Appendix E). These data are plotted in Figure 2.2 as a function of rms deviation for each spectral-envelope/modulation-waveform combination. From these plots one is able to see the relation of the curves among Ss.

There were no significant differences between spectral envelope conditions when the means for a given rms deviation and modulation waveform were compared among the envelopes (two-tailed *t*-tests between flat and -6 dB/oct, flat and vowel, -6 dB/oct and vowel). I would conclude from this that the spectral envelope did not differentially affect judgments of source multiplicity when the tones being compared had the same spectral envelope. Certain apparently aberrant functions are marked with the subject number. Note that most of the variance is due (unsystematically across conditions) to 2 subjects: S4 and S7. In Table 2.2 the means across a) all subjects, and b) all subjects except S4 and S7 are listed.

**TABLE 2.2.** Data summary for Experiment 1. Each cell value is the mean across Ss of the proportion of choices of *CD* tones. The modulation width factor  $\psi$  was constant within a given comparison. The value in parentheses is the unbiased standard deviation for a) means across all Ss, b) means for Ss 1,2,3,5,6,8. At the bottom of each table are listed the means and pooled standard deviations for each rms deviation across spectral envelope and modulation waveform conditions.

a) all subjects ( $N = 8$ )						
Modulation Waveform	Spectral Envelope	Rms Deviation of Modulation ( cents )				
		7	14	28	42	56
vibrato	flat	.58 (.12)	.68 (.17)	.83 (.19)	.87 (.17)	.88 (.20)
	-6 dB/oct	.57 (.06)	.78 (.16)	.88 (.12)	.91 (.14)	.92 (.11)
	vowel /a/	.54 (.13)	.71 (.16)	.88 (.10)	.93 (.06)	.91 (.12)
jitter	flat	.60 (.15)	.70 (.16)	.82 (.12)	.87 (.20)	.88 (.19)
	-6 dB/oct	.65 (.21)	.81 (.18)	.81 (.14)	.87 (.14)	.86 (.16)
	vowel /a/	.67 (.18)	.77 (.12)	.85 (.11)	.87 (.11)	.92 (.08)
<b>overall mean</b>	( $N = 48$ )	.60 (.15)	.74 (.16)	.84 (.13)	.89 (.14)	.89 (.15)
b) without S4, S7 ( $N = 6$ )						
Modulation Waveform	Spectral Envelope	Rms Deviation of Modulation ( cents )				
		7	14	28	42	56
vibrato	flat	.59 (.13)	.69 (.15)	.87 (.06)	.92 (.06)	.93 (.03)
	-6 dB/oct	.57 (.06)	.76 (.16)	.92 (.05)	.97 (.04)	.97 (.04)
	vowel /a/	.57 (.14)	.75 (.16)	.91 (.07)	.96 (.03)	.97 (.03)
jitter	flat	.60 (.17)	.72 (.18)	.84 (.11)	.93 (.12)	.94 (.09)
	-6 dB/oct	.64 (.21)	.82 (.17)	.86 (.10)	.94 (.05)	.94 (.06)
	vowel /a/	.63 (.17)	.78 (.09)	.88 (.10)	.90 (.11)	.96 (.03)
<b>overall mean</b>	( $N = 36$ )	.60 (.15)	.75 (.15)	.88 (.08)	.94 (.08)	.95 (.05)

Likewise, there were no significant differences between modulation waveforms within rms deviation and spectral envelope (two-tailed *t*-tests between vibrato and jitter). I would conclude from this that neither did the modulation waveform

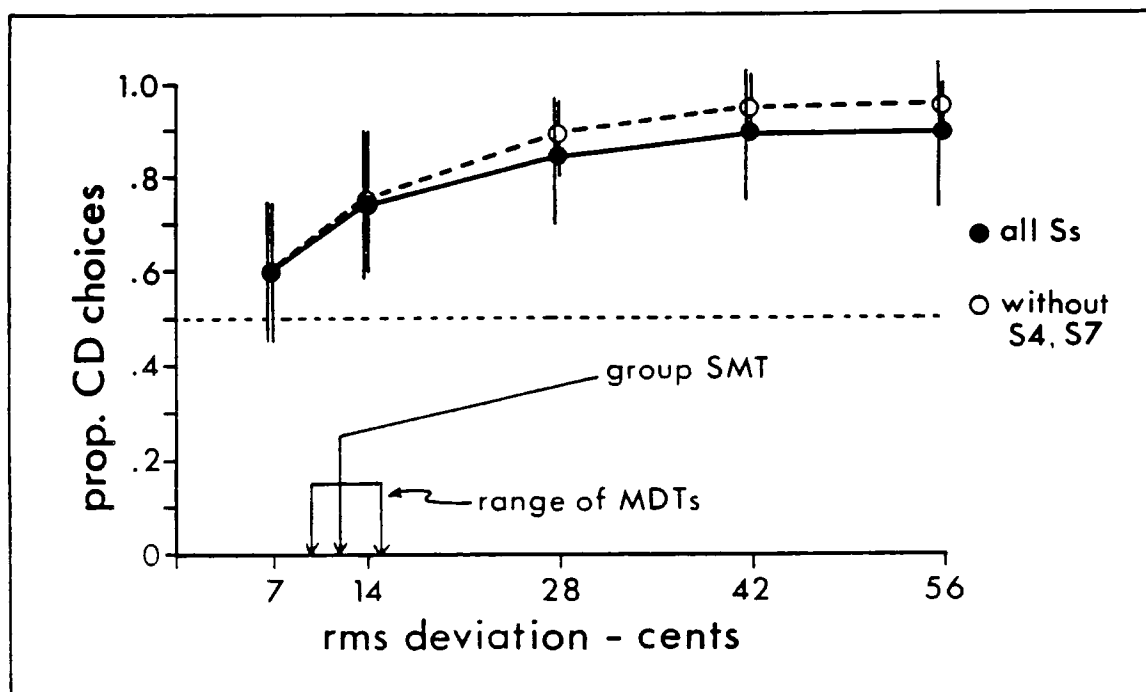
differentially affect judgments of source multiplicity.

The lack of effect of spectral envelope and modulation waveform is with respect to the data averaged across Ss. There was obviously a rather large effect of these stimulus parameters for Ss 4 and 7. For S7, there was a small effect of modulation waveform (vibrato tended to achieve higher values than jitter), but there was a large systematic effect of spectral envelope with the flat envelope attaining the highest values (achieved  $> .90$  at larger modulation widths), followed by the vowel envelope (achieved  $> .75$ ) and then by the  $-6$  dB/oct envelope which never achieved  $> .70$  even at 56 cents deviation. This was the only subject for which there *were* systematic differences for these parameters. The curves for this subject's data were, however, similar to most of the rest of the subjects' curves in being generally monotone increasing with rms deviation. For S4, there were no systematic effects of either parameter, and the data curves are rarely monotonic. Vibrato yields higher values than jitter for most of the  $-6$  dB/oct conditions, but yields lower values for flat and vowel envelopes. For vibrato the  $-6$  dB/oct envelope yields the highest values followed by the vowel and then the flat envelope. With jitter,  $-6$  dB/oct and vowel envelopes yield similar values which are higher than those for the flat envelope. For this subject, the data for the flat envelope are essentially at chance. My guess is that this subject never settled on a criterion for evaluating the source multiplicity.

Aside from these two subjects, it appears that the effects of spectral envelope and modulation waveform are inconsequential with respect to judgments of source multiplicity differences between *CR* and *CD* tones. Accordingly, the overall means<sup>2</sup> within rms deviation across Ss, spectral envelope and modulation waveform conditions are listed in Table 2.2. These values are plotted as a function of rms deviation in Figure 2.3. The effect of removing S4 and S7 is only to increase the slope of the function without perturbing its overall form. From this graph the main result of this experiment may be gleaned: as the rms deviation increases from 7 to 56 cents, *CD* tones are chosen progressively more often as having more sources or more distinguishable sound elements.

---

2. All overall means are significantly different from chance choice at least at the .01 level. This means that even at 7 cents modulation width, the difference in source multiplicity between *CR* and *CD* tones is discernible.



**Figure 2.3.** Experiment 1 data summary. The proportion of times the *CD* tone was chosen as yielding more sources is plotted as a function of the rms deviation of modulation. Data points are averaged over Ss, spectral envelope and modulation waveform. The arrows on the ordinate indicate the group SMT and the range of MDTs found by other investigators for low-frequency sinusoidal carriers (see footnote 3 this chapter). Closed circles ( $N = 48$ ) represent the means across all Ss; open circles ( $N = 36$ ) represent the means across all Ss except S4 and S7.

#### 2.2.4 Discussion

If one accepts that perceived fusion is inversely related to the perceived multiplicity, then the hypothesis of this experiment has been confirmed, at least with respect to the stimuli used here. Namely, predominantly sub-audio frequency modulations maintaining constant frequency ratios (strict harmonicity, here) are perceived as being more fused than is the case with a modulation maintaining constant frequency differences. This holds even when the frequency movement of all the partials is moving in the same direction at all times, harmonicity is being violated. This

evidence for both periodic and random modulations supports and generalizes the findings of Bregman *et. al.* (1978) for frequency sweeps.

It is important to note here, though, that the strength of the effect depends on the rms deviation of the modulation. In Bregman *et. al.*, the glides were on the order of one or two octaves (1070 cents and 2288 cents). In this study, the point at which *CD* tones were chosen at least 71% of the time occurs somewhere between rms deviations of 7 cents and 14 cents (at approximately 12 cents if a cubic spline is fitted to the mean data points). For deviations below 12 cents the difference with respect to source multiplicity is not as evident.<sup>3</sup> For deviations greater than about 40 cents the source multiplicity difference is perfectly clear for most Ss.

Analyses of the natural jitter in instrument tones (see Appendix B) show the rms deviations to be on the order of 7 - 12 cents (for flute, clarinet and trombone) and those of the singing voice are on the order of 7 - 27 cents (unpublished data of X. Rodet, 1982). Fletcher, Blackham & Geersten (1965) reported vibrato widths of 24 - 52 cents for violin tones. (These are most likely peak widths and would correspond approximately to 17 - 37cents<sub>rms</sub>). Seashore (1936, 1938) reported peak vibrato widths in singers varying between 31 cents and 98 cents (approximately 22 - 69cents<sub>rms</sub>). The range of these values includes the range of deviations used in this study.

Some mention should be made of the similarity of the effects of vibrato and jitter. No difference between these two waveforms was observed as far as their effects on source multiplicity judgments. If the measure of frequency deviation had been in terms of peak deviation instead of rms deviation, a difference would have been observed. For the modulating waveforms in this experiment the following relations between rms and peak deviation hold:<sup>4</sup>

3. Frequency modulation detection thresholds for these complex tones with vibrato and jitter were found to be on the order of 2 - 5 cents for vibrato and 1.5 - 5 cents for jitter (see Appendix D). So all rms deviation values used in this experiment are presumed to be above modulation detection threshold. For a 250 Hz sinusoidal carrier, frequency modulation detection thresholds (MDTs) have been found to be on the order of 11 - 15 cents<sub>rms</sub> (Shower & Biddulph, 1931 for  $f_m = 3$  Hz; Groen & Versteegh, 1957, for  $f_m = 4$  Hz; Jesteadt & Sims, 1975, for  $f_m = 8$  Hz). Jitter detection thresholds for low frequency sine carriers have been found to be on the order of 10 - 11 cents<sub>rms</sub> (Pollack, 1968, 1970; Cardozo & Neelen, 1968).



$$\text{vibrato} \quad \Delta f_{peak} = \frac{2}{\sqrt{2}} \Delta f_{rms} \quad (2.6)$$

$$\text{jitter} \quad \Delta f_{peak} = 2.587 \Delta f_{rms} \quad (2.7)$$

The peak deviation values for each waveform are shown in relation to the rms deviation values in Table 2.3. These are determined according to

$$\text{cents}_{peak} = \frac{1200}{\log 2} \log \left( \frac{\Delta f_{peak}}{f} + 1 \right) \approx 4 \times 10^3 \log \left( \frac{\Delta f_{peak}}{f} + 1 \right) \quad (2.8)$$

**TABLE 2.3.** Comparison between rms and peak deviation values for vibrato and jitter waveforms.

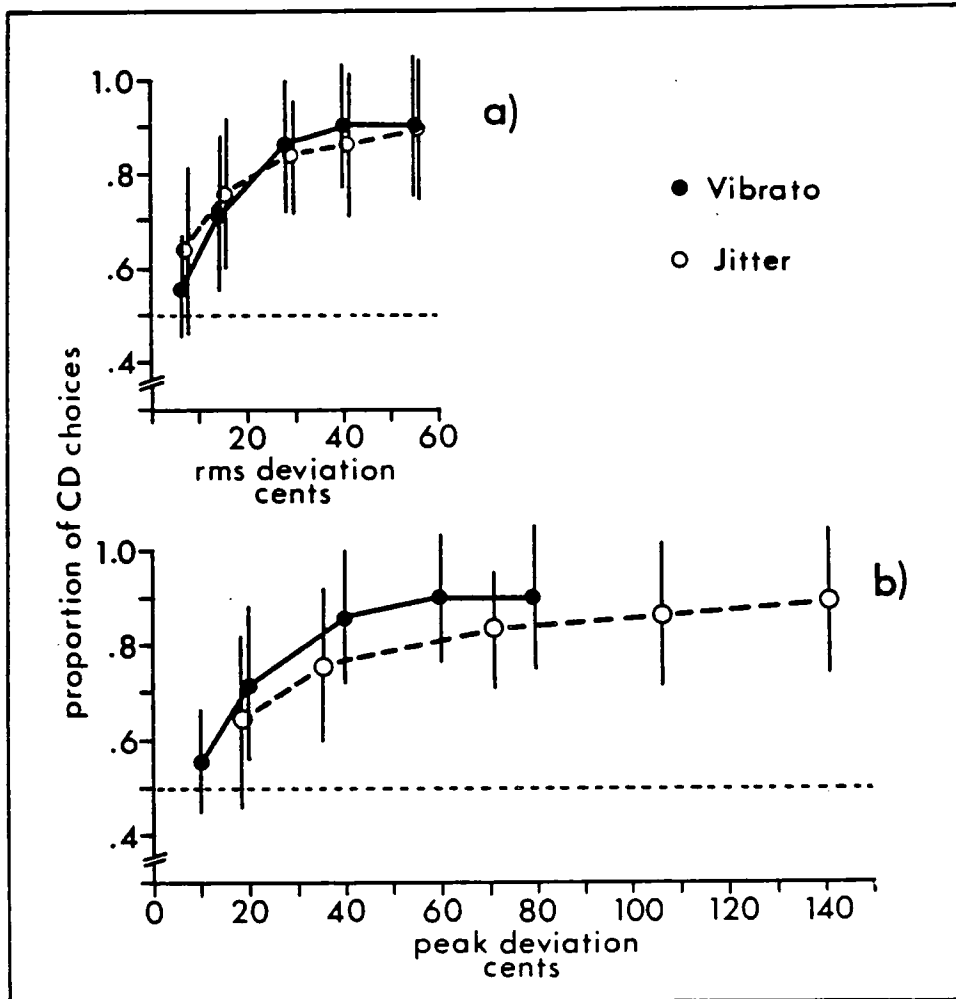
cents <sub>rms</sub>	cents <sub>peak</sub>	
	Vibrato	Jitter
7	9.9	18.0
14	19.8	36.0
28	39.5	71.5
42	59.1	106.6
56	78.7	141.3

The means for each rms deviation within modulation waveform are plotted in Figure 2.4 as functions of both rms deviation and peak deviation. According to the peak deviation measure there seems to be a difference between the data for the two waveforms: comparisons of modulation type would appear not to be as obvious at lower deviation values for jitter as they are for vibrato. However, what is interesting here is that with the rms measure of deviation we have an equivalence between modulation waveforms with respect to source multiplicity perception. This effect may be added to that of vibrato width perception for which Klein & Hartmann (1979) proposed rms deviation detection as one model likely to explain vibrato width matching results for sine carriers.

There is one possible confounding factor in this experiment. It was mentioned earlier that the most prominent percept when listening to *CD* tones is that the fundamental frequency seems to separate from the rest of the tone; that is, at higher

---

4. See the amplitude probability density function for jitter  $J_1$  in Figure B.8.c.



**Figure 2.4.** Experiment 1 data summary. Means within modulation deviation width and modulation waveform are plotted as functions of (a) rms deviation and (b) peak deviation. Data are averaged across all Ss and spectral envelope conditions. The vertical bars represent  $\pm 1$  standard deviation. The closed circles and solid lines represent vibrato data; open circles and dashed lines represent jitter data.

deviations one can clearly hear a low pure tone modulating independently of the rest of the complex. In order to maximize the similarity of criteria used by Ss, they were asked to direct their attention to the region of the  $F_0$ . This effect of a segregated  $F_0$  is not at all present in the *CR* tones; one hears a well-fused rich tone whose pitch is moving according to the modulation function. I would like to believe that what

subjects are reporting is their ability to hear out this pure  $F_0$  in the  $CD$  tones due to the perceptual separation of components. However, if they are trying to listen *only* to the fundamental frequency of either tone, it is entirely possible that they were listening for the tone with a low pitch which moved the most rather than a separate  $F_0$ . Given this were true, we would not expect them to respond preferentially to one tone or the other when the deviation of the  $F_0$  was below modulation detection threshold, or when the difference in  $F_0$  modulation widths was very small. Under these conditions they would either not detect the modulation or, hearing the modulation, not be able to choose one of the tones as having a greater modulation. It should be noted, in light of this argument, that the  $\Delta f_{rms} / \bar{f}$  for the  $CD$   $F_0$  is always a factor of 1.9 greater than that for the  $CR$  tone. Table 2.4 lists the correspondence between the actual rms deviation values on the  $F_0$ 's of  $CR$  and  $CD$  tones.

**TABLE 2.4.** Rms deviation of the fundamental frequencies of  $CR$  and  $CD$  tones. Values are shown in cents<sub>rms</sub> and as  $\Delta f / \bar{f}$ . The column at the right shows the difference in cents between  $CD$  and  $CR$ .

<i>CR</i>		<i>CD</i>		<i>CD - CR</i>
cents	$\Delta f / \bar{f}$	cents	$\Delta f / \bar{f}$	$\Delta$ cents
7	0.00405	13.3	0.00770	6.3
14	0.00812	26.5	0.01543	12.5
28	0.01630	52.8	0.03098	24.8
42	0.02456	78.9	0.04666	36.9
56	0.03288	104.9	0.06246	48.9

As mentioned in footnote 3, MDTs of about 10 - 15 cents have been found for a sinusoidal carrier of approximately the same frequency as the  $F_0$  in this experiment. For the first level of comparisons (7 cents), the modulation of the  $CR$  tone is below these empirical thresholds, while that of the  $CD$  tone is just at threshold. At the second level (14 cents) the  $CD$  tone is well above threshold and the  $CR$  tone is either just at or slightly above threshold. From there on, the differences in modulation width for the two tones are substantial and presumably well above the differential threshold. From these values we would expect subjects that used modulation width as a criterion to not choose preferentially at the first level, but then to choose  $CD$  tones with increasing probability at higher levels. This is fairly well in line with the data

actually obtained and thus represents an alternate possible explanation of the subjects' decisions. To control for this, another study was performed where the  $F_0$  deviations were closer in size. This problem will be addressed in Experiments 3 - 5.

### 2.3 **EXPERIMENT 2:** Perceptual scaling of perceived fusion or multiplicity for harmonic tones with different spectral envelope shapes, frequency modulation waveforms, modulation widths and modulation type.

#### 2.3.1 *Stimuli*

Two of the spectral envelopes from Experiment 1 were used: -6 dB/oct and vowel /a/. Both modulation waveforms (vibrato and jitter) and both modulation types (*CR* and *CD*) were used. For each combination of these 3 parameters, 4 rms deviation values were used: 0, 14, 28, and 42 cents. For 0 cents modulation width (no modulation) the *CR* and *CD* stimuli are identical, and therefore, only one such stimulus is included per combination. Aside from the addition of the no modulation tone, the stimuli are identical to those in Experiment 1.

#### 2.3.2 *Method*

Separate runs were done for each modulation waveform. There were, therefore, 14 stimuli among which comparisons were made: for each spectral envelope there was a stimulus with no modulation, and 3 rms deviations for each modulation type.

To give the subject a sense of the range of differences with respect to fusion in the stimulus set, all stimuli were presented in random succession with a 1.5 sec silence between each tone. Two such random sets were presented in succession. The subject was told to listen carefully to all of the tones, making a quick judgment of the relative multiplicity or fusion of each sound in order to be able to use the scale effectively. After this random presentation, a set of 20 pairs representing the range of expected fusion dissimilarities was presented as practice before the judgments were collected.

To start each trial the subject opened a switch. A pair of tones was presented 500 msec after the switch was raised. A 750 msec silence separated the tones and a 1.2 sec silence followed the last tone. After this interval, the subject could, by pressing one of two buttons, replay either tone at will and as many times as necessary to make

the judgment. The judgment was to decide how dissimilar the two tones were with respect to their perceived fusion, or inversely, with respect to how many sources or distinguishable elements they contained. Subjects were told to ignore the differences between tones due to timbre or perceived modulation width.<sup>5</sup> To make the judgment, the subject was provided with a sliding potentiometer marked only at the top with "very dissimilar, *très dissemblable*" and at the bottom with "very similar, *très semblable*". Subjects were advised to make an initial estimate after the first presentation of the pair, and then to refine their dissimilarity judgment after further presentations, if needed. Once they were satisfied with a judgment, they closed the switch at which point the position of the slider was read and recorded automatically. The positions were translated into a continuous, linear scale between 1 (similar) and 100 (dissimilar). This value represents the subjective difference with respect to fusion between the two tones.

All pairs among the 14 tones (91 pairs) were each presented once. The initial presentation order of the two tones was randomized as was the order of presentation of pairs. The judgments were collected into a lower half-matrix minus diagonal format and analyzed with the KYST multidimensional scaling program (Shepard, 1962a,b, 1963; Kruskal, 1964a,b; Young, 1970, 1972; Kruskal, Young & Seery, 1973). A monotone ascending regression of distances on the data values was used. This means that the regression is non-metric and that large data values will correspond to large distances in the solution and hence to objects that are very different from one another with respect to fusion. An initial configuration for the regression procedure was generated using the TORSCA procedure (Young & Torgerson, 1967).<sup>6</sup> The purpose of this is to

5. This is a very unusual judgment to have to make and always took several practice trials with discussion with the experimenter between each trial before the subject felt confident that he or she could make the judgment. Even then, most Ss reported that it was very difficult to maintain "difference with respect to fusion" as a criterion in the face of the obvious differences in modulation width and timbre of the tones. It should be considered a distinct possibility that these differences, while perhaps correlated with fusion differences, may actually have been the stimulus dimensions being judged as similar or dissimilar.
6. In this procedure the classical Torgerson (1958) scaling technique is used and then Young's quasi-metric method of improving on the resulting configuration is performed. Due to computer memory limitations the latter method cannot be performed on greater than 60 data entries. So, for the group data analysis ( $N = 112$ ) only the Torgerson technique is invoked in the preparation of the initial configuration.

avoid situations where the solution converges on a local minimum in the regression procedure that is not the global minimum being sought. Starting from the initial configuration in a specified number of dimensions, the points are moved bit by bit to reduce the stress (a squared deviation measure of "badness-of-fit") between the configuration and the data. This is iterated until a criterion minimum value is achieved. The resulting configuration is then rotated to principal components which maximizes the spread of points along the different dimensions of the solution space.

Eight Ss participated in the experiment and were paid for their services. Seven of these had participated in the previous experiment. Experiment 2 was always run after Experiment 1 so that the subjects were experienced with the stimuli and with making judgments on the fusion or multiplicity of the stimuli. Instructions were given in English (4 Ss) or French (4 Ss). Two analyses were performed for each subject: one for vibrato and one for jitter stimuli. Group analyses were also performed for vibrato and jitter data separately.

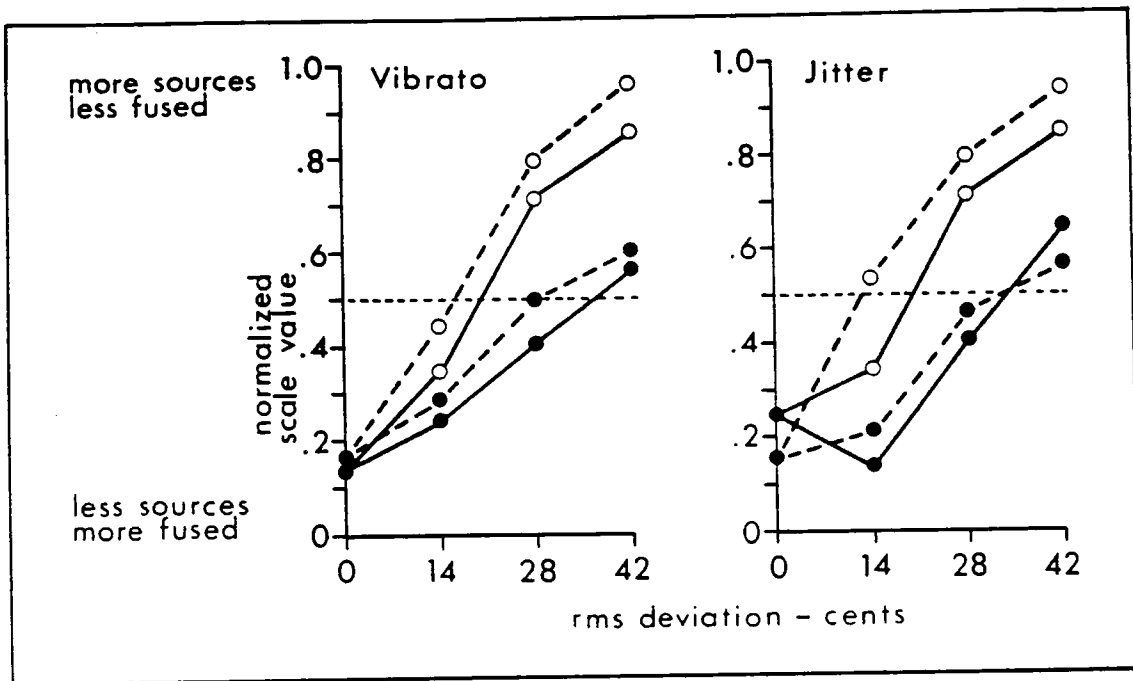
### 2.3.3 *Results*

The one-dimensional solution seemed to give the clearest view of the data structure and was the most easily interpretable. The data structures were very different for vibrato and jitter stimuli in the two-dimensional solution. The jitter stimuli exhibited relationships that had no correlation with the data from Experiment 1 while those from the vibrato stimuli were strongly correlated with Experiment 1 data. With a 1-D solution, the correlations among the data between the experiments were very high and the data structures were very similar in form for both modulation waveforms. Therefore the 1-D solution will be considered to give the best representation of the judgments.

A one-dimensional solution can be interpreted as a scale representing the degree of perceived fusion or perceived dispersion. The stimuli are rank ordered according to their scale values in Table 2.5.<sup>7</sup> In order to visualize the relationships better, the scale values are plotted as a function of rms deviation in Figure 2.5. The 1-D scaling solution can be recovered in this graph by projecting the points onto the vertical axis.

---

7. All scale values fell between  $\pm 2$  in the solution. Here they have been normalized to between 0 and 1.



**Figure 2.5.** Experiment 2 data summary. One-dimensional scaling solutions for judgments on the relative degree of fusion for vibrato and jitter stimuli plotted as a function of rms deviation of modulation. The scale is interpreted as the degree of perceived fusion or perceived source multiplicity. Closed circles represent *CR* tones; open circles represent *CD* tones. Solid lines represent tones with a  $-6$  dB/oct spectral envelope; dashed lines represent tones with a vowel /a/ spectral envelope.

One thing illustrated in Figure 2.5 is that vowel /a/ stimuli almost always have slightly higher scale values than  $-6$  dB/oct stimuli with the same rms deviation and modulation type. Also, *CD* stimuli always have higher scale values than *CR* stimuli with the same rms deviation regardless of spectral envelope type. Finally, with the exception of the *CR* tone with 14 cents rms jitter and a  $-6$  dB/oct spectral envelope, all curves are monotone increasing with rms deviation.

**TABLE 2.5.** Experiment 2 data summary. One-dimensional scaling solution for judgments on the relative degree of fusion for vibrato and jitter stimuli. For each modulation type the stimuli are rank ordered and their normalized scale values are listed. (  $-6 = -6\text{dB/oct}$ ;  $/a/ = \text{vowel } /a/$ .)

Vibrato				Jitter			
Spec. Env.	Mod. Type	Rms Dev.( cents )	Scale Value	Spec. Env.	Mod. Type	Rms Dev.( cents )	Scale Value
-6	-	0	.13	-6	CR	14	.13
/a/	-	0	.17	/a/	-	0	.16
-6	CR	14	.24	/a/	CR	14	.22
/a/	CR	14	.28	-6	-	0	.25
-6	CD	14	.35	-6	CD	14	.34
-6	CR	28	.41	-6	CR	28	.41
/a/	CD	14	.44	/a/	CR	28	.47
/a/	CR	28	.50	/a/	CD	14	.53
-6	CR	42	.56	/a/	CR	42	.57
/a/	CR	42	.60	-6	CR	42	.64
-6	CD	28	.72	-6	CD	28	.72
/a/	CD	28	.79	/a/	CD	28	.79
-6	CD	42	.86	-6	CD	42	.84
/a/	CD	42	.96	/a/	CD	42	.93

#### 2.3.4 Discussion

If we interpret the scale as representing degree of perceived source multiplicity, we can, from the data of Experiment 1, order it with respect to greater and lesser multiplicity. In Experiment 1, the greater the rms deviation, the more often *CD* tones were chosen as yielding more sources. This result is paralleled in the present experiment, i.e. with increasing rms deviation, the pairs of *CD* and *CR* tones presented in Experiment 1 have greater distances between them in the scaling solution. Thus we can say that higher scale values correspond to greater source multiplicity or perceptual dispersion.

According to this orientation, it would be possible to conclude that

1. the vowel sounds are generally perceived as being slightly less fused than the  $-6 \text{ dB/oct}$  sounds,



2. *CD* tones are always perceived as being less fused than *CR* tones, which would confirm the main hypothesis of the experiment, and
3. the greater the rms deviation of the modulation, the less fused the sounds become, even for *CR* tones.

Since all of these deviations are well within musical limits and since musical vibrato is not generally considered to decrease the unity of a given musical source, this result seems a bit surprising and suggests that the structure in the data may not necessarily be related *only* to perceived fusion or source multiplicity.<sup>8</sup> As noted in the Method section, subjects found this task very difficult to do and often caught themselves making dissimilarity judgments on the basis of timbre and modulation width differences. Both of these differences show up in the data structure. In particular, the data structure seems strongly correlated with the actual modulation width of the fundamental frequency. This was noted as a possible confounding influence on judgments in Experiment 1 as well. In fact, the computed correlation coefficients between actual  $F_0$  modulation widths (Table 2.4) and scale values (Table 2.5) are 0.98 and 0.95 for vibrato and jitter, respectively. This means that over 90% of the variation in the data could be accounted for by this physical parameter. Unfortunately, it is impossible to arrange the acoustic parameters so that such confounding effects are entirely eliminated.

In listening to these stimuli, it is obvious to me that the *CD* tones are less fused and more unstable than the *CR* tones. The real problem is three-fold:

1. the task being demanded of subjects is too far removed from the kind of comparisons they are used to making in normal hearing.

---

8. There is, however, the possibility that these relatively simple synthetic stimuli (with a constant spectral envelope that does not exhibit the resonant characteristics of filters) have an amplitude modulation on some partials that is induced by the larger width frequency modulations. This would cause partials on the slope of a sharp formant to oscillate widely in amplitude and possibly make them stand out separately, thus reducing the degree of perceived fusion.

2. several perceptual phenomena (perceived modulation width and perceived multiplicity) are closely coupled to the same physical parameters making it difficult to tell which one is actually being measured, and
3. the subject is instructed to listen for dissimilarity; dissimilarities in timbre and modulation width may be very large even when fusion differences are small and so the subject feels compelled to make a judgment of dissimilarity anyway.

Since the *CR* tones would not *a priori* seem to change in relative fusion in the manner indicated with changing rms deviation, it seems reasonable to conclude that perception of timbre differences and modulation width differences are also entering into the dissimilarity "with respect to fusion" judgments.

#### 2.4 EXPERIMENTS 3 - 5: Corollary Studies to Experiment 1

Given the ambiguity of interpretation of Experiments 1 and 2, at least three things need to be verified in order to clarify the conclusions to be drawn from their results.

1. In the previous experiments, the rms deviation on the fundamental frequency was quite a bit larger for the *CD* tones than for the *CR* tones. One wonders if the relation of multiplicity judgments to the overall rms deviation would be significantly changed if the  $F_0$  modulation widths were approximately equal. If the widths were equal *and* the judgments were made on the basis of modulation width, the ZIFC judgments should fluctuate around random choice. If they were really made on the basis of the perceived multiplicity, we would expect that the proportion of *CD* tones chosen as yielding more sources should increase with increasing rms. Experiment 3 tested this possibility.
2. Experiments 4 and 5 were designed to verify the relation of the multiplicity judgments in Experiments 1 and 3 to judgments of the relative modulation widths of the *CR* / *CD* tone pairs. Experiment 4 used the same stimulus pairs as in Experiment 1, and Experiment 5 used the same stimulus pairs as in Experiment 3.

3. Lastly, it is important, given the possibility that both modulation width and multiplicity judgments are entering into the data, that the subjects be questioned extensively concerning their impressions about the relative salience of both of these perceptual effects.

#### 2.4.1 *Stimuli and Subjects*

All stimuli were selected from those used in Experiment 1. All 3 spectral envelopes were used, but only the vibrato modulation was included since no difference was found between modulation waveforms for these judgments. The rms values of each tone were selected according to the experiment as described below.

Four subjects participated in all 3 experiments and were paid for their time. None reported having any hearing problems. Ss 1 and 3 had participated in Experiment 1 eight months earlier. S1 had also participated in Experiment 2 at the same time. Experimental instructions were given in either English (3 Ss) or French (1 S). All subjects were questioned extensively after each experiment concerning their impressions of the judgment and the stimuli. At the end of all 3 experiments, they were asked to give their impressions of the relation between stimulus sets and the two judgments on the stimuli.

#### 2.4.2 *Method and Results*

##### 2.4.2.1 *Experiment 3: Source multiplicity judgments on CR/CD tone pairs with very small differences in $F_0$ modulation width.*

Tones were selected from Experiment 1 which had the smallest differences between them for  $F_0$  modulation width.<sup>9</sup> These were CR (14 cents) / CD (7 cents), CR (28 cents) / CD (14 cents), CR (56 cents) / CD (28 cents). The actual  $F_0$  deviations and differences for each pair are listed in Table 2.6. Note that the differences in  $F_0$  modulation width are much smaller than and opposite in sign to those used in

9. Unfortunately, the possibility of a confounding effect of  $F_0$  modulation width did not occur to me until after the old computer system, on which all sound synthesis software existed, was removed from service at IRCAM. At the time of this writing, the new system was not yet producing sound. Therefore, I was required to rearrange the old stimuli in as close an approximation to equal  $F_0$  modulation width as possible.

Experiment 1. These 3 pairs of modulation widths were presented with the three spectral envelopes of Experiment 1.

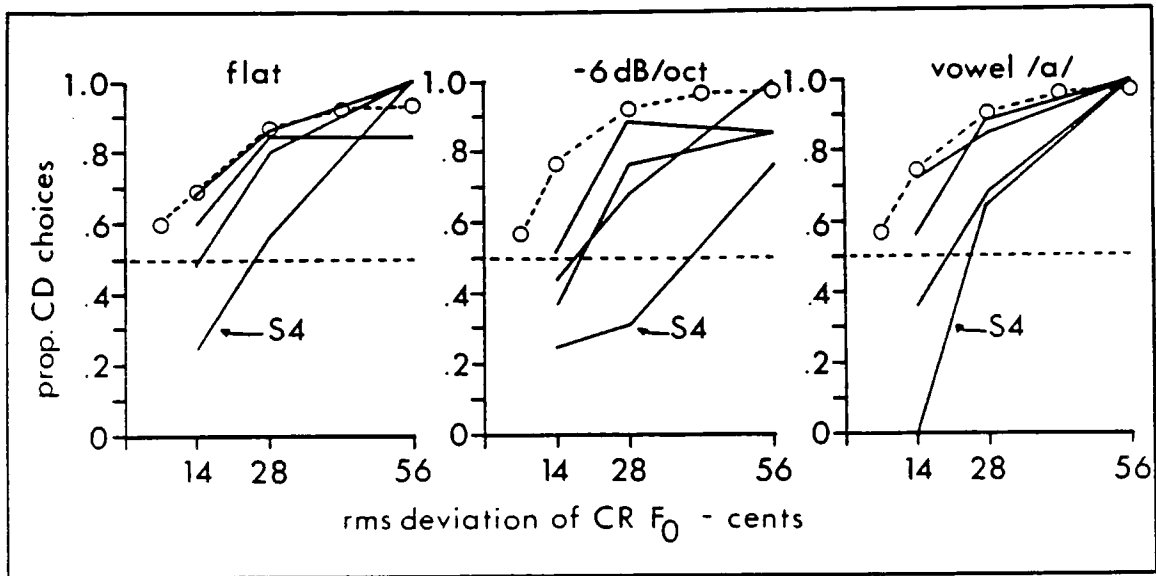
**TABLE 2.6.** Rms deviation (cents) for *CR* and *CD* tones, matched as closely as possible with existing stimuli.  $\Delta\text{cents} = CD - CR$ .

<i>CR</i>	<i>CD</i>	$\Delta\text{cents}$
14	13.3	-0.7
28	26.5	-1.5
56	52.8	-3.2

As in Experiment 1, the subject was to decide which tone in the pair had the most sources or was the most "split apart" perceptually. Each stimulus pair was presented 25 times in a block randomized order. Thus, 225 comparisons, (3 spectral envelopes)  $\times$  (3 rms deviation pairs)  $\times$  (25 repetitions), constituted one experimental session. The collected data represent the proportion of times the *CD* tone was chosen as yielding more sources.

The data for each subject are listed in Table E.2 (Appendix E). These data are plotted as a function of *CR*  $F_0$  modulation width in Figure 2.6 to see the spread due to subjects. Also plotted, for comparison, are the mean data from Experiment 1. Note that all of these curves are monotone ascending, i.e. more *CD* choices are being made at larger rms deviations. This occurs in spite of the fact that with increasing modulation width, the modulation on the  $F_0$ 's of *CR* tones are getting progressively larger than those in *CD* tones (though with much less of an absolute difference than was found in Experiment 1 stimuli). If subjects were making judgments solely on the basis of  $F_0$  modulation width, one would expect these curves to be monotone descending or near random choice, given that the modulation width differences are less than 3.2 cents. (Refer again to Table 2.6 for the  $F_0$  modulation width differences across tones with each stimulus pair.)

The modulation widths on the  $F_0$ 's of all stimuli are listed in Table 2.7. The relative modulation width differences, expressed as a proportion of the actual modulation width of the *CR* tone, are listed in Table 2.8. It is clear from these values that modulation width difference judgments should be much easier to make with the Experiment 1 tone pairs ( $(\Delta f_{CD} - \Delta f_{CR}) / \Delta f_{CR} \approx 0.90$ ) than with the Experiment 3 tone pairs



**Figure 2.6.** Experiment 3 data summary. The proportion of times the *CD* tone was chosen as yielding more sources is plotted as a function of rms deviation of modulation (the modulation width on  $F_0$  in the *CR* tone for each stimulus pair). Each curve represents the data for one subject. Each point represents 25 2IFC judgments. These stimuli have  $F_0$  modulation widths which are very close for each pair. Also plotted, for comparison (open circles), are the mean vibrato data from Experiment 1, where the  $F_0$  modulation widths are different by a factor of 1.9 for *CR* and *CD* tones.

**TABLE 2.7.** Modulation widths on  $F_0$  in Experiments 1,3,4,5.

<i>CR</i> tones		<i>CD</i> tones	
cents	$\Delta f_{rms}$ (Hz)	cents	$\Delta f_{rms}$ (Hz)
7	0.89	13.3	1.70
14	1.79	26.5	3.39
28	3.59	52.8	6.81
42	5.40	78.9	10.26
56	7.23	104.9	13.74

$((\Delta f_{CD} - \Delta f_{CR}) / \Delta f_{CR}) \approx 0.05$ ). I would imagine that if sine tone stimuli with frequencies equal to the  $F_0$  in this experiment were presented to subjects in a differential

modulation width discrimination task, performance would be random.

**TABLE 2.8.** Relative difference of modulation widths ( $\Delta f_{rms}$  in Hz) across *CR* / *CD* pairs in Experiments 1,3,4,5.

<b>Experiments 1 and 4</b>					
$\Delta f_{rms}(CR)$	0.89	1.79	3.59	5.40	7.23
$\Delta f_{rms}(CD)$	1.70	3.40	6.81	10.26	13.74
$\frac{\Delta f_{CD} - \Delta f_{CR}}{\Delta f_{CR}}$	0.91	0.90	0.90	0.90	0.90
<b>Experiments 3 and 5</b>					
$\Delta f_{rms}(CR)$		1.79	3.59		7.23
$\Delta f_{rms}(CD)$		1.70	3.40		6.81
$\frac{\Delta f_{CD} - \Delta f_{CR}}{\Delta f_{CR}}$		-0.05	-0.05		-0.06

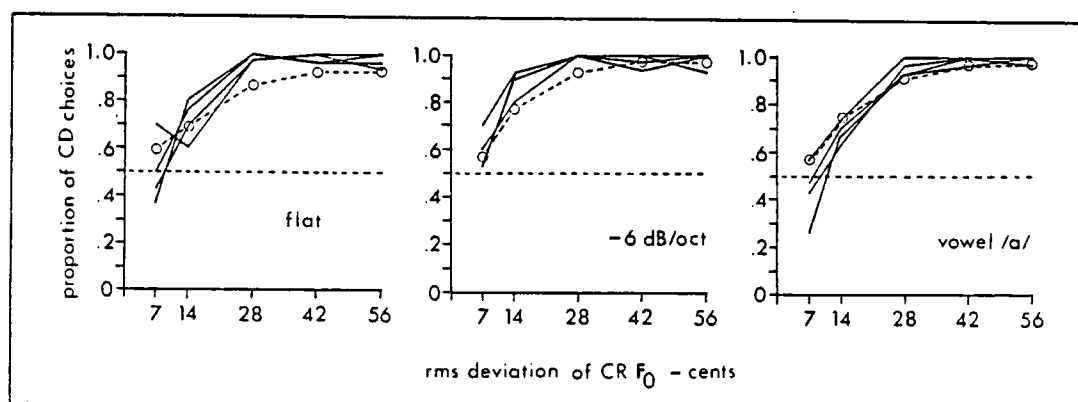
The results seem to indicate large differences due to subjects and moderate differences due to spectral envelope. It is interesting to note that for the smallest modulation width value, there is a tendency for the subjects to be choosing the *CR* tone as yielding more sources than the *CD* tone. This is especially true for S4's judgments on the vowel stimulus. In general, the multiplicity difference between tones is less discernible in this experiment than in Experiment 1, as is illustrated by the higher values from that experiment. The highest values achieved in this experiment were recorded for Ss 1 and 3 who also participated in Experiment 1, so there may be effects of settling into a criterion for making the judgment here.

#### 2.4.2.2 **Experiment 4:** *Modulation width judgments on the vibrato stimulus pairs from Experiment 1.*

The purpose of this experiment was to compare the source multiplicity judgments from Experiment 1 with judgments on the relative modulation widths on the  $F_0$ 's of the same *CR*/*CD* tone pairs used in that latter experiment. Thus the stimuli used in this experiment were identical to those in the vibrato condition in Experiment 1. This time subjects were asked to attempt to listen only to the fundamental

frequency of each complex tone and to decide which one had the largest or widest frequency modulation. As before, the judgment was indicated by pressing a button. Data were collected as the proportion of times the *CD* tone was chosen as having a greater modulation width. The experiment was conducted in one session with 450 comparisons: (3 spectral envelopes)  $\times$  (5 rms deviations)  $\times$  (30 repetitions).

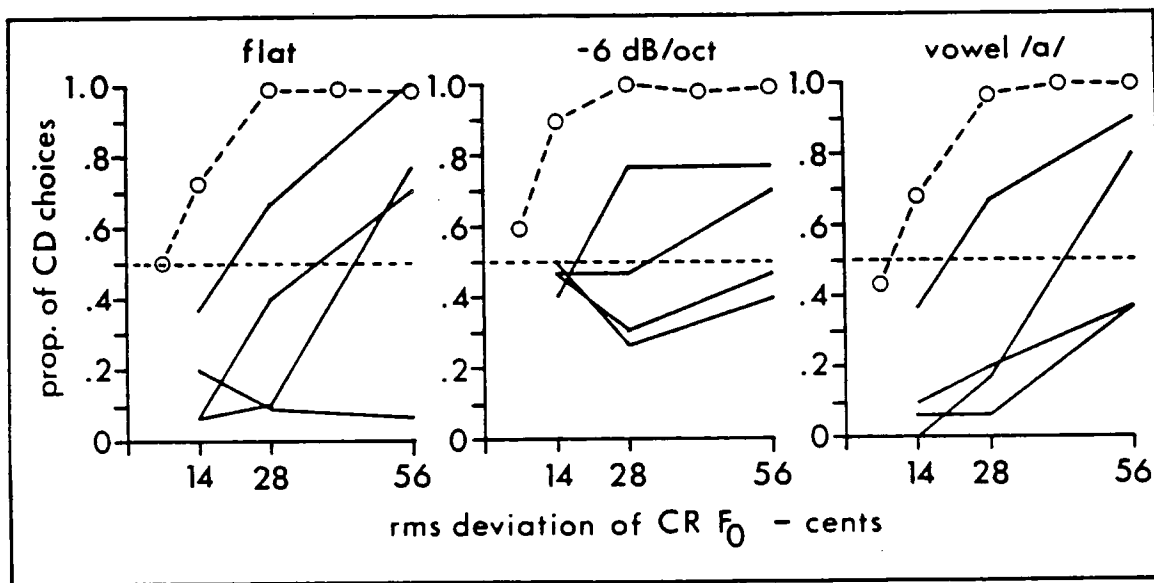
The data for all subjects are listed in Table E.3 (Appendix E). These data are plotted in Figure 2.7, along with the mean data from the multiplicity judgments on the same stimuli from Experiment 1. All curves (except one) are monotone ascending and appear to have a slightly more rapid rise with increasing rms deviation than the curve for Experiment 1. There appear to be no differences due to subjects or spectral envelope. It is quite interesting to note here that very similar results are obtained with either the source multiplicity or modulation width judgment.



**Figure 2.7.** Experiment 4 data summary. The proportion of times the *CD* tone was chosen as having a larger modulation width on the  $F_0$  than that on the *CR* tone is plotted as a function of the modulation width of the *CR*  $F_0$ . Refer to Table 2.4 for the difference in  $F_0$  modulation width between *CR* and *CD* tones. Each curve represents the data for one subject. Each point represents 30 2IFC judgments. For comparison with the source multiplicity judgments on the same stimuli, Experiment 1 vibrato data are plotted as open circles.

### 2.4.2.3 Experiment 5: Modulation width judgments on the vibrato stimulus pairs from Experiment 3.

The purpose of this experiment was also to provide relative  $F_0$  modulation width judgments on a set of stimulus pairs for which source multiplicity judgments had been collected. But this time, the  $F_0$  modulation widths were very similar across the CR/CD tone pair. The stimuli used in this experiment were identical to those in Experiment 3. As in Experiment 4, subjects were asked to judge which tone of the CR/CD pair had the greatest frequency modulation width. Data were collected as the proportion of times the CD tone was chosen as having a greater modulation. The experiment was conducted in one session with 225 comparisons: (3 spectral envelopes)  $\times$  (3 rms deviations)  $\times$  (25 repetitions).



**Figure 2.8.** Experiment 5 data summary. The proportion of times the CD tone was chosen as having a larger modulation width on the  $F_0$  is plotted as a function of CR modulation width. Each curve represents the data for one subject. Each point represents 25 2IFC judgments. For comparison, the mean data from Experiment 4 are plotted (open circles).



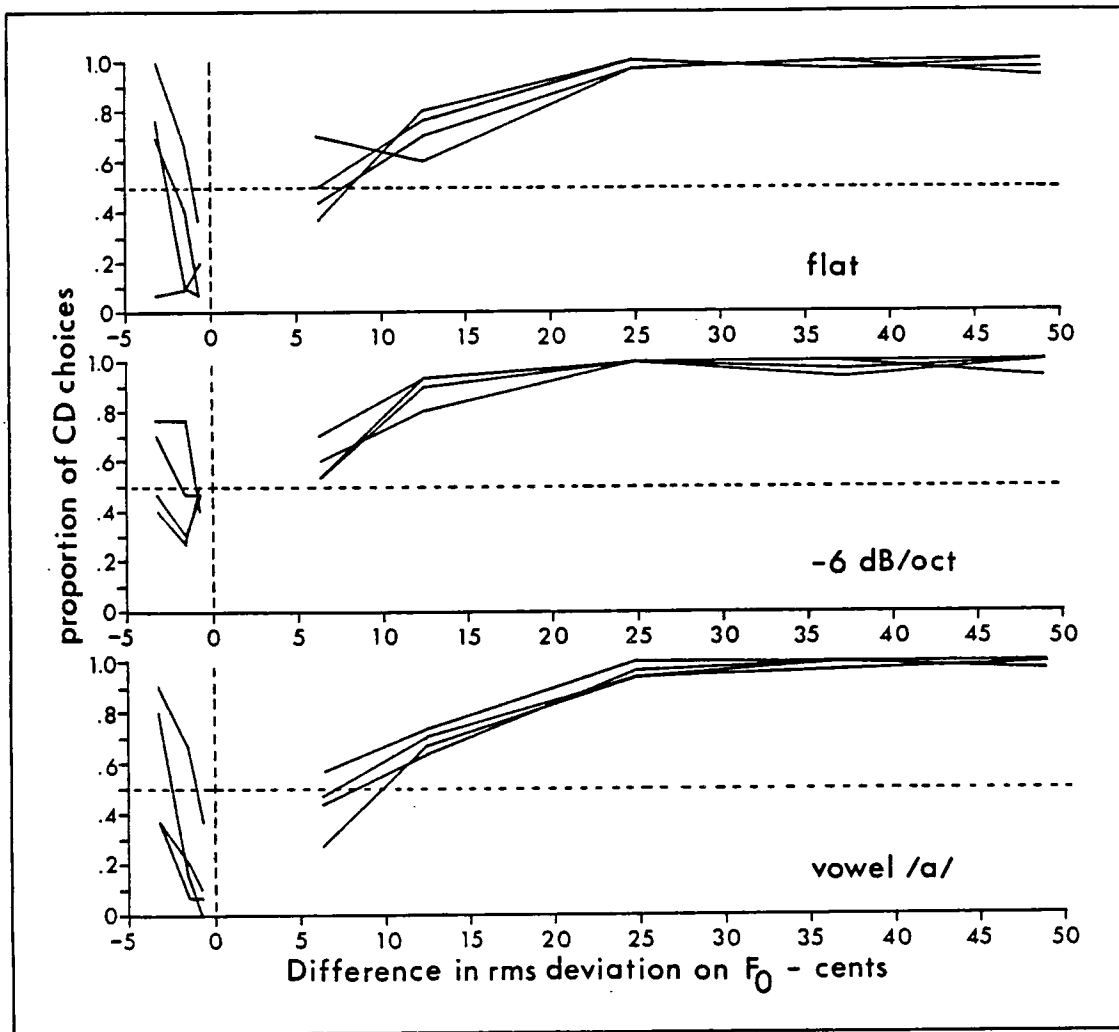
The data for all 4 subjects are listed in Table E.4 in Appendix E. These data are plotted in Figure 2.8. Also plotted are the mean data from Experiment 4 in order to view the difference due to modulation width differences at similar overall rms deviations. The *CR* tones are identical in the two experiments, but the *CD* tones have much smaller  $F_0$  modulation widths than their *CR* partners in the present experiment. Generally, these curves are monotone ascending for the flat and vowel envelopes, except for S4, for whom the curve for the flat envelope is relatively constant at very low values. The responses for the  $-6$  dB/oct spectrum are particularly unsystematic across rms deviation and subjects. Here, as well as in Experiment 3, there are major differences due to subjects. There are even greater differences due to spectral envelope.

The values here are (with one exception) always far below those of Experiment 4. Many of these values are far below chance indicating that the *CR* tone is being chosen as having a greater modulation width, particularly at the 14 cents rms deviation *where the difference between the modulation widths is only  $-0.7$  cents* ! It is important to note that with increasing rms deviation more *CD* choices tend to be made than at smaller rms deviations. In fact, this goes in direct opposition to the direction of change that should occur if subjects were really making judgments on the basis of  $F_0$  modulation width, since at larger deviations, the *CR* tones have larger  $\Delta f_{rms}$ . This raises the doubt as to whether subjects are actually judging  $F_0$  modulation width.

### 2.4.3 Discussion

#### 2.4.3.1 Comparison of the experiments

The data from Experiments 4 and 5 are plotted as a function of the modulation width difference between *CR* and *CD* tones ( $\Delta$ cents) in Figure 2.9. For Experiment 4, at small differences (6.3 cents) and small rms deviation there is no preferential choice of either tone. This is not entirely surprising since the rms deviation is probably below or just at modulation detection threshold. At larger rms deviations, where the  $\Delta$ cents is much larger and the modulations are at suprathreshold widths, subjects reliably chose the tone which actually had the larger modulation width.



**Figure 2.9.** Data from Experiments 4 and 5 (modulation width judgments) plotted as functions of the difference in modulation width ( $CD - CR$ ) of the  $F_0$ 's on the two tones in a trial. The  $\Delta$ cents values in Experiment 4 are positive; those in Experiment 5 are negative. Each curve represents the data for one subject.

For Experiment 5, at small differences and small to medium rms deviations, subjects chose the tone with the larger modulation width (except for the  $-6$  dB/oct spectrum). At slightly larger differences (still much less than the smallest difference in Experiment 4) and larger rms deviations, there was more of a tendency for Ss 1 and 3 to choose the tone with the *smaller* modulation width on  $F_0$  as having a larger modulation. S4 and S2 (for  $-6$  dB/oct and vowel /a/), who generally chose  $CR$  tones as

having larger modulation widths, were an exception to this.

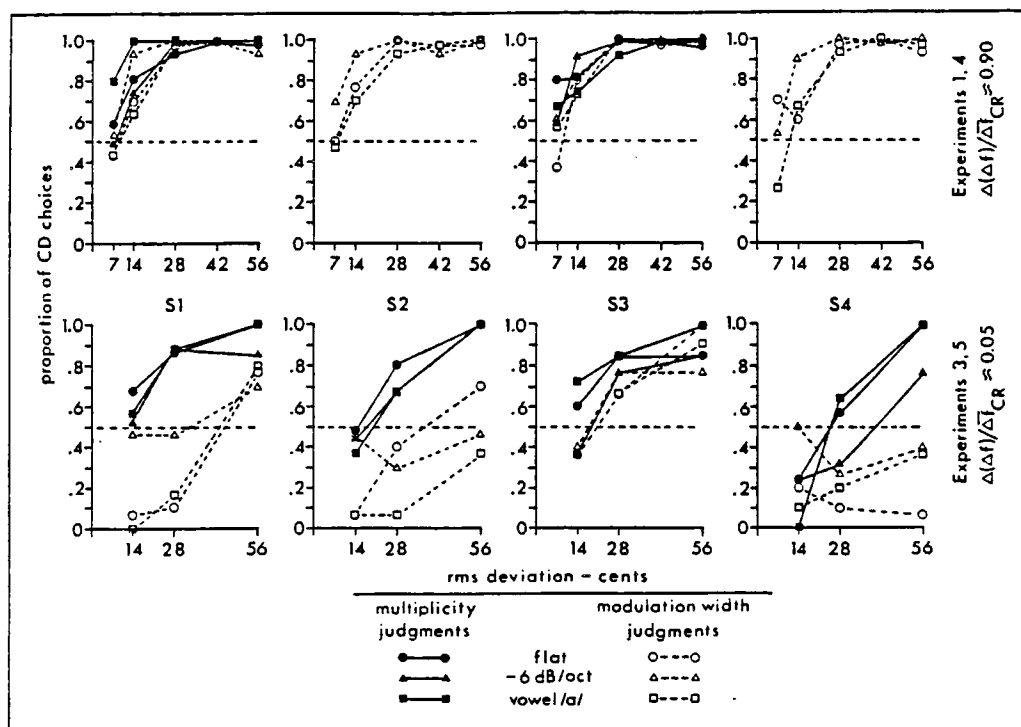
Looking at Table 2.8 we see that all of the differences in Experiment 4 have about the same  $\Delta(\Delta f_{rms})/\Delta f_{CR}$ . The same relation holds among the differences in Experiment 5. However, the values in Expt. 4 are quite large compared to those of Expt. 5 and, as mentioned previously, it is doubtful that the judgments in the latter experiment are actually being made on the basis of modulation width of  $F_0$ .

The individual data for each subject for Experiments 3 - 5 (and for S1 and S3 in Expt. 1) are plotted in Figure 2.10. There is very little difference between fusion and modulation width judgments for Expts. 1 and 4. However, there is a large difference between these judgments for Expts. 3 and 5 for all subjects except S3. It may help clarify the interpretation to consider what the subjects said about the stimuli and the judgments.

#### 2.4.3.2 *Subjects impressions of the stimuli and judgments.*

For the conditions of Expts. 1 and 4 ( $\Delta$ cents large), all subjects said it was relatively easy to make the modulation width judgment. All noted that in the "fused" tone, it was difficult to hear the  $F_0$ , whereas in the "split apart" tone it was very easy to hear the  $F_0$  since it "stood out". Thus the *separation of the  $F_0$  from the rest of the tone* aided in the modulation width judgment on the  $F_0$ . It is clear from the subjects' reports that even though the two effects are perfectly coupled in these conditions due to the construction of the stimuli, they are easily attended to separately.

For the conditions of Expts. 3 and 5 ( $\Delta$ cents small), all subjects felt that the multiplicity judgments were fairly clear, except for the smallest modulation widths. However, all found it very difficult to make the modulation width judgments. Ss 1 and 2 noted that at large widths, it was difficult to hear the  $F_0$  in the fused tone and that they thus tended to choose the audible  $F_0$  of the unfused tone as having a larger width on  $F_0$ . At small rms deviations, the modulation on the *CD* tone is very minor perceptually and the global effect of modulation on the *CR* tone gave the impression of a much greater modulation width. These tendencies are reflected in these subjects' data. S3 remarked that it was very difficult to hear the  $F_0$  in the fused tone and so he almost always chose the unfused tone as having a greater modulation width. It is for this reason, i.e. he used the audibility of  $F_0$  (due to defusion) as a criterion for his



**Figure 2.10.** Individual data for each subject for Experiments 1 (S1 and S3 only), 3, 4 and 5 are plotted. Source multiplicity judgment data are represented as filled/solid and modulation width judgments are represented as open/dashed. The upper plot is for Experiments 1 and 4 ( $\Delta f_{CD} \approx 1.9 \Delta f_{CR}$ ). The lower plot is for Experiments 3 and 5 ( $\Delta f_{CD} \approx \Delta f_{CR}$ ). The values on the abscissa represent the rms deviation on the CR tone's  $F_0$ . See Tables 2.4 and 2.5 for relations between CR and CD  $F_0$ 's. The modulation waveform in all cases was vibrato.

width judgments, that the curves for these judgments and those for the multiplicity judgments are almost identical. S4 claimed that the modulation width judgments were very confusing to make. In the "split apart" tones there were two modulation widths, one on low sounds and a lesser one on higher sounds. She found it difficult to ignore the higher one in making the judgment. In general, she found that the fusion of one tone in the pair (CR) seemed to enhance the degree of overall modulation on that tone and that she thus tended to choose that tone as having a greater modulation. In a sense, she chose the tone that seemed to yield the greatest sensation of "action". These impressions are also clearly reflected in her data.

Objectively, if we were to predict the subjects' performances on the basis of the change in modulation width relative to the total modulation width (Table 2.6), we would predict a very high proportion of *CD* choices for Expt. 4 and random performance for Expt. 5. The lower values actually found at smaller rms in Expt. 4 could be attributed to the rms deviation being near absolute modulation detection threshold. However, the data suggest (and the subjects' impressions confirm) that the fusion or multiplicity of the tones strongly affected their ability to make the modulation width judgments. Therefore, I would conclude that the multiplicity judgment is a more evident one to make perceptually and that the data from Expt. 1 are truly reflective of the relative fusion under these stimulus conditions.

## 2.5 General Discussion and Summary

The main hypothesis of this chapter may be considered as supported by the experimental evidence. Namely, frequency components that are modulated in such a manner that the ratios among them are maintained, tend to fuse more readily into a single source image than components not maintaining such a relation. This holds even if the components are moving in the same direction in frequency at any given time as in the *CD* tones.

The data of Experiments 1 and 3 showed that as the rms deviation of modulation is increased, *CD* tones are more often judged as having more sources. This is hypothesized to be due to at least two factors:

1. As the frequencies are modulated in *CD* tones, they move in and out of a harmonic relation. The greater the modulation width, the greater the departure from harmonicity. Pseudo-harmonic signals of this type have been shown repeatedly to yield multiple pitches or a sensation of dispersion of the pitch. When contrasted with the single-pitched *CR* tone, whose harmonicity is maintained rigorously in the presence of modulation, this multipitched or dispersed pitch nature could induce a judgment of more sources.
2. Because of the manner in which modulation was imposed on the components of *CD* tones, the  $F_0$  moves through a greater range on a log frequency scale (i.e. through a larger pitch interval), and is thus perceived as moving more than the rest of the components. Most subjects reported hearing a  $F_0$  that segregated

perceptually from the less modulated remainder of the complex tone. This separately perceived  $F_0$  is perhaps the strongest cue for the presence of multiple sources in *CD* tones. As with the inharmonicity factor, this separation becomes progressively more apparent as the rms deviation is increased.

One possible confounding element in Experiments 1 and 2 was the fact that the  $F_0$  had a greater modulation width in *CD* tones than in *CR* tones. The predicted judgments based on source multiplicity could also have been considered to be made on the basis of choosing the  $F_0$  that had the widest modulation. Experiment 4 demonstrated that modulation width judgments gave results very similar to those for source multiplicity judgments with the same stimulus set.

However, both judgments were asked of subjects in response to pairs of *CR* and *CD* tones where the  $F_0$  modulation widths were much closer in size. In Experiment 3, the data from multiplicity judgments showed that subjects still chose *CD* tones as having more sources at higher rms deviations when the *CR* tones even had slightly greater modulation widths on the  $F_0$ . In Experiment 5, the data from modulation width judgments were somewhat confusing. Subjects' impressions indicated that the relative fusion or multiplicity of the tones was a strong factor influencing these judgments.

The conclusion to be drawn is that harmonicity-maintaining modulation induces a more unified, less analyzable auditory image than does modulation not maintaining harmonicity, even when the direction of frequency change across components is similar. This extends and supports a similar finding of Bregman *et. al.* (1978). The following chapter will examine the effects on source multiplicity perception of modulation patterns on adjacent components that are completely incoherent with respect to one another.

## CHAPTER 3

### Within-channel and Cross-channel Contributions to Multiple Source Perception

#### 3.1 Introduction

In Chapter 1 it was proposed that two types of mechanisms might be involved with extracting information about source behavior on the basis of frequency modulation coherence:

1. a within-channel mechanism operating on the regularity or change in behavior of the temporal discharge pattern within an auditory channel, and
2. a cross-channel mechanism making comparisons or correlations of the temporal behavior in different auditory channels.

Let us examine in more detail the possible nature of such processes.

##### 3.1.1 *Within channel information*

A within-channel mechanism might use the periodicity or regular pattern of nerve firings to signal the presence of a single source within that channel's effective band of frequency response. Irregularities or perturbations of periodicity may be used to signal the presence of multiple sources. Two partials whose frequencies are close enough to have overlapping excitation patterns on the basilar membrane have been shown to create a complex temporal pattern of neural impulses in the auditory nerve fibers. This response corresponds statistically to a half-wave rectified, band-filtered version of the signal with some phase and amplitude distortion due to

propagation delays in the inner ear and frequency-specific attenuation in the peripheral auditory system, respectively (cf. Hind, Anderson, Brugge & Rose, 1967). Brugge *et al* (1969) recorded a complex temporal pattern in a cat auditory nerve fiber (whose characteristic frequency was 1200 Hz) in response to two partials at 907 Hz and 1814 Hz, a ratio of 2:1 (a separation of more than 4 critical bandwidths in the human auditory system). Another fiber with a characteristic frequency somewhere between the two stimulus frequencies responded with a complex temporal pattern when these frequencies were in a ratio of 3:1 (300 and 900 Hz; a separation of about 5 human critical bandwidths). Note that the characteristic frequencies of the fibers (that is the frequency to which a fiber responds preferentially) were at least 2 critical bandwidths away from the stimulus frequencies, and yet a complex pattern of response was obtained indicating interaction between frequencies at these great distances.

This interaction of components at distances greater than a critical bandwidth has also been investigated psychoacoustically by Plomp (1966). He presented two sinusoidal components that were slightly mistuned from a "consonant" interval (i.e. a small numbered integer ratio,  $m:n$  with  $m < n$ ). At appropriate sound levels of the two tones, subjects reported hearing an auditory beating effect for ratios as large as 12:1. Plomp experimentally ruled out the possibility that these beats were due to interference of combination tones stimulating proximal regions of the cochlea and concluded that the effect results from "some auditory mechanism sensitive to cyclic variations in the compound waveform of the tones in an area where the two excitation patterns overlap . . . . The ear's sensitivity to changes in the compound waveform may be related to the preservation of phase information in the temporal distribution of the discharges of the auditory nerve fibers." (Plomp, 1976; p. 56)

Consider the temporal patterns of response in fibers stimulated by more than one frequency component. If these partials have a harmonic relation, the temporal pattern is statistically periodic. This can be measured by recording a period histogram<sup>1</sup>

1. A period histogram may be considered to represent the probability of occurrence of a neural spike (an action potential) at a certain point during the period of the recording cycle. This period usually corresponds to the period of some component or of some submultiple of two or more components in the stimulus waveform. The histogram is obtained by counting the number of neural spikes that occur during a small time epoch, e.g. 10  $\mu$ sec or 50  $\mu$ sec, at a given phase of the recording period. These are



of the frequency of occurrence of neural spikes over many periods of the signal. One finds prominent peaks at submultiples of the signal period whose actual placement within the recording period depends on the phase relation between the partials. If these frequency components are modulated coherently in frequency at modulation frequencies less than the fundamental, one would obtain a modulation of the periodicity, but the form of the period histogram would presumably remain the same (assuming one changed the histogram period in correlation with the changing signal period).

If the partials have an inharmonic relation, period histograms can still be obtained when the measurement period is synchronized to the frequency of either component (Evans, 1978). In this case the peak in the histogram represents phase-locking of neural fibers to that particular component. This form is not as clear, or the peaks as prominent, as in the harmonic case. However, the overall pattern of response in the fiber varies as the components move in and out of phase with one another since a loss of harmonicity means a loss of phase synchrony between the frequency components. If the components stimulating a given auditory fiber are modulating incoherently, one would also expect an irregular temporal pattern that varied according to the irregularities in the band-filtered stimulating waveform. Such irregularities may be perceived by a listener as auditory roughness if the frequency of variation is between about 10 - 100 Hz or so.

It is important to remember that these regions of overlap and, thus, of complex response constitute only a portion of the response in the auditory nerve fiber array. At other points on either side of the region of interaction one would find fibers responding selectively to either one component or the other. The further apart the frequencies (and thus regions of maximal stimulation) are, the smaller and less significant the areas of compound response in relation to the overall activity.

It is well-established that the extent of excitation in the cochlea of a given sinusoidal stimulus is related to its intensity. At threshold intensities, only a very small, frequency dependent region of the cochlea is stimulated. At greater intensities, the region of stimulation spreads laterally with the greatest spread toward the

---

collected over several thousand periods, i.e. several seconds of continuous presentation of the stimulus.

region sensitive to higher frequencies. The classical "tuning curve" describes the relation between stimulus frequency and the intensity that just barely evokes a measurable response. The area above this threshold curve is called the frequency response area and represents the frequency-intensity combinations of pure tones that will evoke a response. These curves have been measured physically,<sup>2</sup> physiologically,<sup>3</sup> and psychoacoustically.<sup>4</sup> All of these studies suggest that with increased intensity at levels well above response threshold, an ever-increasing range of frequencies is responded to by a given auditory frequency channel.<sup>5</sup> This suggests also that an auditory nerve fiber connected to a particular point on the basilar membrane may be stimulated by only one component of a complex tone at low intensities, but may respond to several frequencies at higher intensities, as their respective excitation regions begin to overlap.

From a consideration of the nature of single channel stimulus encoding, one would expect the following stimulus parameters to affect the regularity of temporal response:

1. proximity of frequency components (affects degree of excitation overlap),
  2. overall stimulus intensity (affects degree of excitation overlap),
  3. harmonicity of components whose excitations overlap (affects periodicity of temporal discharge pattern),
  4. coherence of frequency modulation among overlapping components (affects periodicity of response pattern), and
- 
2. Extent of displacement of the basilar membrane is measured; cf. von Békésy (1960); Johnstone & Boyle (1967); Rhode (1971); Khanna & Leonard (1982).
  3. Response of hair cells: Russell & Sellick (1977, 1978), Sellick & Russell (1979); response of auditory nerve fibers: Kiang (1965), Evans (1970); and response of cells of auditory nuclei in the brainstem and cortex: e.g. Kiang, Morest, Godfrey, Guinan & Kane (1973) for cochlear nucleus; Boudreau & Tsuchitani (1970) for superior olivary complex; Rose, Greenwood, Goldberg & Hind (1963) for inferior colliculus; Hind (1952) for primary auditory cortex.
  4. Masking experiments: cf. Zwicker (1974), Mills & Schmiedt (1983).
  5. In the case of psychoacoustic measurement of masking, this behavioral response is assumed to reflect the physiological fact that the excitation due to the masking signal is able to occlude the excitation due to other frequencies at a greater distance as the masker intensity is increased.

5. extent of modulation (affects degree of excitation overlap).

As the degree of overlap (component proximity) is decreased, the inharmonicity or incoherence among nearby components would have less of an effect on single auditory channels. For more proximal components with overlapping excitation regions, a perturbation of the harmonicity of the components would cause an irregularity in the temporal response pattern. Further, perturbations in regularity would be caused by incoherently modulating the frequencies of proximal components. One would expect in this latter case that a lesser extent of modulation would be necessary to evoke a perceptual response if the components were very close in frequency and were harmonic. The closer the partials, the greater would be the number of fibers responding to multiple components. And if these components are behaving the slightest bit incoherently, this would perturb the regularity of response in each of those channels. If the partials were inharmonic to start with, thus giving rise to irregularity in timing pattern, a greater amount of incoherent modulation would be necessary for a within-channel mechanism to detect a further perturbation of the already irregular temporal response pattern. Indeed, such a task may be too much to ask of this kind of mechanism, in which case the ability of listeners to hear such changes might be accounted for by a cross-channel comparison mechanism.

If it can be shown that local changes in periodicity and degree of excitation overlap are accompanied by changes in the perceived multiplicity of source images, it may be argued that at least some of the information necessary to signal the presence of multiple sources exists at this level of encoding and processing.

### 3.1.2 *Cross-channel information*

To explain some kinds of auditory source imaging, we may need to postulate a cross-channel coherence detection mechanism, where the actual frequency modulation pattern is tracked and a given area of stimulation that is not following the same pattern would be discriminated as such. Such a mechanism would be of a type the Gestalt psychologists called "common fate" (cf. Köhler, 1929), i.e. elements that behave similarly (*coherently*) are more likely to be grouped together than those that behave differently (*incoherently*). This kind of mechanism would be necessarily invoked to explain the detection of incoherence of partials that were too far away from the nearest neighboring partial to create patterns of interaction and

interference in the cochlea and in auditory nerve fiber discharge.

A cross-channel mechanism might perform a kind of cross-correlation on the temporal response patterns of the auditory nerve fiber array. Or (as suggested by Richard Lyon, 1983) a less data-intensive mechanism might cross-correlate the autocorrelation of the cochlear fiber output patterns to select channels with similarly varying periodicities. The grouping of such channels and the extraction of information relative to a given pattern of variation could provide the perceptual system with information concerning the acoustic behavior of a particular source. Thus detection of *coherence* (correlation of frequency behavior) would be a cue for *grouping* of spectral components into auditory images. And detection of *incoherence* (uncorrelated frequency behavior) would be a cue for *separation* of spectral components into auditory images.

There arises the problem of how the auditory system could follow the variation of a frequency component of one source which is very close to frequency components from other sources. There are many facets to this problem. But one aspect that is relevant to Experiment 6 (to follow in this chapter) was addressed by Evans (1978). He reported that recordings of cat auditory nerve fiber responses to simultaneous stimulation by two inharmonically related frequency components showed that both frequencies, within the neurophysiological limits of temporal resolution, were represented in the temporal discharge pattern. Thus, a hypothetical autocorrelator would extract both periods from the fiber's output. And as these varied in frequency, the autocorrelation function would vary accordingly. Some cross-correlation mechanism that was operating on the output of the autocorrelator would have access to the time-varying periodicities of all components stimulating a given fiber. In a sense, this chain of correlation processes could be considered to unshuffle the complex spectrum.

Such a mechanism operating on the autocorrelator output would also be able to group similarly varying groups of frequency components irrespective of their frequency relationships. Accordingly, a cross-channel coherence detector would group coherent inharmonic complexes as well as harmonic complexes. In this case, though, one would still expect within-channel mechanisms to signal irregularity, perhaps confounding the coherence signal given by the cross-channel mechanism. One would also expect that a certain clarity of the autocorrelator output would be necessary for the

cross-correlator to identify similarly varying components, i.e. if the temporal discharge pattern were too noisy it would be difficult to extract a component embedded in the noise.

### 3.2 **EXPERIMENT 6:** Effects of the frequency modulation incoherence, harmonicity and intensity on multiple source perception.

In this experiment several stimulus parameters were varied to investigate the nature of the role played by frequency modulation coherence in auditory source image formation and distinction. Tones where all partials are modulated coherently were compared with tones where one partial was modulated incoherently with respect to the rest. The tones were either harmonic or slightly inharmonic. This allowed a test of the role of overall periodicity in incoherence detection for sustained tones. The number of the partial to be incoherently modulated was varied. This allowed a test of the role of spectral proximity in detection of incoherence, since the excitation patterns in the cochlea of lower partials are further apart than those of higher partials. The overall intensity at which stimuli were presented was varied to provide another way of varying the degree of excitation pattern proximity. And, finally, several different values of frequency modulation width were presented for each stimulus condition to test for sensitivity to incoherence of that particular combination of parameters.

#### 3.2.1 *Stimuli*

Tones were synthesized with 16 equal-amplitude partials. The duration was 1.5 sec with 100 msec raised cosine ramps on the attack and decay.

*Spectral Content:* Two types of spectral content were used: harmonic and inharmonic. The component frequencies and the inter-component distances (in Barks<sup>6</sup>)

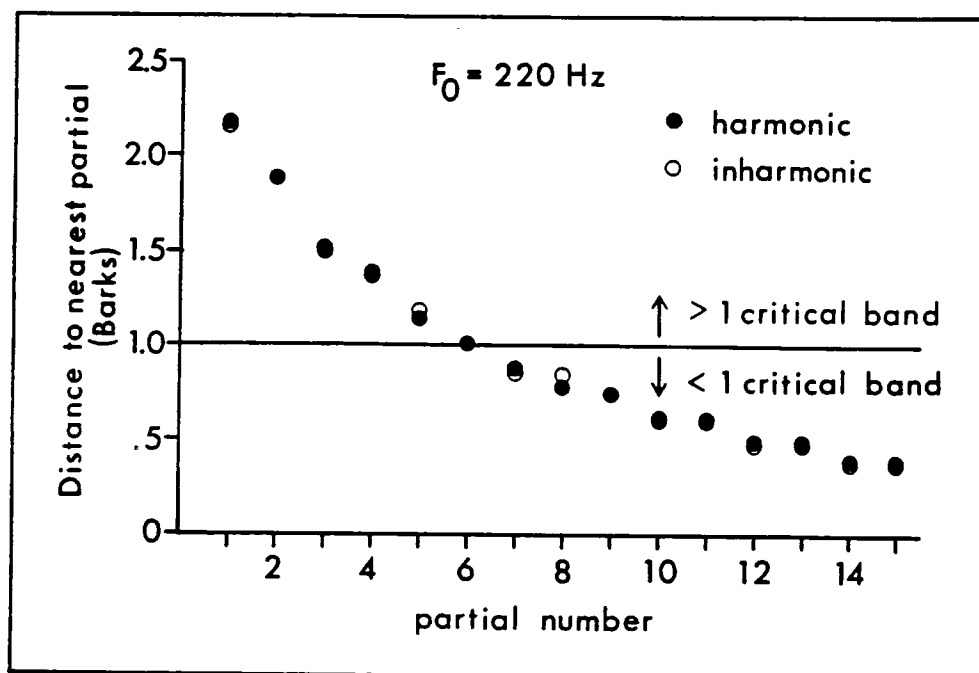
---

6. The Bark is the unit measure of critical band rate. It is meant to describe the frequency scale in terms of a unit range within which frequency components interact to produce perceptual results such as beating, etc. In general, when these components are separated by more than one Bark, such interactions are not reported as producing perceptible results (cf. Scharf, 1970; Zwicker & Terhardt, 1980; though this is contested by Plomp (1976) for perception of the beats of mistuned consonances, sometimes called second-order beats, as mentioned previously).

are listed in Table 3.1.

**TABLE 3.1.** Component frequencies, Bark measures and distance between components in Barks for harmonic and inharmonic stimuli. Bark measures were computed according to the algorithm of Zwicker & Terhardt (1980), implemented at IRCAM by William Hartmann. Maximum displacement from harmonic series ( $f_7$ ): 0.02 Bark. Maximum increase in inter-partial distance ( $f_8 \rightarrow f_9$ ): 0.065 Bark.

Partial Number	Harmonic			Inharmonic		
	Frequency (Hz)	Bark	$\Delta$ Bark	Frequency (Hz)	Bark	$\Delta$ Bark
1	220	2.19		220.43	2.20	
2	440	4.36	2.17	438.98	4.35	2.15
3	660	6.24	1.88	658.24	6.23	1.88
4	880	7.73	1.48	882.50	7.74	1.52
5	1100	9.12	1.39	1098.06	9.11	1.36
6	1320	10.26	1.14	1323.25	10.28	1.17
7	1540	11.27	1.01	1544.44	11.29	1.01
8	1760	12.15	0.88	1755.36	12.13	0.85
9	1980	12.93	0.78	1979.97	12.93	0.85
10	2200	13.67	0.73	2198.11	13.66	0.73
11	2420	14.26	0.59	2424.73	14.27	0.61
12	2640	14.86	0.60	2645.36	14.88	0.60
13	2860	15.34	0.48	2858.62	15.34	0.46
14	3080	15.87	0.52	3087.44	15.88	0.54
15	3300	16.27	0.40	3293.94	16.26	0.38
16	3520	16.71	0.45	3514.65	16.71	0.45



**Figure 3.1.** Plotted here is the distance in Barks to the next nearest partial of a 16-component complex tone. Both harmonic and inharmonic tones are plotted for comparison. The Bark estimates were obtained according to the algorithm of Zwicker & Terhardt (1980).

1. *Harmonic*;  $F_0 = 220$  Hz, center frequencies of all partials were integer multiples of  $F_0$ ;
2. *Inharmonic*; the center frequencies of inharmonic partials differed from the harmonic case by amounts that were selected randomly from a rectangular distribution between  $\pm 5$  cents (see Table 3.1). These slight departures from harmonicity were kept constant for all inharmonic tones in the experiment. The maximum displacement from a harmonic center frequency was 4.99 cents ( $\Delta f / \bar{f} = 0.00289$ ) which is a displacement of 0.02 Bark (i.e. harmonic  $f_7 = 1540$  Hz, inharmonic  $f_7 = 1544.5$  Hz). The maximum increase in inter-partial distance ( $\Delta$ Bark between  $f_8$  and  $f_9$ ) was from 0.783 Bark to 0.848 Bark ( $\Delta$ Bark = 0.065) which is quite small. This yields a spectrum (before modulation) that has roughly the same degree of excitation overlap as the harmonic spectrum when presented at the same intensity. (Figure 3.1 illustrates the

distance in Bark from each partial to the next nearest, usually the next higher, partial.) With this inharmonic signal, however, the periodicity is disturbed considerably.

The waveforms of approximately 7 periods (512 samples) of  $f_1$  of the unmodulated harmonic and inharmonic complexes are presented in Figure 3.2 for comparison. Note the perfect periodicity of the harmonic case.<sup>7</sup> Note also that a vague quasi-periodicity of about the same period as the harmonic waveform can be discerned in the inharmonic waveform.

*Coherent Modulation:* The standard tones were modulated by a pre-determined jitter waveform ( $J_1$ ) whose waveform, spectrum and amplitude probability density function are presented in Appendix B (Figure B.8). This waveform had a mean value of 0, i.e. it is statistically symmetric about 0. The modulation was imposed such as to maintain the original harmonic or inharmonic ratios among the components. As in Chapter 2 this is called "coherent modulation". The resulting signal may be described:

$$S_C(t) = \sum_{n=1}^{18} \sin(2\pi f_n t + \frac{f_n}{f_1} \psi_1 \int_0^t J_1(t') dt'), \quad (3.1)$$

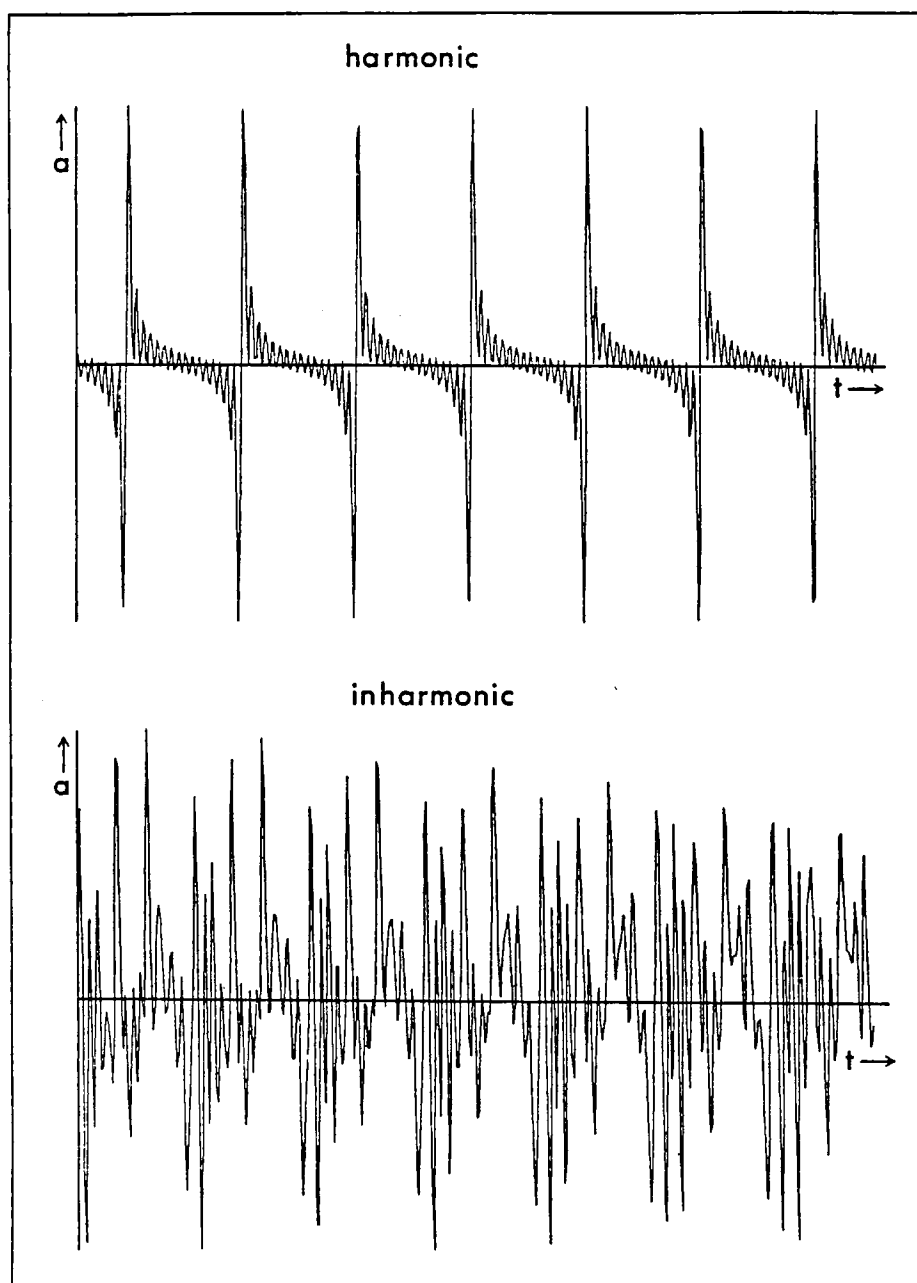
where  $S_C(t)$  is the coherently modulated signal waveform,  $f_n$  are the component center frequencies from Table 3.1, and  $\psi$  is a scalar value chosen to yield a given rms frequency deviation about center frequency with jitter waveform  $J_1$  (see Chapter 2). The factor  $f_n/f_1$  assures the maintenance of frequency ratios, regardless of the harmonicity of the tone complex.

*Incoherent Modulation:* In the case of incoherent modulation, 15 partials were modulated coherently with  $J_1$  and one partial was modulated with  $J_2$ , whose waveform, spectrum and amplitude probability density function are also described in Appendix B (Figure B.9). Note that the spectra of  $J_1$  and  $J_2$  are very similar, but their amplitude probability density functions differ slightly and their waveforms are dissimilar and statistically independent.

---

7. The amplitude variation in the waveform peaks is an artifact of the sampling procedure used in preparing the graphic image.





**Figure 3.2.** Plotted here are 32 msec segments (approximately 7 cycles of  $f_1$ ) from the waveforms of the unmodulated harmonic and inharmonic tones used in Experiment 6.

The equation describing the "incoherent" signal is:

$$S_I(t) = \sin\left(2\pi f_k t + \frac{f_k}{f_1} \psi_2 \int_0^t J_2(t') dt'\right) + \sum_{\substack{n=1 \\ n \neq k}}^{16} \sin\left(2\pi f_n t + \frac{f_n}{f_1} \psi_1 \int_0^t J_1(t') dt'\right)$$

for  $k = 1, 3, 5, 7, 9, 11, 13, 15,$  (3.2)

where  $f_k$  is the frequency of the partial to be modulated incoherently and  $f_n$  are the frequencies from Table 3.1, excluding  $f_k$ . The factor  $f_k/f_1$  assures that the center frequency of  $f_k$  is in the desired relation to  $f_1$ . The values of  $\psi$  are chosen separately for each of the modulating waveforms so that the rms deviations are the same.

**TABLE 3.2.** The 5 rms deviations of modulation used for harmonic and inharmonic stimuli. The number of the incoherently modulated partial is indicated in the column to the left.

Partial Number	Rms Deviation (cents)									
	Harmonic					Inharmonic				
	1	2	3	4	5	1	2	3	4	5
1	2.00	5.00	8.00	11.00	14.00	3.0	6.0	9.0	12.0	15.0
3	0.50	2.00	3.50	5.00	6.50	2.0	4.0	6.0	8.0	10.0
5	0.30	1.50	2.70	3.90	5.10	0.5	1.3	2.1	2.9	3.7
7	0.30	1.50	2.70	3.90	5.10	0.5	2.0	3.5	5.0	6.5
9	0.30	1.20	2.10	3.00	3.90	0.5	2.5	4.5	6.5	8.5
11	0.05	0.30	0.55	0.80	1.05	0.5	2.5	4.5	6.5	8.5
13	0.05	0.30	0.55	0.80	1.05	0.5	2.5	4.5	6.5	8.5
15	0.05	0.30	0.55	0.80	1.05	0.5	2.5	4.5	6.5	8.5

In this experiment the odd partials were selected for incoherent modulation and 5 values of rms deviation were chosen for each partial number (see Table 3.2). These 5 values were chosen from pilot listenings by the experimenter in order to give a range of responses from no perceptual difference to a clearly audible effect for each partial number and for each of the harmonic and inharmonic stimuli. Values ranged from 0.05 - 1.05 cents for high harmonic partials to 3 - 15 cents for  $f_1$  of the inharmonic complex. 40 incoherently modulated stimuli were synthesized for each of the harmonic and inharmonic complexes (8 partial numbers  $\times$  5 rms deviations). The odd partials were chosen to avoid pitch confusion effects at the octaves of the  $F_0$  in the harmonic stimuli (though these still exist for an incoherent  $f_1$ ).

### 3.2.2 Method

Each 2IFC trial contained a coherent tone and an incoherent tone presented in counterbalanced order. The coherent tone had the same rms deviation as the incoherent tone. The tones were separated by a 500 msec silent interval. The observation intervals were marked by differently colored lights on a 2-button box. The subject's task was to decide which tone seemed to have more sound sources in it. No feedback was given after the response. As soon as the subject pressed a button, the computer paused for 500 msec and then presented the next stimulus pair.

There were three main conditions, each presented in separate experimental blocks:

1. harmonic stimuli presented at 75 dbA (H75),
2. harmonic stimuli presented at 50 dbA (H50), and
3. inharmonic stimuli presented at 75 dbA (I75).

These blocks consisted of 10 random series of the 40 stimulus pairs, so 400 comparisons were made per block. Each of the 3 blocks was presented 3 times on separate occasions giving 30 judgments per stimulus pair.

As there were a variety of perceptual effects for the different incoherent partial numbers, the subjects were played the range of possible stimuli before the experiment began in order to demonstrate the range of possible percepts. These varied from a phase rolling or "chorus" effect<sup>8</sup> (partials 7 - 15) to the clear emergence of a pitched sinusoid (1 - 5) and even little melodies on 2 or 3 partials (3 - 7) or an arrhythmic pulsing of auditory roughness (5 - 15)<sup>9</sup>. For this reason and due to the length of the experiment, stimuli were sub-blocked into 4 groups by partial number (1,3; 5,7; 9,11; 13,15) and the presentation order of these sub-blocks was

8. The "chorus" effect is a sensation of many sound sources of the same kind playing at the same pitch as one obtains with many players (violins, for example) trying their best to play in unison.
9. These were the perceptual effects for the small range of rms deviations used in the experiment. If larger deviations are used, e.g. 50 - 85 cents (or 3 - 5 % variation in frequency), the pitch of the incoherent partial is audible up to and including the 16<sup>th</sup>!

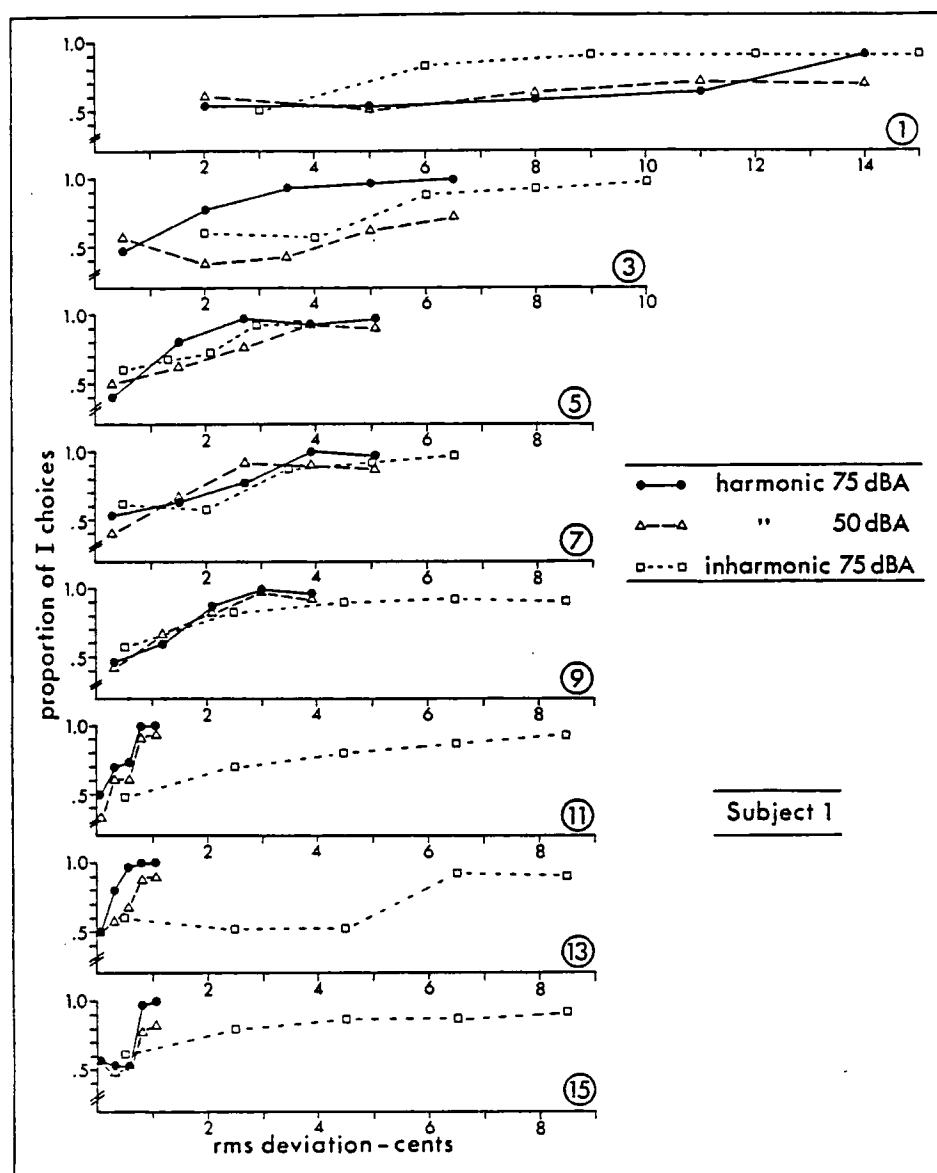
randomized within a block. This allowed subjects to rest between the sub-blocks and to adopt a minimum number of criteria for making the source multiplicity judgments within a group of trials.

Ten subjects were initially tested in the experiment, though only 4 obtained better than random performance on the H75 condition with the rms deviations selected. Subjects were paid for their participation. The others were not continued in the experiment since the other two conditions were even more difficult. Three of the 4 remaining subjects completed all conditions. One subject (S2) completed the 2 harmonic conditions and only 3 of the 8 partial numbers of the inharmonic condition. Ss 1 (the experimenter), 3 and 4 were professional psychoacousticians. S2 was a professional musician and composer.

### 3.2.3 Results

The individual data for the 4 subjects are listed in Tables E.5.1 - E.5.3 (Appendix E) for H75, H50 and I75 conditions, respectively, and are plotted in Figures 3.3 - 3.6 to compare across intensity and harmonicity conditions for each subject. A separate graph is plotted for each partial number. The three curves in each graph represent the data for H75, H50 and I75 conditions. The ordinate represents the percentage of times in 30 presentations of a given stimulus that the tone with the incoherent modulation was chosen as having more sources. The abscissa represents the rms deviation of the modulation. The subjects' response behaviors are quite similar though there are some larger variances in the data for certain conditions (e.g. H75 for partials 1 & 3; H50 for partial 3; I75 for partials 5 & 13). To express graphically the general tendencies across subjects the means were computed for all conditions (listed in Tables E.5.1 - E.5.3) and plotted in Figure 3.7.

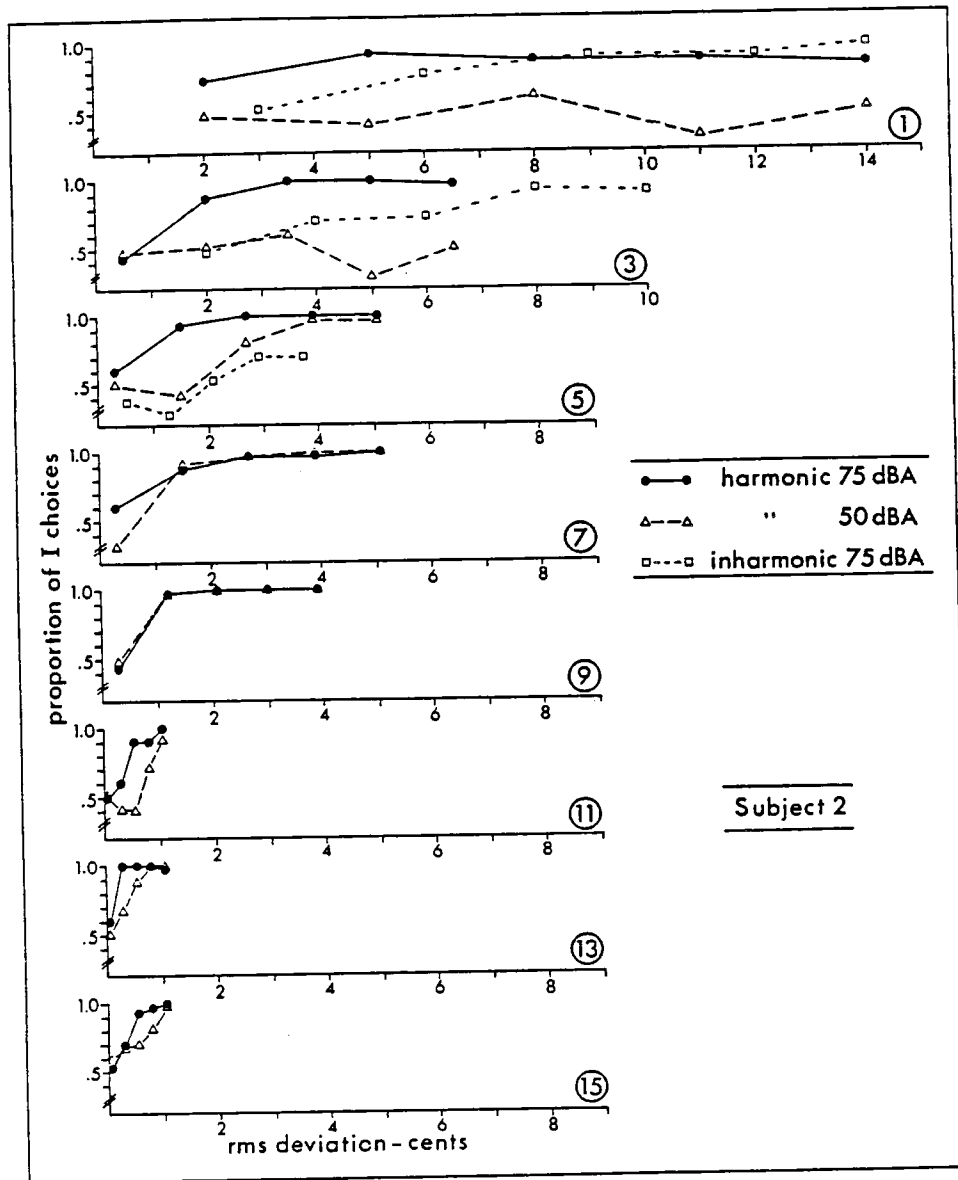
If we draw a smooth curve (cubic spline) through the data points and choose the rms deviation corresponding to 71% choice, we can consider this as a measure of the modulation width necessary to just barely create a perceptual change that subjects judge as indicating multiple sources. This will be called the "source multiplicity threshold" (SMT). Conditions not reaching 71% choice at the largest rms deviation (or having greater than 71% choice at the smallest deviation) are plotted as the maximum (or minimum) deviation used in that condition and are tagged with an arrow indicating that the SMT is higher (or lower) than this value. There are 4 instances (all



**Figure 3.3.** Experiment 6 data summary for Subject 1. The proportion of incoherent tone choices is plotted as a function of rms deviation of modulation (in cents). Each graph represents the data for one incoherently modulated partial (encircled number to right of graph). Shown in each graph are the curves for H75, H50 and I75 conditions (see Key). Each data point represents 30 2IFC comparisons.

in the data of S4) where the curves are non-monotonic and cross the 71% point two or

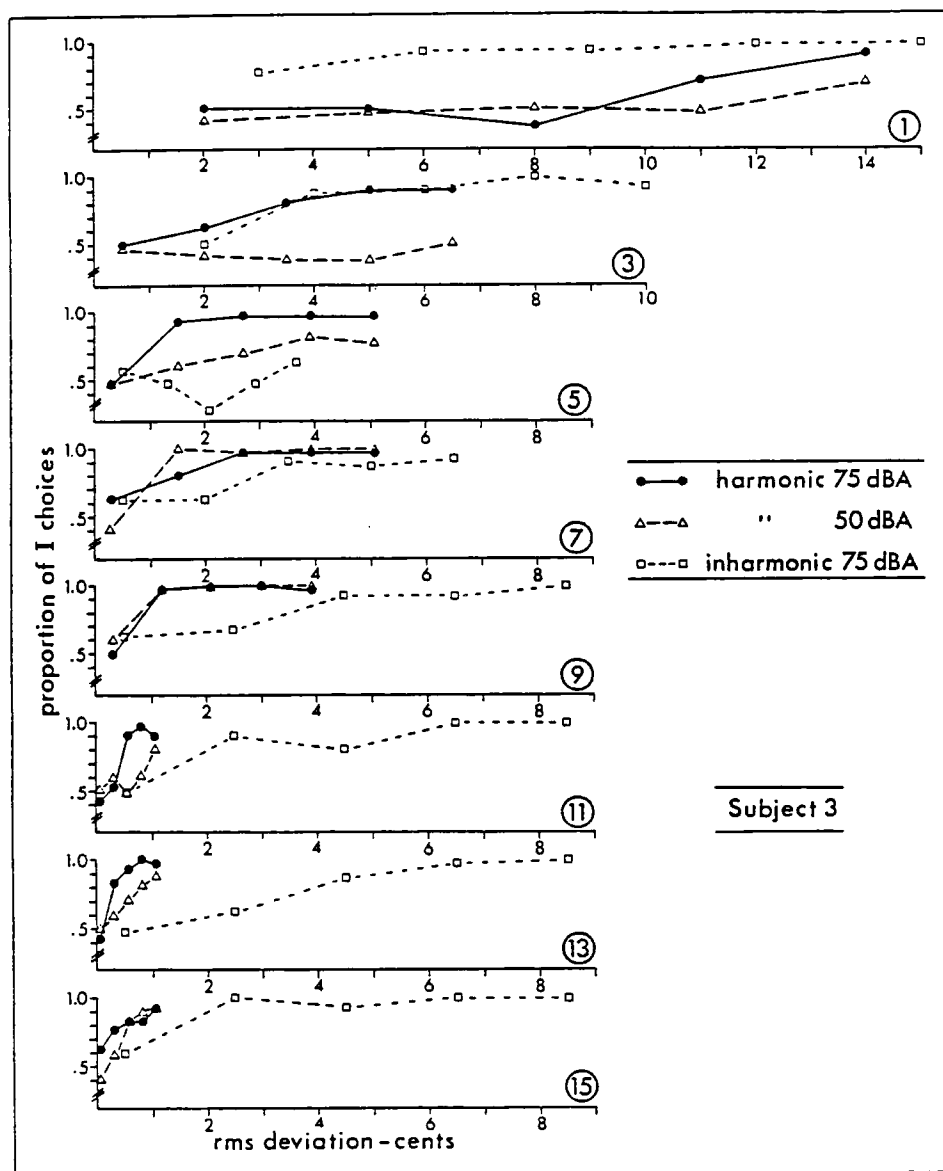
more times. If the curve increased above 71% and then turned down, it was considered that threshold was not reached (H75, H50 for  $f_3$ ; H50 for  $f_5$ ). In cases where



**Figure 3.4.** Experiment 6 data summary for Subject 2. This subject did not complete the I75 condition for partial numbers 7 – 15. (See caption for Fig. 3.3.)

threshold was passed twice, the rms deviation value at the highest positive-going 71%

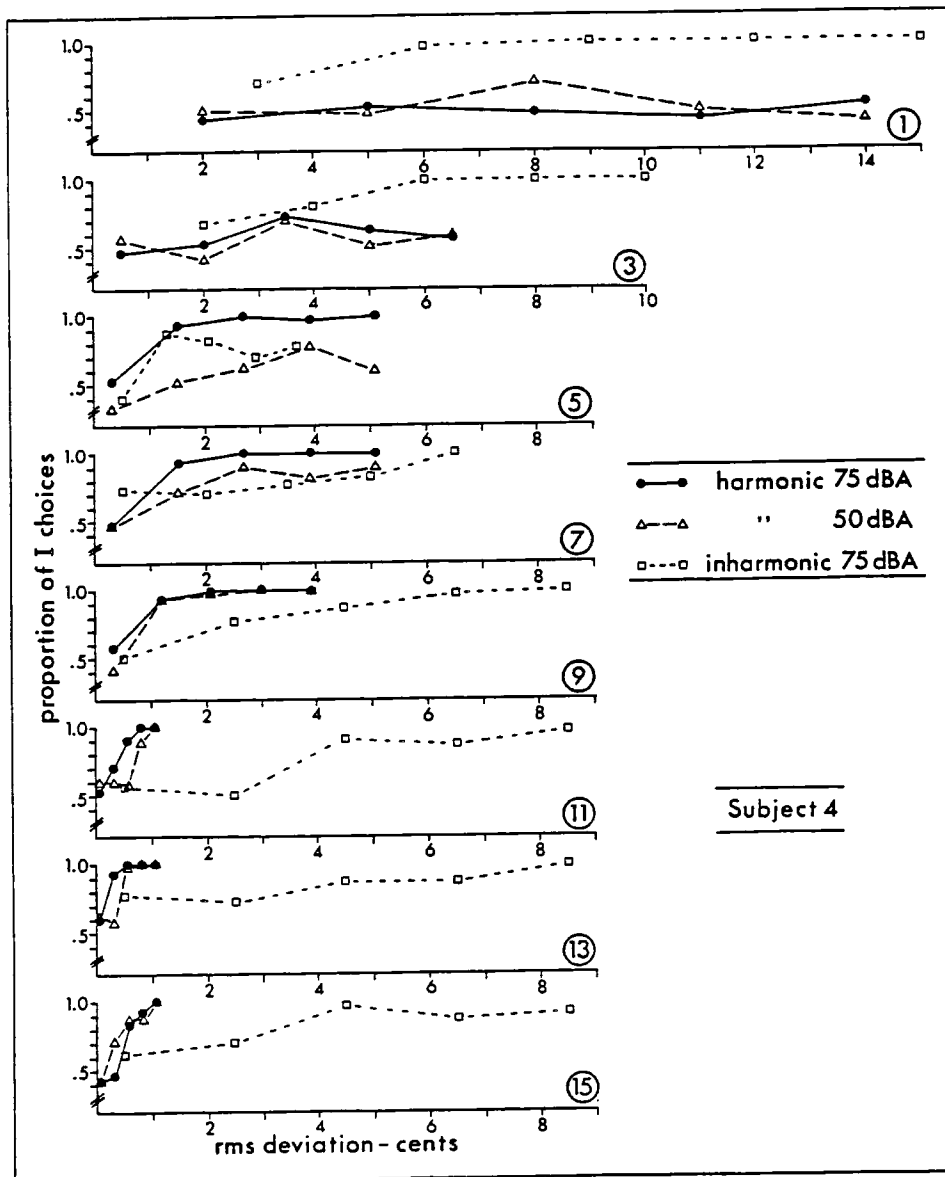
crossing was chosen as the SMT (I75 for  $f_5$ ).



**Figure 3.5.** Experiment 6 data summary for Subject 3. (See caption for Fig. 3.3.)

The individual SMTs for all Ss are listed in Table 3.3 and are plotted in Figures 3.8 - 3.11 to show comparisons of the SMTs for H75 vs. H50 and H75 vs. I75. The hashed areas in these figures are designed to make more visible the regions where H50 SMTs

or I75 SMTs are greater than H75 SMTs. The data are replotted in Figure 3.12 in order to compare the SMT curves across subjects for each condition.

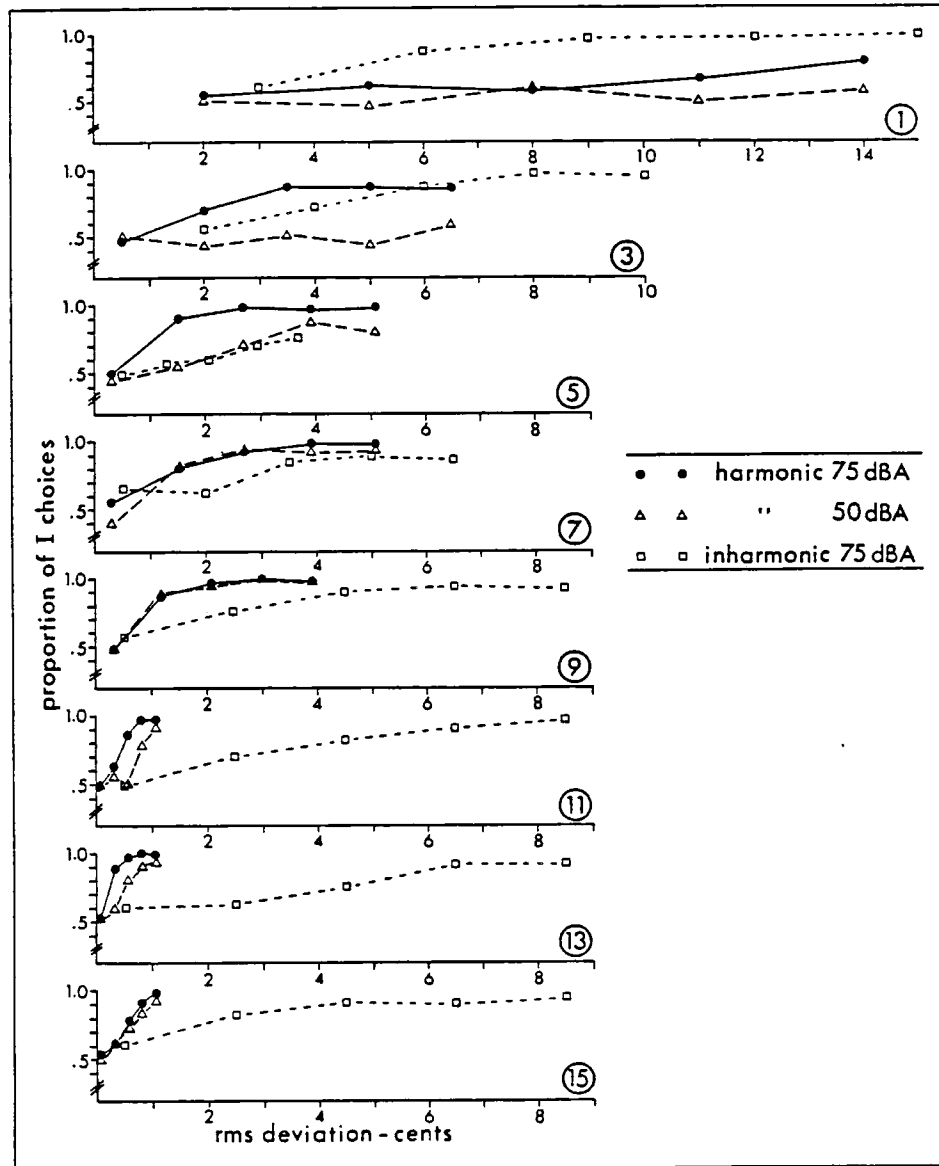


**Figure 3.6.** Experiment 6 data summary for Subject 4. (See caption for Fig. 3.3.)

The group SMTs, extracted from the mean data across subjects (Figure 3.7), are listed in Table 3.4. To get the group SMTs across subjects, the data values at each



rms deviation for each condition were averaged (see Tables E.5.1 - E.5.3, App. E). Then a cubic spline was fitted to the five averaged values and the 71% point determined. These values are expressed in Table 3.4 as cents,  $\Delta f / f$ , and as the time



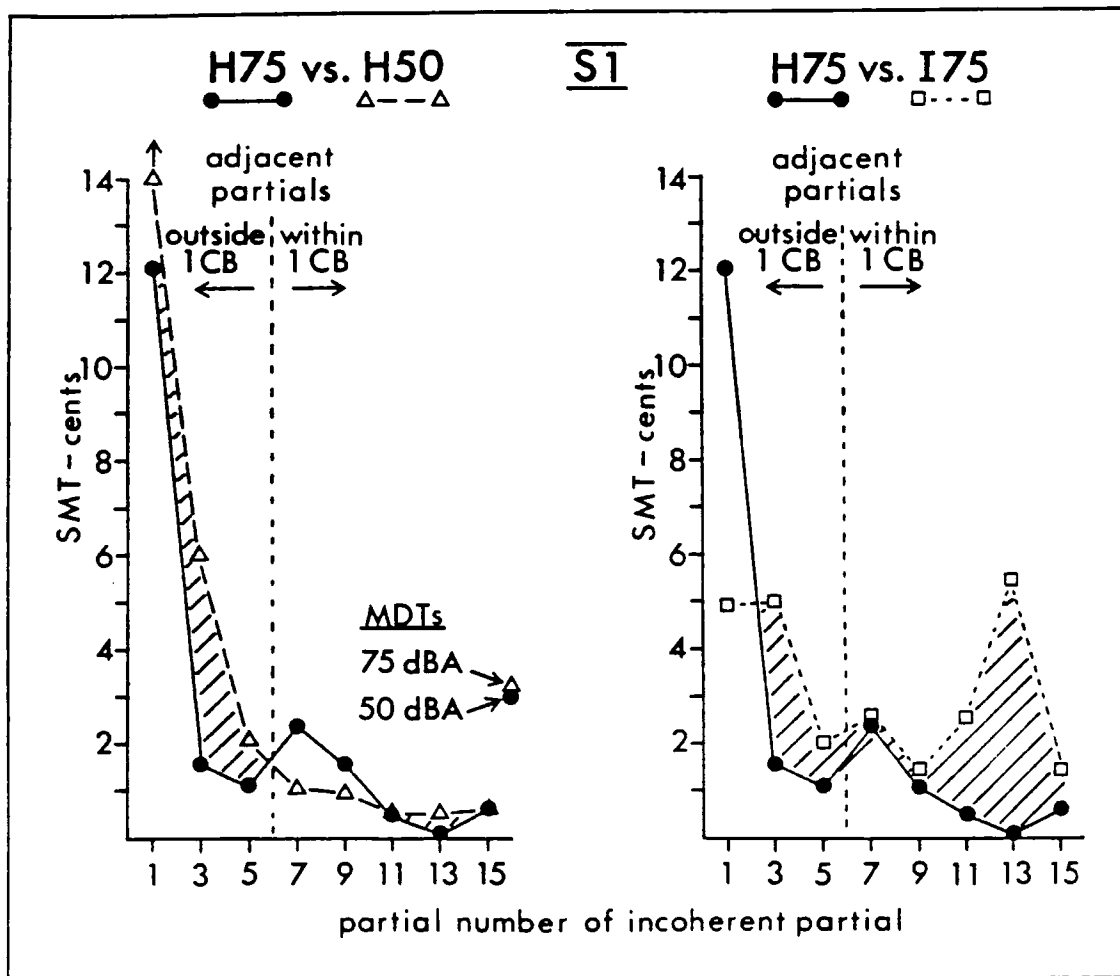
**Figure 3.7.** Mean data for Experiment 6 averaged across 4 subjects. For partials 7 - 15 of the I75 condition, the means are across 3 subjects. (See caption for Fig. 3.3.)

difference,  $\Delta P$ , between the period of  $f$  and that of  $f + \Delta f_{rms}$ :

$$\Delta P = \frac{1}{f_n} - \frac{1}{f_n + \Delta f_{rms}} \quad (3.3)$$

**TABLE 3.3.** Experiment 6 data summary. Source multiplicity thresholds for individual subjects measured from 71% points on cubic spline curves fitted to data points. For curves which started at greater than 71% or never reached that point the smallest or largest values presented in the experiment are listed, respectively.

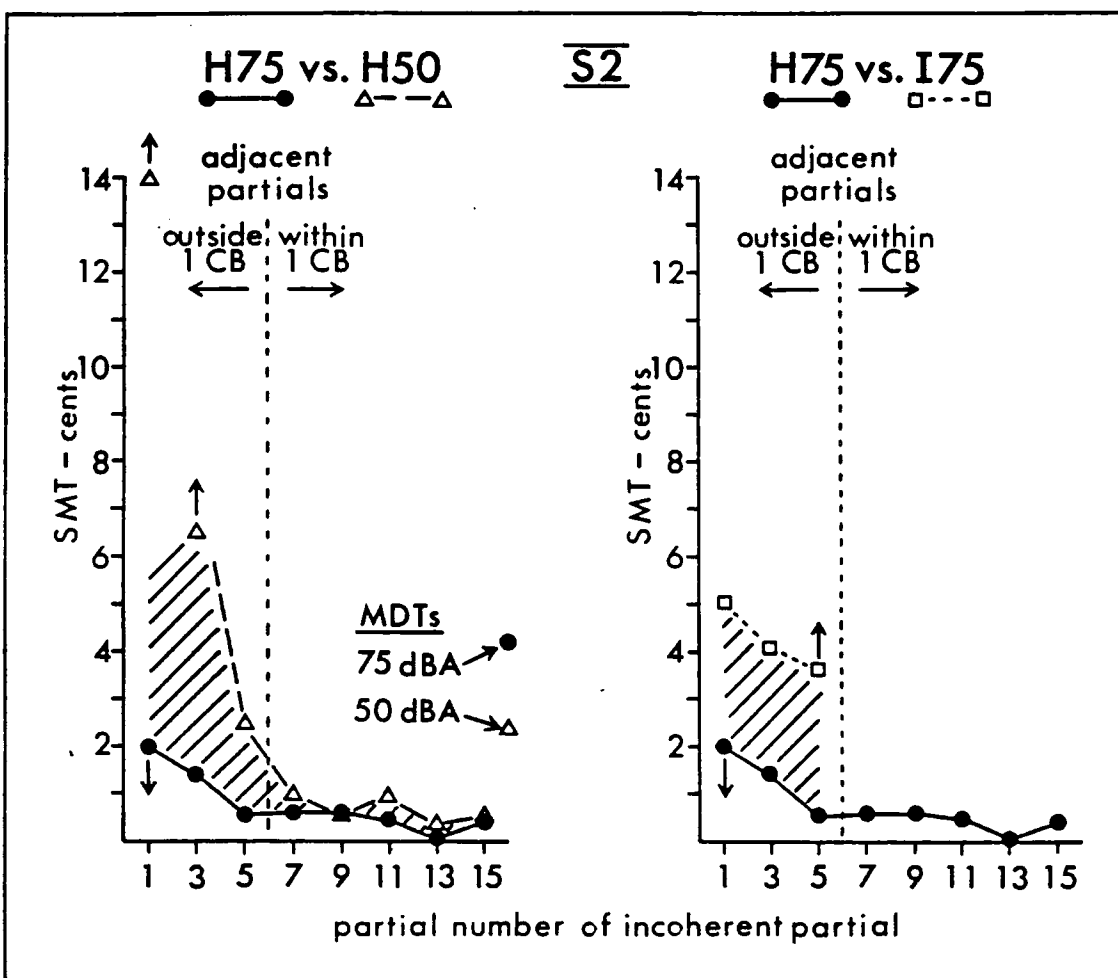
Incoherent Partial	Subject 1			Subject 2		
	Stimulus Condition			Stimulus Condition		
	H75	H50	I75	H75	H50	I75
1	12.1	> 14.0	4.8	< 2.0	> 14.0	5.1
3	1.6	6.0	5.0	1.3	> 6.5	4.2
5	1.2	2.2	2.0	0.6	2.4	> 3.7
7	2.3	1.6	2.7	0.7	0.9	—
9	1.6	1.4	1.4	0.7	0.6	—
11	0.5	0.6	2.6	0.4	0.8	—
13	0.2	0.6	5.4	0.1	0.3	—
15	0.7	0.7	1.3	0.3	0.6	—
Incoherent Partial	Subject 3			Subject 4		
	Stimulus Condition			Stimulus Condition		
	H75	H50	I75	H75	H50	I75
1	11.1	> 14.0	< 3.0	> 14.0	> 14.0	3.1
3	2.7	> 6.5	3.0	> 6.5	> 6.5	2.7
5	0.8	2.8	> 3.7	0.8	> 5.10	3.0
7	0.9	0.8	2.5	0.8	1.4	2.3
9	0.6	0.5	2.8	0.6	0.7	1.9
11	0.4	0.9	1.3	0.3	0.7	3.5
13	0.2	0.5	3.1	0.1	0.4	< 0.5
15	0.2	0.3	0.9	0.5	0.3	2.6



**Figure 3.8.** Source multiplicity thresholds (SMTs) for subject 1. The SMT (in cents rms deviation) is plotted as a function of the partial number of the frequency component receiving incoherent modulation. H75 vs. H50 is plotted on the left to see the effect on SMT of intensity difference. H75 vs. I75 is plotted on the right to see the effect of difference in harmonic of the center frequencies of the partials. Also indicated for comparison are this subject's modulation detection thresholds (MDTs) for a 16-harmonic tone at 75 and at 50 dBA (from Experiment 10, Appendix D).

3.2.3.1 *Effects of rms deviation of modulation*

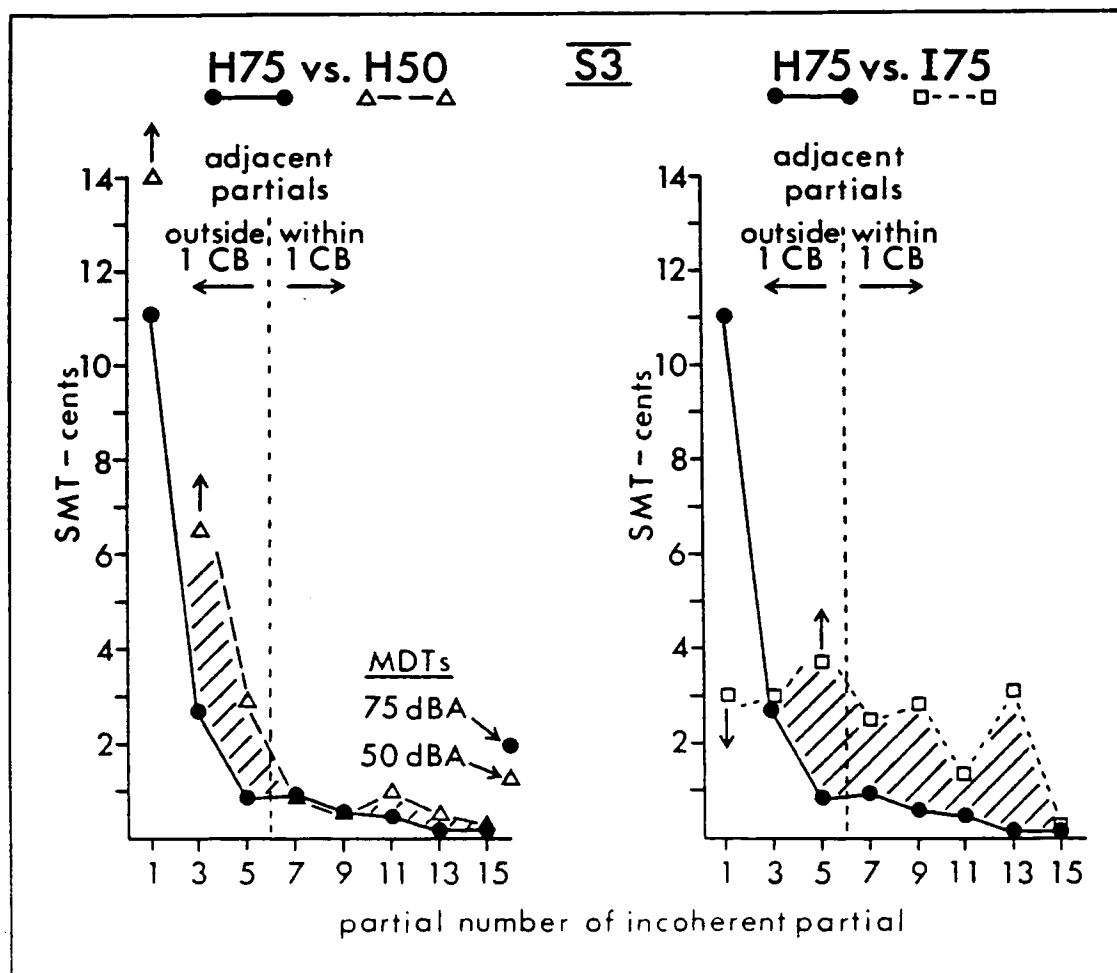
Almost all curves in Figs. 3.3 - 3.7 are approximately monotone ascending: as the rms deviation increased for a given stimulus configuration, Ss more often chose the incoherent tone as having more sources. There are two exceptions to this generalization. Some curves never really depart from a fluctuation around random performance. For these cases, it is probable that the subject never discerned an effect interpretable as the presence of multiple sources.



**Figure 3.9.** SMTs and complex tone MDTs for subject 2. (See caption for Fig. 3.8.)

Also, 4 curves for S4 were non-monotonic. Three of these were shaped like an

inverted U, and one actually increased, decreased and then increased again. These probably reflect a degree of uncertainty as to the judgement being made. But as can be seen in the mean curves (Fig. 3.7) the general trend is to have an increasing proportion of incoherent tone choices with increasing rms deviation.



**Figure 3.10.** SMTs and complex tone MDTs for subject 3. (See caption for Fig. 3.8.)

### 3.2.3.2 Effect of the number of the partial being modulated incoherently

For harmonic stimuli, there is a more rapid rise in the curves (Figs. 3.3 - 3.7) for higher partial numbers. This suggests that incoherence is judged as indicating

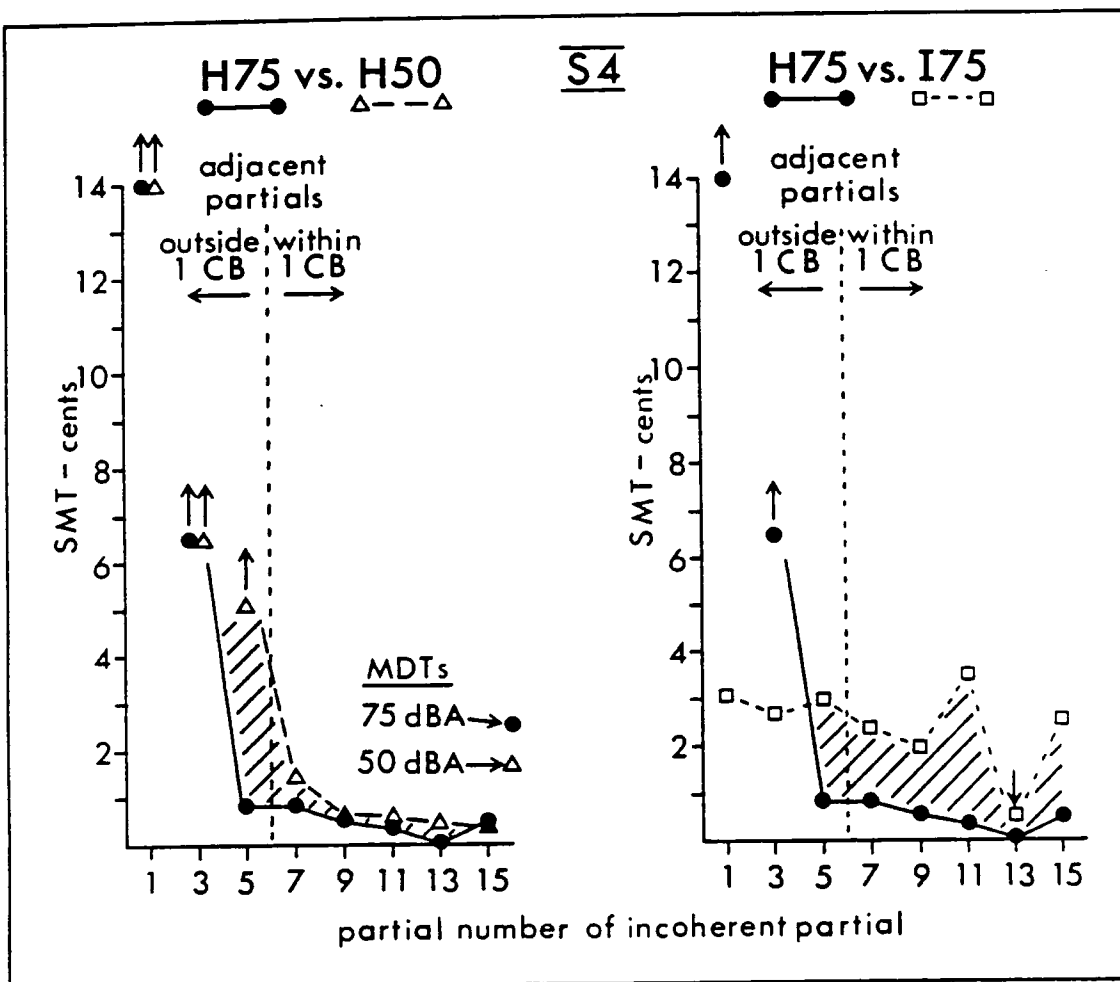


Figure 3.11. SMTs and complex tone MDTs for subject 4. (See caption for Fig. 3.8.)

more sources at lower rms deviations for higher partials than for lower partials, i.e. less deviation is necessary to create the effect.<sup>10</sup> This trend is reflected in Figs. 3.8 -

10. This is generally true for all subjects though S2 has a much more rapidly rising curve for incoherent  $f_1$  in H75 than all of the other subjects (by a

3.12 as a decrease in SMT with partial number. Note that the SMT falls very rapidly with partial number for partials below the 5<sup>th</sup> and then declines more gradually. The behavior of these curves is very similar above  $f_5$  for H75 and H50 stimuli. It is instructive to note that for these stimuli, all partials above  $f_5$  have at least one neighboring partial within one critical bandwidth (CB). Once inside this CB distance the SMT curve (and slope of the data curve) begins to approach an asymptote. However, it is important to remark that (except for S1) there is no break or discontinuity in the SMT curves at the single-CB border. The SMT curve is a relatively smooth function of component proximity.

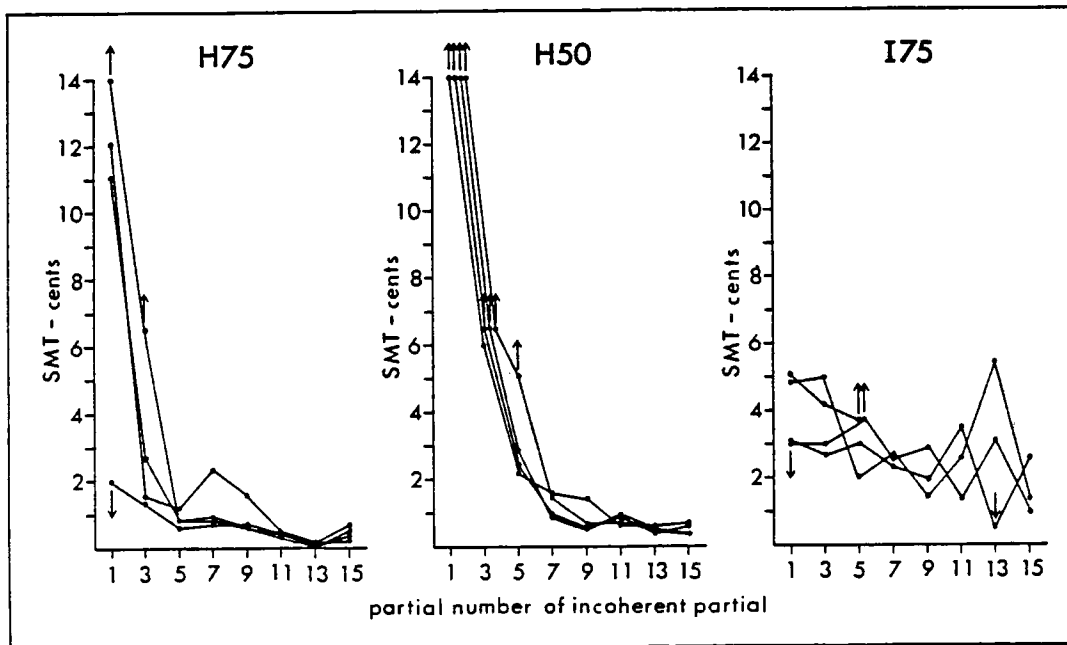
There are some notable differences between subjects in the effect of partial number. There is, in the SMT curve for S1 (Fig. 3.8), a sizeable bump (increased SMTs) at  $f_7$  and  $f_9$  for the H75 condition. This is less noticeable in H50, though this subject's SMTs are still higher than those of the other subjects. Also, S4's SMT curves (Fig. 3.11) do not descend as rapidly as other subjects due to his higher SMTs for the lower partials.

**TABLE 3.4.** Group SMTs across subjects expressed as  $\text{cents}_{rms}$ ,  $\Delta f_{rms}/\bar{f}$  and  $\Delta P_{rms}$  (change in period of incoherent component that is just noticeable as yielding multiple sources, see Eq. 3.3).

Incoherent Partial	$\text{cents}_{rms}$			$\frac{\Delta f_{rms}}{\bar{f}} (\times 10^{-3})$			$\Delta P_{rms} (\mu\text{sec})$		
	Stimulus Condition			Stimulus Condition			Stimulus Condition		
	H75	H50	I75	H75	H50	I75	H75	H50	I75
1	12.1	> 14.0	3.9	7.01	—	2.26	31.66	—	10.23
3	2.0	> 6.5	3.7	1.16	—	2.14	1.75	—	3.23
5	0.8	2.6	3.0	0.46	1.50	1.73	0.42	1.36	1.57
7	0.9	1.1	2.6	0.52	0.64	1.50	0.34	0.41	0.97
9	0.7	0.7	1.9	0.40	0.40	1.10	0.20	0.20	0.55
11	0.4	0.7	2.6	0.23	0.40	1.50	0.09	0.17	0.62
13	0.2	0.4	3.8	0.12	0.23	2.20	0.04	0.08	0.77
15	0.4	0.5	1.2	0.23	0.29	0.69	0.07	0.09	0.21

factor of at least 6).

In contrast to the harmonic stimuli the behavior of the data for inharmonic stimuli with respect to partial number are very erratic (see Fig. 3.12). There is a slight tendency for data curve slopes to increase, and for SMTs to decrease with partial number up to about  $f_9$ . But beyond this point, the data are wildly unsystematic for each subject as well as across subjects. This reflects the report of all subjects that the judgments were very difficult to make on these stimuli, since the inharmonicity gave an impression of multiplicity even at very small modulations. The task became one of discerning some changing pattern in one of the tones that could be interpreted as a differently behaving source.



**Figure 3.12.** Individual SMTs for 4 subjects plotted as a function of partial number. Each graph represents a different experimental condition as labeled.



### 3.2.3.3 *Effect of intensity*

The effect of intensity is most pronounced at low harmonic numbers which are greater than one CB from their neighbors. In the data curves (Fig. 3.3 - 3.7), the H50 curve is almost always below the H75 curve; the rise in proportion of incoherent tone choices with rms deviation is slower. Again, this is reflected in higher SMTs in Figs. 3.8 - 3.11. Note the hashed areas in those figures which represent the regions within which H75 stimuli have lower SMTs than H50 stimuli. For most subjects this difference becomes very small at  $f_7$  and  $f_9$  (and even reverses for S1 due to the bump in the H75 curve). A small difference ( $<0.5$  cents) reappears for higher partial numbers. For the higher harmonics ( $>f_8$ ) we would expect less of an effect of intensity (at the intensity values used here) due to the heavy degree of excitation overlap already present in H50. If a much greater intensity difference were used, we might see this effect extend into some of the higher harmonics. The disappearance of the effect at harmonics 7 and 9 is puzzling and no explanation is immediately apparent, except to note that the inter-partial distances here border on one Bark. It may be that there is a change in the criteria and cues used for detecting incoherence at these proximities.

### 3.2.3.4 *Effect of harmonicity (phase synchrony of adjacent partials)*

The initial harmonicity of a tone complex has a profound effect on subjects' ability to detect incoherence and interpret that as indicating multiple sources. As mentioned before, the effect of the partial number is much less systematic in the inharmonic condition than with harmonic tones. Subjects are also much less consistent with respect to each other with inharmonic stimuli than with the harmonic stimuli. The data curves (Figs. 3.3 - 3.7) for I75 stimuli are almost always below those for H75 stimuli, indicating that a greater rms deviation of modulation is necessary in the former to detect a difference in source multiplicity. A notable exception is for  $f_1$ , where the I75 curves are generally above those for H75. For Ss 1, 3 and 4, the SMT for  $f_1$  is much lower for I75 than for H75. S2 had an SMT for the I75  $f_1$  that was similar to that of the other subjects, but had an unusually low SMT for the H75  $f_1$ , so the placement of these curves is reversed with respect to the other subjects.

### 3.2.4 Discussion

All of the stimulus parameters had an effect on judgments of source multiplicity. Under certain conditions incoherent FM on one partial of a 16-component tone generates a perceptual effect that listeners can judge as indicating the presence of multiple sources. This occurs at values of rms deviation of the modulation that are dependent on the partial number of the component being jittered. Generally, as the partial number is increased, the rms deviation necessary to generate the effect 71% of the time decreases. This dependence of source multiplicity judgments on modulation width indicates that a certain proximity of excitation (however fleeting) of the incoherent components is necessary to create enough of an interaction to be detectable, since these components will move closer together at larger modulation widths.

The amount of modulation creating a perceptual difference between coherently and incoherently modulated tones also depends on the partial number of the incoherent component. As the number of the incoherent partial increases, there is a decrease in the distance from one area of excitation to the excitation area on the basilar membrane stimulated by the next nearest partial. Correlated with this decrease in distance is a decrease in the SMT. This can be taken to be a measure of the smallest amount of incoherent modulation necessary to generate a multiple source perception. When this value is very small, it means that the excitation patterns of adjacent stimuli are already very close and the smallest incoherence in their modulation patterns generates enough of an aperiodicity in the stimulation of auditory fibers in that region to create the effect. As distances between partials become greater, SMTs become larger since a greater amount of modulation is necessary to move the respective regions of excitation by adjacent components into the same area.<sup>11</sup> This is supported by the highly significant correlation between the group SMTs and the distance in Barks to the next nearest partial for harmonic stimuli (H75  $r = .88$ , H50  $r = .96$ ).

---

11. A notable exception to the degree of effect of partial number is found in S2's data. Although his SMTs still decreased with increasing partial number, he had very low SMTs for  $f_1$  and  $f_3$  of H75 compared to the rest of the subjects. This is puzzling in view of the fact that his SMTs for H50 and I75 stimuli at these partials were similar to those of the other subjects. I find no apparent explanation for this result.

The effect of partial number is not as clear for the inharmonic stimuli. The correlation between group SMTs and the distance to the nearest partial (in Barks) was not significantly different from zero. Also, the I75 SMT curve is non-monotonic for all subjects who completed this portion of the experiment. Except for  $f_1$ , the SMTs for the partials of inharmonic stimuli are, on the average, 4 to 5 times greater than the SMTs of harmonic partials. This result is, then, an additional indication that proximity of incoherent partial excitation patterns plays an important role in these effects (at least for harmonic tones).

For harmonic stimuli the overall effect is qualitatively similar for intensities of 50 and 75 dBA. However, the SMTs are considerably higher for 50 dB stimuli when the "resolved" harmonics, lower than the 6<sup>th</sup>, are jittered incoherently. By decreasing the intensity of the tone complex, the extent of excitation due to a given partial is reduced and the distance between areas of maximum excitation on the basilar membrane is increased. To get a similar degree of perceptual effect at the lower intensity, a greater rms deviation is required. This is primarily true for harmonics lower than the 6<sup>th</sup> which are presumably outside of a critical band. There is a substantial difference for partials 1, 3 and 5 between the SMTs for H75 and H50. This difference practically disappears once the nearest partial is less than a critical bandwidth distant. This may mean that once the excitation patterns are *already* overlapping at the lesser intensity, a further increase in overlap does not play a strong role.

There is a possibility of confounding effects of the loudness of individual partials in the two intensity conditions. In Table 3.5, the partial loudnesses (in phons), calculated according to the procedure of Zwicker (1960) are listed. With a decrease of 25 dB in the overall rms amplitude, there is an increase in the slope of the loudness by partial function by a factor of approximately 1.5. A marked decrease in relative loudness of the lower harmonics between the two intensity conditions is also evident ( -28 phons for harmonics 14, 15, 16 compared to -30.5 phons for  $f_3$  and  $f_5$ , and -35 phons for  $f_1$ ). This may increase the difficulty of the task, particularly in the case of the fundamental. For the harmonic and inharmonic stimuli at 75 dBA with  $f_1$  incoherent, the most salient difference between the coherent and incoherent tones is that the  $f_1$  seems to stand out and separate from the rest of the complex in the incoherent tone. If in H50 stimuli, the independently modulating  $f_1$  is much weaker than the pitch at the "virtual"  $F_0$  due to the rest of the harmonics, it may be very difficult to detect and thus may not be heard as a separate source.

**TABLE 3.5.** Loudnesses (in phons) of individual partials for harmonic tones. Partial loudnesses calculated according to procedure of Zwicker (1960), implemented at IRCAM by Clarence Barlow.

Partial Number	Overall Rms Amplitude	
	75 dbA	50 dbA
1	49.29	14.33
2	50.52	18.07
3	50.94	19.35
4	51.21	20.50
5	51.54	21.05
6	51.86	21.58
7	52.30	22.57
8	52.97	23.50
9	53.71	24.38
10	54.58	25.60
11	55.55	27.10
12	56.51	28.42
13	57.28	29.36
14	57.80	29.97
15	57.97	30.25
16	57.75	29.97

However, the partial loudness values in Table 3.5 indicate that the lower harmonics should still be above auditory threshold even at 50 dbA overall amplitude. And informal listening indicates that  $f_1$  and  $f_3$  can be separately resolved at 50 dbA presentation at low modulation widths, and further, that they can be discerned as modulating incoherently if the modulation width is large enough. So I would conclude that the effect of intensity is more related to excitation proximity than to relative loudness differences between the partials.

Another element that seems to have a strong effect on the SMT is the overall periodicity. As mentioned previously, the I75 SMTs are 4 to 5 times greater than those for H75 stimuli. If the judgments are being made on the basis of an irregularity in the temporal patterns of firings in some nerve fibers compared to the regularity found in others, then it would make sense that perturbing the general periodicity would make such a decision more difficult. This result points to the role played by periodicity or phase synchrony of spectral components in judgments on source multiplicity.

**TABLE 3.6.** Summary of subjects' descriptions of perceived effects of introducing an incoherently modulating partial into a complex tone, for the 3 main conditions (H75, H50, I75). Note that these effects occur at modulation widths just above the source multiplicity threshold within the range of modulation widths used in the experiment.

Partial Number	H75	H50	I75
1	$f_1$ slightly audible as separately modulating.	no effect of incoherence	$f_1$ audible as separately modulating spectral pitch
3	$f_3$ audible	$f_3$ slightly audible for one subject	$f_3$ audible
5	irregular alternation between pitches of $f_4$ and $f_5$ (sometimes $f_3$ )	alternation between pitches of $f_4$ and $f_5$	$f_5$ audible; rhythmic pattern of roughness heard at $f_5$
7	rhythmic roughness around $f_7$ ; weak "chorus" effect	alternation between pitches of $f_6$ and $f_7$ ; rhythmic roughness around $f_7$	$f_7$ only slightly audible; rhythmic roughness around $f_7$
9	rhythmic roughness around $f_9$ ; weak "chorus" effect	rhythmic roughness around $f_9$ ; difficult to discern pitch; weak "chorus" effect	rhythmic roughness around $f_9$
11	rhythmic roughness around $f_{11}$ ; weak "chorus" effect	rhythmic roughness around $f_{11}$ ; weak "chorus" effect	high frequency roughness around $f_{11}$ ; roughness relatively continuous
13	rhythmic roughness around $f_{13}$ ; "chorus" effect stronger	rhythmic roughness around $f_{13}$ ; "chorus" effect stronger	continuous roughness around $f_{13}$
15	strong "chorus" effect; roughness around $f_{15}$ ; rhythmic pattern barely audible	strong "chorus" effect; continuous roughness around $f_{15}$	continuous roughness around $f_{15}$

### 3.2.4.1 *Nature of the Experimental Task and Subjective Impressions of the Stimuli*

Before discussing possible mechanisms involved in these aspects of perception it may be illuminating to consider more carefully at this point the nature of the task being performed by the subjects and their reports of the perceptual effects they found themselves attending to in order to make the judgment.

The task is to decide which of two sequentially presented tones appears to have more sources. Since the perceptual effects of the different conditions were so varied, the stimuli were blocked in order to allow subjects to focus into a smaller spectral region within which differences between the tones were occurring. And since the task was 2IFC, this reduces to detecting something in one of the tones that may be labeled as indicating more sources.<sup>12</sup> In this case, I presume that once an additional irregularity is perceptible in one of the tones, and the subject can come to recognize it as such, it may be labelled as the indicator of more sources. The theoretical question then becomes one of understanding the nature of the perception and detection of these irregularities and of understanding the auditory processes that encode, detect and interpret them.

As described in the stimulus section previously, the many perceptual effects of introducing an incoherent partial ranged from the appearance of a separately audible spectral pitch, to trills and melodies on partials, to auditory roughness that increased and decreased in a discernible rhythmic pattern. The perceptual effects are summarized in Table 3.6.

Some general effects of modulation width and inharmonicity should be noted. At very small modulation widths, it is very easy to begin to hear these flat-spectrum complex tones as chords and hear out the lower partials. This phenomenon is well-known from psychophysical antiquity (cf. Helmholtz, 1877/1885, and more recently: Plomp, 1976). At times, for example, one imagines oneself hearing out harmonics up

---

12. This is, of course, quite far removed from the kinds of tasks we perform in relation to the distinguishing of auditory sources in the environment. However, as with most such psychoacoustic studies, the aim is to learn something about the limits of sensitivity to certain stimulus parameters that may contribute to such perception in the real world, and to understand something about the auditory processes underlying that perception to the extent to which such extrapolations are viable.

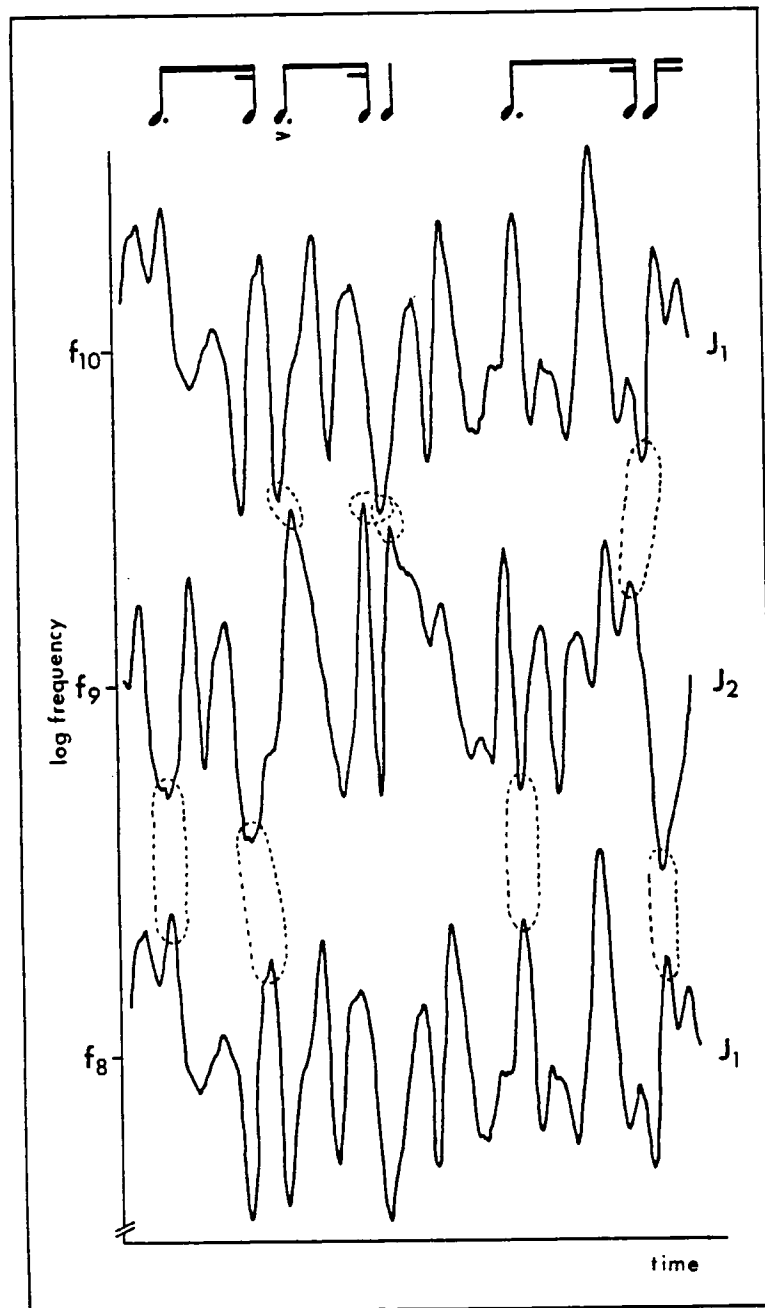
to the 16<sup>th</sup> though this has not been investigated rigorously with these stimuli and is countered by existing evidence (Plomp, 1976). With sufficient modulation widths, however, it becomes increasingly difficult to hear out the lower harmonics even if one know what the pitches are. This suggests a changeover from analytic to synthetic listening as the partials are made to move coherently. Accordingly, one might consider spectral fusion as an increasing tendency toward synthetic listening, i.e. toward attending to more global or collective features of groups of spectral components. The presence of incoherent modulation would stimulate analytic or parsing processes. The question arises, following this line of thinking, about whether spectral fusion is an active or passive (i.e. default) grouping process. This question will be entertained in Chapter 6.

With the inharmonic sounds, it should be noted that they have a virtual pitch roughly equal to that of the harmonic stimuli though this pitch is quite dispersed and rough. Remember that this is a slightly perturbed harmonic series and, as can be seen in Fig. 3.2, may be considered quasi-periodic. However, the lack of phase synchrony between the partials introduces a wide-spread roughness. (These tones sound somewhat like an airplane engine.) It does not seem very easy to hear out the partials (above about  $f_3$ ) in the inharmonic tones due to this roughness.

Now let us consider the percepts in Table 3.6. Note that when spectral pitches become audible and indicate a "new" source it is when their modulation can be perceived as being incoherent with respect to the other modulation pattern present, e.g.  $f_1$  to  $f_5$  for H75 and  $f_1$  to  $f_7$  for I75. The mechanism operating in these cases requires a detectable modulation to make the decision.

The next kind of "new source" percept involves the apparent alternation of spectral pitches, e.g.  $f_5$  for H75 and  $f_5$  and  $f_7$  for H50. This is a surprising percept and I can only imagine that somehow the very slight interactions between adjacent incoherent partials leads the listeners attention from one partial to the next according to when the different modulation patterns move into closer proximity. This effect occurs primarily for partials bordering on a critical band.

A prominent effect in medium to high partials is the "chorus" effect and a rhythmic modulation of roughness. I notated the rhythm (which starts just after the beginning of the tone) as follows:



**Figure 3.13.** Highly exaggerated schematic diagram of the frequency modulation patterns of two coherently modulated components on either side of an incoherently modulated component. The components  $f_8$  and  $f_{10}$  were modulated with jitter function  $J_1$ , and component  $f_9$  was modulated with  $J_2$ . The regions of possible excitation pattern overlap are circled. These moments are notated at the top of the graph as the rhythm of perceived roughness modulation.



If we examine the moments when the frequency movements of adjacent incoherent partials come into greater proximity we can see that this beating of roughness corresponds quite closely to their positions in time. In Figure 3.13 an exaggerated frequency-by-time plot is shown for an incoherent partial surrounded by two coherent partials. Regions where the peaks of the jitter waveform come into proximity between two partials are circled. At the top of the graph these are localized as points in time by the note-heads and the rhythmic figure is indicated immediately underneath. Note the good correspondence between moments of frequency (excitation pattern) proximity and the notated rhythm. In these cases, I believe subjects judged the appearance of this rhythm as indicating a "new source". It is easy to imagine that at very small modulation widths, these interactions would not be as prominent. Also, as incoherent partials are made closer together (by increasing the partial number of the incoherent partial), smaller deviations in frequency would suffice to create the effect. In the case of the alternating spectral pitches, it may be that as the modulation patterns move into proximity, analytic pitch perception follows the pitch trajectory to the other component. Interestingly, the previously described pitch alternations conform more or less to certain parts of this rhythm as well.

This rhythmic pattern is not as prominent for the highest partial for harmonic stimuli and for partials  $f_{11}$  to  $f_{15}$  for the inharmonic stimuli. For the harmonic stimuli, I imagine that the excitation overlap is already sufficient to modulate the roughness with all of the peaks in the jitter waveform and is thus perceived as being more continuous, i.e. as a chorus effect. For the inharmonic sounds, there is a large degree of roughness already present. The incoherent modulation would just increase this roughness.

There is one more thing to note about the roughness that occurs with modulation incoherence. It is perceptually localized in the spectral region of the incoherent partial and appears to be very weakly pitched. This is most apparent when, for example, a trial with a supra-SMT  $f_{13}$  is followed by a supra-SMT  $f_{15}$  trial. One notices that the roughness is "brighter" or higher spectrally in the  $f_{15}$  trial. This suggests that the irregularities of auditory nerve output are being "labeled" with the spectral channel within which they occur. The subjects can then focus into a particular spectral region when attempting to detect the incoherence.

As outlined above, there are several perceptual effects that indicate an "otherness" that subjects judge as the presence of more sources. We might group these effects into two categories:

1. perceived interactions between components, and
2. separately audible components.

I would like to propose that these are due to separate mechanisms, within- and cross-channel mechanisms, respectively.

#### 3.2.4.2 *Within- and Cross-channel Mechanisms of Incoherence Detection*

So let us consider the involvement of within- and cross-channel mechanisms in the processing of frequency modulation incoherence. All of the harmonic data are consistent with the existence of a mechanism that detects within-channel irregularities. This mechanism, as proposed in the introduction, would be sensitive to phase synchrony and excitation overlap of nearby spectral components. If adjacent components are modulated incoherently with a sufficient modulation width, the periodicity (or, more generally, regularity) of the neural discharge pattern within an auditory channel is perturbed and subjects judge such a stimulus as deriving from more sources than a similar stimulus with no incoherent modulation. As will be discussed below, the amount of modulation necessary to detect an incoherence from within-channel information is surprisingly small when compared with coherent modulation detection data. Secondly, the modulation width necessary to make the incoherence detectable was found to be a function of excitation pattern proximity. When the patterns due to adjacent incoherent components were more distant (due to a decrease in overall intensity or to a low partial number) a greater modulation width was required to elicit a "more sources" judgment. In this case one imagines that a minimal overlap of incoherently modulating partials is necessary to make detectable the irregularity that the overlap creates in the auditory channel (or small group of adjacent channels).

The perturbation of the harmonicity of the component center frequencies is accompanied by an increase in the SMTs for higher partial numbers. This reflects the importance of phase synchrony or temporal pattern regularity for within-channel

mechanisms. Under the present condition of a slight initial perturbation of harmonicity, a greater modulation width is necessary to further perturb the regularity of the temporal discharge pattern. When subjects compare across a coherent and incoherent tone, both give irregular cochlear channel output, but the output of the incoherent tone is more irregular. It is important to recall that these "inharmonic" sounds are slight perturbations of a harmonic series and quasi-periodicities are visible in the waveform and audible as well. One might expect that with inharmonic tones that were very different from harmonic tones, i.e. with no discernible quasi-periodicity, incoherent modulation would be even more difficult to detect. Even with the slight inharmonicity used in this study, the difficulty of the task is demonstrated by the large variance in the data (cf. Fig. 3.12).

Perceptually, one would expect the effects of a within-channel mechanism to be related to spectral component interactions in the cochlea, e.g. roughness effects, etc. And these were indeed reported for incoherent partials that were within one critical bandwidth from their neighbors. Although the hypothesis that component proximity plays a role in incoherence detection is consistent with the data for lower partial numbers, the resulting percepts for those stimuli were of an entirely different character, being separately audible spectral pitches. It may be that in those cases a cross-channel comparison process is more prominent in its influence on the perceptual result.

Other data and perceptual effects are also not so easily attributed to a within-channel mechanism. The comparatively low SMTs of  $f_1$  and  $f_3$  for inharmonic stimuli would necessarily involve some mechanism that could compare frequency modulation patterns across auditory channels, as proposed in the introduction section. The lack of phase synchrony between  $f_1$  and the rest of the complex, in addition to its incoherent frequency movement would make it more easily separated than in the harmonic case where a much larger deviation is necessary to override its harmonic relation to the rest of its tone complex. This comparison points to the possible influence of a harmonic spectral pattern-matching process in the decisions made on the temporal information in frequency modulation patterns by a cross-channel incoherence detection mechanism.

Another feature of a cross-channel mechanism indicated by these data is that this mechanism is much less sensitive than the within-channel mechanism. Larger modulations are necessary to be detectable as incoherent when they are being compared across auditory channels, rather than interacting within a single channel. Given that such a mechanism would most likely be at a higher level in the auditory system, more noise might enter the information stream as it passed through various stages of processing. It is likely that certain fine variations would be lost in that case. The relative sensitivities of the two mechanisms will be discussed below in relation to modulation detection ability of the listeners.

The percepts obtained by incoherently modulating the lower partials are more likely due to a cross-channel mechanism. For these partials, in both harmonic and inharmonic stimuli, the incoherent partial was heard separately as a spectral pitch. Informal listenings have demonstrated that if comparable modulation widths (3 - 8 cents) are applied to the higher partials, they too are separately audible. This suggests that if a cross-spectral mechanism is responsible for this effect, it requires greater modulation width to be stimulated and, most important, that it is this kind of mechanism that is responsible for the distinction of separate auditory source images.

As noted previously, the data indicate that the within-channel mechanism is much more sensitive than the cross-channel mechanism. The information for both kinds of processing would be present in the temporal firing pattern of single auditory fibers. And perhaps the synchrony or correlation of firing across the array of fibers could be detected further along in the nervous system. The use of fiber firing synchrony, or periodicity, for detection of multiple sources could be added to that of supra-saturation representation of spectral form by the temporal discharge pattern in the auditory nerve fiber array (Young & Sachs, 1979; Delgutte, 1980; Sachs & Young, 1980; Voigt, Sachs & Young, 1981).

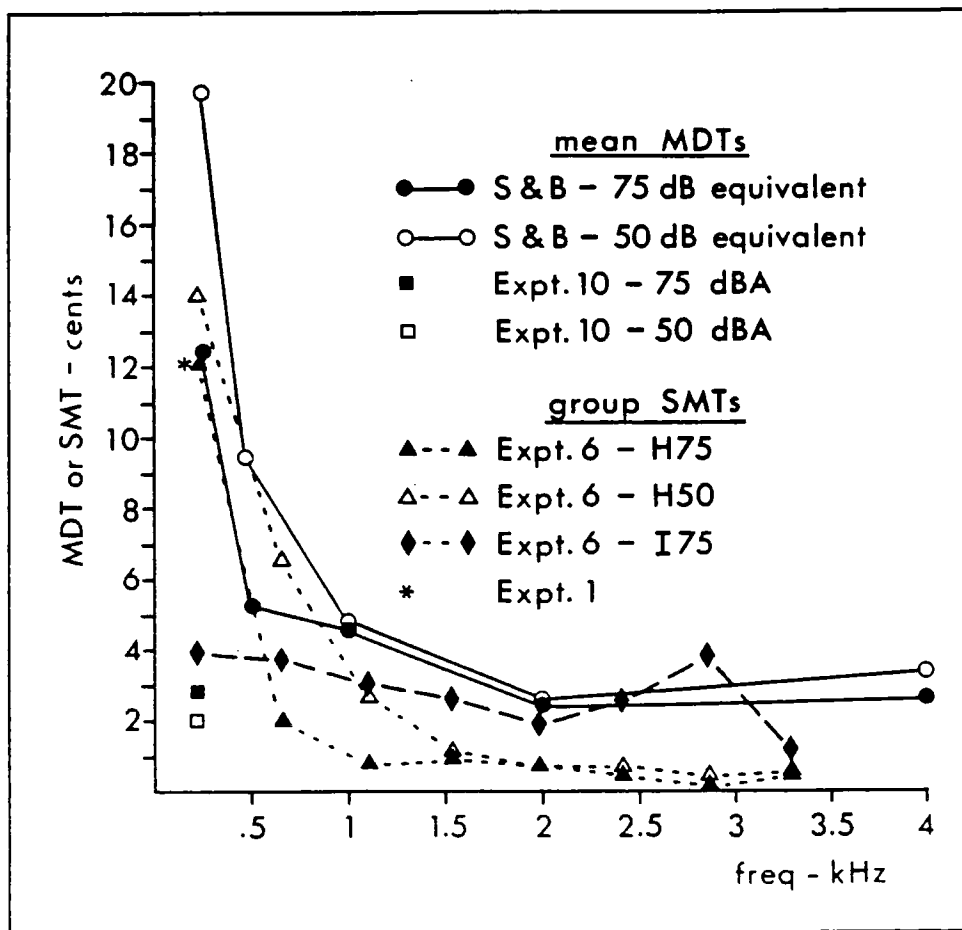
It is interesting to examine the relative sensitivities of these mechanisms in relation to the sensitivity to frequency modulation of sinusoidal and complex harmonic tones. Figure 3.14 presents the modulation detection threshold (MDT) data from Shower & Biddulph (1931) for FM detection with a periodic carrier, and those from Experiments 1 (SMTs), 6 (SMTs) & 10 (MDTs for coherent, harmonic complex tones).

For the Shower & Biddulph data, the computed partial loudnesses at different frequencies (see Table 3.5) for the 75 and 50 dB stimuli of this experiment were used in place of the  $\alpha_s$  (sensation level) measure in their study. If this value fell between the values of  $\alpha_s$  they used, the MDT was linearly interpolated between the corresponding  $\Delta f_{peak}/f$  measures in their table (p. 279). The rms deviation was calculated from the measured frequency deviations (marked "X" in their Fig. 2, p. 277). The relation between  $\Delta f_{peak}$  and  $\Delta f_{rms}$  was determined to be  $\Delta f_{rms} = 0.724\Delta f_{peak}$ . Then the  $\Delta f_{peak}/f$  was converted to cents<sub>rms</sub>. These values are listed in Table 3.7 and plotted in Figure 3.14.

**TABLE 3.7.** Transformation of Shower & Biddulph's (1931) data for comparison with those from Experiments 1, 6 & 10. Plotted below are the carrier frequencies, sensation levels (equivalent to partial loudnesses of components in the present study), MDTs expressed as  $\Delta f_{peak}/f$ , and MDT expressed as cents<sub>rms</sub>. The modulation waveform they used was a rounded trapezoid and the ratio  $\Delta f_{rms}/\Delta f_{peak}$  was 0.724.

Equivalent to partial loudnesses of 75 dBA tone.			
Carrier	Sensation	MDT	
Frequency	Level	$\Delta f_{peak}/f$	cents <sub>rms</sub>
250	50	.0099	12.4
500	50	.0042	5.2
1000	50	.0036	4.5
2000	54	.0019	2.4
4000	58	.0020	2.5
Equivalent to partial loudnesses of 50 dBA tone.			
Carrier	Sensation	MDT	
Frequency	Level	$\Delta f_{peak}/f$	cents <sub>rms</sub>
250	15	.0158	19.7
500	18	.0075	9.4
1000	20	.0039	4.9
2000	24	.0020	2.5
4000	30	.0027	3.4

The mean MDTs from Experiment 10 (Appendix D) are for the coherent harmonic stimuli used in this experiment at both 75 and 50 dBA. These were determined using an adaptive up-down procedure that estimates the detection threshold at 71% correct choice. Thus, these represent sensitivity to coherent frequency modulation applied



**Figure 3.14.** Comparison of SMTs with MDTs for sinusoidal and complex carriers. Data for sinusoidal carriers are taken from Shower & Biddulph (1931; see Table 3.6). MDTs for the harmonic complex carrier are taken from Experiment 10 (App. D, Table D.1). Also plotted is the group SMT from data on the jittered flat-envelope stimuli in Experiment 1. This value is plotted at 220 Hz since the prominent perceptual effect was the segregation of the fundamental frequency from the rest of the tone complex.

to a complex harmonic tone. The individual MDTs for the subjects in this experiment are also indicated in Figures 3.8 - 3.11 for comparison with Expt. 6 data. The SMT value from Experiment 1 was obtained from the mean data of the flat spectral envelope stimuli with jitter modulation. In all cases, the threshold measure is plotted as a function of the frequency of either the sinusoidal carrier (Shower & Biddulph),

the  $F_0$  of the complex tone (Expts. 1 and 10) or the frequency of the incoherently modulated partial (Expt. 6).

Several things are worth remarking. The differences between "75 dB" and "50 dB" curves are similar for both the sinusoidal MDTs and incoherent partial SMTs for harmonic stimuli. Specifically, at lower frequencies the threshold is substantially higher for the lower intensity tone, whereas the two curves are very close to one another at higher frequencies. This may indicate that the effect of intensity on SMTs is other than an effect of changing the proximity of excitation area and may be more closely related to the limited spectral resolution in the lower frequencies.

Note also that for frequencies higher than the fundamental, the SMTs are much lower than the sine tone MDTs. This is also true for the relation between SMTs and complex tone MDTs. In the latter case the SMTs are lower than the MDT of the coherent signal for incoherent partials greater than the 3<sup>rd</sup> for 75 dB and the 5<sup>th</sup> for 50 dB. In general, here we may conclude that our sensitivity to incoherent modulation of partials is much more acute than is our sensitivity to the modulation itself in a coherent source. This presents a rather remarkable monaural precision of either temporal or spectral encoding.

As can be seen in Table 3.4, the largest  $\Delta P$  about  $30\mu\text{sec}$  is one order of magnitude smaller than the finest time resolution classically reported for monaural listening (cf. Pollack, 1968; Green, 1971; Schubert, 1979, editor's comments, p. 259). The smallest  $\Delta P$ s are again three orders of magnitude smaller than that. These values cannot really be compared directly with the other studies of temporal acuity since the tasks and stimuli are completely different. However, these results do indicate a remarkable acuity of incoherent modulation detection on the part of the auditory system.

The group SMT found in comparisons of coherent (constant-ratio) modulation and constant-difference modulation in Experiment 1 is almost identical to that for the incoherent  $F_0$  in the present experiment and for modulation detection with a rounded trapezoidal carrier by Shower & Biddulph (1931). This would suggest the possibility that in both of these experiments the judgment was based on detecting the modulation of a perceptually segregated  $F_0$ . It was certainly not the detection of the modulation of the virtual pitch of the harmonic complex as is evidenced by the much lower

MDTs of coherent harmonic complexes. It is still difficult, however, to explain the low SMT of the  $f_1$  for inharmonic stimuli, which is a factor of 3 smaller than the rest of these values. My initial interpretation of this effect was that the coherent partials in the inharmonic tone are not generating a strong virtual pitch as is the case in the harmonic tones. And, consequently, there would be less confusion perceiving an independently modulating  $f_1$  as a separate source.<sup>13</sup> Given that this tone is already much less fused or unified in its pitch quality, the  $f_1$  may already be independently perceptible and then detection of incoherence as a cue for multiple sources required much less modulation to make the  $f_1$  stand out even more by further perturbing its frequency relation to the rest of the complex. It should be noted that the amount of modulation necessary to achieve this is still much less than is detectable when applied to a sinusoid at the same frequency. However, apparently the amount needed to make it stand out is below the MDT for this frequency, which makes the argument untenable.

Perhaps the difference in SMT of incoherent  $f_1$  for harmonic and inharmonic tones is the combined effect of a cross-channel extraction and pitch confusion. In this case, it is not the detection of a modulating  $f_1$  *per se* that makes it audible but rather that it is being "defused" from the rest of the inharmonic complex, whereas for the harmonic complex there is a tendency to subsume the  $f_1$  into the complex by virtue of its harmonic relation and a great deal of modulation is necessary to override this tendency for most subjects. Note again that this is not the case for S2 who had an SMT for the H75  $f_1$  that was much lower.

### 3.3 Summary

This experiment has demonstrated that the auditory system is capable of using incoherence among the frequency modulations of spectral components as a cue for the presence of multiple sound sources. The sensitivity to incoherent modulation among the higher partials of a harmonic tone is at modulation widths 3 to 10 times smaller than the sensitivity to modulation of a sinusoid at the same frequency and approximate loudness. The sensitivity to the incoherent modulation of a single partial in a harmonic tone is at modulation widths 4 to 5 times smaller than the sensitivity in

---

13. Work in progress by Hartmann (1983) suggests that under certain conditions, the inharmonicity of a single partial *alone* can cause it to stand out as separately audible, even without modulation.



an inharmonic tone.

We are looking for mechanisms by which source components are collected across the spectrum and are fused into images which are distinguished perceptually from other simultaneously occurring images. From these data it appears that one mechanism of distinction can operate at a local spectral level, using peripheral interactions of spectral components with overlapping excitation patterns, i.e. a within-channel mechanism. Coherent harmonic sources generate relatively slowly modulating periodicities in these regions of overlap. When the excitation patterns of incoherently modulating partials overlap, this generates an aperiodic response in auditory fibers connected to that region, thus destroying synchrony among them.

The within-channel mechanism is most sensitive to small incoherent modulations in heavily overlapping excitations, i.e. due to components falling within a critical bandwidth. It is also most sensitive to departures from periodicity. Since, the system is also capable of detecting incoherence in inharmonic tones and incoherence between spectral components that are presumably minimally or non-overlapping in the peripheral auditory system, another, higher-level process needs to be postulated which detects coherence in either the firing patterns or coherent movement of activity across the tonotopic array of neural elements, i.e. a cross-channel mechanism.

The cross-channel mechanism requires greater modulation to determine which partials are modulating coherently with respect to one another and which are not. From informal listenings it appears that this mechanism is capable of separating even "unresolved" partials that are embedded in a tone complex if their modulation is incoherent and of sufficient width. This more global process would certainly be necessary to explain the ability of the auditory system to separate multiple, spectrally intertwined sources as is possible in multi-speaker and musical environments. The problem of multi-source identification is treated in Chapter 5.

Of course both mechanisms contribute to the final perceptual result and in an experimental situation their separate indications can be made to converge on a given parsing solution (as is the case in many everyday encounters), in which case the perceptual result is unambiguous as with the parsing of harmonic stimuli. Or their indications can be made to diverge as in the coherent inharmonic stimuli where the

within-channel "reading" of irregularity of temporal discharge that signals multiple sources competes with the cross-channel "reading" of frequency modulation coherence that signals a single source, etc. What this indicates is that there are limits to the distinction of and attention to multiple sources.

## CHAPTER 4

### Fixed Spectral Structure and Spectral Fusion

#### 4.1 Introduction

In Chapter 1, it was proposed that imposing a coherent frequency modulation on a complex harmonic tone such that the amplitudes of the spectral components remained constant rather than tracing the spectral envelope would result in a less stable or less fused sound image than would be the case when the amplitude changed with frequency by tracing the spectral envelope. It was also proposed that more complex spectral forms, such as a vowels, would be more susceptible to such distortion than simple ones, such as a  $-6$  dB/oct slope. Experiment 7 tested both of these hypotheses in a 2IFC task where subjects were asked to judge which tone in a pair was less fused or had more sources in it. One tone had a constant spectral envelope while the other had constant component amplitudes.

#### 4.2 EXPERIMENT 7: Effects of fixed spectral envelope on perceived source multiplicity.

##### 4.2.1 Stimuli

Two spectral envelopes were used: vowel /a/ and  $-6$  dB/oct slope. These were imposed on 16-harmonic complex tones with a 220 Hz  $F_0$ . The two modulation waveforms of Chapter 2 were used: vibrato and jitter. There were two conditions with different relations between the spectral envelope and frequency modulation: 1) constant spectral envelope (*CSE*), where the amplitudes of the harmonics were made to conform to the spectral contour when being modulated in frequency, and 2) constant component amplitudes (*CCA*), where the initial amplitude values of the harmonics

were derived from the spectral envelope at their center frequencies, but these amplitudes remained constant with frequency modulation. These waveforms may be described as follows:

$$S_{CSE}(t) = \sum_{n=1}^{16} A(f_{ni}) \sin(2\pi n F_0 t + n \psi \int_0^t Mod(t') dt'), \quad (4.1)$$

and

$$S_{CCA}(t) = \sum_{n=1}^{16} A(\bar{f}_n) \sin(2\pi n F_0 t + n \psi \int_0^t Mod(t') dt'), \quad (4.2)$$

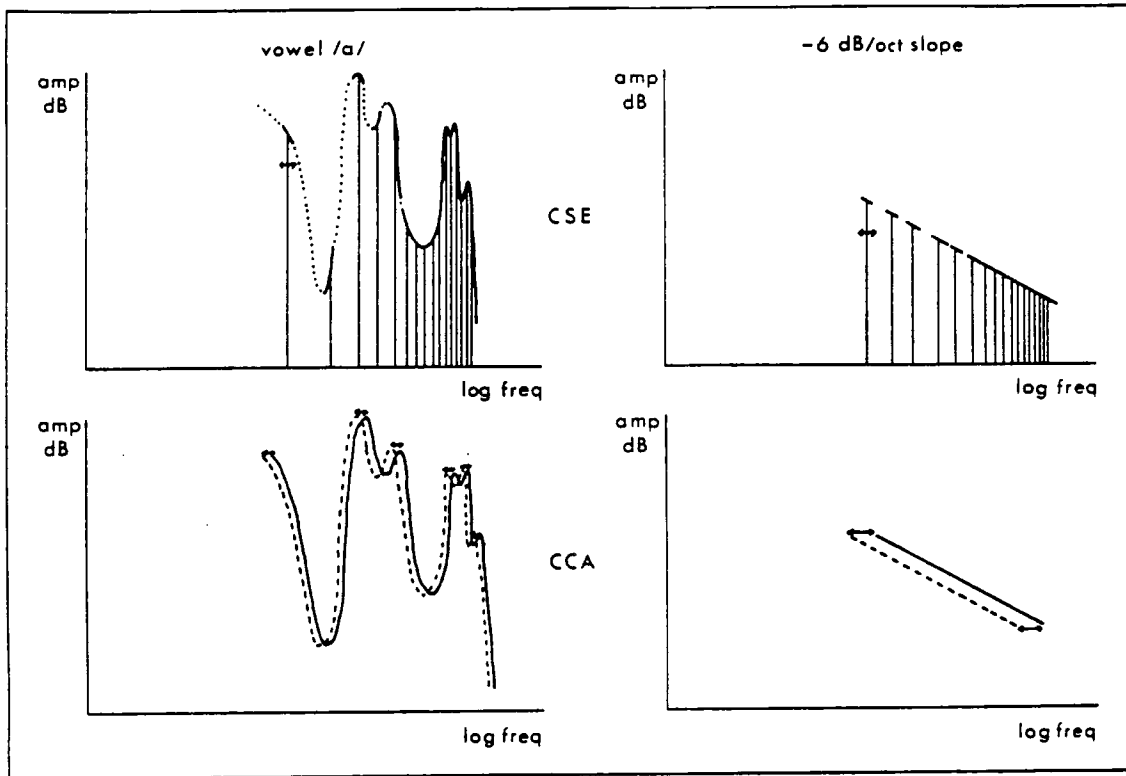
where  $f_{ni}$  is as in equation (2.2);  $A(f_{ni})$  is the instantaneous amplitude of partial  $n$  dependent on that partial's instantaneous frequency,  $f_{ni}$ ;  $A(\bar{f}_n)$  is the constant amplitude of partial  $n$  dependent only on an amplitude value related to the partial's center frequency,  $\bar{f}_n$ . Figure 4.1 presents a schematic representation of these two types of signals for the two spectral envelopes. Note that for the vowel spectrum and a 220 Hz  $F_0$ , at least two harmonics are found within each formant contour.

This yields 6 basic tone types: (2 spectral envelopes)  $\times$  (2 modulation waveforms)  $\times$  (2 AM/FM coupling conditions). For each of these tone types, 6 values of modulation width (rms deviation) were used: 7, 14, 28, 42, 56, 70 cents. Each tone was 1.5 sec in duration with 100 msec raised cosine ramps. The *CSE* stimuli with modulation widths up to 56 cents were identical to the *CR* stimuli in Chapter 2. All stimuli were matched for loudness. Vowel /a/ stimuli were amplified by approximately 1.2 dB in order to equalize loudness with the -6 dB/oct stimuli. Stimuli were presented over headphones at approximately 75 dBA in a sound-treated room (see Appendix A).

#### 4.2.2 Method

In each trial, subjects were presented one *CSE* tone and one *CCA* tone in succession and in counterbalanced order. Each tone presentation was accompanied by a light of a different color to signal the two intervals. The tones were separated by a 500 msec silence. Both tones had the same spectral envelope shape, modulation waveform and rms deviation. The task was to decide which tone seemed to have more sources in it, potentially derived from more sources, or was perceptually more analyzable into separate sounds, i.e. seemed to split apart into distinguishable sound elements. This choice was indicated by pressing the appropriate button on a 2-button

box. Once the response was received, an additional 500 msec silence occurred before presentation of the next pair.



**Figure 4.1.** Schematic illustration of the effect of either tracing (*CSE*) or of modulating (*CCA*) the spectral envelopes of the vowel /a/ or the -6 dB/oct slope. For the purposes of illustration a modulation width twice the size of the largest used in the experiment is shown (133 cents,  $\Delta f / \bar{f} = 0.08$ ). Note that the spectral envelope is completely defined above the 5<sup>th</sup> partial.

Stimuli were blocked according to modulation waveform. Each run consisted of one such block with 120 comparisons: (2 spectral envelopes)  $\times$  (6 rms deviations)  $\times$  (10 repetitions). 5 vibrato and 5 jitter runs were presented to each subject in counterbalanced order across subjects. Thus, 50 2IFC judgments were collected for each stimulus pair. An experimental session consisted of 2 - 4 runs. Data values represent the proportion of times the *CCA* tone was chosen as yielding more sources or being more split apart perceptually. The greater the differential effect of the

frequency modulation-spectral envelope coupling in the direction expected, the higher this value will be.

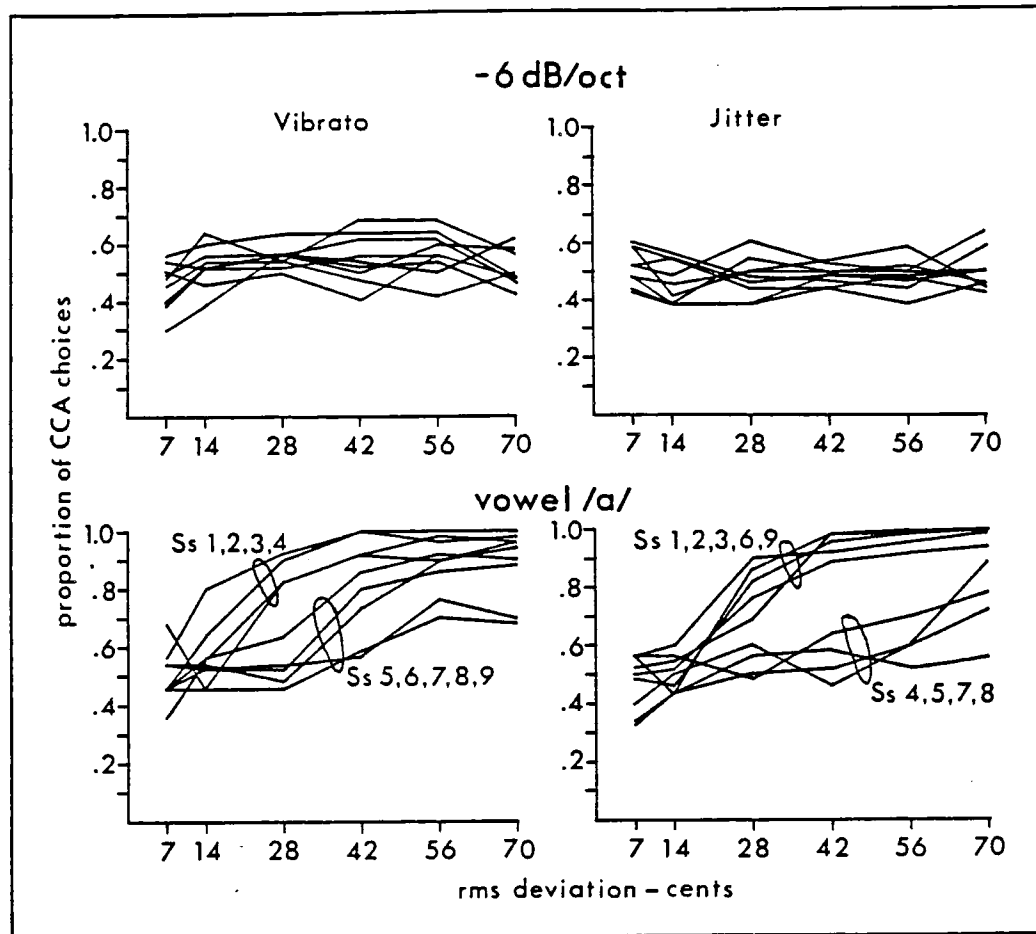
Nine subjects participated in the experiment and were paid for their time. None reported having any serious hearing problems. Five of these subjects had participated in Experiments 1 and 2, and 2 of them in Experiments 1 - 6. These subjects were informed that though the judgment protocol was the same, the perceptual effect to be listened for was quite different, so they would have to adopt a new criterion for making the judgments. Five subjects were musicians with several years of formal musical training (4 of these were professional musicians or composers and one was the author) and 4 subjects reported having no such training.

#### 4.2.3 Results

The data for all subjects and the means and standard deviations across subjects are listed in Table E.6 (Appendix E). These data are plotted in Figure 4.2 in order to show the spread of responses for different subjects. From this graph it appears that the data fluctuate around random choice for the  $-6$  dB/oct spectrum, and are roughly monotone increasing with modulation width for the vowel spectrum.

##### 4.2.3.1 Effects of spectral envelope

Mean data values across subjects (see Table 4.1) were tested to see if they were significantly different from chance (two-tailed  $t$ -test for  $\bar{x} = 0.5$ ). For the  $-6$  dB/oct spectrum with vibrato modulation, two means were significantly higher than 0.5, i.e. at 28 cents and 56 cents; for the jitter waveform no means were different from 0.5. I cannot find any reasonable interpretation for the significance of these two values given that the largest modulation width, and that in between these two are at random choice. I would prefer to conclude that these are Type I errors, i.e. the  $H_0$  is rejected when true, and that with the  $-6$  dB/oct spectrum, subjects do not preferentially choose either *CSE* or *CCA* tones as more often yielding an impression of multiple sources. In fact, *every* subject mentioned the extreme difficulty of making a choice with this spectral envelope, claiming that they heard no difference whatsoever, regardless of modulation width.



**Figure 4.2.** Experiment 7 data summary. The proportion of CCA tones chosen as yielding more sources is plotted as a function of rms deviation of modulation for 2 modulation waveforms and 2 spectral envelopes. Each curve represents the data for one subject. Each point represents 50 2IFC judgments.

For the vowel spectrum, there appear to be two groups of responses with different slopes in the data curves. For vibrato, these groups include Ss 1, 2, 3, 4 vs. Ss 5, 6, 7, 8, 9. For jitter, the groups include Ss 1, 2, 3, 6, 9 vs. Ss 4, 5, 7, 8. It is interesting to note that Ss 1, 2, 3, 6, 9 are musicians, while the others are non-musicians. Accordingly, the data are grouped by subject's musical ability and are averaged separately for judgments on vowel /a/ stimuli (see Table 4.1).<sup>1</sup> The values for each group that

are significantly different from chance are marked in Table 4.1. For both groups, the values are roughly monotone increasing with rms deviation, but the slopes are much less for the non-musicians than for the musicians. For musicians, the difference between *CSE* and *CCA* stimuli was discernible at 14 cents modulation width ( $p(\bar{x} = .5) < .05$ ), while for non-musicians a 42 cents modulation width was necessary for discrimination (see Table 4.2).

**TABLE 4.1.** Experiment 7 data summary. Means and unbiased standard deviations (in parentheses) across subjects for each spectral envelope, modulation waveform and rms deviation. The data for the vowel /a/ spectrum are split into two groups according to the musical ability of the subject. <sup>a</sup>  $p(\bar{x} = .50) < .05$  <sup>b</sup>  $p(\bar{x} = .50) < .01$ .

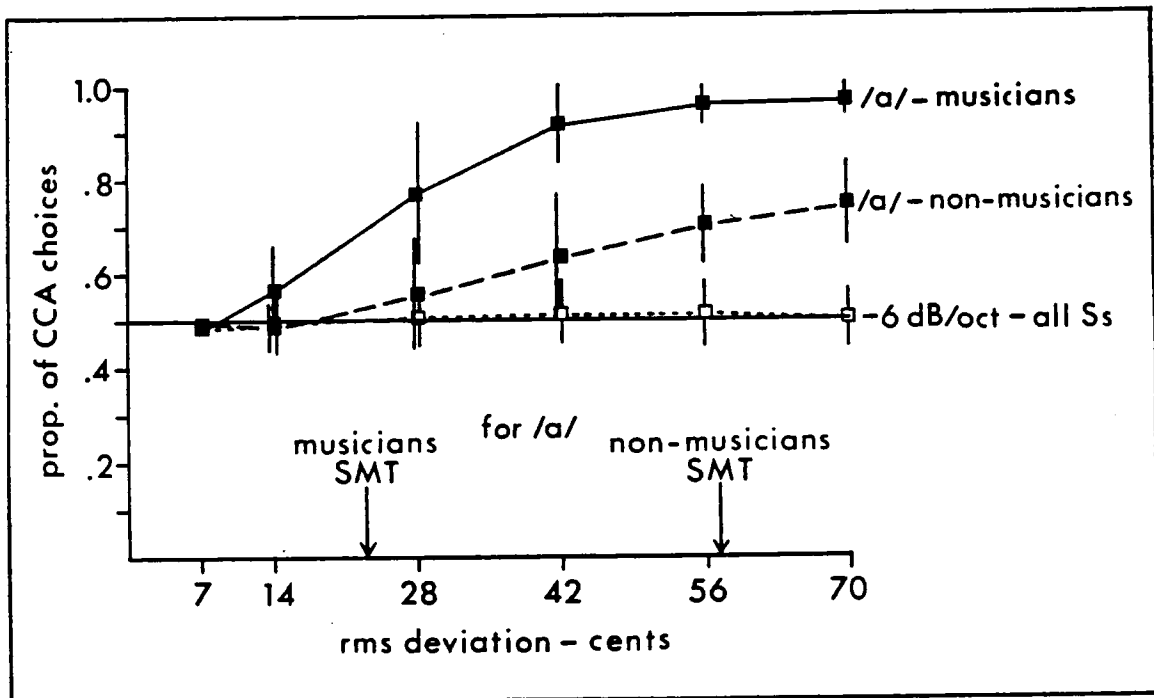
Rms Deviation of Modulation ( cents )						
	7	14	28	42	56	70
<b>-6 dB/oct spectrum. All subjects (<math>N = 9</math>)</b>						
vibrato	.46 (.08)	.53 (.07)	.55 <sup>b</sup> (.04)	.55 (.09)	.57 <sup>a</sup> (.08)	.51 (.06)
jitter	.51 (.06)	.46 (.07)	.48 (.07)	.49 (.04)	.48 (.05)	.50 (.07)
<b>vowel /a/ spectrum. Ss 1,2,3,6,9 (musicians; <math>N = 5</math>)</b>						
vibrato	.48 (.08)	.63 (.11)	.75 <sup>a</sup> (.19)	.90 <sup>b</sup> (.11)	.95 <sup>b</sup> (.04)	.96 <sup>b</sup> (.04)
jitter	.48 (.09)	.51 (.06)	.80 <sup>b</sup> (.09)	.94 <sup>b</sup> (.04)	.97 <sup>b</sup> (.03)	.98 <sup>b</sup> (.03)
<b>vowel /a/ spectrum. Ss 4,5,7,8 (non-musicians; <math>N = 4</math>)</b>						
vibrato	.53 (.10)	.49 (.03)	.58 (.16)	.71 (.17)	.80 <sup>b</sup> (.09)	.80 <sup>a</sup> (.13)
jitter	.46 (.11)	.48 (.06)	.53 (.05)	.55 (.08)	.60 (.07)	.73 <sup>a</sup> (.13)

1. There was no significant difference between these groups for the -6 dB/oct stimuli.



#### 4.2.3.2 Effects of modulation waveform

To test for the effect of modulation waveform, the means for each spectral envelope (within each group for vowel /a/) and each rms deviation were compared across modulation waveforms (two-tailed  $t$ -test for  $x_i - x_j = 0$ ). The only comparisons that reached a significance level of .05 were those for the -6 dB/oct spectrum at 28 cents and 56 cents. These are the same stimuli mentioned previously that



**Figure 4.3.** Experiment 7 data summary. Means and pooled standard deviations ( $\pm\sigma_p$  indicated by vertical bar) across subjects and modulation waveform are plotted as a function of rms deviation. The judgments on vowel /a/ are split into groups of musical and non-musical subjects. The 71% *source multiplicity thresholds* (SMTs) are indicated for the vowel stimuli for both groups of subjects.

appeared anomalously different from chance. Likewise, here, I will attribute this to chance and conclude that there are no significant effects of modulation waveform. This adds to similar results in Experiment 1, reinforcing the contention that rms

deviation is an appropriate measure of modulation width for many perceptual effects. The overall means and pooled standard deviations (across subjects and modulation waveforms) are collected in Table 4.2 and plotted in Figure 4.3.

**TABLE 4.2.** Overall means and pooled standard deviations across subjects and modulation waveforms for Experiment 7. <sup>a</sup>  $p(\bar{x} = 0.5) < .05$   
<sup>b</sup>  $p(\bar{x} = 0.5) < .01$ .

	Rms Deviation of Modulation (cents)					
	7	14	28	42	56	70
<b>-6 dB/oct spectrum.</b>						
All Ss ( $N = 18$ )	.48 (.07)	.49 (.07)	.51 (.06)	.52 (.07)	.52 (.07)	.50 (.06)
<b>vowel /a/ spectrum.</b>						
Musicians ( $N = 10$ )	.48 (.09)	.57 <sup>a</sup> (.09)	.77 <sup>b</sup> (.15)	.92 <sup>b</sup> (.08)	.96 <sup>b</sup> (.04)	.97 <sup>b</sup> (.03)
<b>vowel /a/ spectrum.</b>						
Non-musicians ( $N = 8$ )	.49 (.11)	.48 (.05)	.55 (.12)	.63 <sup>a</sup> (.13)	.70 <sup>b</sup> (.08)	.76 <sup>b</sup> (.09)

As in the previous chapters, a source multiplicity threshold (SMT) for the judgments on vowel /a/ stimuli was calculated for the two groups of subjects. A cubic spline was fitted to the six points and the 71% point was determined. For musicians, the SMT was at 23.1 cents ( $\Delta f_{rms} / \bar{f} = 0.0134$ ); for non-musicians the SMT was at 57.4 cents ( $\Delta f_{rms} / \bar{f} = 0.0337$ ), approximately twice the value for the musicians.

#### 4.2.4 Discussion

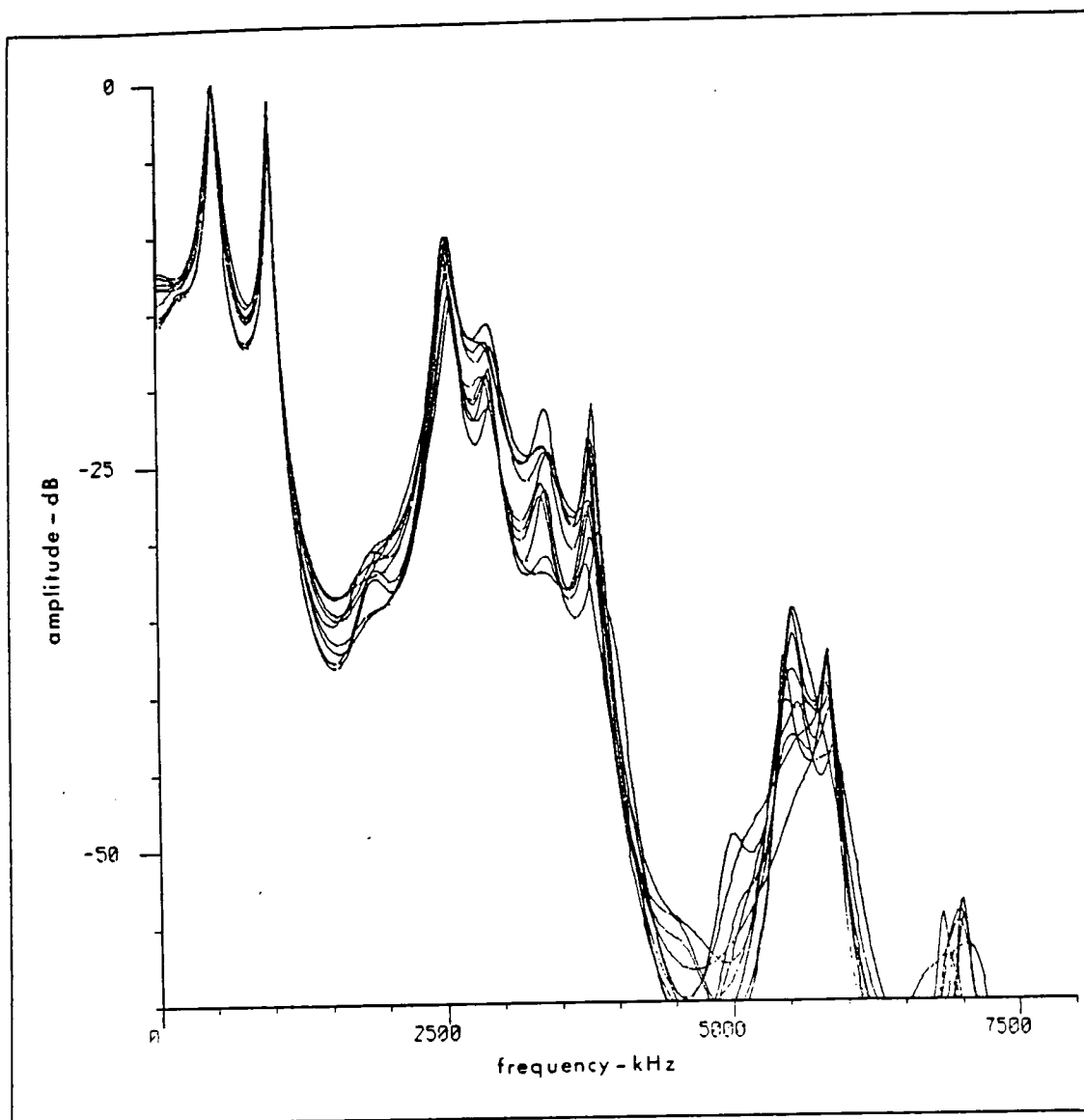
The data show that the vowel-like tones with a sub-audio frequency modulation that were not accompanied by an amplitude modulation following the contour of a constant spectral envelope were perceived as less fused than similar tones where the spectral contour *was* followed. This effect was dependent on the nature of the spectral envelope, on the modulation width and on the musical ability (or perhaps listening skill) of the subject. A vowel /a/ spectrum yields this effect quite clearly, but the effect is altogether absent for a simple spectral slope of -6 dB/oct. For musicians, the modulation width at which the CCA tones just begin to "defuse" is approximately 23 cents. This value for non-musicians is approximately twice that of the musicians. There was no appreciable effect of the modulation waveform: periodic and (fixed) aperiodic waveforms gave the same results when their modulation widths were

expressed as rms deviation from center frequency.

The effect of spectral envelope shape has interesting implications for the global perception of sound sources with complex resonance structures. The results may be interpreted as indicating that our perceptual strategies for organizing the environment are somehow derived from, or closely tied to the normal behavior of the physical world. Normal sources do not have resonances that vary so radically with the natural modulation of the fundamental frequency. While certainly the voice has an enormous flexibility in the variation of its resonance structure, the changes in formant frequencies rarely occur with the speed and to the same extent as was synthesized in the *CCA* tones in this experiment.

Rodet (1983) performed a period by period spectral analysis of a vowel /a/ sung by a professional soprano. This analysis is illustrated in Figure 4.4. The transfer functions extracted from 10 successive periods are superimposed in the graph. Note that while there is considerable variation in the amplitudes of the higher formants, their frequencies change relatively little. In the stimuli in the present experiment, the first 4 formants are represented (the 16th harmonic has a frequency of 3520 Hz). Note also in the Rodet data that there is very little fluctuation in the first 2 formants' amplitudes and no appreciable fluctuation in their frequencies. Rodet resynthesized this vowel sound and remarked that when the amplitude fluctuations were removed and a frequency modulation was included, a kind of whistling sound was heard in the region of the higher formants. This whistling disappeared when the higher formant amplitudes were jittered slightly.

The perceptual result of moving the formant frequencies with the modulation is a reduction in the quality of the vowel as well as the addition of a high frequency whistling which modulates with the same pattern as the modulation waveform. This whistling becomes more apparent with increasing modulation width. Most subjects reported basing their multiplicity judgments on the appearance of this separate whistling sound. However, none reported (and I could not detect) any whistling in the constant spectral envelope sound. According to Rodet's report the lack of modulation on the formant amplitudes should result in some abnormal effects. However, as the data suggest the greatest abnormality with respect to source multiplicity occurred for the *CCA* tones. Informal pitch matchings by the experimenter and one musical subject placed the pitch of this whistling sound in the region of  $F_3/F_4$  (2450 - 2750 Hz). It



**Figure 4.4.** Superimposed transfer functions extracted from 10 successive periods of the vowel /a/ sung by a soprano. Note the amplitude variation of the formant peaks. Note also the relative lack of variation in the formant frequencies. Data from Rodet (1983).

seems imaginable that the rapid movement of this double formant, coupled with its narrow bandwidth, created rapid amplitude fluctuations for auditory fibers being stimulated in that region. If we hypothesize that there are special processes for detecting and following formant trajectories (as would need be the case for the

perception of liquid and glide consonants, for instance), it is possible that the output from such a process is responsible for this percept (cf. Sapozhkov, 1973). As for explaining why this is *not* then integrated into the whole, I am drawn back into an explanation that proposes that perception is constructive and that it reconstructs the world from the analyzed information it receives *according to models of how the world normally behaves*. This perceiving has an active relation with the world, where it is constantly constructing hypotheses about how the world is behaving and then modifying these hypotheses based on the information it continues to accumulate as the perceiving organism actively explores the world. With the CCA vowel sound in this experiment, the formant structure is modulating abnormally and the formants begin to be perceived independently rather than as a group.<sup>2</sup> Conversely, there is very little change in the spectral structure with the -6 dB/oct spectrum and the difference between following the slope and not following the slope is minimal if at all perceptible.

Concerning the difference between musical and non-musical listeners, I would be inclined to attribute the difference to skill in analytic spectral listening. Instrumentalists, for example, often need to be able to hear "into" a complex spectrum and to hear out single sources (their own instrument, or another with which they are trying to coordinate their playing) using whatever spectral/temporal cues they can. One might object that the same is the case with normal listening to speech in a noisy environment. But in this case there is a much more redundant and rich context from which to retrieve lost information. The case of this experiment is closer to the abstract, immediate listening that takes place in music (though it is still a far cry from being a "musical" situation).

### 4.3 Summary

This experiment has shown that the perception of the unity of more complex spectral structures is sensitive to the coupling of frequency modulation with an amplitude modulation that defines that spectral structure over time. It would be interesting from this starting point to investigate the role that formant perception plays in such perceptual organization, to know if the actual identity or familiarity of a

---

2. This technique of independent formant frequency modulation for defusion of voice sounds into individual formants has been used elegantly in a recent composition by Jean-Baptiste Barrière (1983). This will be discussed in Chapter 6.

spectral structure is important, or whether the important factor for such perception is merely the complexity of the spectral form.

As concerns music synthesis procedures that do not maintain constant spectral structure, this study suggests that, particularly for voice synthesis, vibrato or jitter modulations in excess of  $\pm 1$  quarter-tone will deform the identity and perceptual unity of the sound. Following this, it may prove useful to examine ways of making such syntheses a bit more malleable or to find parameter manipulations that override these tendencies toward spectral deformation of the musical source.

various conditions. In contrast to previous chapters, subjects were making judgments on the identifiability of the sources under different stimulus conditions that are known to affect source image formation and distinction. One would expect that when no vowels are modulated, it would be difficult to separate them as source images and that the judged prominence would be low. Here the effect of frequency masking and spectral overlap between the vowels would be the limiting factor in source identification. Similarly, when all vowels are modulated coherently (in an identical manner) in frequency, it may also be difficult to separate them. However, with frequency modulation (and as a direct result of the synthesis technique to be used in generating the stimuli), each modulating component would trace the spectral envelope of the vowel to which it belonged, thus providing additional information about the resonance structure that was contributing to the multi-source complex. This information may make the identification (prominence) judgment easier. Here the relative influences of these two stimulus parameters can be tested against one another. Finally, one would expect that when a vowel is modulated independently of the others, it would be more easily formed into a source image and judged to be more prominent.

## 5.2 **EXPERIMENT 8:** Effects of sub-audio frequency modulation and fixed resonance structure on the perceived prominence of vowel sources embedded in a complex (multi-source) spectrum.

This experiment owes its basic ideas to two people. John Chowning demonstrated the emerging source image effect by embedding the spectra of many voices in a bell-like sound and then causing them to emerge by modulating the frequency components coherently. David Wessel suggested that a good way to test the notion that fusion is aided by coherent frequency modulation is to create a situation where identifiability or perceptual salience of a source in a sound complex was dependent on its being modulated.

5.2.1 *Pre-test*

Before sense can be made out of judgments of the prominence (implying identifiability) of vowels embedded in a complex spectrum, it must be ascertained that the component vowels are identifiable in isolation. This also allows Ss to have prior experience with the stimuli that are to be identified under more difficult circumstances later.

**TABLE 5.1.** Parameters for vowel synthesis with the program CHANT.

<b>vowel /a/</b> (rms amp = -3.8 dB re: /o/)		
formant frequency (Hz)	bandwidth (Hz)	amplitude (dB re: most intense formant)
600	77.6	0.00
1050	88.4	-6.15
2400	122.9	-11.98
2700	127.8	-11.04
3100	137.7	-23.77

<b>vowel /o/</b>		
formant frequency (Hz)	bandwidth (Hz)	amplitude (dB re: most intense formant)
360	51.0	0.00
750	61.0	-11.49
2400	167.8	-29.34
2675	183.5	-26.41
2950	198.4	-35.36

<b>vowel /i/</b> (rms amp = -4.4 dB re: /o/)		
formant frequency (Hz)	bandwidth (Hz)	amplitude (dB re: most intense formant)
238	73.4	0.00
1741	107.5	-19.61
2450	122.9	-16.46
2900	131.6	-19.61
4000	150.0	-31.65



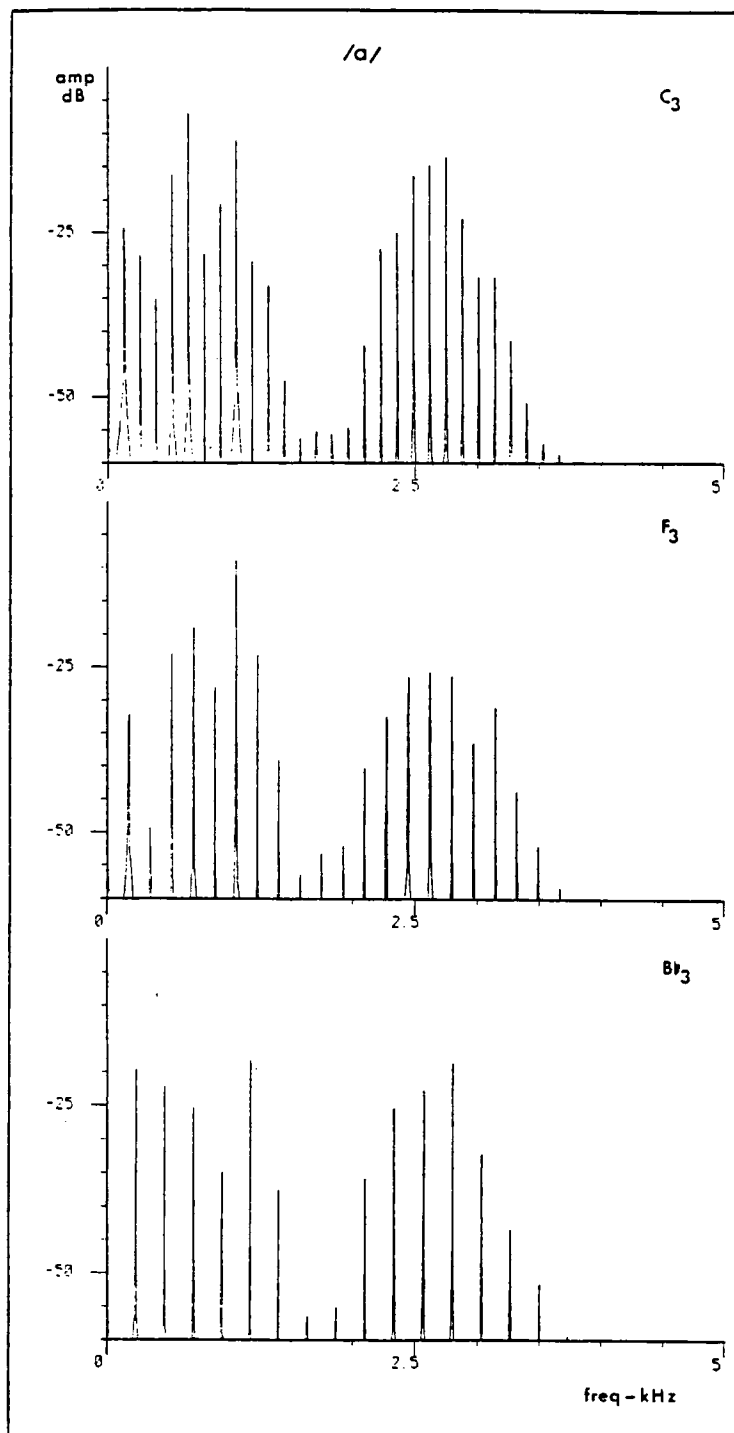
### 5.2.1.1 Stimuli

Tones were 2 sec in duration with 150 msec trapezoidal attack and decay ramps. The three vowels /a/, /o/ and /i/ were used. Since Ss were drawn from a multilingual pool, it was felt that these vowels were the closest to being common across languages. They are also quite common in the Western classical singing repertoire. In addition, these vowels are well separated in the classical "vowel space" that plots first formant frequency ( $F_1$ ; related to the closedness or openness of the mouth) vs second formant frequency ( $F_2$ ; related to the position of the tongue controlling the size of the mouth cavity).<sup>1</sup> The vowels were derived from a male singing voice and synthesized by the computer program CHANT according to a time-domain formant-wave-function synthesis algorithm developed by Rodet (1980a,b; Rodet & Bennett, 1980; see Appendix A for a brief description). The formant frequencies, bandwidths and relative amplitudes are shown in Table 5.1. Each vowel was synthesized at three pitches (fundamental frequencies):  $C_3$  (130.8 Hz),  $F_3$  (174.6 Hz) and  $Bb_3$  (233.1 Hz). The spectra of three vowels at each pitch are shown in Figures 5.1 - 5.3. Each of the stimuli was synthesized both with and without sub-audio frequency modulation. Due to the synthesis algorithm, any modulation in frequency was coupled to a modulation in amplitude such that a constant resonance structure was maintained throughout.

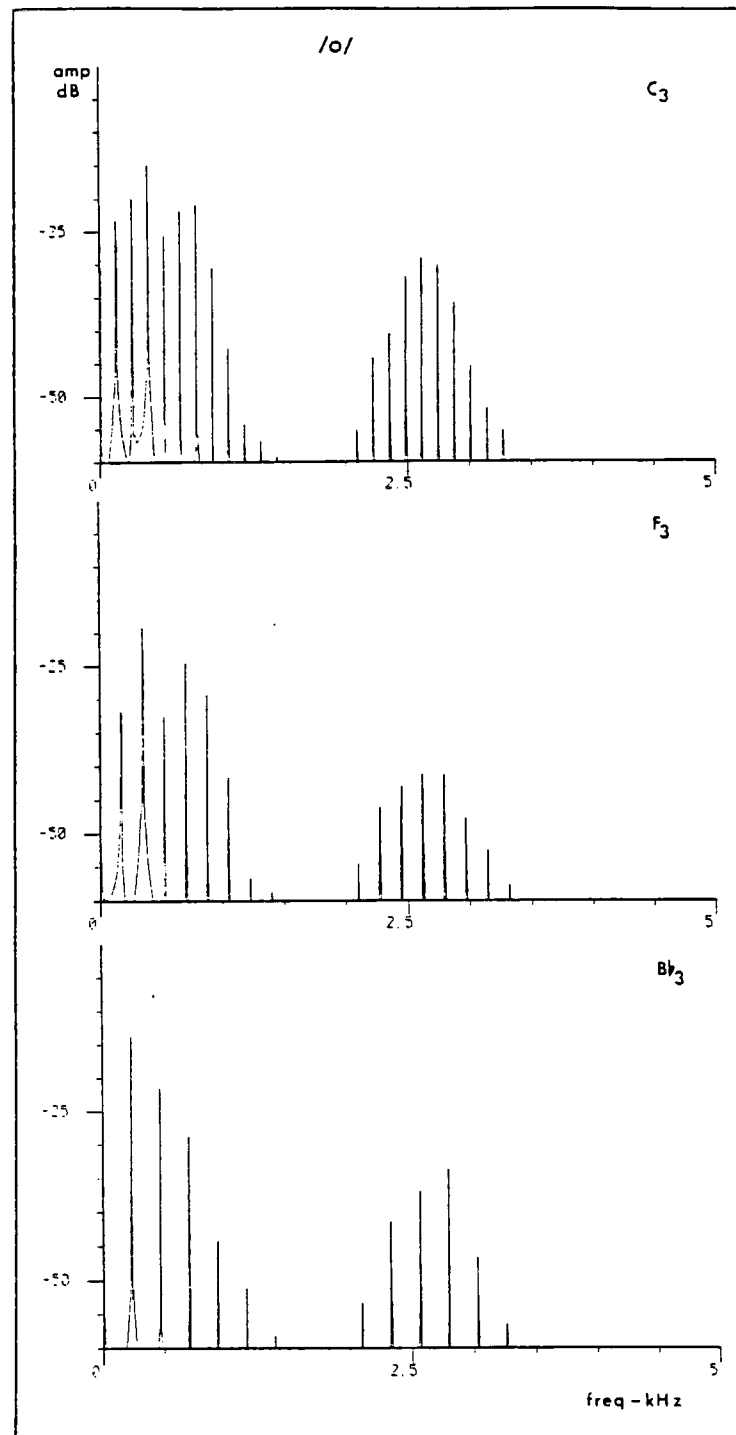
The modulation waveform was composed of a *vibrato* component (periodic) and a *jitter* component (aperiodic). The vibrato component was sinusoidal with a frequency of 5.1 or 6.3 Hz and an amplitude yielding a maximum frequency excursion of 1.5% of the frequency of a given partial. The jitter component had an rms amplitude of 0.8% of the center frequency and had a spectral content resembling those of the jitters described in Appendix B. When this compound modulation was imposed, it was scaled over time, beginning with no modulation for the first 300 msec, followed by a linear growth to maximum modulation width at 700 msec and a constant modulation width thereafter.

The amplitudes of the stimuli were adjusted for equal loudness by the experimenter. Their rms amplitudes (in dB re: most intense vowel stimulus) are listed in Table 5.2. The presence of modulation had very little effect on the perceived

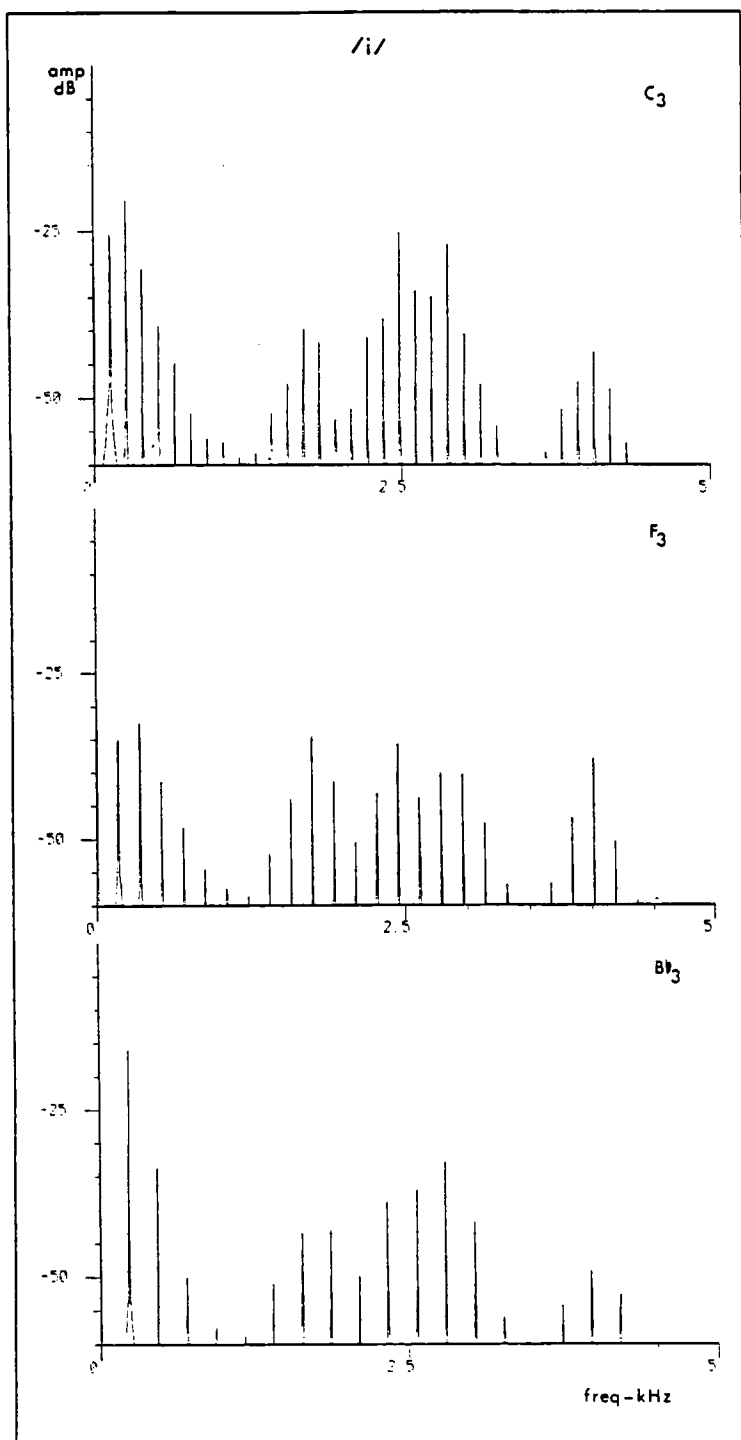
1. /a/ is an open-middle vowel. /o/ is a medium-back vowel. /i/ is a closed-front vowel.



**Figure 5.1.** Spectra from the steady state portion of the unmodulated vowel /a/ with  $F_0$  at  $C_3$ ,  $F_3$ ,  $Bb_3$ .



**Figure 5.2.** Spectra from the steady state portion of the unmodulated vowel /o/ with  $F_0$  at  $C_3$ ,  $F_3$ ,  $Bb_3$ .



**Figure 5.3.** Spectra from the steady state portion of the unmodulated vowel /i/ with  $F_0$  at  $C_3$ ,  $F_3$ ,  $Bb_3$ .

**TABLE 5.2.** Rms amplitudes (dB re: the most intense stimulus) of vowel stimuli at 3 pitches both with and without modulation.  $\overline{\Delta A}_{rms}$  is the average rms amplitude for each vowel across pitches and modulation states.

pitch	vowel			
	/a/	/o/	/i/	
$C_3$	-5.4	-1.9	-5.7	no modulation
$F_3$	-3.1	0.0	-5.5	
$Bb_3$	-5.0	-0.2	-3.8	
$C_3$	-5.8	-2.2	-6.1	modulation
$F_3$	-4.0	0.0	-5.8	
$Bb_3$	-4.9	-0.9	-4.8	
$\overline{\Delta A}_{rms}$	-4.7	-0.9	-5.3	

loudness;<sup>2</sup> modulated vowels were attenuated 0.4 dB on the average in relation to the unmodulated vowels. However, adjustments in intensity on the order of 2 dB were sometimes necessary to equalize the loudness at different pitches of a given vowel. On the average (across pitches) vowel /a/ was attenuated 3.8 dB and vowel /i/ 4.4 dB relative to vowel /o/. The /o/ stimuli were presented at an average rms sound pressure level of 75 dBA.

#### 5.2.1.2 Method and Results

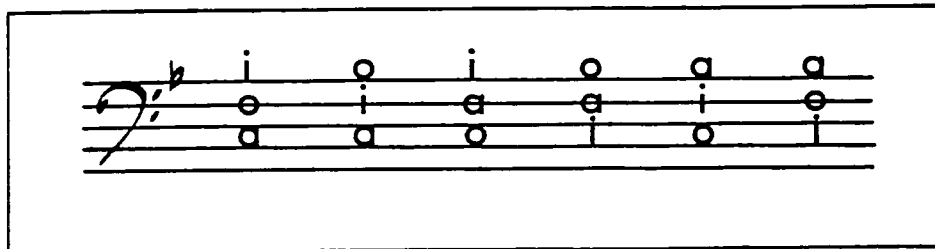
For "pre-calibration" of the vowel examples used, each of the 18 stimuli (3 vowels)  $\times$  (3 pitches)  $\times$  (2 modulation conditions: with and without) were presented 5 times in a randomized block to each subject. The subject's task was to identify the tone as /a/, /o/ or /i/ and to flip one of three switches accordingly. All subjects identified all vowels with perfect accuracy regardless of pitch and presence or absence of modulation. Some Ss felt the sound labeled as /o/ was closer to /u/. They were told that whenever they heard this sound in the main experiment, they were to label it /o/.

2. Loudness in the context of multiple sources will, of course, be changed by partial masking. Though the loudness masking procedure may appear superfluous at this point, an analysis of the perceptual data will be performed which, in principle, tests for mutual partial masking.

## Main Experiment

### 5.2.2 Stimuli

The stimuli from the pre-test were then combined to form chords of 3 different simultaneous vowels at the 3 different pitches; the chord always consisted of the pitches  $C_3$ ,  $F_3$ , and  $Bb_3$  with some permutation of the 3 vowels /a/, /o/ and /i/. This gives 6 permutations as illustrated below:



These permutations were included to test the effects of fundamental frequency on each vowel's perceptual prominence (when unmodulated, the spectral form of the vowel is less well defined at higher  $F_0$ 's since the formants are not as well filled out; see Figs. 5.1 - 5.3), and to test the effects of masking between the different vowels. The pitches were chosen in order to minimize coincidence of partials between simultaneous sources and to minimize dissonance and roughness due to first and second order beats (cf. Plomp, 1976). The interval of a perfect 4<sup>th</sup> serves this purpose. It is also beyond the 1 - 3 semit pitch separation range found to be a limiting factor in voice separation by Brokx & Nooteboom (1982) and Scheffers (1983).

The interest here is in the *salience* or *prominence* of the vowel sounds under various conditions of modulation of a given vowel (the "figure") and of modulation of the rest of the complex: the two other vowels (the "ground"). A figure condition is defined as the imposing of an independent vibrato/jitter on the frequency components of a given vowel. Four conditions were used: /a/ modulated independently (*A*), /o/ modulated independently (*O*), /i/ modulated independently (*I*), no vowel modulated independently of the others (*N*). The "ground" comprises the vowels that are not chosen as figure. Two ground conditions were used with each figure condition: ground steady or unmodulated (*S*), and ground modulated (*V*). All of the modulation possibilities for a permuted chord are listed in Table 5.3. In each cell, the modulation

**TABLE 5.3.** Modulation state of the vowels under different figure-ground combinations

		Vowel Modulated Independently ("figure")			
		<i>N</i>	<i>A</i>	<i>O</i>	<i>I</i>
State of Other Vowels ("ground")	<i>S</i>	/a/ = none	/a/ = Mod1	/a/ = none	/a/ = none
		/o/ = none	/o/ = none	/o/ = Mod1	/o/ = none
		/i/ = none	/i/ = none	/i/ = none	/i/ = Mod1
	<i>V</i>	/a/ = Mod2	/a/ = Mod1	/a/ = Mod2	/a/ = Mod2
		/o/ = Mod2	/o/ = Mod2	/o/ = Mod1	/o/ = Mod2
		/i/ = Mod2	/i/ = Mod2	/i/ = Mod2	/i/ = Mod1

(see text for description of Mod1 and Mod2)

specifications for each vowel (regardless of pitch position) are shown.

**Mod1** and **Mod2** in Table 5.3 represent two independent modulation functions. **Mod1** was as described in section 5.2.1.1 with a 5.1 Hz, 1.5% vibrato and 0.8% jitter. **Mod2** had a 6.3 Hz, 1.5% vibrato and 0.8% jitter. The two jitter waveforms had similar statistical characteristics (spectrum and amplitude probability density function), but their cross-correlation for  $\tau = 0$  is very close to 0, i.e. they are temporally independent.

With ground = *S* and figure = *N*, no vowels were modulated. Otherwise, only one vowel was modulated with **Mod1** according to the value of figure. With ground = *V* and figure = *N*, all vowels were modulated coherently with **Mod2**, thus maintaining the frequency ratios among all of the partials in the complex. Otherwise, one vowel, as specified by figure, was modulated independently of the other two with **Mod1**.

These two ground conditions were included to investigate the differences between

1. moving vs. steady ground with no figure; is a coherently moving ground perceived differently from a non-moving ground with respect to the perception of its constituent sources or with respect to the target vowel?
2. vowels modulated against a steady ground vs. vowels modulated against a moving ground (as would be more the case in music); does the modulation state of the ground affect the prominence of the figure?

Following the discussion in section 5.1 one would expect for the conditions *NS* and *NV*, that it would be difficult to hear out any vowels that are spectrally obscured and that judgments of the salience or prominence of these vowels would be low. Any difference between these two conditions would most likely be attributable to the reduction in ambiguity of the spectral forms provided by the coupled amplitude and frequency modulations. For conditions with figure = *A, O, I* a significant increase in the salience judgment for that particular vowel is expected whereas less increase would be expected for the other two which make up the ground.

### 5.2.3 *Method*

Ten subjects with 4 different native tongues (English, French, Finnish and Roumanian) were run in the experiment and were paid for their participation. All Ss were at least bilingual with English as either a first or second language. Experimental instructions were given in either French (5 Ss) or English (5 Ss) according to the wish of the subject. Five Ss were highly trained musicians and five Ss reported having no formal training in music, though one considered himself an accomplished amateur pianist. All Ss reported having no hearing problems.

Each of the 48 stimuli (6 permutations)  $\times$  (4 "figure" conditions)  $\times$  (2 "ground" conditions) was presented 5 times in a randomized block design. Each stimulus was presented once before any stimulus was repeated. A trial consisted of a single chord presented repeatedly at approximately 75 dBA over headphones in a sound-treated studio (see Appendix A for a description of the sound presentation system). Ss were allowed to listen to the chord as long as necessary to make the judgments. The subject was informed that a complex tone would be heard with 3 pitches in a chord and that any or all or none of the sounds at these pitches might be the vowels /a/, /o/ or /i/ as heard in the pretest. The task was to adjust a sliding potentiometer to estimate on a linear scale the degree of salience or prominence of a given vowel, or their certainty that the given vowel was present. For each stimulus tone, three judgments were made on three separate potentiometers - one each for /a/, /o/ and /i/. The top position indicated that the vowel was "very prominent" or that the subject was "perfectly certain" that the vowel *was* present. This was coded with a value of 100 in the data. The bottom position, encoded as 0, indicated "not at all prominent", or "perfectly certain" that the vowel was *not* present. The Ss were advised to use the following strategy in order to make the judgments quickly:



1. focus on one vowel at a time and try to hear out that vowel at the different pitches (the clarity of these pitches depended strongly on the modulation context),
2. judge that vowel's prominence and then
3. focus on the next vowel, and so on.

This ensured that the subject was listening for, and trying to hear, the vowel currently being judged. Once all three judgments were made, a switch was closed and the positions of the sliders were registered. The switch was opened again for presentation of the next trial. The experimental session lasted 100 to 150 min depending on the self-pacing of the subject. Subjects were allowed to take breaks as they needed between trials when they felt their concentration lagging or felt fatigued.

The values of the five judgments for each stimulus were averaged for each subject and these mean prominence ratings were used as data for further analysis. To compensate for the fact that Ss used the range of the potentiometers differently, the data for each subject were normalized with respect to the mean and standard deviation over all of their judgments, i.e. over /a/, /o/ and /i/ judgments for all conditions. (This assumes, of course, that all Ss were using a linear scale.) Then the individual normalized data were scaled and translated so that the data in its final form for all Ss fell between values of 0 and 100. The data transform is expressed as:

$$D'_{si} = \frac{(D_{si} - \bar{D}_s)}{\sigma_s} \quad (5.1)$$

$$D''_{si} = \frac{(D'_{si} - D_{\min})}{(D_{\max} - D_{\min})} \cdot 100 \quad (5.2)$$

where,

- $D_{si}$  is the individual datum  $i$  for subject  $s$ ,
- $\bar{D}_s$  is the mean over all data for subject  $s$ ,
- $\sigma_s$  is the unbiased standard deviation over all data of subject  $s$ ,
- $D'_{si}$  is the datum  $i$  normalized with respect to mean and standard deviation of subject  $s$ ,

- $D_{\min}$  is the minimum value  $D'_{si}$  for all  $s$  and  $i$ .
- $D_{\max}$  is the maximum value  $D'_{si}$  for all  $s$  and  $i$ .
- $D''_{si}$  is the final normalized datum  $i$  for subject  $s$ .

This transform operation preserves the pattern of ratio relations among data for any given  $S$  but reduces the standard deviations within cells across  $Ss$  by approximately a factor of 2. This allows more sensitive comparisons across conditions to be made and allows patterns in the data common to all subjects to emerge.

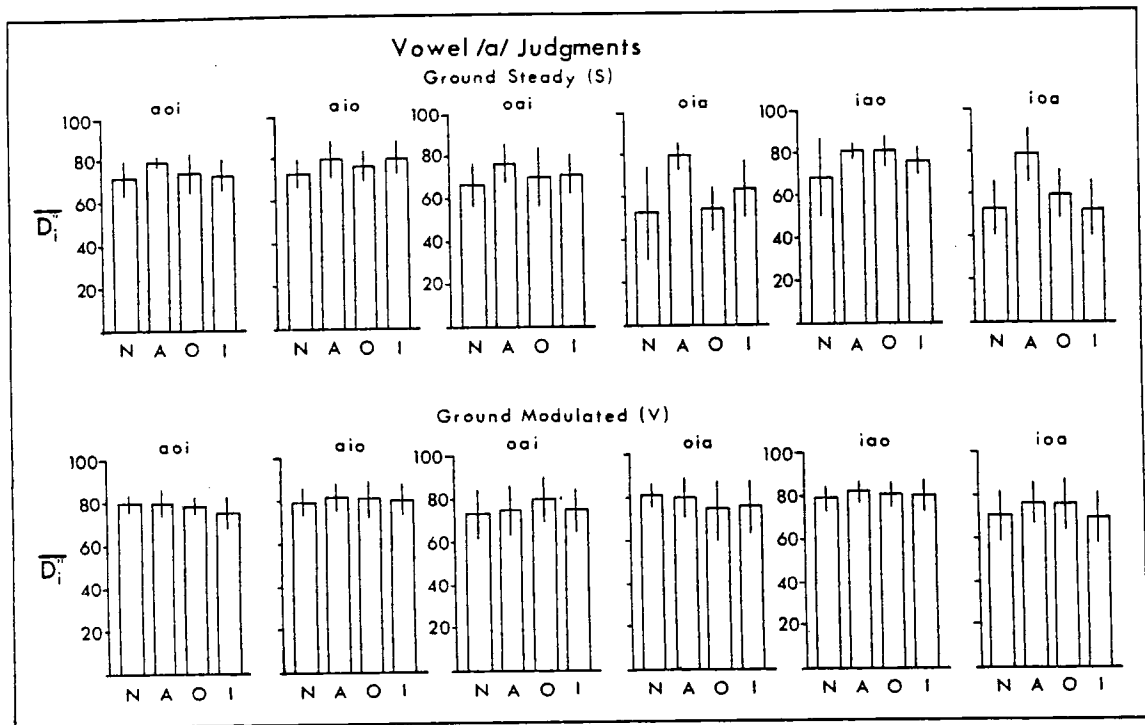
#### 5.2.4 Results

The means of the normalized data are listed in Table E.7 (App. E). These 48 means for each vowel are plotted in Figures 5.4 - 5.6 for a comparison of the effect of modulation condition within each vowel-pitch permutation.<sup>3</sup> Each graph represents one figure-ground combination. The permutations are notated to indicate the ascending pitch order of the vowels, e.g. *a oi* indicates /a/ at  $C_3$ , /o/ at  $F_3$ , /i/ at  $Bb_3$ . Each set of 12 graphs are the mean prominence judgments for one target vowel. The abscissa is  $\overline{D_i''}$ , the average  $D_{si}''$  across subjects (see Eq. 5.2).

It is important to remember in examining and interpreting these results that the data represent judgments made on a complex sound with all three vowels present while focusing on a single vowel. This means that the results for any stimulus are influenced by the stimulus as well as an intent on the part of the listener, i.e. trying to hear /a/ and then /o/ and then /i/.

3. A comparison of the 48 means for each vowel judgment between musically trained and untrained  $Ss$  yielded only two statistically significant comparisons (two-tailed  $t$ -test;  $H_0: \bar{x}_1 = \bar{x}_2$ ). This can be attributed to chance. Therefore, it is concluded that there is no difference between the two groups.

A possible criticism of the method would question whether  $Ss$  could make unbiased judgments, given that they guessed that /a/, /o/ and /i/ were actually always present. Since one subject,  $S_1$ , was the experimenter, who *knew* that this was the case, we can examine his data in relation to the means and standard deviations of the group. Only 30 of the 144 judgements had absolute  $z$ -scores greater than 1. No absolute  $z$ -scores were greater than 2. Of these 30, one half were positive and the other half were negative, indicating there was no tendency for the prominence judgments to be generally higher as a result of the prior knowledge of the stimulus set. We can conclude, therefore, that such knowledge has no effect on the ability to make an unbiased judgment on a given vowel's prominence in this context.

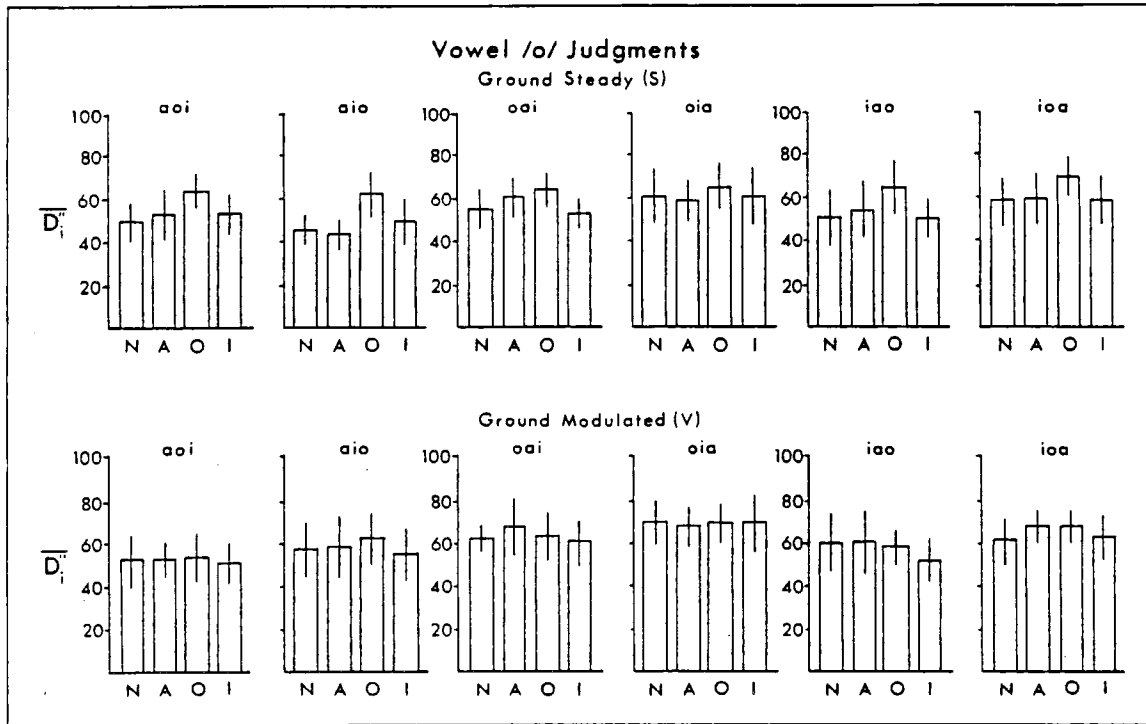


**Figure 5.4.** Experiment 8 data summary for prominence judgments on the vowel /a/. Each graph represents the data for a vowel-pitch permutation at a given *ground* modulation state. Each bar in the histogram represents the mean across subjects' normalized data for that particular *figure* modulation state. The vertical bar represents  $\pm 1\sigma$  from the mean.

#### 5.2.4.1 *Effect of modulation state of non-target vowels on prominence ratings of the target vowel.*

In examining the data for each vowel, it appears that the pattern of vowel prominence judgments for a given permutation is similar across figure-ground combinations in which the target vowel is not modulated. For example, /a/ judgments on *NS*, *OS*, *IS* stimuli are all similar for a given permutation. In all of these conditions, the /a/ is unmodulated. Also, the data patterns are similar for combinations in which the target vowel is modulated, regardless of whether it is the "figure" or part of the "ground", e.g. *AS*, *NV*, *AV*, *OV*, *IV* for /a/ judgments. In all of these conditions, /a/ is being modulated. To test for the equality of means across the figure-ground

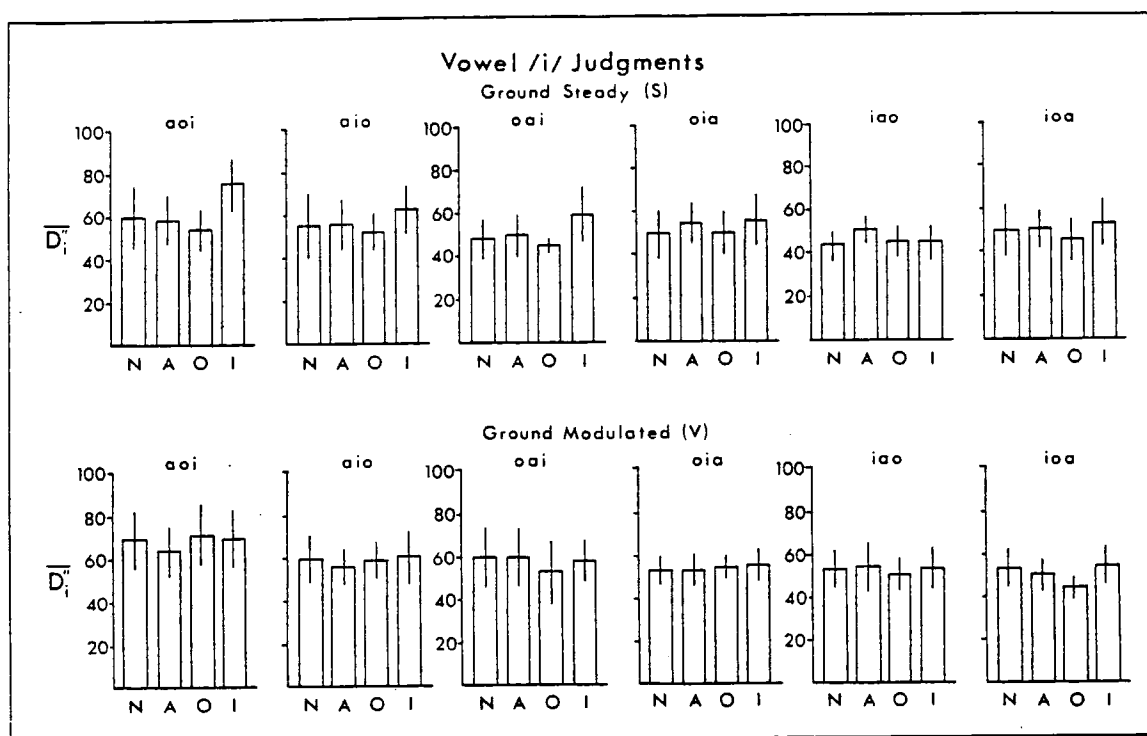
conditions within each of these two groups (modulated and unmodulated targets) two-tailed  $t$ -tests were performed ( $H_0: \bar{x}_1 = \bar{x}_2$ ) on all pairs.<sup>4</sup>



**Figure 5.5.** Experiment 8 data summary for prominence judgments on the vowel /o/. (See caption Figure 5.4.)

For the target unmodulated conditions only 3 out of 54 comparisons were found to be significant at the .05 level. This may be attributed to random variation, and thus, we may say that on the whole there is no essential difference between these conditions for a given permutation and target vowel. This indicates that whatever is being done to non-target vowels has no effect on the prominence of the target vowel when it is not modulated.

4. Normally, Student's  $t$ -test assumes equality of variances for the two samples. In these data, this condition was not always met, in which case a modified  $t$ -statistic is used whose criterion value is more difficult to meet (Pollard, 1977; pp. 161-163).



**Figure 5.6.** Experiment 8 data summary for prominence judgments on the vowel /i/. (See caption Figure 5.4.)

Similarly, only 10 out of 210 comparisons across all pairs of target modulated conditions within permutation are found to be significant at the .05 level. This may also be attributed to random variation. These results may be interpreted as indications that if a vowel is modulated at all, its judged prominence is, for the most part, unaffected by the behavior of the other vowels, whether they are part of "figure" or "ground" as defined for the stimulus manipulations in this experiment. It appears that the modulation state of the non-target vowels (whether modulating or no and whether coherent with the target or not) had no effect on the prominence ratings for a target vowel in a given permutation.

### 5.2.4.2 Effect of permutation of non-target vowels on prominence ratings of the target vowel.

Another question one might ask is whether the permutation of the two non-target vowels has any effect on the prominence of the target vowel. A difference might indicate that masking is playing a role in the perceived prominence since one permutation of the non-attended vowels may mask certain features of the attended vowel's spectrum better than the reverse. In fact, evidence for this effect is rather sparse across conditions. If we compare the means of permutation conditions within a given target vowel's pitch position (e.g. *aoiNS* vs. *aioNS* for /a/ at  $C_3$ , *oaiNS* vs. *oiaNS* for /o/ at  $C_3$ , *iaoNS* vs. *ioaNS* for /i/ at  $C_3$ , etc.), a significant difference is found in less than 17% of the comparisons. Table 5.4 shows the results of a series of two-tailed *t*-tests for those comparisons whose probability of occurrence is less than .05.

**TABLE 5.4.** Statistically significant comparisons of means for permutations within a given pitch position of the target vowel. In all cases  $n_1 = n_2 = 10$ .

Target vowel	Comparison	Modulation state	<i>t</i> ( <i>v</i> )	$p(\bar{x}_1 = \bar{x}_2)$
/a/	<i>aoi</i> vs. <i>aio</i>	<i>IS</i>	2.14 (18)	< .05
"	<i>oia</i> vs. <i>ioa</i>	<i>NV</i>	2.71 (10)	< .05
/o/	<i>aio</i> vs. <i>iao</i>	<i>AS</i>	2.24 (18)	< .05
"	<i>oai</i> vs. <i>oia</i>	<i>NV</i>	2.11 (18)	< .05
"	<i>aoi</i> vs. <i>ioa</i>	<i>AV</i>	4.16 (18)	< .01
"	<i>aoi</i> vs. <i>ioa</i>	<i>OV</i>	3.26 (18)	< .01
"	<i>aoi</i> vs. <i>ioa</i>	<i>IV</i>	2.82 (18)	< .05
/i/	<i>aoi</i> vs. <i>oai</i>	<i>NS</i>	2.10 (18)	< .05
"	<i>aoi</i> vs. <i>oai</i>	<i>IS</i>	2.84 (18)	< .05
"	<i>iao</i> vs. <i>ioa</i>	<i>OV</i>	2.14 (18)	< .05
"	<i>aoi</i> vs. <i>oai</i>	<i>OV</i>	2.69 (18)	< .05
"	<i>aoi</i> vs. <i>oai</i>	<i>IV</i>	2.18 (18)	< .05

If the effect of permutation relative to the pitch position of the target vowel were strong, yielding frequent differences between permutations that were independent of the modulation conditions, one might be inclined to interpret this as a simple masking effect, i.e. certain arrangements of the adjacent vowels resulted in more masking and obscured features essential for perception of that vowel. The absence of these differences in all cases would not necessarily discount this possibility, though. If

there was never any effect of permutation, regardless of modulation condition or pitch position of the target vowel, a possible interpretation might be that there was a homogeneity of masking due to the dense spectrum, and that such masking was unaffected by the presence of modulation. As the data indicate, neither of these extremes is true, though the results tend more toward the latter. One notes that for a given permutation comparison, certain modulation patterns are accompanied by significant differences between the mean prominence judgments, while others are not. This suggests that modulating a vowel may either increase or decrease its effect as a masker differentially with respect to its relation to the target vowel. There appear to be some systematic tendencies in the data (i.e. one permutation generally receives higher prominence judgments than the other for a given target vowel), however, comparisons are seldom significant statistically. For the purposes of clarification of more important effects, this effect of permutation of non-target vowels for a given pitch position of the target vowel will be considered of minor significance aside from pointing to the small role that frequency modulation may play in temporal masking and unmasking among multiple, simultaneous sources.

**TABLE 5.5.** Modulation forms for each vowel under different figure-ground combinations. ( $\Sigma_V = NV + AV + OV + IV$ ).

		Target Vowel		
		/a/	/o/	/i/
Pitch	$C_3$	$aoi + aio$	$oai + oia$	$iao + ioa$
	$F_3$	$oai + iao$	$aoi + ioa$	$aio + oia$
Position	$Bb_3$	$oia + ioa$	$aio + iao$	$aoi + oai$
Modulation	$U$	$NS + OS + IS$	$NS + AS + IS$	$NS + AS + OS$
State	$M$	$AS + \Sigma_V$	$OS + \Sigma_V$	$IS + \Sigma_V$

5.2.4.3 *Regrouping of the data*

The relative insignificance of these two kinds of comparisons (effect of non-target modulation states and of non-target permutations on prominence ratings of the target vowel) allows a collection of the data into categories among which more meaningful comparisons can be made. All of these categories are related to the parameters of the target vowel within a given stimulus. Accordingly, data are collected into target modulation state ( $U$  = unmodulated,  $M$  = modulated) by target pitch by target vowel judgment. Since the collections under modulation state and pitch position are with respect to the target vowel, a given combination is a collected datum from different stimuli for each target vowel. The stimulus conditions included in each datum are listed for each vowel in Table 5.5. In Table 5.6 the collected means ( $\bar{x}$ ) and pooled standard deviations ( $\sigma_p$ ) are shown. In each cell the  $\bar{x}$  and  $\sigma_p$  are derived from the stimulus combinations listed in Table 5.5. To obtain the pooled data values for a given target vowel, each pitch position combination is crossed with each modulation state combination in the same column. For example, the cell /a/  $U-C_3$  represents the  $\bar{x}$  and  $\sigma_p$  of 6 stimuli: 2 permutations ( $aoi$ ,  $aio$ ) under 3 modulation conditions ( $NS$ ,  $OS$ ,  $IS$ ).

**TABLE 5.6.** Means and standard deviations for data pooled according to target vowel's pitch position and modulation state for each target vowel. For  $U$ ,  $n=60$ ; for  $M$ ,  $n=100$ .

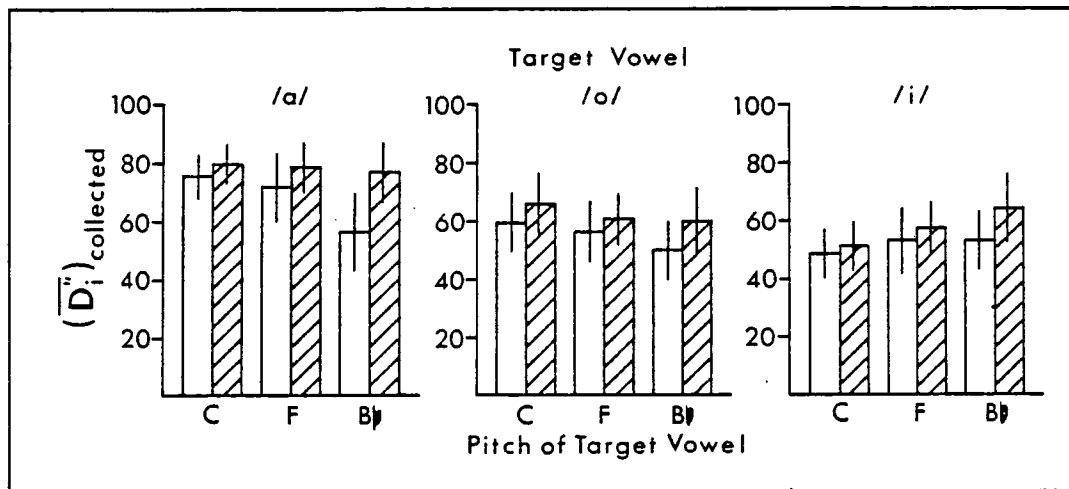
		/a/		/o/		/i/	
		$U$	$M$	$U$	$M$	$U$	$M$
$C_3$	$\bar{x}$	75.	79.	59.	66.	48.	51.
	$\sigma_p$	7.4	5.8	10.0	10.0	8.4	8.2
$F_3$	$\bar{x}$	72.	78.	56.	61.	53.	57.
	$\sigma_p$	11.4	8.0	10.5	9.0	10.8	8.8
$Bb_3$	$\bar{x}$	56.	77.	50.	60.	53.	64.
	$\sigma_p$	13.8	10.0	10.0	11.6	10.0	12.0

These data are plotted in Figure 5.7 to compare among pitch positions and modulation states.



#### 5.2.4.4 *Effect of the modulation state of the target vowel.*

One notes that in every case the mean normalized prominence rating of a modulated target vowel is greater than that for a similar unmodulated target. Every comparison (within target vowel and pitch position) between *U* and *M* conditions is significant at least at the .05 level (see Table 5.7 for the values of the *t*-statistic). The effect of modulating the target vowel is to increase significantly its prominence in a complex spectral background (either modulating or steady background).



**Figure 5.7.** Summary for data collected under pitch and modulation state of the target vowel. Bars for unmodulated conditions are clear, those for modulated conditions are hashed. The vertical bar represents  $\pm 1 \sigma_p$ .

#### 5.2.4.5 *Effects of pitch position of the target vowel.*

It is clear from studying Figure 5.7 that the relative prominence of a target vowel at the different pitches changes with the vowel's modulation state. It is also clear that the different vowels are affected differently by the pitch position they are placed in. For vowels /a/ and /o/, prominence always decreases with increasing fundamental frequency, but this decrease is less when the vowel is being modulated. For these vowels, one effect of modulation is to reduce the differences due to pitch position. For the vowel /i/, prominence tends to increase with increasing fundamental

**TABLE 5.7.** Values of the  $t$ -statistic for comparisons within pitch position for means collected into pitch and modulation state of the target vowel. In all cases,  $n_1 = 60$  ( $U$ ) and  $n_2 = 100$  ( $M$ ). The value in parentheses is  $v$ . (The variation in the computed value of  $v$  is related to the inequality of variance between samples being compared.) For all of these  $t$ -values  $p(\bar{x}_U = \bar{x}_M) < .01$ .

Pitch	Vowel		
	/a/	/o/	/i/
$C_3$	3.78 (100)	4.56 (158)	2.21 (158)
$F_3$	3.56 (92)	3.42 (158)	2.34 (105)
$Bb_3$	9.97 (94)	5.55 (158)	6.03 (158)

frequency. An effect of modulation for this vowel is to increase the difference due to pitch position. For all three vowels, the greatest change in prominence with modulation occurs with the  $Bb_3$  position (the highest note, where the partials have the greatest spread). Table 5.8 lists the differences between means for modulated and unmodulated target vowels at each pitch. The changes due to modulation are relatively small for  $C_3$  and  $F_3$  compared to those for  $Bb_3$ .

**TABLE 5.8.** Differences between means for modulated and unmodulated target vowels at each pitch. Each cell value represents  $M - U$ .

Pitch	Vowel		
	/a/	/o/	/i/
$C_3$	4.	7.	3.
$F_3$	6.	5.	4.
$Bb_3$	21.	10.	11.

The spread of spectral energy (and thus presumed maskability) is quite different for each vowel, with /a/ having the least spread followed by /o/ and then /i/ whose energy is very widely distributed across the 4 kHz range. It should be recalled that these vowels were matched for equal loudness *in isolation*. Therefore, it is likely that the major differences between the 3 vowels can be attributed to masking effects.

The results of this experiment may be summarized as follows:

1. The judged prominence of a target vowel at a given pitch is unaffected by the pitch position and modulation state of the other two vowels in the chord, regardless of the modulation state of the target.
2. The judged prominence increases significantly when the target vowel is modulated over when it is not modulated.
3. The degree of increase in judged prominence is a function of the vowel's fundamental frequency (or position in the chord). For all vowels, the greatest increase is found for the highest position ( $Bb_3$ ).
4. There are significant differences between the mean prominence ratings at different pitches for each vowel. For /a/ and /o/, prominence decreases with increasing fundamental, whereas for /i/, prominence increases with increasing fundamental.
5. Vowel /a/ is always judged more prominent than /o/ and /i/ at all pitches. Vowel /o/ is judged more prominent than /i/ at the two lowest fundamentals, but less prominent at the highest fundamental.

#### 5.2.5 Discussion

Some of these results are surprising, given the *a priori* assumptions, and several issues are brought into question.

##### 5.2.5.1 Pitch Position and Modulation State of Non-target Vowels

Neither of these parameters of non-target vowels had any systematic effect on the judged prominence of a given target vowel. If, for example, /i/ was positioned at  $Bb_3$  and was not modulated, its judged prominence was unaffected by whether or not /a/ and /o/ were modulated, though there was a slight tendency for it to be judged more prominent if /a/ was at  $C_3$  and /o/ at  $F_3$  than vice versa. Two things are implied here:

1. Masking effects specifically due to the pitch arrangement of the non-targets are very slight, i.e. masking tends to be relatively homogeneous for the vowel-pitch combinations used in this study. This supports claims by Brokx & Nooteboom (1982) and Scheffers (1983) that once the pitch separation of two voice-like sources is greater than 1 – 3 semitones (5 semitones in this case), pitch separation has reached its maximum effect as a cue for source distinction. Although Darwin (1981) and Scheffers (1983) claim that when vowels are heavily overlapped spectrally, listeners use  $F_0$  differences to extract them, evidence to the contrary is provided by the present study.
2. No additional release from masking occurs due to modulation of the non-target vowels which would increase the prominence of the target vowel. This is somewhat surprising since one might imagine *a priori* that modulation of a non-target, resulting in its emergence as a more clear source image would simplify the situation, making the target vowel more easily extractable. Obviously, the evidence from the present study indicates the contrary. Further studies, varying the potential masking relations among the several simultaneous sources, are needed to follow out this result.

One positive aspect of these results is that the judged prominence of a given target vowel would seem to be independent of the specific position or behavior of the other vowels, as well as of the prominence judgments made on those vowels.

#### 5.2.5.2 *Modulation of the Target Vowel and Coherence with Non-targets*

The data indicate quite clearly that regardless of other stimulus conditions present, when a target vowel is modulated, its judged prominence increases significantly over when it is not modulated. One could conclude that some cue or cues associated with coherent frequency modulation can be used for grouping decisions. However, the relative coherence of modulation of one vowel with another had little or no effect. If coherent frequency modulation alone were responsible for grouping, we would have expected "ground" vowels (those modulated coherently with respect to one another) to be difficult to separate from one another. But the fact that coherent modulation across source sub-groups does *not* reduce separation with these stimuli suggests that the coupling of frequency modulation with a resonance structure is playing a stronger role than previously considered. This is evidenced by

the fact that the prominence of all ground vowels always increases for *V* conditions compared with similar *S* conditions.

### 5.2.5.3 *Spectral Form Stability*

It was suggested in Chapter 1 that constant spectral envelope was likely to be a cue important in grouping. It is certainly the case that the spectral form is the information from which vowel identity is derived. We assume that vowel prominence judgments are closely tied to the ability of a listener to extract a spectral form from the complex spectrum and identify it as such. Factors that may inhibit this extraction would include

1. lack of definition of the spectral form by the frequency components composing the vowel, as is the case, for example, with a higher  $F_0$ .
2. inability to separate the components "carrying" the spectral form because they were grouped with other elements that then distorted this form, or
3. masking of features essential for vowel recognition and identification, i.e. the lower two or three formant peaks (Carlson, Fant & Granström, 1975; Karnickaya, Mushnikov, Slepokurova & Zhukov, 1975).

With the synthesis algorithm used, each group of components representing a vowel is modulated under a constant spectral envelope, i.e. the resonance structure is unchanging. With no modulation, the nature of the resonance structure may be ambiguous depending on the number of partials contained in each formant band. As modulation is added, each partial's frequency-amplitude motion is potentially providing information about the slope of the spectral envelope in that region. When these components are taken as an ensemble, this slope information greatly reduces the ambiguity of identity. We would expect a reduction in ambiguity to be accompanied by an increase in prominence judgments. This is confirmed by the large increase in prominence with modulation for the highest pitch of each vowel. With no modulation and a higher fundamental, there are fewer components within the formant bands and the spectral form is thus less well defined. With modulation, this structure is more clearly implied by the coupled frequency-amplitude motions. For adjacent partials belonging to separate vowels, these motions might be incompatible and indicating

separate formant structures. In essence, each sub-group of partials is tracing its own spectral envelope. The extent to which these envelopes could be separated would influence the judged prominence of each of them. Given that the tracing of a spectral envelope by a group of components provides enough information to identify the vowel, one is led to conclude that vowel identification for these stimuli can take place independently of the coherence or lack of coherence of modulation *across* the separate vowels. Indeed, this suggests (following Huggins, 1952, 1953) that a constant spectral form, implying a stable resonance structure, gives rise both to information that such a stable structure is present as well as to information about its nature, i.e. spectral identity. Therefore, these two aspects, source structure and spectral quality, are perfectly tied to one another. But other, more temporal aspects, more closely related to the excitation of the source structure, may be processed separately. This is supported by the findings of Cutting (1976) and others that certain aspects of speech perception (or, more generally, voice perception) result from processes that are separate from those involved with other aspects of auditory perception.

#### 5.2.5.4 *Pitch Position of the Target Vowel*

Without modulation, one notes that the prominence of the different vowels depends a great deal on the pitch position. For vowels /a/ and /o/ there is a decrease in prominence with increasing  $F_0$ . This is most likely due to a decrease in spectral definition of the vowel formants which would logically result in a degraded perception of the vowel quality. This result is not found for the vowel /i/. It may be in this case that since the spectral energy for /i/ is so spread out, it is almost always heavily masked by the other vowels. However, with modulation, all vowels show the greatest increase in prominence for the highest pitch position,  $Bb_3$ . This supports the hypothesis that in light of ambiguity of spectral definition, frequency modulation (vibrato and/or jitter) coupled with a resonance structure reduces the ambiguity since the component amplitudes trace the spectral form. As argued in Chapter 1 (section 1.6.2), this challenges the claims of Carlson, Fant & Granström (1975) and of Sundberg (1977) who reported increases in uncertainty of vowel identity with frequency modulation. Clearly, the opposite is demonstrated here.

#### 5.2.5.5 *Perceived Pitch of the Vowel Sources*

There is yet another cue coupled with frequency modulation which may also be playing a role in source perception: the harmonicity of the frequency components of each vowel. There are many auditory models and theories based on much data that supports the hypothesis that there is some predisposition toward processing of harmonic sound sources by mammalian auditory systems. Certainly, any ethological theory about predisposition toward perception of species-specific vocalizations (cf. Worden & Galambos, 1971) would have to include the harmonicity of the signal in the specification of auditory "templates" or "feature-detectors" of human speech (whether learned or innate). Of interest here is the possibility that the harmonicity of a subset of partials may be a cue for separating it from the rest of the spectrum. There is some suggestion in the previously cited work of Chowning (1980) and McNabb (1981) that coherent sub-audio frequency modulation increases the apparent fusedness and naturalness of synthetic voice and instrument sounds. It seems possible for a sensory system most often processing dynamic, rather than steady-state, signals that a coherently modulating harmonic series is somehow less ambiguously harmonic (or less ambiguously a harmonic *group*) than a steady-state harmonic series. This is lent support in the present experiment by reports of some musical Ss that there was a vast difference between the various stimuli with respect to the pitches perceived. They were told to expect three pitches (there were, in fact, three harmonic series), but reported sometimes hearing four to six pitches. In verifying this with different sets of finely tuned musical ears and certain stimuli, it appeared that this occurred most often with stimuli where two or three of the vowels were steady. When all vowels were modulating, subjects reported that the three pitches were more clear and unequivocal. Modern pitch theories would all have us believe (and our ears usually agree in the right context) that the most unambiguous or unequivocal pitch sensation in complex tones is perceived with the harmonic series. If FM coupled with a harmonic series reduced the ambiguity that any sub-group was or was not harmonic, we would expect a decrease in ambiguity or equivocality of pitch perception. Several authors have noted that the virtual pitches of certain stimuli with little or no energy at the  $F_0$  are better heard when the complex is modulated coherently than when it is stable (Thurlow & Small, 1955; Plomp, 1976). That this appears to be the case in the present study as well supports the hypothesis that perception of at least some (not specifically vocal) source qualities are dependent on the properties of the ensemble of elements collected as a group.

It is difficult to determine from this experimental design the extent to which each of the factors of coherent FM, spectral form and harmonicity are separately contributing to source image formation and separation, though it seems quite likely that the processing of a vowel spectral form is independent of the processes of other source grouping cues. In the stimuli used in this experiment, the subgroups within which all three cues are perfectly coupled and operating together constitute the three vowel spectra. For example, even when all three vowels are being coherently modulated together, the whole ensemble is inharmonic. Also, there are incompatibilities in the amplitude modulation patterns of adjacent partials belonging to separate vowels that arise as each partial follows its own spectral form. Conversely, within each vowel there is harmonicity and coherence of frequency modulation under a single, constant, familiar spectral structure. It may be illustrative to consider the consequences of stimulus compositions in which these cues are not working so closely together or are even opposing one another in the organizations they suggest.

It is not possible to uncouple FM from stability of spectral form, though one could have *incoherent* FM on all partials which were tracing the same vowel formants. This would test whether the coherence of FM was necessary for the vowel shape being traced to be selected as a group or whether the tracing itself would provide enough information independent of any source formation processes for the vowel to be identified. For example, in Cutting (1976) certain stimulus configurations were constructed where the identity remained constant even though the number of sources reported changed. In demonstration examples (McAdams & Wessel, 1981), isolated vowels were constructed with incoherent amplitude modulation on each partial and with modulation widths between 1% and 3%. The effect is one of many voices singing the same vowel at the same pitch (a "chorus" effect); more evidence for the separate processing of sources and speech properties.

It would be possible to uncouple spectral form stability and harmonicity. Inharmonic frequencies could be made to conform to a certain vowel shape while being modulated coherently or incoherently. Cohen (1980) has shown that "stretched" inharmonic tones are perceived as being less fused than harmonic tones. If harmonicity were necessary as a grouping cue, the vowel would be difficult to hear, whereas, again, the tracing of the vowel envelope itself might provide sufficient information for identification indicating that vowel identity perception is not necessarily tied to perceived fusion.<sup>5</sup>



One test that would make the present experiment more complete would uncouple the spectral envelope tracing from the frequency modulation, as was done in Experiment 7. In this case, particularly with coherent modulation of all 3 vowels, one would expect vowel prominence to be degraded since the tracing information believed to be responsible for the increase in prominence would be perturbed, while at the same time the relative amplitudes of the harmonics would remain the same. Since the modulation widths used in this experiment are below those accompanied by defusion of the source (Expt. 7) with constant component amplitudes, this combination would allow a test of the relative roles of spectral form tracing and coherent modulation. Of course, the synthesis method used in this experiment would not be suited for such stimulus manipulations since the spectral form tracing is an inherent part of the synthesis algorithm.

What has been lacking in previous work (and lacks also in the present experiment) is a simultaneous testing of both source multiplicity judgments and identification and association of perceived qualities to the various sources. This kind of experimentation would help sort out many of the questions about the apparent independence of some perceived qualities from the source groupings themselves. At present, it appears that voice sounds are independent of the grouping processes, whereas non-speech timbres and pitch are dependent on these processes.

Relevant to these considerations is a notion that will be entertained in Chapter 6 on the "multi-leveled" nature of auditory image formation. According to this notion, an "image" may be defined at an arbitrary level of complexity (and, by implication, multiplicity). The hope is to extend this metaphor in order to be able to encompass some of the apparent conflicts presented in this chapter by claiming that one process (such as speech sound perception) is operating at a higher level of image formation, which may take as input the collection of outputs from lower level processes such as source groupings on the basis of frequency modulation coherence and harmonicity or periodicity.

One complaint that may be leveled against these considerations of the uncoupling of the behavior of certain parameters is that these kinds of uncouplings never occur

---

5. Pilot studies have demonstrated that the latter is true (McAdams, 1980, 1982, App. F).

in the "real" (or "everyday") world and are thus not of interest for an understanding of "normal" perception. I would counter with a theoretical and a practical argument. Theoretically, I am operating from a constructivist position whose belief is that the perceptual systems analyze the incoming sensory input, extract important features based on both innate and learned criteria, and then reconstruct a salient representation of the behavior of the world based on current features of the world, ongoing expectations of what it is doing, and prior knowledge of the normal behavior of the world. This assumption is dealt with to some extent in Chapter 6, but I am not prepared to go into lengthy detail on the philosophical and theoretical implications of this position within the framework of this dissertation.

The more practical argument, which will be dealt with in some detail in Chapter 6 concerns an envisioned application of these principles to music composition and synthesis by electroacoustic means. In such a situation, one *does* have, and is compelled to exploit, the uncoupling of the various stimulus parameters that contribute to the formation of a musical image or images. Here, ambiguous and polyvalent perceptions are often strived for. Having a theoretical framework to guide one's approach to microcomposition of the musical image, then, holds the possibility of being a powerful tool in the hands of an ingenious composer, and provides a good argument against those who propose a greater emphasis on "mere" physical modeling of musical sound sources in music synthesis. In the latter case, one would not have the fine control over image formation that is proposed by the present experimental results.

## CHAPTER 6

### The Auditory Image: A Metaphor for Musical and Psychological Research on Auditory Organization

Let us now consider the application of the auditory image metaphor to musical and psychological research on auditory organization. I will select several pertinent examples to circumscribe the nature of sequential and simultaneous organizing processes and to illustrate the essential differences between them. Then I will return to the notions of the auditory image and the coherence of behavior of a sound entity to see how far we can push the metaphor at this stage.

#### 6.1 Sequential Organization

Research on sequential organization of sound is concerned with how the structure of a sequence of events affects the perceived continuity of the sequence. That is, under what conditions is a sequence of sounds heard as one or more "streams"? Bregman & Campbell (1971) employed the metaphor "stream" to denote a psychological representation of a sequence of sounds that can be interpreted as a "whole", since it displays an internal consistency, or continuity. Van Noorden (1975) termed this continuity "temporal coherence," i.e. the events in the sequence cohere as a perceptual structure through time. In general, we may consider that a stream represents the behavior of a real or virtual source of sound. This is consistent with the notion of image - a stream is an image of a source whose emanations are extended across several events in time, i.e. a melody is a stream is an image. Implied here is the possibility that a single sequence of tones can be organized (grouped) as more than one stream. This case is particularly common in music for solo instruments of the Baroque period (cf. the violin partitas of J.S. Bach). In these compositions, the soloist sometimes alternates rapidly between registers or strings on successive notes, and

what one hears is two melodies that appear to overlap in time.

It is also possible to have a situation where a listener can switch between hearing a sequence as one or two streams by changing attentional focus. Many of the more interesting instances of this in music have such multiple perceptual possibilities. It is an important point psychologically to note that in such cases a listener may hear one organization or the other but not both at once. In other words, I can hear the sequence as one stream or as two streams (and switch my attention between each of the two streams at will), but I cannot hear the sequence as both one *and* two streams simultaneously. These are mutually exclusive organizations.

There are several important properties exhibited by a stream. These are discussed more fully elsewhere (McAdams & Bregman, 1979), so I will merely summarize them.

1. *It is possible to focus one's attention on a given stream and follow it through time*; this means that a stream, by definition, exhibits temporal coherence.
2. *The parsing of a sequence into smaller streams takes a certain amount of time to occur*: it generally takes several notes into a compound melody line until the separate registers are relegated to different streams. It appears that the perceptual organizing processes assume things are coming from one source until they accumulate enough information to suggest a different interpretation of how the world is behaving.
3. *It is easily possible to order the events of a stream in time, but it is more difficult to determine the relative order of events across streams*. Two streams resulting from the same sequence of notes appear to overlap in time, but it is hard to say exactly how they are related temporally. Since temporal ordering of notes is an essential determinant of a melody, this means that a melody is, by definition, a stream, i.e. a melody has a perceptual unity (temporal coherence). This also implies that not just any arbitrary sequence of tones constitutes a melody; if the sequence is not temporally coherent, it is not heard as a melody (but maybe as two or more melodies!). This property refers simply to a single sequence and is not meant to imply, of course, that we cannot perceive the temporal relations between rhythmically, synchronized melodic lines. The

simultaneous occurrences of events in the separate streams can serve as temporal anchors to relate the two sequential structures.

4. *A given event can be a member of one or of another concurrent stream but not both simultaneously.* As mentioned above, one might switch between hearing an event belonging to one organization and then to another. The important point here is that several parsing schemes cannot be used at the same time.

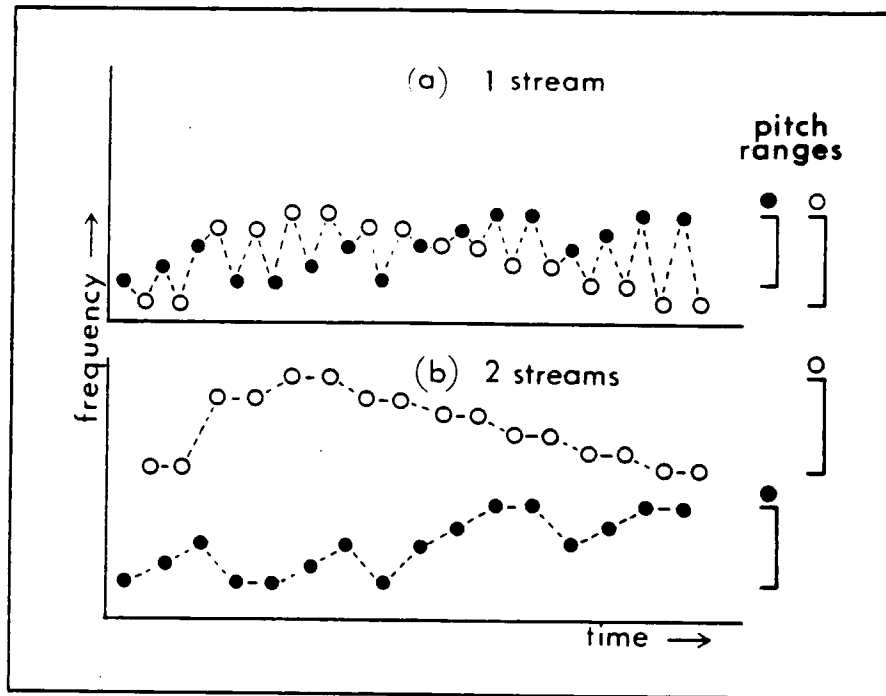
The main acoustic factors which have been found to be used by the perceptual system to build descriptions of streams include frequency, rate of occurrence of events (tempo), amplitude, and spectral content and form, i.e. the frequencies present in a complex tone and their respective amplitudes. It is not possible to go into great detail about all of these factors. Simple illustrations will be given here and the reader is referred to the review articles (cf. Bregman, 1978a,b).

#### 6.1.1 *Frequency Separation*

It has been shown repeatedly with sine tone sequences that the relative frequency separation between tones influences the formation of stream organizations (Bozzi & Vicario, 1960; Vicario, 1965, 1982; Bregman & Campbell, 1971; Dowling, 1973; Deutsch, 1975; van Noorden, 1975, 1977; Bregman, 1978a). At a given tempo, tones that are further apart in frequency are more likely to be heard in separate streams than those that are closer together. Also, at a given frequency separation the role of tempo is such that faster sequences have more of a tendency to split into multiple streams than slower sequences. There is a kind of trade-off between tempo and frequency separation.

A compelling example of the frequency separation effect is illustrated in Figure 6.1. This example is based on an experiment by Jay Dowling (1973) where he interleaved (alternated) the notes of two familiar nursery rhyme melodies. When the ranges of the pitches of the two melodies are the same (see Figure 6.1a) it is difficult to hear out the separate melodies and one melody is heard which is a combination of the two. However, when the frequency ranges are sufficiently separated, one easily discerns the two melodies as is indicated in Figure 6.1b. In this and similar succeeding figures, the dashed lines between tones indicate temporal coherence. For example, in Fig. 6.1b, the third tone is perceived as following the first tone rather than the

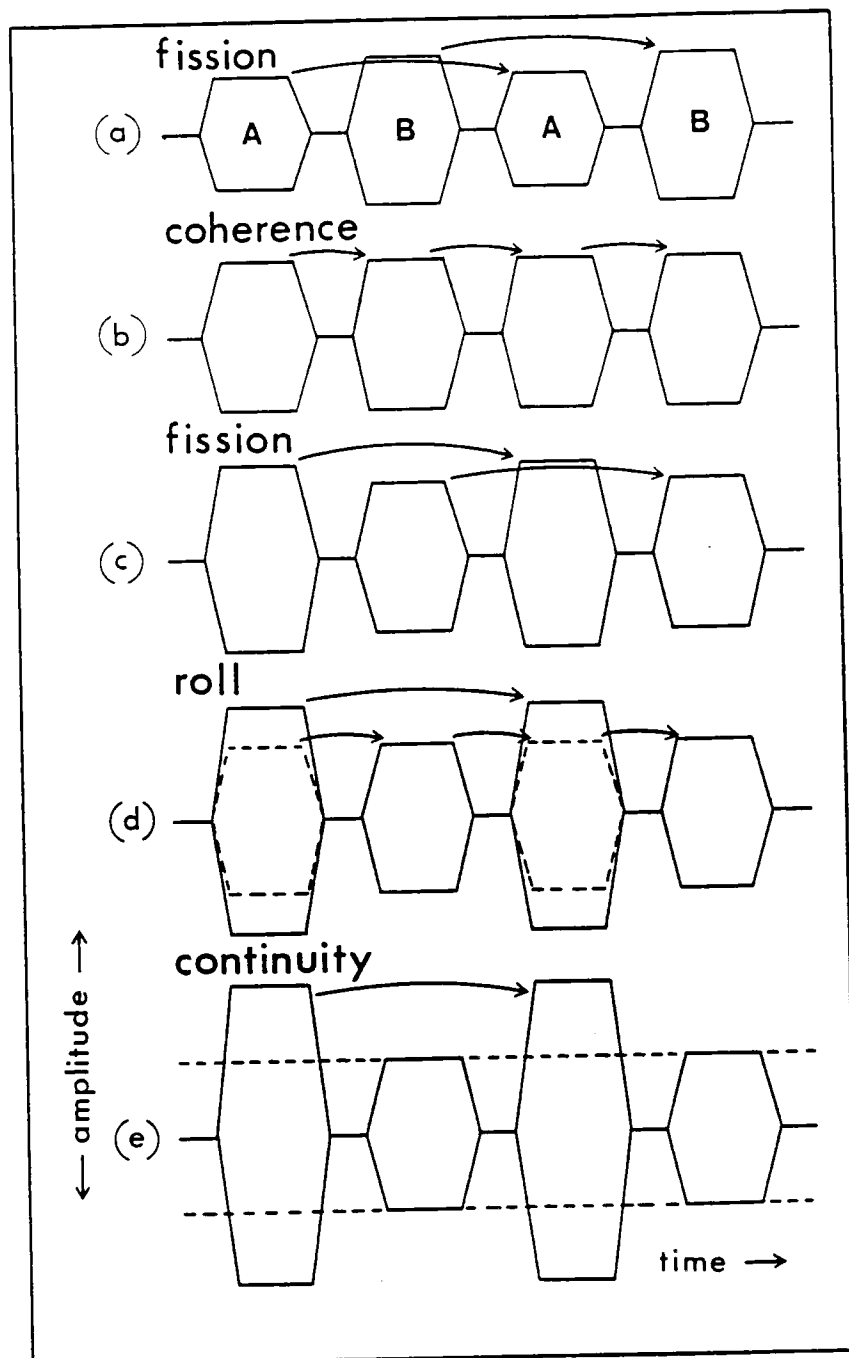
second tone. In Taped Example 1,<sup>1</sup> the two melodies are played at four different separation values: the first and last are as shown in Figure 6.1. Here the identification



**Figure 6.1.** The tones of parts of two common nursery rhyme melodies are interleaved. In (a) the frequency ranges of the two melodies are similar and the sequence is heard as one, unfamiliar melody. In (b) the frequency ranges of the melodies are non-overlapping and each melody is heard independently. The dotted lines indicate temporal coherence (perceived sequential organization). [derived from Dowling, 1973]

of the melody as a whole entity is dependent on being able to separate its elements from the other melody and to hear them as a group. The streams that are formed on the basis of frequency separation *are* the melodies.

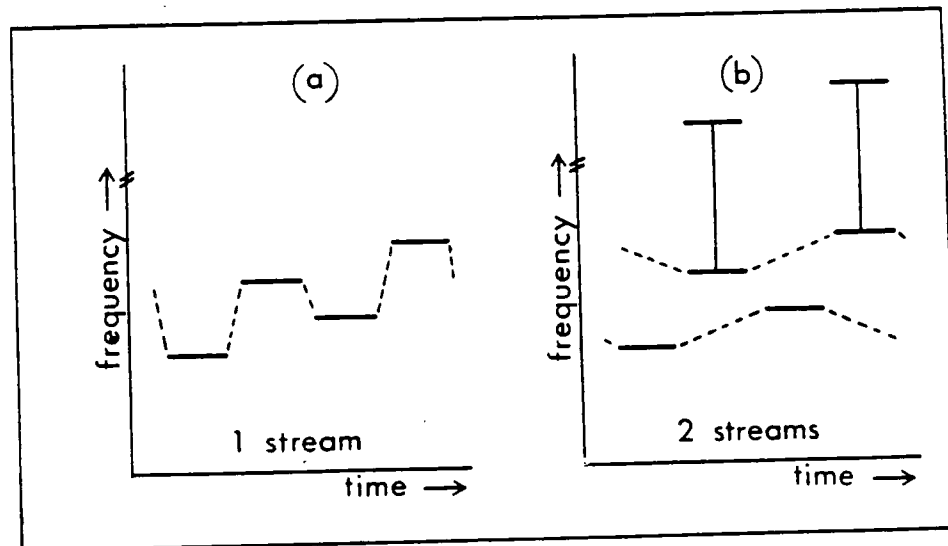
1. A description of the sound examples is to be found in Appendix G.



**Figure 6.2.** Illustration of the different percepts resulting from the alternation of two sinusoidal tones of identical frequency and duration. The amplitude of tone A is varied relative to that of tone B. The shapes in the figure represent the amplitude envelopes of the tones. [derived from van Noorden, 1975]

### 6.1.2 Amplitude Differences

Another factor that can contribute to stream formation is the relative amplitude of the tones, though this is a much weaker effect than the rest. Van Noorden (1975), for example, has demonstrated many perceptual effects resulting from the alternation of two identically pitched pure tones that differ in amplitude. These effects depend on the degree of difference in amplitude, and they range from hearing (1) a fission of the sequence into two pulsing streams (one soft and one loud; Fig. 6.2 a,c), to (2) to a coherent stream at twice the tempo (Fig. 6.2b), to (3) to a single loud stream at one tempo plus a soft stream at twice that tempo ("roll"; Fig. 6.2d), to (4) to a loud pulsing stream plus a continuous soft tone ("continuity"; Fig. 6.2e). These amplitude-based effects are also dependent on tempo and frequency separation.

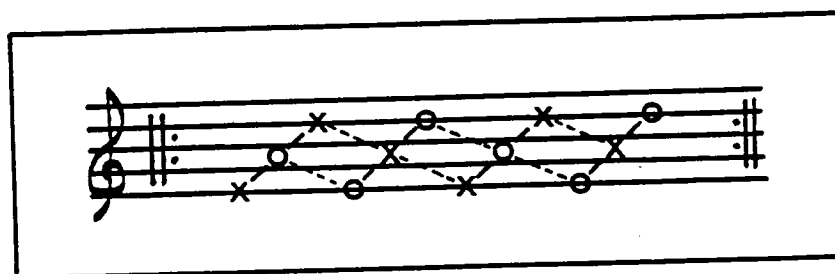


**Figure 6.3.** Effect of differences in spectral composition on sequential organization. In (a) all tones are sinusoidal and the frequency separation between them is adjusted so that a single stream percept may be heard, as indicated by the dotted lines. In (b) the third harmonic is added to one pair and the spectral difference causes two streams to form, each with a different timbre. Each of these 4-tone patterns was recycled continuously. [from McAdams & Bregman, 1979]



### 6.1.3 Spectral Form and Content

The last factor to be discussed that contributes to stream formation is spectral form and content. A stimulus sequence can be constructed where the spectral composition is very similar (all tones are sinusoidal, as in Fig. 6.3a) and the frequency variation from tone to tone is small enough so that the sequence can be heard as one stream. By adding a harmonic (the third, in this case) to certain tones in this sequence, those tones are made to form a separate stream (Fig. 6.3b). The solid vertical bar denotes the fusion of the spectral components into one percept. In Taped Example 2, you may hear first the stimulus cycle in Fig. 6.3a and then the cycle in Fig. 6.3b. Note also that the tempo of the new streams is half that of the original stream. This illustrates that perceived rhythm is also dependent on the stream organization, i.e. rhythm may be considered as a quality of a given stream.



**Figure 6.4.** When all of these tones are played by the same instrument, ascending pitch triplets are heard (solid lines). But when two instruments with very different spectral forms each play the X's and O's, respectively, descending triplets are heard (dotted lines). [from Wessel, 1979]

Another example of stream formation based on spectral form is illustrated in Figure 6.4 which is taken from a paper by David Wessel (1979). In the first part of Taped Example 3 you may hear the ascending three-note sequence played by one instrument. Then, in the second part the tones marked X are played by one instrument and those marked O by another. After a couple of cycles of the three-note figure (and as the sequence is sped up) the percept splits into two overlapping sets of *descending* triplets being played by the two separate instruments with very different spectral characteristics. Note here that not only the rhythm, but also the direction of the

triplet, changes when the sequence is parsed into multiple images.

#### 6.1.4 *Spectral Continuity and Sequential Organization*

At first consideration, there would appear to be three basic factors used to organize monodic (single-voiced) sequences of sound. One might try to explain the parsing as being based on the differences in *perceptual qualities* of the separate organizations. For example, a series of sine tones that form two separate auditory streams may be said to be parsed on the basis of pitch differences. A series of complex tones that form separate streams when they have different spectral forms but do not stream when they are sinusoidal may be said to be parsed on the basis of timbre differences. A sequence of tones of equal pitch and timbre which differ in amplitude and form separate streams may be said to be parsed on the basis of loudness differences. However, Bregman has proposed (cf. Bregman & Pinker, 1978) that the perceptual qualities themselves are derived from the stream organizations, or source image groupings. That is, the auditory system first groups the complex acoustic array into source sub-groups, and then the qualities of these sub-groups are derived from their respective properties. We then hear a continuity or proximity of those qualities within a given stream.

I have proposed (McAdams, 1981, 1983b; McAdams & Wessel, 1981) that sequential organization is based on a context-dependent criterion of spectral continuity. All of the three acoustic factor criteria proposed earlier in this section may be reduced to this one criterion. Particularly for experiments done with sine tones or complex tones with constant amplitude relations among the partials, spectral continuity and pitch-height continuity are perfectly correlated (Wessel, 1983). But van Noorden (1975) and Bregman (1982) have shown that when one constructs stimuli with a sequence of alternating tones where the pitch sensations are identical but the spectral compositions are very different, they form separate perceptual streams due to the discontinuity of the spectral change, or of the place of stimulation in the auditory periphery. For experiments with complex tones whose spectral structure changes from tone to tone, the spectral discontinuity and timbral discontinuity are well-correlated. In Taped Example 4, composed by David Wessel (1979), you may hear the effects of continuity of spectral form on the organization of a sequence of tones. This sequence has a different instrument playing each note. In the first case, the instruments are chosen to maximize the spectral discontinuity and, not surprisingly, it

sounds discontinuous, like a series of melodic fragments strung haphazardly together. In the second part, the instruments are chosen to maximize spectral continuity while still changing instruments from note to note. (The pitches and apparent spatial location of the notes were also varied to make the example more musically interesting.)

Any musical passage that is changing in pitch, timbre and loudness on a note to note basis is creating spectral discontinuities all the time. And yet we rarely have trouble following melodies or other kinds of musical figures. We cannot rule out the influence of higher level musical constructs such as rhythmic and harmonic function on our organization of sequential material. Certain rhythmic figures can be especially strong "groupers" of events with diverse spectral compositions as anyone who has heard the marvelous complexity, and yet perceptual unity, of a brazilian *batucada* will testify. To my knowledge, there have been no systematic investigations of the effect of strength of metric field or rhythmic pattern on sequential organization.

Of musical interest is the suggestion that the principal factor for sequential organization can result in several different perceptual qualities which can then set up interesting paradoxes in musical streams. The ear follows spectral continuity and not necessarily a given sound source that is being composed with (though most musical sources tend to be relatively continuous spectrally as used in common practice). One might compose for example in a polyphonic setting certain patterns that jump around in pitch for individual instruments. But these may be reorganizable by the ear into several meaningful melodic patterns, each being a different *klangfarbenmelodie* (hear Taped Example 5, arranged at IRCAM by Marco Stroppa).<sup>2</sup> The important fact is that while the principal of spectral continuity is simple, spectral organization in music implies a vast complexity of musical possibilities, particularly as concerns the possible functional roles of timbre in musical composition.

---

2. A score for this Taped Example is included with the description in Appendix G.

## 6.2 Simultaneous Organization

Spectral continuity of event sequences is not the only source image organizing principal. This dissertation investigated how it is that complex tones such as those produced by musical instruments or voice are heard as single sound images and not as compounds of many sinusoids. Also, how are we able to separate complexes from one another that are sounding at the same time?

The present work has been concerned with determining the processes that contribute to the formation and distinction of concurrent source images, and of particular musical interest, the relation between these processes and the derivation of the perceptual qualities of sound sources. As discussed in Chapter 1, at least six classes of acoustic cues may be shown to contribute to auditory image formation:

1. localization in space of a spectral sub-group,
2. harmonicity of a spectral sub-group,
3. pitch separation of two or more spectral sub-groups,
4. coherence of frequency modulation across a sub-group of spectral components belonging to the same source,
5. coherence of amplitude modulation across a spectral sub-group, and
6. stable resonance structure forming the amplitudes of a spectral sub-group.

Items 2 - 6 have been shown to contribute to *spectral fusion*, i.e. the perceptual fusion of spectral components into a unified percept or source image. Let us re-examine 2, 4 and 6 in light of the present experimental results.

### 6.2.1 Frequency Modulation Coherence and Harmonicity

In all natural, sustaining vibration sources, any perturbation, periodic or otherwise, of the fundamental frequency is imparted proportionally to all of the harmonics. There are, of course, minor departures due to various non-linearities in such acoustic systems, but in general as the fundamental frequency changes, all of the harmonics change with it maintaining their harmonic relations. Thus, what I call coherent

frequency modulation is modulation maintaining the frequency ratios of the partials. With computer synthesis, this can be applied to sustained inharmonic tones as well as to harmonic tones.

It is difficult to do an experiment to show that frequency modulation actually fuses a tone complex. But it is certain that for music synthesis it adds a liveliness and naturalness to otherwise dead and electronic sounding images as discussed in Chapter 1. This effect can be heard in Taped Example 6 where a vowel-like sound is unmodulated, then modulated, then unmodulated and then modulated again with a spectral change to a different vowel. Note that the natural voice quality goes away, and one can actually hear out individual harmonics, when the modulation is not present.

It can be shown that if a frequency modulation (vibrato or jitter) is imposed on the partials of a harmonic tone complex such that the ratios are not maintained, the complex "defuses". In Chapter 2, the question posed was whether coherence could be achieved simply by moving all the harmonics in the same direction at the same time or whether it was really necessary to maintain the frequency ratios. Subjects were asked to compare among harmonic complex tones with different modulation schemes. One tone had a modulation that maintained constant frequency ratios among the 16 harmonics. The other tone had a modulation that maintained constant frequency differences among the harmonics. Note that on a logarithmic scale, constant ratios maintain a constant distance, whereas constant differences do not (see Fig. 2.1, Chap. 2). We know that the basilar membrane resolves frequency components in the inner ear roughly on a log frequency continuum. Thus a constant ratio modulation would maintain the relative distances between the places of maximum stimulation on the membrane due to the various harmonics.

When the rms modulation width was at least 12 cents, listeners more often chose the constant difference tone as having more sources, or images, or distinguishable entities in it. In these tones, one experiences a modulating fundamental frequency with the rest of the tone being relatively stationary, particularly at larger modulation widths. It should be noted also that the frequencies of the components making up these tones move in and out of a harmonic relation. For the constant ratio tones, the percept is very unified, even at rather large modulation widths. In Taped Example 7 you may hear one series of each type of modulation (constant ratios, constant

differences). The series starts with no modulation and then progressively increases the modulation width up to 56 cents (3.3%, a frequency excursion of about a quarter tone on either side of the center frequency). These experiments demonstrated that the maintenance of constant ratio is an important part of the definition of *coherence* for frequency modulation.

In Chapter 3, 15 of the harmonics of a 16-component tone were modulated coherently and one harmonic was modulated incoherently. In these tones, a jitter modulation was used. The statistical characteristics of the modulation on the 15 coherent harmonics and that on the incoherent harmonic were very similar, but the random waveforms were independent. Several perceptual effects resulted depending on which harmonic was modulated and on what the overall modulation width was. Either certain partials stand out as separately audible (lower partials), or a kind of "choral effect" results where an illusion of multiple sources is heard (higher partials). A similar effect may be heard in Taped Example 8. A large vibrato modulation is added to a single harmonic and then removed while the rest of the harmonics are unmodulated. This is done for each harmonic in turn from the lowest up to the 16<sup>th</sup>. Note that even the pitch of the 16<sup>th</sup> harmonic can be heard when it is modulated independently of the rest of the tone complex, if the modulation width is great enough.

The choral effect in the higher partials is not surprising if we stop to imagine the behavior of several instruments playing sustained tones simultaneously: five violins, for example. Each acoustic source has its own independent jitter modulating all of its harmonics. When we add all of the sources together we get these random movements of the frequencies beating against one another creating quite a complex situation acoustically. In addition, as one moves into the higher harmonics, the patterns of stimulation on the basilar membrane move closer and closer together until they are very heavily overlapping. In these regions, the incoherent movement of adjacent harmonics is creating a very complex stimulation for any given auditory nerve fiber. I am sure you can imagine that if enough of these violins are playing the same pitch, there is a limit to how many sources you can pick out. The difference between 15 and 16 violins is very small indeed and after about 8 to 10, we generally just hear "many".

It was suggested previously that frequency modulation serves not only to group simultaneous components into a source image, but serves as a cue to distinguish concurrent sources as well. The presence of independent modulation patterns on separate sub-groups gives two types of cues for the presence of multiple sources:

1. adjacent partials belonging to separate sources are incoherently modulating with respect to one another, and
2. the modulation across the partials belonging to a single source is coherent.

It seems likely that the auditory system makes use both of the local incoherence between partials to detect the presence of multiple sources and the global coherence among partials to accumulate the appropriate spectral components into a source image. This may be one reason soloists, particularly opera singers, use vibrato to the extent they do, i.e. to separate themselves from the rest of the ensemble. Of course, as mentioned before, if things are too crowded temporally and spectrally the system may have trouble distinguishing individual source images and tracking their behavior. This would be due to the limitations of spectral and temporal resolution in the auditory system.

Another cue related to frequency that interacts to a certain extent with modulation coherence is the harmonicity of the frequency components. This is particularly evident with sustained sounds. A harmonic series, in most contexts, gives an unambiguous pitch sensation. Sustained inharmonic sounds tend to elicit a perception of multiple pitches. In many cases, the perception of multiple pitches can be interpreted as the presence of multiple sources (Brokx & Nooteboom, 1982; Cutting, 1976; Darwin, 1981; Scheffers, 1983). This, however, has been shown by Cohen (1980) to be dependent to some extent on the form of the amplitude envelope. Inharmonic (and by implication, multi-pitched) sounds are most often heard as fused, single sources when they have exponentially decaying amplitude envelopes as one finds with many kinds of struck sound sources, e.g. strings, bars, tubes, plates, etc.

In a laboratory situation, unmodulated, sustained harmonic sounds can be perceptually analyzed into their harmonics for harmonic numbers up to the fifth, sixth or seventh depending on the fundamental frequency. This means listeners can reliably hear out individual harmonics and identify their pitches. But a pilot study I have

recently performed suggests that when these tones are modulated coherently, listeners are no longer able to hear out the harmonics, indicating that the image is fused and unanalyzable perceptually. In this case the only pitch heard is the virtual pitch of the fundamental. This may be interpreted as support for the notion that the grouping processes (including *spectral fusion*) influence our perception of the qualities of source images (including pitch). There remains, however, the possibility that pitch detection also influences source image formation under certain conditions. It is still unclear at this point whether it is the presence of a number of pitches that indicates the number of sources, or whether the presence of *multiple harmonic series* indicates multiple sources *and* gives rise to multiple pitches. I am more inclined toward the latter interpretation given the preliminary result of reduction of perceptual analyzability of harmonic complexes in the presence of frequency modulation.

### 6.2.2 *Spectral Form*

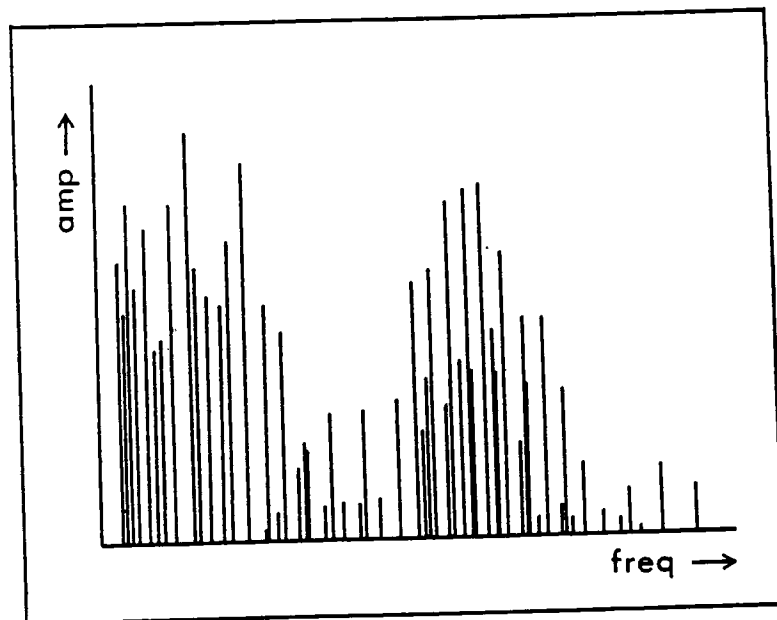
Most sustaining musical sound sources have resonance structures that are relatively stable, or very slowly changing, compared to the frequency fluctuations mentioned in the previous section. These structures are due to resonant cavities that filter the sound before it radiates into the air, e.g. vocal cavities, body resonance for string instruments and tube resonance for winds. Each resonance has a particular frequency to which it responds the greatest. This frequency is related to the volume of the cavity and the size of the opening. Other frequencies are attenuated (made less intense), or are not passed as easily, relative to this resonant frequency. Also, different shapes and the nature of the walls of the resonant cavities influence which frequencies near the resonant frequency are allowed to pass. When it allows a larger number of frequencies around the resonant frequency to pass we say it has a larger bandwidth.

These resonance regions are called formants in voice sciences. And the placement of the formant (center) frequencies, their relative amplitudes and their bandwidths are thought to determine which vowel is perceived. This is particularly true for the arrangement of the first three formants.

Now let us imagine what happens when a singer sings with vibrato. All of the frequencies are moving back and forth in a coherent manner. And what happens to their relative amplitudes? Well, since the resonances come after the point in the system



where the vibrato is introduced, the amplitudes must follow the contour of the formant structure. This was illustrated schematically in Figure 1.3 (Chap. 1). In a sense, we can consider that as the frequencies modulate, their amplitudes change such that each partial *traces* a small portion of the *spectral envelope*, i.e. the frequency-amplitude curve describing the overall spectral form. This complex coupling of frequency and amplitude modulation serves to define the spectral contour and in certain cases may actually reduce the ambiguity of the resonant identity of the sound source. This has been verified experimentally, particularly for higher fundamentals where definition of spectral form is lacking (Chap. 5).



**Figure 6.5.** Complex spectrum resulting from several, unmodulated sustaining harmonic sources. [from McAdams (1984)]

Experiment 7 (Chap. 4) showed that if the spectral envelope moves with the frequency modulation, i.e. the amplitudes of the components remain constant, a kind of timbral modulation occurs. With vowel envelopes one hears a whistling sound which seems associated with the perceptual decomposition of the higher formants. This occurs for modulation widths in excess of  $1/8$  to  $1/4$  tone (25 to 50 cents). This result has some interesting musical possibilities for sound synthesis methods based on formant structures. This will be discussed later.

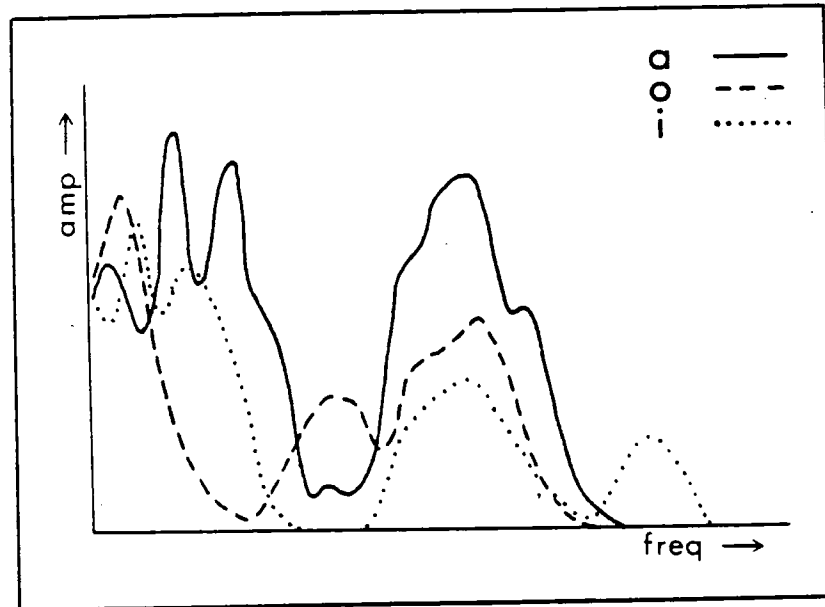
But now let us imagine the following perception problem. The ear receives a complex spectrum as shown in Figure 6.5. There is no modulation on any of the components. Listeners sometimes report hearing certain vowels embedded in this complex and report as many as 6 - 8 different pitches. You may hear six such configurations in Taped Example 9. Without some cue to help us group the elements of this complex, it sounds like a tone mass. If, however, we add some frequency modulation coupled to the resonance structures, the elements are grouped perceptually. We hear the images more clearly as being a certain vowel at a certain pitch. What the auditory system may then have access to spectrally is represented in Figure 6.6. It now knows there are three sources each with a different vowel quality and pitch. The stimuli actually presented are the same ones used in Chapter 5 (notated in section 5.2.2.1). In Taped Example 10, you may hear the individual vowels at the 3 pitches, first without modulation and then with modulation, to verify that they are identifiable in isolation. Then you may hear these same configurations of three vowels at 3 pitches as in Taped Example 9, but with vibrato on one single vowel at a time, in Taped Example 11.

It is important to remark here that without the grouping information to select a certain subset of spectral components, one does not have access to the particular spectral form which gives the vowel quality. The overall spectral form is heard, which does not really correspond to any vowel.<sup>3</sup> But when the modulation is added and the partials trace the individual spectral envelopes, both the coherent harmonic behavior and the reduced ambiguity of spectral form can be used to hear out the vowels.

In Experiment 8, listeners judged the vowels to be more prominent and the pitches less ambiguous when the "sources" were modulating. But there was also a surprising result. In some conditions, all three vowels at their respective pitches were modulated coherently, maintaining the exact ratios between all harmonics of all vowels. Given the putative criterion of frequency modulation coherence for grouping,

- 
3. For many listeners, the vowel /a/ was always more prominent than the vowels /o/ and /i/, even when there was no modulation. This was most likely due to the fact that the individual vowels were equalized for loudness before mixing into the complexes. The formants for the vowel /a/ are grouped into two spectral regions and thus their energy is more concentrated and the harmonics in these regions stand above those of the other two vowels whose spectra are more spread out. This is easily seen in Figure 6.6. Listeners, thus, have access to several formant features for the vowel /a/ even when it is not being modulated.

I expected it to be more difficult to hear out the vowels in this situation. However, there was no difference in prominence between stimuli with vowels that were modulated either coherently or incoherently with respect to one another. But remember



**Figure 6.6.** Spectral forms extracted by the auditory system when the individual sources are modulated, thereby increasing the information pertaining to the number and nature of the sources. [from McAdams (1984)]

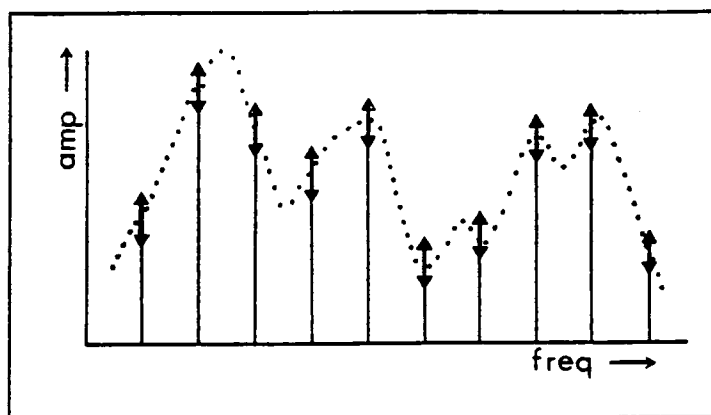
that even when their frequencies are moving coherently, the amplitudes of the partials of each vowel are tracing the spectral envelope of that vowel *alone*. Thus each vowel is still being unambiguously defined by the amplitude movement. Listeners' results indicated that there was no effect of the coherence of modulation of the 3 vowels. As long as they were being modulated at all, they were more prominent perceptually. In Taped Example 12, you may hear the same 6 configurations with all vowels being modulated. This suggests the possibility noted in Chapter 1 that perception of vowel identity is independent of a source forming process, that vowel identification (or speech sound identification in general) is performed in parallel with source image processing.

I would now like to present a musical example where the same kind of effect takes place. Namely, the behavior of the overall spectral form is extracted as meaningful speech information, while the actual spectral contents are perceived as being several sources. Taped Example 13 is a fragment composed by Alain Louvier for the dance theater piece, "Casta Diva", by Maurice Béjart and is taken from a recent record of extracts from compositions and research done at IRCAM (1983). The example was realized by Andy Moorer using linear predictive coding techniques for analysis and resynthesis of voice. In the analysis phase, the voice is modeled as a source of acoustic excitation (a periodic sound produced by the vocal cords and the noise produced by breath, etc.) and a series of filters (the vocal cavities) which change in time. These can be resynthesized exactly as analyzed in which case one recovers a sound very much like the original. Or one can perform a resynthesis (called "cross-synthesis") where the normal vocal cord excitation stimulating the vocal tract is replaced by a more complex, computer-synthesized waveform. Both kinds of resynthesis can be heard in the Taped Example.

What is fascinating musically and psychologically in this kind of example is the demonstration of the multipotentiality of the imaging process. One can perceptually *synthesize* the behavior of the overall spectral form and hear intelligible speech. At the same time one can *analyze* the spectral contents into multiple source images.

### 6.2.3 *Coherent Behavior of Sound Objects and Simultaneous Organization*

All of the factors discussed above contribute to a general coherence of the elements belonging to a physical sound source. And this coherence may be considered in turn as a by-product of the behavior of the physical system producing the sound. It has been proposed that much of the organization that our perceptual systems perform is based on (but by no means limited to) a learning of the normal behavior of physical objects in the world around us. One would not want to limit the possibilities of perceiving to such object-based learning, but there is suggestive evidence that this whole realm of normal perceiving heavily influences our perception of music. I have found, in my own perceptual analyses of several pieces of music, that extensions of these various criteria of "behavioral coherence" have proven useful in predicting when different kinds of reorganization of the physical objects are possible, e.g the recombination of several instruments into a fused composite timbre, and so on.

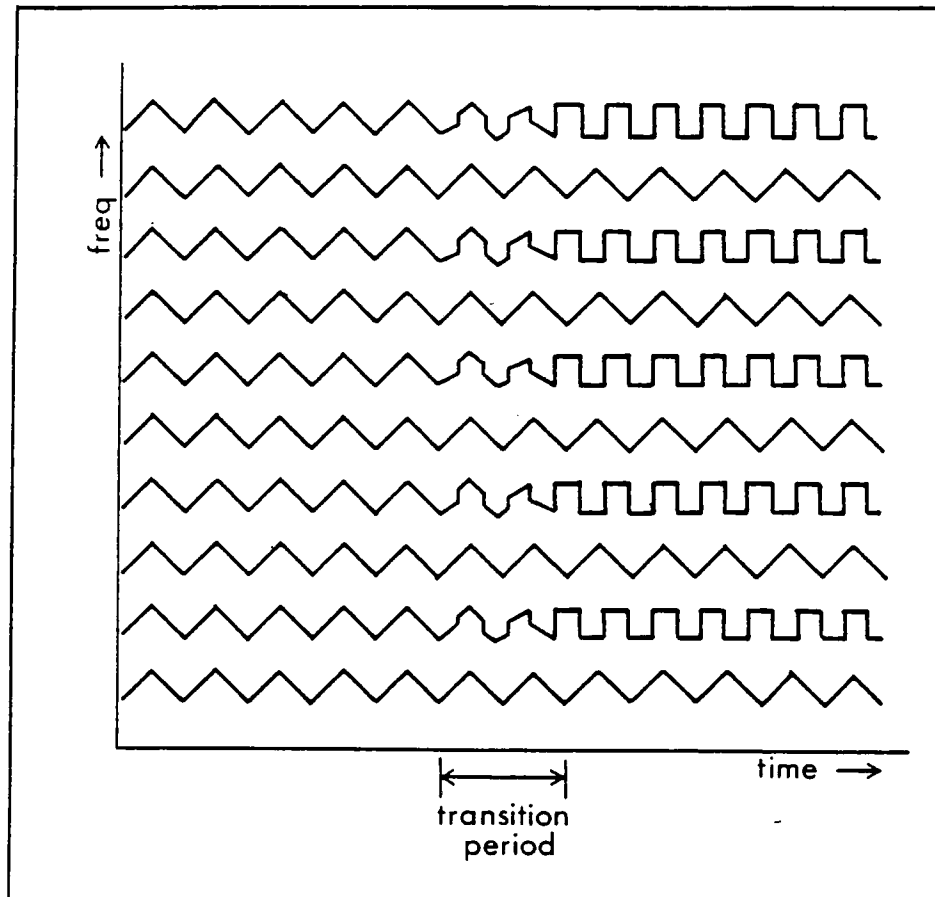


**Figure 6.7.** The amplitudes of each harmonic are modulated randomly above and below their central value, that is defined by the vowel spectral envelope. The modulation pattern on each harmonic is independent of that on any other harmonic. In Taped Example 14 the modulation depth is varied. As the modulation depth is increased, the central spectral form is deformed to a greater degree. Note also that in the Taped Example, the central spectral form is actually evolving as well and that this evolution is not depicted here. [from McAdams (1984)]

With computer music synthesis one can independently control the degree of coherence for any of the grouping factors and even play them against one another. In Taped Example 14 a slowly evolving, but stable vocal spectral form is pitted against incoherent random amplitude modulation on each of the harmonics of the complex tone. This modulation occurs around the main spectral form as illustrated in Figure 6.7. Three different versions are played, each with a progressively greater modulation depth. Note that the effect moves from a single voice to a kind of chorus effect and then to a crowd-like image. Since the average spectral form is the same for each condition, the vowel sounds are maintained. But the incoherence of amplitude behavior gives the impression of many sources and so produces an image of a crowd trying to say the same vowels.

For Taped Example 15, the sound of an oboe was analyzed by phase vocoder. You will first hear the original oboe sound on the tape. The output of this analysis describes the amplitude and frequency behavior of each harmonic. From this data the sound can be resynthesized either exactly as analyzed or with certain modifications. In this case, the even and odd harmonics are sent to separate

channels in order to be played over different loudspeakers. Initially the same vibrato and jitter patterns are imposed on the two groups of sounds. Then, slowly, the frequency modulation pattern on the even harmonics is decorrelated from the pattern on the odd harmonics. This is illustrated in Figure 6.8.



**Figure 6.8.** The parsing of even and odd harmonics of an oboe sound. Initially, all harmonics are modulated coherently. Then the even harmonics are slowly decorrelated from the odd harmonics until they have an independent modulation pattern. The triangle and square waves are used only for easy visualization. Vibratos of different rates and independent jitter functions were used in Taped Example 15. [from McAdams (1984)]

As you may hear, the initial image of an oboe between the speakers gradually pulls apart into two images in the two speakers: one of a soprano-like sound an octave higher (the even harmonics) and one of a hollow, almost clarinet-like sound at the original pitch (the odd harmonics). Following this, each channel is played separately and then the two channel version is played again. It is extremely important that the levels of the two channels be properly adjusted for the effect to work. This example was used in a composition by Roger Reynolds ("Archipelago", 1983) and was realized at IRCAM with the assistance of Thierry Lancino. Here we have a case where the coherence of frequency modulation at the beginning of the sound overrides the spatial separation of the two subsets of harmonics and one hears a single image of the oboe, more or less localized between the speakers. But as the modulations become incoherent, the images move to their rightful places and the sounds now appear to come from where they were originally coming from.

As discussed in the section on spectral form, the auditory system is very sensitive to the behavior of the overall spectral structure. With Xavier Rodet's (1980a,b) synthesis program CHANT (from the French word for "sing") one has flexible and independent control over the behavior of each formant. Jean-Baptiste Barrière used this capability in a series of studies for his piece "Chréode" (1983), realized at IRCAM. He manipulated the way the individual formants changed in time to make the spectral forms either coalesce into vowels or disintegrate into the several formants as individual images. In Taped Example 16 you will hear some voices modeled after Tibetan chant that are slowly disintegrated by decorrelating the formant movements until, at the end of the fragment, individual formants can be heard whistling around across the harmonics.

All of these examples are intended to show that the constellation of factors contributing independently to the organization of simultaneous elements into auditory source images is quite complex. They also demonstrate that the factors can interact and that some factors can override the effects of others, as was demonstrated in the split oboe example.

The important similarities among these factors are that they are dynamic, i.e. they change with time, and the coherence of change indicates a common source origin while incoherence of change indicates diverse source origins. I consider the spatial location factor to be dynamic because in normal listening both the head and the

sound source are moving to some extent and the dynamic coherence of the result, with respect to the two ears, becomes a relatively unambiguous cue for place of origin of the sound source. Indeed, "place" becomes an invariant quality of the sound image when this coherence is maintained.

What I think we need as a general and subdividable principle is the notion of the *coherence of behavior* of the elements belonging to a source. Again, as an explanatory metaphor, this notion can be applied at several levels of description and defined with respect to the factor being considered, as I have demonstrated in the previous sections. It is also important to consider that the "meaning" of coherence can be dependent on the previous experience of a listener. We can learn the behavior of a certain sound source. And we can incorporate the instances and relations between instances of its sound emanations into a model of coherence for that object and for physically similar objects.

I find myself returning often to consider the behavior of physical objects for reasons of both "ecological validity" (perceptual systems are "meant" to operate in the physical world) and personal experience with electroacoustic music. In listening to several hundred hours of electronic and computer music I have often been struck by a particular, natural (almost default) mode of listening which remarks that electronic sounds most often *sound like something*. Which is to say that my perception, being always influenced by memory and learned patterns of categorizing and identifying, tries to hear with respect to the *already heard*. I will return to this point shortly, but let me pass on to consider the interactions of sequential and simultaneous organization.

### 6.3 Interactions Between Sequential and Simultaneous Organization

It is no news to musicians that there is some essential distinction to be made between these two types of organization, which are traditionally denoted as *horizontal* and *vertical* in reference to the page of the musical score. In music we see different kinds of compositional principles in operation for writing that tends more toward homophony and that which tends more toward polyphony. But of course, the most interesting music arises where these come into counterplay – either converging on similar propositions of perceptual organization, or proposing separate, conflicting organizations. It is the creation of tension and functional ambiguity that, among



many other things, brings an exhilaration to me as a listener.

In a sense, there are two separate propositions before our ears in this kind of counterplay. It appears that sequential and simultaneous organization are determined by separate criteria: sequential elements are organized according to spectral continuity, simultaneous elements are organized according to coherence of dynamic cues. That they are organized by separate criteria suggests the possibility that they may conflict, even compete, with one another. This was tested experimentally by Bregman & Pinker (1978), and described in Chapter 1. The two extreme cases noted in Figure 1.1 may be heard in Taped Example 17.

This was an important experiment in two respects. It demonstrated the separate kinds of organization and their interaction in the final perceptual result. And it also demonstrated that a perceived quality of a source, e.g. the timbre of *C*, was dependent on how the elements were organized. When tone *B* was not grouped (fused) with *C*, the latter was more pure. But when they were fused, *C* was perceived as being rich. Thus timbre, and, as I demonstrated with the oboe in Taped Example 15, pitch, are properties of source images that are derived *after* the concurrent elements have been organized into those images. Demonstrations such as this support the proposition of Bregman (1977, 1980) that perception is an active process of composing the sensory data into some kind of interpretation of the way the world is behaving. The composition process draws from a large number of elements in the perceptual field that interact in complex ways to produce a final percept (Bregman & Tougas, 1979).

Where this becomes interesting musically is in its implication that, with computer synthesis techniques, processes of horizontal and vertical musical organization can be carried into the sound microstructure. The composer can play with the processes of perceptual organization that underlie the heard musical surface. This sets up the possibility of composing situations where sequential and simultaneous organizations compete for individual spectral components to be part of the structure of a musical image. With a careful consideration (or better yet, embodiment and subsequent intuitive use) of these principles of perceptual organization, the composer has access to a whole realm of mutability of the heard "image". Convergences and divergences of the musical functionalities of individual elements allow the development of microstructural (and pre-perceptual) ambiguity. These possibilities are evidenced in the last Taped Example (no. 18), created by Xavier Rodet with the CHANT computer-synthesis

program.

#### 6.4 Toward a Theory of Auditory Image Formation and Source Perception

A theory of auditory source perception would be a subset of a more general theory of auditory image formation. It has been stated several times throughout this dissertation that higher-level groupings of several sources into coherent meaningful units are also considered as images. And that a multiplex perception is possible with respect to these many-tiered perceptual structures, one level of which may include a perception of the actual physical sources of sound (refer to the Prologue). The necessary elements for a theory of auditory image formation would include:

1. a part concerned with *a description of coherent sound source behavior*; what is the nature of the coherence of the sound environment that is reflected in our accurate and rich perceptual organization of it?
2. a part concerned with *the processes of organization* which would in turn include:
  - a. acquisition and representation of knowledge about the coherent behavior of the world (in terms of schemata and conceptual structures), and
  - b. processes of grouping and separation of image elements based on criteria of image coherence; this would necessarily reflect the predominant accuracy and consensuality of perceptual organization as well as its flexibility and differentiation among different perceivers; involved here also would be processes of dynamic pattern recognition, the patterns again reflecting, but not necessarily being bound to the predominant pattern of coherence in the environment.
3. a part concerned with *the derivation or emergence of qualities of auditory images*; this would require some clarification of the relation between grouping processes and perceived qualities of groups; this is particularly important for an understanding of perceived qualities of complex perceptual structures where different properties derive from (or are associated with) different levels

of structuring.

4. a part concerned with *the processes of attention and selection*; in particular, there is the perceived problem of what controls attention as well as what attention operates on with respect to its influence on perceptual organization; this is of paramount importance given that what is grouped depends in many instances both on what the perceiver is inclined to perceive as a group and on what the perceiver is trying to "pay attention to".

#### 6.4.1 *Behavioral Coherence of Sound Objects*

The first problem here is to define what is meant by "behavioral coherence", a term used metaphorically up to this point. This is certainly a part of the problem for which we can take Gibson as an example of how to proceed. It has been suggested by some (Heyser, 1973a,b, 1974, 1976a,b; Balzano, 1983) that the classical pair of "spaces" used to represent sound mathematically are insufficient. This pair comprises, of course, the "time" and "frequency" domain representations related by the Fourier transform, with which concepts much struggle has taken place in the preceding pages concerning which mechanisms are using "temporal" information and which are using "spectral" information. It is Heyser's contention (and one that resonates remarkably with that of Gibson) that a theory of the stimulus is a proper part of a theory of perception. Accordingly, one should search for an alternate system (of which there are an infinite number) which more closely represents the properties one observes in perception. In the case of perception of physical sources this may imply something more like the wave equation representation (physical process modeling) proposed by Hiller & Ruiz (1971, cited in Balzano, 1983). Following this line of thinking, Balzano proposed that many of the perceived invariances of source objects with changes in pitch, intensity, acoustic environment etc. and many of the tight couplings among "cues" for source image formation that have been proposed in previous sections might reduce to simple manipulations of parameters in this different representation. Indeed if such representations can be found which keep the richness of perceptual change as well, and if we can demonstrate that perceptual processes reflect this representation, then many apparent paradoxes may be resolved. Nonetheless the enormous (if tedious) flexibility possible with sound synthesis based on imaging cues that have been shown to be reasonable predictors of perceptual results above must also be representable by such a new set of "sound

coordinates". At present, the Fourier representation serves very well and gives us a control over the decomposition and recomposition of auditory images that mere physical models cannot attain, such as the oboe split example. Thus, I would caution that by advocating a closer consideration of the behavioral coherence of sound objects, I am *not* advocating a complete abandoning of the Fourier representation. I do, however, agree that certain aspects of the representation of coherent sound behavior may be better represented in other coordinate systems. The challenge is to find possibilities for manipulating this representation in perceptually meaningful and imaginative ways. And even then, I do not think having a more suitable representation of the stimulus will obviate the necessity for theories of grouping and attention as well. A better theory of the stimulus might simplify but could not replace a theory of perception.

#### 6.4.2 *The Processes of Auditory Organization*

The concept of the schema appears to be a likely candidate upon which to build a theory of auditory organization. For one thing, this notion easily fits the requirement of a means of representing knowledge about the behavior of the world. The proposed nature of schemata as ordered structures satisfies the constraint that the regularities and order of the world be preserved. Also, since schemata are built either of basic units (concrete or abstract concepts) or of other schemata and are constructed as experience and further knowledge are accrued, their current state of development represents one's current, incomplete knowledge of the world. In the face of conflicting evidence, schematic structures can be disassembled and reconstructed to fit a new comprehension, and then further tested against one's continuing experience.

Where this notion becomes more difficult is with respect to grouping processes that pull the elements together for a completed image. For very familiar images, this represents no problem since a schema may be considered to exist already. For sequences of complex and unfamiliar images, some process needs to be proposed that assembles the units. One also needs to develop the nature of its mode of operation as concerns (a) the criteria it uses for grouping and assembling, (b) the context dependence of the relative strengths of the various criteria, and (c) how conflicts between possible solutions can result in either suppression of some solutions or existence of multiple solutions as in the case of various illusions. This *assembler* of schemata (i.e.

of images) is proposed as such by Sowa (1984). It is the assembler that generates the working model from previously stored schemata matching incoming sensory information.

In Sowa's scheme, the assembler (concerned with reconstruction) is independent of a kind of *associative comparator* that matches incoming sensory information to previously stored schemata. Thus the processes of pattern recognition and grouping of elements are independent. This feature is very important for modeling some of the results reported which indicate that spectral form perception appears to be independent of source grouping processes. In normal situations these converge, in experimental conditions (e.g. Cutting, 1976 with dichotic presentation of a single speech source; Chapter 5 with perfect frequency modulation coherence of 3 sources) they may diverge and the perceived qualities result accordingly. Again, these kinds of processes must reflect the heuristic, multi-staged, heterarchical and multiplex nature of image perception.

#### 6.4.3 *Derivation of Image Qualities*

Throughout this dissertation, it has been shown that there is a rather complex relation between image grouping and image quality perception. Many of the properties claimed to be used in grouping decisions (such as harmonicity and spectral form) also play a strong role in the generation of perceptual qualities. For these it seems likely at this point that the contributions to quality derivation and grouping are concurrent. Harmonicity gives an unequivocal pitch and unified image under simple conditions. Spectral form can trigger recognition of a familiar quality as well as indicate the presence of a stable resonance structure in the environment. Other more dynamic cues which contribute most strongly to a dynamic coherence, can at times influence groupings and perceived source qualities. Frequency components in a harmonic relation and conforming to a vowel spectral form can be made to form multiple groups on the basis of frequency and amplitude modulation incoherence. However, common onset time, and perfect vibrato coherence can fail to fuse 3 separate vowels. All of this points to a possible system of constraints that limit which cues can affect which perceptual processes under various conditions.

#### 6.4.4 *Attentional Processes*

Finally we come to the kind of *central controller* of perceptual activity, that, in ambiguous or not-perfectly-clear situations, can have a large degree of influence on the organization and subsequent interpretation of the environment. Attention can either be controlled by previous expectancies set up by currently active schemata, or drawn passively by the structure of the perceived world (an interaction between the exploring of the environment and the assembly of current schemata) or it can be directed by "conscious will" on the part of the perceiver (again perhaps by the intermediary of active anticipatory schemata).

But given that multi-leveled perceptual structures are possible and are representable by nested schematic structures, what does a model of attention look like and how are specific subsets and supersets selected to occupy consciousness? For instance, in a simple case, one can attend to the facts that 4 instruments are playing (countable number of sources), each at a different pitch (separate image qualities), they are oboe, clarinet, flute and bassoon (separate image qualities and identities) in descending order of pitch (assignment of quality to identity), and they are playing a major sixth chord with the sixth in the lowest pitch position and it sounds like a major sixth chord and not a minor seventh of the relative minor key (they are isomorphic with respect to pitch content) because of the preceding harmonic context (emergent quality of a higher-level group dependent on the contextual information from preceding qualities of the same grouping level). In this case the hearing of a flute playing its low  $C_4$  and the hearing of a major sixth chord of which the flute note is the tonic are not incompatible. They are at different levels of the structure, or at least the two schemata are overlapping and not mutually exclusive. In the Bregman & Pinker (1978) example (Fig. 1.1) one could not hear tone  $B$  both as being in a sequential stream with  $A$  and a fused timbre organization with  $C$ . If it was fused, it was no longer available as a separate element to stream with  $A$ . Here the sequential grouping and simultaneous grouping solutions are mutually exclusive in the given context. In the Cutting (1976) "spectral/temporal" fusion (Fig. 1.2) one can presumably hear a /da/ in one ear and a chirp in the other ear. This is a tough illusion to explain and requires giving special license to speech sound perception in the present scheme. Spectral form perception in the service of speech recognizers is a central process presumably receiving information from both ears (and perhaps indiscriminating about where it gets its information from). It recognizes all the necessary elements to claim the presence of

a /da/. Meanwhile, spatial location criteria for grouping are proposing two stimuli, one in the left ear and one in the right. Grouping decisions separate the two and the  $F_2$  transition is heard as a "chirp". All of the perceptible elements have somehow to be assigned to an image somewhere, perceptual qualities, in general, don't exist unattached to something. The /da/, according to information from the grouping processes, is assigned to the ear from which comes most of the spectrum that resembles what is believed to have been heard. This explanation is not unlikely within the model of perception being considered.

What is implied in the previous paragraph is that there are a set of rules and constraints on the possible logical structures that schemata can assume in relation to the incoming sensory information. These include subset relations, mutual exclusion, overlapping relations (as in the Cutting example), and perhaps even logical interactions of a more complex nature. Any development of a theory in the direction proposed above must take such complexities into account.

## EPILOGUE

I have proposed that there are separate groups of criteria that determine the way one organizes acoustic information sequentially and simultaneously. This distinction, perhaps, reflects the involvement of different types of perceptual mechanisms. Sequential information is organized according to criteria of spectral continuity. A sequence of events that maintains spectral continuity is more easily followed as a source image than a sequence that is discontinuous. In the latter case one is more likely to reorganize the sequence into two or more streams. What the actual limits of spectral continuity are acoustically are probably as much determined by the environmental or musical context as they are by the previous experience of the listener. Simultaneous information is organized according to criteria of coherence of dynamic cues such as amplitude and frequency modulation that indicate common source origin, the tracing of spectral form that indicates a common resonance structure, and the coherent binaural disparities that indicate a common spatial origin. Spectral components that behave coherently in these ways are more likely to be heard as originating from a common source and will thus form a unified, fused auditory image. As this coherence of behavior is maintained across time, the image can be followed in time as well. However, since there are several criteria for sequential and simultaneous organization, it is possible to construct situations where they come into conflict. In situations of conflict, either one criterion overrides another and the organization follows accordingly, or situations of organizational ambiguity result, or multiple percepts result.

Given the extensibility of the auditory image metaphor and the principles of continuity and coherence, it should be possible to develop a psychologically relevant theory of musical attention and organization that covers the range from the formation of the image of a single event to the accumulation of the "image" of a musical form, passing through many intermediate levels of organizational polyvalence (each



element is potentially a member of several concurrent organizations) and the construction of composite musical objects that have a complex evolution through time. I feel that structural and functional ambiguities are a very important part of musical organization and if well understood can be used with great effectiveness and power in musical composition.

So where do we stand with respect to the initial questions of the Prologue? As concerns what might possibly be attended to as a musical image, a good start has been made with an understanding of certain basic principles of sequential and simultaneous organization, at least at the level of source organization. There remains much work to be done on the effects of higher level organizing principles, such as underlying metric and rhythmic structure and underlying harmonic structure, on what can be followed through time as a coherent entity and on what can be grouped as musically meaningful conglomerates or composite images. With respect to the nature of processes and cues involved in the perception of complex situations, the principles outlined here certainly point to an important group of processes that are involved in the act of auditory organization. Again, what needs to be further researched (as much in musical as in psychological paradigms) is the extent to which the active, creative involvement of the listener's attention can play a role in the organizing of complex constellations of sound events into musical images.

Pattern recognition processes may be conceived of as "templates" of classes of sources that are encountered in the environment. This kind of "template" does not represent a static "object" with which it can be directly compared. Rather, it represents a class of rules of dynamic relations among and coordinated transformations of elements of a source characterizing (i.e. schematizing) a family of particular instances of that type of "object". In this sense, what we deal with as a particular object in the environment really implies a process of relationship between the object and its perceiver. The richness of these rules of source formation allows for the perception of invariance. A visual image of a clarinet remains a clarinet image regardless of its orientation in space, i.e. rigid spatial rotation is one kind of transformation rule used in the visual system (cf. Bregman & Mills, 1982; Shepard & Cooper, 1982). And an image of a clarinet tone remains a clarinet image regardless of the register or the intensity being played, i.e. with experience, we learn the kinds of transformations in spectral relations within a clarinet tone that can accompany register changes, loudness changes, etc.

But these rules also imply certain limits to our perception with respect to what we "allow" ourselves to perceive. People have the tendency to confine their perception within the bounds of these learned rules rather than "stretching" their perceptual abilities to meet with novel acoustic organizations such as are found in "new" music or "old" music of other cultures. *Listening to these musics requires an evolution of one's perceptual patterns.*

Another area of richness for music that is implied by these grouping processes concerns the notion of the "boundaries" of auditory objects. These boundaries may be considered in terms of the perceptual/cognitive structures underlying their perception (Miller & Carterette, 1975). The dimensional structures underlying the perception of timbre and pitch as determined by multidimensional scaling techniques have been described by Grey for timbre (1977; Grey & Gordon, 1978) and by Shepard and Krumhansl for pitch (Shepard, 1964, 1982; Krumhansl, 1979; Krumhansl & Shepard, 1979). For example, Grey has described a three-dimensional structure for timbre where each dimension is considered to represent an orthogonal perceptual dimension contributing to the perceived similarity of the timbres of different sounds. Grey used the tones of common Western musical instruments. Shepard claims that "distance" in this geometrical representation of perceptual/cognitive space represents the degree of dissimilarity among sounds such that identical sounds have the same location in the structure and very dissimilar sounds are far removed from one another. An instrument such as the clarinet, would be defined by a region occupied by the sounds of this instrument in a combined pitch/timbre space. This would represent all of the pitch/timbre possibilities of that instrument. No acoustic instrument could encompass the entire, possible space. Each instrument would have a bounded region, though there is certainly a strong probability that different instrument's regions would overlap in certain places. In these overlap regions, the instruments' sounds would be more easily confused with one another. Those instruments with a wider pitch range and more timbral versatility would occupy larger or even multiple regions (if one were to include some of the multiphonic instrumental techniques developed in recent times).

These instruments usually elicit unitary images (when playing normally), so these regions represent a range of possible pitch/timbre percepts for fused sources. Consider now the possibility of having elements of one pitch/timbre "point" embedded in those of another. The resulting percept could be transformed by the context of a

parsing process such that a gradual transition from one point in the space to two other points could be realized. This would correspond to a transformation from the combined image to the two individual images and would represent a sort of non-linearity in the structure. The transformation is not perceptually continuous, but neither is it noticeably discontinuous. At some point after a *pre-conscious transformation* has taken place, one becomes aware of new presences in the music. This would imply the possibility of musically dissolving "instrument" boundaries or "voice" boundaries by means of sound synthesis. With versatile and sensitive performers, similar transitions can be made acoustically as has been attempted by many contemporary composers. In these cases, new spectral information is added with the addition of another instrument, whereas, with sound synthesis, the long-term spectrum can remain constant and yet be dynamically rearranged into new images by the listener.

The possibilities of continual transformation of the spectrum in ways such as this present the opportunity to move away from the limited (bounded) conceptions of pitch and timbre. Changes in the grouping of a complex spectral "field" can result in very rich perceptual changes. Myriad fleeting images can continually be emerging from and submerging into one another.

In music we move away from concepts of simple or complex successions of pitches and/or timbres. We move toward complex evolving organisms given life by the active organizing of the time-varying spectrum by the listener. We may never have the ability to predict the experience of any listener since a lot of what is perceived by an actively attentive listener ultimately depends on what the listener brings to the music as well as what the music brings to the listener. The composer creates a universe within which the listener can create musical worlds and forms. The composer enfolds a universe, the listener unfolds a new music and is unfolded by a new music with each coming-into-being of the sounds of that universe.

Hearing is a dynamic process which engages a listener actively in interacting with the environment, not merely with the analyzing and storing of acoustic information. We hear a world *as if it were a given property of the mind / that certain bounds hold against chaos* (Robert Duncan). We can limit our hearing of the world it is possible to hear, or we can allow ourselves to hear the possible in the world. "What is the realm of *allowable perception*?" is a question beyond the scope of this dissertation, but one well worth contemplating.

As I am prone to reiterate *ad nauseum*, each listener still carries into the musical situation "normal" tendencies of hearing that are going to act as defaults in the organization of musical sound. However, what is most compelling as a result of all of the research on auditory organization is the fact that the will and focus of the listener play an extraordinarily important role in determining the final perceptual results. Musical listening (as well as viewing visual arts or reading poetry) is, and must be considered seriously by any artist as, a creative act on the part of the participant. As mentioned previously, perceiving is an act of composition, and perceiving a work of art can involve conscious and willful acts of composition. What this proposes to the artist is the creation of forms that contain many possibilities of "realization" by a perceiver, to actually compose a multipotential structure that allows the perceiver to compose a new work within that form at each encounter. This proposes a relation to art that demands of perception that it be creative *in essence*.

## APPENDIX A

### Synthesis Procedures and Sound Presentation System

Two basic synthesis procedures were used: an additive (Fourier) synthesis algorithm for individual control of the amplitudes and frequencies of a bank of digital oscillators (Expts. 1-7, 9-10), and a time-domain formant-wave-function synthesis algorithm (CHANT) for realistic voice sounds (Expt. 8). In all cases a sampling rate of 16129 Hz was used. Signals were synthesized on a PDP-10 computer and stored on computer disks. They were then transferred to the disks of a PDP-11/34 computer which was used to run the experiments.

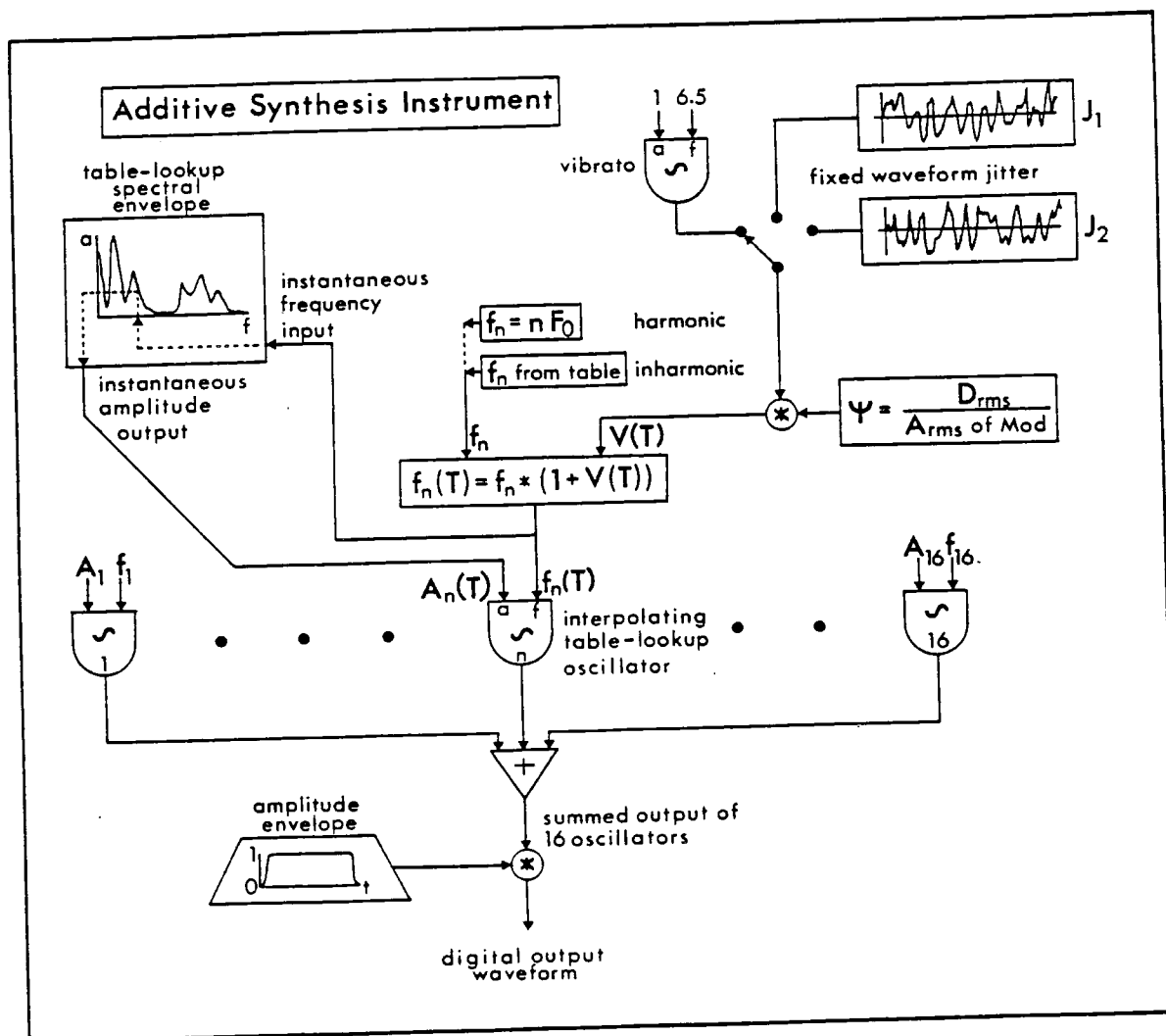
#### A.1 Additive Synthesis Procedures

The various procedures of this type were realized as computer "instruments" in the MUS10 sound synthesis language (Tovar & Smith, 1977). Essentially, each instrument consisted of 16 oscillators with independent time-varying control over the amplitude and frequency of each. This allows one to adjust the coherence of frequency modulation among the components, to choose a desired modulating waveform, and to choose whether or not the amplitude functions are coupled to the frequency fluctuations by a desired spectral envelope. The basic structure of the instrument is schematized in Figure A.1. The converted waveform  $S(t)$  may be described mathematically as

$$S(t) = A_{global}(t) \sum_{n=1}^{16} A_{spectral}(f_{ni}) \sin(2\pi f_n t + q \psi \int_0^t Mod(t') dt'), \quad (A.1)$$

where  $A_{global}(t)$  is the global amplitude envelope,  $A_{spectral}(t)$  is the instantaneous

amplitude of partial  $n$  dependent on its instantaneous frequency  $f_{ni}$ .  $f_n$  is the central frequency of partial  $n$ ,  $q$  is a factor controlling the maintenance of harmonicity among the partials during modulation, and  $\psi$  is a factor related to the width of frequency excursion of the modulation waveform  $Mod(t)$ .



**Figure A.1.** Schematic of sound synthesis procedure for generation of 16-component complex tones.

The basic element in the synthesis scheme is an interpolating table-lookup oscillator which accesses a 512 point table containing one period of a digitized sine function in 36-bit floating-point representation. In synthesizing a frequency,  $f$ , sampled at a

rate of  $SR$  Hz, with table size  $TS$ , a pointer is incremented through the table at

$$I = \frac{f \cdot TS}{SR} \quad \text{steps / sample.} \quad (\text{A.2})$$

If  $I \bmod 512$  is non-integer, a linear interpolation is performed between the flanking table values. The rate of incrementation can be changed at each sample, i.e. frequency can be updated on a sample-to-sample basis. The output of the oscillator can also be scaled in amplitude on a sample-to-sample basis. These instantaneous amplitude and frequency values are shown as  $A_n[T]$  and  $f_n[T]$  for oscillator  $n$  in the schematic.

#### A.1.1 *Amplitude Control*

As seen in Eq. A.1 there are two terms controlling the instantaneous amplitude of a given partial: the global amplitude envelope ( $A_{global}(t)$ ) and the fluctuation of a given partial's amplitude with frequency according to a pre-defined spectral envelope ( $A_{spectral}(f_{ni})$ ).

##### A.1.1.1 *Global Amplitude Function*

The global amplitude envelope is a multiplicative factor read at each sample from a 512 word table by an interpolating table-lookup procedure. In all of the experiments using this instrument, the duration of a tone was 1.5 sec. The entire envelope from 0 to 1.5 sec is represented in this table. In the first 34 and last 34 points, a half-cycle raised cosine is represented. This yields a rise and decay time of 99.6 msec. Between these portions, the table value stays at 1. The global amplitude function can thus be described as

$$A_{global}(T) = \begin{cases} \frac{1}{2} \cos(2\pi \frac{1}{2t_a} T + \pi) + \frac{1}{2} & \text{for } 0 \leq T < t_a \\ 1 & \text{for } t_a \leq T \leq D - t_a \\ \frac{1}{2} \cos(2\pi \frac{1}{2t_a} (T - D + t_a)) + \frac{1}{2} & \text{for } D - t_a < T \leq D \end{cases} \quad (\text{A.3})$$

where  $T$  is the sampling interval,  $t_a$  is the attack/decay time and  $D$  is the duration. This function is applied to the combined output of all 16 frequency components.

#### A.1.1.2 *Component Amplitude Determined by Spectral Form*

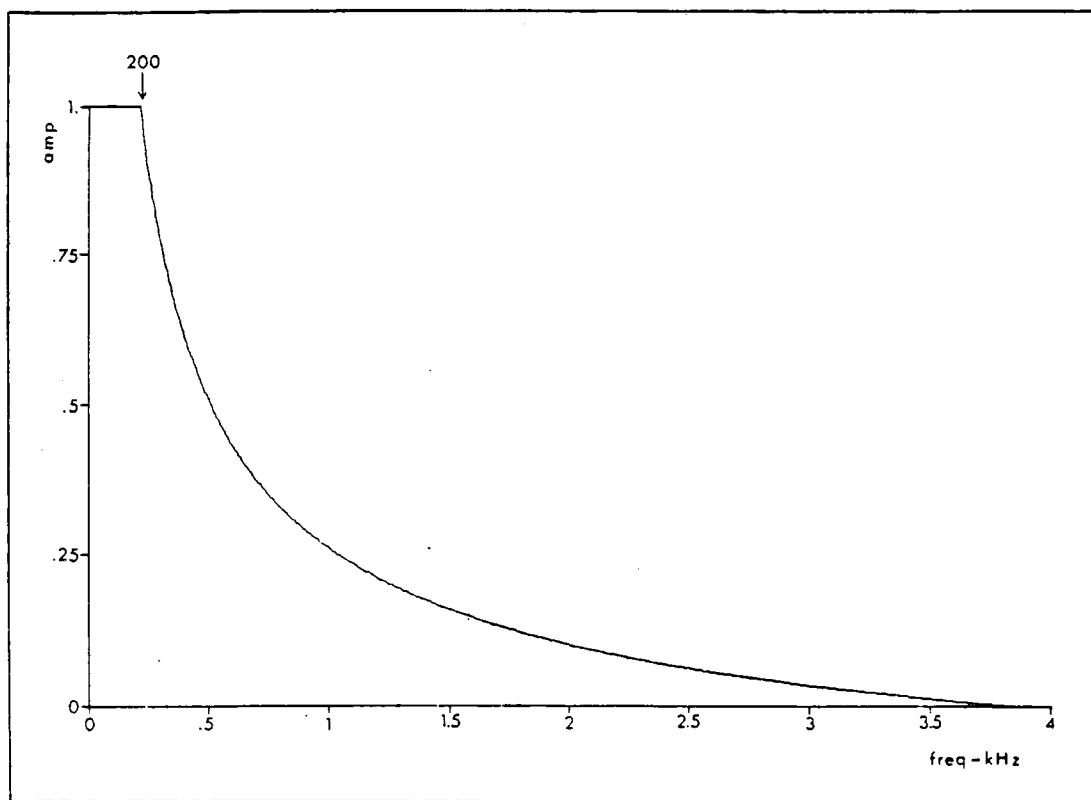
In most of the experiments (Expts. 1-5, 9-10) the amplitude of the components followed a spectral form when the components were frequency modulated. For some of the stimuli in Experiment 7, the initial component amplitudes were chosen according to the spectral form, but then remained constant afterwords. Three different spectral envelopes describing the relative amplitude (values of 0 - 1) as a function of frequency (0 - 4 kHz) were stored in 512 word tables. Thus the value in the 512<sup>th</sup> word is the amplitude at 4 kHz, that in the 256<sup>th</sup> word is the amplitude at 2 kHz, etc. The envelopes were:

1. a flat spectrum: all components had equal amplitudes,
2. a -6 db/oct spectrum: the amplitude of a component one octave above another was 6 dB less in amplitude (see Figure A.2). The flat portion below 200 Hz never affects any stimulus component since the lowest component used was 220 Hz and at the largest modulation width used this descends only to 211 Hz.
3. a spectrum derived from a vowel /a/ sung by a male voice (see Figure A.3) taken from data of Rodet at IRCAM.

A special algorithm accepts an instantaneous frequency value at each sample, converts it to a table address and returns the amplitude value at that point in the table. If the floating point address value falls between possible integer address values, a linear interpolation between the appropriate values is performed. Therefore, as the frequency changes on a sample-to-sample basis, the amplitude of that component continuously conforms to the stored spectral envelope function, thereby tracing the form. In a sense, this is like a filter with spectral conformity characteristics but no resonating effects. This spectral envelope table is accessed by each frequency component at each sample.



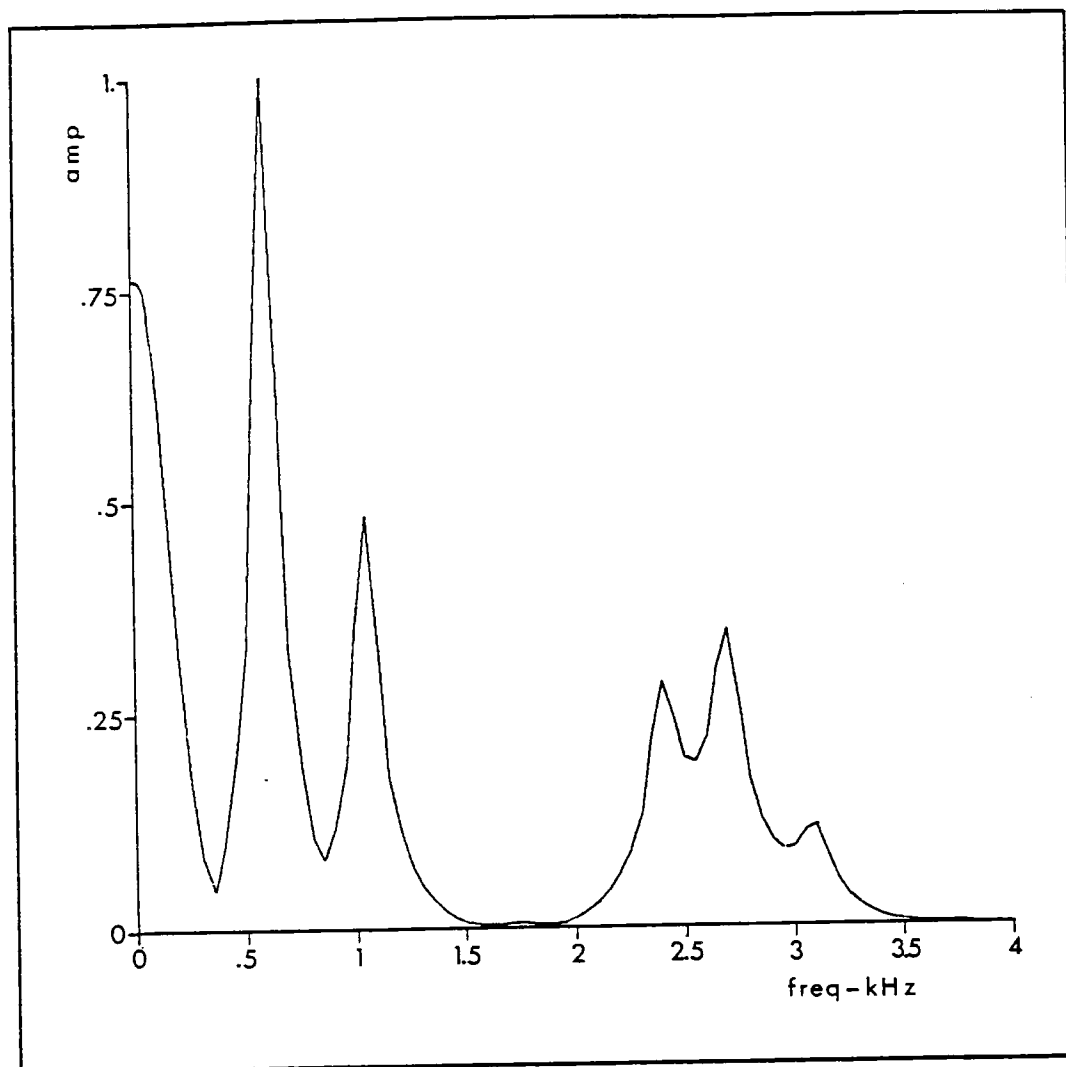
For the constant component amplitude stimuli of Experiment 7, these tables were accessed only before synthesis was begun to determine the relative amplitudes of the components. Then the component amplitudes remained in this relation for the entire duration of the tone irrespective of any frequency movement.



**Figure A.2.** The -6 dB/oct spectral envelope. The flat portion below 200 Hz never affects the amplitude of any stimulus component.

#### A.1.2 Frequency Control

The sixteen center values of  $f_n$  were determined either according to a harmonic series ( $f_n = nF_0$ ; Expts. 1-7, 9-10) or from a fixed table of inharmonic frequencies (the inharmonic stimuli of Expt. 6).



**Figure A.3.** The vowel /a/ spectral envelope from a singing male voice.

In most cases these frequencies were modulated according either to a sinusoid function (vibrato) or one of two fixed low-frequency random functions ( $J_1$  and  $J_2$ ; see Appendix B). The vibrato function was generated with a digital oscillator of the same type as those for the frequency components. Its amplitude was fixed at 1 and its frequency at 6.5 Hz (Expts. 1-5, 7, 9-10). Each of the 3 modulating waveforms had a different rms amplitude. Therefore, to have control over the desired rms deviation in frequency, the factor  $\psi$  was adjusted such that

$$\psi = \frac{D_{rms}}{A_{rms} \text{ of } Mod(t)}, \quad (A.4)$$

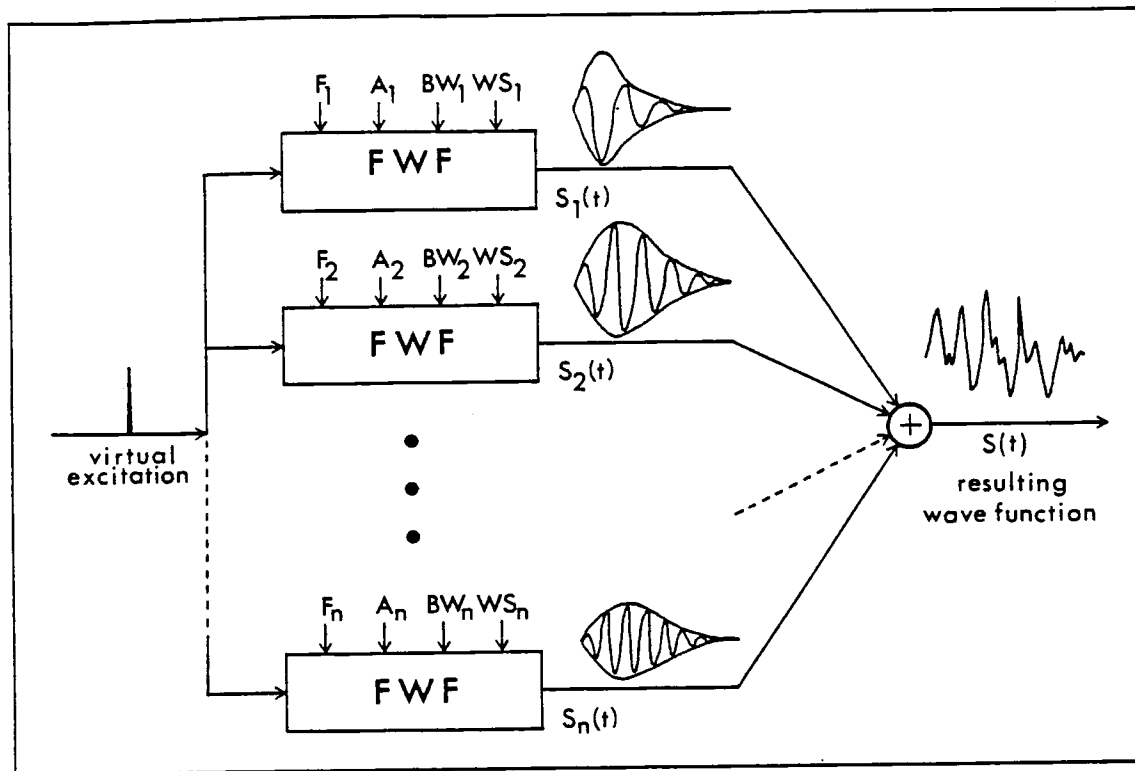
where  $D_{rms}$  is the desired rms deviation and  $A_{rms}$  is the actual rms deviation of the waveform in question. Each modulating waveform was approximately symmetric about 0 (in linear frequency) over the duration of the tone.

The factor  $q$ , if set equal to the ratio of the partial's frequency to that of  $f_1$ , assures that the frequency ratios are maintained in the presence of frequency modulation. For example, for harmonic components,  $q$  is set equal to the partial number ( $n$  in Eqs. 1.3, 2.2, 4.1;  $f_n/f_1$  in Eq. 3.1). This is the case for harmonic constant-ratio (coherent) stimuli. For the inharmonic stimuli, the ratio is not an integer value but is still adjusted for each partial such that the ratios remain constant ( $f_k/f_1$  in Eq. 3.2). For the constant frequency difference modulations of Experiments 1 - 5,  $q$  is set to a constant value ( $k$  in Eq. 2.3). This maintains a constant difference but the ratios are not maintained.

## A.2 Time-domain Formant-wave-function Synthesis Procedure

The music synthesis program CHANT developed by Xavier Rodet (Rodet, 1980a,b; Rodet & Bennett, 1980) replaces the classic representation of the voice (as an excitation source followed by a series of filters) with a unique formula describing the output of the filters - a kind of impulse response of the resonance structure. This response is called the formant-wave-function (FWF). This function is derived from a specification of the frequencies ( $F_i$ ), amplitudes ( $A_i$ ), bandwidths ( $BW_i$ ) and widths of the skirts ( $WS_i$ ) of a desired multi-formant resonance structure. The output of a transfer function modeled on these specifications, as stimulated with a pulse, generates a waveform that is a sum of as many damped sinusoids as there are formants (see Figure A.4). The damping parameters of each sinusoid depend on the formant parameters.

This compound waveform is then repeated periodically at a rate equal to the desired fundamental frequency ( $F_0$ ). As  $F_0$  is modulated, what is modulated in the synthesis algorithm is the period between repetitions of the FWF. This generates a harmonic series with the harmonic ratios being maintained as the frequencies



**Figure A.4.** Structure of a parallel formant-wave-function (FWF) synthesizer (after Fig. 5 in Rodet, 1980b).  $F$  = formant frequency,  $A$  = formant amplitude,  $BW$  = formant bandwidth,  $WS$  = width of formant skirt.

modulate. Since the FWF determines the relative amplitudes of the frequency components, the result with frequency modulation is an amplitude modulation of each component coupled perfectly with the spectral form.

Finally, a global amplitude envelope function can be applied to the resulting signal as in the previous synthesis method. All calculations are carried out in 36-bit floating point format and the resulting signal is stored on disk in 18-bit integer format.

### A.3 Experimental Equipment and Sound Presentation

All experiments were conducted on a PDP-11/34 minicomputer with 18-bit digital-to-analog converters (DACs). The sound from the DACs was routed through a Neeve professional mixing console. This signal was sent out of two channels to a Revox A740 stereo power amplifier and then to AKG headphones. All experiments were run diotically (the same signal to both ears). The sound level was determined separately for each earphone prior to each experimental session. For this purpose a B & K sound level meter (set to A-weighting) with a flat-plate coupler was used. All experiments took place in an acoustically treated sound studio.

Two types of response boxes were used: a 2-button box with an LED associated with each button (built by William Hartmann), and a 3-switch, 3-slider box (built by Peter Easty). The blinking of lights and reading of the state of buttons, switches and sliders was done with subroutines (written by Bennett Smith) callable from the main experimental program. Thus, once calibration was completed each experimental session conducted itself automatically at a pace determined by the subject's responses.

## APPENDIX B

### Analysis and Synthesis of Jitter Functions

#### B.1 Introduction

Jitter is to be investigated as a potential contributor to spectral fusion of complex tones. The hypothesis is that frequency modulation maintaining coherence (constant ratios) across the spectral components belonging to a real or virtual sound source contributes to the simultaneous grouping of those components into a source image. In order to use jitter waveforms that are characteristic of various musical sources (e.g. voices and instruments) and to compare the ranges of jitter parameters found to have perceptual significance with those typically heard in musical situations, some kind of analysis procedure is necessary to extract the jitter.

Ideally, one would like to be able to specify the random (or otherwise) variation of frequency components, i.e. describe the time-varying instantaneous frequency. However, with standard spectral and/or temporal analysis procedures, one can never arise at an instantaneous frequency due to the time-frequency reciprocity inherent in Fourier-based analysis procedures. Different types of limitations arrive with different types of analysis techniques.

This appendix will first describe a zero-crossing estimation technique used to extract period by period variations in the frequency of sinusoidal components. Zero-crossing analysis has been used successfully for frequency fluctuation analyses for voice and bowed strings (Bjørklund, 1961; Lieberman, 1961; Cardozo & van Noorden, 1968; Baker, 1975; Kohut, Mathews, Miller & Zukovsky, 1981; MacIntyre, Schumacher & Woodhouse, 1981, 1982; Mathews & Miller, 1981). Also to be described, are means of characterizing the resulting jitter waveform and various artifacts, errors and

limitations of the analysis procedure. Then the procedure used to create the jitter waveforms for Experiments 1-7, 9-10 will be described and those jitter waveforms will be characterized according to the procedures of the first section.

## B.2 Jitter Analysis

For acoustic sources, the sounds were recorded on audio magnetic tape at 15 ips with DBX encoding for noise reduction. They were then digitized at a sampling rate of 32 kHz. The low-pass filters on the analog-to-digital converters attenuated any frequencies at and above the Nyquist frequency by at least 90 dB.

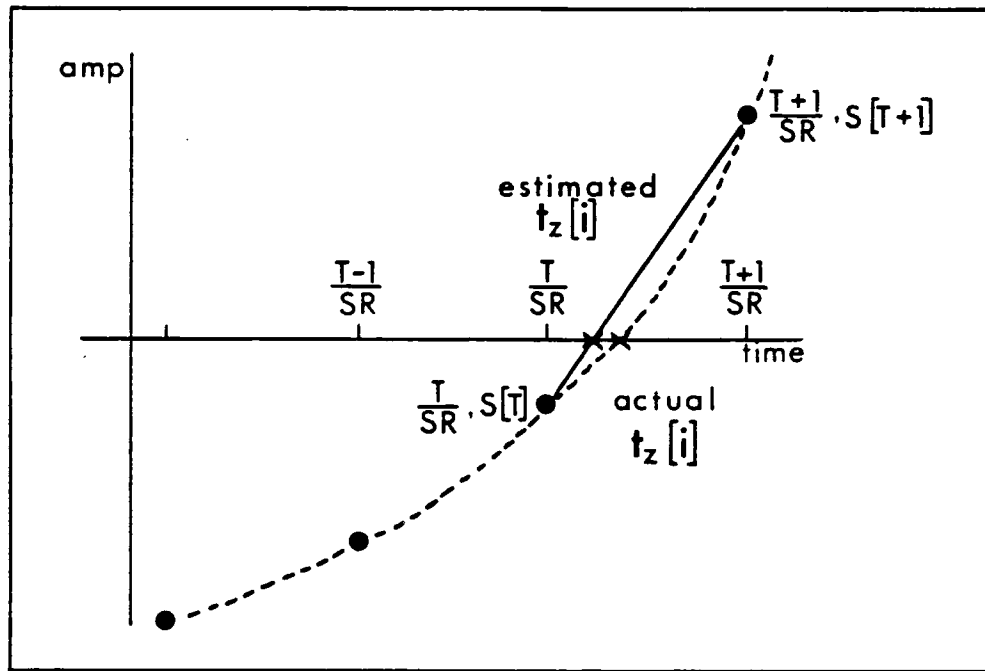
### B.2.1 *Preparation of Sounds*

An FFT was taken of the signal to determine the approximate mean value of the fundamental frequency ( $F_0$ ). Initially, the interest was in characterizing perturbations in the frequency of the driving force of these acoustic instruments, i.e. variation in the  $F_0$ . The signal was then filtered with an 8<sup>th</sup> order Butterworth band-pass filter with cutoff frequencies at  $1/2 F_0$  and a value just below  $2F_0$ . The lower cutoff was to remove any low frequency noise and any dc component in the signal. The upper cutoff frequency was to attenuate the upper harmonics so that their intensities were significantly less than that of the  $F_0$  to prevent multiple zero-crossings within a period. For sounds such as the male voice at 220 Hz  $F_0$ , where both the 2<sup>nd</sup> and 3<sup>rd</sup> harmonics are more intense than the  $F_0$ , it was necessary to use an upper cutoff frequency of  $1.65F_0$ . In general, the upper cutoff was adjusted to be only as low as necessary to remove most multiple zero-crossings and to minimize phase distortion in the region of the  $F_0$ . It was for this reason also that such a low-order filter was used. The zero-crossing analysis program took the filtered version of the digital sound file and searched for pairs of adjacent samples which were different in sign with the first being negative and the second positive, i.e. located samples surrounding a positive-going zero-crossing. The zero was then located in time by performing a linear interpolation between the samples. This is not as desirable perhaps as a sine wave interpolation. However, given that the amplitude of the signal was not known in advance and was varying anyway, the sine interpolation would be much more complicated to implement. The estimation error due to linear interpolation is a function of the frequency of the tone and the sampling rate. For lower frequencies the error is more than 2 orders of magnitude less than the range of interest. Thus, if an element

of an array,  $S[T]$ , contains the amplitude of a sound sample at sample number  $T$  just before the zero, and  $SR$  is the sampling rate, then the time of the  $i^{\text{th}}$  zero-crossing,  $t_z[i]$  can be determined as follows (see Figure B.1):<sup>1</sup>

the slope of the line connecting  $(S[T], \frac{T}{SR})$  and  $(S[T+1], \frac{T+1}{SR})$  is

$$\lambda = \frac{S[T+1] - S[T]}{\frac{T+1}{SR} - \frac{T}{SR}} = SR(S[T+1] + S[T]) \quad (\text{B.1})$$



**Figure B.1.** Determination of the time of the  $i^{\text{th}}$  positive-going zero-crossing,  $t_z[i]$ , from samples flanking that moment.

Since we are locating a positive-going zero-crossing, we know  $S[T]$  is negative and the slope is positive, therefore after some time,  $\Delta t$ , the amplitude will reach zero, i.e.

1. The curve in Figure B.1 has been drawn to exaggerate the error. A curve approximating a sine wave that was symmetrical about zero would have a point of inflection at the x-axis and the error would be much smaller.



$$S[T] + \lambda \Delta t = 0 \quad \text{or} \quad \Delta t = \frac{-S[T]}{\lambda} \quad (\text{B.2})$$

Therefore

$$t_z[i] = \frac{T}{SR} + \Delta t = \frac{T}{SR} - \frac{S[T]}{SR (S[T+1] - S[T])} \quad (\text{B.3})$$

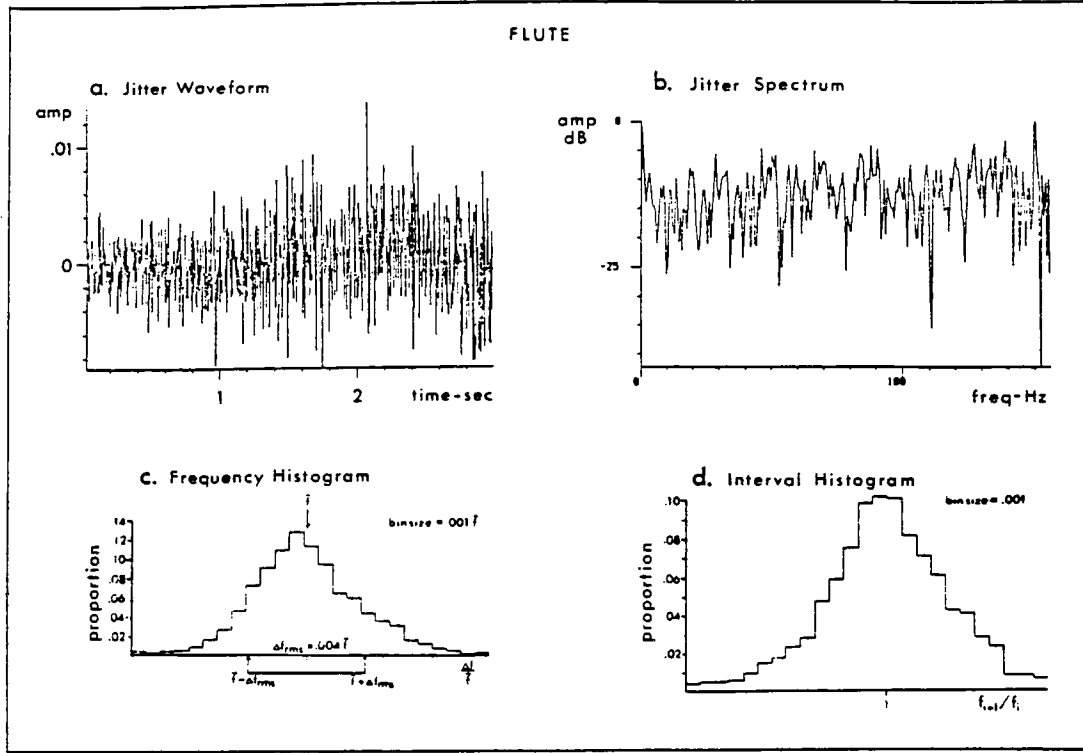
where  $t_z[i]$  is in seconds.

If  $i > 1$ , we can assign a period ( $P[i] = t_z[i] - t_z[i-1]$ ) and "cycle-frequency" ( $F_{cyc} = 1/P$ ; Baker, 1975) at time  $t_z[i]$ . In this way we obtain a series of cycle-frequency values. The program accepts a range of acceptable frequency values and then proceeds to discard values that are out of range. Generally, only two values are thrown out: those at the very beginning and end due to the presence of partial periods. Sometimes, however, one obtains multiple zero-crossings within a period when the 2<sup>nd</sup> harmonic is relatively strong. To account for this, the program checks to see if the sum of two adjacent periods is within the range of acceptable values and then replaces those two values with their sum. One obvious limit of this technique is that the program has difficulty with large excursions in  $F_0$ .

### B.2.2 Characterization of the Jitter Function

The series of frequency values was averaged to obtain the mean cycle-frequency,  $\bar{f}$ . Then the series was converted to proportion deviation from  $\bar{f}$  giving a modulation function  $Mod[i] = 1/\bar{f} F_{cyc}[i]$ . Setting the sampling rate as  $\bar{f}$  this function could be viewed as a representation of the jitter waveform imposed on the signals  $F_0$ . By taking the FFT of this time series, the spectral properties could be examined as well. (In this case the ordinate (in dB) would represent  $10 \Delta \log f$ .) Performing a sampling rate conversion with linear interpolation between samples, a digital sound file was created which could then be scaled and imposed on spectral components to obtain a desired rms deviation. Or it could be compared with an original, pre-determined jitter waveform for verification of the procedure. Verification of this process showed the technique to be sensitive to deviations one order of magnitude smaller than the smallest deviation used in the experiments. The error at a sampling rate of 16129 Hz and a signal frequency at 220 Hz is about  $\Delta f / \bar{f} = 2.8 \times 10^{-5}$ . In general, the error rises with signal frequency as would be expected due to a reduction in

the number of samples representing each period.

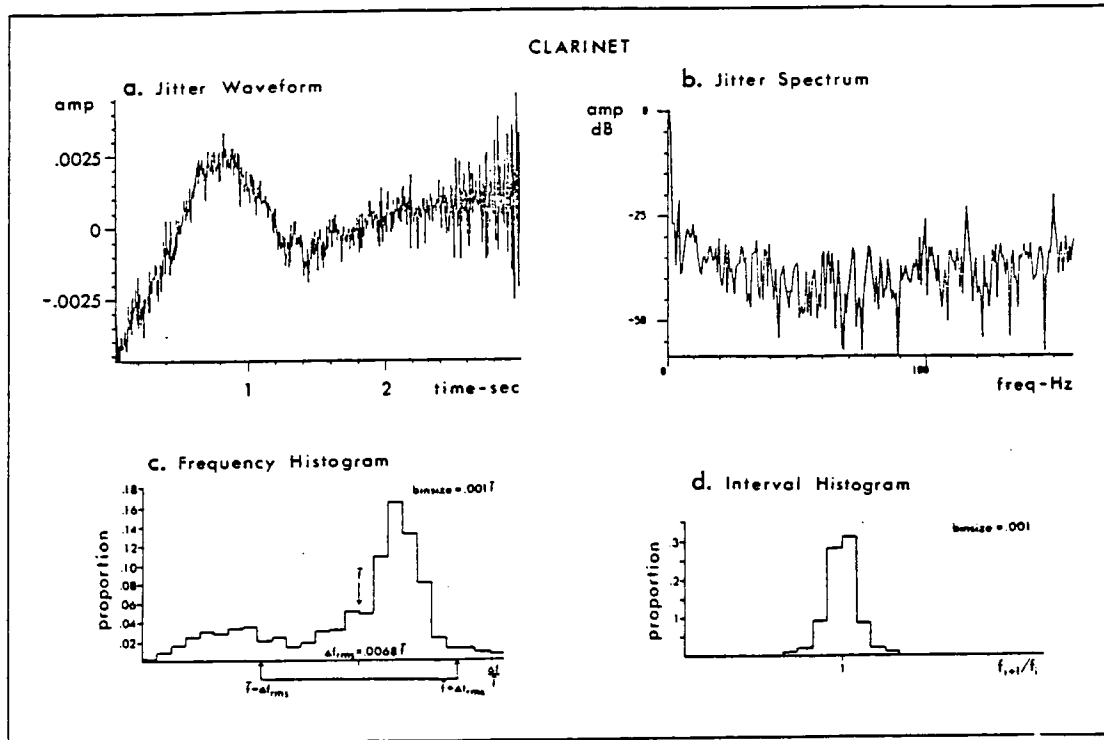


**Figure B.2.** Jitter data for flute playing an  $Eb_4$  at  $mf$ ; (a) time series of period deviations (jitter waveform), (b) spectrum of (a), (c) cycle-frequency histogram, (d) interval histogram.

The rms deviation of the function  $Mod[i]$  was determined according to

$$\frac{\Delta f_{rms}}{f} = \left( \frac{1}{N} \sum_{i=1}^N Mod^2[i] \right)^{\frac{1}{2}}. \quad (B.4)$$

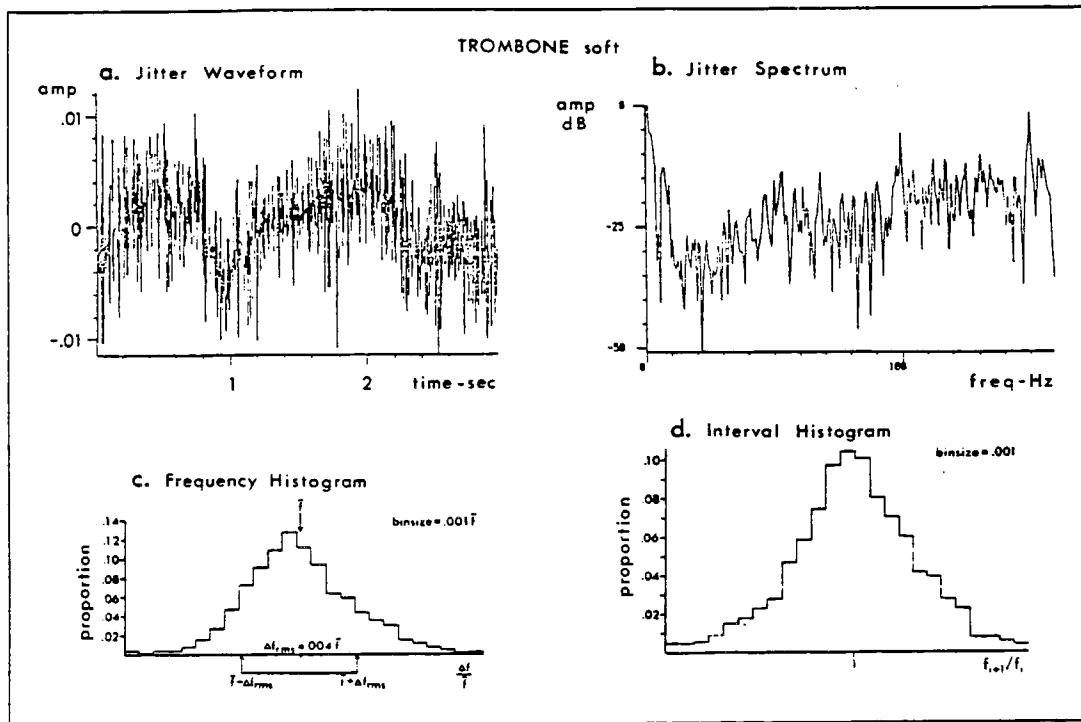
This represents the rms deviation as a proportion of the center frequency ( $\Delta f_{rms} / \bar{f}$ ) and can be converted to  $\Delta f_{rms}$  or to cents<sub>rms</sub> (see Eq. 2.1, Chap. 2).



**Figure B.3.** Jitter data for clarinet playing an  $Eb_4$  at  $mf$ ; (a) time series of period deviations (jitter waveform), (b) spectrum of (a), (c) cycle-frequency histogram, (d) interval histogram.

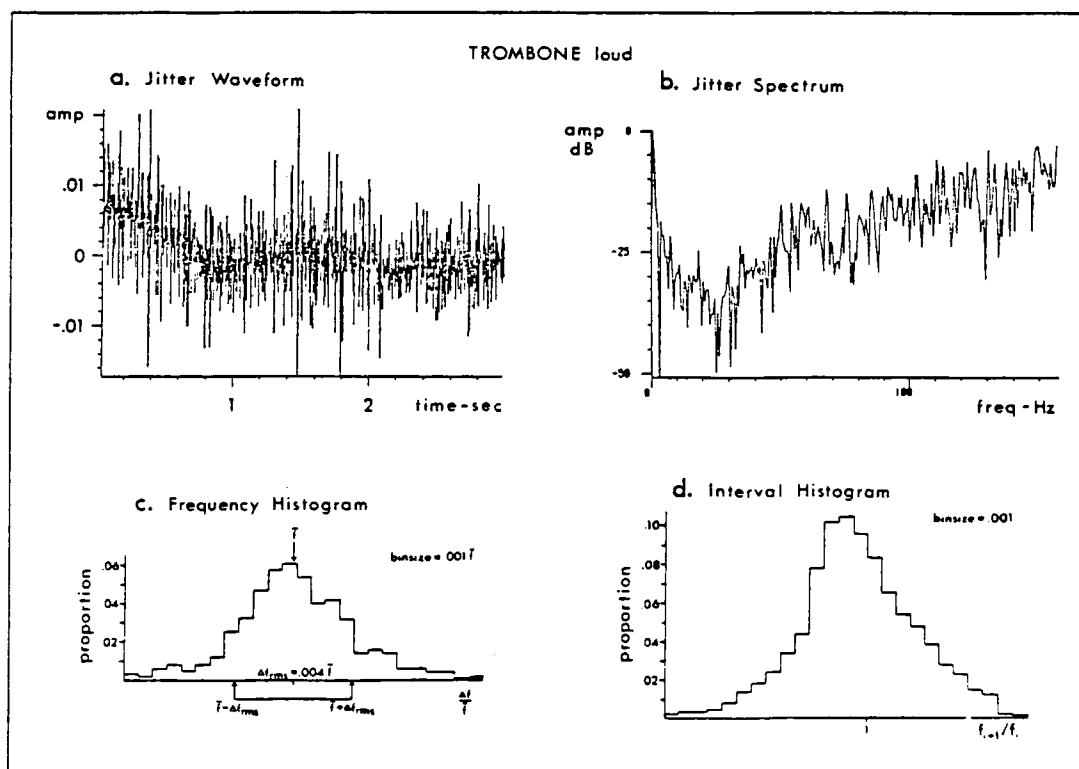
Two other characterizations involve constructing histograms of frequency deviation and of interval of change in frequency between periods. In the former, each cycle-frequency is placed in an appropriate bin and the result shows the frequency distribution of the  $F_0$  (this may also be viewed as the amplitude probability density function of the jitter waveform itself). This histogram shows how much time the

function spends in a certain region around  $\bar{f}$ . In the interval distribution, the ratio of cycle-frequencies between consecutive values in the time series is placed in an appropriate bin and the result shows the interval distribution of the deviation function. This shows the tendency of the function to be smooth or jumpy. The smoother the function, the greater will be the number of intervals close to 1.



**Figure B.4.** Jitter data for trombone playing an  $Eb_4$  at  $mf$ ; (a) time series of period deviations (jitter waveform), (b) spectrum of (a), (c) cycle-frequency histogram, (d) interval histogram.

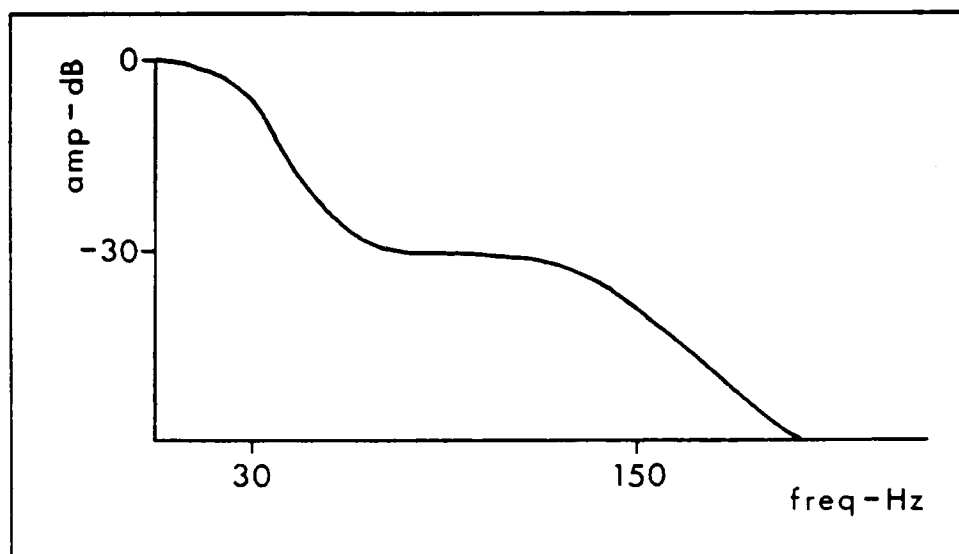
Four instrument sounds were analyzed with this technique: flute at  $Eb_4$  (311 Hz) playing  $mf$ , clarinet at  $Eb_4$  playing  $mf$ , trombone at  $Eb_4$  playing  $mf$  and  $ff$ . In Figure B.2 are shown the jitter (a) waveform, (b) spectrum, (c) frequency deviation histogram, and (d) frequency interval histogram. Those for the clarinet, trombone  $mf$  and trombone  $ff$  are shown in Figures B.3 - B.5, respectively. Note that the jitter



**Figure B.5.** Jitter data for trombone playing an  $Eb_4$  at  $ff$ : (a) time series of period deviations (jitter waveform), (b) spectrum of (a), (c) cycle-frequency histogram, (d) interval histogram.

spectrum of the flute shows it to be relatively flat, indicating that this may be mostly breath noise that fell within the filter band around the  $F_0$ , but with some very low

frequency fluctuations in the  $F_0$  as well. For the clarinet and trombone, we see much more energy below 10 Hz which is easily visible as slow fluctuations of the center frequency in the jitter waveform. For the loud trombone sound, more high frequency energy is present, perhaps because of increased turbulence due to greater air speed in the tube. The frequency interval histograms of all instruments are uniformly unimodal and reasonably symmetrical except for the clarinet's frequency histogram. The displaced peak may be related to the obvious slow fluctuation that starts at about 0.5% below  $\bar{f}$  and glides upward to about 0.3% above  $\bar{f}$  over a period of about 750 msec. We note also in the clarinet jitter waveform that the random fluctuations around the lower frequency intonational fluctuations are much smaller than in the other instruments. This is also reflected in the concentration of values around 1 in the interval histogram.

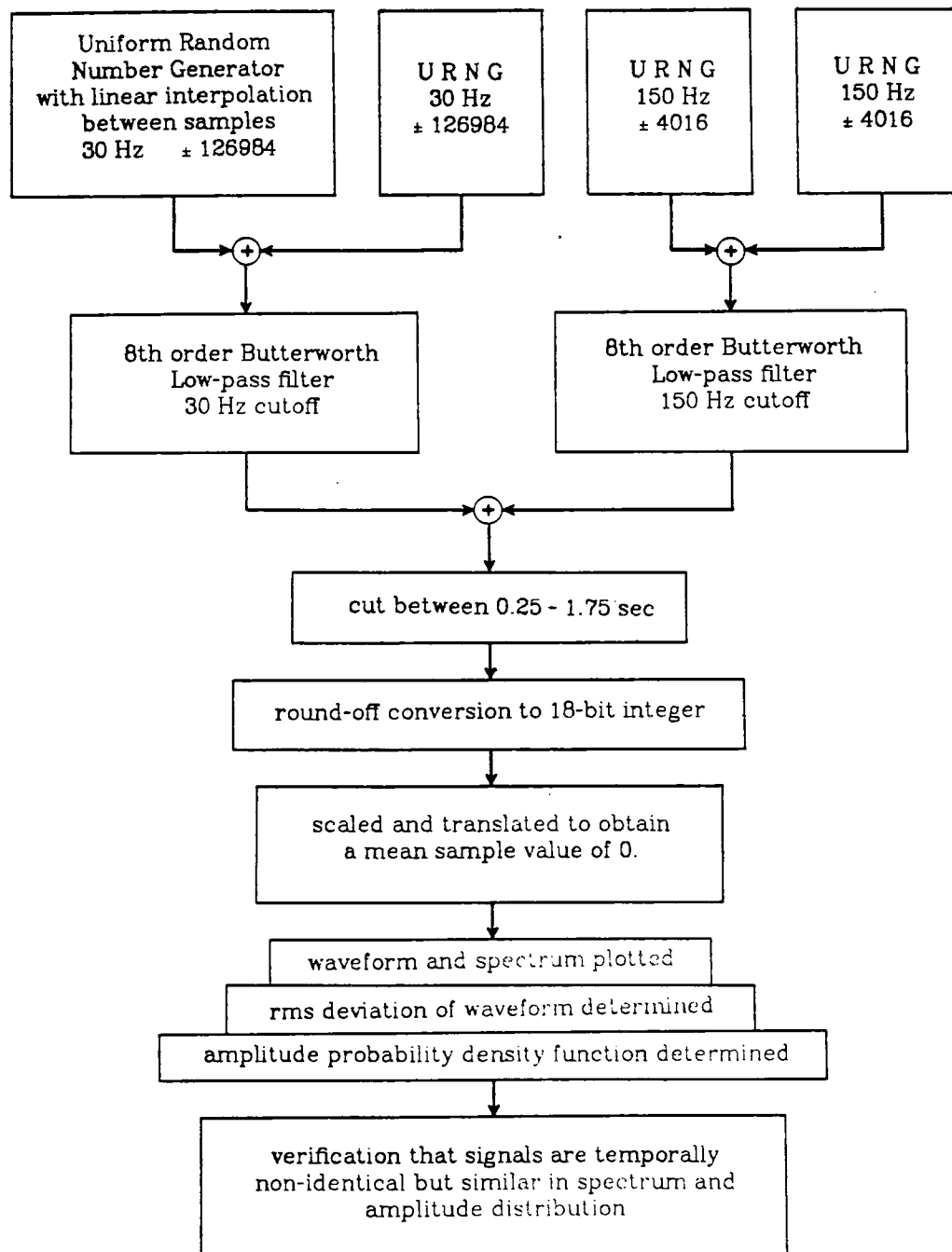


**Figure B.6.** Schematic diagram of a typical spectrum for the jitter function found on the fundamental frequency of a professional male singer singing a steady note without vibrato.

### B.3 Synthesis of Jitter Functions

Two independent jitter waveforms were synthesized and analyzed for use in these experiments. Analysis of the jitter in a trained male voice reveals a typical spectrum as schematized in Figure B.6. A jitter with this kind of spectrum is desirable for these experiments because it can be imposed on complex harmonic tones at rms deviations of over 3% frequency excursion without sounding noisy. This jitter can be modeled as the sum of 2 separate band-limited waveforms with different low-pass cutoff frequencies; one at  $\approx 30$  Hz and another at  $\approx 150$  Hz. The flatter portion of the 150 Hz band is  $\approx 30$ –40 dB below the level of the 30 Hz band. Each band was synthesized according to the procedure outlined in Figure B.7. The output of two uniform random number generators (choosing real 36-bit floating point values between  $\pm 1$ ) were added to give an amplitude probability density function that was more peaked near zero (this according to the Central Limit Theorem). When such a waveform is imposed as a frequency modulation function, this means there is a higher probability of an instantaneous frequency value near the center frequency. The fact that no values are to be found outside the specified range means that no sudden long-range leaps can occur in the waveform's amplitude (i.e. no sudden large frequency excursions are made when this jitter is imposed upon the partials. The generator is supplied an amplitude scaling factor ( $A$ ) and a "bandwidth" factor ( $BW$ ). The generator chose random values between  $\pm A$  at equal time intervals which are determined by  $BW$ . Then the amplitudes of all the samples between these points in time were linearly interpolated between the randomly chosen values. The outputs of the random generators were summed and the result was written to disk as a 2-second 18-bit integer file. Due to the procedure of sampling the random series at equal time intervals, spectral bands are generated above the specified  $BW$ , with peaks at odd multiples of  $BW$  and valleys at even multiples. The peaks decay as the frequency increases according to a  $\sin x/x$  function.

Consequently, this waveform was low-pass filtered at  $BW$  with an 8<sup>th</sup>-order Butterworth filter to attenuate the higher frequency spectral bands. This had the effect of smoothing the angularity of the linearly interpolated waveform. The filtered waveform was written as a 36-bit floating point file.



**Figure B.7.** Flow diagram of the process of jitter function synthesis and analysis



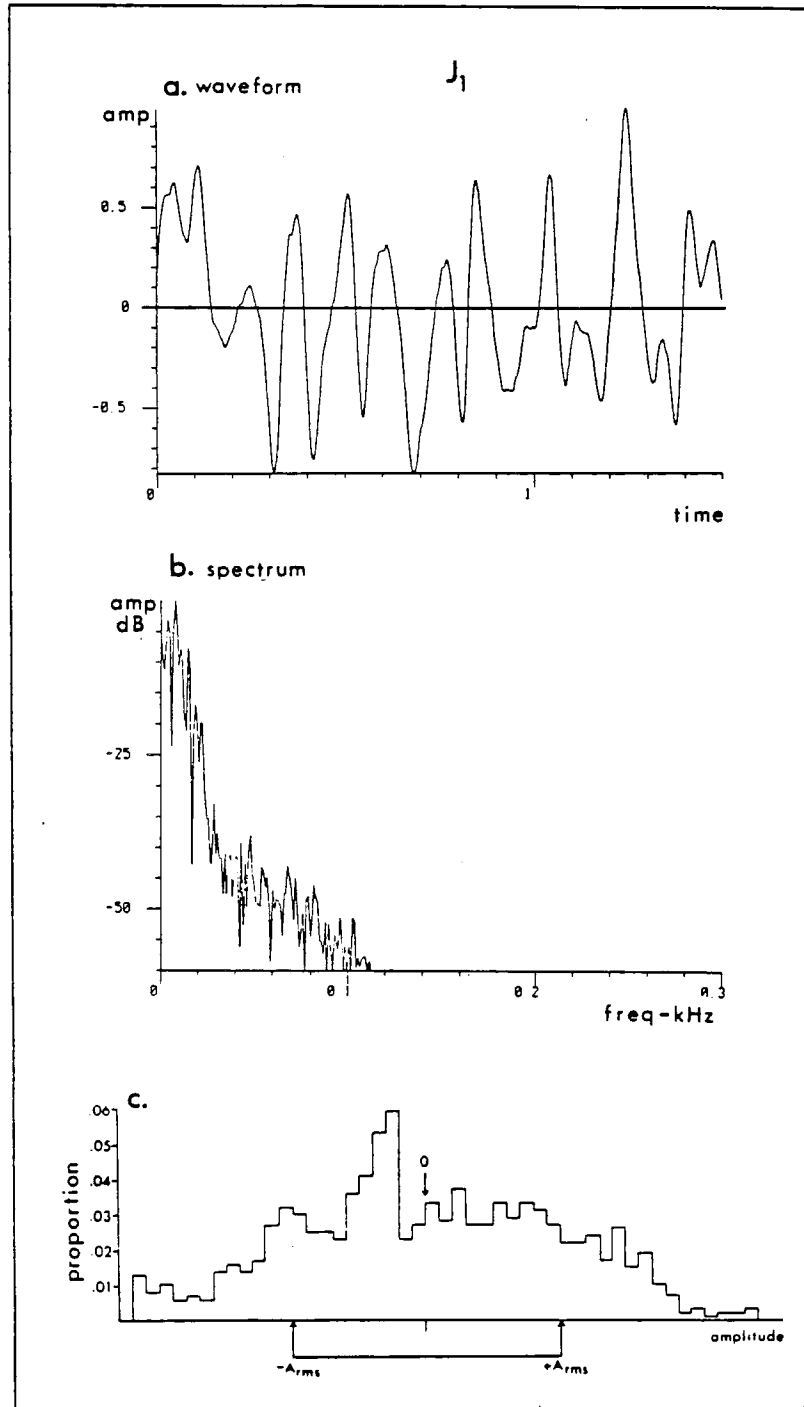
Two such files were created for each jitter file desired, one with a 30 Hz bandwidth and one with a 150 Hz bandwidth. The original maximum amplitudes specified to the random number generator for the 30 and 150 Hz *BWs* were chosen to give a level difference of approximately 30 - 40 dB between the two. At this point the samples of the waveforms were simply added together yielding a 36-bit floating point file which was converted by a rounding procedure to 18-bit integer format. The waveform was cut to a 1.5 sec duration between 0.25 and 1.75 sec to remove any disturbances at the onset and offset due to the filtering process.

The minimum, maximum and mean amplitudes of the waveform were determined and the waveform was translated to a mean of zero and scaled so that the absolute maximum amplitude was 131000 (maximum possible amplitudes are +131072 and -131071). All calculations were performed in floating point and the result was written in integer format. This constitutes the resultant jitter file.

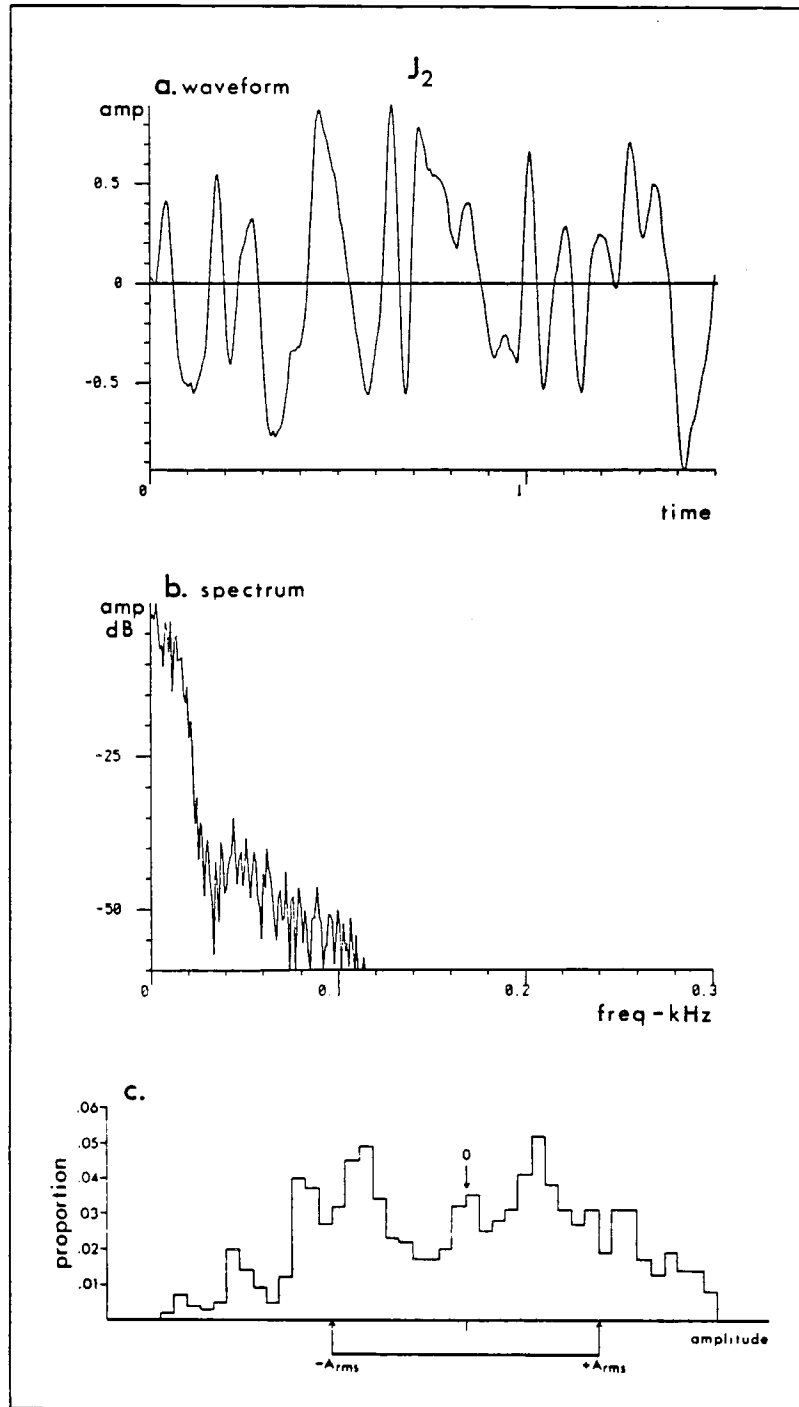
Each such jitter waveform was analyzed to determine its rms amplitude and amplitude distribution function. Its waveform and FFT were plotted. The FFT was taken over 16384 ( $2^{14}$ ) of the 24194 samples and was windowed with a class III 2<sup>nd</sup>-order windowing function.

Finally, both jitter waveforms were compared to verify that they conformed to the following criteria:

1. their waveforms were non-identical and thus temporally independent,
2. their spectral envelopes did not generally differ in overall shape,
3. their amplitude probability density functions were reasonably similar in overall shape.



**Figure B.8.** Characterization of jitter waveform  $J_1$ . Plotted below are (a) the 1.5 sec waveform, (b) the spectrum, and (c) the amplitude probability density function.



**Figure B.9.** Characterization of jitter waveform  $J_2$ . Plotted below are (a) the 1.5 sec waveform, (b) the spectrum, and (c) the amplitude probability density function.

Any jitter waveform not meeting these criteria was rejected and another was synthesized. An attempt was made to obtain two jitter waveforms which were different in their temporal fine structure and yet statistically similar in other respects. In Figure B.8 are shown (a) the waveform, (b) the spectrum and (c) the amplitude probability density function of waveform  $J_1$ . The same are shown for waveform  $J_2$  in Figure B.9. These meet all of the criteria above with the possible exception of number 3: amplitude probability density function. However, after sixteen different versions, these two were the closest in this respect and it was decided to use them.

## APPENDIX C

**EXPERIMENT 9:** Matching of perceived modulation width and loudness for complex tones with a frequency modulation maintaining constant frequency ratios or constant frequency differences among the partials.

In the experiments in Chapter 2, modulation was imposed on complex tones in two ways. These two modulations were designed to maintain either constant frequency ratios (*CR*) or constant frequency differences (*CD*) among 16 harmonic partials. If two tones are synthesized with identical spectral envelopes, rms amplitudes and modulation waveforms, and these two modulation types are imposed on the tone such that the modulation widths of the fundamental frequencies are equal and are well above modulation detection threshold, several differences are perceptible. One of the effects addressed by Experiments 1 and 2 is the difference in perceived fusion between the tones. Two other differences include unequal loudnesses and unequal perceived modulation widths. The purpose of this study was to match tones with various spectral envelopes and modulation waveforms for these two perceptual dimensions. Minimizing differences along these dimensions is required for Experiments 1 and 2.

### C.1 Modulation Width Matching

#### C.1.1 *Stimuli and Method*

All tones were composed of 16 harmonic partials of a 220 Hz fundamental frequency. Three spectral envelopes were used as in Chapter 2: flat (equal amplitude harmonics), -6 dB/octave spectral slope, and the vowel /a/. Two modulation waveforms were used: 6.5 Hz sinusoid (vibrato) and a low-frequency random waveform (jitter  $J_1$ , described in Appendix B). The two modulation types, *CR* and *CD*, are described in Chapter 2. All tones were synthesized with 14 cents rms deviation, i.e.

$D_{rms} = 0.00812$  in equations 2.2 and 2.3.

For *CD* tones, a series of  $k$  values ( $k = 1, 1.5, 2, 2.5, \dots$ ; see Eq. 2.3) were used to synthesize tones of each stimulus type (spectral envelope and modulation waveform). The subject was presented the *CR* tone of a particular stimulus type and could switch between this tone and any of the series of *CD* tones at will. The *CD* tones were labeled with their  $k$  values. The subject was to give a value (perhaps interpolated between those actually presented) for which the overall modulation width for *CR* and *CD* tones was perceived as being equal. Two subjects, both highly experienced listeners (and one being the author), participated in the experiment.

Stimuli were presented over headphones at 75 dBA in a sound insulated room (see Appendix A). Each stimulus was 1.5 sec in duration with 100 msec raised cosine ramps. Two matches were made to each *CR* tone by each subject.

### C.1.2 Results and Discussion

The means of the  $k$  values chosen by the subjects are listed in Table C.1. There was no significant difference between these mean values among the different spectral envelopes or modulation waveforms. Accordingly, the overall mean  $k$  value of 1.9 was chosen for the loudness study and for Experiments 1 and 2.

**TABLE C.1.** Data summary for modulation width matches. Each cell contains the mean values of  $k$  chosen to yield an equivalent overall modulation width between *CR* and *CD* tones.

	flat	-6 dB/oct	vowel /a/
vibrato	1.8	1.9	2.2
jitter	1.9	1.8	1.9

overall mean = 1.9

This result raises an issue about what is being attended to in listening to "overall" modulation width in *CD* tones. One subject remarked that if he were to listen "analytically", he would match the modulation width of the fundamental of the *CD* tone with that of the fundamental of the *CR* tone and thus choose smaller values. However, when he listened "synthetically" or "holistically", as he did in this study, he chose

larger  $k$  values. It seems that some kind of averaging of perceived modulation width weighted across the partials may be used in making the match. More research, beyond the scope of this dissertation, would be needed to open this area up for more careful investigation. To date, no research on modulation width perception of which I am aware has addressed problems with modulation of multi-component tones. The models that have been proposed for modulation of sinusoidal carriers with modulating signals of various degrees of complexity would not appear to apply, in their present form, to these results (cf. Klein & Hartmann, 1979).

In this experiment, both subjects noted that the *CR* tones were always louder than the *CD* tones even though both were presented at equal intensity. The purpose of the next study was to equalize the loudness differences due to spectral envelope, modulation waveform and modulation type.

## C.2 Loudness Matching

### C.2.1 Stimuli and Method

The stimuli were similar to those used in the modulation width experiment: 3 spectral envelopes (flat,  $-6$  dB/oct, vowel /a/), 2 modulation waveforms (vibrato, jitter - exactly as before), and 2 modulation types (*CR*, *CD*). For *CR* tones the modulation width was 14 cents ( $\Delta f / \bar{f} = 0.00812$ ). For *CD* tones the value of  $k$  was set to 1.9 thus yielding an rms modulation width of  $\pm 3.4$  Hz for all partials. Stimuli were synthesized at different amplitudes for subjects to compare among them.

The subject's task was to choose from among the tones of varying amplitudes the one that had an equal loudness with the reference tone. First, matches were made among the six *CD* stimuli to equalize for loudness differences due to modulation waveform and spectral envelope. Then a loudness match was made between *CR* and *CD* tones, each with the same modulation waveform and spectral envelope. Finally, the equality among *CR* tones was verified.

C.2.2 *Results and Discussion*

The attenuations (re: the most intense stimulus) necessary to achieve equal loudnesses are listed in Table C.2.

**TABLE C.2.** Data summary for loudness matches. Cell values represent attenuation in dB relative to the rms amplitude of the *CD* stimulus with vowel /a/ spectral envelope and sinusoidal modulation.

	constant ratio			constant difference		
	flat	-6 dB/oct	vowel /a/	flat	-6 dB/oct	vowel /a/
vibrato	-4.0	-4.7	-3.6	-1.2	-1.2	0
jitter	-4.2	-5.5	-4.2	-1.7	-1.7	-0.5

On the average, tones with jitter modulation were attenuated 0.5 dB relative to those with vibrato modulation. Tones with flat or -6 dB/oct spectral slopes were attenuated an average of 0.7 and 1.2 dB, respectively, relative to tones with the vowel /a/ spectrum. Most important of all, it was necessary to attenuate *CR* tones by an **average 3.3 dB** to equalize their loudnesses with respect to *CD* tones. This is an unusual finding. One thing that is different between *CR* and *CD* tones is their periodicity. *CD* tones are slightly aperiodic. Young & Sachs (1979) proposed that the degree of synchrony (a measure of periodicity of neural firing pattern across auditory nerve fibers) is used as a measure of intensity for nerve fibers whose mean firing rate is saturated. If this is so, then we might expect periodic sounds to be perceived as louder than aperiodic sound of similar physical intensity. However, this subject is beyond the scope of this dissertation and will not be pursued further here.



## APPENDIX D

**EXPERIMENT 10:** Frequency modulation detection of periodic and aperiodic waveforms imposed on complex harmonic tones.

The purpose of this experiment is to determine the frequency modulation detection thresholds (MDT) for jitter and vibrato waveforms imposed on harmonic complexes. This will be done in order to compare these thresholds with the ranges of modulation depth found to elicit different perceptual effects of sub-audio frequency modulation in other experiments.

### D.1 Stimuli

#### *Carrier signal:*

Harmonic tone complexes (16 consecutive equal-amplitude harmonics of 220 Hz) were synthesized with a duration of 1.5 sec and with 100 msec raised cosine attack and decay functions.

#### *Modulating signal:*

*Vibrato:* a sinusoidal signal of 6.5 Hz was used. This frequency is well within the range of musical vibrato frequencies. The sine wave was always started in sine phase at the beginning of the tone. All vibrato stimuli were presented at 75 dBA.

*Jitter:* two jitter waveforms ( $J_1(t)$ ,  $J_2(t)$  described in Appendix B) were used. These were approximately equivalent with respect to spectral content. They differed very slightly in amplitude probability density function and differed significantly in temporal fine structure. Jitter stimuli were presented at both 50 and 75 dBA.

These three waveforms were imposed on the tone complexes such as to maintain the ratios among the harmonics (see Eq. 3.1, Chapter 3).

## D.2 Method

For each trial, two tones were presented in succession with a 100 msec silence between them. One tone was modulated and the other was unmodulated. The order of the tones was chosen randomly and counterbalanced. The subject held a 2-button response box with a light associated with each button. With the presentation of each tone the appropriate button was illuminated. The subject's task was to decide which interval contained the modulated tone and press the corresponding button. Feedback was given by flashing the light corresponding to the correct response button 300 msec after a button was pressed. All stimuli were presented binaurally over headphones (audio system described in Appendix A).

The Levitt (1971) 2IFC 1-up/2-down tracking procedure was used. The rms deviation of the modulation was varied in 0.25 cent steps. This step size, when measured as peak deviation is 0.35 cents for vibrato and 0.65 and 0.55 cents for jitter waveforms 1 and 2, respectively. After an incorrect response,  $\Delta$ cents was increased one step. After two consecutive correct responses,  $\Delta$ cents was decreased one step. Each change in direction of  $\Delta$ cents was counted as a turnaround and 16 turnarounds composed a run. An estimate of the modulation detection threshold was calculated as the mean of the  $\Delta$ cents values from the last 12 turnarounds. Five consecutive runs for each modulation waveform were collected. The mean of these five estimates was taken as a measure of the MDT (71% choice).

Six subjects were tested but not all subjects completed all conditions. MDTs were determined for all subjects that participated in Experiment 6 for both jitter waveforms and both at 50 and 75 dBA levels. The presentation order of these intensity conditions was counterbalanced across subjects.

## D.3 Results

The results for all Ss are presented in Table D.1. The threshold values are expressed as rms deviation in both cents<sub>rms</sub> and as  $\Delta f_{rms} / \bar{f}$ .

**TABLE D.1.** Experiment 10 data summary. Cell values represent the 71% MDT for a complex harmonic carrier determined with an adaptive tracking procedure. MDTs are listed as both  $\Delta$ cents<sub>rms</sub> and as  $\Delta f_{rms} / \bar{f}$ . The means across subjects for jitter stimuli (in  $\Delta$ cents<sub>rms</sub>) are also indicated.

Modulation Waveform	Subject	Stimulus Intensity			
		50 dBA		75 dBA	
		$\Delta$ cents <sub>rms</sub>	$\frac{\Delta f_{rms}}{\bar{f}} \times 10^{-3}$	$\Delta$ cents <sub>rms</sub>	$\frac{\Delta f_{rms}}{\bar{f}} \times 10^{-3}$
Vibrato	1	—	—	3.25	1.88
	2	—	—	2.27	1.31
	3	—	—	2.10	1.21
	4	—	—	5.06	2.93
	$\bar{x}$	—		3.17	
	s	—		1.36	(N = 4)
Jitter ( $J_1$ )	1	2.88	1.66	3.12	1.80
	2	2.63	1.54	3.60	2.08
	3	—	—	2.37	1.37
	4	1.67	0.97	3.86	2.23
	5	1.64	0.95	2.31	1.34
	6	1.68	0.97	2.47	1.43
	$\bar{x}$	2.10	(N = 5)	2.96	(N = 6)
	s	0.60		0.67	
Jitter ( $J_2$ )	1	3.13	1.81	3.42	1.98
	2	1.99	1.15	4.85	2.81
	3	—	—	2.09	1.21
	5	0.90	0.52	1.49	0.86
	6	1.23	0.71	1.87	1.08
	$\bar{x}$	1.81	(N = 4)	2.74	(N = 5)
	s	0.99		1.38	
Jitter overall mean	$\bar{x}$	1.97	(N = 9)	2.86	(N = 11)
	s	0.76		1.00	

On the average, vibrato MDTs are not significantly different from those for jitter waveforms at 75 dBA. Nor is there a difference between the MDTs for the two jitter waveforms. Again, as expressed in Chapters 2 and 4, rms deviation seems to be a highly appropriate measure of frequency modulation width for many perceptual effects.

There was a significant difference between the pooled means of the jitter stimuli for the two intensity conditions ( $p < .05$ ). The MDTs for 50 dBA stimuli were significantly lower than those for 75 dBA stimuli. This cannot be attributed to learning since the presentation order was counterbalanced. I find no apparent reason for this effect, which is consistent across subjects.

The vibrato detection data agree reasonably well with those obtained by Hartmann and Klein (1980) for 4 Hz vibrato imposed on 800 Hz sine carriers. These investigators found thresholds in the range of 1.3 to 3.2 cents<sub>rms</sub>. However, the jitter detection data (when expressed as rms deviation) are 20 - 30% of the thresholds found by Pollack (1968; approximately 10 cents<sub>rms</sub> under similar conditions) and Cardozo and Neelen (1968; 8.5 - 11.2 cents<sub>rms</sub>). Both of these studies used pulse trains where the interpulse interval was randomly varied about the mean interpulse interval. In Pollack's stimuli the jitter had a rectangular distribution. Cardozo & Neelen did not specify the frequency content of their jitter.

## APPENDIX E

### DATA TABLES

**Table E.1: Experiment 1**

**Table E.2: Experiment 3**

**Table E.3: Experiment 4**

**Table E.4: Experiment 5**

**Table E.5: Experiment 6**

**Table E.6: Experiment 7**

**Table E.7: Experiment 8**

VIBRATO							JITTER						
Rms Deviation (cents)							Rms Deviation (cents)						
Spec-							Spec-						
trum	S #	7	14	28	42	56	trum	S #	7	14	28	42	56
F	1	.58	.82	.90	.98	.96	F	1	.52	.72	.94	1.00	1.00
	2	.80	.82	.98	.98	.96		2	.54	.90	.88	1.00	1.00
	3	.62	.80	.88	.88	.90		3	.94	.96	.88	1.00	1.00
	4	.50	.46	.38	.46	.40		4	.62	.66	.64	.44	.46
	5	.52	.46	.84	.96	.94		5	.48	.70	.86	.92	.90
	6	.64	.60	.83	.86	.88		6	.50	.50	.62	.70	.77
	7	.56	.86	1.00	.94	1.00		7	.62	.62	.88	.92	.96
	8	.40	.62	.80	.86	.96		8	.60	.56	.86	.96	.96
	$\bar{x}$	.58	.68	.83	.87	.88		$\bar{x}$	.60	.70	.82	.87	.88
	$\sigma$	.12	.17	.19	.17	.20		$\sigma$	.15	.16	.12	.20	.19
without S4, S7													
	$\bar{x}$	.59	.69	.87	.92	.93		$\bar{x}$	.60	.72	.84	.93	.94
	$\sigma$	.13	.15	.06	.06	.03		$\sigma$	.17	.18	.11	.12	.09
6	1	.48	.74	.98	1.00	1.00	6	1	.50	.58	.92	.96	.92
	2	.58	.92	.98	1.00	1.00		2	.44	.84	.76	1.00	1.00
	3	.62	.68	.88	.98	1.00		3	1.00	1.00	.98	1.00	1.00
	4	.64	.98	.94	.86	.88		4	.88	1.00	.82	.74	.72
	5	.64	.70	.88	.90	.90		5	.70	.94	.88	.92	.92
	6	.58	.98	.92	.98	.96		6	.68	.90	.72	.88	.84
	7	.50	.66	.60	.58	.68		7	.48	.60	.54	.58	.54
	8	.52	.56	.88	.96	.94		8	.50	.64	.88	.90	.94
	$\bar{x}$	.57	.78	.88	.91	.92		$\bar{x}$	.65	.81	.81	.87	.86
	$\sigma$	.06	.16	.12	.14	.11		$\sigma$	.21	.18	.14	.14	.16
without S4, S7													
	$\bar{x}$	.57	.76	.92	.97	.97		$\bar{x}$	.64	.82	.86	.94	.94
	$\sigma$	.06	.16	.05	.04	.04		$\sigma$	.21	.17	.10	.05	.06

**Table E.1 (continued)**

VIBRATO							JITTER						
Rms Deviation (cents)							Rms Deviation (cents)						
Spec-	S #	7	14	28	42	56	Spec-	S #	7	14	28	42	56
A	1	.80	1.00	1.00	1.00	1.00	A	1	.52	.80	.94	.94	.96
	2	.66	.74	.92	.98	.98		2	.42	.70	.76	.98	.98
	3	.44	.72	.90	.96	.98		3	.92	.92	.98	1.00	1.00
	4	.50	.62	.68	.82	.64		4	.94	.90	.84	.82	.86
	5	.52	.56	.88	.92	.94		5	.60	.82	.90	.92	.94
	6	.58	.86	.94	.96	.98		6	.58	.78	.94	.70	.92
	7	.40	.54	.88	.88	.84		7	.64	.54	.68	.74	.76
	8	.44	.62	.80	.92	.92		8	.72	.68	.76	.86	.94
	$\bar{x}$	.54	.71	.88	.93	.91		$\bar{x}$	.67	.77	.85	.87	.92
	$\sigma$	.13	.16	.10	.06	.12		$\sigma$	.18	.12	.11	.11	.08
without S4, S7													
	$\bar{x}$	.57	.75	.91	.96	.97		$\bar{x}$	.63	.78	.88	.90	.96
	$\sigma$	.14	.16	.07	.03	.03		$\sigma$	.17	.09	.10	.11	.03

**TABLE E.2.** Experiment 3 data summary. Each cell value represents the proportion of 25 2IFC judgments where the constant frequency difference tone was chosen as yielding more sources than the constant frequency ratio tone. All tones had a vibrato modulation waveform.

Spectrum	S #	Rms Deviation on $F_0$ (cents <sub>CR</sub> /cents <sub>CD</sub> )		
		14/13.3	28/26.5	56/52.8
Flat	1	.68	.86	1.00
	2	.48	.80	1.00
	3	.60	.84	.84
	4	.24	.56	1.00
	$\bar{x}$	.50	.76	.96
	$\sigma$	.19	.14	.08
-6 dB/oct	1	.52	.88	.84
	2	.44	.68	1.00
	3	.36	.76	.84
	4	.24	.32	.76
	$\bar{x}$	.39	.66	.86
	$\sigma$	.12	.24	.10
vowel /a/	1	.56	.88	1.00
	2	.36	.68	1.00
	3	.72	.84	1.00
	4	.00	.64	1.00
	$\bar{x}$	.41	.76	1.00
	$\sigma$	.31	.12	.00



**TABLE E.3.** Experiment 4 data summary. Each cell value represents the proportion of 30 2IFC judgments where the constant frequency difference tone was chosen as having a larger modulation width on  $F_0$  than the constant frequency ratio tone. All tones had a vibrato modulation waveform. The value in parentheses under the rms deviation is the difference in modulation on  $F_0$  between  $CR$  and  $CD$  tones:  $\Delta\text{cents} = \text{cents}_{CD} - \text{cents}_{CR}$ . The rms deviation of modulation refers to the value of  $\psi$ , given that  $k = 1.9$  for  $CD$  tones (see Eqs. 2.2, 2.3).

Spectrum	S #	Rms Deviation of Modulation (cents)				
		7 (6.3)	14 (12.5)	28 (24.8)	42 (36.9)	56 (48.9)
flat	1	.43	.70	.97	1.00	1.00
	2	.50	.77	1.00	.97	.97
	3	.37	.80	1.00	.97	1.00
	4	.70	.60	.97	1.00	.93
	$\bar{x}$	.50	.72	.98	.98	.97
	$\sigma$	.14	.09	.02	.02	.03
-6 dB/oct	1	.53	.93	1.00	1.00	.93
	2	.70	.93	1.00	.93	1.00
	3	.60	.80	1.00	1.00	1.00
	4	.53	.90	1.00	.97	1.00
	$\bar{x}$	.59	.89	1.00	.97	.98
	$\sigma$	.08	.06	.00	.03	.03
vowel /a/	1	.43	.63	.97	1.00	1.00
	2	.47	.70	.93	.97	1.00
	3	.57	.73	1.00	1.00	1.00
	4	.27	.67	.93	1.00	.97
	$\bar{x}$	.43	.68	.96	.99	.99
	$\sigma$	.12	.04	.03	.01	.01

**TABLE E.4.** Experiment 5 data summary. Each cell value represents the proportion of 25 2IFC judgments where the constant frequency difference tone was chosen as having a larger modulation on  $F_0$  than the constant frequency ratio tone. All tones had a vibrato modulation waveform. The value in parentheses under the rms deviation is the difference in modulation on  $F_0$  between  $CR$  and  $CD$  tones:  $\Delta\text{cents} = \text{cents}_{CR} - \text{cents}_{CD}$ .

Spectrum	S #	Rms Deviation ( $\text{cents}_{CR} / \text{cents}_{CD}$ )		
		14/13.3 (0.7)	28/26.5 (1.5)	56/52.8 (3.2)
flat	1	.06	.10	.76
	2	.06	.40	.70
	3	.36	.66	1.00
	4	.20	.10	.06
	$\bar{x}$	.17	.31	.63
	$\sigma$	.14	.27	.40
-6 dB/oct	1	.46	.46	.70
	2	.46	.30	.46
	3	.40	.76	.76
	4	.50	.26	.40
	$\bar{x}$	.45	.44	.58
	$\sigma$	.04	.23	.18
vowel /a/	1	.00	.16	.80
	2	.06	.06	.36
	3	.36	.66	.90
	4	.10	.20	.36
	$\bar{x}$	.13	.27	.60
	$\sigma$	.16	.27	.29

**TABLE E.5.** Experiment 6 data summary. Individual data are the proportion of times in 30 2IFC judgments that the tone with incoherent modulation on one partial was chosen as having more sources. Means and unbiased standard deviations across subjects are also listed.

**Table E.5.1:** Data for harmonic stimuli presented at 75 dBA (H75).

Incoherent Partial	Subject Number	Rms Deviation of Modulation (cents)				
		2.00	5.00	8.00	11.00	14.00
1	1	.53	.53	.57	.63	.90
	2	.73	.93	.87	.87	.83
	3	.50	.50	.37	.70	.90
	4	.43	.53	.47	.43	.53
	$\bar{x}$	.55	.62	.57	.66	.79
	$\sigma$	.13	.19	.22	.18	.18
		0.50	2.00	3.50	5.00	6.50
3	1	.47	.77	.93	.97	1.00
	2	.43	.87	1.00	1.00	.97
	3	.50	.63	.80	.90	.90
	4	.47	.53	.73	.63	.57
	$\bar{x}$	.47	.70	.87	.87	.86
	$\sigma$	.03	.15	.12	.17	.20
		0.30	1.50	2.70	3.90	5.10
5	1	.40	.80	.97	.93	.97
	2	.60	.93	1.00	1.00	1.00
	3	.47	.93	.97	.97	.97
	4	.53	.93	1.00	.97	1.00
	$\bar{x}$	.50	.90	.98	.97	.98
	$\sigma$	.08	.06	.02	.03	.02
		0.30	1.50	2.70	3.90	5.10
7	1	.53	.63	.77	1.00	.97
	2	.60	.87	.97	.97	1.00
	3	.63	.80	.97	.97	.97
	4	.47	.93	1.00	1.00	1.00
	$\bar{x}$	.56	.81	.93	.98	.98
	$\sigma$	.07	.13	.11	.02	.02

Table E.5.1: H75 Data (continued)

Incoherent Partial	Subject Number	Rms Deviation of Modulation (cents)				
		0.30	1.20	2.10	3.00	3.90
9	1	.47	.60	.87	1.00	.97
	2	.43	.97	1.00	1.00	1.00
	3	.50	.97	1.00	1.00	.97
	4	.57	.93	1.00	1.00	1.00
	$\bar{x}$	.49	.87	.97	1.00	.98
	$\sigma$	.06	.18	.06	0.00	.02
		0.05	0.30	0.55	0.80	1.05
11	1	.50	.70	.73	1.00	1.00
	2	.50	.60	.90	.90	1.00
	3	.43	.53	.90	.97	.90
	4	.53	.70	.90	1.00	1.00
	$\bar{x}$	.49	.63	.86	.97	.97
	$\sigma$	.04	.08	.08	.05	.05
		0.05	0.30	0.55	0.80	1.05
13	1	.50	.80	.97	1.00	1.00
	2	.60	1.00	1.00	1.00	.97
	3	.43	.83	.93	1.00	.97
	4	.60	.93	1.00	1.00	1.00
	$\bar{x}$	.53	.89	.97	1.00	.98
	$\sigma$	.08	.09	.03	.00	.02
		0.05	0.30	0.55	0.80	1.05
15	1	.57	.53	.53	.97	1.00
	2	.53	.70	.93	.97	1.00
	3	.63	.77	.83	.83	.93
	4	.43	.47	.83	.93	1.00
	$\bar{x}$	.54	.62	.78	.92	.98
	$\sigma$	.08	.14	.17	.07	.03
		0.05	0.30	0.55	0.80	1.05

**Table E.5.2:** Data for harmonic stimuli presented at 50 dBA (H50).

Incoherent Partial	Subject Number	Rms Deviation of Modulation (cents)				
		2.00	5.00	8.00	11.00	14.00
1	1	.60	.50	.63	.70	.67
	2	.47	.40	.60	.30	.50
	3	.43	.47	.50	.47	.70
	4	.50	.47	.70	.50	.43
	$\bar{x}$	.50	.46	.61	.49	.57
	$\sigma$	.07	.04	.08	.16	.13
		0.50	2.00	3.50	5.00	6.50
3	1	.57	.37	.43	.63	.73
	2	.47	.53	.60	.27	.50
	3	.47	.43	.37	.37	.53
	4	.57	.43	.70	.53	.60
	$\bar{x}$	.52	.44	.52	.45	.59
	$\sigma$	.06	.07	.15	.16	.10
		0.30	1.50	2.70	3.90	5.10
5	1	.50	.63	.77	.93	.90
	2	.50	.43	.80	.97	.97
	3	.47	.60	.70	.83	.77
	4	.33	.53	.63	.77	.60
	$\bar{x}$	.45	.55	.72	.87	.81
	$\sigma$	.08	.09	.08	.09	.16
		0.30	1.50	2.70	3.90	5.10
7	1	.40	.67	.93	.90	.87
	2	.33	.93	.97	1.00	1.00
	3	.40	1.00	.97	1.00	1.00
	4	.47	.73	.90	.83	.90
	$\bar{x}$	.40	.83	.94	.93	.94
	$\sigma$	.06	.16	.03	.08	.07

Table E.5.2: H50 Data (continued)

Incoherent Partial	Subject Number	Rms Deviation of Modulation (cents)				
		0.30	1.20	2.10	3.00	3.90
9	1	.43	.67	.83	.97	.93
	2	.47	.97	1.00	1.00	1.00
	3	.60	.97	1.00	1.00	1.00
	4	.43	.93	.97	1.00	1.00
	$\bar{x}$	.48	.88	.95	.99	.98
	$\sigma$	.08	.14	.08	.01	.03
		0.05	0.30	0.55	0.80	1.05
11	1	.33	.60	.60	.90	.93
	2	.53	.43	.40	.73	.93
	3	.53	.60	.47	.63	.80
	4	.60	.60	.57	.87	1.00
	$\bar{x}$	.50	.56	.51	.78	.91
	$\sigma$	.12	.08	.09	.13	.08
		0.05	0.30	0.55	0.80	1.05
13	1	.50	.57	.67	.87	.90
	2	.50	.67	.87	1.00	1.00
	3	.50	.60	.73	.83	.87
	4	.63	.57	.97	1.00	1.00
	$\bar{x}$	.53	.60	.81	.92	.94
	$\sigma$	.06	.05	.14	.09	.07
		0.05	0.30	0.55	0.80	1.05
15	1	.57	.47	.53	.77	.83
	2	.60	.67	.70	.83	.97
	3	.40	.67	.83	.90	.93
	4	.43	.73	.87	.87	1.00
	$\bar{x}$	.50	.63	.73	.84	.93
	$\sigma$	.10	.11	.15	.06	.07
		0.05	0.30	0.55	0.80	1.05

Table E.5.3: Data for inharmonic stimuli presented at 75 dBA (I75).

Incoherent Partial	Subject Number	Rms Deviation of Modulation (cents)				
		3.00	6.00	9.00	12.00	15.00
1	1	.50	.83	1.00	1.00	1.00
	2	.53	.77	.90	.90	.97
	3	.77	.93	.93	.97	.97
	4	.70	.97	1.00	1.00	1.00
	$\bar{x}$	.62	.87	.96	.97	.98
	$\sigma$	.13	.09	.05	.05	.02
		2.00	4.00	6.00	8.00	10.00
3	1	.60	.57	.87	.93	.97
	2	.47	.70	.73	.93	.90
	3	.50	.87	.90	1.00	.93
	4	.67	.80	1.00	1.00	1.00
	$\bar{x}$	.56	.73	.87	.97	.95
	$\sigma$	.09	.13	.11	.04	.04
		0.50	1.30	2.10	2.90	3.70
5	1	.60	.67	.73	.93	.93
	2	.37	.27	.53	.70	.70
	3	.57	.47	.27	.47	.63
	4	.40	.87	.83	.70	.77
	$\bar{x}$	.48	.57	.59	.70	.76
	$\sigma$	.12	.26	.25	.19	.13
		0.50	2.00	3.50	5.00	6.50
7	1	.63	.57	.87	.93	.97
	2	—	—	—	—	—
	3	.63	.63	.90	.87	.93
	4	.73	.70	.77	.83	1.00
	$\bar{x}$	.66	.63	.85	.88	.87
	$\sigma$	.06	.06	.07	.05	.03

Table E.5.3: I75 Data (continued)

Incoherent Partial	Subject Number	Rms Deviation of Modulation (cents)				
		0.50	2.50	4.50	6.50	8.50
9	1	.57	.83	.90	.93	.90
	2	—	—	—	—	—
	3	.63	.67	.93	.93	1.00
	4	.50	.77	.87	.97	1.00
	$\bar{x}$	.57	.76	.90	.94	.93
	$\sigma$	.06	.08	.03	.02	.06
		0.50	2.50	4.50	6.50	8.50
11	1	.47	.70	.80	.87	.93
	2	—	—	—	—	—
	3	.47	.90	.80	1.00	1.00
	4	.57	.50	.90	.87	.97
	$\bar{x}$	.50	.70	.83	.91	.97
	$\sigma$	.06	.20	.06	.07	.03
		0.50	2.50	4.50	6.50	8.50
13	1	.60	.53	.53	.93	.90
	2	—	—	—	—	—
	3	.47	.63	.87	.97	1.00
	4	.77	.73	.87	.87	1.00
	$\bar{x}$	.61	.63	.76	.92	.93
	$\sigma$	.15	.10	.20	.05	.06
		0.50	2.50	4.50	6.50	8.50
15	1	.63	.80	.87	.87	.93
	2	—	—	—	—	—
	3	.60	1.00	.93	1.00	1.00
	4	.63	.70	.97	.87	.93
	$\bar{x}$	.62	.83	.92	.91	.95
	$\sigma$	.02	.15	.05	.07	.04
		0.50	2.50	4.50	6.50	8.50



**TABLE E.6.** Experiment 7 data summary. Each value represents the proportion of times in 50 2IFC judgments the subject chose the *CCA* tone as yielding more sources than the *CSE* tone. For the spectral envelopes, **6** is the -6 dB/oct spectral slope, and **A** is the vowel /a/ spectrum. The means and unbiased standard deviations across subjects are shown at the bottom of each stimulus category.

VIBRATO							
Spec- trum	S #	Rms Deviation of Modulation (cents)					
		7	14	28	42	56	70
<b>6</b>	1	.52	.46	.50	.40	.56	.48
	2	.48	.64	.54	.48	.42	.50
	3	.46	.54	.54	.68	.68	.56
	4	.54	.52	.52	.56	.56	.48
	5	.38	.52	.56	.54	.50	.62
	6	.50	.56	.56	.50	.60	.58
	7	.40	.52	.56	.52	.54	.42
	8	.30	.38	.56	.62	.62	.46
	9	.56	.60	.64	.64	.64	.48
	$\bar{x}$	.46	.53	.55	.55	.57	.51
	$\sigma$	.08	.07	.04	.09	.08	.06
<b>A</b>	1	.46	.68	.90	1.00	.96	.98
	2	.56	.80	.92	1.00	1.00	1.00
	3	.36	.56	.82	.92	.98	.96
	4	.68	.46	.82	.92	.90	.94
	5	.46	.52	.54	.56	.76	.70
	6	.54	.54	.48	.74	.90	.96
	7	.46	.46	.46	.58	.70	.68
	8	.54	.52	.52	.80	.86	.88
	9	.46	.56	.64	.86	.92	.90
	$\bar{x}$	.50	.57	.68	.82	.89	.89
	$\sigma$	.09	.11	.19	.16	.10	.12

Table E.6 (continued)

Spec- trum	JITTER						
	S #	Rms Deviation of Modulation (cents)					
		7	14	28	42	56	70
6	1	.48	.46	.50	.50	.52	.46
	2	.44	.38	.54	.50	.50	.64
	3	.52	.54	.46	.48	.48	.50
	4	.42	.38	.38	.44	.48	.42
	5	.60	.56	.48	.46	.44	.58
	6	.52	.48	.60	.54	.48	.50
	7	.48	.38	.38	.48	.46	.50
	8	.58	.42	.50	.54	.58	.44
	9	.58	.54	.44	.44	.38	.46
	$\bar{x}$	.51	.46	.48	.49	.48	.50
	$\sigma$	.06	.07	.07	.04	.05	.07
A	1	.50	.52	.76	.88	.92	.94
	2	.32	.44	.86	.98	.98	1.00
	3	.52	.54	.68	.98	1.00	1.00
	4	.40	.50	.60	.46	.60	.72
	5	.56	.44	.50	.52	.60	.88
	6	.48	.46	.82	.96	.98	1.00
	7	.34	.44	.56	.58	.52	.56
	8	.56	.56	.48	.64	.70	.78
	9	.56	.60	.90	.92	.96	.98
	$\bar{x}$	.47	.50	.68	.77	.81	.87
	$\sigma$	.09	.06	.16	.21	.20	.16

**TABLE E.7.** Experiment 8 data summary. For each stimulus condition a separate vowel prominence rating (0 - 100) was made for each of the target vowels /a/, /o/ and /i/. The data represent the means ( $\bar{x}$ ) and unbiased standard deviations ( $\sigma$ ) of normalized data across 10 Ss. (See Chapter 5 for a discussion of the normalization procedure.) The permutation is the pitch ordering of the 3 vowels with the lowest pitch first. The vibrato state of the *figure* and *ground* sources are specified as described in Chapter 5.

Permu- tation	Vibrato fig gnd		TARGET VOWEL					
			/a/		/o/		/i/	
			$\bar{x}$	$\sigma$	$\bar{x}$	$\sigma$	$\bar{x}$	$\sigma$
aoi	N	S	72.	8.0	51.	8.4	60.	14.2
aio	N	S	73.	6.0	47.	6.9	56.	15.2
oai	N	S	67.	9.9	57.	9.0	49.	8.5
oia	N	S	53.	21.2	62.	12.7	51.	11.3
iao	N	S	68.	17.9	51.	13.0	45.	6.6
ioa	N	S	53.	11.5	58.	10.6	50.	11.7
aoi	A	S	79.	2.3	54.	11.7	58.	11.2
aio	A	S	80.	7.7	45.	7.4	57.	10.7
oai	A	S	77.	8.9	61.	8.9	51.	10.2
oia	A	S	79.	5.2	59.	8.8	54.	9.3
iao	A	S	81.	2.8	55.	12.0	52.	6.0
ioa	A	S	78.	11.9	59.	11.7	51.	8.2
aoi	O	S	75.	8.8	65.	7.6	54.	9.4
aio	O	S	78.	6.7	64.	10.9	52.	7.9
oai	O	S	71.	14.0	65.	8.1	47.	2.7
oia	O	S	54.	10.4	67.	11.1	50.	8.9
iao	O	S	81.	6.7	65.	12.3	46.	7.2
ioa	O	S	60.	10.9	70.	9.1	47.	9.5
aoi	I	S	73.	7.5	54.	9.0	76.	12.3
aio	I	S	80.	7.1	50.	11.2	63.	10.9
oai	I	S	71.	9.2	54.	6.8	60.	12.9
oia	I	S	65.	12.7	61.	12.7	56.	11.6
iao	I	S	76.	6.0	50.	8.1	46.	8.4
ioa	I	S	53.	13.1	59.	11.1	53.	10.7

Table E.7: (continued)

Permu- tation	Vibrato fig gnd		TARGET VOWEL					
			/a/		/o/		/i/	
			$\bar{x}$	$\sigma$	$\bar{x}$	$\sigma$	$\bar{x}$	$\sigma$
aoi	N	V	79.	3.7	53.	10.7	69.	13.4
aio	N	V	80.	5.5	58.	11.6	59.	9.8
oai	N	V	73.	11.2	63.	6.2	60.	13.5
oia	N	V	81.	4.9	70.	9.1	53.	6.2
iao	N	V	79.	4.8	60.	12.5	54.	7.3
ioa	N	V	71.	10.6	63.	10.5	53.	7.6
aoi	A	V	79.	5.6	54.	8.0	64.	10.9
aio	A	V	82.	5.7	59.	13.8	57.	7.7
oai	A	V	75.	11.2	68.	12.9	60.	13.1
oia	A	V	80.	9.3	68.	9.2	53.	7.1
iao	A	V	82.	4.4	61.	14.1	54.	10.6
ioa	A	V	77.	8.5	68.	7.0	51.	7.3
aoi	O	V	78.	4.1	55.	10.5	71.	14.4
aio	O	V	82.	7.6	63.	11.6	59.	8.1
oai	O	V	80.	10.4	64.	10.8	54.	13.9
oia	O	V	77.	13.1	69.	8.7	55.	5.5
iao	O	V	81.	4.8	58.	7.8	51.	7.3
ioa	O	V	77.	10.7	68.	7.0	45.	5.0
aoi	I	V	75.	6.8	52.	8.9	70.	12.8
aio	I	V	80.	6.8	57.	11.7	61.	11.8
oai	I	V	75.	9.4	61.	9.6	59.	9.5
oia	I	V	77.	11.6	70.	13.0	56.	6.8
iao	I	V	80.	6.9	52.	8.5	53.	8.9
ioa	I	V	70.	11.0	64.	10.1	55.	7.6

## APPENDIX F

### Pilot Studies on Spectral Fusion and Spectral Parsing

#### F.1 Informal Preliminary Studies

The following is a brief summary of informal pilot work reported in McAdams (1982b). All of the stimuli reported were generated using early versions of the additive synthesis instrument described in Appendix A. The instrument was realized at IRCAM during the summer of 1980.

##### F.1.1 *Fusion Investigations*

1. *Familiarity of the spectral envelope.* Four envelopes were used. Two were vowels, /a/ as in father and /o/ as in rope. (These were derived from analyses of a male voice by X. Rodet at IRCAM.) Two were unnatural (triangular) envelopes, one with a +6 dB/oct slope and one with a -6 dB/oct slope. Stimuli were 2 sec in duration and contained 16 harmonics of a 200 Hz fundamental frequency ( $F_0$ ). Listeners reported that the vowel shapes seemed more fused than the triangular spectra.
2. *Harmonicity of the frequency content.*  $F_0$  of 200 Hz with 16 partials and 2 sec duration. Spectral envelope of /a/ as in the previous demonstration. Four frequency content types were used:
  - a. harmonic ( $f_n = n F_0$ ).
  - b. shifted harmonic ( $f_n = (n + .5) F_0$ ).
  - c. stretched harmonic ( $f_n = n^{1.07} F_0$ ), and
  - d. compressed harmonic ( $f_n = n^{0.93} F_0$ ).

Most listeners found the order of relative fusedness to be: harmonic compressed, stretched and shifted, proceeding from most to least fused. The differences in fusion between stretched and compressed stimuli were minor.

3. *Vibrato and jitter.* Random jitter and periodic vibrato were combined with relative amplitudes yielding 0.8% random and 2% periodic modulation of the component frequencies. These values were chosen to result in strong fusion in the harmonic case and to sound relatively natural. The random jitter had a spectral distribution roughly corresponding to -6 dB/oct spectral slope and was band-limited between 8 and 25 Hz. The frequency of the vibrato was 6.5 Hz. Due to the construction of the synthesis instrument, which controls the amplitude of a given partial based on its instantaneous frequency, there was an amplitude modulation of the component accompanying the frequency modulation such that the partials followed the resonance peaks in the vowel spectrum. The vowel /a/ envelope was superimposed on the 4 frequency content types used in the previous demonstration. The same ordering of relative fusion was found for the content types, but it was further found that the degree of fusion was much greater with modulated tones than with unmodulated tones, even for inharmonic complexes. The two triangular-envelope harmonic stimuli were also subjected to vibrato and, although they seemed more fused than without vibrato, the increase in fusion was less than with the vowel spectral shape.
4. *Portamento, with constant ratios maintained among the partials.* Vowel /a/ spectral envelope, 16 partials, 2 sec duration. Four frequency content types as in the two previous demonstrations. A simple frequency glide was used such that the fundamental glided linearly from 100 to 300 Hz. Again, the spectral envelope was independent of the instantaneous frequency content. Gliding tones were found to fuse very strongly, maintaining the order of fusion among frequency content types. It was striking that even with the inharmonic complexes there was still a compelling vocal quality that resulted as the amplitudes of the partials "traced out" the spectral envelope of the vowel /a/.

In general, it was found that all of these parameters contribute to fusion but that their combined, interactive contributions created the most life-like images.

F.1.2 *Parsing Investigations*

1. *Changing simultaneous source interpretations.* 16 harmonics with  $F_0$  at 250 Hz; vowel /a/ spectral envelope; 4 sec duration. Initially all spectral components had the same combined random and periodic vibrato. Between 1.5 to 2.5 sec in the stimulus, the original vibrato function was faded out from either the even or the odd harmonics while an independent vibrato function was gradually introduced (see Figure 6.8, Chap.6). The new function had an independent jitter and a periodic component that differed from that of the original by 0.5 Hz. With both stimuli (change in odd or in even harmonics), the impression is that a "new" voice enters at about halfway through the stimulus and has a pitch corresponding to an octave above the original. Two things should be noted:
  - a. only very minute changes to the long term spectrum are being produced by this change and yet a very compelling appearance of a voice at the octave results from the parsing of the even and odd harmonics.
  - b. The second, more surprising, result is that the timbre of the original voice does not appear to change despite the fact that as far as pitch processing is concerned, half of its harmonics have been parsed into a different source, to the extent that they give rise to a new pitch sensation. This occurs whether the change is introduced to the odd or the even harmonics. The image is one of a "new" voice entering at the octave, the original remaining undisturbed. This has implications for pitch and timbre processing as context dependent perceptual attributes and for auditory grouping as a dynamic, adaptive process of modeling the source based on both previously established criteria and on the incoming, dynamic stimulus information.
2. *Multiple source superimposition and vibrato onset asynchrony.* Three voices (all vowel /a/), each with 16 partials, and 3 different  $F_0$ 's arranged in an augmented chord, i.e.  $F_0' = 1.25F_0$ ,  $F_0'' = 1.25F_0'$  for harmonic stimuli, and  $F_0' = 1.27F_0$ ,  $F_0'' = 1.27F_0'$  for the stretched inharmonic stimuli. In other words, the chord was stretched by the same amount as the spectral content so that all of the spectral relations between overlapping spectral subsets were

stretched as well. The lowest partial of the lowest voice was 200 Hz in both cases. In both stimuli, all 48 partials onset synchronously without vibrato. At 1.5 sec, the vibrato of the  $F_0$  subset was introduced (again combined random and periodic modulation was used to simulate a singing voice). At 3 sec, that for  $F_0'$  and at 4.5 sec that for  $F_0''$  were introduced. All modulation functions were independent.

The initial impression is of a complex tone mass without much harmonic structure to it and a great deal of beating-partial interactions were present. Very shortly after the introduction of a given vibrato, that subset fused into a singing male voice against the remaining tone mass. Then another voice appeared at the major 3<sup>rd</sup> and finally a third voice (somewhat more weakly fused) appeared an augmented 5<sup>th</sup> above the lowest voice. The same strong fusion of individual voices and parsing of separate voices occurred even with the stretched stimulus, indicating that such a context may play a strong role in the fusion and parsing of inharmonic complexes. Though the top voice was weak, there was a fairly good preservation of three pitches and three voice timbres.

3. *Fusion of adjacent partials and increased "spectral resolution" with vibrato.* A long steady harmonic tone complex with 16 harmonics was presented. A melody was created by synchronously jittering two adjacent partials at a time. All partials remained at the same amplitude, only sometimes they were jittered. Against the steady tone, a melody emerged (harmonized by the drone provided by the steady components), and elements of the melody were still prominent when even the 15<sup>th</sup> and 16<sup>th</sup> partials were jittered. Thus pairs of harmonics were made to stand out from the rest solely by virtue of their being jittered. This last example demonstrated that it may be possible to increase the degree of resolution of partials in regions of a harmonic complex (above harmonic number 8 or 9) previously considered to be "uresolved" by the auditory system (cf. Plomp, 1976).



## F.2 Formal Pilot Studies

Three pilot studies were conducted to test the efficacy of experimental measures in elucidating the role of certain fusion and parsing cues in the formation of auditory images and in the association of perceptual attributes such as pitch and timbre to those images. In these studies, combinations of different types of spectral content and spectral envelope were used in conjunction with a combined periodic and random frequency modulation waveform in order to illuminate to some extent the role of strength of fusion of single images in the parsing of multiple images. It is expected that more strongly fused images, such as are formed from familiar, harmonic, modulated signals would also be more easily separable.

The informal preliminary studies suggested the importance of the correlation of frequency modulation functions between separate sets of spectral components belonging to different sources. In those studies, the tone had the spectral shape of a vowel and a combination of vibrato and jitter was used to obtain a natural quality of a singing voice when the tone complex was harmonic. If, for example, all harmonics had the same modulating function, the image perceived was that of a single male voice singing a given vowel. However, if the even and odd harmonics of the same spectrum received different, independent modulating functions, a double image often resulted with two male voices singing the same vowel one octave apart. This has two implications that need to be investigated:

1. Is the lack of correlation between modulating functions an important aid in parsing the spectrum into separate sources?
2. Are the pitch and timbre that are assigned to an image derived from the spectrum of a given parsed subset?

Experiments A and B investigated the influence of the degree of correlation between modulating functions of two subsets of the spectrum on the perceived (reported) number of sources. Experiment C investigated the relation of the correlation to the number of pitches perceived and the relation between the pitches perceived and the manner of parsing harmonic and inharmonic spectra. In performing some early exploratory studies it became obvious that a synchronous onset of the two sets of partials was stronger as a cue for a fused image than the uncorrelated

modulation functions were used as cues for separate images when the amplitudes of the modulation were small (i.e. the modulation deviation was small). Accordingly, an asynchrony was introduced to all stimuli regardless of the degree of correlation of the modulation functions. What one expects to observe in such cases is an increase in separability with modulation presence.

### F.2.1 *Experiment A*

This experiment investigated the effects of spectral envelope, harmonicity, onset asynchrony and modulation correlation across separate spectral subgroups on the perceived number sources in a 16-component stimulus.

#### F.2.1.1 *Stimuli*

36 stimuli, each containing 16 partials, were generated combining:

1. *4 spectral envelopes* of varying degrees of presumed familiarity and complexity. These were /a/, /o/, -6 dB/oct slope, and flat spectral slope. The data for the vowel spectra were obtained from Rodet at IRCAM. Here the interest was in the difference in response to spectral envelopes that are frequently encountered in everyday life and those that are usually confined to the laboratory.
2. *3 types of spectral content* of varying degrees of harmonicity. These were harmonic, shifted harmonic, and stretched harmonic as described in the informal preliminary studies above (section F.1.1). One rationale for using inharmonic tones such as these was to address the question of whether the number of pitches influences the number of source images. Each spectral content type was divided into two sets of 8 partials. The parsing scheme was different for each type:
  - a. *harmonic* — divided into even and odd harmonics,
  - b. *shifted* — divided 1,2,3,4,10,11,12,13 and 5,6,7,8,9,14,15,16,
  - c. *stretched* — divided 1,2,5,6,9,10,13,14 and 3,4,7,8,11,12,15,16.

These last two were chosen by trial and error such that some pitch change was noticeable between the condition when the modulation functions for the two

subsets were perfectly correlated and when they were independent.

3. *3 degrees of correlation* between the frequency modulation functions imposed on the two subsets of the spectral components. The modulation functions were composed of a random component (rectangular amplitude distribution, bandlimited between 0 and 32 Hz, with a modulation excursion of 0.2%) and a periodic component (frequency of 6 or 6.5 Hz at 0.5% modulation width) to simulate the natural quality of a singing voice with the vowel spectral envelopes (Chowning, 1980). This combined modulation function was found in preliminary studies to increase the perceived fusion of harmonic and inharmonic tones. When the spectral subsets were perfectly correlated they both had the same modulation function  $A(f, t)$  superimposed on all frequency components such that constant ratios among the partials were maintained. This has been shown to be an important factor in the fusion of harmonic, complex tones (Bregman, McAdams & Halpern, 1978). When the subsets were uncorrelated, the periodic components differed in frequency by 0.5 Hz and the random components were independent. Thus, the components of one spectral subset were modulated by a function  $A(f, t)$  while those of the other subset were modulated by a function given by

$$c A(f, t) + (1 - c) B(f, t), \quad \text{for } 0 \leq c \leq 1, \quad A \text{ and } B \text{ independent.} \quad (\text{F.1})$$

For  $c = 1$ , the functions are identical for the two groups of partials; for  $c = 0$ , the functions are independent; for  $c = 0.5$ , the functions are partially correlated.

The onset of the two subsets of partials were asynchronous by 70 msec for the uncorrelated stimuli. The group of partials with the lowest frequency always started first. As discussed previously, this asynchrony has been shown to facilitate the parsing of sources. All partials onset synchronously for correlated stimuli. The offsets of all partials were synchronous for all stimuli. In these stimuli, a combination of onset asynchrony and lack of correlation between the two groups of partials is used to induce a separation of the source images.

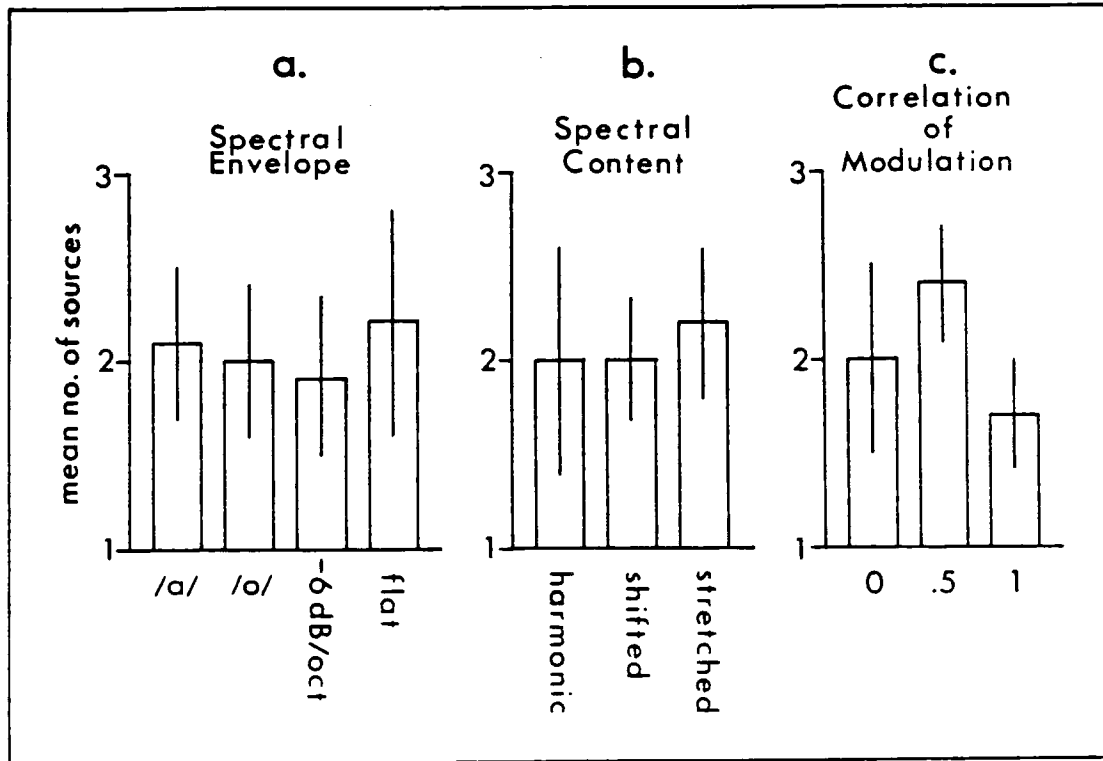
All tones had a duration of 1.5 sec. This relatively long duration was chosen in order to allow the subject time to listen into the sound to determine how many "sources" were present (see below). The  $F_0$  was 196 Hz ( $G_3$ ). Each tone had an amplitude envelope with 10 msec raised cosine attack and decay functions to minimize transients. The stimuli were synthesized on the Systems Concepts Digital Synthesizer at CCRMA (Dept. of Music, Stanford University) at a sampling rate of 37,710 Hz. The highest frequency present in any stimulus was 3808 Hz, well below the Nyquist frequency of 18,855 Hz.

Stimuli were randomized in a block and four such blocks were recorded on tape directly from the digital-to-analog converters of the synthesizer. Tones were recorded on one track of the tape and a stop-tone for each trial was recorded on the other track. This stop-tone was relayed to a remote control device which stopped the tape player. The tape was then restarted when the subject pressed a button after having recorded the response. The stimuli were presented over a loudspeaker approximately 1 m from the listener's head at a comfortable listening level (approximately 75 dB SPL) in an IAC sound isolation chamber.

#### F.2.1.2 *Method*

Nine adult subjects of both sexes participated in the experiment and were paid for their services. Five of the subjects were musicians and/or composers. All four blocks were presented in one session. The first block was used as practice and data were collected from the last 3 blocks. Subjects were instructed to listen to each tone and decide how many separate "voices", "instruments", "sources" or "images" they heard. No attempt was made to restrict what each subject might consider a "source image" to be. This was one informal aim of the study. Subjects were questioned afterward as to their response strategies. For each stimulus they were asked to write down a number. If the subject felt that there were more sources present than they were able to distinguish and count, they were instructed to respond "many" and also to write down a number as a rough approximation. This was to allow for the expression of ambiguity of perception expected to result either with cases of partial correlation or with inharmonic stimuli. In these studies, an attempt was made to establish the range and richness of possible perception and the tendencies of perception rather than restricting what the subject was allowed to perceive in order to quantify their responses to their perceptions. At this point, the possibilities and

predilections of perception were more interesting than the rigorous quantification of its limits.



**Figure F.1.** Experiment A data summary. Average number of sources reported for (a) 4 types of spectral envelope, (b) 3 types of spectral content, and (c) 3 degrees of correlation between modulating functions of two simultaneously-present spectral subsets.

### F.2.1.3 Results and Discussion

There were some predictable and some surprising results. The average number of sources reported are collected under each type of main category in Figure F.1. Figure F.1a shows the results for spectral envelope shape averaged across subjects, replications, spectral content and correlation. There is not much difference in response to the envelopes for /a/, /o/ and -6 dB/oct slope. A greater average number of sources was reported for the flat spectrum. This is a bit surprising at first,

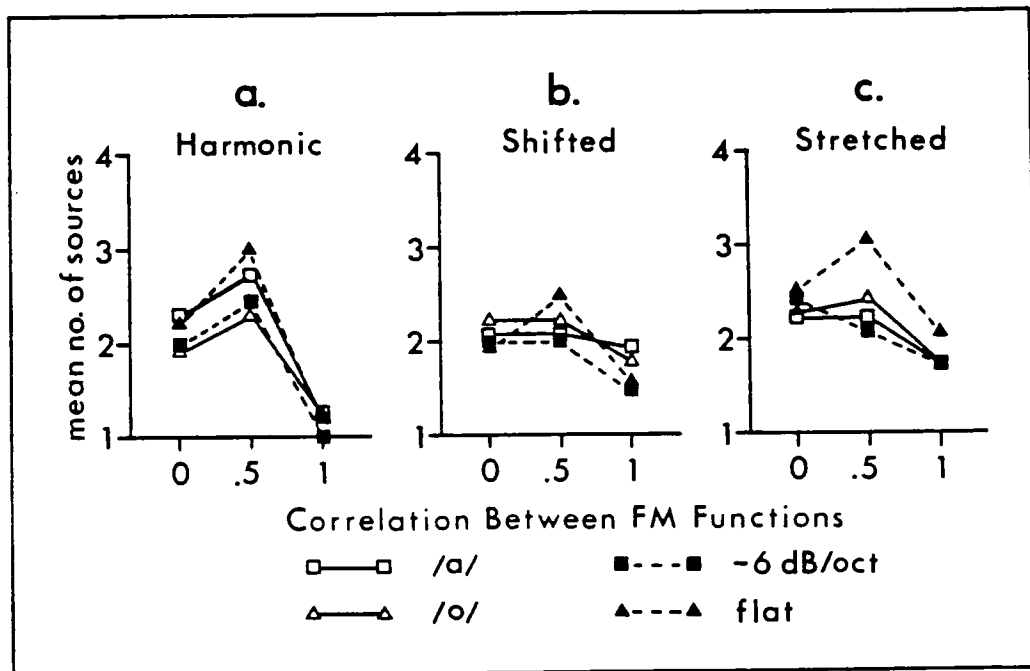
since familiarity of envelope was expected to induce better fusion and thus decrease the number of perceived sources; and the  $-6$  dB/oct slope was initially considered as an "unfamiliar" envelope along with the flat spectrum. However, due to the sawtoothed nature of excitation of bowed strings (catch-release-etc.) the spectral shapes of certain instruments, e.g. violin and viola, have overall slopes somewhere between  $-3$  to  $-6$  dB/oct (Hall, 1980). Several subjects, when presented with the harmonic  $-6$  dB/oct stimulus, remarked afterward that it had a quality like that of a bowed string, though it did not sound exactly like any familiar Western stringed instrument. The important point is that it has a more familiar quality along with the vowel sounds and its perceptual response seems to group with the vowels as will be shown below. It may be, then, that familiarity or recognizability (or at least complexity, to some degree) of the spectral envelope plays a role in fusion and that fusion affects the number of perceived sources.

Figure F.1b shows the results for spectral content averaged across subjects, replications, spectral shape and correlation. This is a very mild effect though it shows a tendency for more sources to be reported for stretched stimuli than for harmonic or shifted ones. It is somewhat surprising that fewer sources were not reported for harmonic stimuli, but this result may be obscured by the fact that there seems to be an interaction between spectral content and correlation (discussed below, Figure F.3). This interaction may be due to the fact that an increase in both harmonicity and correlation increases the degree of temporal synchrony among spectral components.

Figure F.1c shows the results for correlation between modulation functions averaged across subjects, replications, spectral shape and spectral content. This was expected to be the most important effect and is certainly the strongest. More sources were reported when the modulation functions were uncorrelated (and asynchronous) than when they were perfectly correlated (i.e. identical). This supports the hypothesis that the auditory system parses the spectrum based on some criterion of commonality (in this case certain subsets of the spectrum are changing coherently in frequency with respect to one another) and then constructs source images based on these subsets. Also of interest is the fact that even more sources were reported on the average when the modulation functions were only partially correlated. This produces a condition of ambiguity for the auditory system in which the source images are less distinguishable. This is supported by the number of "many" responses

reported for these stimuli (to be discussed below).

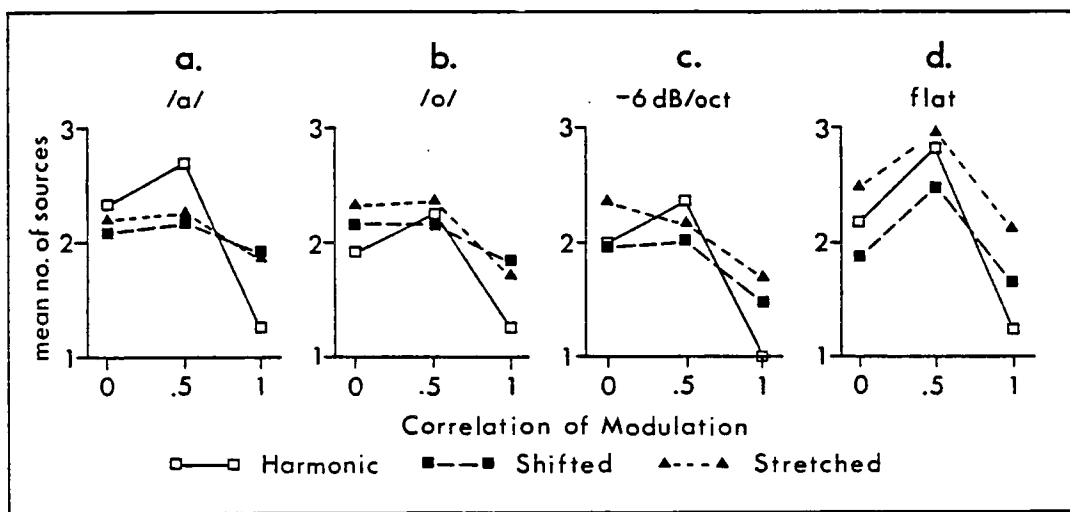
These results should be interpreted with some caution since a combined effect of asynchrony and correlation is likely to be important in the different responses to correlated and partial or uncorrelated stimuli. However, this does not account for the differences (where found) between uncorrelated and partially correlated stimuli, so we may presume that correlation is at least partially contributing to the separation of sources.



**Figure F.2.** Experiment A data summary. The average number of sources reported is plotted to show the interaction between the effects of spectral envelope and correlation for (a) harmonic, (b) shifted harmonic, and (c) stretched harmonic tone complexes.

Figures F.2a-c are plotted in order to show the interaction between spectral shape and correlation of modulation. While the effect is mild, it seems that the familiar envelopes (/a/, /o/, -6 dB/oct) behave similarly with changes in correlation. These conform to the ordering shown in Figure F.1c except for a deviation by the -6 dB/oct slope with stretched partials. The effect is much more pronounced for the unfamiliar

flat spectrum, while the ordering of correlation with respect to number of sources reported remains the same, particularly for the 0.5 correlation. This may imply that more tolerance is allowed in the perception of familiar things, which would have implications for the way we categorize events in the world and for notions concerning the socialization of perception. The culture we live in certainly has its influence in determining the "boundaries" of our perceptual tendencies. But this interpretation also contradicts much psychological research that implies that discrimination of small differences is better for familiar objects than for unfamiliar objects (e.g. "all martians look alike"!).

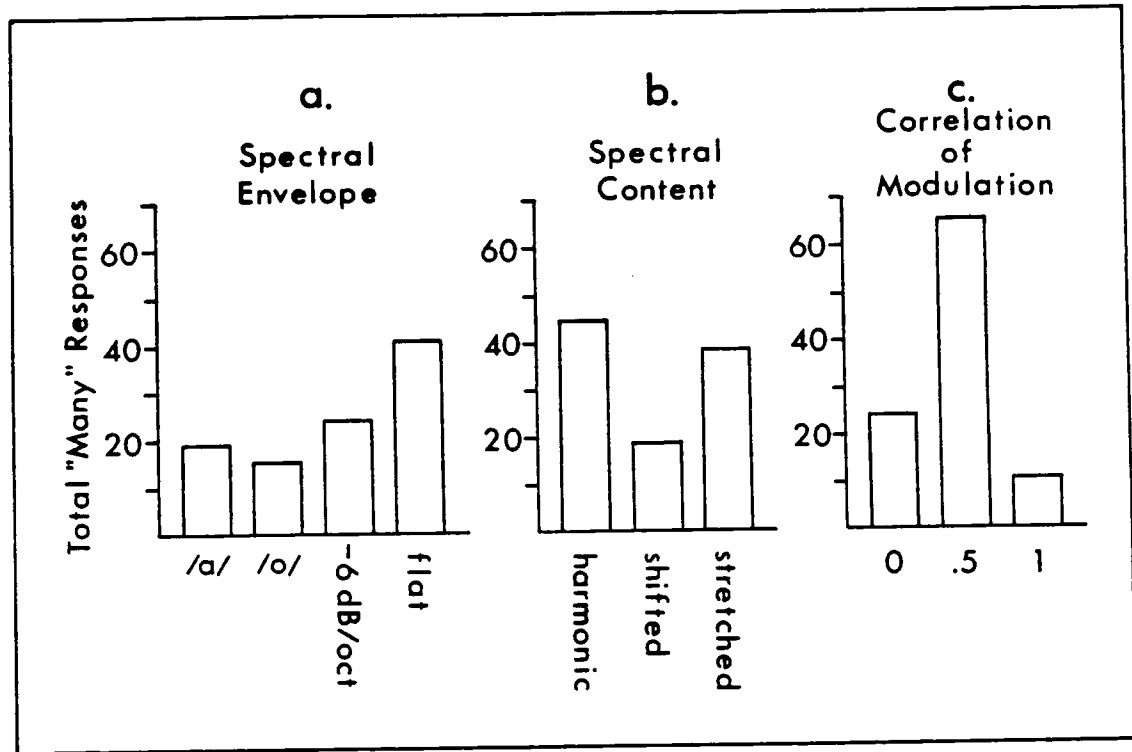


**Figure F.3.** Experiment A data summary. The average number of sources reported is plotted to show the interaction between the effects of spectral content and correlation for the different spectral envelopes: (a) vowel /a/, (b) vowel /o/, (c) -6 dB/oct slope, and (d) flat spectral slope.

Figures F.3a-d are plotted to show the interaction between spectral content and correlation. The curves appear to group into harmonic vs. inharmonic stimuli. The effect of correlation is more exaggerated for harmonic stimuli. This may be related to the regularity of temporal pattern (periodicity) in the signal. In harmonic signals, there is already a higher degree of inherent correlation in the waveform since all of the partials in the complex relate to a common integer divisor, the fundamental. It



remains the case for individual spectral subsets though imperfections in the periodicity are introduced. When the subsets are parsed via independent modulating functions, there is a maintaining of synchrony



**Figure F.4.** Experiment A data summary. Total number of *many* responses collected across subjects and repetitions for (a) spectral envelope, (b) spectral content, and (c) correlation.

(phase-locking) within a given subset while this relation is obscured across subsets. And since each subset remains harmonic, there is a strongly fused and an unequivocal pitch percept. If the degree of fusion of an image is somehow related inversely to the equivocality of the resulting pitch of that image, then one would expect the perception of source images to be more confused for inharmonic tones, even when the partials have identical modulating functions. Parsing by imposing independent modulating functions may just increment the confusion rather than reorder the source interpretation.

Figures F.4a-c show the total number of "many" responses for spectral shape, spectral content and correlation, respectively. In Figure F.4a it is apparent that more ambiguity results from the unfamiliar shape than from the more familiar shapes.

Figure F.4b presents a rather interesting picture. Both harmonic and stretched stimuli are perceived more ambiguously with respect to the number of discernible sources than is the case with shifted stimuli. This result may imply that the degree of spectral proximity of two sources is an important factor in the separation of those sources. We might well predict that the degree to which the excitation patterns from sources overlap in the auditory system affects how much their distinguishability is obscured. Note that with the parsing for the shifted stimuli, there is the least possibility of spectral overlap between the two subsets. This is followed in degree of overlap by the stretched stimuli and then harmonic stimuli (in which each harmonic is surrounded by harmonics from the other subset, except for the fundamental and the highest harmonic). Indeed, the ordering of ambiguity would bear this hypothesis out since harmonic stimuli had the greatest degree of ambiguity followed by stretched and then shifted stimuli.

The data plotted in Figure F.4c strongly support the hypothesis that partial correlation among sources presents an ambiguous situation to the auditory system. The ambiguity present in the condition of zero correlation may be a by-product of the rather arbitrary scheme used to parse the inharmonic spectra.

### F.2.2 *Experiment B*

This experiment investigated the effects of spectral envelope, harmonicity and modulation correlation between separate spectral subgroups on the perceived number of sources in a 16-component stimulus presented in a 2IFC situation.

### F.2.2.1 *Stimuli*

24 tones, each containing 16 partials as before, were generated combining the same four spectral envelopes (/a/, /o/, -6 dB/oct, flat) and 3 three types of spectral content (harmonic, shifted, stretched as in Expt. A). Only two values of correlation among modulation functions were used:  $c = 1$  (correlated) and  $c = 0$  (uncorrelated). Tones were again 1.5 sec in duration with 10 msec raised cosine attack and decay ramps. The previous study varied onset asynchrony and correlation simultaneously and used a rather extreme value of asynchrony shown by many researchers to generate multi-source percepts (Bregman & Pinker, 1978; Dannenbring & Bregman, 1978; Rasch, 1978). Some informal explorations prior to this study suggested that the interaction between onset asynchrony and correlation parameters is at a minimum for asynchrony values between 5-20 msec. So, for this study, a 15 msec asynchrony was used between the onsets of the two groups of partials in each tone regardless of correlation value. In the correlated tones, this is not perceptible as two onsets, but more as a change in quality of the attack when compared with a synchronous onset. All partials decayed simultaneously.

Also, to enhance the effect of correlation differences, greater amplitudes of modulation were used: for periodic frequency modulation, an amplitude yielding a 1.5% excursion in frequency of a partial was used (this is within the range of musical vibrato found for voice, Bjørklund, 1961, and violin, Small, 1936); and for random frequency modulation, an amplitude yielding an excursion of 0.8% was used (this is within the range of 0.4 - 1% random deviation from perfect periodicity in a bowed cello string found by Cardozo & van Noorden, 1968 and in flute, clarinet and trombone found in Appendix B). Additionally, the difference in frequency between the vibrato for uncorrelated tones was increased to 2 Hz. So for uncorrelated tones the group of partials with the earliest onset had a vibrato frequency of 5.5 Hz and the other group had a vibrato frequency of 7.5 Hz. For correlated tones the vibrato frequency was set at 6.5 Hz. These vibrato frequencies are within the range of musical vibrato. The three types of spectral content were parsed as in Experiment A.

Tones were presented in sequential pairs separated by a 0.5 sec silence and followed by an automatic stopping of the recorder and restarting by the subject as in Experiment A. Tone pairs were all possible combinations of correlated (*C*) and uncorrelated (*U*) tones of the 12 spectral-envelope/spectral-content combinations:

*CU, UC, CC, UU*. Note that in *CC* and *UU* pairs, the two tones were not identical. The stimulus parameters were identical, but the jitter would be slightly different in temporal fine structure for each tone. Stimuli were blocked according to spectral content type since exploratory studies suggested that the difference between content types has a tendency to override other effects in large blocks of stimuli, thereby creating a large stimulus uncertainty which fatigues the subject and makes it difficult to attend to finer perceptual details. So it was decided that stimuli would be blocked by content type to reduce the stimulus uncertainty. There were 64 stimuli per block with each of 16 stimuli (4 spectral envelope  $\times$  4 tone pair combinations) being presented four times such that each stimulus was presented in random order before any stimulus was presented a second time, etc. The blocks were presented in random order, one time each, and were separately recorded on tape directly from the digital-to-analog converters of the synthesizer.

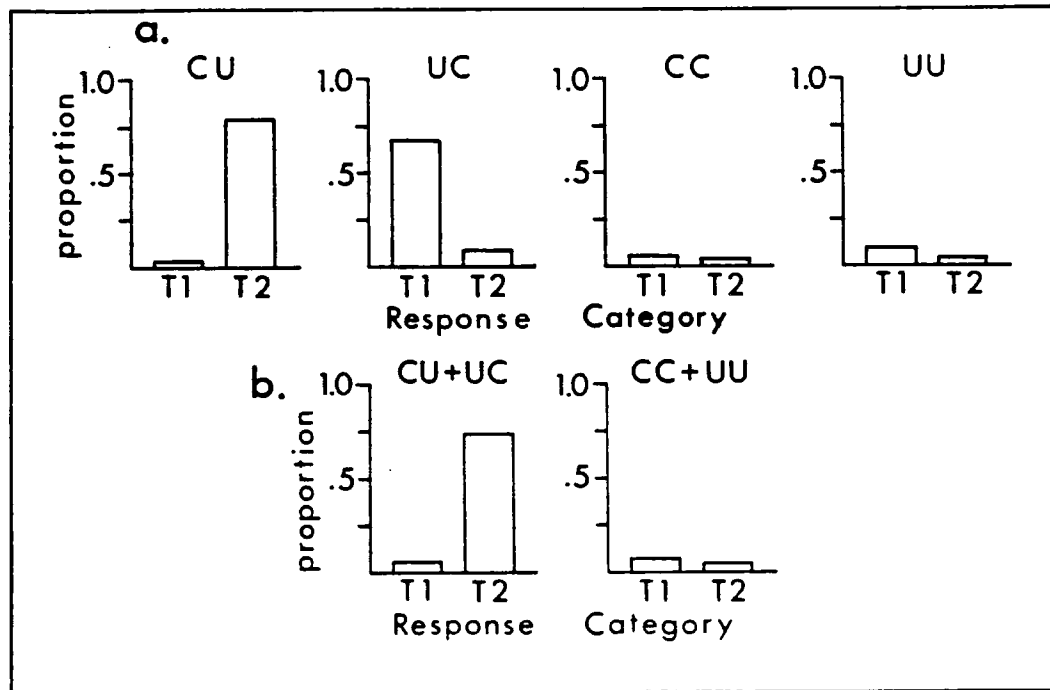
#### F.2.2.2 *Method*

Eight adults of both sexes participated in the experiment and were paid for their services. Four of these had participated in Experiment A and 5 had had extensive experience with computer music. Subjects were instructed to listen to the tone pair and decide which of the two tones might have resulted from more "sources", "voices" or "instruments". They were told there was a 50% probability that the tones were different. If no difference was heard with respect to the number of perceived sources, they were to respond "0"; if a difference was heard, they were to respond with the interval number containing the tone perceived as resulting from more sources.

#### F.2.2.3 *Results and Discussion*

Figure F.5 illustrates the main effect of tone pair combinations. These results are averaged across subjects, replications, spectral content and spectral envelope. It is obvious from Figure F.5a that when the correlation value is changed from 1 to 0 or vice versa across the tone pair, the uncorrelated tone is perceived as having more sources by a margin of at least 69% of the time overall. (In all of the figures that follow, the percentage of "0" responses are not recorded, but can easily be determined in each case.) Note the small percentage of response "reversals" overall, i.e. cases in which a "more sources" responses was reported for a correlated stimulus. Note also

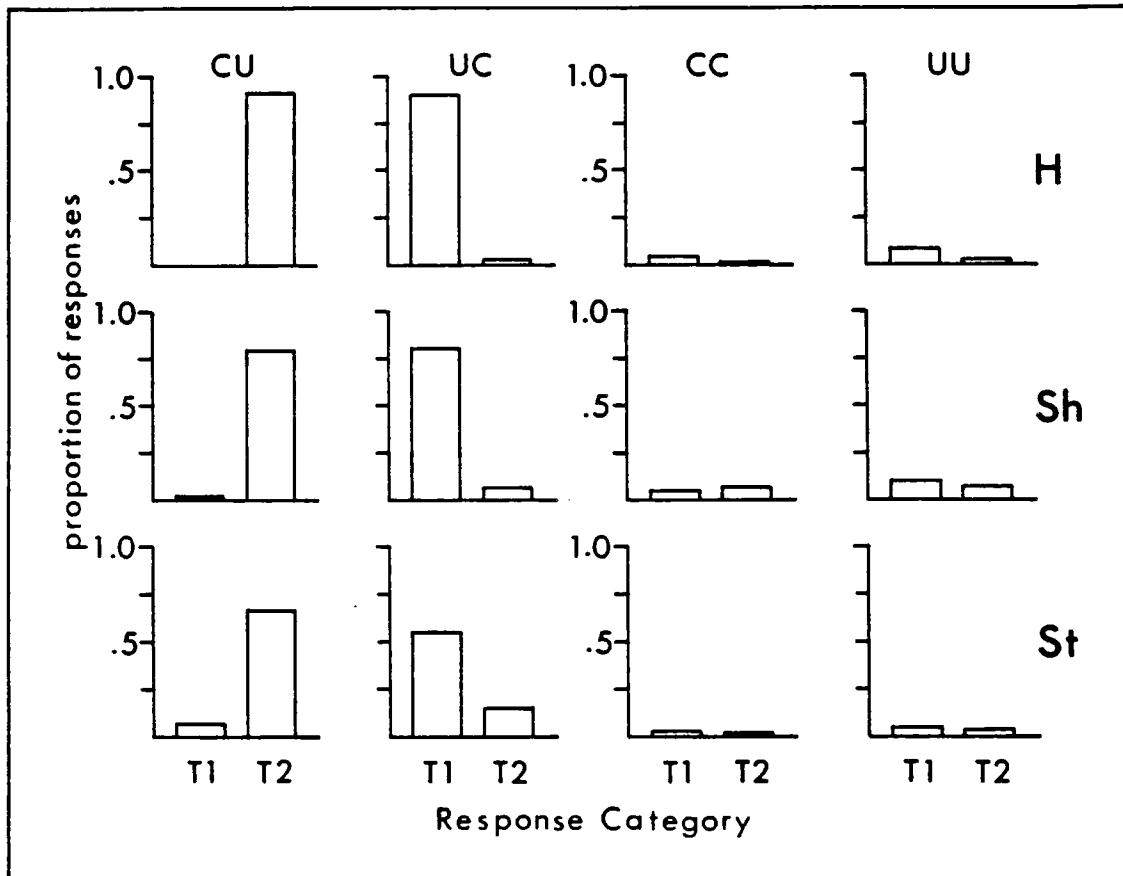
the small percentage (10% overall) of "false alarms", i.e. reports of difference in the number of perceived sources when the parameters of the stimuli did not change.



**Figure F.5.** Experiment B data summary. (a) Proportion responses within response categories averaged across spectral content, spectral envelope and subjects for each of the four tone pair combinations. (*C* = correlated, *U* = uncorrelated.) (b) Proportion responses averaged across change and no-change tone pairs. For *UC*, the value for the first interval was average with the value for the second interval of *CU* and vice versa (see text).

In Figure F.5b, the *CC* and *UU* responses are averaged together for "1" and "2" responses. The "1" values for *UC* are averaged with the "2" values for *CU* and vice versa in order to reflect the overall difference between change and no-change stimuli. These results, though compelling, should be interpreted with caution since there appear to be interaction effects between some of the stimulus dimensions which may be obscuring the magnitude of the correlation effect.

Figure F.6 illustrates the way the responses to correlation value vary as a function of spectral content type. For harmonic stimuli, when the correlation changed subjects reported a change in the number of sources over 90% of the time. Uncorrelated stimuli were always judged as having more sources. This trend holds true for



**Figure F.6.** Experiment B data summary. Proportion responses within response category averaged across spectral envelope and subjects. H = harmonic; Sh = shifted; St = stretched.

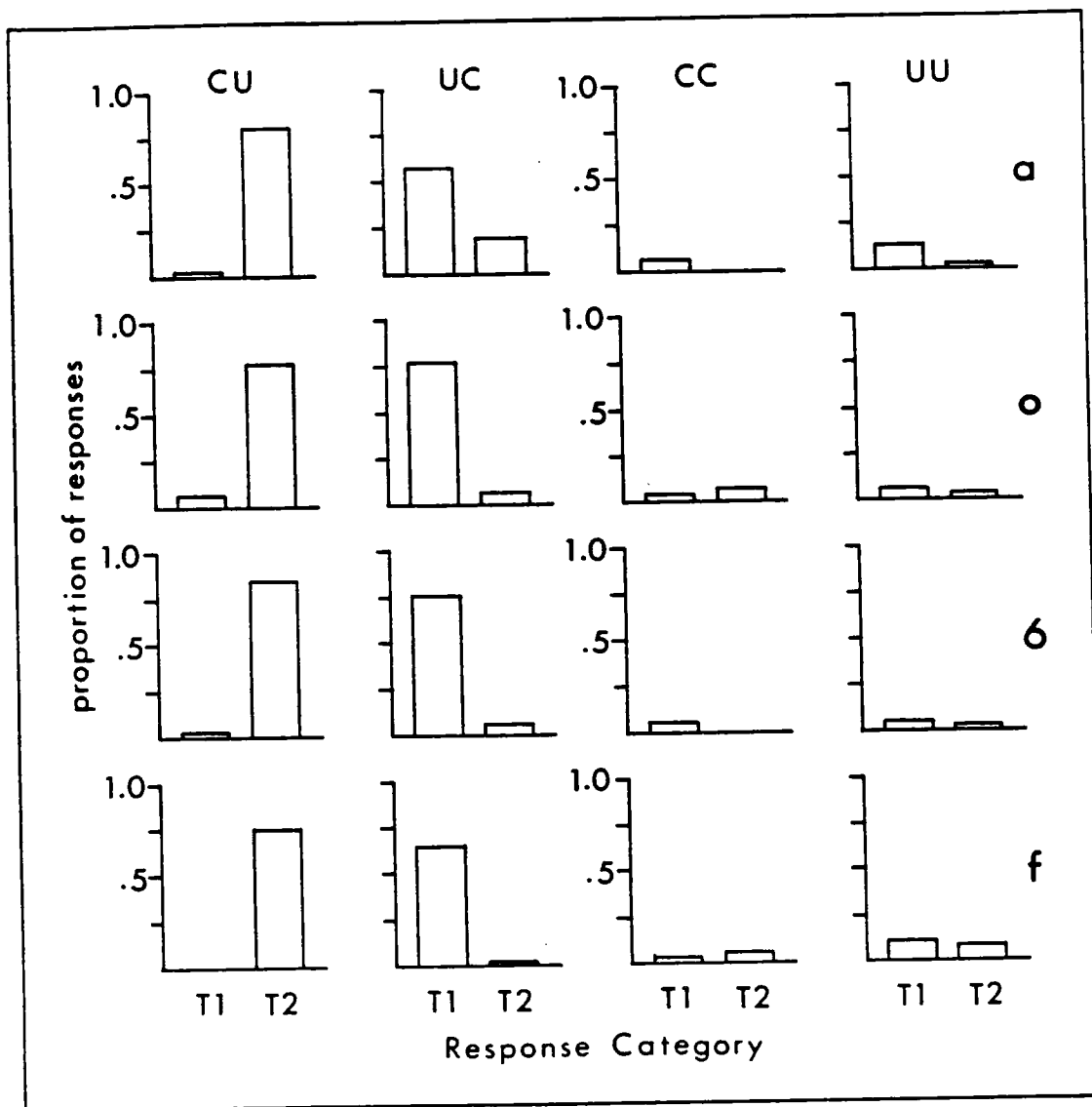
both shifted and stretched stimuli though there is an increase in both the number of "0" responses and in the equivocality of response as indicated by the percentage of response reversals. This is most marked for the stretched stimuli. One possible explanation for this lies in the temporal fine structure of the stimuli. The most

unequivocal periodicity results from harmonic stimuli. For shifted stimuli, the waveform has a period half that of the harmonic signal and has a periodic amplitude envelope with a period equal to that of the harmonic signal. The peaks of the waveform are slightly out of phase with the peaks of the amplitude envelope (see de Boer, 1976, p. 512). There is much less apparent pseudo-periodicity in the stretched stimulus waveform. If we postulate some mechanism looking for (even imperfect) periodicities in order to distinguish sources it would have the easiest time discovering them in harmonic sounds, a somewhat rougher time with shifted stimuli and the most difficult time with stretched signals. But this cannot be the whole explanation because a temporal model would have a difficult time explaining why any consistent results emerge with stretched sounds. So some kind of time-varying spectral analysis must be coming into play. If source distinction is somehow involved with pitch distinction, this might help account for the decreases in distinguishability for more inharmonic sounds which elicit multi-pitched percepts in most cases. This possibility is suggested by the increase in "0" responses and in the increase of response reversals for stretched stimuli.

There is very little systematic effect of spectral envelope (Figure F.7) except for the odd result that there is a strong hysteresis effect for the /a/ envelope. The strength of response is less (and is more equivocal) for the *UC* condition than for the *CU* condition. Examining Figure F.8 reveals that the major part of this effect is due to the inharmonic spectral contents with the /a/ envelope. The response is clear for *CU* presentations and is remarkably unclear with *UC* presentations which yield an increase in no-change responses and substantial response reversals. One may hypothesize that in receiving an uncorrelated stimulus first, the listener is biased toward analytic listening in the subsequent correlated stimulus. Still, the fact that this is found mostly for one spectral envelope is puzzling.

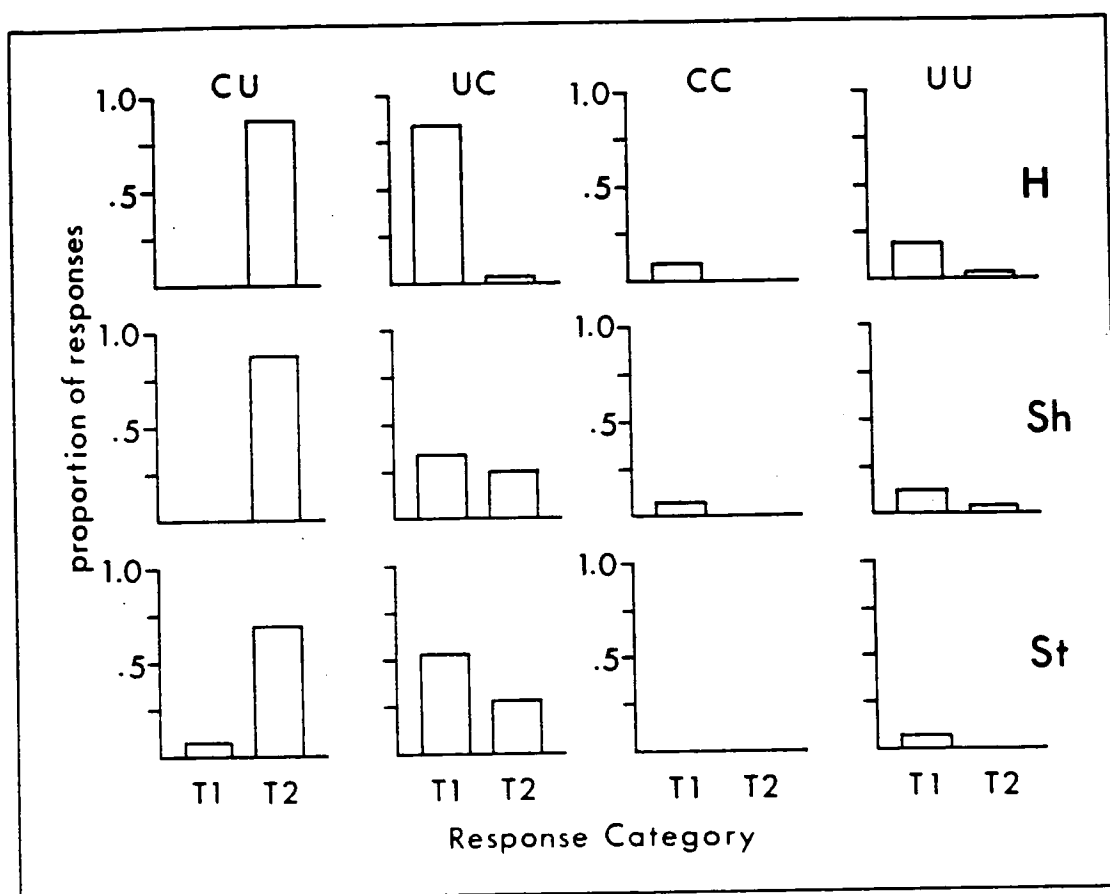
### F.2.3 *Experiment C*

This experiment investigated the effects of spectral envelope, harmonicity and modulation correlation between separate spectral subgroups on the perceived number of pitches in a 16-component stimulus presented in a 2IFC task.



**Figure F.7.** Experiment B data summary. Proportion responses within response categories averaged across spectral content and subjects. **a** = /a/; **o** = /o/; **6** = -6 dB/oct; **f** = flat.



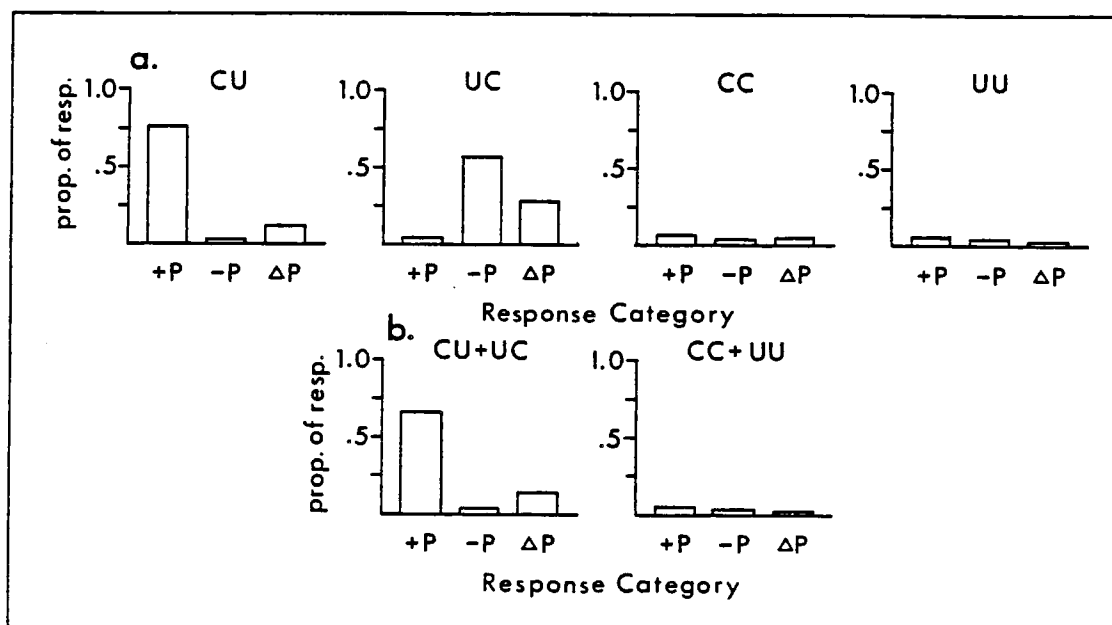


**Figure F.8.** Experiment B data summary. Proportion responses within response categories averaged across subjects for three spectral content types with a vowel /a/ spectral envelope. H = harmonic; Sh = shifted; St = stretched.

#### F.2.3.1 Stimuli and Method

The stimuli were identical to those in Experiment B. Seven male adults who all participated in Experiment B participated in this study and were paid for their services. Both experiments were run in a single 1-hour session. This time subjects were asked to listen to the pitch(es) of the two tones and decide if there were the same pitch(es) in both tones, again with a 50% probability that the tones would be different. In the event they were different, one of three responses was appropriate: *+P* if pitches not present in interval 1 (I1) were present in interval 2 (I2); *-P* if pitches present in I1 were not present in I2; *P* if the number of pitches was the same in both tones but one or more of them had changed to a higher or lower pitch. This is

designed to test the notion that pitch is a derived property of a source and is computed after the available spectrum has been parsed into source spectra. If the number of sources changes, the number of pitches would be expected to change. Or in the case of ambiguous or equivocal pitch percepts as with inharmonic tones, one might expect the change of information being sent to some pitch processor to create a shift in the relative salience of several potential pitches.



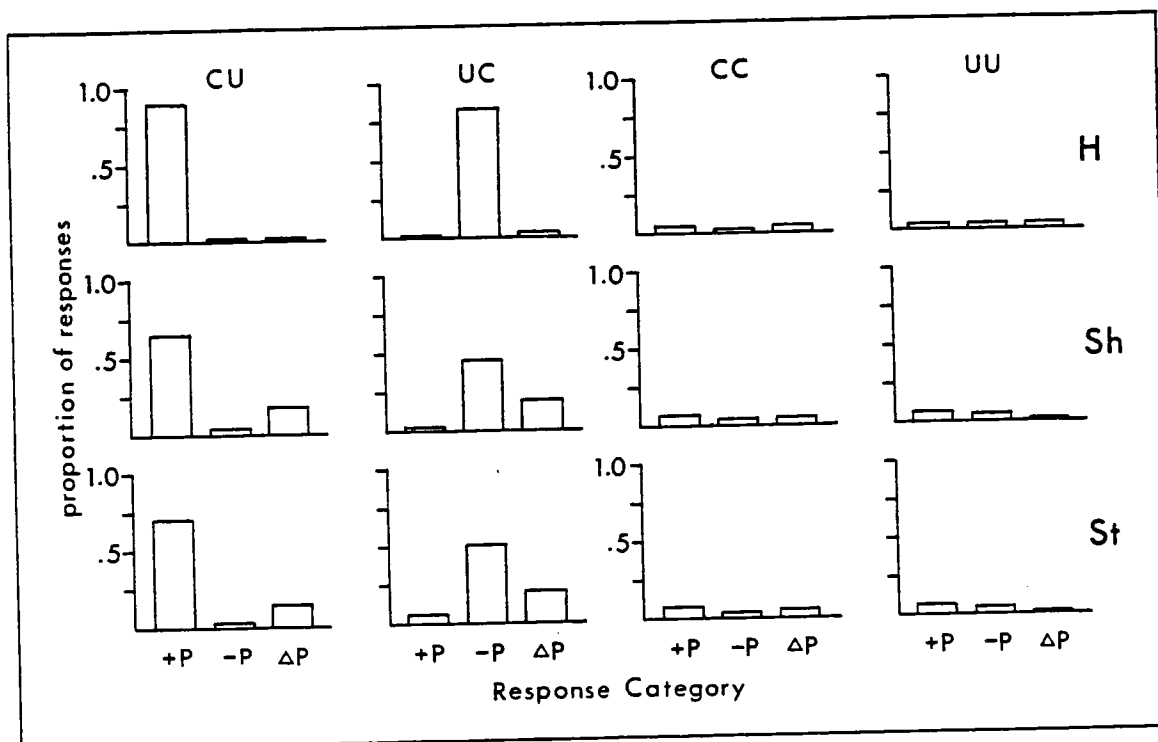
**Figure F.9.** Experiment C data summary. (a) Proportion responses within response categories averaged across spectral content, spectral envelope and subjects for each of the four tone pair combinations. (b) Proportion responses averaged across change and no-change tone pairs. For *UC*, the +*P* value was averaged with -*P* of *CU* and vice versa.

### F.2.3.2 Results and Discussion

Figure F.9 illustrates the main effect of tone pair combination with % responses being averaged over subjects, replications, spectral content and spectral envelope. The main observation of the effect of correlation value corresponds well with the number of sources responses in Figure F.5. Some of the changes are being heard as pitch shifts rather than as new pitches. Again, note the relatively infrequent

occurrence of response reversals and false alarms.

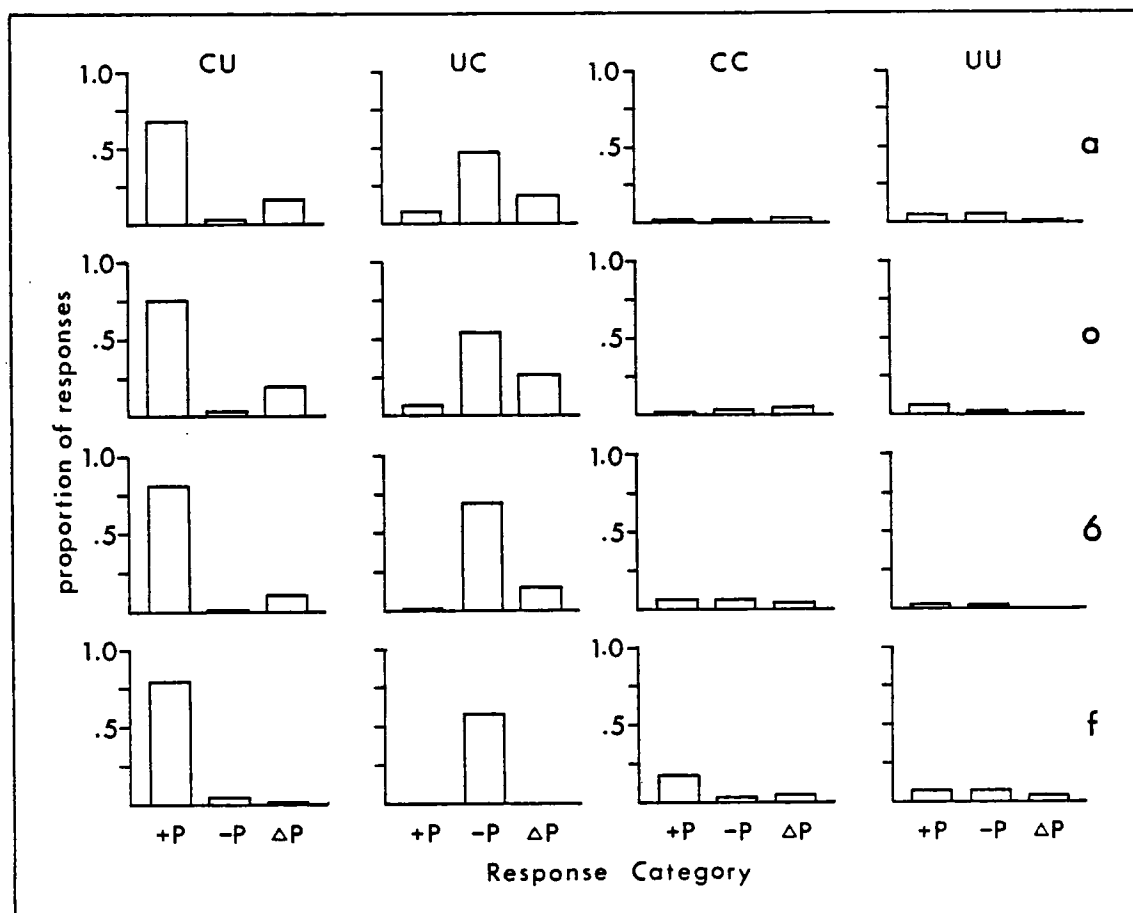
The results in Figure F.10 correspond well with those of Figure F.6. In general, pitch additions and deletions are more readily reported with harmonic signals. With inharmonic signals, a small portion of the responses were heard more as pitch shifts which may be accounted for by the ambiguous nature of inharmonic pitch percepts.



**Figure F.10.** Experiment C data summary. Proportion responses within response categories averaged across spectral envelope and subjects. H = harmonic; Sh = shifted; St = stretched.

The results in Figure F.11 correspond fairly well with those of Figure F.7. There seems to be no main effect of envelope shape. This may well be due to the method of synthesis used for these studies. When a modulation function is imposed on one of these signals, it is not subsequently transformed by the spectral envelope, which serves to smear the shape of the envelope somewhat, particularly at higher partials where the linear frequency excursion is greater for any given partial. In physical

sources such as the voice, the vibrato function would be modified by the resonance cavities thus maintaining the spectral shape to some extent (Lewis, 1936). Given this situation, the hypothesis about the effect of familiarity of the spectral envelope cannot be adequately tested in these experiments.



**Figure F.11.** Experiment C data summary. Proportion responses averaged across spectral content and subjects. a = /a/; o = /o/; 6 = -6 dB/oct; f = flat.

## APPENDIX G

### Description of Taped Examples for Chapter 6<sup>1</sup>

#### **Taped Example 1**

The tones of two familiar nursery rhyme melodies are initially interleaved in the same frequency range, i.e. one tone of Melody 1 is followed by a tone of Melody 2 and then the second tone of Melody 1, etc. Then, in successive cycles of the melodies, one of them is progressively transposed into a higher frequency range until the pitch ranges do not overlap at all. All tones are sinusoidal. Note that once the melodies are perceptually separable it is easy to identify them, a task that is somewhat difficult when they are formed into a single unfamiliar tune. This example is a demonstration of the experiments of Dowling (1973). (See Figure 6.1.)

#### **Taped Example 2**

A cycle of four sinusoidal tones is adjusted in its frequency range and tempo so that it may be heard as a single stream or sequential image. Note also that it is possible, by adjusting one's attentional focus to hear this stimulus as two tone pairs in separate frequency ranges. In the second part of the example, the upper tone pair is enriched timbrally by adding the 3<sup>rd</sup> harmonic. Since these simultaneous components are harmonically related and synchronous they are fused perceptually and a richer timbre results. Note also that the upper pair segregates very easily from the lower pair. This example is taken from McAdams & Bregman (1979). (See Figure 6.3.)

---

1. These sound examples are available from the author by writing to I.R.C.A.M., 31 rue Saint-Merri, F-75004 Paris, France.

**Taped Example 3**

A triplet pattern of ascending perfect fourths is at first played with the same instrument. Then every other note is played by a different instrument (i.e. the X's and O's in Figure 6.4 are played with different timbres). In the second case, the tempo of a stream is half that of the original stream organization and the triplets are descending. This example is taken from Wessel (1979).

**Taped Example 4**

In this example a *klangfarbenmelodie* was composed by David Wessel (1979) to be either continuous with respect to spectral change or to be discontinuous. For the two versions the pitch, rhythm, intensity and spatial location relations are the same; only the timbral relations change. A different instrument is chosen to play each note. In the first part the spectral change is very discontinuous and the melody sounds fragmented. In the second part the spectral change is as continuous as possible while still satisfying the constraint that the instrument change with each note. Note that various rhythmic features emerge which were not heard in the discontinuous version.

**Taped Example 5**

This timbral polyphony was arranged by Marco Stroppa. There are five melodic lines in the structure. One of them is a familiar nursery rhyme melody. These are notated in Figure G.1 and are each played in turn at the beginning of the Taped Example. They are then arranged in a polyphonic structure so that they are completely intertwined with respect to pitch range and (after the addition of a few complementary notes to complete the structure) form four other cross-melodies, each with a constant rhythmic pattern. These lines are notated in Figure G.2 and are each played in turn in the Taped Example. Then the whole structure is played twice with all of the notes having the same timbre (a sinusoidal waveform). Note that the arrangement is adjusted such that these rhythmic streams are organized across the original melodies. After the uni-timbre version is played twice various timbral manipulations are arranged so that the rhythmic lines emerge each with a different timbre. Then the original 5 melodies are made to emerge by changing their timbres and finally the

$\text{♩} = 60$       Original Melodies

The musical notation is presented in two systems, each consisting of five staves. The tempo is indicated as  $\text{♩} = 60$ . The title is "Original Melodies".

The first system (top) shows a melody in the upper staff with eighth notes and triplets (marked with '3' and a bracket). The second staff contains a bass line with 'x' marks. The third staff contains a melody with eighth notes and triplets. The fourth staff contains a bass line with eighth notes and triplets. The fifth staff contains a bass line with eighth notes and triplets.

The second system (bottom) continues the notation, ending with a double bar line and a repeat sign.

**Figure G.1.** Notation of *Original Melodies* for Taped Example 5

single timbre version is replayed to note the difficulty in hearing the embedded

melodies.

$\text{♩} = 60$       **Cross Melodies**

The musical score for 'Cross Melodies' is written for four staves. The tempo is marked as  $\text{♩} = 60$ . The score is divided into two systems, each containing two measures. The notation includes various rhythmic values, including eighth and sixteenth notes, and rests. There are several triplet markings (indicated by a '3' over a bracket) and a 'sim' (simultaneous) marking. The score uses a key signature of one flat (B-flat) and a common time signature (C). The notation is dense, with many notes and rests, and includes some 'x' marks, possibly indicating specific performance techniques or editing points.

**Figure G.2.** Notation of *Cross Melodies* for Taped Example 5


Order of events in Taped Example 5.


Original Melodies


OM 1 (popular tune  $\text{♩}$  )

OM 2 (descending chromatic scale  $\text{♩}$  )



OM 3 (diminished seventh arpeggios )

OM 4 (drone on F# )

OM 5 (12-tone series from Webern's String Quartet opus 28 )


Cross Melodies

CM 1 (  )

CM 2 (  )

CM 3 (  )

CM 4 (  )

Whole structure with complementary notes (  ) and same timbre on all notes: played twice.

Timbre change on CM1, on CM2, on CM3, on CM4, on all CM simultaneously.

Timbre change on OM1, on OM2, on OM3, on OM4, on OM5, on all OM simultaneously.

Whole structure with single timbre: played twice.

### **Taped Example 6**

To illustrate the effect of frequency modulation on spectral fusion, a vowel-like sound is initially presented without modulation. Note the unnatural quality and the ability to hear out the components of the complex tone. Then a vibrato is slowly added to the tone causing it to fuse, and then the vibrato is removed again to allow the possibility of analytic listening again. Finally the vibrato is brought back while the spectral form changes to another vowel. This example is taken from McAdams & Wessel (1981).

### **Taped Example 7**

The two series of stimuli from Chapter 2 are presented in this example. In both series the tones start with no modulation and then succeeding tones are played with 7, 14, 28, 42 and then 56 cents modulation width. In the first series the tones with constant frequency ratio maintaining modulation are played. In the second series the tones with a constant frequency difference maintaining modulation are played. (See Figure 2.1.)

**Taped Example 8**

An exaggerated example of the type of stimulus presented in Chapter 3 is played in this example. A continuous, 16-component, flat spectrum tone is played. A rather wide vibrato is introduced to each of the partials one at a time starting with the fundamental and moving up to the 16<sup>th</sup>. Unlike the stimuli in Experiment 6, none of the other partials are moving in this case. Note that even the pitch of the highest partials can be heard when a wide enough frequency modulation is introduced.

**Taped Example 9**

All 6 of the chords with 3 vowels at three pitches are played in succession as notated in Chapter 5, section 5.2.2.1. No modulation is present in any of these stimuli. Try to listen for the vowels /a/, /o/ and /i/ in each tone. Pretty difficult, eh? (See Figure 6.5)

**Taped Example 10**

In this example each of the vowels /a/, /o/, /i/ are played at each of the 3 pitches. First they are played without frequency modulation and then with modulation. These stimuli are exactly the ones mixed into the chords. Note that they are easily identifiable in isolation. Note however the improvement of the quality of the vowel, particularly at the highest pitch, when modulation is introduced.

**Taped Example 11**

In this example all of the vowel chords with one vowel modulating and two vowels steady are presented. They are ordered such that the vowel in question is played twice at each pitch, starting from the lowest. The difference between two stimuli with the modulated vowel at the same pitch is the arrangement of the other two vowels at the other two pitches. For example, /a/ at  $C_3$  is presented once with /o/ and /i/ at  $F_3$  and  $Bb_3$ , respectively, and once in the reverse order. The six permutations are presented once for each vowel being modulated, i.e. /a/, /o/, /i/ in turn. Note the increased ease of hearing the vowels when they are modulated. Note also the difference in perceived prominence of the three target vowels.

**Taped Example 12**

A series of vowel chords is now presented in which all three vowels are modulated perfectly coherently. The order of the stimuli is the same as in Taped Example 9. (See Figure 6.6)

**Taped Example 13**

A fragment from *Casta Diva*, composed by Alain Louvier and realized by Andy Moorer at IRCAM. In this example, linear predictive coding and resynthesis is used initially to reproduce the original speaking voice and then to transform the nature of the excitation source, leaving the behavior of the vocal tract intact. Note that at several points it is possible to hear the excitation as being multi-sourced and yet it is also possible to understand the speech signal (if you understand French, that is).

**Taped Example 14**

The basis of this example is a voice-like sound which is constantly changing in spectral structure, though very slowly, between /a/, /o/ and /i/. In each of the three presentations a different amount of random amplitude modulation is introduced to the partials. In all cases, the modulation waveforms are independent for each partial. Note that with increasing modulation depth, the apparent number of contributing sources increases. (See Figure 6.7)

**Taped Example 15**

This example is taken from a piece by Roger Reynolds for orchestra and computer-generated tape called *Archipelago* realized at IRCAM with the musical assistance of Thierry Lancino. The original sound, an unassuming oboe note, is played first. A phase vocoder analysis was performed and then the tone was resynthesized with the odd harmonics in one channel and the even harmonics in the other channel. It is important that the channels be appropriately balanced for this example to work. At the outset, there is an identical frequency modulation function on the partials of the two channels and a fused oboe image between the two speakers results. As the tone progresses, the modulation functions in the two channels are decorrelated and the image slowly splits into two images with different pitches and different timbres. (See Figure 6.8)

**Taped Example 16**

This example was composed by Jean-Baptiste Barrière with the CHANT synthesis program. At the outset, the changing vocal formants remained fused in rather grumbly chanting voices. Near the end, the formants begin to move around rapidly and the voice images defuse into whistling formants even though the excitation is still harmonic.

**Taped Example 17**

This example illustrates the experiments of Bregman & Pinker (1978). The first part presents a situation where tones *A* and *B* are close in frequency and tones *B* and *C* are asynchronous. There is a tendency to hear the sequential stream *AB* and to hear tone *C* as separate and pure. The second part presents a situation where tones *A* and *B* are greatly separated in frequency and tones *B* and *C* onset synchronously. There is a tendency to hear *A* in a segregated stream and to hear *C* as being rich, i.e. *B* is fused with *C* and difficult to hear on its own. (See Figure 1.1)

**Taped Example 18**

The last example illustrates the possibilities of spectral and temporal fusion with the means of modern music synthesis. This fragment was composed by Xavier Rodet with the CHANT program. I leave your ears to describe it to you.

## REFERENCES

- Aitken, L.M. & Webster, W.R.** (1971) Tonotopic organization in the medial geniculate body of the cat, *Brain Res.*, **26**:402-405.
- Baker, J.M.** (1975) A new time-domain analysis of human speech and other complex waveforms, PhD. dissertation, Carnegie-Mellon University, Pittsburgh, Penn.
- Balzano, G.J.** (1983) Changing conceptions of pitch and timbre: A modest proposal, *J. Acous. Soc. Am.*, **74** (S1):S18 (A).
- Barrière, J.-B.** (1983) *Chréode* for computer-generated tape (IRCAM, Paris).
- von Békésy, G.** (1960) *Experiments in Hearing*, McGraw-Hill: New York.
- von Békésy, G.** (1963) Three experiments concerned with pitch perception, *J. Acous. Soc. Am.*, **35**:602-606.
- Bennett, G.** (1981) Singing synthesis in electronic music, *in: Research Aspects on Singing: Autoperception, Computer Synthesis, Health, Voice Source*, publ. no. 33, Royal Swedish Academy of Music, pp. 34-50.
- Berger, K.W.** (1964) Some factors in the recognition of timbre, *J. Acous. Soc. Am.*, **36**:1888-1891.
- von Bismarck, G.** (1974) Sharpness as an attribute of the timbre of steady sounds, *Acustica*, **30**:159-172.
- Björklund, A.** (1961) Analysis of soprano voices, *J. Acous. Soc. Am.*, **33**:575-582.
- Blauert, J.** (1981) Lateralization of jittered tones, *J. Acous. Soc. Am.*, **70**:694-698.
- de Boer, E.** (1956) Pitch of inharmonic signals, *Nature (London)*, **178**:535-536.
- de Boer, E.** (1976) On the "residue" and auditory pitch perception, *in: W.D. Keidel & W.D. Neff (eds.), Handbook of Sensory Physiology*, vol. V/3, Springer-Verlag:Vienna, pp. 479-583.
- Boudreau, J.C. & Tsuchitani, C.** (1970) Cat superior olive S-segment cell discharge to tonal stimuli, *in: W.D. Neff (ed.), Contributions to Sensory Physiology*, vol. 4, Academic:New York, pp. 143-213.

- Bozzi, P. and Vicario, G., (1960) Due fattori di unificazione fra note musicali: La vicinanza temporale e la vicinanza tonale, *Rivista de Psicologia*, 54:235-258, cited in Vicario (1982).
- Bregman, A.S., (1977) Perception and behavior as compositions of ideals, *Cog. Psych.* 9:250-292.
- Bregman, A.S., (1978a) Asking the 'what for' question in auditory perception, *in*: Kubovy, M. and Pomerantz, J. (eds.), *Perceptual Organization*, Lawrence Erlbaum:Hillsdale, New Jersey.
- Bregman, A.S., (1978b) The formation of auditory streams, *in*: Requin, J. (ed.), *Attention and Performance VII*, Lawrence Erlbaum:Hillsdale, New Jersey.
- Bregman, A.S., (1980) The conceptual basis of perception and action, *in*: *Perception and Cognition II: Presentations on Art Education Research*, vol. 4, Concordia Univ.:Montreal.
- Bregman, A.S., (1982) Two-factor theory of auditory organization, *J. Acous. Soc. Am.* 72:S10(A).
- Bregman, A.S., Abramson, J. & Darwin, C., (1983) Spectral integration based on common amplitude modulation, unpublished manuscript, McGill Univ., Montreal, Canada, [reported in *J. Acous. Soc. Am.*, 74 (S1):S9 (A)].
- Bregman, A.S. & Campbell, J., (1971) Primary auditory stream segregation and the perception of order in rapid sequences of tones, *J. Exp. Psych.*, 89:244-249.
- Bregman, A.S., McAdams, S. & Halpern, L. (1978) Auditory segregation and timbre, presented at the meeting of the Psychonomic Society, Nov. 1978, San Antonio, Texas.
- Bregman, A.S. & Mills, M.I. (1982) Perceived movement: the Flintstone constraint, *Perception*, 11:201-206.
- Bregman, A.S. & Pinker, S. (1978) Auditory streaming and the building of timbre, *Can. J. Psych.*, 32:19-31.
- Bregman, A.S. & Tougas, Y., (1979) Propagation of constraints in auditory organization, unpublished manuscript, McGill Univ., Montreal.
- Broadbent, D.E. (1955) A note on binaural fusion, *Quart. J. Exp. Psych.*, 7:46-47.
- Broadbent, D.E. & Ladefoged, P. (1957) On the fusion of sounds reaching different sense organs, *J. Acous. Soc. Am.*, 29:708-710.
- Brokx, J.P.L. & Nootboom, S.G. (1982) Intonation and the perceptual separation of simultaneous voices, *J. Phonetics*, 10:23-36.
- Brugge, J.F., Anderson, D.J., Hind, J.E. & Rose, J.E. (1969) Time structure of discharges in single auditory-nerve fibers of the squirrel monkey in response to

- complex periodic sounds, *J. Neurophysiol.*, **32**:386-401.
- Cardozo, B.L. & Neelen, J.J.M.** (1968) Audibility of jitter in pulse trains as affected by filtering. *IPO Prog. Rep.*, **3**:13-15, Institute for Perception Research, Eindhoven, The Netherlands.
- Cardozo, B.L. & van Noorden, L.P.A.S.** (1968) Imperfect periodicity in the bowed string, *IPO Prog. Rep.* **3**:23-28, Institute for Perception, IPO, Eindhoven, The Netherlands.
- Carlson, R., Fant, G. & Granström, B.** (1975) Two-formant models, pitch and vowel perception, in: G. Fant & M.A.A. Tatham (eds.), *Auditory Analysis and Speech Perception*. Academic: London, pp. 55-82.
- Charbonneau, G.R.** (1981) Timbre and the perceptual effects of three types of data reduction, *Comp. Mus. J.*, **5**(2):10-19.
- Cherry, E.C.** (1953) Some experiments on the recognition of speech with one and with two ears, *J. Acous. Soc. Am.*, **25**:975-979.
- Chowning, J.M.** (1973) The synthesis of complex audio spectra by means of frequency modulation, *J. Audio Eng. Soc.*, **21**:526-534.
- Chowning, J.M.** (1980) Computer synthesis of the singing voice, in: *Sound Generation in Winds, Strings, Computers*. Royal Swedish Academy of Music, publ. no. 29, Kungl. Musikaliska Akademien, Stockholm.
- Chowning, J.M.** (1982) The synthesis of sung vowel tones, in: *Numero e Suono* (Proc. 1982 Int'l. Comp. Mus. Conf., Venice, Italy, September, 1982), p. 250.
- Cohen, E.** (1979) Fusion and consonance relations for tones with inharmonic partials. *J. Acous. Soc. Am.*, **65**:S123 (A).
- Cohen, E.** (1980) The influence of non-harmonic partials on tone perception, Ph.D. dissertation, Stanford University, Stanford, California.
- Craik, K.J.W.** (1943) *The Nature of Explanation*, Cambridge University:Cambridge.
- Cutting, J.E.** (1976) Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening, *Psych. Rev.*, **83**:114-140.
- Dannenbring, G.L. and Bregman, A.S.** (1976) Stream segregation and the illusion of overlap, *J. Exp. Psych./Human Perc. Perf.*, **2**:544-555.
- Darwin, C.J.** (1981) Perceptual grouping of speech components differing in fundamental frequency and onset-time, *Quart. J. Exp. Psych.*, **33A**:185-207.
- Darwin, C.J.** (1983) Auditory processing and speech perception, in: H. Bouma & H. Bouwhuis (eds.) *Attention and Performance X*, Erlbaum:Longdale, N.J., [cited in Hall, Haggard & Fernandes (1984)].
- Delgutte, B.** (1980) Representation of speech-like sounds in the discharge patterns of

- auditory-nerve fibers, *J. Acous. Soc. Am.*, **68**:843-857.
- Deutsch, D.**, (1975) Two-channel listening to musical scales, *J. Acous. Soc. Am.*, **57**:1156-1160.
- Dolson, M.** (1983) Modification of musical sounds by means of the phase vocoder, *J. Acous. Soc. Am.*, **74** (S1):S19 (A).
- Donders, F.C.** (1864) Zur Klangfarbe der Vocale, *Ann. Phys. Chem.*, **123**:527-528, [cited in Plomp (1970)].
- Dowling, W.J.**, (1973) The perception of interleaved melodies, *Cog. Psych.*, **5**:322-337.
- Duncan, R.** (1960) *The Opening of the Field*, New Directions:New York.
- Durlach, N.I.** (1972) Binaural signal detection: Equalization and cancellation theory, in: Tobias, J.V. (ed.) *Foundations of Modern Auditory Theory*, vol. II, Academic:New York.
- Ehresman, D.** (1977) A parallelogram model of timbral analogies, M.A. thesis, Michigan State University, East Lansing, Michigan.
- Ehresman, D. & Wessel, D.** (1978) Perception of timbral analogies, *Rapport IRCAM*, no. 13, Paris, France.
- Evans, E.F.** (1970) Narrow tuning of cochlear nerve fibers, *J. Physiol. (London)*, **206**:14P,15P.
- Evans, E.F.** (1971) Central mechanisms relevant to the neural analysis of simple and complex sounds, in: R. Klinke & O.J. Grüsser (eds.), *Pattern Recognition in Biological and Technical Systems*, Springer-Verlag:Berlin, pp. 328-343.
- Evans, E.F.** (1975) Cochlear nerve and cochlear nucleus, in: W.D. Keidel & W.D. Neff (eds.) *Handbook of Sensory Physiology*, vol. V/3, Springer-Verlag:Berlin.
- Evans, E.F.** (1978) Place and time coding of frequency in the peripheral auditory system: Some physiological pros and cons, *Audiol.*, **17**:369-420.
- Fairbanks, G. & Grubb, P.** (1961) A psychophysical investigation of vowel formants, *J. Speech Hearing Res.*, **4**:203-219.
- Fant, G.** (1959) Acoustic analysis and synthesis of speech with applications to Swedish, *Ericsson Tech.*, **15**:3-108.
- Fant, G.** (1971) Distinctive features and phonetic dimensions, in: G.E. Perron & J.L.M. Trim (eds.), *Applications of Linguistics. Selected Papers of the 2nd Int'l Cong. of Applied Linguistics, Cambridge 1969*, Cambridge Univ.:Cambridge, pp. 219-239.
- Fant, G.** (1973) *Speech sounds and features*, MIT Press:Cambridge, Mass.
- Flanagan, J.L.** (1972) *Speech Analysis, Synthesis and Perception*, Springer-Verlag:Berlin.



- Fletcher, H., Blackham, E.D. & Geersten, O.N.** (1965) Quality of violin, viola, 'cello, and bass-viol tones. I, *J. Acous. Soc. Am.*, **37**:851-863.
- Fletcher, H. & Sanders, L.** (1967) Quality of violin vibrato tones, *J. Acous. Soc. Am.*, **41**:1534-1544.
- Gibson, E.J.** (1969) *Principles of Perceptual Learning and Development*, Appleton-Century-Crofts:New York.
- Gibson, J.J.** (1966) *The Senses Considered as Perceptual Systems*, Houghton:Boston.
- Gibson, J.J.** (1974) A note on ecological optics, in: E.C. Carterette & M.P. Friedman (eds.) *Handbook of Perception, vol. I: Historical and Philosophical Roots of Perception*, Academic:New York.
- Gibson, J.J. & Gibson, E.J.** (1955) Perceptual learning: differentiation or enrichment? *Psych. Rev.*, **62**:32-41.
- Goldstein, J.L.** (1966) An investigation of monaural phase perception, PhD. dissertation, University of Rochester, University Microfilms (66-6852): Ann Arbor, Michigan [cited in Schubert & Nixon (1970)].
- Goldstein, J.L.** (1973) An optimum processor theory for the central formation of the pitch of complex tones, *J. Acous. Soc. Am.*, **54**:1496-1516.
- Goldstein, J.L.** (1978) Mechanisms of signal analysis and pattern perception in periodicity pitch, *Audiol.*, **17**:421-445.
- Grassman, H.** (1877) Über die physikalische Natur der Sprachlaute, *Ann. Phys. Chem.*, **1**:606-629, [cited in Plomp (1970)].
- Green, D.M.** (1971) Temporal auditory acuity, *Psychol. Rev.*, **78**:540-551.
- Green, D.M. & Kidd, G.** (1983) Further studies of auditory profile analysis, *J. Acous. Soc. Am.*, **73**:1261-1265.
- Green, D.M., Kidd, G. & Mason, C.R.** (1983) Profile analysis and the critical band, *J. Acous. Soc. Am.*, **73**:S92(A).
- Green, D.M., Kidd, G. & Picardi, M.C.** (1983) Successive vs. simultaneous comparisons in auditory intensity discrimination, *J. Acous. Soc. Am.*, **73**:639-643.
- Green, D.M. & Mason, C.R.** (1983) Phase effects and profile analysis, *J. Acous. Soc. Am.*, **74**:S71(A).
- Gregory, R.L.** (1974) Choosing a paradigm for perception, in: E.C. Carterette & M.P. Friedman (eds.) *Handbook of Perception, vol. I: Historical and Philosophical Roots of Perception*, Academic:New York.
- Gregory, R.L.** (1981) *Mind in Science: A History of Explanations in Psychology and Physics*, Cambridge University:Cambridge.
- Grey, J.M.** (1975) An exploration of musical timbre, PhD. dissertation, Stanford

- University, publ. as *Dept. of Music Technical Report* STAN-M-2.
- Grey, J.M.** (1977) Multidimensional perceptual scaling of musical timbres, *J. Acous. Soc. Am.*, **61**:1270-1277.
- Grey, J.M. & Gordon, J.G.** (1978) Perceptual effects of spectral modifications on musical timbres, *J. Acous. Soc. Am.*, **63**:1493-1500.
- Grey, J.M. & Moorer, J.A.** (1977) Perceptual evaluations of synthesized musical instrument tones, *J. Acous. Soc. Am.*, **62**:454-462.
- Groen, J.J. & Versteegh, R.M.** (1957) Frequency modulation and the human ear, *Acta Otolaryngol.*, **47**:421-430.
- Guinan, J.J., Norris, B.F. & Guinan, S.S.** (1972) Single auditory units in the superior olivary complex. II: Locations of unit categories and tonotopic organization, *Int'l. J. Neurosci.*, **4**:147-166.
- Hall, D.E.** (1980) *Musical Acoustics: An Introduction*, Wadworth:Belmont, Calif.
- Hall, J.W. & Fernandes, M.A.** (1983) Temporal integration, frequency resolution and off-frequency listening in normal-hearing and cochlear-impaired listeners, *J. Acous. Soc. Am.*, **74**:1172-1175.
- Hall, J.W., Haggard, M.P. & Fernandes, M.A.** (1983) Detection in noise by spectro-temporal pattern analysis, unpublished manuscript, Univ. of Nottingham, U.K.
- Hartmann, W.M.** (1983) Personal communication: conversations December 1983.
- Hartmann, W.M. & Klein, M.A.** (1980) Theory of frequency modulation detection for low modulation frequencies, *J. Acous. Soc. Am.*, **67**:935-946.
- von Helmholtz, H.L.F.** (1859) Über die Klangfarbe der Vocale, *Ann. Phys. Chem.*, **18**:280-290, [cited in Plomp (1970)].
- von Helmholtz, H.L.F.** (1867) *Treatise on Physiological Optics*, vol. III, section 26: Concerning the perceptions in general, trans. from the 3<sup>rd</sup> German ed., Southall:New York, pp. 1 - 37.
- von Helmholtz, H.L.F.** (1877/1885) *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, republ. 1954, from 1885 edition of the English translation by A.J. Ellis, Dover:New York.
- Heyser, R.C.** (1973a) The delay plane, objective analysis of subjective properties: Part I, *J. Audio Eng. Soc.*, **21**:690-701.
- Heyser, R.C.** (1973b) The delay plane, objective analysis of subjective properties: Part II, *J. Audio Eng. Soc.*, **21**:786-791.
- Heyser, R.C.** (1974) Geometrical considerations of subjective audio, *J. Audio Eng. Soc.*, **22**:674-682.
- Heyser, R.C.** (1976a) Perspectives in audio analysis: Changing the frame of reference,

- Part I, *J. Audio Eng. Soc.*, **24**:660-667.
- Heyser, R.C.** (1976b) Perspectives in audio analysis: Changing the frame of reference, Part II, *J. Audio Eng. Soc.*, **24**:742-751.
- Hiller, & Ruiz,** (1971) *J. Audio Eng. Soc.*, **19**:462-470.
- Hind, J.E.** (1952) An electrophysiological determination of tonotopic organisation in auditory cortex of cat, *J. Neurophysiol.*, **16**:475-489.
- Hind, J.E., Anderson, D.J., Brugge, J.F. & Rose, J.E.** (1967) Coding of information pertaining to paired low-frequency tones in single auditory-nerve fibers of the squirrel monkey, *J. Neurophysiol.*, **30**:794-816.
- Hoffman, D.D.** (1983) The interpretation of visual illusions, *Sci. Am.*, **249** (6):137-145.
- Houtsma, A.J.M.** (1983) Perception of harmonic intervals made by simultaneous complex tones, *J. Acous. Soc. Am.*, **73**:S77(A), and unpublished manuscript from A.S.A. talk.
- Houtsma, A.J.M. & Goldstein, J.L.** (1972) The central origin of the pitch of complex tones: Evidence from musical interval recognition, *J. Acous. Soc. Am.*, **51**:520-529.
- Huggins, W.H.** (1952) A phase principle for complex-frequency analysis and its implications in auditory theory, *J. Acous. Soc. Am.*, **24**:582-589.
- Huggins, W.H.** (1953) A theory of hearing, in: W. Jackson (ed.) *Communication Theory*, Butterworths:London, pp. 303-379 [cited in Schubert & Nixon (1970)].
- IRCAM** (1983) *IRCAM: Un Portrait*, recorded disk with accompanying text, IRCAM, Paris.
- Isenberg, D. & Liberman, A.M.** (1979) The use of duplex perception to study silence as a cue for stop consonants, *J. Acous. Soc. Am.*, **65**:S79(A).
- Jeffress, L.A.** (1972) Binaural signal detection: Vector theory, in: Tobias, J.V. (ed.) *Foundations of Modern Auditory Theory*, vol. II, Academic:New York.
- Jesteadt, W. & Sims, S.L.** (1975) Decision processes in frequency discrimination, *J. Acous. Soc. Am.*, **57**:1161-1168.
- Johnstone, B.M. & Boyle, A.J.T.** (1967) Basilar membrane vibration examined with the Mössbauer technique, *Science*, **158**:390-398.
- Julesz, B.** (1971) *Foundations of Cyclopean Perception*, Univ. of Chicago:Chicago.
- Julesz, B. & Hirsh, I.J.** (1972) Visual and auditory perception — An essay of comparison, in: E.E. David & P.B. Denes (eds.) *Human Communication: A Unified View*, McGraw-Hill:New York, pp. 283-340.
- Karnickaya, E.G., Mushnikov, V.N., Slepokurova, N.A. & Zhukov, S.J.** (1975) Auditory processing of steady-state vowels, in: G. Fant & M.A.A. Tatham (eds.) *Auditory Analysis and Speech Perception*. Academic:London pp. 37-53.

- Katsuki, Y.** (1961) Neural mechanisms of auditory sensation in cats, *in*: W.A. Rosenblith (ed.) *Sensory Communication*, Wiley:New York.
- Keidel, W.D.** (1974) Information processing in the higher parts of the auditory pathway, *in*: E. Zwicker & E. Terhardt (eds.) *Facts and Models in Hearing*, Springer-Verlag:Berlin.
- Kersta, L.G., Bricker, P.D. & David, E.E.** (1960) Human or machine? A study of voice naturalness, *J. Acous. Soc. Am.*, **32**:1502(A).
- Khanna, S.M. & Leonard, D.G.B.** (1982) Basilar membrane tuning in the cat cochlea, *Science*, **215**:305-306.
- Kiang, N.Y.S.** (1965) *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve*, Res. Monogr. 35, M.I.T.:Cambridge, Mass.
- Kiang, N.Y.S., Morest, D.K., Godfrey, D.A., Guinan, J.J. & Kane, E.C.** (1973) Stimulus coding at caudal levels of the cat's auditory nervous system: I. Response characteristics of single units, *in*: A. Møller (ed.) *Basic Mechanisms in Hearing*, Academic:New York, pp. 455-478.
- Klatt, D.H.** (1980) Speech perception: a model of acoustic-phonetic analysis and lexical access, *in*: R. Cole (ed.) *Perception and Production of Fluent Speech*, Erlbaum:Hillsdale, New York, pp. 243-288, [cited in Scheffers (1983)].
- Klatt, D.H.** (1982) Speech processing strategies based on auditory models, *in*: R. Carlson & B. Granström (eds.) *The Representation of Speech in the Peripheral Auditory System*, Elsevier:Amsterdam, pp. 181-196 [cited in Scheffers (1983)].
- Klein, M.A. & Hartmann, W.M.** (1979) Perception of vibrato width, *Research Symposium on the Psychology & Acoustics of Music*, Lawrence, Kansas.
- Koffka, K.** (1935) *Principles of Gestalt Psychology*, Harcourt and Brace:New York.
- Kohler, W.** (1929) *Gestalt Psychology*, Horace Liveright:New York.
- Kohut, J., Mathews, M.V., Miller, J.E. & Zukovsky, P.** (1981) Violin pitch detection, *J. Acous. Soc. Am.*, **69** (S1):S88 (A).
- Kosslyn, S.M.** (1980) *Image and Mind*, Harvard University:Cambridge, Mass.
- Krumhansl, C.L.** (1979) The psychological representation of musical pitch in a tonal context, *Cog. Psych.*, **11**:346-374.
- Krumhansl, C.L. & Shepard, R.N.** (1979) Quantification of the hierarchy of tonal functions within a diatonic context, *J. Exp. Psych./Human Perc. Perf.*, **5**:579-594.
- Kruskal, J.B.** (1964a) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, **29**:1-27.
- Kruskal, J.B.** (1964b) Nonmetric multidimensional scaling: A numerical method, *Psychometrika*, **29**:115-129.

- Kruskal, J.B., Young, F.W. & Seary, J.B. (1973) How to use KYST, a very flexible program to do multidimensional scaling and unfolding, Bell Labs, Murray Hill, N.J.
- Kubovy, M. and Jordan, R. (1979) Tone-segregation by phase: On the phase sensitivity of the single ear, *J. Acous. Soc. Am.*, **66**:100-106.
- Levitt, H. (1971) Transformed up-down procedures in psychoacoustics, *J. Acous. Soc. Am.*, **49**:467-477.
- Lewis, D. (1936) Vocal resonance, *J. Acous. Soc. Am.*, **8**:91-99.
- Liberman, A.M. & Studdert-Kennedy, M. (1978) Phonetic perception, in: *Handbook of Sensory Physiology, vol. VIII: Perception*, Springer-Verlag:Berlin, pp. 143-178.
- Lieberman, P. (1961) Perturbations in vocal pitch,
- Ligeti, G. (1968) *Atmosphères*, Universal Editions:Vienna.
- Lindsay, P. & Norman, D. (1972) *Human Information Processing*, Academic:New York.
- Lyon, R.F. (1983) Personal communication: conversation 6 July 1983.
- MacIntyre, M.E., Schumacher, R.T. & Woodhouse, J. (1981) Aperiodicity in bowed string motion, *Acustica*, **49**:13-32.
- MacIntyre, M.E., Schumacher, R.T. & Woodhouse, J. (1982) Aperiodicity in bowed string motion: on the differential-slipping mechanism, *Acustica*, **50**:294-295 (L).
- Mathews, M.V. & Miller, J.E. (1981) A study of the pitches produced by violinists while playing a short piece of music, *J. Acous. Soc. Am.*, **70** (S1):S23 (A).
- Mathews, M.V. & Pierce, J.R. (1980) Harmony and nonharmonic partials, *J. Acous. Soc. Am.*, **68**:1252-1257.
- Mattingly, I.G., Liberman, A.M., Syrdal, A. & Halwes, T. (1971) Discrimination in speech and non-speech modes, *Cog. Psych.*, **2**:131-157.
- McAdams, S. (1980) The effects of spectral fusion on the perception of pitch for complex tones, *J. Acous. Soc. Am.*, **68**:S109 (A).
- McAdams, S., (1981) Auditory perception and the creation of auditory images, paper presented at Informatica e Composizione Musicale, Festival Internazionale di Musica Contemporanea, La Biennale di Venezia, Venice, Italy, Sept. 1981.
- McAdams, S. (1982a) Contributions of sub-audio frequency modulation and spectral envelope constancy to spectral fusion in complex harmonic tones, *J. Acous. Soc. Am.*, **72**:S11 (A).
- McAdams, S. (1982b) Spectral fusion and the creation of auditory images, in: M. Clynes (ed.), *Music, Mind and Brain: The Neuropsychology of Music*, Plenum:New York, pp. 279-298.
- McAdams, S. (1983a) Acoustic cues contributing to spectral fusion, *Proc. 11th Int'l.*

- Cong. Acous.*, Paris, **3**:127-130.
- McAdams, S.** (1983b) L'image auditive: Un métaphore pour la recherche musicale et psychoacoustique, *in*: [25].
- McAdams, S.** (1984) The auditory image: A metaphor for musical and psychological research on auditory organization, *in*: R. Crozier & A. Chapman (eds.), *Cognitive Processes in the Perception of Art*, North-Holland:Amsterdam.
- McAdams, S. and Bregman, A.** (1979) Hearing musical streams, *Comp. Mus. J.*, **3**(4):26-43.
- McAdams, S. and Wessel, D.** (1981) A general synthesis package based on principles of auditory perception, Int'l. Comp. Mus. Conf., Denton, Texas (Nov. 1981).
- McNabb, M.** (1981) Dreamsong: The composition, *Comp. Mus. J.*, **5**(4):36-53 (recording available from 1750 Arch St. Records, Berkeley, CA).
- Merzenich, M.M., Knight, P.A. & Roth, G.L.** (1975) Representation of cochlea within primary auditory cortex in cat, *J. Neurophys.*, **38**:231-249.
- Miller, J.R. & Carterette, E.C.** (1975) Perceptual space for musical structures, *J. Acous. Soc. Am.*, **58**:711-720.
- Miller, R.L.** (1953) Auditory tests with synthetic vowels, *J. Acous. Soc. Am.*, **25**:114-121.
- Mills, J.H. & Schmiedt, R.A.** (1983) Frequency sensitivity: Physiological and psychophysical tuning curves and suppression, *in*: J.V. Tobias & E.D. Schubert (eds.) *Hearing Research and Theory*, vol. 2, Academic:New York, pp. 234-336.
- Moore, B.C.J.** (1982) *An Introduction to the Psychology of Hearing*, 2<sup>nd</sup> ed., Academic:London.
- Moorer, J.A.** (1978) How does a computer make music? *Comp. Mus. J.*, **2**(1):32-37.
- Neisser, U.** (1967) *Cognitive Psychology*, Appleton-Century-Crofts:New York.
- Neisser, U.** (1976) *Cognition and Reality: Principles and Implications of Cognitive Psychology*, Freeman:San Francisco.
- van Noorden, L.P.A.S.** (1975) Temporal coherence in the perception of tone sequences, Doctoral dissertation, Tech. Hogeschool, Eindhoven, The Netherlands.
- van Noorden, L.P.A.S.** (1977) Minimum differences of level and frequency for perceptual fission of tone sequences ABAB, *J. Acous. Soc. Am.*, **61**:1041-1045.
- Nordmark, J.O.** (1976) Binaural time discrimination, *J. Acous. Soc. Am.*, **60**:870-880.
- Pastore, R.E., Schmuckler, M.A., Rosenblum, L. & Szczesiul, R.** (1983) Duplex perception with musical stimuli, *Perc. & Psychophys.*, **33**:469-474.
- Penderecki, K.** (1961) *Threnody: To the Victims of Hiroshima*, Belwin-Mills:New York.
- Peterson, G.E. & Barney, H.L.** (1952) Control methods used in the study of vowels, *J.*

- Acous. Soc. Am.*, **24**:175-184.
- Plomp, R.** (1964) The ear as frequency analyzer, *J. Acous. Soc. Am.*, **36**:1628-1636.
- Plomp, R.** (1966) *Experiments on Tone Perception*, Institute for Perception RVO-TNO:Soesterberg, The Netherlands.
- Plomp, R.** (1967) Pitch of complex tones, *J. Acous. Soc. Am.*, **41**:1526-1533.
- Plomp, R.** (1970) Timbre as a multidimensional attribute of complex tones, in: R. Plomp & G. Smoorenburg (eds.), *Frequency Analysis and Periodicity Detection in Hearing*, Sijthoff: Leiden, pp. 398-414.
- Plomp, R.** (1976) *Aspects of Tone Sensation*, Academic:London.
- Plomp, R. & Mimpen, A.M.** (1968) The ear as frequency analyser. II, *J. Acous. Soc. Am.*, **43**:764-767.
- Plomp, R. & Steenecken, H.J.M.** (1971) Pitch versus timbre, *Proc. 7th Int'l. Cong. Acous.*, Budapest, **3**:377-380.
- Polanyi, M.** (1966) *The Tacit Dimension*, Doubleday:Garden City, N.Y.
- Pollack, I.** (1968) Detection and relative discrimination of auditory jitter, *J. Acous. Soc. Am.*, **43**:308-315.
- Pollack, I.** (1970) Jitter detection for repeated pulse trains, in: R. Plomp & G. Smoorenburg (eds.), *Frequency Analysis and Periodicity Detection in Hearing*, Sijthoff:Leiden, pp. 329-335.
- Pollard, J.H.** (1977) *A Handbook of Numerical and Statistical Techniques*, Cambridge Univ. Press:Cambridge.
- Potter, R.K. & Steinberg, J.C.** (1950) Toward the specification of speech, *J. Acous. Soc. Am.*, **22**:807-820.
- Rand, T.C.** (1974) Dichotic release from masking for speech, *J. Acous. Soc. Am.*, **55**:678-680 (L).
- Rasch, R.** (1978) The perception of simultaneous notes such as in polyphonic music, *Acustica*, **40**:21-33.
- Rasch, R.** (1979) Synchronization in performed ensemble music, *Acustica*, **43**:121-131.
- Reynolds, R.** (1983) *Archipelago*, for orchestra and computer-generated tape, C.F. Peters:N.Y.
- Rhode, W.S.** (1971) Observations of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique, *J. Acous. Soc. Am.*, **49**:1218-1231.
- Ritsma, R.J.** (1962) Existence region of the tonal residue I. *J. Acous. Soc. Am.*, **34**:1224-1229.
- Ritsma, R.J.** (1963) Existence region of the tonal residue II. *J. Acous. Soc. Am.*, **35**:1241-1245.

- Ritsma, R.J.** (1967) Frequencies dominant in the perception of the pitch of complex sounds, *J. Acous. Soc. Am.*, **42**:191-198.
- Rodet, X.** (1980a) CHANT manual, IRCAM, Paris, France.
- Rodet, X.** (1980b) Time-domain formant-wave-function synthesis, in: J.C. Simon (ed.), *Spoken Language Generation and Understanding*, Reidel:Dordrecht, Holland. pp. 429-441.
- Rodet, X.** (1982) Projet "Analyse et synthèse de la voix": Rapport d'activité 1981 - Perspectives 1982, *IRCAM Report*, Paris, France.
- Rodet, X.** (1983) Unpublished data from research in progress, IRCAM, Paris, France.
- Rodet, X. & Bennett, G.** (1980) Synthèse de la voix chantée par ordinateur, Conférences des journées d'études, Festival International du Son, Paris, France.
- Rose, J.E., Greenwood, D., Goldberg, J. & Hind, J.E.** (1963) Some discharge characteristics of single neurons in the inferior colliculus of the cat: I. Tonotopical organisation, relation of spike-counts to tone intensity and firing patterns of single elements, *J. Neurophysiol.*, **26**:294-320.
- Roth, G.L., Aitken, L.M., Andersen, R.A. & Merzenich, M.M.** (1978) Some features of the spatial organization of the central nucleus of the inferior colliculus of the cat, *J. Comp. Neurophys.*, **182**:661-680.
- Russell, I.J. & Sellick, P.M.** (1977) The tuning properties of cochlear hair cells, in: E.F. Evans & J.P. Wilson (eds.), *Psychophysics and Physiology of Hearing*, Academic: London, pp. 71-84.
- Russell, I.J. & Sellick, P.M.** (1978) Intracellular studies in the mammalian cochlea, *J. Physiol. (London)*, **284**:261-290.
- Sachs, M.B. & Young, E.D.** (1980) Effects of nonlinearities on speech encoding in the auditory nerve, *J. Acous. Soc. Am.*, **68**:858-875.
- Saldanha, E.L. & Corso, J.F.** (1964) Timbre cues for the recognition of musical instruments, *J. Acous. Soc. Am.*, **36**:2021-2026.
- Sapozhkov, M.A.** (1973) Some factors determining speech perception at cochlea level, presented at the Acoustical Society of the U.S.S.R., June 1973, [cited in Fant (1975)].
- Scharf, B.** (1970) Critical bands, in: J.V. Tobias (ed.) *Foundations of Modern Auditory Theory*, vol. 1, Academic:New York, pp. 159-202.
- Scheffers, M.T.M.** (1983) Sifting vowels: Auditory pitch analysis and sound segregation, Doctoral thesis, University of Groningen, The Netherlands.
- Schubert, E.D.** (1975) The role of auditory perception in language processing, in: *Reading, Perception and Language*, York: Baltimore, pp. 97-130.



- Schubert, E.D. (ed.) (1979) *Psychological Acoustics*, Dowden, Hutchinson & Ross:Stroudsburg, Penn.
- Schubert, E.D. & Nixon, J.C. (1970) On the relation between temporal envelope patterns at two different points in the cochlea, Technical Report, Hearing Science Laboratories, Stanford University.
- Seashore, C.E. (1936) *Psychology of the Vibrato in Voice and Music*, Univ. Press:Iowa City, Iowa.
- Seashore, C.E. (1938) *Psychology of Music*, McGraw-Hill:New York, reprinted 1977 Dover:New York.
- Sellick, P.M. & Russell, I.J. (1979) Two-tone suppression in cochlear hair cells, *Hear. Res.*, 1:227-236.
- Selz, O. (1913) *Über die Gesetze des geordneten Denkverlaufs*, Spemann:Stuttgart, [cited in Sowa (1984)].
- Selz, O. (1922) *Zur Psychologie des produktiven Denkens und des Irrtums*, Cohen:Bonn, [cited in Sowa (1984)].
- Shepard, R.N. (1962a) Analysis of proximities: Multidimensional scaling with an unknown distance function. I, *Psychometrika*, 27:125-140.
- Shepard, R.N. (1962b) Analysis of proximities: Multidimensional scaling with an unknown distance function. II, *Psychometrika*, 27:219-246.
- Shepard, R.N. (1963) Analysis of proximities as a technique for the study of information processing in man, *Human Factors*, 5:19-34.
- Shepard, R.N. (1964) Circularity in judgments of relative pitch, *J. Acous. Soc. Am.*, 36:2346-2353.
- Shepard, R.N. (1982) Structural representations of musical pitch, in: D. Deutsch (ed.) *The Psychology of Music*, Academic:New York.
- Shepard, R.N. & Cooper, L.A. (1982) *Mental Images and their Transformations*, MIT:Cambridge, Mass.
- Shower, E.G. & Biddulph, R. (1931) Differential pitch sensitivity of the ear, *J. Acous. Soc. Am.*, 3:275-287.
- Slaymaker, F.H. (1970) Chords from tones having stretched partials, *J. Acous. Soc. Am.*, 47:1569-1571.
- Small, A.M. (1936) An objective analysis of artistic violin performance, in: C.E. Seashore (ed.) *University of Iowa Studies in the Psychology of Music*, University of Iowa:Iowa City, pp.172-231.
- Sowa, J.F. (1984) *Conceptual Structures: Information Processing in Mind and Machine*, Addison-Wesley:Reading, Mass.

- Spiegel, M.F. & Green, D.M.** (1982) Signal and masker uncertainty with noise maskers of varying duration, bandwidth and center frequency, *J. Acous. Soc. Am.*, **71**:1204-1210.
- Speigel, M.F., Picardi, M.C. & Green, D.M.** (1981) Signal and masker uncertainty in intensity discrimination, *J. Acous. Soc. Am.*, **70**:1015-1019.
- Stevens, K.N. & House, A.S.** (1972) Speech perception, in: J.V. Tobias (ed.), *Foundations of Modern Auditory Theory*, vol. 2, Academic:New York, pp. 1-62.
- Stumpf, C.** (1890) *Tonpsychologie*, S. Hirzel-Verlag:Leipzig (reissued 1965 by Knuf-Bonset:Hilversum-Amsterdam) [cited in Brokx & Nootboom (1982)].
- Stumpf, C.** (1926) *Die Sprachlaute*, Springer:Berlin [cited in Plomp (1970) and Sundberg (1977)].
- Sundberg, J.** (1975) Formant technique in a professional female singer, *Acustica*, **32**:89-96.
- Sundberg, J.** (1977) Vibrato and vowel identification, *Archives of Acoustics, Polish Academy of Sciences*, **2**(2):257-266.
- Sundberg, J.** (1978) Synthesis of singing, *Swedish J. Musical.*, **60**. [cited in Sundberg (1982)].
- Sundberg, J.** (1982) Perception of singing, in: D. Deutsch (ed.), *The Psychology of Music*, Academic:New York, pp.59-98.
- Terhardt, E.** (1974) Pitch, consonance and harmony, *J. Acous. Soc. Am.*, **55**:1061-1069.
- Thurlow, W.R. & Small, A.M.** (1955) Pitch perception for certain periodic auditory stimuli, *J. Acous. Soc. Am.*, **27**:132-137.
- Torgerson, W.S.** (1958) *Theory and Methods of Scaling*, Wiley:New York.
- Tovar & Smith, L.** (1977) *MUS10 Manual*, unpubl. user's manual, CCRMA, Dept. of Music, Stanford University, Stanford, Calif.
- Vicario, G.B.**, (1965) Vicinanza spaziale e vicinanza temporale nella segregazione di eventi, *Rivista di Psicologia*, **59**:843-863 [cited in Vicario (1982)].
- Vicario, G.B.**, (1982) Some observations in the auditory field, in: Beck, J. (ed.) *Organization and Representation in Perception*, Erlbaum:Hillsdale, New Jersey.
- Voigt, H.F., Sachs, M.B. & Young, E.D.** (1981) Representation of whispered vowels in temporal patterns of auditory-nerve fiber discharges, *J. Acous. Soc. Am.*, **69**:S53 (A).
- Wedin, L. & Goude, G.** (1972) Dimension analysis of the perception of instrumental timbre, *Scand. J. Psych.*, **13**:228-240.
- Wessel, D.L.** (1979) Timbre space as a musical control structure, *Comp. Mus. J.*, **3**(2):45-52.

- Wessel, D.L.** (1983) Timbral control in research on melodic patterns, presented at the 4th Int'l. Workshop on the Physical and Neuropsychological Foundations of Music, Ossiach, Austria, August, 1983.
- Wever, E.G.** (1949) *Theory of Hearing*, republ. 1970, Dover:New York.
- Wightman, F.L.** (1973) The pattern-transformation model of pitch, *J. Acous. Soc. Am.*, **54**:407-416.
- Willis, R.** (1830) On the vowel sounds and on reed organ-pipes, *Trans. Cambr. Philos. Soc.*, **3**:231-268, [cited in Plomp (1970)].
- Worden, F.G. & Galambos, R.** (1971) Auditory processing of biologically significant sounds, *Neurosciences Res. Prog. Bul.*, **10**(1).
- Young, E.D. & Sachs, M.B.** (1979) Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers, *J. Acous. Soc. Am.*, **66**:1381-1402.
- Young, F.W.** (1970) Nonmetric multidimensional scaling: Recovery of metric information, *Psychometrika*, **35**:455-473.
- Young, F.W.** (1972) A model for polynomial conjoint analysis algorithms, in: R.N. Shepard, A.K. Romney & S.B. Nerlove (eds.), *Multidimensional Scaling: Theory and Application in the Behavioral Sciences*, vol. 1, Seminar Press:New York, pp.69-102.
- Young, F.W. & Torgerson, W.S.** (1967) TORSCA: A Fortran-4 program for Shepard-Kruskal multidimensional scaling analysis, *Beh. Sci.*, **12**:498.
- Zwicker, E.** (1960) Ein Verfahren zur Berechnung der Lautstärke, *Acustica*, **1**:304-308.
- Zwicker, E.** (1974) On a psychoacoustical equivalent of tuning curves, in: E. Zwicker & E. Terhardt (eds.), *Facts and Models in Hearing*, Springer-Verlag:Berlin, pp. 132-141.
- Zwicker, E. & Terhardt, E.** (1980) Analytic expression for critical-band rate and critical bandwidth as a function of frequency, *J. Acous. Soc. Am.*, **68**:1523-1525 (L).

