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

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

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 resulsts 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 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.

to
Pa &
Gloria

Spectral Fusion, Spectral Parsing and the Formation of Auditory Images

Stephen McAdams, Stanford University, 1984

ABSTRACT

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 or 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 of 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.

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 Universtiy 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

LIST OF TABLES

LIST OF FIGURES

~ 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.

~ II ~

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 [1] 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 pitfall 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. Köhler, 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, in 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 effeots 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 readinesses) 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. [2] 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 Helmoltz' (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*. [3] (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

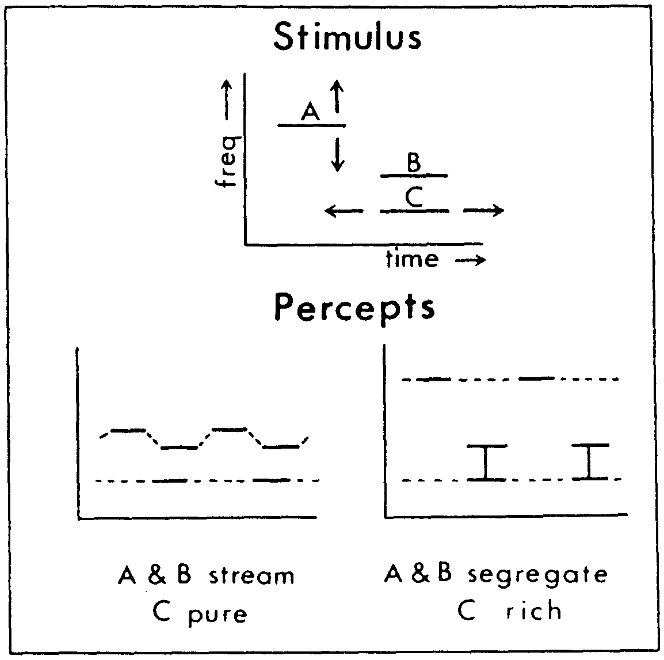

Figure 1.1. Schematic representation of the stimuli used by Bregman & Pinker (1978) and two common perceptual results. Each horizontal line represents a sinusoidal frequency component. The dashed lines represent perceived sequential organization and the vertical solid lines represent perceptual fusion.

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

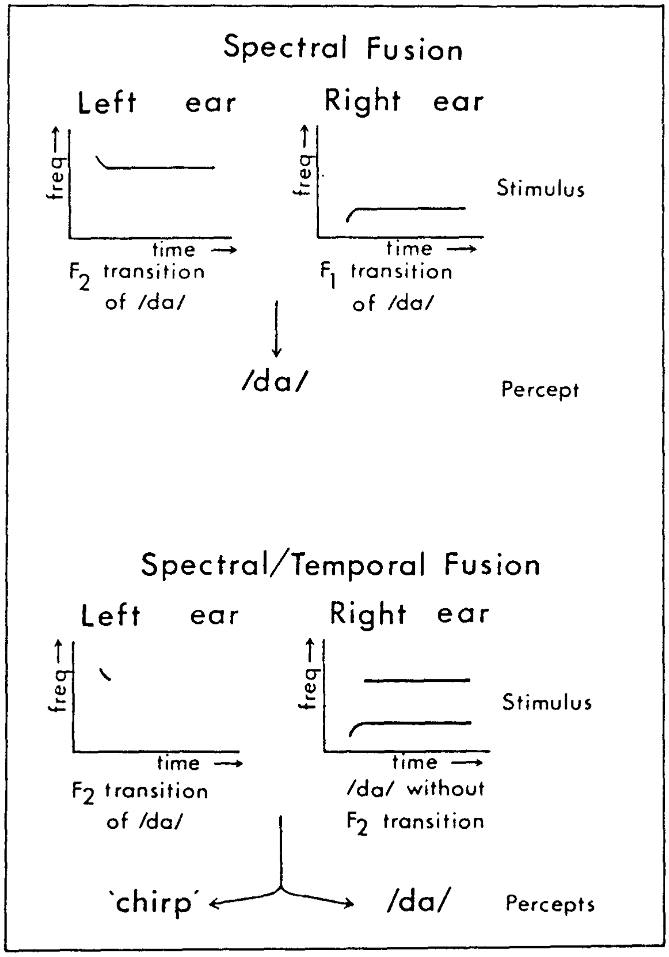

Figure 1.2. Schematic representation of the stimuli used by Cutting (1976) and the most common perceptual results. The lines represent the frequency trajectories of the first (F_1) and second (F_2) speech formants.

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)[4] 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.[5] 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 investigate 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 16th 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 5th (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 3rd 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.[6] 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 $f_n = n^s F_0$ (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 submultiples 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,[7] 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.[8] 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.[9] 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.[10] 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.[11]

[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.

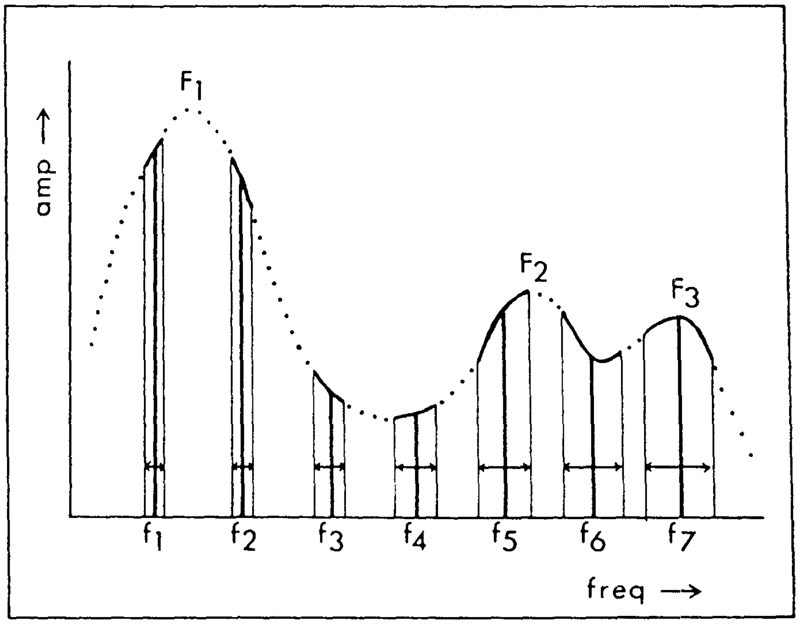

Figure 1.3. The spectral form created by a 3-formant resonance structure is represented by the dotted line. As the harmonics are modulated in frequency, their respective amplitudes fluctuate as a function of the spectral envelope. This is indicated by the solid portions on the dotted line. Note that 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.

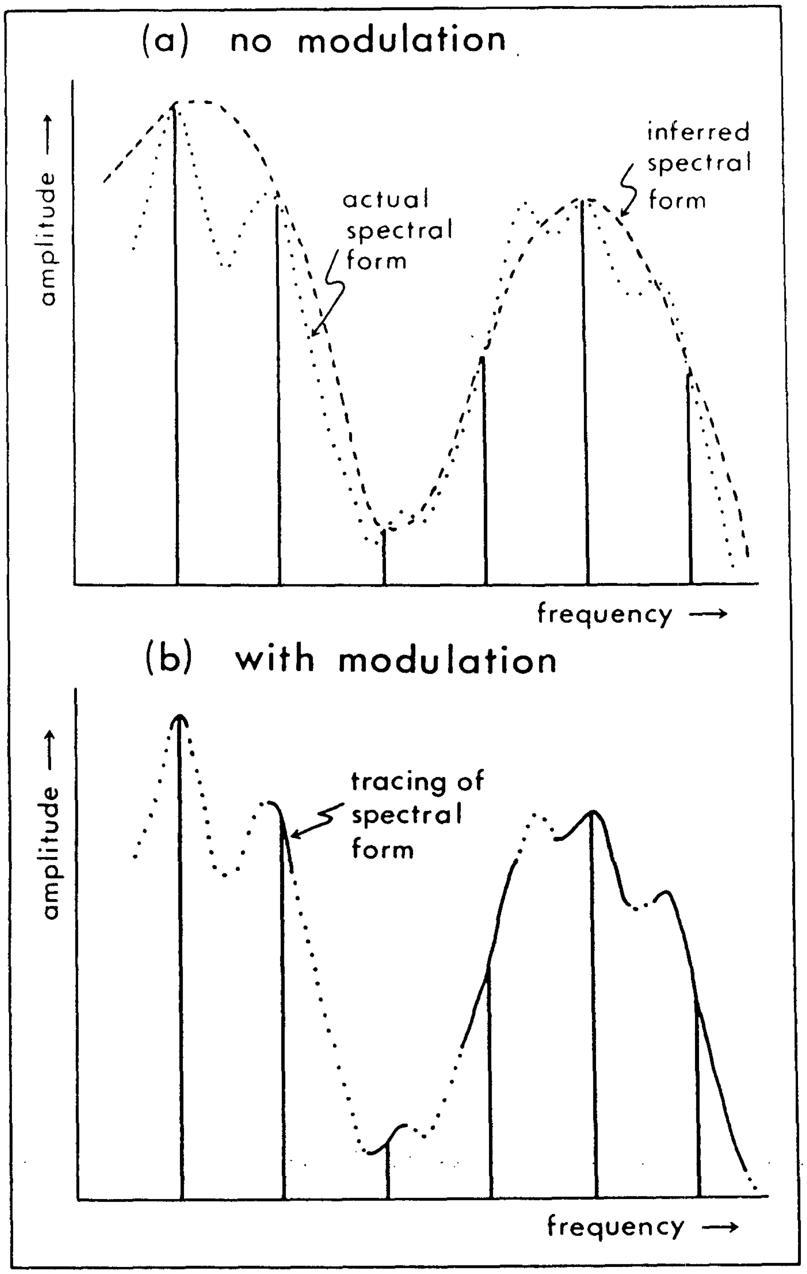

Figure 1.4. The vowel /a/ is plotted with a high fundamental frequency where there are few harmonics present. Without modulation (a), the inferred spectral form (dashed line) would be very different from the actual spectral form (dotted line). With modulation (b), the spectral slopes give a much clearer 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 ($\overline{M}_k$) of the "average response vowel", expressed in Mels,[12] were calculated as the average of a given formant across all responses,

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

   where $M_{kRi} \equiv$ the k^{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:

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

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. $M = 1000 \log_2(1+F/1000)$,[13] 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/, /ʉ/ and /ø/ had smaller D than response /u/. For stimulus /i/, response vowels /e/, /y/, /ae/, /ε/, /ʉ/, /ø/, and /oe/ had smaller D than response /i/ For stimulus /y/, response vowels /ʉ/ 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 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

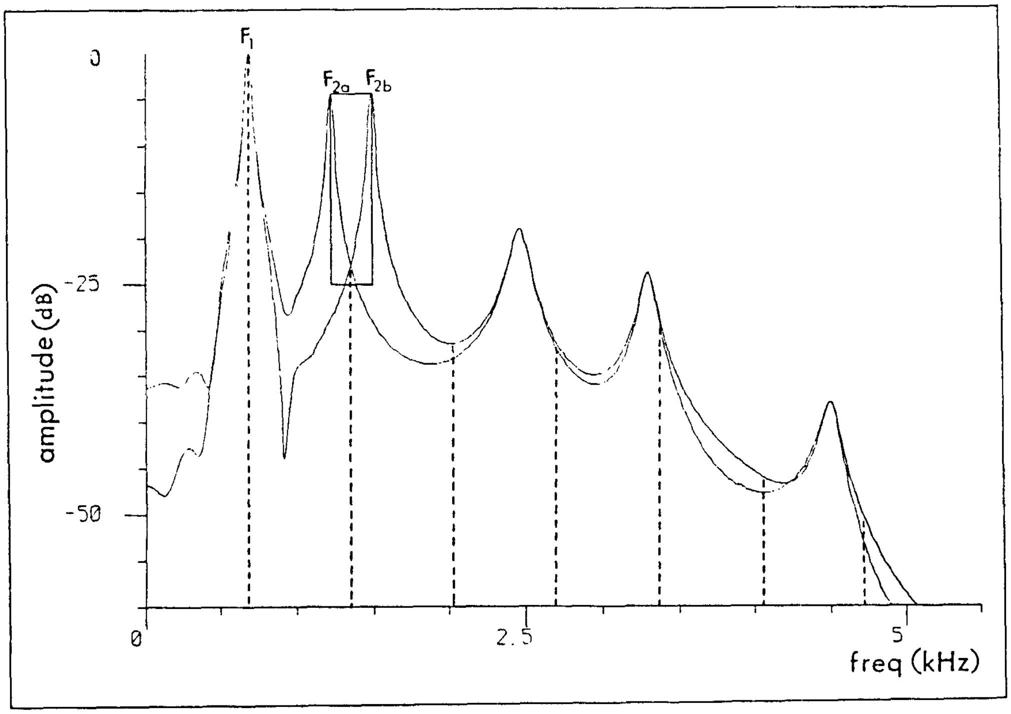

Figure 1.5. Two spectral envelopes used by Rodet (1983). The second harmonic falls at the intersection of the two possible F_2's. For F_{2a} its amplitude slope with frequency modulation is negative. For F_{2b}, it is positive. Figure 1.6 shows an enlargement of the box.

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.

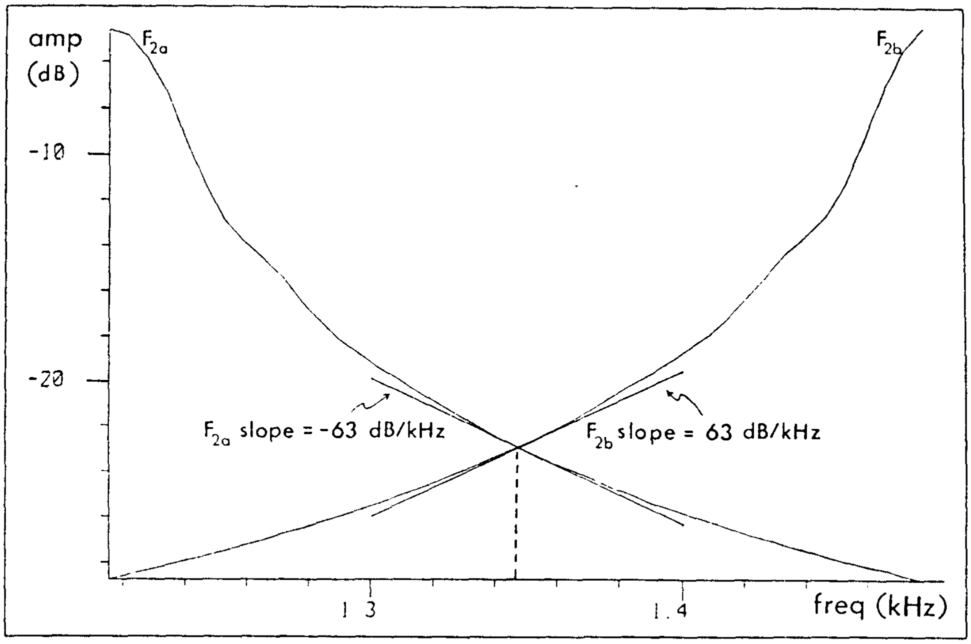

Figure 1.6. Enlargement of the region of the second harmonic's spectral slopes from Figure 1.5 (from Rodet, 1983).

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 (Bjørklund, 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

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

where n is the harmonic number, A_n is the amplitude of harmonic n, $Mod(t')$ is the modulating waveform representing the frequency perturbation, and φ_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.[14] 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 & Nooteboom (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 (1977/1885), 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.[15]

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 asychrony.

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 nonlinear. 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 [16] 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 discriminibility between coherently or incoherently modulated sinusoidal carriers, where the modulating waveforms were narrow-band noise (75 Hz or 300 Hz). The lack of discriminibility 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 [17] 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.[18] 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*.[19] 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 & Nooteboom, 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) that 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 [1] of the modulation were used. These values, expressed as both cents (1cent = 1/100th of a semitone) and as $\Delta f / \overline{f}$, are listed in Table 2.1. The relation between the cents measure and $\Delta f / \overline{f}$ 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
$\frac{\Delta f}{f}$	0.00405	0.00812	0.01630	0.02456	0.03288

$$\frac{\Delta f}{f} = 2^{\frac{\text{cents}}{1200}} - 1. \tag{2.1}$$

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:

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

and

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

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; ψ is the desired rms deviation, D_{rms}, divided by the actual rms deviation, A_{rms}, of $Mod(t)$ (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 $Mod(t)$, then

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

and

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

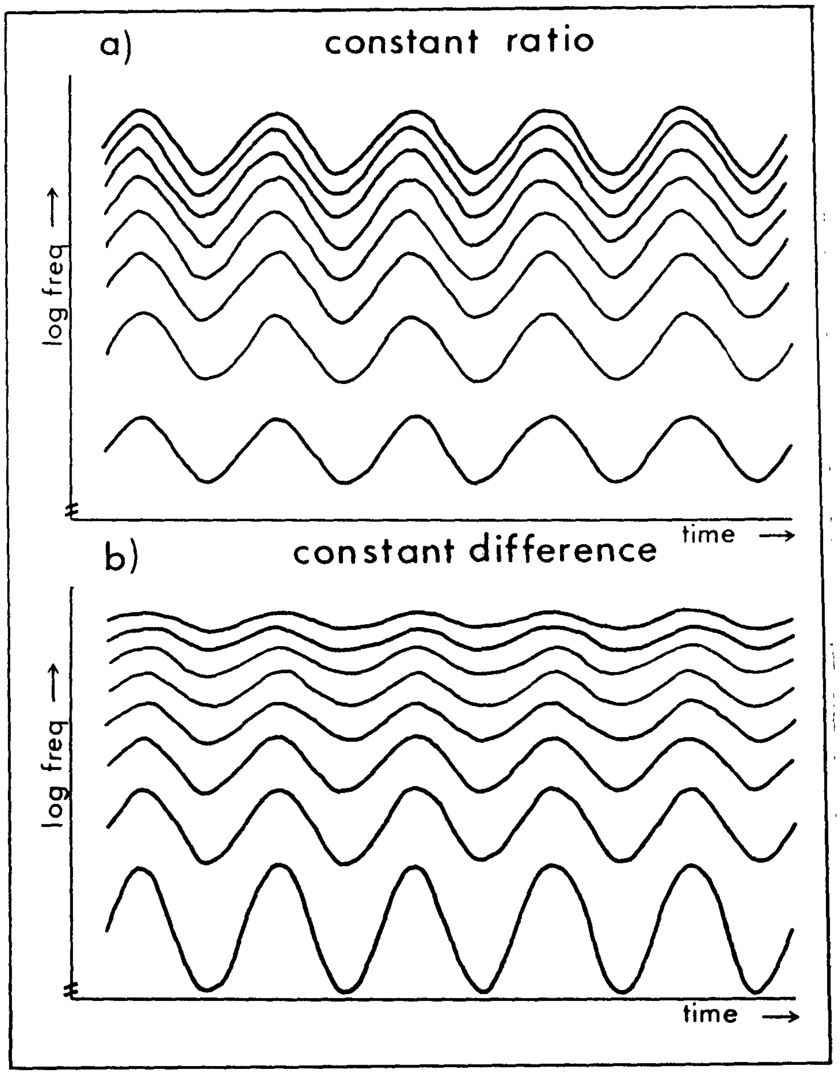

Figure 2.1. Exaggerated spectrographic diagram of *CR* and *CD* modulations plotted on a log frequency scale for the first 8 harmonics.

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, Δf_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} = 14cents, the rms deviations on the fundamentals are equal, but the maximum frequency excursion for the 16th 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 ψ (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.

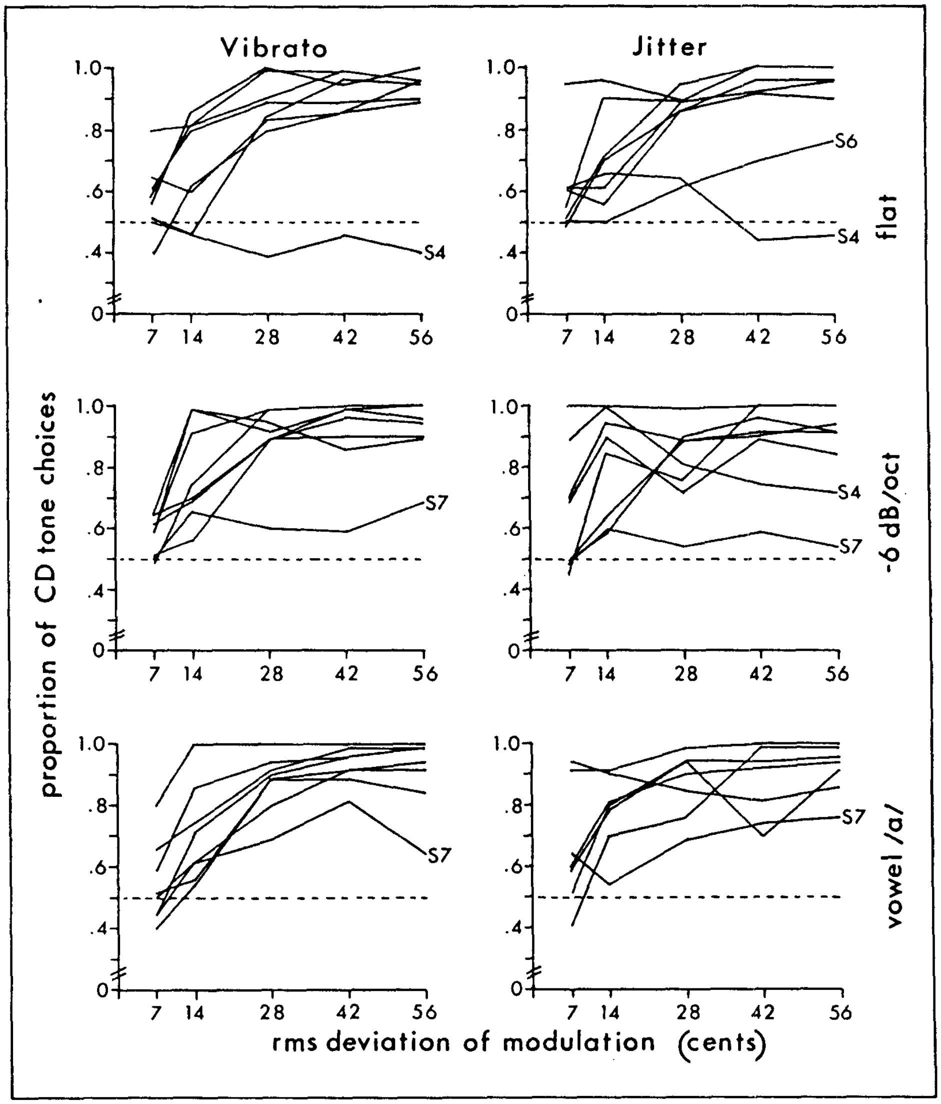

Figure 2.2. Experiment 1 data summary. Each graph shows the proportion of times the constant-difference tone was chosen as having more sources than the constant-ratio tone as a function of the rms deviation (expressed in cents). Within each graph a separate function is shown for each subject. The graphs 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) × (5 rms deviations) × (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 ψ 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)
overall mean	(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)
overall mean	(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 [2] 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.

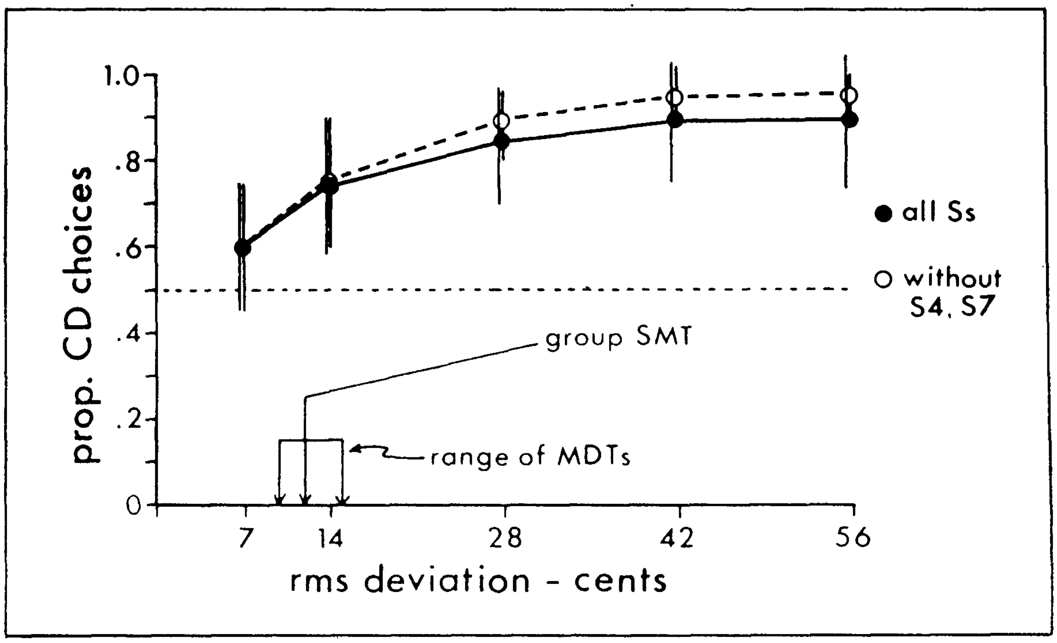

Figure 2.3. Experiment 1 data summary. The proportion of times the *CD* tone was chosen as yielding more sources is plotted as a function of the rms deviation of modulation. Data points are averaged over Ss, spectral envelope and modulation waveform. The arrows on the ordinate indicate the group SMT 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.[3] 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$_{rms}$). Seashore (1936, 1938) reported peak vibrato widths in singers varying between 31 cents and 98 cents (approximately 22 - 69cents$_{rms}$). 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: [4]

[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$_{rms}$ (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$_{rms}$ (Pollack, 1968, 1970; Cardozo & Neelen, 1968).

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

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

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

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

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

cents_{rms}	cents_{peak}	
	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.

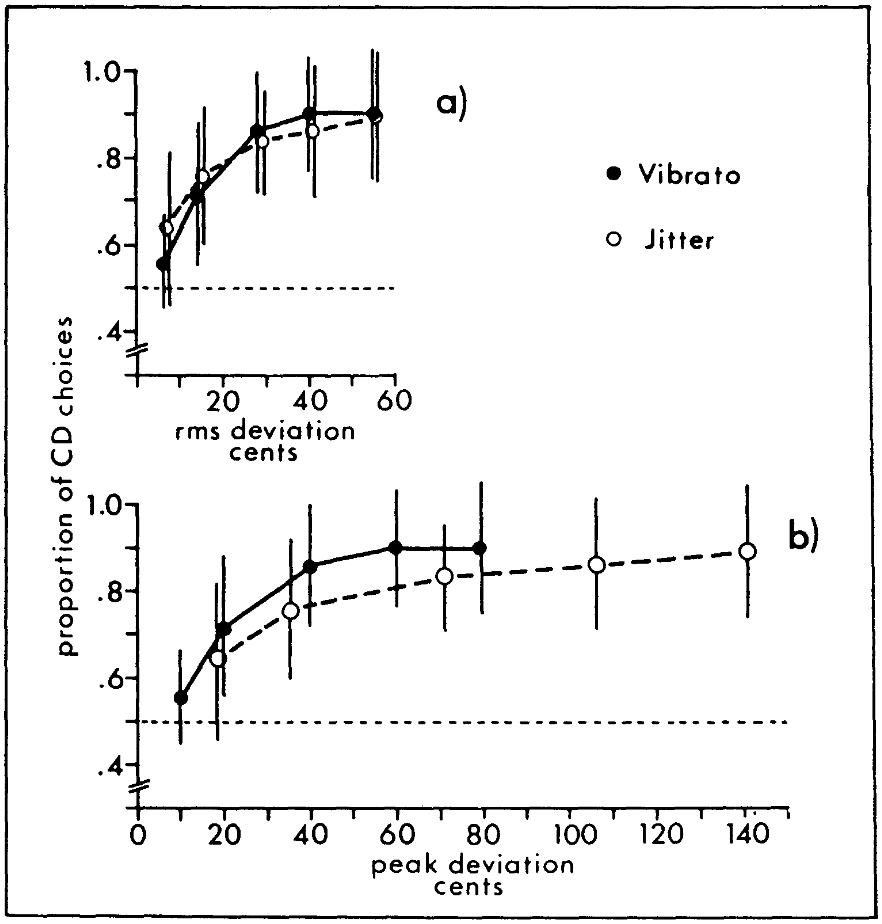

Figure 2.4. Experiment 1 data summary. Means within modulation deviation width and modulation waveform are plotted as functions of (a) rms deviation and (b) peak deviation. Data are averaged across all Ss and spectral envelope conditions. The vertical bars represent ±1 standard deviation. The closed 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_{rms} and as $\Delta f / \bar{f}$. The column at the right shows the difference in cents between *CD* and *CR*.

CR		*CD*		*CD - CR*
cents	$\Delta f / \bar{f}$	cents	$\Delta f / \bar{f}$	Δcents
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 wàs 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.[5] 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).[6] 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.[7] 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 ±2 in the solution. Here they· have been normalized to between 0 and 1.

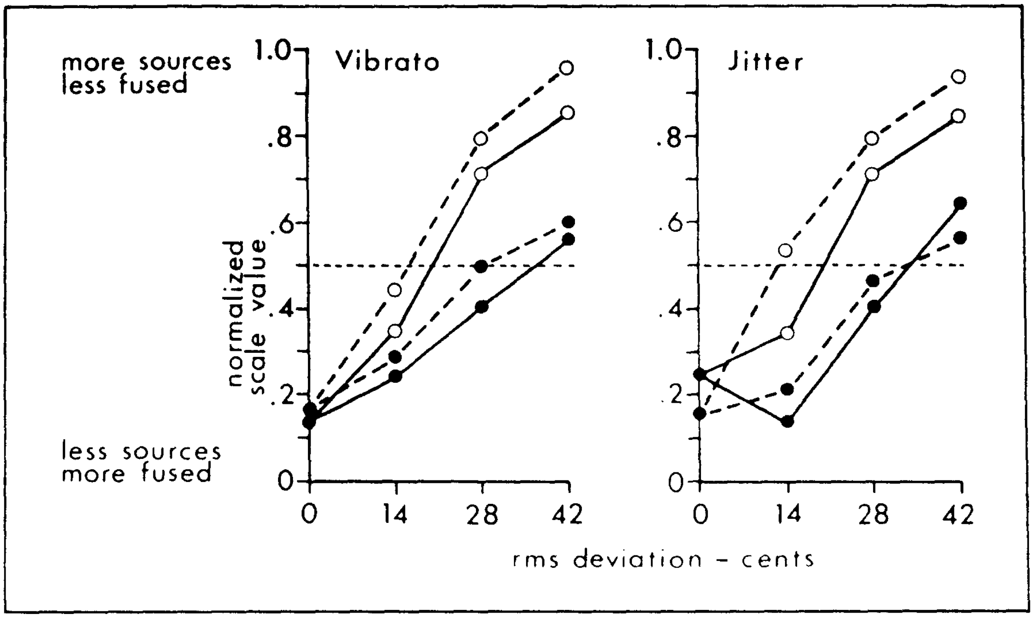

Figure 2.5. Experiment 2 data summary. One-dimensional scaling solutions for judgments on the relative degree of fusion for vibrato and jitter stimuli plotted as a function of rms deviation of modulation. The scale is interpreted as the degree of perceived fusion or perceived source multiplicity. Closed 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 = −6dB/oct; /a/ = 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 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.[8] 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; dissimiliarities 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 2IFC 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.[9] 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. Δcents = *CD* − *CR*.

CR	*CD*	Δcents
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) × (3 rms deviation pairs) × (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

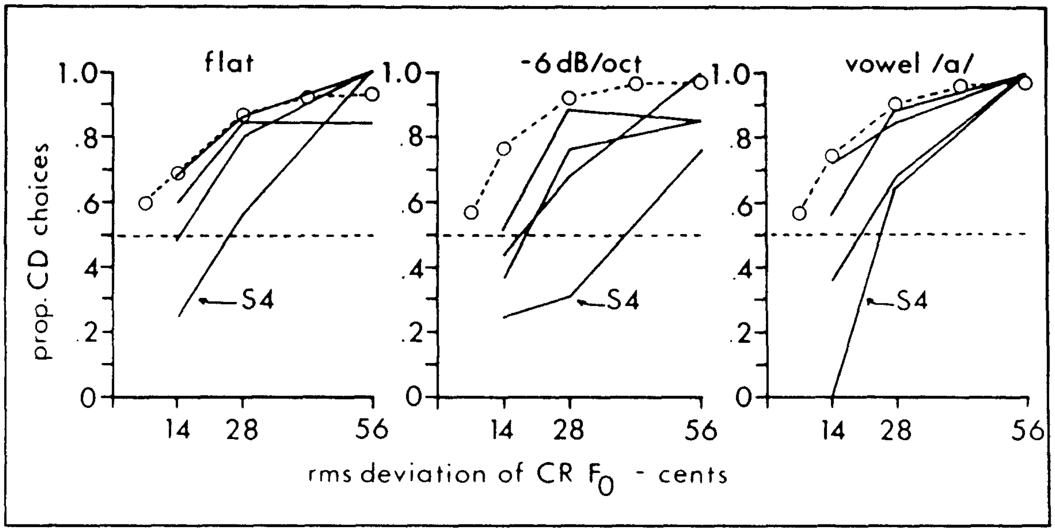

Figure 2.6. Experiment 3 data summary. The proportion of times the *CD* tone was chosen as yielding more sources is plotted as a function of rms deviation of modulation (the modulation width on F_0 in the *CR* tone for each stimulus pair). Each curve represents the data for one subject. Each point represents 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.

CR tones		*CD* tones	
cents	Δf_{rms} (Hz)	cents	Δf_{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 (Δf_{rms} in Hz) across *CR/ CD* pairs in Experiments 1,3,4,5.

Experiments 1 and 4					
$\Delta f_{rms}(CR)$	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
Experiments 3 and 5					
$\Delta f_{rms}(CR)$		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) x (5 rms deviations) x (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 stimili 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.

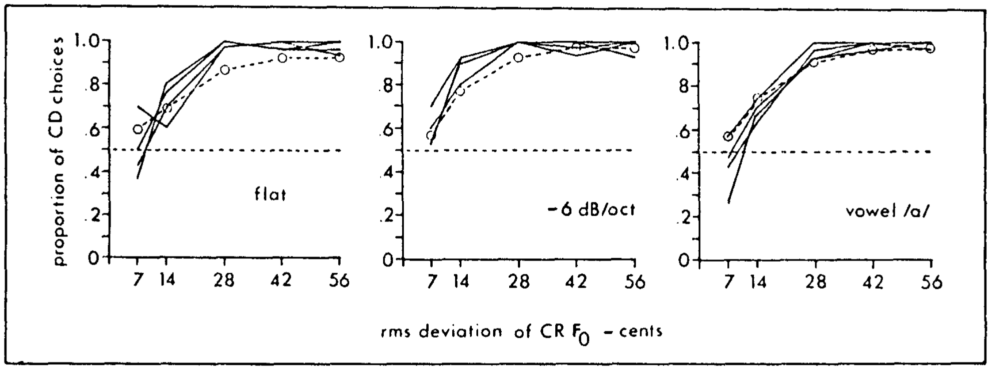

Figure 2.7. Experiment 4 data summary. The proportion of times the *CD* tone was chosen as having a larger modulation width on the F_0 than that on the *CR* tone is plotted as a function of the modulation width of the *CR* F_0. Refer to Table 2.4 for the difference in F_0 modulation width between *CR* 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) x (3 rms deviations) x (25 repetitions).

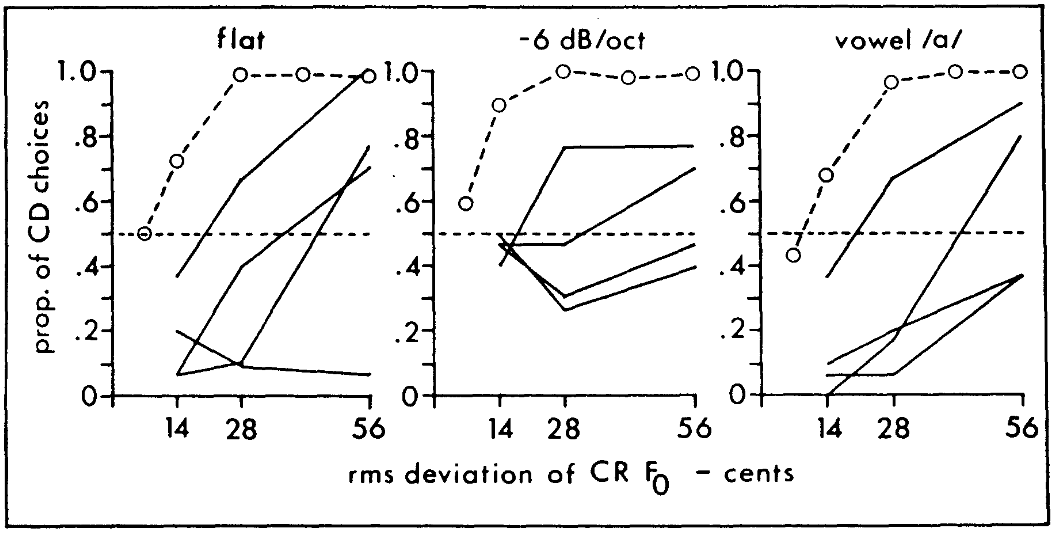

Figure 2.8. Experiment 5 data summary. The proportion of times the *CD* tone was chosen as having a larger modulation width on the F_0 is plotted as a function of *CR* modulation width. Each curve represents the data for one subject. Each point represents 25 2IFC judgments. For comparison, the mean data from Experment 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 devation ***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 Δf_{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 (Δcents) 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 Δcents is much larger and the modulations are at suprathreshold widths, subjects reliably chose the tone which actually had the larger modulation width.

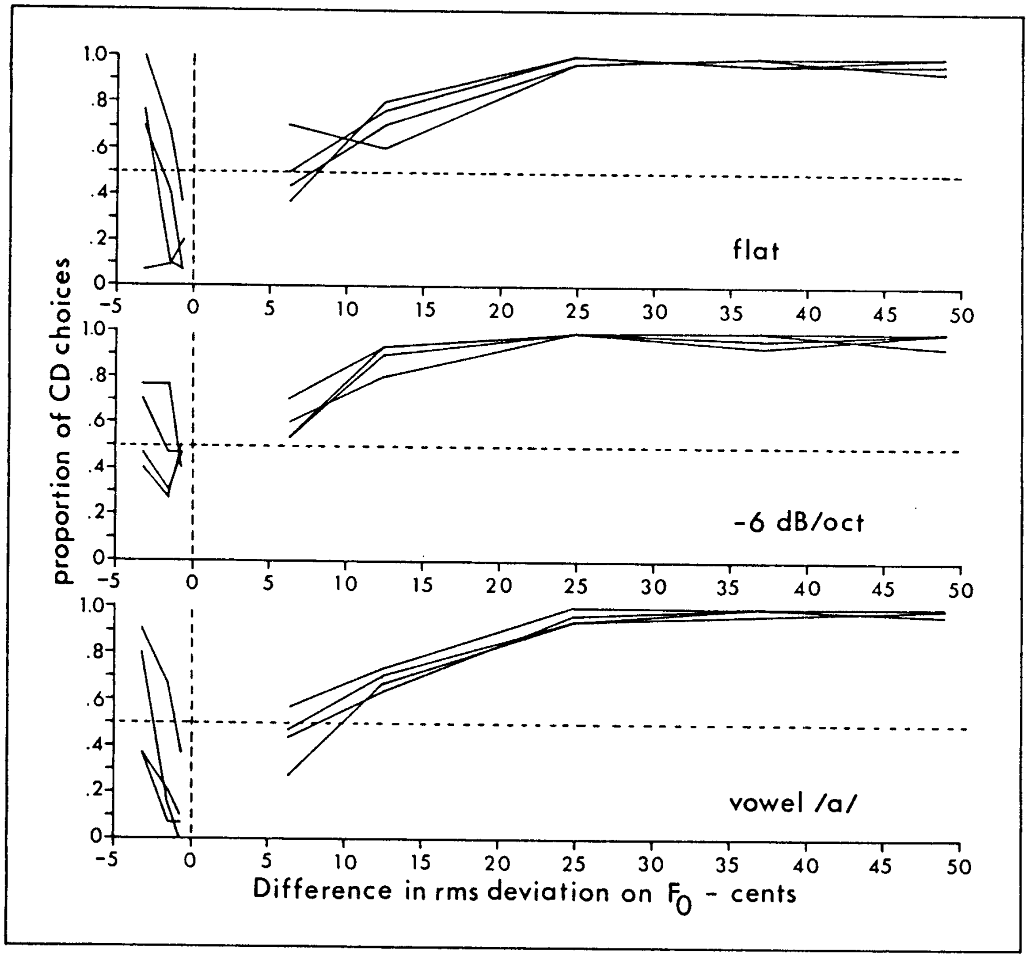

Figure 2.9. Data from Experiments 4 and 5 (modulation width judgments) plotted as functions of the difference in modulation width (*CD−CR*) of the F_0's on the two tones in a trial. The Δcents values in Experiment 4 are positive; those in Experiment 5 are negative. Each curve represents the data for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

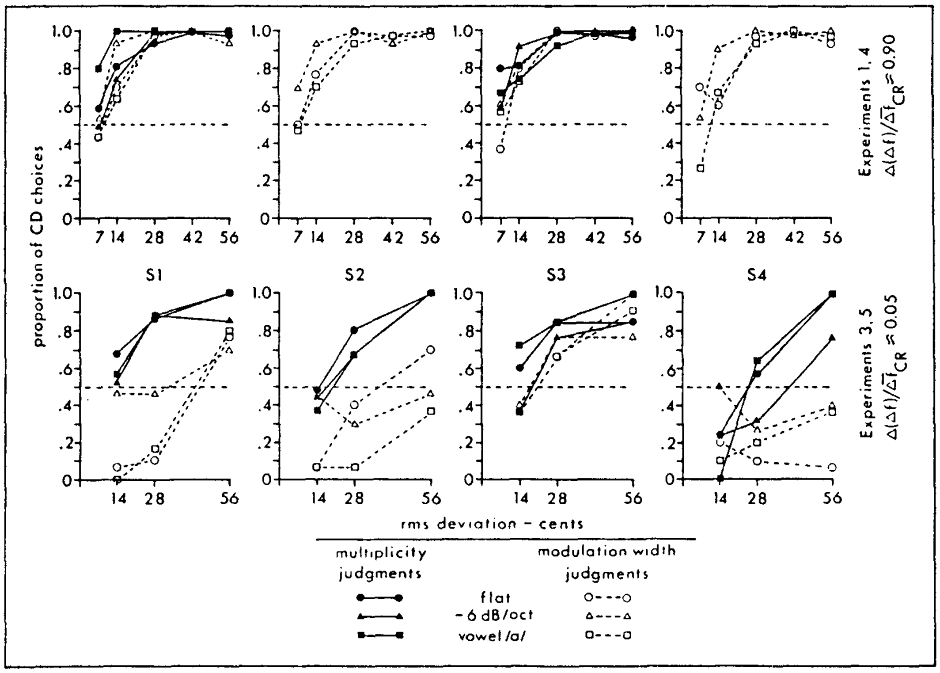

Figure 2.10. Individual data for each subject for Experiments 1 (S1 and S3 only), 3, 4 and 5 are plotted. Source multiplicity judgment data are represented as filled/solid and modulation width judgments are represented as open/dashed. The upper plot is for Experiments 1 and 4 ($\Delta f_{CD} \approx 1.9\Delta f_{CR}$). 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[1]

[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 μsec or 50 μsec, 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,[2] physiologically,[3] and psychoacoustically.[4] 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.[5] 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[6])

[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

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

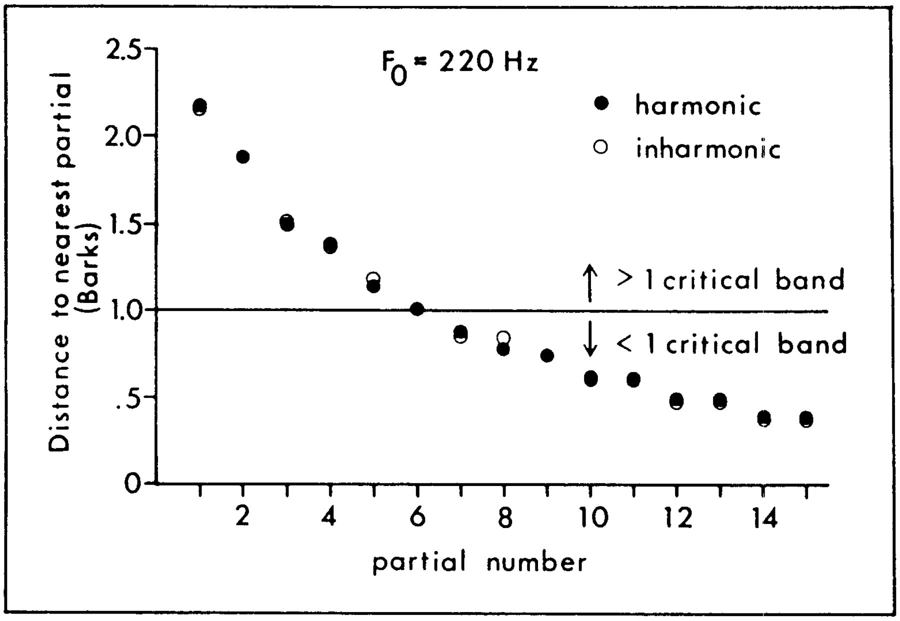

Figure 3.1. Plotted here is the distance in Barks to the next nearest partial of a 16-component complex tone Both harmonic and inharmonic tones are plotted for comparison. The Bark estimates were obtained according to the algorithm of Zwicker & Terhardt (1980).

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 ±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 (ΔBark between f_8 and f_9) was from 0.783 Bark to 0.848 Bark (ΔBark = 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.[7] 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:

$$S_C(t) = \sum_{n=1}^{16} \sin(2\pi f_n t + \frac{f_n}{f_1}\psi_1 \int_0^t J_1(t')\,dt'), \qquad (3.1)$$

where $S_C(t)$ is the coherently modulated signal waveform, f_n are the component center frequencies from Table 3.1, and ψ 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.

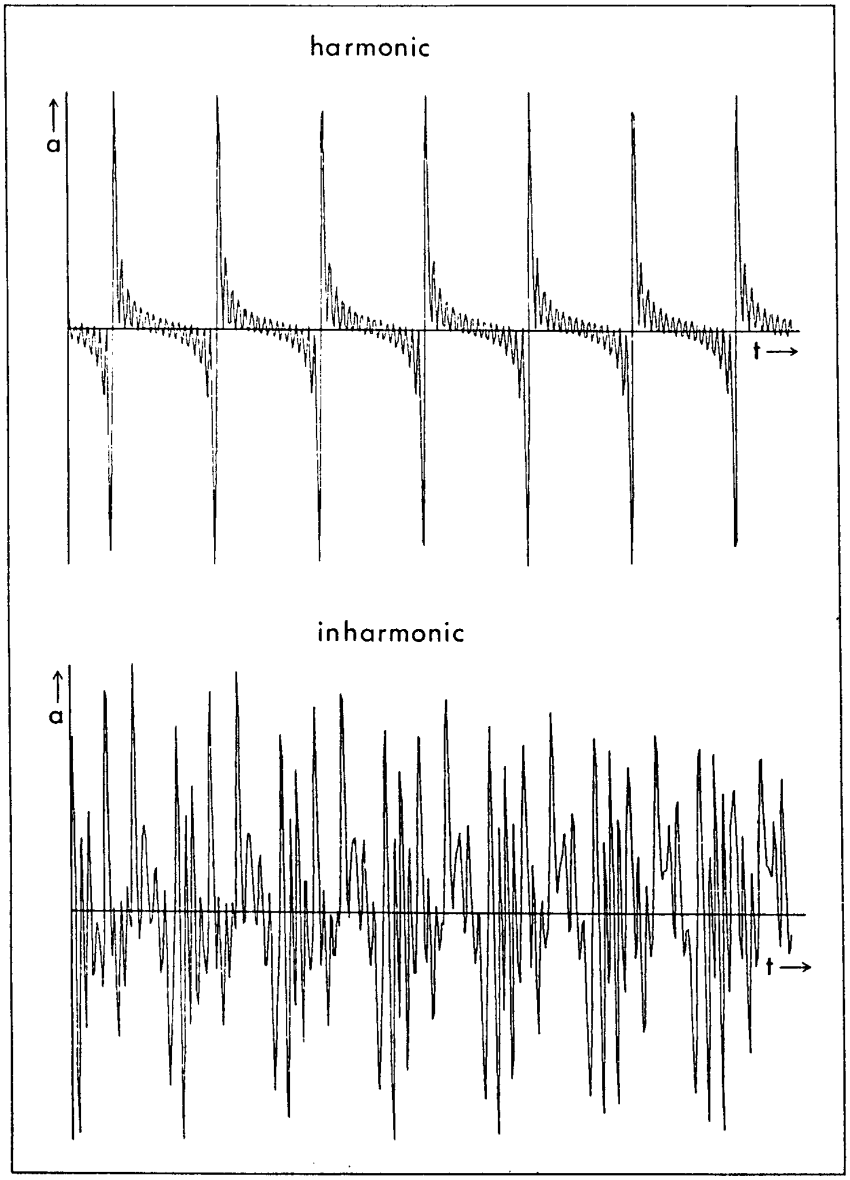

Figure 3.2. Plotted here are 32 msec segments (approximately 7 cycles of f_1) from the waveforms of the unmodulated harmonic and inharmonic tones used in Fxperiment 6

The equation describing the "incoherent" signal is:

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

$$\text{for} \quad k = 1, 3, 5, 7, 9, 11, 13, 15, \tag{3.2}$$

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 ψ 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.

	Rms Deviation (cents)									
Partial	**Harmonic**					**Inharmonic**				
Number	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 × 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 [8] (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 arhythmic pulsing of auditory roughness (5 - 15)[9]. 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 16th!

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

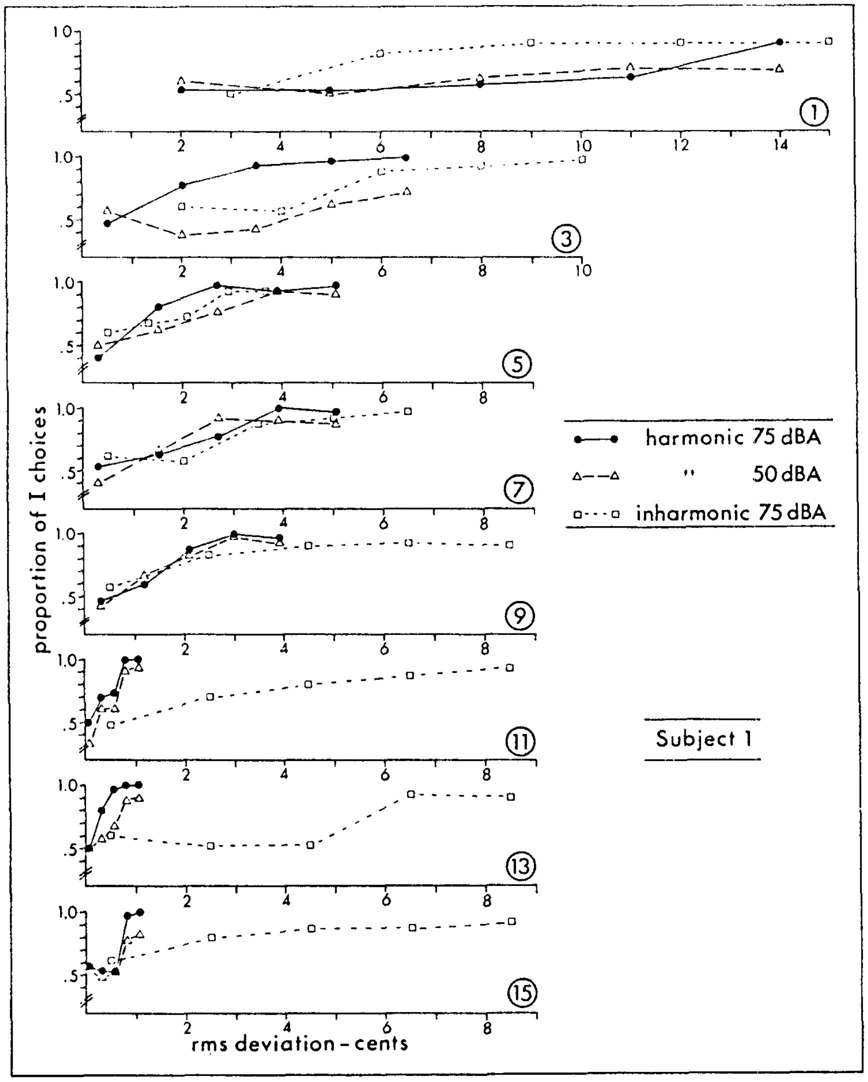

Figure 3.3. Experiment 6 data summary for Subject 1 The proportion of incoherent tone choices is plotted as a function of rms deviation of modulation (in cents). Each graph represents the data for one incoherently modulated partial (encircled number to right of graph). Shown in each graph are the curves 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 mores 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

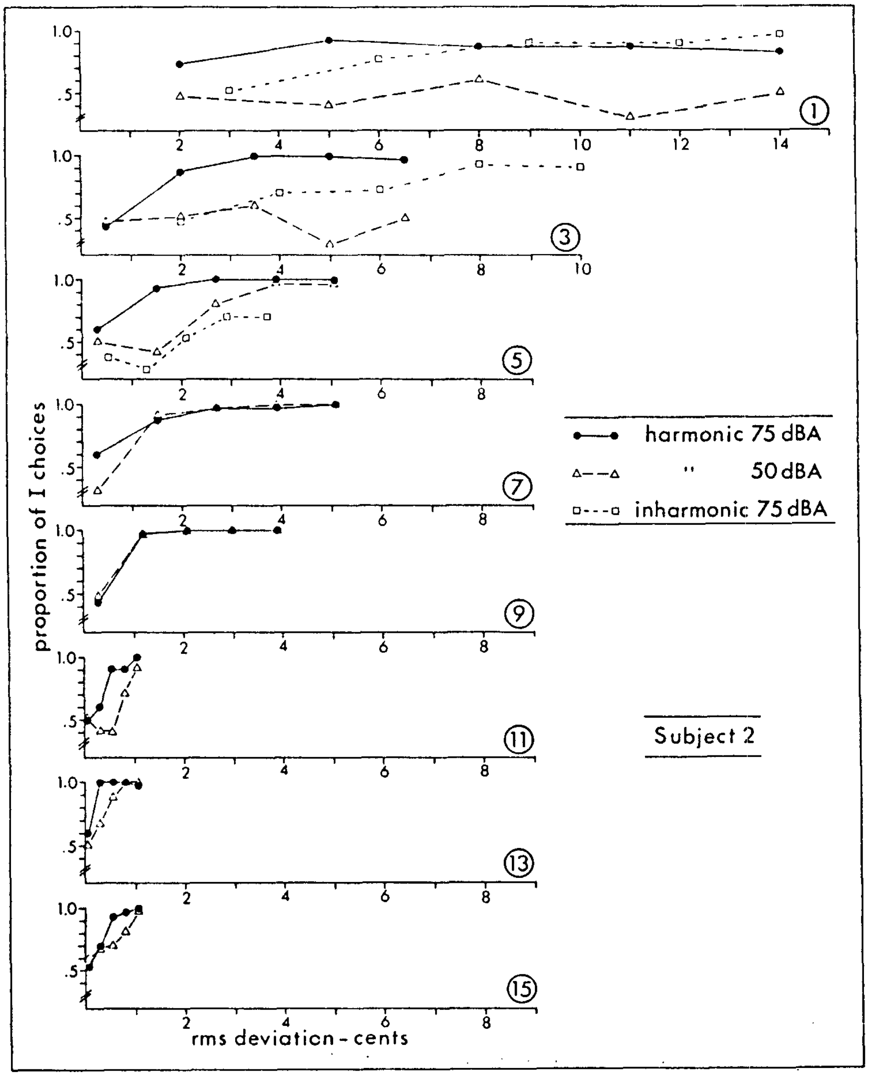

Figure 3.4. Experiment 6 data summary for Subject 2. This subject did not complete the I75 condition for partial numbers 7 – 15. (See caption for Fig. 3.3.)

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

crossing was chosen as the SMT (I75 for f_5)

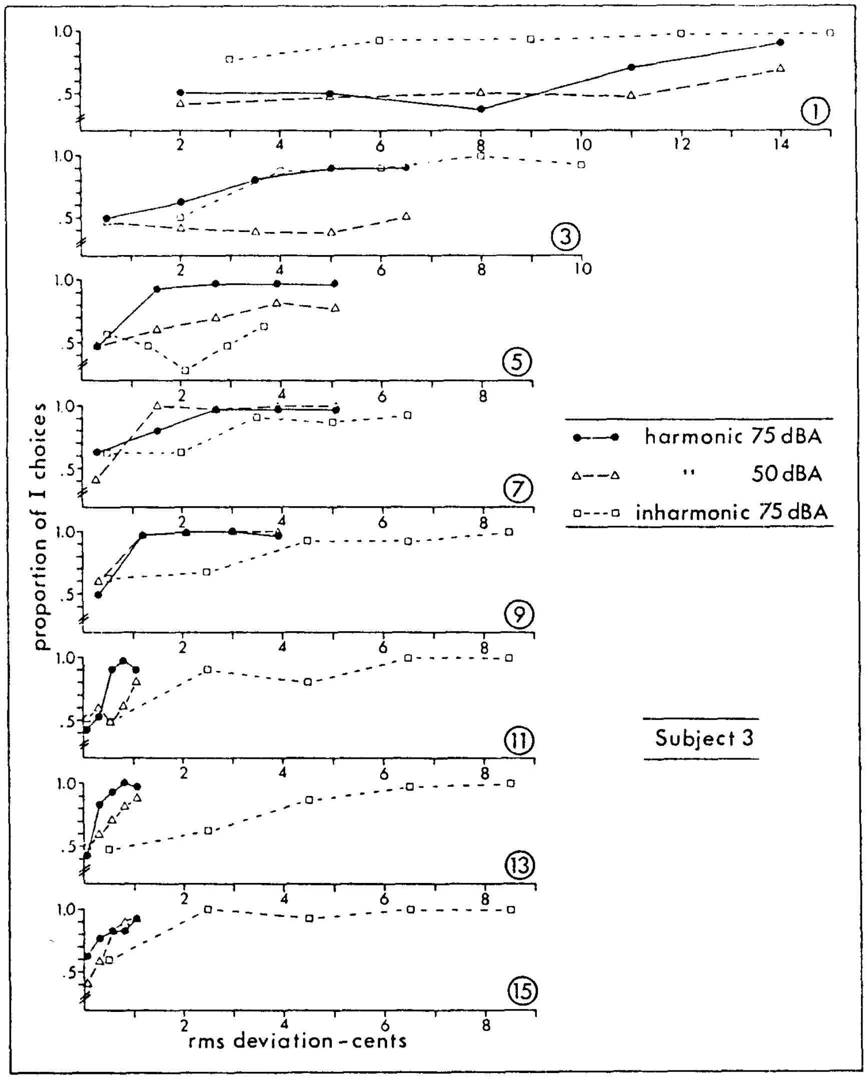

Figure 3.5. Experiment 6 data summary for Subject 3. (See caption for Fig. 3.3.)

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

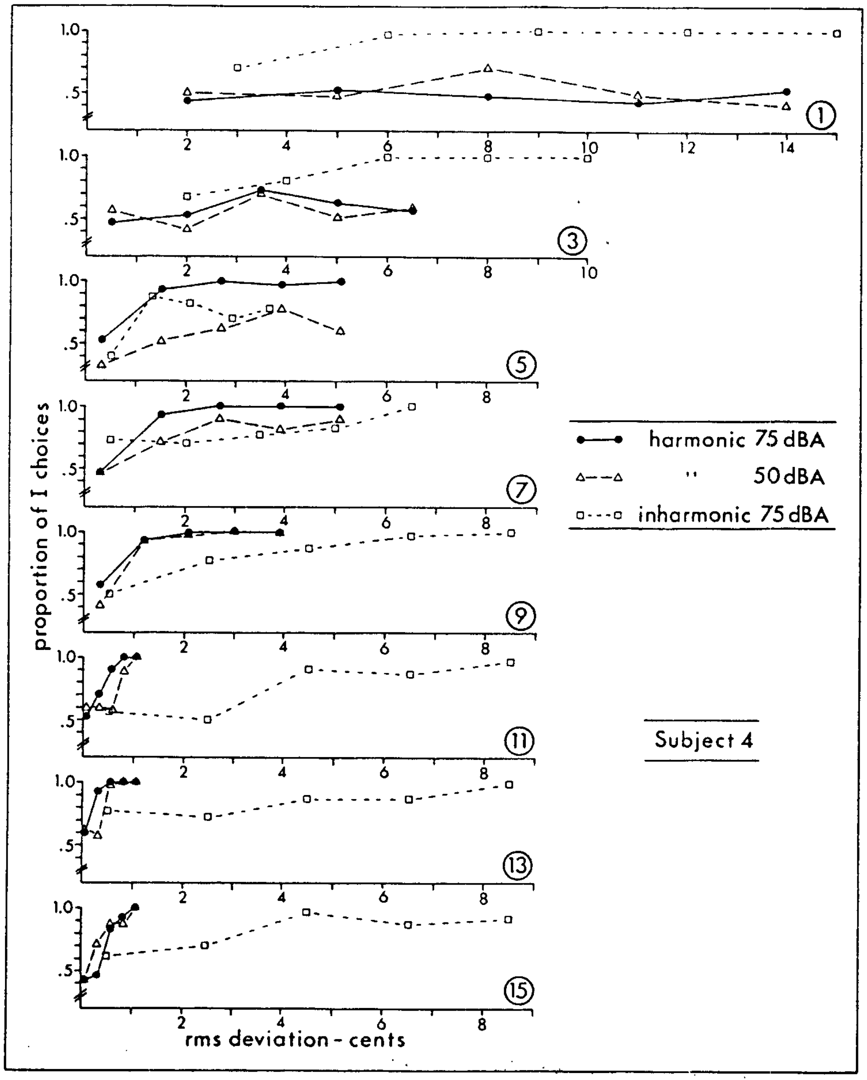

Figure 3.6. Experiment 6 data summary for Subject 4. (See caption for Fig. 3.3.)

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

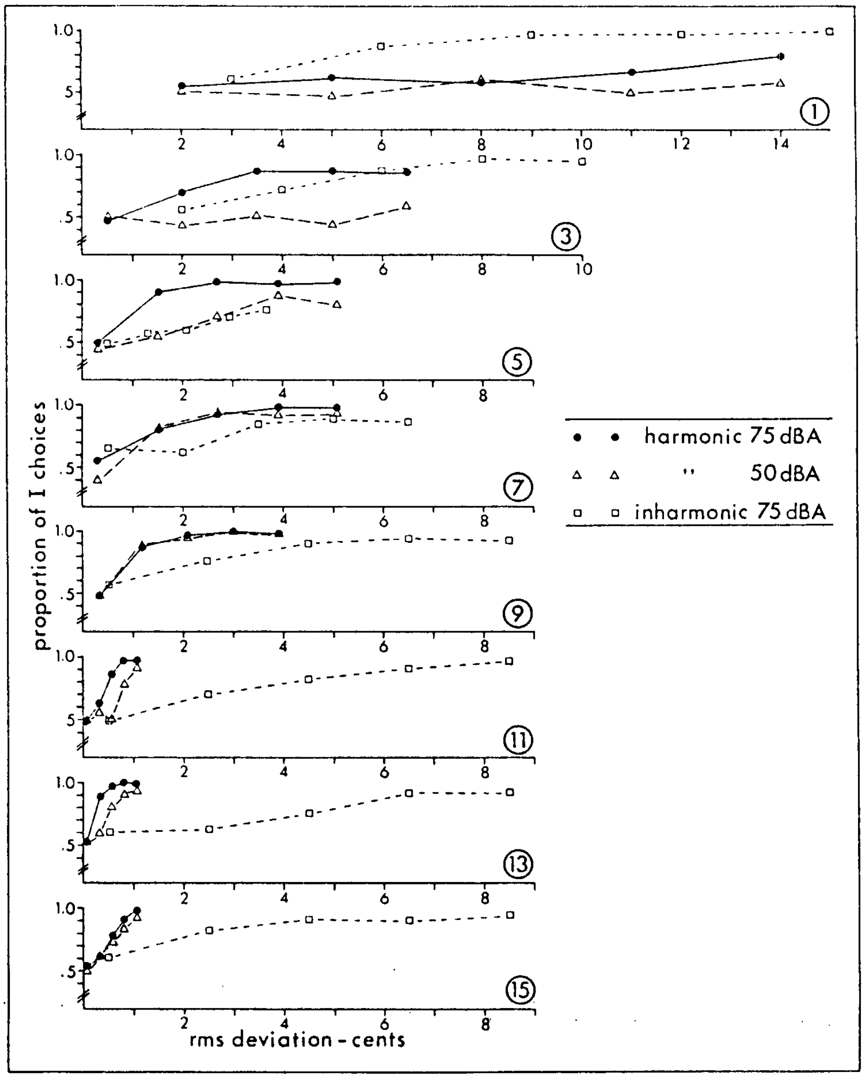

Figure 3.7. Mean data for Experiment 6 averaged across 4 subjects. For partials 7 – 15 of the I75 condition, the means are across 3 subjects (See caption for Fig. 3.3.)

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

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

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.

	Subject 1			**Subject 2**		
Incoherent Partial	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	–

	Subject 3			**Subject 4**		
Incoherent Partial	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

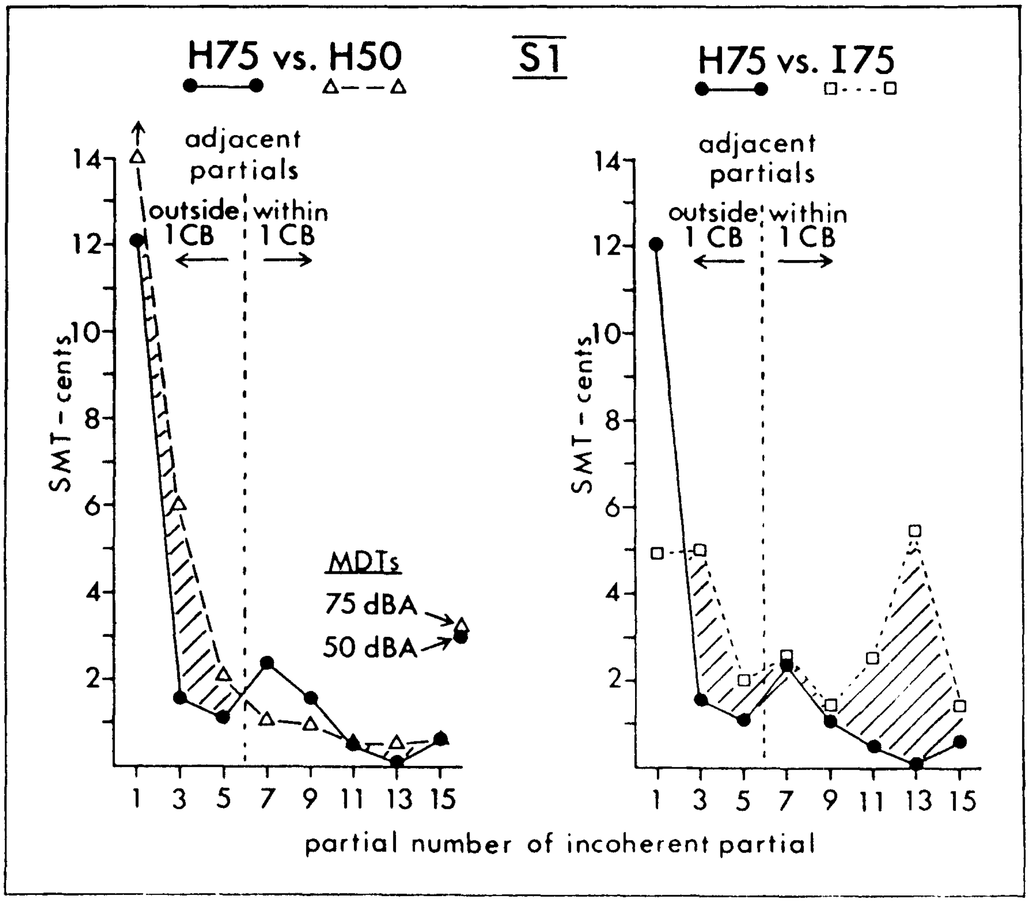

Figure 3.8. Source multiplicity thresholds (SMTs) for subject 1. The SMT (in cents rms deviation) is plotted as a function of the partial number of the frequency component receiving incoherent modulation. H75 vs. H50 is plotted on the left to see the effect on SMT of intensity difference. H75 vs. I75 is plotted on the right to see the effect of difference in harmonicity 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 than the subject never discerned an effect interpretable as the presence of multiple sources

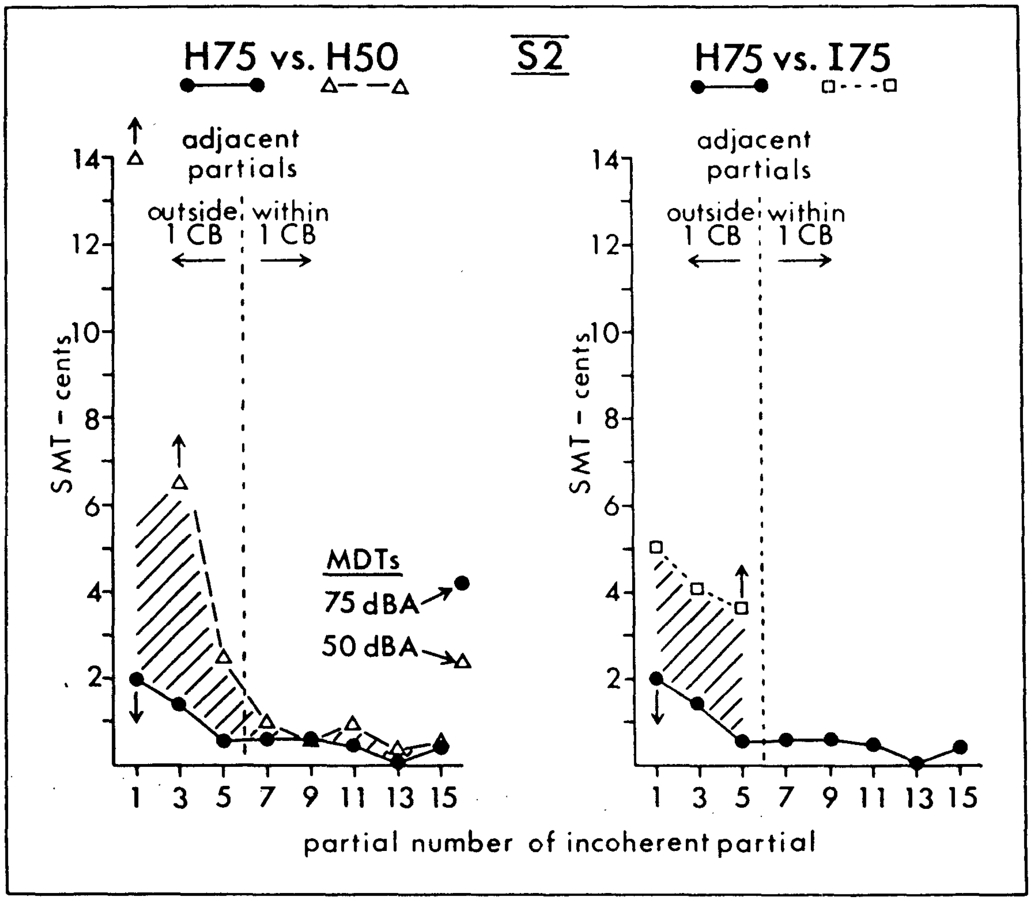

Figure 3.9. SMTs and complex tone MDTs for subject 2. (See caption for Fig. 3.8.)

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

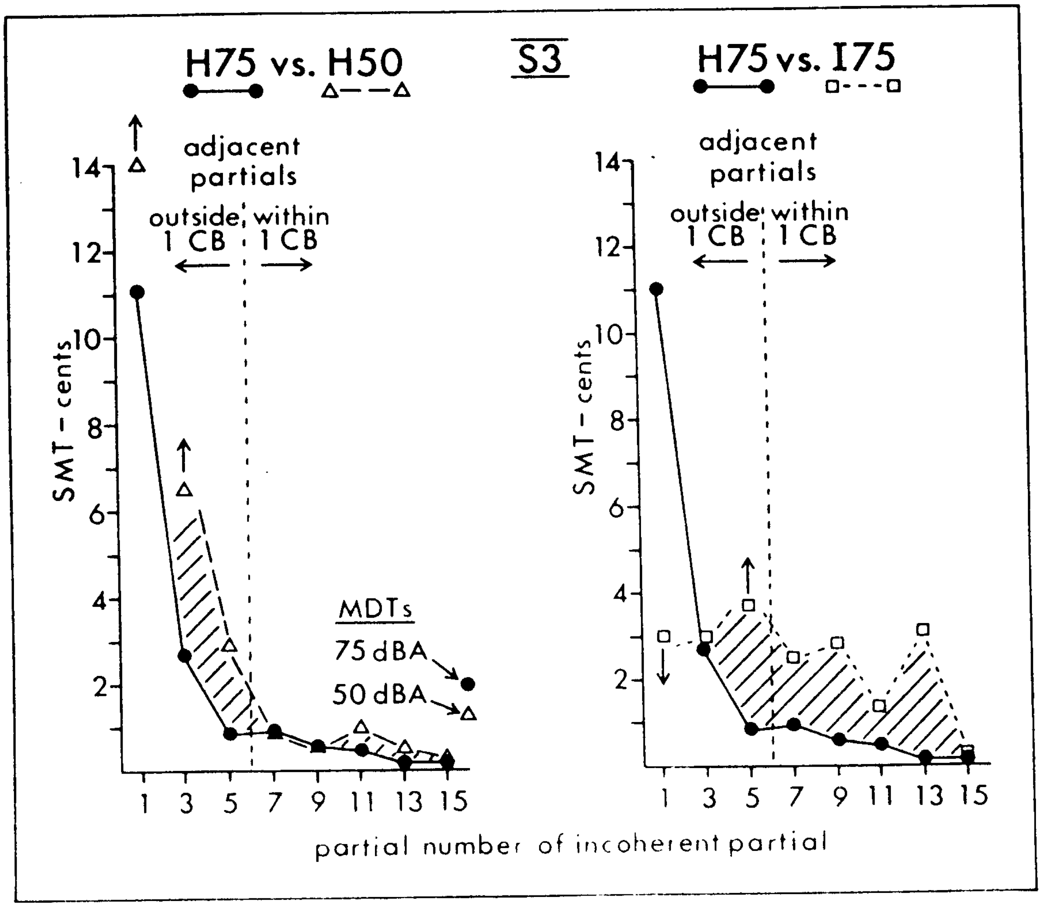

Figure 3.10. SMTs and complex tone MDTs for subject 3. (See caption for Fig. 3.8.)

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

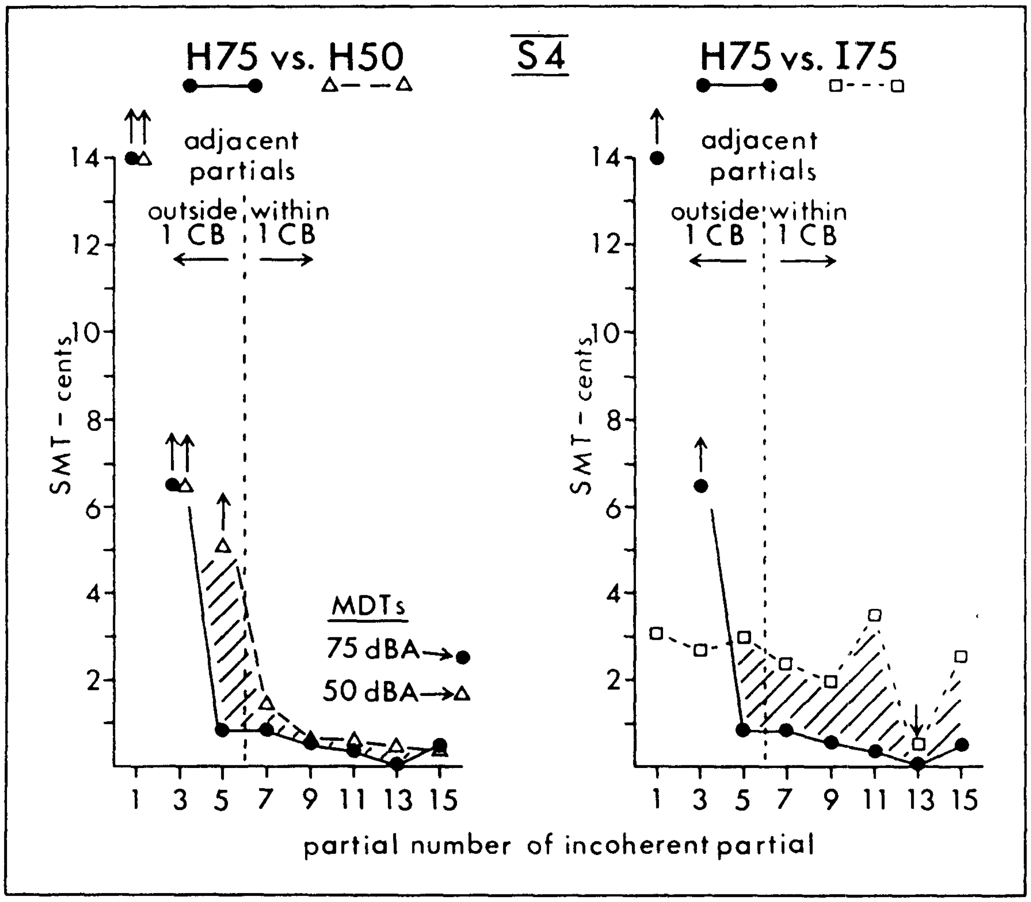

Figure 3.11. SMTs and complex tone MDTs for subject 4. (See caption for Fig. 3.8.)

[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 5th 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 cents_{rms}, $\Delta f_{rms}/\bar{f}$ and ΔP_{rms} (change in period of incoherent component that is just noticeable as yielding multiple sources, see Eq. 3.3).

	cents_{rms}			$\frac{\Delta f_{rms}}{\bar{f}} (\times 10^{-3})$			ΔP_{rms} (μsec)		
Incoherent Partial	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.

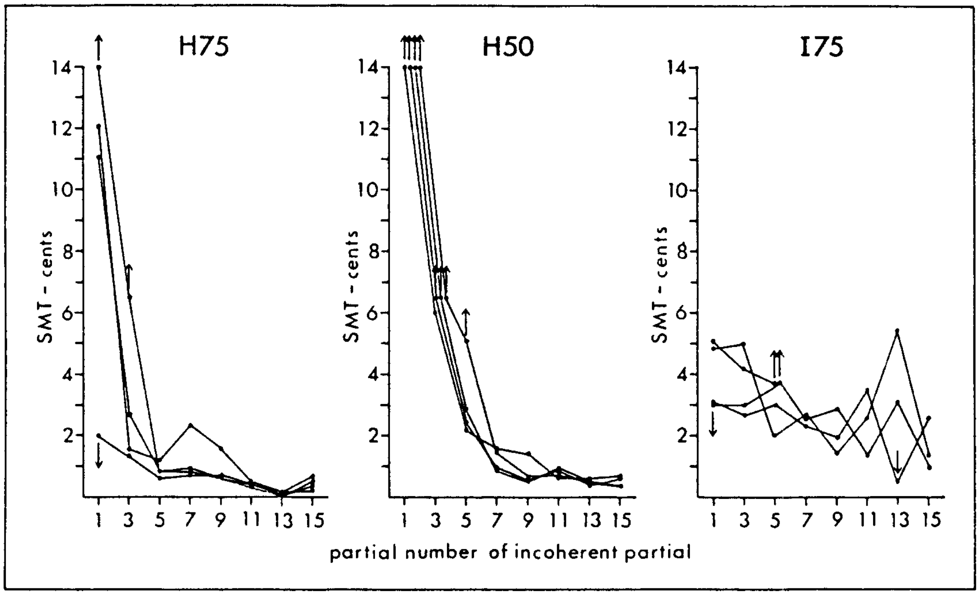

Figure 3.12. Individual SMTs for 4 subjects plotted as a function of partial number. Each graph represents a different experimental condition as labeled.

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 ($>f_6$) 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.[11] 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 SMT's 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 SMT's 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 6th, 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 6th 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.